НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-797-806
УДК 004.932.2:681.7:548.5

Применение машинного зрения для автоматического контроля процесса выращивания монокристаллов галогенидов таллия по методу Бриджмена–Стокбаргера

Максим Игоревич Кузьмин[1]✉, Максим Сергеевич Ельников[2], Давид Ильич Кушнирук[3], Максим Витальевич Морозов[4], Михаил Сергеевич Кузнецов[5]

[1,2,3,4,5] Государственный научно-исследовательский и проектный институт редкометаллической промышленности «Гиредмет» имени Н.П. Сажина, Москва, 111524, Российская Федерация
[2] Московский государственный технический университет им. Н.Э. Баумана, Москва, 105005, Российская Федерация

[1] mimikatz@mail.ru✉, https://orcid.org/0000-0001-6265-9052
[2] elnikow.max@gmail.com, https://orcid.org/0009-0001-8149-5302
[3] k.davjd@gmail.com, https://orcid.org/0009-0006-4419-6983
[4] MViMorozov@yandex.ru, https://orcid.org/0009-0007-1782-9623
[5] gradan@mail.ru, https://orcid.org/0000-0002-8441-4424

Аннотация

Введение. Рассмотрена проблема управления процессом выращивания монокристаллов галогенидов таллия методом Бриджмена–Стокбаргера. Определена важность обеспечения робастного управления температурным градиентом в зоне кристаллизации, оказывающего прямое влияние на качество получаемого монокристалла. Предложено и научно обосновано применение методов машинного зрения для определения положения границы расплав–кристалл и последующего автоматического управления температурным режимом выращивания. **Метод.** Для автоматизированного управления температурным градиентом предлагается использовать алгоритм, основанный на визуальном отслеживании положения кристаллизующейся границы (фронта). Распознавание фронта осуществляется посредством применения инструментов машинного зрения, позволяющих производить расчет корректирующего управляющего воздействия на верхнюю зону нагрева установки. **Основные результаты.** Представлено описание ключевых шагов алгоритма, приведена его блок-схема. На примере одной итерации производственного цикла проанализирована во времени динамика изменения высоты границы расплав–кристалл и температуры верхней печи. Соответствие полученного на опытной установке продукта принятым техническим условиям подтверждает эффективность предлагаемого подхода в стабилизации температурного профиля. **Обсуждение.** Разработанный алгоритм позволяет отказаться от ручного регулирования параметров на каждой установке и обеспечивает возможности для горизонтального масштабирования производства. Подход демонстрирует преимущества по сравнению с традиционными методами управления в контексте повышения повторяемости и качества выращиваемых монокристаллов. Предложенный алгоритм может быть использован при проектировании и модернизации установок, работающих по методу Бриджмена–Стокбаргера. Основным ограничением предлагаемого подхода является его применимость только к процессам, в которых осуществляется выращивание монокристаллов, обладающих характерной окраской.

Ключевые слова

машинное зрение, автоматическое управление, галогениды таллия, метод Бриджмена–Стокбаргера

Ссылка для цитирования: Кузьмин М.И., Ельников М.С., Кушнирук Д.И., Морозов М.В., Кузнецов М.С. Применение машинного зрения для автоматического контроля процесса выращивания монокристаллов галогенидов таллия по методу Бриджмена–Стокбаргера // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 797–806. doi: 10.17586/2226-1494-2025-25-5-797-806

Application of machine vision for automatic control of the process of growing monocrystals of thallium halides using the Bridgman-Stockbarger method

Maksim I. Kuzmin[1]✉, Maksim S. Elnikov[2], David I. Kushniruk[3], Maksim V. Morozov[4], Mikhail S. Kuznetsov[5]

[1,2,3,4,5] JSC N.P. Sazhin State Scientific Research and Design Institute of Rare Metal Industry "Giredmet", Moscow, 111524, Russian Federation

[2] Bauman Moscow State Technical University, Moscow, 105005, Russian Federation

[1] mimikatz@mail.ru✉, https://orcid.org/0000-0001-6265-9052

[2] elnikow.max@gmail.com, https://orcid.org/0009-0001-8149-5302

[3] k.davjd@gmail.com, https://orcid.org/0009-0006-4419-6983

[4] MViMorozov@yandex.ru, https://orcid.org/0009-0007-1782-9623

[5] gradan@mail.ru, https://orcid.org/0000-0002-8441-4424

Abstract

The article discusses the issue of controlling the growth process of monocrystals of thallium halides using the Bridgman-Stockbarger technique. The significance of maintaining a stable temperature gradient in the crystallization zone, which has a direct effect on the quality of the final monocrystal, is determined. The use of machine vision techniques to determine the position of the melt-crystal interface and subsequently automatic control of the temperature regime is proposed and scientifically justified. To control the temperature gradient automatically, it is suggested to utilize an algorithm that relies on visual tracking of the crystallization front. This front is identified using machine vision techniques, that allow calculating the corrective action on the upper heating zone of the apparatus. A brief overview of the main steps of the algorithm is provided, and a flowchart illustrating the process is included. Using the example of one iteration of the production cycle, the over time dynamics of changes in the height of the melt-crystal interface and the temperature of the upper furnace are analyzed. The compliance of the product obtained at the pilot apparatus with the accepted technical conditions confirms the effectiveness of the proposed approach in stabilizing the temperature profile. The developed algorithm eliminates manual parameter control at each apparatus, providing opportunities for horizontal scaling of production. It demonstrates advantages over traditional control methods, increasing the repeatability and quality of grown monocrystals. It can be used in the design and modernization of Bridgman-Stockbarger apparatuses. The main limitation of proposed approach is that it can only be applied to processes involving the growth of monocrystals with specific coloration.

Keywords

machine vision, automatic control, thallium halides, Bridgman-Stockbarger method

For citation: Kuzmin M.I., Elnikov M.S., Kushniruk D.I., Morozov M.V., Kuznetsov M.S. Application of machine vision for automatic control of the process of growing monocrystals of thallium halides using the Bridgman-Stockbarger method. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 797–806 (in Russian). doi: 10.17586/2226-1494-2025-25-5-797-806

Введение

Получение монокристаллов различных веществ и их смесей с заданными характеристиками является одной из ключевых задач материаловедения и кристаллохимии. Повышенный интерес к данному вопросу объясняется широким применением кристаллов с высокой структурной совершенностью вследствие проявления ими уникальных свойств, актуальных для использования в ряде высокотехнологических областей, претерпевающих интенсивное развитие в последние десятилетия: микроэлектронике [1], оптике [2, 3], медицине [4, 5], ядерной физике и многих других [6–8].

В настоящей работе в качестве исходного вещества для проведения процесса выращивания монокристаллов рассматриваются различные галогениды таллия (TlCl, TlBr, TlI) (рис. 1, *a*) как в индивидуальном виде, так и виде смеси. В зависимости от состава исходной шихты и соотношения входящих в нее компонентов, после проведения процесса выращивания и последующей механической обработки [9], могут быть получены

Рис. 1. Общий вид ампулы, заполненной шихтой (*a*) и заготовки после механической обработки полученного в ходе процесса монокристалла (*b*)

Fig. 1. General view of the ampoule filled with mixture (*a*) and of the workpiece after machining of the monocrystal obtained during the process (*b*)

заготовки (рис. 1, *b*), обладающие свойствами, специфическими для дальнейшего производства изделий в области инфракрасной техники, лазерной техники, акустооптики, волоконной оптики, приборов регистрации ионизирующего излучения.

На сегодняшний день разработано и опробовано в промышленности множество методов выращивания монокристаллов. Все они главным образом подразделяются на группы в зависимости от фазы (расплав, раствор, газ), из которой осуществляется кристаллизация целевого вещества. Выбор конкретного метода мотивируется физической и химической природой соединения, технико-экономическим расчетом, требованиями к структуре получаемого монокристалла. Применительно к рассматриваемым соединениям таллия, ввиду их плохой растворимости в воде и органических растворителях [10–14], процесс кристаллизации наиболее рационально проводить из фазы расплава.

Методы кристаллизации из расплава различаются между собой по ряду аспектов: объему кристаллизуемого материала, способу ведения процесса, сложности аппаратурного оформления. Одним из наиболее эффективных и широко используемых из них является метод Бриджмена–Стокбаргера [15]. Лежащая в основе выбранного метода идея заключается в направленной кристаллизации расплава. Конструктив установки (рис. 2), реализующей данный метод, включает в себя две высокотемпературные печи, разделенные термоизолирующим материалом для обеспечения градиента температур и создания двух функциональных зон, механизма перемещения ампулы и контура датчиков для мониторинга параметров течения процесса, а также управляющих устройств. В начале процесса ампулу заполняют шихтой заданного состава и запаивают таким образом, чтобы один из ее концов имел форму остроконечного конуса. Данная операция позволяет обеспечить образование затравочных кристаллов в нижней части сосуда. Подготовленную ампулу закрепляют на подвижном подвесе в зоне верхней печи и проводят первый этап процесса — плавление шихты. После полного расплавления содержимого активируется подъемно-опускной механизм, который с установленной постоянной скоростью обеспечивает движение ампулы вниз через зону температурного градиента. Перепад температуры ниже точки плавления шихты способствует началу процесса зародышеобразования, а затем, в соответствии с законом геометрического отбора, полноценному росту одного монокристалла. Постепенное снижение температуры приводит к упорядоченному росту кристаллической структуры по высоте ампулы, а скорость перемещения и профиль температурного градиента определяют качество получаемого изделия [16, 17]. После полного прохождения ампулы в зону нижней печи производится последний этап процесса — конечный отжиг. Для этого температурным регулятором осуществляется постепенное понижение температуры в зоне нижней печи до комнатных значений в соответствии с задаваемой температурной программой. Данный процесс позволяет избежать резкого охлаждения полученного монокристалла и снизить образование внутренних напряжений.

Основными достоинствами рассматриваемого метода являются простота конструкции установки и экономичность. С другой стороны, незначительные колебания температуры, нестабильность градиента или отклонения в скорости движения ампулы могут приводить к образованию дефектов, внутренних напряжений, неоднородности состава или росту поликристаллических структур вместо монокристаллических. Все вышеперечисленные негативные факторы способствуют растрескиванию кристалла во время процесса выращивания или при извлечении из ампулы. В соответствии с этим, главным недостатком метода является большая чувствительность к ошибкам управления. Другая немаловажная практическая проблема

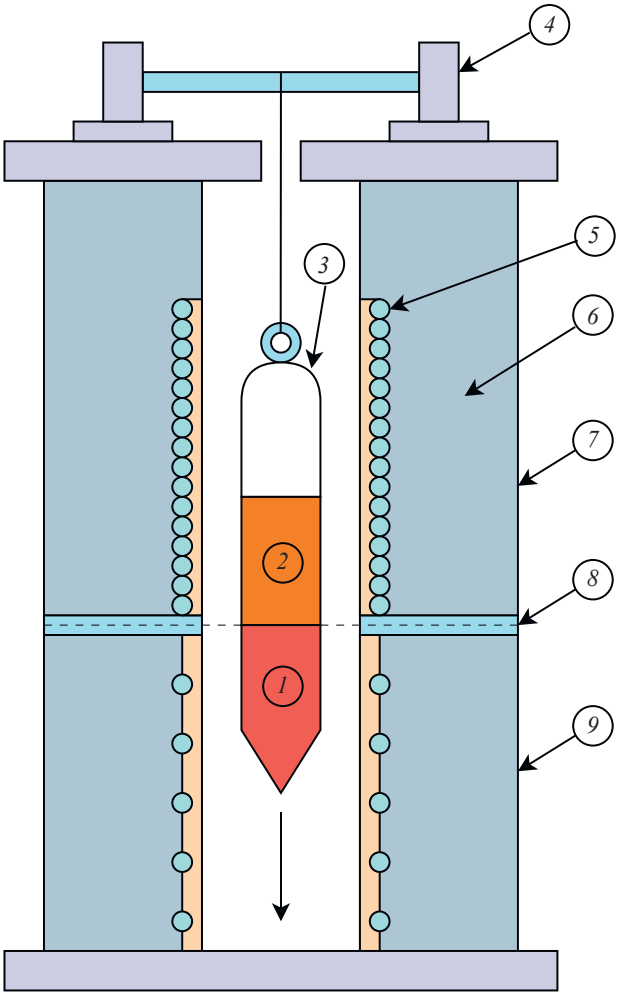

Рис. 2. Схема установки выращивания кристаллов методом Бриджмена–Стокбаргера.

1 — монокристалл; *2* — расплав; *3* — ампула на подвижном подвесе; *4* — привод с подъемно-опускным механизмом; *5* — элементы резистивного нагревателя; *6* — огнеупорный материал; *7* — верхняя печь; *8* — термоизоляция; *9* — нижняя печь

Fig. 2. Scheme of the apparatus for crystal growing by the Bridgman-Stockbarger method.

1 — monocrystal; *2* — melt; *3* — ampoule on a movable suspension; *4* — drive with lifting and lowering mechanism; *5* — elements of a resistive heater; *6* — refractory material; *7* — upper furnace; *8* — thermal insulation; *9* — lower furnace

с учетом недостатков — ручной контроль выращивания кристаллов в установках, число которых измеряется десятками. В каждой из таких установок может выращиваться монокристалл, имеющий отличный от других исходный состав шихты и потому требующий обеспечения индивидуального температурного профиля для оптимального протекания процесса. По этой причине одновременный ручной контроль выращивания во всех установках является затруднительной задачей. Вследствие этого особую актуальность приобретает автоматизация процесса выращивания кристаллов, позволяющая минимизировать влияние человеческого фактора и обеспечить стабильность условий кристаллизации.

Целью работы является программная реализация настраиваемого алгоритма автоматизации процесса выращивания монокристалла с использованием средств машинного зрения и апробация его работы.

Задача автоматизации процесса

Попытки оптимального управления рассматриваемым процессом предпринимались исследователями ранее. Для этого вносились изменения в конструкцию установки [18] и нагревателей [19], изучалось влияние величины температурного градиента на форму границы расплав–кристалл [20], а также проверялась возможность отслеживания положения этой границы с помощью вихретокового контроля [21].

Глобальной задачей автоматизации процесса является обеспечение постоянной оптимальной скорости роста монокристалла в области перепада температур (*8*, рис. 2). В настоящей работе, исходя из необходимости визуального наблюдения границы расплав–кристалл и недопущения стихийного протекания процесса, фронт кристаллизации поддерживался несколько выше этой области. Расстояние от начала зоны термоизоляции до визуально наблюдаемой границы расплав–кристалл в ампуле, находящейся в верхней печи, называется высотой фронта кристаллизации. Она является регулируемым параметром и, как правило, ее целевое значение устанавливается в интервале 2–5 см. Критерий оптимальной скорости роста монокристалла предполагает такое ее значение, которое позволяет фронту кристаллизации находиться в пределах установленного целевого значения высоты при постоянной скорости движения ампулы вниз.

Так как в установке имеется две функциональные зоны нижней и верхней печи, то возможны два варианта организации управления процессом.

Вариант 1. Регулировка изменением нагрева верхней печи, при этом в нижней поддерживается постоянная температура немного ниже точки кристаллизации расплава.

Вариант 2. Регулировка одновременным изменением нагрева обоих печей.

В настоящей работе сделан акцент на программной реализации алгоритма для варианта 1. В соответствии с ним управляемым параметром является значение мощности нагревателя верхней печи.

Используемые соединения таллия и их смеси (за исключением чистого хлорпроизводного) имеют естественную окраску. В дополнение к этому внутренняя структура расплава в отличии от монокристалла является неупорядоченной, вследствие чего при сквозном просвечивании прожектором он имеет вид темной области. Совокупность перечисленных факторов позволяет разработать метод определения высоты фронта используя методы машинного зрения. Для программной реализации и отладки алгоритма опытная установка стандартной конструкции (рис. 2) была модернизирована и дополнена (рис. 3) необходимым оборудованием.

Сзади установки на целевой высоте размещалась осветительная система в виде светодиодного прожектора AECLight АЭК-ДСП44-020-001 со световым потоком 2800 лм, а спереди на том же уровне устанавливалась IP-камера Beward SV2018M с функцией автофокусировки и расширенным динамическим диапазоном (рис. 4). В качестве температурного регулятора для каждой печи использовался ОВЕН ТРМ251, ко входу которого подключалась термопара типа «К». Выход температурного регулятора подсоединялся к твердотельному реле, управляющему мощностью, подаваемой на нагревательные элементы соответствующей печи. Спай термопары выводился во внутреннее пространство, по которому осуществлялось движение ампулы. Управляющий компьютер технолога подключался к интерфейсу RS-485 температурного регулятора верхней печи посредством витой пары. Связь с прибором и смена температурной уставки осуществлялась с помощью протокола Modbus RTU. Единоразовое задание фиксированной уставки в температурном регуляторе нижней печи производилось вручную через встроенный в прибор задатчик в начале процесса.

В теплоизолирующей стенке опытной установки организовывался вертикальный паз минимальной ширины, достаточной для наблюдения содержимого. В ходе проведения процесса выращивания монокристалла из непрерывного видеопотока с установленным интервалом захватывался единичный кадр и отправлялся на вход реализованному алгоритму. Интервальный захват единичного кадра объясняется временем реакции системы (30–60 мин) на управляющее воздействие и инертностью рассматриваемой системы в целом. Слишком большой интервал срабатывания алгоритма и, соответственно, регулирования приводит к неконтролируемому уходу фронта, слишком малый — к нарушению процесса, так как система реагирует на корректирующее воздействие с запаздыванием. Время запаздывания системы определялось экспериментально заданием вручную небольшого ступенчатого воздействия путем смены температурной уставки и отслеживания начала изменения высоты фронта от исходного положения.

Выходным результатом работы алгоритма является корректировка уставки температуры в температурном регуляторе верхней печи в ответ на визуально отслеживаемое изменение положения фронта кристаллизации. Правильность работы алгоритма характеризуется корректным определением высоты пограничной области между расплавом и монокристаллом. Стабильность работы алгоритма основывается на получении сопоставимых по величине значений положения границы при

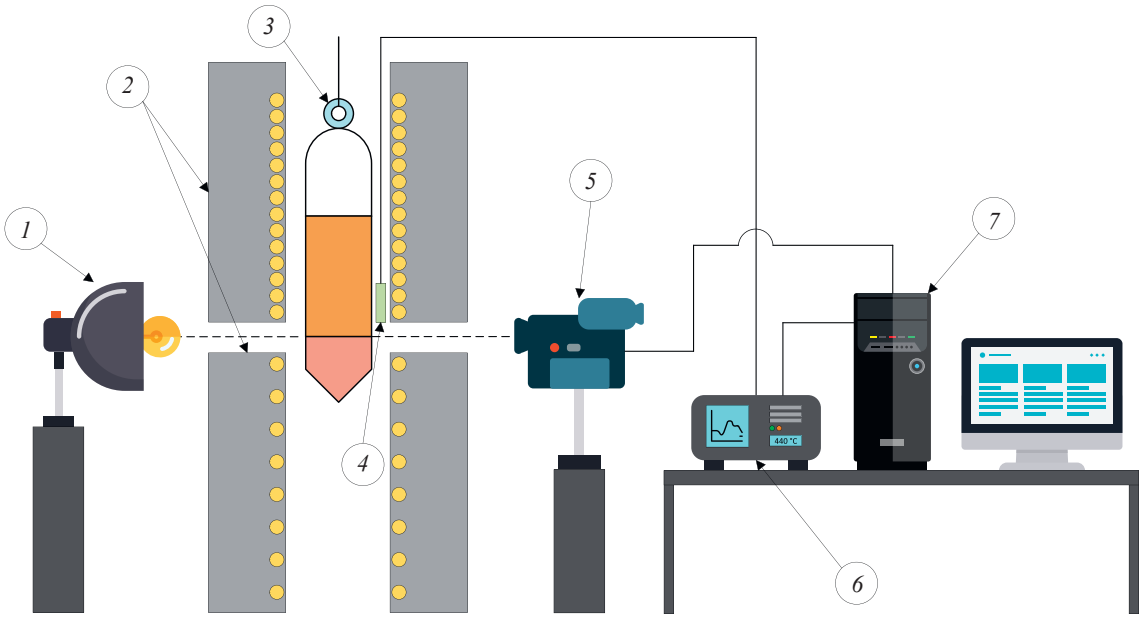

Рис. 3. Схема модернизированной и дополненной установки.

1 — светодиодный прожектор; *2* — верхняя и нижняя печь; *3* — ампула на подвижном подвесе; *4* — термопара типа «К»; *5* — IP-камера; *6* — терморегулятор; *7* — компьютер

Fig. 3. Scheme of the upgraded and augmented apparatus.

1 — light-emitting diode spotlight; *2* — upper and lower furnace; *3* — ampoule on a movable suspension; *4* — type “K” thermocouple; *5* — IP camera; *6* — thermoregulator; *7* — computer

неизменной реальной высоте фронта кристаллизации в ходе нескольких последовательных замеров в условиях воздействия негативных факторов. Среди подобных факторов в настоящей работе выступали: изменение условий внешнего освещения ввиду круглосуточного проведения производственного процесса; блики света, дефекты и загрязнения на поверхности кварцевой стенки установки.

Тестовое применение алгоритма показало, что определяемая высота фронта в некоторых случаях может варьироваться в пределах 1–2 мм, в связи с чем был введен в качестве одного из его внутренних параметров показатель буферной границы, позволяющий нивелировать подобную неточность определения и избежать ложных срабатываний операции регулирования. Последующие испытания алгоритма проводились в течение полного цикла выращивания, поэтому точность выходных регулирующих воздействий напрямую влияла на качество получаемого продукта. Вследствие этого дополнительным критерием робастности алгоритма являлось соответствие полученного кристалла установленным производственным требованиям.

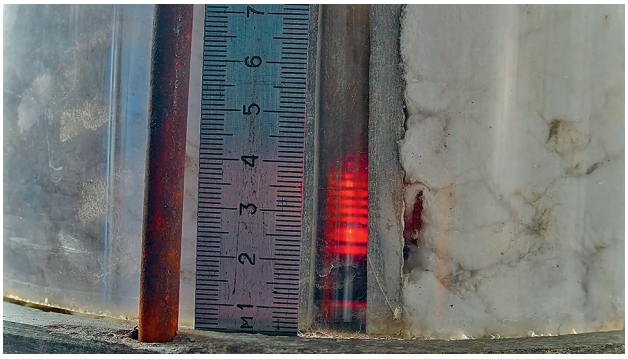

Рис. 4. Вид на установку с IP-камеры

Fig. 4. View of the apparatus from the IP camera

Структура алгоритма и особенности реализации

Программная реализация и отладка алгоритма осуществлялась на языке программирования Python с применением библиотеки компьютерного зрения OpenCV. Структурно алгоритм подразделяется на две последовательно идущие части (рис. 5).

Блок 1. Блок определения высоты фронта.

Блок 2. Блок расчета регулирующего воздействия.

В самом начале работы алгоритма кадр, захваченный из видеопотока, поступает на вход блока 1, где осуществляется этап его предварительной обработки. Он заключается в повороте и обрезке полнокадрового изображения (рис. 4) до области интереса, представляющей собой зону технологического окна. К полученному изображению применяется размытие по Гауссу с ядром (3, 3). Размер ядра подбирался вручную таким образом, чтобы добиться легкого сглаживания без существенных искажений. Стандартное отклонение по Гауссу рассчитывалось автоматически на основании размера ядра. Цель предварительной обработки — нивелировать возможные шумы и подготовить изображение к последующим операциям.

Следующим этапом блока 1 алгоритма является фильтрация полезного сигнала. Основная задача состоит в отделении и усилении области изображения, относящейся непосредственно к монокристаллу. Как

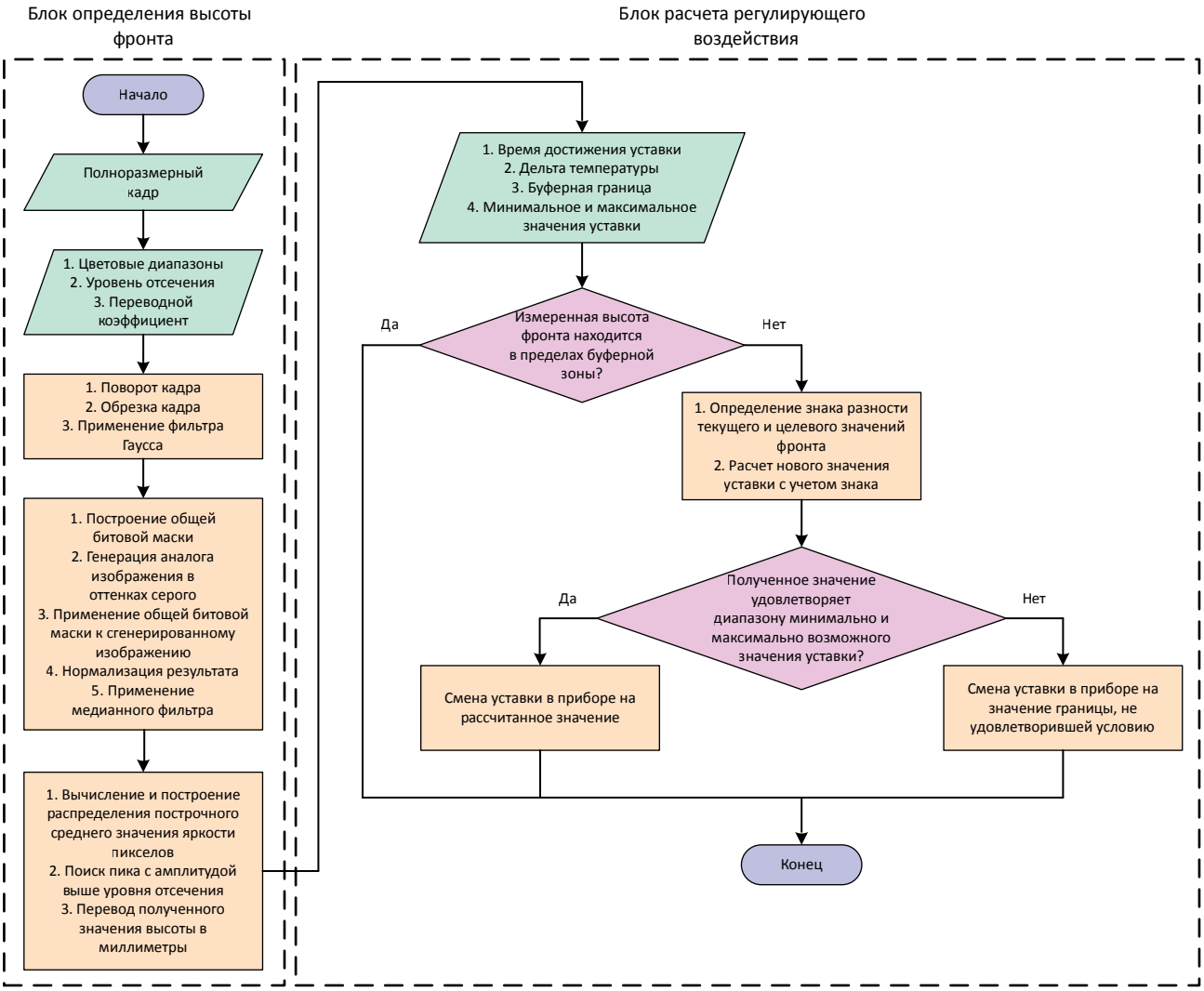

Рис. 5. Блок-схема алгоритма

Fig. 5. The block diagram of the algorithm

описывалось в разделе «Задача автоматизации процесса», в зависимости от используемых веществ и состава шихты цвет соответствующего монокристалла может варьироваться. В соответствии с рис. 4 монокристалл имеет красно-оранжевый цвет. В цветовой модели HSV (Hue, Saturation, Value — тон, насыщенность, значение) интересующие цвета лежат в двух прямоугольных областях: (0, 10) — (20, 255) и (165, 10) — (180, 255) (рис. 6). Вследствие этого для качественного отделения полезного сигнала потребуется использовать две маски, объединенные в одну.

Таким образом, этап фильтрации включает в себя следующую последовательность операций:

— построение двух битовых масок на основе выбранных диапазонов от (0, 10, 10) до (20, 255, 255) и от (165, 10, 10) до (180, 255, 255) с помощью бинаризации;
— построение общей битовой маски на основе двух отдельных путем их логического сложения;
— генерация аналога исходного изображения в оттенках серого. Для этого определялась разность цветов пикселов красного канала и максимума из двух других. Полученные значения нормировались в диапазоне 0–255;
— применение общей битовой маски, сгенерированной бинаризацией цветного изображения, к полученному изображению в оттенках серого посредством логического умножения;
— нормализация результирующего изображения;
— применение медианной фильтрации с окном равным 9, для нивелирования возможного шума, созданного наложением битовой маски. Значение окна фильтрации подбиралось экспериментально аналогично ядру для размытия по Гауссу.

Завершающим этапом работы блока 1 алгоритма является непосредственное определение высоты границы расплав–кристалл на основе обработанного изображения. Для этого осуществляется вычисление построчного среднего значения яркости пикселов на изображении в оттенках серого, что позволяет построить характерный график распределения (рис. 7).

Определение высоты границы расплав–кристалл осуществляется путем поиска первого единичного пика, амплитуда которого больше некоторого уровня

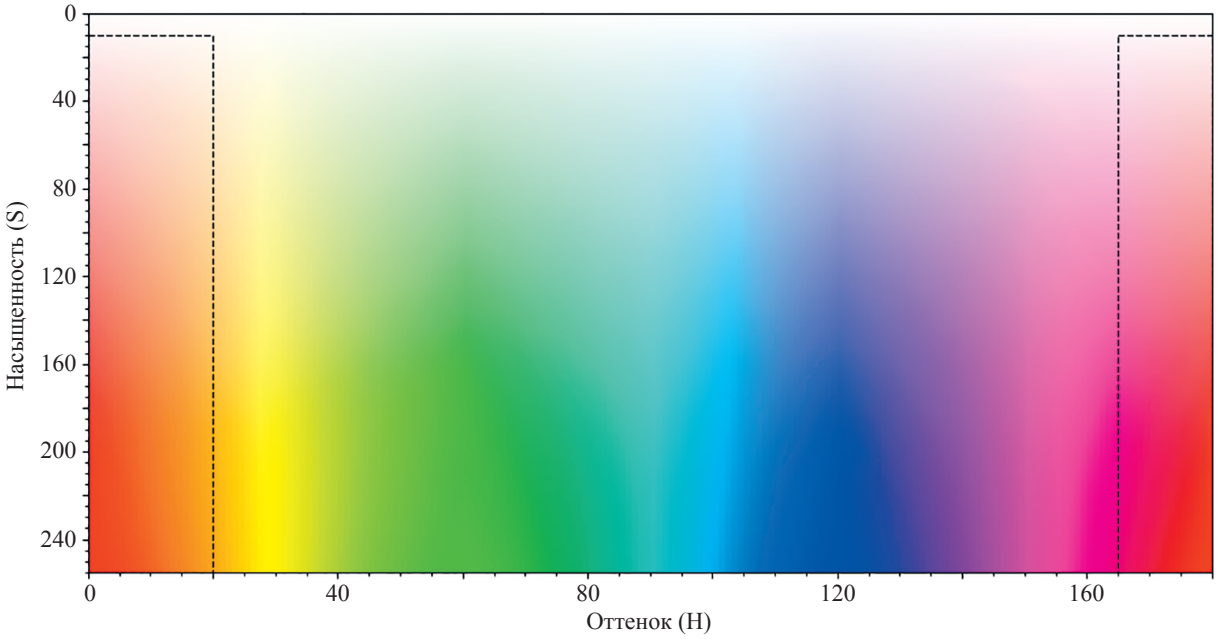

Рис. 6. Представление цветовой модели HSV в двух измерениях при фиксированном значении яркости равном 255

Fig. 6. Representation of the HSV color model in two dimensions with a fixed brightness value of 255

отсечения. Достоверное и математически обоснованное значение уровня отсечения может быть получено путем статистической обработки амплитуд пиков, извлеченных из массива кадров, снятых при данных условиях. Результирующая высота фронта кристаллизации определяется путем деления полученного значения высоты в пикселах на переводной коэффициент и подается на вход блоку 2. Для определения переводного коэффициента рядом с технологическим окном во время тестового процесса выращивания устанавливается линейка (рис. 4). По линейке оценивается высота фронта кристаллизации относительно зоны термоизоляции. Делается снимок и обрабатывается алгоритмом до момента осуществления перевода значения. Полученное алгоритмом значение высоты в пикселах делится на наблюдаемое значение в миллиметрах. Полученный коэффициент (в пикс./мм) используется в дальнейшем для перевода значений.

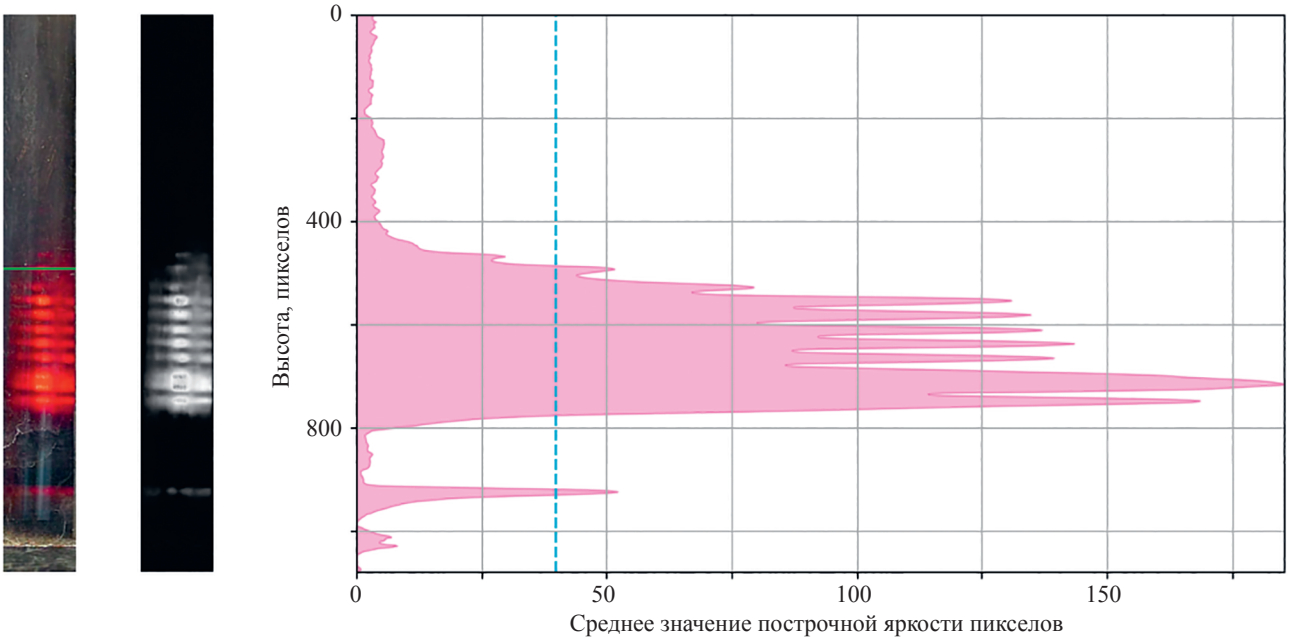

Рис. 7. График зависимости среднего значения построчной яркости пикселов на изображении в оттенках серого. Синим пунктиром показан уровень отсечения, зеленой линией — результирующее положение фронта кристаллизации

Fig. 7. Graph of the dependence of the average line-by-line brightness of pixels in an image in shades of gray. The blue dotted line shows the clipping level. The green line shows the resulting position of the crystallization front

Блок 2 принимает на вход измеренную текущую высоту фронта кристаллизации и сравнивает ее с целевым значением с учетом ряда параметров, задаваемых технологом. К внутренним параметрам блока 2 относятся: время достижения уставки терморегулятором, дельта температуры, минимально и максимально допустимые значения уставки, буферная граница. Если текущее значение высоты фронта кристаллизации располагается в диапазоне буферной зоны, рассчитываемой на основе значений целевой высоты и буферной границы, то смену уставки температуры производить не требуется. Температурный регулятор продолжает поддерживать текущее значение. В случае если измеренная высота границы находится за пределами — происходит расчет новой уставки. Для этого к ее текущему значению прибавляется с соответствующим знаком параметр дельты температуры. Если полученное новое значение попадает в заданный интервал минимально и максимально допустимых уставок температур, то оно отправляется в температурный регулятор верхней печи без изменения. В противном случае в прибор отправляется значение той границы, которой не удовлетворило рассчитанное значение. Если в результате регулирования высота фронта кристаллизации больше целевой — температура повышается, в противном случае снижается.

Обсуждение

Результат работы алгоритма в течение единичной производственной итерации показан на рис. 8. Низкая температура в верхней печи в начале процесса обуславливает высокое положение границы между расплавом и кристаллом относительно зоны термоизоляции (*8*, рис. 2). Ступенчатые и линейные изменения температуры являются следствием смены соответствующих целевых уставок и работы температурного регулятора. Увеличение мощности, подаваемой на нагревательные элементы (*5*, рис. 2) верхней печи приводит к усилению плавления и смещению фронта кристаллизации в меньшую сторону по высоте.

Главной практической ценностью разработанного алгоритма является возможность полного исключения потребности в ручном контроле процесса. Это в свою очередь позволяет горизонтально масштабировать производство, значительно повысить количество и ассортимент выпускаемой продукции, а также увеличить положительный экономический эффект за счет снижения себестоимости. Установки для выращивания кристаллов по методу Бриджмена–Стокбаргера могут проектироваться и создаваться в модернизированном варианте с учетом аспектов работы алгоритма, предложенного в настоящей работе. Сам алгоритм в таком случае может быть реализован в адаптированном варианте с привлечением желаемого инструментария и языка программирования. Среди дополнительных особенностей его работы можно выделить возможность использования профилей, содержащих массив внутренних параметров, под индивидуальный состав шихты и условия окружающей среды.

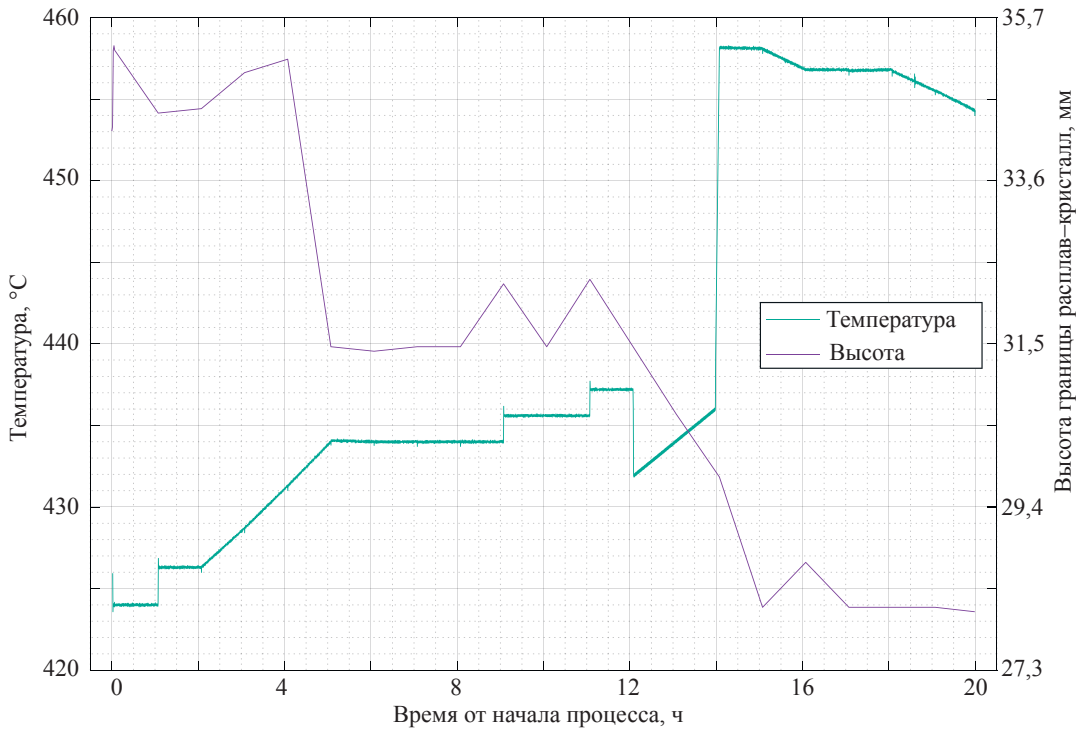

Рис. 8. Изменение температуры и высоты границы расплав–кристалл на протяжении единичной производственной итерации процесса

Fig. 8. Change in temperature and height of the melt-crystal boundary during a single production iteration of the process

Заключение

В работе рассмотрена задача автоматизации процесса выращивания монокристаллов галогенидов таллия с применением методов машинного зрения. С использованием языка программирования Python и библиотеки OpenCV реализован и отлажен алгоритм, осуществляющий определение высоты границы расплав–кристалл и последующее регулирование процесса. Работоспособность алгоритма проверена в рамках единичных итераций производственного процесса.

В дальнейшем планируется разработка системы управления множеством параллельных процессов выращивания монокристаллов различного состава с использованием полученных наработок и практического опыта.

Литература

1. Lang W. Silicon microstructuring technology // Materials Science and Engineering: R: Reports. 1996. V. 17. N 1. P. 1–55. https://doi.org/10.1016/0927-796X(96)00190-8
2. Kaplunov I.A., Kolesnikov A.I., Gavalyan M.Y., Belotserkovskiy A.V. Optical properties of large germanium monocrystals // Optics and Spectroscopy. 2016. V. 120. N 4. P. 654–659. https://doi.org/10.1134/S0030400X16030139
3. Isaenko L., Yelisseyev A., Tkachuk A., Ivanova S. New monocrystals with low phonon energy for mid-IR lasers // NATO Science for Peace and Security Series B: Physics and Biophysics. 2008. P. 3–65. https://doi.org/10.1007/978-1-4020-6463-0_1
4. Taubin M.L., Yaskolko A.A. Improvement of medical X-ray tube performance // Biomedical Engineering. 2010. V. 44. N 2. P. 73–75. https://doi.org/10.1007/s10527-010-9159-8
5. Ababiy I., Aramă E. Advantages of applications UV detectors based on stratified crystals in medicine // Proc. of the Professional Education and Economic Needs of the Black Sea Region. 2015. P. 127–133.
6. Ganesh V., Shkir M., Maurya K.K., Yahia I.S., AlFaify S. Phenol red dyed bis thiourea cadmium acetate monocrystal growth and characterization for optoelectronic applications // Journal of Materials Research. 2018. V. 33. N 16. P. 2364–2375. https://doi.org/10.1557/jmr.2018.235
7. Wischnewski M., Delibas B., Wischnewski A., Pertsch P. Microscale monocrystal ultrasonic actuators for miniature optical systems // Proc. of the International Conference and Exhibition on New Actuator Systems and Applications. 2022. P. 1–4.
8. Wang D., Chen J.S. Progress on the applications of piezoelectric materials in sensors // Materials Science Forum. 2016. V. 848. P. 749–756. https://doi.org/10.4028/www.scientific.net/MSF.848.749
9. Лапшин В.В., Захаревич Е.М., Кузнецов М.С., Зараменских К.С., Осипов А.В. Технология обработки оптических деталей из кристаллов КРС-5 методом алмазного точения и фрезерования // Фотоника. 2021. Т. 15. № 1. С. 18–28. https://doi.org/10.22184/1993-7296.FRos.2021.15.1.18.28
10. Жукова Л.В., Китаев Г.А., Козлов Ф.Н. Растворимость TlBr, TlI и их твердых растворов в воде // Журнал физической химии. 1978. Т. 52. № 7. С. 1692–1695.
11. Китаев Г.А., Жукова Л.В., Козлов Ф.Н. Растворимость галогенидов таллия(I) и их твёрдых изоморфных смесей в полярных растворителях // Журнал физической химии. 1980. Т. 54. № 8. С. 2032–2036.
12. Жукова Л.В., Китаев Г.А., Козлов Ф.Н. Растворимость галогенидов одновалентного таллия в воде и неводных растворителях. Справочник по продуктам растворимости. Новосибирск: Наука, 1983. 191 с.
13. Козлов Ф.Н., Китаев Г.А., Жукова Л.В. Растворимость и кристаллизация галогенидов таллия(I) из водных растворов // Журнал неорганический химии. 1983. Т. 28. № 2. С. 482–486.
14. Haynes W.M. CRC Handbook of Chemistry and Physics. CRC Press, 2016. 2670 p.
15. Bridgman P.W. Crystals and their manufacture. Patent US1793672A. 1931.
16. Ерохин С.В., Зараменских К.С., Кузнецов М.С., Пилюшко С.М. Оптимизация процесса роста монокристалла КРС-5 с помощью расчета градиента температуры методом конечных элементов // Тонкие химические технологии. 2025. Т. 20. № 1. С. 55–62. https://doi.org/10.32362/2410-6593-2025-20-1-55-62
17. Potts H., Wilcox W.R. Thermal fields in the Bridgman–Stockbarger technique // Journal of Crystal Growth. 1985. V. 73. N 2. P. 350–358. https://doi.org/10.1016/0022-0248(85)90312-4

References

1. Lang W. Silicon microstructuring technology. *Materials Science and Engineering: R: Reports*, 1996, vol. 17, no. 1, pp. 1–55. https://doi.org/10.1016/0927-796X(96)00190-8
2. Kaplunov I.A., Kolesnikov A.I., Gavalyan M.Y., Belotserkovskiy A.V. Optical properties of large germanium monocrystals. *Optics and Spectroscopy*, 2016, vol. 120, no. 4, pp. 654–659. https://doi.org/10.1134/S0030400X16030139
3. Isaenko L., Yelisseyev A., Tkachuk A., Ivanova S. New monocrystals with low phonon energy for mid-IR lasers. *NATO Science for Peace and Security Series B: Physics and Biophysics*, 2008, pp. 3–65. https://doi.org/10.1007/978-1-4020-6463-0_1
4. Taubin M.L., Yaskolko A.A. Improvement of medical X-ray tube performance. *Biomedical Engineering*, 2010, vol. 44, no. 2, pp. 73–75. https://doi.org/10.1007/s10527-010-9159-8
5. Ababiy I., Aramă E. Advantages of applications UV detectors based on stratified crystals in medicine. *Proc. of the Professional Education and Economic Needs of the Black Sea Region*, 2015, pp. 127–133.
6. Ganesh V., Shkir M., Maurya K.K., Yahia I.S., AlFaify S. Phenol red dyed bis thiourea cadmium acetate monocrystal growth and characterization for optoelectronic applications. *Journal of Materials Research*, 2018, vol. 33, no. 16, pp. 2364–2375. https://doi.org/10.1557/jmr.2018.235
7. Wischnewski M., Delibas B., Wischnewski A., Pertsch P. Microscale monocrystal ultrasonic actuators for miniature optical systems. *Proc. of the International Conference and Exhibition on New Actuator Systems and Applications*, 2022, pp. 1–4.
8. Wang D., Chen J.S. Progress on the applications of piezoelectric materials in sensors. *Materials Science Forum*, 2016, vol. 848, pp. 749–756. https://doi.org/10.4028/www.scientific.net/MSF.848.749
9. Lapshin V.V., Zakharevich E.M., Kuznetsov M.S., Zaramenskikh K.S., Osipov A.V. Technology of machining optical parts made of KRS-5 crystals by diamond turning and milling. *Photonics Russia*, 2021, vol. 15, no. 1, pp. 18–28. (in Russian). https://doi.org/10.22184/1993-7296.FRos.2021.15.1.18.28
10. Zhukova L.V., Kitaev G.A., Kozlov F.N. On solubility of TlBr, TlI and their solid solutions. *Zhurnal Fizicheskoj Himii*, 1978, vol. 52, no. 7, pp. 1692–1695. (in Russian)
11. Kitaev G.A., Zhukova L.V., Kozlov F.N. Solubility of thallium(I) halides and their solid isomorphic mixtures in polar solvents. *Zhurnal Fizicheskoj Himii*, 1980, vol. 54, no. 8, pp. 2032–2036. (in Russian)
12. Zhukova L.V., Kitaev G.A., Kozlov F.N. *Solubility of Monovalent Thallium Halides in Water and Non-Aqueous Solvents*. Handbook of Solubility Products. Novosibirsk, Nauka Publ., 1983, 191 p. (in Russian)
13. Kozlov F.N., Kitaev G.A., Zhukova L.V. Solubility and crystallization of thallium(i) halides from aqueous-solutions. *Zhurnal Neorganicheskoi Khimii*, 1983, vol. 28, no. 2, pp. 482–486. (in Russian)
14. Haynes W.M. *CRC Handbook of Chemistry and Physics*. CRC Press, 2016, 2670 p.
15. Bridgman P.W. Crystals and their manufacture. *Patent US1793672A*. 1931.
16. Erohin S.V., Zaramenskikh K.S., Kuznetsov M.S., Pilyushko S.M. Optimization of KRS-5 single crystal growth process by calculation of temperature gradient using finite element method. *Fine Chemical Technologies*, 2025, vol. 20, no. 1, pp. 55-62. (in Russian). https://doi.org/10.32362/2410-6593-2025-20-1-55-62
17. Potts H., Wilcox W.R. Thermal fields in the Bridgman–Stockbarger technique. *Journal of Crystal Growth*, 1985, vol. 73, no. 2, pp. 350–358. https://doi.org/10.1016/0022-0248(85)90312-4

18. Mouchovski J.T., Penev V.T., Kuneva R.B. Control of the growth optimum in producing high-quality CaF2 crystals by an improved Bridgman–Stockbarger technique // Crystal Research and Technology. 1996. V. 31. N 6. P. 727–737. https://doi.org/10.1002/crat.2170310603
19. Nicoară D., Nicoară I. An improved Bridgman–Stockbarger crystal-growth system // Materials Science and Engineering: A. 1988. V. 102. N 2. P. L1–L4. https://doi.org/10.1016/0025-5416(88)90584-8
20. Chang C.E., Wilcox W.R. Control of interface shape in the vertical Bridgman–Stockbarger technique // Journal of Crystal Growth. 1974. V. 21. N 1. P. 135–140. https://doi.org/10.1016/0022-0248(74)90161-4
21. Rosen G.J., Carlson F.M., Thompson J.E., Wilcox W.R., Wallace J.P. Monitoring vertical Bridgman–Stockbarger growth of cadmium telluride by an eddy current technique // Journal of Electronic Materials. 1995. V. 24. N 5. P. 491–495. https://doi.org/10.1007/bf02657952

18. Mouchovski J.T., Penev V.T., Kuneva R.B. Control of the growth optimum in producing high-quality CaF2 crystals by an improved Bridgman–Stockbarger technique. *Crystal Research and Technology*, 1996, vol. 31, no. 6, pp. 727–737. https://doi.org/10.1002/crat.2170310603
19. Nicoară D., Nicoară I. An improved Bridgman–Stockbarger crystal-growth system. *Materials Science and Engineering: A*, 1988, vol. 102, no. 2, pp. L1–L4. https://doi.org/10.1016/0025-5416(88)90584-8
20. Chang C.E., Wilcox W.R. Control of interface shape in the vertical Bridgman–Stockbarger technique. *Journal of Crystal Growth*, 1974, vol. 21, no. 1, pp. 135–140. https://doi.org/10.1016/0022-0248(74)90161-4
21. Rosen G.J., Carlson F.M., Thompson J.E., Wilcox W.R., Wallace J.P. Monitoring vertical Bridgman–Stockbarger growth of cadmium telluride by an eddy current technique. *Journal of Electronic Materials*, 1995, vol. 24, no. 5, pp. 491–495. https://doi.org/10.1007/bf02657952

Авторы

Кузьмин Максим Игоревич — руководитель направления по разработке программного обеспечения, Государственный научно-исследовательский и проектный институт редкометаллической промышленности «Гиредмет» имени Н.П. Сажина, Москва, 111524, Российская Федерация, sc 59374615000, https://orcid.org/0000-0001-6265-9052, mimikatz@mail.ru

Ельников Максим Сергеевич — стажер-исследователь, Государственный научно-исследовательский и проектный институт редкометаллической промышленности «Гиредмет» имени Н.П. Сажина, Москва, 111524, Российская Федерация; студент, Московский государственный технический университет им. Н.Э. Баумана, Москва, 105005, Российская Федерация, https://orcid.org/0009-0001-8149-5302, elnikow.max@gmail.com

Кушнирук Давид Ильич — начальник группы, Государственный научно-исследовательский и проектный институт редкометаллической промышленности «Гиредмет» имени Н.П. Сажина, Москва, 111524, Российская Федерация, https://orcid.org/0009-0006-4419-6983, k.davjd@gmail.com

Морозов Максим Витальевич — ведущий инженер-технолог, Государственный научно-исследовательский и проектный институт редкометаллической промышленности «Гиредмет» имени Н.П. Сажина, Москва, 111524, Российская Федерация, https://orcid.org/0009-0007-1782-9623, MViMorozov@yandex.ru

Кузнецов Михаил Сергеевич — начальник лаборатории, Государственный научно-исследовательский и проектный институт редкометаллической промышленности «Гиредмет» имени Н.П. Сажина, Москва, 111524, Российская Федерация, sc 55421893200, https://orcid.org/0000-0002-8441-4424, gradan@mail.ru

Authors

Maksim I. Kuzmin — Head of the Software Development Department, JSC N.P. Sazhin State Scientific Research and Design Institute of Rare Metal Industry “Giredmet”, Moscow, 111524, Russian Federation, sc 59374615000, https://orcid.org/0000-0001-6265-9052, mimikatz@mail.ru

Maksim S. Elnikov — Research Intern, JSC N.P. Sazhin State Scientific Research and Design Institute of Rare Metal Industry “Giredmet”, Moscow, 111524, Russian Federation; Student, Bauman Moscow State Technical University, Moscow, 105005, Russian Federation, https://orcid.org/0009-0001-8149-5302, elnikow.max@gmail.com

David I. Kushniruk — Head of Group, JSC N.P. Sazhin State Scientific Research and Design Institute of Rare Metal Industry “Giredmet”, Moscow, 111524, Russian Federation, https://orcid.org/0009-0006-4419-6983, k.davjd@gmail.com

Maksim V. Morozov — Leading Process Engineer, JSC N.P. Sazhin State Scientific Research and Design Institute of Rare Metal Industry “Giredmet”, Moscow, 111524, Russian Federation, https://orcid.org/0009-0007-1782-9623, MViMorozov@yandex.ru

Mikhail S. Kuznetsov — Head of Laboratory, JSC N.P. Sazhin State Scientific Research and Design Institute of Rare Metal Industry “Giredmet”, Moscow, 111524, Russian Federation, sc 55421893200, https://orcid.org/0000-0002-8441-4424, gradan@mail.ru

Статья поступила в редакцию 20.05.2025
Одобрена после рецензирования 20.07.2025
Принята к печати 23.09.2025

Received 20.05.2025
Approved after reviewing 20.07.2025
Accepted 23.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-807-816

Optical spin currents in chiral optical fibers

Ilya A. Deriy[1]✉, Danil F. Kornovan[2], Mihail I. Petrov[3], Andrey A. Bogdanov[4]

[1,4] Harbin Engineering University, Qingdao, 266000, China
[1,2,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] ilya.deriy@metalab.ifmo.ru✉, https://orcid.org/0000-0002-6515-9325
[2] d.kornovan@metalab.ifmo.ru, https://orcid.org/0000-0002-4851-0697
[3] m.petrov@metalab.ifmo.ru, https://orcid.org/0000-0001-8155-9778
[4] a.bogdanov@metalab.ifmo.ru, https://orcid.org/0000-0002-8215-0445

Abstract

This paper is devoted to the study of optical chiral cylindrical waveguides from the point of view of their application in optical spintronics. In the paper, it is proposed to use a chiral optical cylindrical waveguide as an optical spin diode. The mode structure of the waveguide under consideration is calculated and the dispersion equation for fundamental modes of the waveguide with an azimuthal number $m = \pm 1$ is numerically solved for various values of the chirality parameter of the waveguide material. Expressions for the energy flux and the optical spin current inside the waveguide are derived. It is shown that in the single-mode regime, the direction of the optical spin currents in the waveguide is determined exclusively by the sign of the chirality parameter of the waveguide material, regardless of the azimuthal number and the direction of mode propagation. Due to this, the superposition of $m = 1$ and $m = -1$ modes propagating in opposite directions will have a zero energy flux, but a nonzero optical spin current. Our results expand the element base of optical spintronics and open up new ways for creating energy-efficient optical computing systems.

Keywords

optical spin, optical spintronics, chirality, optical information transfer, optical waveguides, optical nanofibers

Acknowledgements

The studies were supported by the Russian Science Foundation (Project 23-72-10059). Ilya Deriy acknowledges the BASIS Foundation for valuable support.

For citation: Deriy I.A., Kornovan D.F., Petrov M.I., Bogdanov A.A. Optical spin currents in chiral optical fibers. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 807–816. doi: 10.17586/2226-1494-2025-25-5-807-816

УДК 535.131

Оптические спиновые токи в хиральных оптоволокнах

Илья Александрович Дерий[1]✉, Данил Феодосьевич Корновaн[2], Михаил Игоревич Петров[3], Андрей Андреевич Богданов[4]

[1,4] Харбинский Инженерный Университет, Циндао, 266000, Китай
[1,2,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] ilya.deriy@metalab.ifmo.ru✉, https://orcid.org/0000-0002-6515-9325
[2] d.kornovan@metalab.ifmo.ru, https://orcid.org/0000-0002-4851-0697
[3] m.petrov@metalab.ifmo.ru, https://orcid.org/0000-0001-8155-9778
[4] a.bogdanov@metalab.ifmo.ru, https://orcid.org/0000-0002-8215-0445

Аннотация

Представлено исследование оптических хиральных цилиндрических волноводов с точки зрения их применения в оптической спинтронике. Предложено использование хирального оптического цилиндрического волновода в роли оптического спинового диода. Рассчитана модовая структура рассматриваемого волновода и численно решено дисперсионное уравнение для фундаментальных мод волновода с азимутальным числом $m = \pm 1$ для

различных значений параметра хиральности материала волновода. Получены выражения для потока энергии и оптического спинового тока внутри волновода. Показано, что в одномодовом режиме работы волновода, направление протекания оптических спиновых токов в волноводе определяется исключительно знаком параметра хиральности материала волновода, вне зависимости от азимутального числа и направления распространения моды. Следовательно, суперпозиция $m = 1$ и $m = -1$ мод, распространяющихся в разных направлениях, будет иметь нулевой поток энергии, но ненулевой оптический спиновой ток. Полученные результаты расширяют элементную базу оптической спинтроники и открывают новые пути для создания энергоэффективных оптических вычислительных систем.

Ключевые слова
оптический спин, оптическая спинтроника, хиральность, оптические методы переноса информации, оптические волноводы, оптоволокно

Благодарности
Исследования поддержаны Российским научным фондом (проект 23-72-10059). Илья Дерий выражает благодарность Фонду развития теоретической физики и математики «БАЗИС» за неоценимую поддержку.

Ссылка для цитирования: Дерий И.А., Корнован Д.Ф., Петров М.И., Богданов А.А. Оптические спиновые токи в хиральных оптоволокнах // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 807–816 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-807-816

Introduction

Photonics has rapidly advanced as a transformative technology providing numerous benefits compared to conventional electronics, including lower material losses, enhanced bandwidth capabilities, and significantly faster data transmission [1, 2]. While traditional electronic systems depend on the movement of electrons, photonics leverages the intrinsic characteristics of photons, notably their extremely high propagation speed and negligible mass. This intrinsic advantage ensures that photonic devices operate with substantially lower heat generation, resulting in higher operational frequencies and markedly improved energy efficiency relative to electronic devices [3–5]. However, the operation principle of most photonic and optoelectronic devices is based on electromagnetic energy which introduces severe difficulties into the engineering process of nonreciprocal and nonlinear optical devices, and imposes significant limitations on energy consumption and modulation frequencies.

Spintronics, similarly poised as a promising alternative to traditional electronics, utilizes the intrinsic spin of electrons rather than their charge for data storage and processing [6–8]. The utilization of spin degrees of freedom in electrons allows for energy-efficient data processing and storage, reduced heat generation, and increased device performance [9, 10]. Analogously, photons possess intrinsic angular momentum, comprised of Spin Angular Momentum (SAM), related to polarization, and Orbital Angular Momentum, arising from spatial phase distributions [11–13]. The manipulation of SAM and Orbital Angular Momentum of photons provides additional degrees of freedom for information encoding, transmission, and manipulation, enhancing the capabilities of optical communication systems and quantum computing technologies [14–16]. And recently it was shown that utilization of electromagnetic SAM as an information carrier allows for the implementation of the low-energy nonreciprocal optical devices such as optical diode and circulator without the violation of the Lorentz reciprocity [17]. However, the real-life implementation of an optical spin diode here was purely conceptual and far from real photonic systems.

Optical nanofibers have become particularly important in the field of photonics due to their exceptional ability to guide and confine light in extremely small volumes [18, 19]. These fibers are characterized by sub-wavelength diameters, enabling strong evanescent fields and efficient interaction between guided modes and surrounding materials or quantum emitters. Nanofibers offer significant advantages for applications in optical communications, sensing, and quantum information processing due to their high field intensity, low loss, flexibility, and ease of integration into existing photonic platforms [20, 21]. The convenience, versatility, and widespread availability of optical nanofibers have made them ubiquitous in modern photonic technologies.

In this work, we extend the ideas developed in [17] on the use of optical spin currents for optical communication and information processing. We show that a chiral optical nanofiber can operate in the optical spin diode regime. To confirm this fact, we solve the eigenstate problem for a chiral optical waveguide, derive expressions for the energy flux and the SAM flux in such a waveguide, and show that in the nanofiber regime, the electromagnetic energy can propagate in either direction, while the direction of propagation of the optical spin currents is strictly specified, and is determined by the sign of the chirality parameter of the waveguide material. We note that while all results presented in this paper are for chiral waveguides made of isotopic chiral materials, the same physics can be observed for structurally chiral waveguides, e.g. like in [22–24].

Eigenmodes of a Chiral Cylindrical Waveguide

Consider a cylindrical waveguide with radius a whose axis is directed along the z-axis (Fig. 1, *a*). The waveguide under consideration is made of an isotropic chiral material with permittivity ϵ, permeability $\mu = 1$, and chirality parameter χ. For simplicity, we consider air to be the surrounding medium. Following the default approach for the solution of an eigenmode problem for a cylindrical waveguide, we find eigenmodes by solving Maxwell's equations inside and outside of the waveguide, and then use boundary conditions to find unknown complex coefficients.

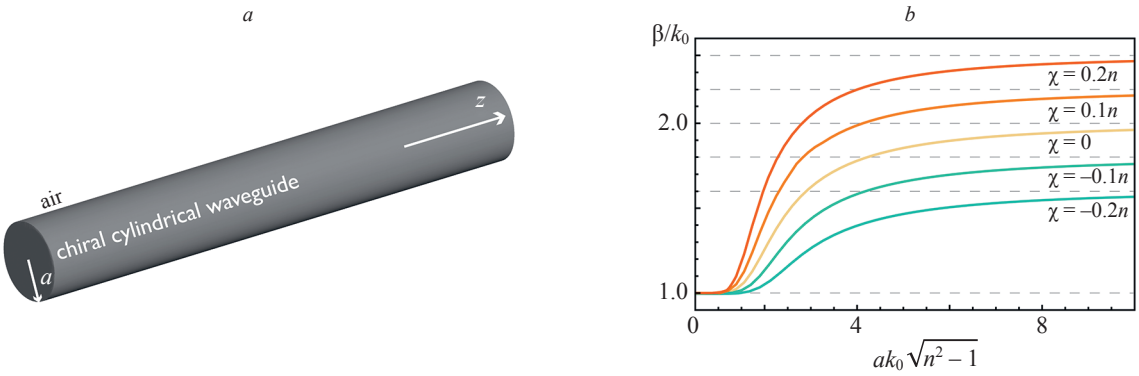

Fig. 1. Isotropic chiral waveguide. Geometry of the waveguide (*a*). Dispersion of the fundamental mode with azimuthal number $m = 1$ for the value of the refractive index $n = 2$, and different values of χ (*b*). Variables k_0 and β encode freespace wavenumber and propagation constant accordingly

In an arbitrary homogeneous chiral medium with material parameters ϵ, μ, and χ, in the basis of Riemann-Silberstein vectors (see section "Supplementary Materials"), Maxwell's equations can be written as

$$\nabla\times\mathbf{F}_{\pm}(\mathbf{r}) = \pm k_0 n_{\pm}\mathbf{F}_{\pm}(\mathbf{r}),\ n_{\pm} = n \pm \chi, \tag{1}$$

where $\mathbf{F}_{\pm}(\mathbf{r}) = \mathbf{E}(\mathbf{r}) \pm i\sqrt{\mu/\epsilon}\mathbf{H}(\mathbf{r})$ are the Riemann-Silberstein vectors [25, 26]; $n = \sqrt{\epsilon\mu}$ is the refractive index of the media; $k_0 = \omega/c$ is the freespace wavenumber.

Riemann-Silberstein vectors separates electromagnetic fields into two physically independent chiral components with sign $\pm$ encoding the helicity of the field, which makes such a basis a very natural choice for the chiral materials. In addition, due to the cylindrical symmetry of the problem, we can use the following ansatz [27]

$$\mathbf{F}_{\pm}(\mathbf{r}) = \mathbf{F}_{\pm,m}(\rho, \beta)e^{im\varphi + i\beta z}, \tag{2}$$

where m is the angular momentum number; β is the propagation constant; ρ and φ are the radial and angular coordinates in the cylindrical system of coordinates.

Thus, in cylindrical coordinates, Eq. (1) can be written as (omitting index m and argument β for the sake of brevity)

$$\frac{im}{\rho}F_{\pm,z}(\rho) - i\beta F_{\pm,\varphi}(\rho) = \pm n_{\pm}k_0 F_{\pm,\rho}(\rho), \tag{3}$$

$$i\beta F_{\pm\rho}(\rho) - \partial_\rho F_{\pm,z}(\rho) = \pm n_{\pm}k_0 F_{\pm,\varphi}(\rho), \tag{4}$$

$$\frac{1}{\rho}[\partial_r(\rho F_{\pm\varphi}(\rho)) - imF_{\pm,\rho}(\rho)] = \pm n_{\pm}k_0 F_{\pm,z}(\rho), \tag{5}$$

from which it is possible to derive the following relations,

$$\begin{gathered}\left[\frac{\pm imn_{\pm}k_0}{\rho} + i\beta\partial_\rho\right]F_{\pm,z}(\rho) = \kappa_{\pm}^2 F_{\pm,\rho}(\rho),\\ \left[\frac{m\beta}{\rho} \mp n_{\pm}k_0\partial_\rho\right]F_{\pm,z}(\rho) = \kappa_{\pm}^2 F_{\pm,\varphi}(\rho),\end{gathered} \tag{6}$$

where $\kappa_{\pm}^2 = n_{\pm}^2 k_0^2 - \beta^2$, and the equation for $F_{\pm,z}$

$$\rho\partial_\rho[\rho\partial_\rho F_{\pm,z}(\rho)] + [\rho^2\kappa_{\pm}^2 - m^2]F_{\pm,z}(\rho) = 0. \tag{7}$$

Eq. (7) is the Bessel equation, which solution are Bessel functions. To satisfy the regularity condition for the solution at the origin $\rho = 0$, and to satisfy the outgoing boundary conditions, we choose a Bessel function of the first kind $J_m(\kappa_{\pm}^{\text{in}}\rho)$ for the fields inside the waveguide, and a Hankel function of the first kind $H_m^{(I)}(\kappa_{\pm}^{\text{out}}\rho)$ for the fields outside of the waveguide

$$\begin{aligned}F_{\pm,z}^{\text{in}}(\rho) &= a_{\pm}\kappa_{\pm}^{\text{in}2}J_m(\kappa_{\pm}^{\text{in}}\rho),\\ F_{\pm,z}^{\text{out}}(\rho) &= b_{\pm}\kappa_{\pm}^{\text{out}2}\rho H_m^{(\text{I})}(\kappa_{\pm}^{\text{out}}\rho),\end{aligned} \tag{8}$$

where indices "in" and "out" encode fields inside and outside of the waveguide; $a_{\pm}$ and $b_{\pm}$ are the complex amplitudes yet to be determined.

To find complex amplitudes $a_{\pm}$ and $b_{\pm}$, we substitute the obtained solution into a set of boundary conditions, which reads as

$$\begin{gathered}F_{+,\tau}^{\text{in}}(a) + F_{-,\tau}^{\text{in}}(a) = F_{+,\tau}^{\text{out}}(a) + F_{-,\tau}^{\text{out}}(a),\\ \frac{F_{+,\tau}^{\text{in}}(a) - F_{-,\tau}^{\text{in}}(a)}{\sqrt{\mu^{\text{in}}/\epsilon^{\text{in}}}} = \frac{F_{+,\tau}^{\text{out}}(a) - F_{-,\tau}^{\text{out}}(a)}{\sqrt{\mu^{\text{out}}/\epsilon^{\text{out}}}}.\end{gathered} \tag{9}$$

where τ encodes tangential component of the vector field; $\mu^{\text{in/out}}$, $\epsilon^{\text{in/out}}$ are the permeability and permittivity of the material of the core (in) and claddings (out) of the waveguide.

Expressions for the angular component of the fields can be derived from Eqs. (6), and (8)

$$\begin{aligned}F_{\pm,\varphi}^{\text{in}}(\rho) &= Q_{m,\pm}(\kappa_{\pm}^{\text{in}}\rho) =\\ &= a_{\pm}\left[-\frac{m\beta}{\rho}J_m(\kappa_{\pm}^{\text{in}}\rho) \mp n_{\pm}^{\text{in}}\kappa_{\pm}^{\text{in}}k_0 J_m'(\kappa_{\pm}^{\text{in}}\rho)\right],\\ F_{\pm,\varphi}^{\text{out}}(\rho) &= P_{m,\pm}(\kappa\rho) =\\ &= b_{\pm}\left[-\frac{m\beta}{\rho}H_m(\kappa_{\pm}^{\text{out}}\rho) \mp n_{\pm}^{\text{out}}\kappa_{\pm}^{\text{out}}k_0 H_m'(\kappa_{\pm}^{\text{out}}\rho)\right].\end{aligned} \tag{10}$$

Combining Eqs. (8)–(10), and introducing dimensionless parameters

$$\begin{gathered}\tilde{\kappa}_{\pm} \equiv a\kappa_{\pm}^{\text{in}} = \sqrt{n_{\pm}^2\tilde{k}^2 - \tilde{\beta}^2},\ \tilde{\kappa} \equiv a\kappa_{\pm}^{\text{out}} = \sqrt{\tilde{k}^2 - \tilde{\beta}^2},\\ \tilde{k} = ak_0,\ \tilde{\beta} = a\beta,\end{gathered} \tag{11}$$

and recalling that

$$\epsilon^{\text{in}} = \epsilon,\ \epsilon^{\text{out}} = 1,\ \mu^{\text{in}} = \mu^{\text{out}} = 1,\ \chi^{\text{out}} = 0, \tag{12}$$

boundary conditions become

$$\begin{aligned} a_+\tilde{\kappa}_+^2 J_m(\tilde{\kappa}_+) + a_-\tilde{\kappa}_-^2 J_m(\tilde{\kappa}_-) &= (b_+ + b_-)\tilde{\kappa}^2 H_m(\tilde{\kappa}),\\ a_+\tilde{\kappa}_+^2 n J_m(\tilde{\kappa}_+) - a_-\tilde{\kappa}_-^2 n J_m(\tilde{\kappa}_-) &= (b_+ - b_-)\tilde{\kappa}^2 H_m(\tilde{\kappa}),\\ a_+ Q_{m,+}(\tilde{\kappa}_+) + a_- Q_{m,-}(\tilde{\kappa}_-) &= b_+ P_{m,+}(\tilde{\kappa}) + b_- P_{m,-}(\tilde{\kappa}),\\ a_+ n Q_{m,+}(\tilde{\kappa}_+) - a_- n Q_{m,-}(\tilde{\kappa}_-) &= b_+ P_{m,+}(\tilde{\kappa}) - b_- P_{m,-}(\tilde{\kappa}), \end{aligned} \tag{13}$$

where $n = \sqrt{\epsilon^{\text{in}}}$. The solvability condition $\det \hat{\mathcal{D}} = 0$ gives the dispersion equation for the eigenmodes of the system. These equations can be simplified if we notice that

$$\begin{aligned} &b_+ P_{m,+}(\tilde{\kappa}) \pm b_- P_{m,-}(\tilde{\kappa}) =\\ &= b_+[-m\tilde{\beta} H_m(\tilde{\kappa}) - \tilde{k}\tilde{\kappa} H'_m(\tilde{\kappa})] \pm\\ &\pm b_-[-m\tilde{\beta} H_m(\tilde{\kappa}) + \tilde{k}\tilde{\kappa} H'_m(\tilde{\kappa})] =\\ &= -(b_+ \pm b_-)m\tilde{\beta} H_m(\tilde{\kappa}) - (b_+ \mp b_-)\tilde{k}\tilde{\kappa} H'_m(\tilde{\kappa}). \end{aligned} \tag{14}$$

Introducing

$$c_\pm = (b_+ \pm b_-)H_m(\tilde{\kappa}), \tag{15}$$

we rewrite the boundary conditions in a matrix form as

$$\begin{bmatrix} \tilde{\kappa}_+^2 J_m(\tilde{\kappa}_+) & \tilde{\kappa}_-^2 J_m(\tilde{\kappa}_-) & -H_m(\tilde{\kappa}) & 0\\ n\tilde{\kappa}_+^2 J_m(\tilde{\kappa}_+) & -n\tilde{\kappa}_-^2 J_m(\tilde{\kappa}_-) & 0 & -H_m(\tilde{\kappa})\\ Q_{m,+}(\tilde{\kappa}_+) & Q_{m,-}(\tilde{\kappa}_-) & m\tilde{\beta} H_m(\tilde{\kappa}) & \tilde{k}\tilde{\kappa} H'_m(\tilde{\kappa})\\ nQ_{m,+}(\tilde{\kappa}_+) & \underbrace{-nQ_{m,-}(\tilde{\kappa}_-)}_{\hat{\mathcal{D}}} & \tilde{k}\tilde{\kappa} H'_m(\tilde{\kappa}) & m\tilde{\beta} H_m(\tilde{\kappa}) \end{bmatrix} \times \begin{bmatrix} a_+\\ a_-\\ c_+\\ c_- \end{bmatrix} = 0. \tag{16}$$

It is possible to show that

$$\begin{aligned} (m, \beta) \to (-m, -\beta) \Rightarrow (F_{\pm,\rho}, F_{\pm,\varphi}, F_{\pm,z}) \to\\ \to (-1)^m(-F_{\pm,\rho}, -F_{\pm,\varphi}, F_{\pm,z}), \end{aligned} \tag{17}$$

which follows from Eqs. (6), (8), and expressions for the Bessel and Hankel functions of the negative order $J_{-m}(z) = (-1)^m J_m(z)$, $H_{-m}(z) = (-1)^m H_m(z)$. In addition,

$$\begin{aligned} (m, \chi) \to (-m, -\chi) \Rightarrow (-1)^m(F_{\pm,\rho}, F_{\pm,\varphi}, F_{\pm,z}) \to\\ \to (F_{\mp,\rho}, -F_{\mp,\varphi}, F_{\mp,z}). \end{aligned} \tag{18}$$

And, from Eqs. (17), and (18) it follows that

$$\begin{aligned} (\beta, \chi) \to (-\beta, -\chi) \Rightarrow (F_{\pm,\rho}, F_{\pm,\varphi}, F_{\pm,z}) \to\\ \to (-F_{\mp,\rho}, -F_{\mp,\varphi}, F_{\mp,z}). \end{aligned} \tag{19}$$

Symmetries in Eqs. (17)–(19) follow from the properties of the physical system, namely reciprocity, chirality, and time-reversibility. It is possible to show that each of the symmetries corresponds to the multiplication of $\hat{\mathcal{D}}$ rows/columns by a constant or to the row/column permutations. It is well-known that such transformations leave the determinant of the matrix invariant, and, as a consequence, lead to the corresponding degeneracies in the eigenmodes structure of the system.

By taking $\hat{\mathcal{D}} = 0$, we arrive at the dispersion equation which can be written as

$$2n\tilde{k}^2\tilde{\kappa}^2 H_m^2(\tilde{\kappa}) J_m(\tilde{\kappa}_+) J_m(\tilde{\kappa}_-) \mathcal{F}_m(\tilde{k}, \tilde{\beta}, n, \chi) = 0, \tag{20}$$

where

$$\begin{aligned} \mathcal{F}_m(\tilde{k}, \tilde{\beta}, n, \chi) = \tilde{\kappa}_+^2\tilde{\kappa}_-^2 \mathcal{H}_m^2(\tilde{\kappa}) +\\ + \frac{\tilde{\kappa}(n^2+1)}{2n} \mathcal{H}_m(\tilde{\kappa})[4\tilde{k}\tilde{\beta} mn\chi - \tilde{\kappa}_+\tilde{\kappa}_-(n_+\tilde{\kappa}_- \mathcal{J}_m(\tilde{\kappa}_+))] +\\ + n_-\tilde{\kappa}_- \mathcal{J}_m(\tilde{\kappa}_-) + \left[n_+\tilde{\kappa}\tilde{\kappa}_+ \mathcal{J}_m(\tilde{\kappa}_+) - \frac{\tilde{k}\tilde{\beta}(n_+^2 - 1)}{\tilde{\kappa}}\right] \times\\ \times \left[n_-\tilde{\kappa}\tilde{\kappa}_- \mathcal{J}_m(\tilde{\kappa}_-) + \frac{\tilde{k}\tilde{\beta}(n_-^2 - 1)}{\tilde{\kappa}}\right], \end{aligned} \tag{21}$$

where

$$\mathcal{H}_m(x) = \frac{H'_m(x)}{H_m(x)},\ \mathcal{J}_m(x) = \frac{J'_m(x)}{J_m(x)}. \tag{22}$$

We note that for waveguide modes, $\tilde{\kappa}^2 < 0$, which corresponds to the total internal reflection regime. In this case, in the waveguide cladding, fields become evanescent instead of propagating. Mathematically, this means that a more natural choice for the fields outside of the waveguide is the modified Bessel function (Macdonald function). The conversion from the Hankel to the Macdonald function can be done via the introduction of the parameter ζ as

$$\tilde{\zeta} = \sqrt{\tilde{\beta}^2 - \tilde{\kappa}_0^2} \in \mathbb{R},\ \tilde{\kappa} = i\tilde{\zeta}. \tag{23}$$

Then, the Hankel functions will be expressed in terms of the modified Bessel functions as

$$H_m(\tilde{\kappa}) = \frac{2}{\pi i^{m+1}} K_m(\tilde{\zeta}),\ H'_m(\tilde{\kappa}) = \frac{2}{\pi i^{m+2}} K'_m(\tilde{\zeta}), \tag{24}$$

$$\mathcal{H}_m(\tilde{\kappa}) = -i\frac{K'_m(\tilde{\zeta})}{K_m(\tilde{\zeta})} \equiv -i\mathcal{K}_m(\tilde{\zeta}). \tag{25}$$

Eq. (20) is a quadratic equation for $\mathcal{H}_m$, thus it has two solutions, corresponding to EH and HE modes, as in the dielectric waveguide. We note that due to the presence of the reciprocity and time-reversal symmetry, the dispersion equation is invariant under each of the following transformations

$$\begin{aligned} (\chi, \beta) \to (-\chi, -\beta),\ (m, \beta) \to (-m, -\beta),\\ (m, \chi) \to (-m, -\chi). \end{aligned} \tag{26}$$

Dispersion in terms of the effective refractive index $n_{\text{eff}} = \beta/k_0$ and size parameter $V = ak_0\sqrt{n^2 - 1}$ for the fundamental mode ($m = \pm 1$), calculated from Eq. (20), is

presented in Fig. 1, *b* for $n = 2$ and different values of χ. The dispersion is plotted only for $m = 1$ and $\beta > 0$, however all other cases can be obtained from the symmetrical considerations. One can see that the dispersion of the fundamental mode for $\chi \neq 0$ is similar to the dispersion of a dielectric waveguide with dielectric permittivity $\epsilon = n_{\pm}^2$. Dispersion curves for the chiral waveguide have similar behavior to the dispersion curves for the dielectric waveguide. Namely, the effective refractive index of the fundamental waveguide mode with $m = 1$ lies in the range $n_{\text{eff}} \in [n_0, n_+]$, where $n_0 = 1$ is the refractive index of the waveguide cladding. This fact suggests that one can treat κ_+ as the transversal quasi-wavevector component of the mode, similar to the $\sqrt{n_{\text{in}}^2 k_0^2 - \beta^2}$ in the dielectric waveguides. From the symmetry of the system, one can show that the dispersion of the fundamental mode with $m = -1$ is the same as for the fundamental mode with $m = 1$, but for the chirality parameter equal to $-\chi$. In particular, the effective refractive index of the mode with $m = -1$ will lie in the range $n_{\text{eff}} \in [n_0, n_-]$, and the transversal quasi-wavevector component will be κ_-. Fig. 2 shows the dependence of the effective refractive index of the fundamental eigenmodes (with $m = \pm 1$) on the chirality parameter in the different regions of the dispersion curve. Near the cutoff ($ak_0\sqrt{\epsilon - 1} = 1.01$, Fig. 2, *a*), the effective refractive index shows nonlinear dependence on the chirality parameter, however the difference $\Delta n_{\text{eff}} = n_{\text{eff}}(\chi) - n_{\text{eff}}(-\chi)$ is negligible. Far from cut-off, as expected, the behaviour of the effective refractive index is close to a straight line $n_{\text{eff}} \approx n_{\pm}$.

Power and SAM Flow

Energy flow and optical SAM flow inside the waveguide can be calculated as (see section "Supplementary Materials")

$$\langle P_z^{\text{in}} \rangle = \langle P_{z,+}^{\text{in}} \rangle + \langle P_{z,-}^{\text{in}} \rangle, \langle J_{zz}^{\text{in}} \rangle = \langle J_{zz,+}^{\text{in}} \rangle + \langle J_{zz,-}^{\text{in}} \rangle, \quad (27)$$

where

$$\langle P_{z,\pm}^{\text{in}} \rangle = \mp \frac{cn}{8} \int_0^a \text{Im}\{F_{\pm,\rho}^{\text{in}}(\rho) F_{\pm,\varphi}^{\text{in}*}(\rho)\} \rho \, d\rho, \quad (28)$$

and

$$\langle J_{zz,+}^{\text{in}} \rangle = \pm \frac{cn}{32k_0} \int_0^a \left[|F_{\pm,\rho}^{\text{in}}(\rho)|^2 + |F_{\pm,\varphi}^{\text{in}}(\rho)|^2 - |F_{\pm,z}^{\text{in}}(\rho)|^2 \right] \rho \, d\rho. \quad (29)$$

Each of the integrals above can be presented in the following form

$$I = c_1 \int_0^a \frac{J_m^2(\kappa\rho)}{\rho} d\rho + c_2 \int_0^a J_m(\kappa\rho) J_m'(\kappa\rho) \, d\rho + c_3 \int_0^a J_m'^2(\kappa\rho) \rho \, d\rho + c_4 \int_0^a J_m^2(\kappa\rho) \rho \, d\rho. \quad (30)$$

We note that all the integrals above are calculated for the case of $m \geq 0$, however, since $J_{-m}(x) = (-1)^m J_m(x)$ and $H_{-m}(x) = (-1)^m H_m(x)$, all quadratic forms of Bessel functions of same m will be invariant under the transformation $m \to -m$. The integral

$$I_4 = \int_0^a J_m^2(\kappa\rho) \rho \, d\rho = \frac{a^2}{2} [J_m^2(a\kappa) - J_{m-1}(a\kappa) J_{m+1}(a\kappa)] \quad (31)$$

can be easily evaluated using the Lommel's integral. The integral

$$I_2 = \int_0^a J_m(\kappa\rho) J_m'(\kappa\rho) \, d\rho = \frac{1}{\kappa} \int_0^{\kappa a} J_m(x) \frac{dJ_m(x)}{dx} dx = \frac{J_m^2(a\kappa) - J_m^2(0)}{\kappa}, \quad (32)$$

is straightforward. Integrals

$$I_1 = \int_0^a \frac{J_m^2(\kappa\rho)}{\rho} d\rho = \frac{(a\kappa)^{2m}}{4^m} \Gamma(2m) {}_2\tilde{F}_3\left(m, \frac{1}{2} + m; 1 + m, 1 + m, 1 + 2m; -a^2\kappa^2\right), \quad (33)$$

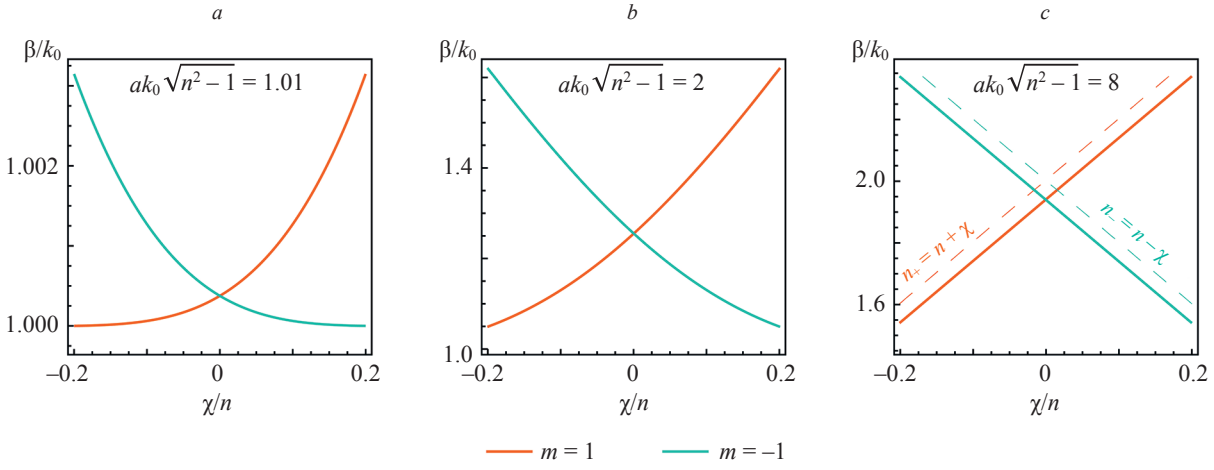

Fig. 2. Dependence of the effective refractive index $n_{\text{eff}} = \beta/k_0$ of the fundamental eigenmodes of the waveguide ($m = \pm 1$) vs. the normalized chirality parameter χ/n for different values of the dimensionless size parameter $V = ak_0\sqrt{\epsilon - 1} = 1.01$ (*a*); $V = 2$ (*b*); $V = 8$ (*c*)

where $\Gamma(x)$ is the Gamma function, and

$$I_3 = \int_0^a J_m'^2(\kappa\rho)\rho\, d\rho = \frac{1}{4}\int_0^a [J_{m-1}^2(\kappa\rho) + J_{m+1}^2(\kappa\rho) - 2J_{m-1}(\kappa\rho)J_{m+1}(\kappa\rho)]\rho\, d\rho = \frac{a^2}{8}[J_{m-1}^2(a\kappa) - J_{m-2}(a\kappa)J_m(a\kappa)] + \frac{a^2}{8}[J_{m+1}^2(a\kappa) - J_m(a\kappa)J_{m+2}(a\kappa)] + \frac{a^2(a\kappa)^{2m} m\, m!}{2^{2m+1}}\Gamma(2m)_3\tilde{F}_4\left(\frac{1}{2}+m, 1+m, 1+m; m, 2+m, 2+m, 1+2m; -a^2\kappa^2\right), \quad (34)$$

can be evaluated with the help of the regularized hypergeometric function ${}_p\tilde{F}_q$ [28]. Finally,

$$\langle P_{z,\pm}^{\text{in}}\rangle = \mp|a_\pm|^2\frac{cn}{8}\,\text{Im}\{\mp im^2 n_\pm k_0\beta I_1 - im\kappa_\pm(\beta^2 + n_\pm^2 k_0^2)I_2 \mp \mp i\beta n_\pm k_0\kappa_\pm^2 I_3\} = \frac{cn}{8}[n_\pm k_0\beta(m^2 I_1 + \kappa_\pm^2 I_3) \pm \pm(\beta^2 + n_\pm^2 k_0^2)\kappa_\pm I_2], \quad (35)$$

$$\langle J_{zz,+}^{\text{in}}\rangle = \pm|a_\pm|^2\frac{cn}{32k_0}[m^2(n_\pm^2 k_0^2 + \beta^2)I_1 \pm 4mn_\pm k_0\beta\kappa_\pm I_2 + \kappa_\pm^2(n_\pm^2 k_0^2 + \beta^2)I_3 - \kappa_\pm^4 I_4] = \pm|a_\pm|^2\frac{cn}{32k_0}[(n_\pm^2 k_0^2 + \beta^2)(m^2 I_1 + \kappa_\pm^2 I_3) \pm 4mn_\pm k_0\beta\kappa_\pm I_2 - \kappa_\pm^4 I_4]. \quad (36)$$

For the fundamental mode with $m \pm 1$,

$$I_1 = \frac{1 - J_0^2(a\kappa) - J_1^2(a\kappa)}{2},\ \kappa I_2 = J_1^2(a\kappa),$$
$$\kappa^2 I_3 = \frac{-1 + (a^2\kappa^2+1)J_0^2(a\kappa) + (a^2\kappa^2 - 1)J_1^2(a\kappa)}{2}, \quad (37)$$
$$\kappa^2 I_4 = \frac{a^2\kappa^2}{2}[J_1^2(a\kappa) - J_0(a\kappa)J_2(a\kappa)],$$

and

$$m^2 I_1 + \kappa^2 I_3 = \frac{a^2\kappa^2 J_0^2(a\kappa) + (\kappa^2 - 2)J_1^2(a\kappa)}{2}. \quad (38)$$

Thus, for the fundamental mode

$$\langle P_{z,\pm}^{\text{in}}\rangle = |a_\pm|^2\frac{cn}{8a^2}\left[\frac{n_\pm\tilde{k}_0\tilde{\beta}}{2}[\tilde{\kappa}_\pm^2 J_0^2(\tilde{\kappa}_\pm) + (\kappa_\pm^2 - 2)J_1^2(\tilde{\kappa}_\pm)] \pm \pm(\tilde{\beta}^2 + n_\pm^2\tilde{k}_0^2)J_1^2(\tilde{\kappa}_\pm)\right], \quad (39)$$

$$\langle J_{zz,+}^{\text{in}}\rangle = \pm|a_\pm|^2\frac{cn}{32a^2k_0}\left[\frac{n_\pm^2\tilde{k}_0^2 + \tilde{\beta}^2}{2}[\tilde{\kappa}_\pm^2 J_0^2(\tilde{\kappa}_\pm) + (\kappa_\pm^2 - 2)J_1^2(\tilde{\kappa}_\pm)] \pm 4n_\pm\tilde{k}_0\tilde{\beta}J_1^2(\tilde{\kappa}_\pm) - \frac{\tilde{\kappa}_\pm^2}{2}[J_1^2(\tilde{\kappa}_\pm) - J_0(\tilde{\kappa}_\pm)J_2(\tilde{\kappa}_\pm)]\right]. \quad (40)$$

Optical spin diode regime

In the nanofiber regime, when $ak_0 << 1$, by doing a series expansion in ak and leaving only the leading terms, the eigenmode equation for $m = 1$ mode becomes

$$\begin{bmatrix} \frac{\tilde{\kappa}_+^3}{2} & \frac{\tilde{\kappa}_-^3}{2} & \frac{2\tilde{\kappa}}{\pi} & 0 \\ \frac{n\tilde{\kappa}_+^3}{2} & -\frac{n\tilde{\kappa}_-^3}{2} & 0 & \frac{2\tilde{\kappa}}{\pi} \\ -\tilde{\kappa}_+\frac{n_+\tilde{k}+\tilde{\beta}}{2} & \tilde{\kappa}_-\frac{n_-\tilde{k}-\tilde{\beta}}{2} & -\frac{2i\tilde{\beta}}{\pi\tilde{\kappa}} & \frac{2i\tilde{k}}{\pi\tilde{\kappa}} \\ -n\tilde{\kappa}_+\frac{n_+\tilde{k}+\tilde{\beta}}{2} & -n\tilde{\kappa}_-\frac{n_-\tilde{k}-\tilde{\beta}}{2} & \frac{2i\tilde{k}}{\pi\tilde{\kappa}} & -\frac{2i\tilde{\beta}}{\pi\tilde{\kappa}} \end{bmatrix}\begin{bmatrix} a_+ \\ a_- \\ c_+ \\ c_- \end{bmatrix} = 0, \quad (41)$$

with two solutions

$$a_+ = \frac{2i}{n\pi\tilde{\beta}},\ a_- = \frac{2i}{n\pi\tilde{\beta}},\ c_+ = 1,\ c_- = 0, \quad (42)$$

and

$$a_+ = \frac{2i}{n\pi\tilde{\beta}},\ a_- = -\frac{2i}{n\pi\tilde{\beta}},\ c_+ = 0,\ c_- = 1. \quad (43)$$

In addition, expressions for the energy and SAM flow became

$$\langle P_{z,\pm}^{\text{in}}\rangle = |a_\pm|^2\,\tilde{\kappa}_\pm^2\frac{cn}{8a^2}\frac{n_\pm\tilde{k}_0\tilde{\beta}}{4},$$
$$\langle J_{zz,\pm}^{\text{in}}\rangle = \pm|a_\pm|^2\,\tilde{\kappa}_\pm^2\frac{cn}{32a^2k_0}\frac{n_\pm^2\tilde{k}_0^2 + \tilde{\beta}^2}{4}. \quad (44)$$

Using Eqs. (42), and (43), and noting the fact that near the cutoff, $\beta \approx \eta k$, $\eta = \pm 1$, one can write that

$$\langle P_z^{\text{in}}\rangle = \frac{c\eta}{4a^2\pi^2}(n^2 + 3\chi^2 - 1),$$
$$\langle J_{zz}\rangle = \frac{c\chi}{4a^2\pi^2k_0}(n^2 + \chi^2). \quad (45)$$

Eqs. (45) are the key result of the paper. They show that the direction of the energy flow in the waveguide is governed by the direction of the propagation of the waveguide mode as in the traditional dielectric nanofibers. However, the direction of an optical spin current is governed by the sign of the chirality parameter.

From the orthogonality of the complex exponents $e^{im\varphi}$ and $e^{i\beta z}$ it follows that contribution of the different waveguide modes into the total energy and SAM flow is additive. In addition, it follows from Eq. (17) that transformation $(m, \beta) \to (-m, -\beta)$ leads to a transformation $\hat{\mathcal{D}} \to (-1)^m\hat{\mathcal{D}}$ since only φ and z components of the fields participate in the boundary conditions. Therefore, the eigenmode equation $\hat{\mathcal{D}}\mathbf{v} = 0$ is unchanged, which means

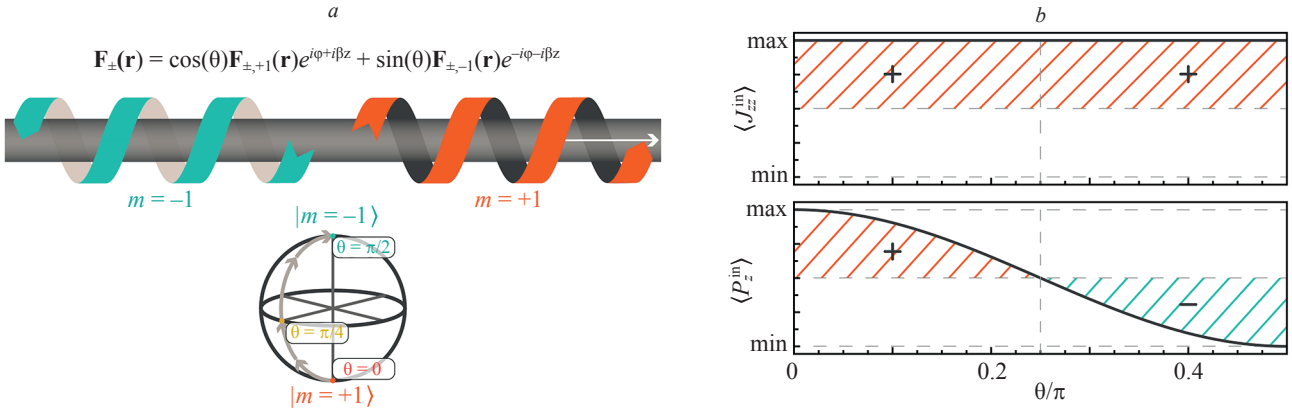

Fig. 3. Spin currents in an isotropic chiral nanofiber. Schematical representation of the propagation of the superposition of $m = 1$ and $m = -1$ modes in the optical spin diode regime (*a*). Spin Angular Momentum flux (spin current) and energy flux as a function of the parameter θ (*b*)

that both the eigenfrequencies and the eigenvectors of coefficients are identical for $m = 1$ and $m = -1$ eigenmodes.

This suggests that combination of two modes with opposite SAM (i.e. modes with $m = 1$ and $m = -1$) propagating in the opposite directions (Fig. 3, *a*) will lead to the emergence of the standing wave with zero energy transfer but nonzero optical spin current. Indeed, for a following configuration of fields

$$\mathbf{F}_{\pm}(\mathbf{r}) = a_{+1}\mathbf{F}_{\pm,+1}(\mathbf{r})e^{i\varphi+i\beta z} + a_{-1}\mathbf{F}_{\pm,-1}(\mathbf{r})e^{-i\varphi-i\beta z}, \quad (46)$$

total Poynting vector and total optical spin current in the waveguide are given by the expressions

$$\langle P_z^{\text{in}} \rangle = (|a_{+1}|^2 - |a_{-1}|^2)\frac{c}{4a^2\pi^2}(n^2 + 3\chi^2 - 1),$$

$$\langle J_{zz} \rangle = (|a_{+1}|^2 - |a_{-1}|^2)\frac{c\chi}{4a^2\pi^2 k_0}(n^2 + \chi^2). \quad (47)$$

One can see that when amplitudes of both waves are equal, there is no energy flux. This happens because two counter-propagating modes create a standing wave in the waveguide. However, optical spin current is non-zero for non-zero χ. Moreover, its direction and intensity is governed by the value of χ, which means that optical spin current is strictly positive for positive χ and strictly negative otherwise. And, the higher the value of χ, the higher the intensity of the optical spin current.

Indeed, both energy and optical spin current for such a superposition of an $m = 1$ and $m = -1$ modes with $a_{+1} = \cos\theta$ and $a_{-1} = \sin\theta$ is shown in Fig. 3, *b*. One can see that when amplitudes of both modes are equal, they form a standing wave. However, optical spin current is nonzero for any θ. Thus, it is possible to find such a regime, when energy flux is zero inside the waveguide, while optical spin current is not.

Conclusions

In this paper, we have shown that optical circular waveguide made of isotropic chiral material is a promising candidate for creating optical spin diodes and other opto spintronic devices with an asymmetric transfer characteristic. We have solved the eigenmode problem for such a waveguide and have shown that, by analogy with a circular dielectric waveguide, the fundamental mode of a circular chiral waveguide is the mode with an azimuthal number $m = \pm 1$. The main result of the paper are you the equations showing that the direction of the optical spin current in such waveguides is determined exclusively by the sign of the chirality parameter of the material from which the waveguide is made. Moreover, due to the symmetry of the system and the orthogonality of the waveguide modes, the chiral optical nanofiber supports configurations of standing electromagnetic waves with a non-zero optical spin current. We believe that our findings open new avenues for creating new energy efficient opto-spintronic devices for processing and transmitting information, which can serve as a good platform for further development of optical computing systems.

Supplementary Materials

We consider electromagnetic field in isotropic homogeneous chiral media. We start from Maxwell's equations

$$\begin{aligned} \nabla\times\mathbf{E} &= ik_0[\mu\mathbf{H} - i\chi\mathbf{E}], \\ \nabla\times\mathbf{H} &= -ik_0[\epsilon\mathbf{E} + i\chi\mathbf{H}], \\ \nabla\cdot[\epsilon\mathbf{E} + i\chi\mathbf{H}] &= 0, \\ \nabla\cdot[\mu\mathbf{H} - i\chi\mathbf{E}] &= 0. \end{aligned} \quad (48)$$

It is easy to show that the transversality of **E** and **H** fields still hold. From Eq. (48) it is possible to show that

$$\nabla\cdot[\epsilon\mu\mathbf{E} + i\chi\mu\mathbf{H} - i\chi\mu\mathbf{H} - \chi^2\mathbf{E}] = (\epsilon\mu - \chi^2)\nabla\cdot\mathbf{E} = 0,$$

and

$$(\epsilon\mu - \chi^2)\nabla\cdot\mathbf{H} = 0.$$

In chiral media, the basis of **E** and **H** fields is much less convenient than in usual dielectrics. We are looking for fields $\mathbf{F} = a\mathbf{E} + b\mathbf{H}$, and $\mathbf{G} = c\mathbf{E} + d\mathbf{H}$, such that

$$\nabla\times\mathbf{F} = ik_0\beta\mathbf{G}, \nabla\times\mathbf{G} = -ik_0\alpha\mathbf{F}. \quad (49)$$

Using Eqs. (48), and (49), we calculate

$$\nabla\times\mathbf{F} = \nabla\times[a\mathbf{E} + b\mathbf{H}] = ik_0[a(\mu\mathbf{H} - i\chi\mathbf{E}) + b(-\epsilon\mathbf{E} - i\chi\mathbf{H})] = \\ = ik_0[(-ai\chi - b\epsilon)\mathbf{E} + (a\mu - bi\chi)\mathbf{H}), \quad (50)$$

$$\nabla\times\mathbf{G} = \nabla\times[c\mathbf{E} + d\mathbf{H}] = -ik_0[c(-\mu\mathbf{H} + i\chi\mathbf{E}) + d(\epsilon\mathbf{E} + i\chi\mathbf{H})] = \\ = -ik_0[(ci\chi + d\epsilon)\mathbf{E} + (-c\mu + di\chi)\mathbf{H}). \quad (51)$$

Comparing Eqs. (50), and (51) equations with Eq. (49) results in the following system of equations.

$$\begin{bmatrix} -i\chi & -\epsilon & -\beta & 0 \\ \mu & -i\chi & 0 & -\beta \\ -\alpha & 0 & i\chi & \epsilon \\ 0 & -\alpha & -\mu & i\chi \end{bmatrix}\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = 0.$$

Solvability condition gives

$$\alpha\beta = (\sqrt{\epsilon\mu} \pm \chi)^2,$$

and coefficients are

$$b = \pm iaZ,\ c = \mp iaV,\ d = aZV,$$

$$Z = \sqrt{\frac{\mu}{\epsilon}},\ V = \sqrt{\frac{\alpha}{\beta}}.$$

Therefore

$$\begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix} = \begin{bmatrix} 1 & \pm iZ \\ \mp iV & ZV \end{bmatrix}\begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix}.$$

One can see that

$$\mathbf{G}_\pm = -iV\mathbf{F}_\pm.$$

Transition back to electric and magnetic fields is possible by the following transformation

$$\mathbf{E} = \frac{\mathbf{F}_+ + \mathbf{F}_-}{2},\ \mathbf{H} = \frac{\mathbf{F}_+ - \mathbf{F}_-}{2iZ}.$$

In such a basis, Maxwell equations become

$$\nabla\times\mathbf{F}_\pm = \pm k_0 n_\pm \mathbf{F}_\pm,\ n_\pm = n \pm \chi. \quad (52)$$

Such vectors are called Riemann-Silberstein vectors [25, 26]. We also note that from Eq. (52), $\nabla\times\mathbf{F}_\pm = 0$ automatically if $n_\pm \neq 0$.

Boundary conditions for electric and magnetic fields at the interface are $E_\tau^{\text{in}} = E_\tau^{\text{out}}$, $H_\tau^{\text{in}} = H_\tau^{\text{out}}$ where τ encodes tangential w.r.t interface component. One can derive boundary conditions for $\mathbf{F}_\pm$,

$$F_{+,\tau}^{\text{in}} + F_{-,\tau}^{\text{in}} = F_{+,\tau}^{\text{out}} + F_{-,\tau}^{\text{out}},\ \frac{F_{+,\tau}^{\text{in}} - F_{-,\tau}^{\text{in}}}{Z^{\text{in}}} = \frac{F_{+,\tau}^{\text{out}} - F_{-,\tau}^{\text{out}}}{Z^{\text{out}}}.$$

Time-averaged energy density in chiral media, in terms of Riemann-Silberstein vectors is

$$\langle W\rangle = \frac{\text{Re}\{\mathbf{E}\cdot\mathbf{D}^*\} + \text{Re}\{\mathbf{H}\cdot\mathbf{B}^*\}}{8\pi} = \\ = \frac{\text{Re}\{\epsilon|\mathbf{E}|^2 + i\chi\mathbf{E}\cdot\mathbf{H}^*\} + \text{Re}\{\mu|\mathbf{H}|^2 + i\chi\mathbf{H}\cdot\mathbf{E}^*\}}{8\pi} = \\ = \frac{\epsilon|\mathbf{E}|^2 + \mu|\mathbf{H}|^2 - \chi\text{Im}\{\mathbf{E}\cdot\mathbf{H}^*\} - \chi\text{Im}\{\mathbf{H}\cdot\mathbf{E}^*\}}{8\pi} = \\ = \frac{\epsilon|\mathbf{E}|^2 + \mu|\mathbf{H}|^2}{8\pi} = \frac{\epsilon}{32\pi}[|\mathbf{F}_+|^2 + |\mathbf{F}_-|^2].$$

Similarly, one can derive the time-averaged Poynting vector and the time-averaged optical spin current as

$$\langle\mathbf{S}\rangle = \frac{c}{8\pi}\text{Re}\{\mathbf{E}\times\mathbf{H}^*\} = \frac{c}{32\pi Z}\text{Re}\left\{\frac{(\mathbf{F}_+ + \mathbf{F}_-)\times(\mathbf{F}_+^* + \mathbf{F}_-^*)}{-i}\right\} = \\ = \frac{c}{32\pi Z}\text{Im}\{\mathbf{F}_-\times\mathbf{F}_-^* - \mathbf{F}_+\times\mathbf{F}_+^*\}$$

and

$$\langle\hat{\mathrm{J}}_s\rangle = \frac{c}{16\pi\omega}\text{Im}\{\hat{\mathbb{1}}\mathbf{E}\cdot\mathbf{H}^* - \mathbf{E}\otimes\mathbf{H}^* - \mathbf{H}^*\otimes\mathbf{E}\} = \\ = \frac{c}{64\pi\omega Z}\text{Re}\{\hat{\mathbb{1}}(|\mathbf{F}_+|^2 - |\mathbf{F}_-|^2) - 2[\mathbf{F}_+\otimes\mathbf{F}_+^* - \mathbf{F}_-\otimes\mathbf{F}_-^*].$$

References

1. Li Y., Monticone F. Exploring the role of metamaterials in achieving advantage in optical computing. *Nature Computational Science*, 2024, vol. 4, no. 8, pp. 545–548. https://doi.org/10.1038/s43588-024-00657-w
2. McMahon P.L. The physics of optical computing. *Nature Reviews Physics*, 2023, vol. 5, no. 12, pp. 717–734. https://doi.org/10.1038/s42254-023-00645-5
3. Chanana A., Larocque H., Moreira R., Carolan J., Guha B., Melo E.G., et al. Ultra-low loss quantum photonic circuits integrated with single quantum emitters. *Nature Communications*, 2022, vol. 13, no. 1, pp. 7693. https://doi.org/10.1038/s41467-022-35332-z
4. Dong T., Liang J.J., Camayd-Muñoz S., Liu Y., Tang H., Kita S., et al. Ultra-low-loss on-chip zero-index materials. *Light: Science & Applications*, 2021, vol. 10, no. 1, pp. 10. https://doi.org/10.1038/s41377-020-00436-y
5. Blundell S., Radford T.W., Ajia I.A., Lawson D., Yan X.Z., Banakar M., et al. Ultracompact programmable silicon photonics using layers of low-loss phase-change material Sb_2Se_3 of increasing thickness. *ACS Photonics*, 2025, vol. 12, no. 3, pp. 1382–1391. https://doi.org/10.1021/acsphotonics.4c01789

Литература

1. Li Y., Monticone F. Exploring the role of metamaterials in achieving advantage in optical computing // Nature Computational Science. 2024. V. 4. N 8. P. 545–548. https://doi.org/10.1038/s43588-024-00657-w
2. McMahon P.L. The physics of optical computing // Nature Reviews Physics. 2023. V. 5. N 12. P. 717–734. https://doi.org/10.1038/s42254-023-00645-5
3. Chanana A., Larocque H., Moreira R., Carolan J., Guha B., Melo E.G., et al. Ultra-low loss quantum photonic circuits integrated with single quantum emitters // Nature Communications. 2022. V. 13. N 1. P. 7693. https://doi.org/10.1038/s41467-022-35332-z
4. Dong T., Liang J.J., Camayd-Muñoz S., Liu Y., Tang H., Kita S., et al. Ultra-low-loss on-chip zero-index materials // Light: Science & Applications. 2021. V. 10. N 1. P. 10. https://doi.org/10.1038/s41377-020-00436-y
5. Blundell S., Radford T.W., Ajia I.A., Lawson D., Yan X.Z., Banakar M., et al. Ultracompact programmable silicon photonics using layers of low-loss phase-change material Sb2Se3 of increasing thickness // ACS Photonics. 2025. V. 12. N 3. P. 1382–1391. https://doi.org/10.1021/acsphotonics.4c01789

6. Bader S.D., Parkin S.S.P. Spintronics. *Annual Review of Condensed Matter Physics*, 2010, vol. 1, pp. 71–88. https://doi.org/10.1146/annurev-conmatphys-070909-104123
7. Pulizzi F. Spintronics. *Nature Materials*, 2012, vol. 11, no. 5, pp. 367. https://doi.org/10.1038/nmat3327
8. Žutić I., Fabian J., Sarma S.D. Spintronics: Fundamentals and applications. *Reviews of Modern Physics*, 2004, vol. 76, no. 2, pp. 323–410. https://doi.org/10.1103/RevModPhys.76.323
9. Qin J., Sun B., Zhou G., Guo T., Chen Y., Ke C., et al. From spintronic memristors to quantum computing. *ACS Materials Letters*, 2023, vol. 5, no. 8. pp. 2197–2215. https://doi.org/10.1021/acsmaterialslett.3c00088
10. Yang H., Valenzuela S.O., Chshiev M., Couet S., Dieny B., Dlubak B., et al. Two-dimensional materials prospects for non-volatile spintronic memories. *Nature*, 2022, vol. 606, no. 7915, pp. 663–673. https://doi.org/10.1038/s41586-022-04768-0
11. Bliokh K.Y., Bekshaev A.Y., Nori F. Dual electromagnetism: helicity, spin, momentum and angular momentum. *New Journal of Physics*, 2013, vol. 15, pp. 033026. https://doi.org/10.1088/1367-2630/15/3/033026
12. Bliokh K.Y., Rodríguez-Fortuño F.J., Nori F., Zayats A.V. Spin–orbit interactions of light. *Nature Photonics*, 2015, vol. 9, no. 12, pp. 796–808. https://doi.org/10.1038/nphoton.2015.201
13. Bliokh K.Y., Nori F. Transverse and longitudinal angular momenta of light. *Physics Reports*, 2015, vol. 592, pp. 1–38. https://doi.org/10.1016/j.physrep.2015.06.003
14. Marrucci L., Manzo C., Paparo D. Optical spin-to-orbital angular momentum conversion in inhomogeneous anisotropic media. *Physical Review Letters*, 2006, vol. 96, no. 16, pp. 163905. https://doi.org/10.1103/PhysRevLett.96.163905
15. Erhard M., Fickler R., Krenn M., Zeilinger A. Twisted photons: new quantum perspectives in high dimensions. *Light: Science & Applications*, 2018, vol. 7, pp. 17146. https://doi.org/10.1038/lsa.2017.146
16. Nagali E., Giovannini D., Marrucci L., Slussarenko S., Santamato E., Sciarrino F. Experimental optimal cloning of four-dimensional quantum states of photons. *Physical Review Letters*, 2010, vol. 105, no. 7, pp. 073602. https://doi.org/10.1103/PhysRevLett.105.073602
17. Deriy I., Kornovan D., Petrov M., Bogdanov A. Optical spintronics: towards optical communication without energy transfer. *arXiv*, 2025, arXiv:2505.10489. https://doi.org/10.48550/arXiv.2505.10489
18. Wu X., Tong L. Optical microfibers and nanofibers. *Nanophotonics*, 2013, vol. 2, no. 5-6, pp. 407–428. https://doi.org/10.1515/nanoph-2013-0033
19. Nayak K.P., Kien F.L., Nakajima K., Miyazaki H.T., Sugimoto Y., Hakuta K. Nano-structured optical nanofibers for cavity-QED. *Proc. of the Conference on Lasers and Electro-Optics (CLEO)*, 2011, pp. QFC2. https://doi.org/10.1364/qels.2011.qfc2
20. Le Kien F., Rauschenbeutel A. Nanofiber-based all-optical switches. *Physical Review A*, 2016, vol. 93, no. 1, pp. 013849. https://doi.org/10.1103/PhysRevA.93.013849
21. Guo J., Liu X., Jiang N., Yetisen A., Yuk H., Yang C., et al. Highly stretchable, strain sensing hydrogel optical fibers. *Advanced Materials*, 2016, vol. 28, no. 46, pp. 10244–10249. https://doi.org/10.1002/adma.201603160
22. Russell P.S.J., Beravat R., Wong G.K.L. Helically twisted photonic crystal fibres. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2017, vol. 375, no. 2087, pp. 20150440. https://doi.org/10.1098/rsta.2015.0440
23. Machnev A.A., Pushkarev A.P., Tonkaev P., Noskov R.E., Rusimova K.R., Mosley P.J., et al. Modifying light–matter interactions with perovskite nanocrystals inside antiresonant photonic crystal fiber. *Photonics Research*, 2021, vol. 9, no. 8, pp. 1462–1469. https://doi.org/10.1364/PRJ.422640
24. Kolchanov D.S., Machnev A., Blank A., Barhom H., Zhu L., Lin Q., et al. Thermo-optics of gilded hollow-core fibers. *Nanoscale*, 2024, vol. 16, no. 29, pp. 13945–13952. https://doi.org/10.1039/D3NR05310E
25. Bialynicki-Birula I., Bialynicka-Birula Z. The role of the Riemann–Silberstein vector in classical and quantum theories of electromagnetism. *Journal of Physics A: Mathematical and Theoretical*, 2013, vol. 46, no. 5, pp. 053001. https://doi.org/10.1088/1751-8113/46/5/053001
26. Belkovich I.V., Kogan B.L. Utilization of Riemann-Silberstein vectors in electromagnetics. *Progress in Electromagnetics Research B*, 2016, vol. 69, no. 1, pp. 103–116. https://doi.org/10.2528/pierb16051809

6. Bader S.D., Parkin S.S.P. Spintronics // Annual Review of Condensed Matter Physics. 2010. V. 1. P. 71–88. https://doi.org/10.1146/annurev-conmatphys-070909-104123
7. Pulizzi F. Spintronics // Nature Materials. 2012. V. 11. N 5. P. 367. https://doi.org/10.1038/nmat3327
8. Žutić I., Fabian J., Sarma S.D. Spintronics: Fundamentals and applications // Reviews of Modern Physics. 2004. V. 76. N 2. P. 323–410. https://doi.org/10.1103/RevModPhys.76.323
9. Qin J., Sun B., Zhou G., Guo T., Chen Y., Ke C., et al. From spintronic memristors to quantum computing // ACS Materials Letters. 2023. V. 5. N 8. P. 2197–2215. https://doi.org/10.1021/acsmaterialslett.3c00088
10. Yang H., Valenzuela S.O., Chshiev M., Couet S., Dieny B., Dlubak B., et al. Two-dimensional materials prospects for non-volatile spintronic memories // Nature. 2022. V. 606. N 7915. P. 663–673. https://doi.org/10.1038/s41586-022-04768-0
11. Bliokh K.Y., Bekshaev A.Y., Nori F. Dual electromagnetism: helicity, spin, momentum and angular momentum // New Journal of Physics. 2013. V. 15. P. 033026. https://doi.org/10.1088/1367-2630/15/3/033026
12. Bliokh K.Y., Rodríguez-Fortuño F.J., Nori F., Zayats A.V. Spin–orbit interactions of light // Nature Photonics. 2015. V. 9. N 12. P. 796–808. https://doi.org/10.1038/nphoton.2015.201
13. Bliokh K.Y., Nori F. Transverse and longitudinal angular momenta of light // Physics Reports. 2015. V. 592. P. 1–38. https://doi.org/10.1016/j.physrep.2015.06.003
14. Marrucci L., Manzo C., Paparo D. Optical spin-to-orbital angular momentum conversion in inhomogeneous anisotropic media // Physical Review Letters. 2006. V. 96. N 16. P. 163905. https://doi.org/10.1103/PhysRevLett.96.163905
15. Erhard M., Fickler R., Krenn M., Zeilinger A. Twisted photons: new quantum perspectives in high dimensions // Light: Science & Applications. 2018. V. 7. P. 17146. https://doi.org/10.1038/lsa.2017.146
16. Nagali E., Giovannini D., Marrucci L., Slussarenko S., Santamato E., Sciarrino F. Experimental optimal cloning of four-dimensional quantum states of photons // Physical Review Letters. 2010. V. 105. N 7. P. 073602. https://doi.org/10.1103/PhysRevLett.105.073602
17. Deriy I., Kornovan D., Petrov M., Bogdanov A. Optical spintronics: towards optical communication without energy transfer // arXiv. 2025. arXiv:2505.10489. https://doi.org/10.48550/arXiv.2505.10489
18. Wu X., Tong L. Optical microfibers and nanofibers // Nanophotonics. 2013. V. 2. N 5-6. P. 407–428. https://doi.org/10.1515/nanoph-2013-0033
19. Nayak K.P., Kien F.L., Nakajima K., Miyazaki H.T., Sugimoto Y., Hakuta K. Nano-structured optical nanofibers for cavity-QED // Proc. of the Conference on Lasers and Electro-Optics (CLEO). 2011. P. QFC2. https://doi.org/10.1364/qels.2011.qfc2
20. Le Kien F., Rauschenbeutel A. Nanofiber-based all-optical switches // Physical Review A. 2016. V. 93. N 1. P. 013849. https://doi.org/10.1103/PhysRevA.93.013849
21. Guo J., Liu X., Jiang N., Yetisen A., Yuk H., Yang C., et al. Highly stretchable, strain sensing hydrogel optical fibers // Advanced Materials. 2016. V. 28. N 46. P. 10244–10249. https://doi.org/10.1002/adma.201603160
22. Russell P.S.J., Beravat R., Wong G.K.L. Helically twisted photonic crystal fibres // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2017. V. 375. N 2087. P. 20150440. https://doi.org/10.1098/rsta.2015.0440
23. Machnev A.A., Pushkarev A.P., Tonkaev P., Noskov R.E., Rusimova K.R., Mosley P.J., et al. Modifying light–matter interactions with perovskite nanocrystals inside antiresonant photonic crystal fiber // Photonics Research. 2021. V. 9. N 8. P. 1462–1469. https://doi.org/10.1364/PRJ.422640
24. Kolchanov D.S., Machnev A., Blank A., Barhom H., Zhu L., Lin Q., et al. Thermo-optics of gilded hollow-core fibers // Nanoscale. 2024. V. 16. N 29. P. 13945–13952. https://doi.org/10.1039/D3NR05310E
25. Bialynicki-Birula I., Bialynicka-Birula Z. The role of the Riemann–Silberstein vector in classical and quantum theories of electromagnetism // Journal of Physics A: Mathematical and Theoretical. 2013. V. 46. N 5. P. 053001. https://doi.org/10.1088/1751-8113/46/5/053001
26. Belkovich I.V., Kogan B.L. Utilization of Riemann-Silberstein vectors in electromagnetics // Progress in Electromagnetics Research B. 2016. V. 69. N 1. P. 103–116. https://doi.org/10.2528/pierb16051809
27. Snyder A.W., Love J. Optical Waveguide Theory. Chapman and Hall, 1983. 746 p.

27. Snyder A.W., Love J. *Optical Waveguide Theory*. Chapman and Hall, 1983, 746 p.
28. Seaborn J.B. *Hypergeometric Functions and Their Applications*. Springer, 2013, 268 p.

28. Seaborn J.B. Hypergeometric Functions and Their Applications. Springer, 2013. 268 p.

Authors

Ilya A. Deriy — Junior Researcher, Harbin Engineering University, Qingdao, 266000, China; Junior Reseacher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57221052856, https://orcid.org/0000-0002-6515-9325, ilya.deriy@metalab.ifmo.ru

Danil F. Kornovan — PhD (Physics & Mathematics), Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56644703300, https://orcid.org/0000-0002-4851-0697, d.kornovan@metalab.ifmo.ru

Mihail I. Petrov — PhD (Physics & Mathematics), Associate Professor, Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, Researcher ID K-5924-2012, https://orcid.org/0000-0001-8155-9778, m.petrov@metalab.ifmo.ru

Andrey A. Bogdanov — PhD (Physics & Mathematics), Senior Researcher, Harbin Engineering University, Qingdao, 266000, China; Senior Researcher, Associate Professor of Practice, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56393877900, https://orcid.org/0000-0002-8215-0445, a.bogdanov@metalab.ifmo.ru

Авторы

Дерий Илья Александрович — младший научный сотрудник, Харбинский Инженерный Университет, Циндао, 266000, Китай; младший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57221052856, https://orcid.org/0000-0002-6515-9325, ilya.deriy@metalab.ifmo.ru

Корновaн Данил Феодосьевич — кандидат физико-математических наук, инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56644703300, https://orcid.org/0000-0002-4851-0697, d.kornovan@metalab.ifmo.ru

Петров Михаил Игоревич — кандидат физико-математических наук, доцент, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Researcher ID K-5924-2012, https://orcid.org/0000-0001-8155-9778, m.petrov@metalab.ifmo.ru

Богданов Андрей Андреевич — кандидат физико-математических наук, старший научный сотрудник, Харбинский Инженерный Университет, Циндао, 266000, Китай; старший научный сотрудник, доцент практики, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56393877900, https://orcid.org/0000-0002-8215-0445, a.bogdanov@metalab.ifmo.ru

Received 24.07.2025
Approved after reviewing 18.08.2025
Accepted 15.09.2025

Статья поступила в редакцию 24.07.2025
Одобрена после рецензирования 18.08.2025
Принята к печати 15.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-817-824

Geometric modeling and compensation of cutting tool positioning errors for eliminating protrusion in large-radius spherical surface machining

Muhamad Albani Rizki[1]✉, Yuri V. Fedosov[2], Maxim Y. Afanasiev[3], Anastasia A. Krylova[4]

[1,2,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] muhamadalbanirizki@gmail.com✉, https://orcid.org/0000-0001-7502-1699
[2] Yf01@yandex.ru, https://orcid.org/0000-0003-1869-0081
[3] amax@niuitmo.ru, https://orcid.org/0000-0003-4061-1407
[4] ananasn94@gmail.com, https://orcid.org/0000-0002-5822-6702

Abstract

The production of optical components with a large radius of spherical surfaces requires exceptionally high surface profile accuracy. Minor deviations in the positioning of the cutting tool caused by factors, such as mechanical backlash, thermal deformation, and incorrect tool positioning, can result in dimensional errors of the machined surface, particularly in the form of protrusions that indicate processing defects. Despite a wide range of studies focused on tool wear and general machining errors, insufficient attention has been given to the geometric modeling and correction of defects caused by tool positioning errors. This study presents a comprehensive approach to geometrically modeling the impact of cutting tool positioning errors on the machined surface profile. A mathematical model has been developed to model the interaction between the tool and the spherical surface, enabling precise estimation of the radial machining error. Based on these data, a new error compensation method is proposed, allowing for the correction of errors by modifying the tool movement trajectory. The proposed model accurately predicts the formation and characteristics of protrusions resulting from tool displacement during the machining of spherical surfaces with a large radius. Implementation of the compensation method significantly reduces the defect rate, improves geometric accuracy, and decreases the need for additional processing. Addressing defects caused by positioning errors enables the proposal of a new method that has not previously been considered in precision machining research. The proposed model and tool positioning error compensation method offer an effective and practical solution for improving the surface profile accuracy of optical components, thereby enhancing the precision and efficiency of manufacturing processes. The proposed method contributes to the advancement of high-precision optical component manufacturing with minimal post-processing costs, providing a novel approach in the fields of instrument engineering and precision mechanical engineering.

Keywords

machining errors, tool wear, precision manufacturing, concave surface milling, calibration techniques, mathematical modeling

Acknowledgements

The work was supported by the Ministry of Science and Higher Education of the Russian Federation, Agreement No. 075-11-2023-015, 10.02.2023, “Creation of high-tech serial production of energy-efficient synchronous electric motors with integrated intelligent position sensor and self-diagnosis functions for robotics and digital automation systems”.

For citation: Rizki M.A., Fedosov Yu.V., Afanasiev M.Y., Krylova A.A. Geometric modeling and compensation of cutting tool positioning errors for eliminating protrusion in large-radius spherical surface machining. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 817–824. doi: 10.17586/2226-1494-2025-25-5-817-824

УДК 681.7.023.42

Геометрическое моделирование и компенсация ошибок позиционирования режущего инструмента для устранения выступов при обработке оптических поверхностей с большим радиусом сферы

Мухамад Албани Ризки[1]✉, Юрий Валерьевич Федосов[2], Максим Яковлевич Афанасьев[3], Анастасия Андреевна Крылова[4]

[1,2,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] muhamadalbanirizki@gmail.com✉, https://orcid.org/0000-0001-7502-1699
[2] Yf01@yandex.ru, https://orcid.org/0000-0003-1869-0081
[3] amax@niuitmo.ru, https://orcid.org/0000-0003-4061-1407
[4] ananasn94@gmail.com, https://orcid.org/0000-0002-5822-6702

Аннотация

Введение. Производство оптических компонентов с большими радиусами сферических поверхностей требует исключительно высокой точности профиля поверхности. Незначительные отклонения в позиционировании режущего инструмента, вызванные такими факторами, как механический люфт, тепловая деформация и ошибочное позиционирование инструмента, могут привести к ошибке в размере обработанной поверхности — в частности, выступам, которые свидетельствуют об ошибке обработки. Несмотря на широкий спектр исследований, посвященных износу инструмента и общим ошибкам обработки, недостаточно внимания уделено геометрическому моделированию и коррекции дефектов, вызванных ошибками позиционирования инструмента. **Метод.** Представлен комплексный подход к геометрическому моделированию влияния ошибок позиционирования режущего инструмента на профиль обработанной поверхности. Разработана математическая модель, имитирующая взаимодействие инструмента и сферической поверхности, что позволяет точно оценить ошибку радиального размера обработки. На основе этих данных предложен новый метод компенсации ошибки, позволяющий корректировать ошибки путем изменения траектории перемещения инструмента. **Основные результаты.** Предложенная модель с высокой точностью предсказывает формирование и характеристики выступов, возникающих вследствие смещения инструмента при обработке сферических поверхностей с большим радиусом. Внедрение метода компенсации существенно снижает количество брака, улучшая геометрическую точность и уменьшая потребность в дополнительной обработке. **Обсуждение.** Рассмотрение дефектов, вызванных ошибками позиционирования, позволяет предложить новый метод, ранее не рассмотренный в исследованиях по точной обработке. Предлагаемая модель и метод компенсации погрешности позиционирования инструмента демонстрируют эффективное и практическое решение для повышения точности профиля оптических компонентов, что способствует увеличению точности и эффективности производственных процессов. Разработанный метод вносит вклад в развитие высокоточного производства оптических изделий с минимальными затратами на постобработку, обеспечивая новый подход в области приборостроения и прецизионного машиностроения.

Ключевые слова

погрешности обработки, износ инструмента, высокоточное производство, фрезерование вогнутых поверхностей, методы калибровки, математическое моделирование

Благодарности

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации, соглашение № 075-11-2023-015 от 10.02.2023, «Создание высокотехнологичного серийного производства энергоэффективных синхронных электродвигателей со встроенным интеллектуальным датчиком положения и функциями самодиагностики для робототехники и цифровых систем автоматизации»

Ссылка для цитирования: Ризки М.А., Федосов Ю.В., Афанасьев М.Я., Крылова А.А. Геометрическое моделирование и компенсация ошибок позиционирования режущего инструмента для устранения выступов при обработке оптических поверхностей с большим радиусом сферы // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 817–824 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-817-824

Introduction

The instrumentation and precision manufacturing industries have witnessed remarkable advancements in recent years, primarily driven by the increasing demand for ultra-high accuracy components in fields, such as optics, aerospace, and microelectronics. Within these domains, particularly in the fabrication of optical components, geometric precision and surface integrity, are paramount. Optical elements, such as concave mirrors, spherical lenses, and freeform surfaces, must exhibit strict dimensional fidelity and surface smoothness to achieve their intended optical performance. One of the persistent and critical challenges in achieving such accuracy lies in the occurrence of cutting tool positioning errors which introduce localized surface anomalies such as protrusions and profile distortions, especially in the machining of large-radius spherical surfaces.

Cutting tool positioning error refers to the deviation of the actual tool location from its programmed position during the machining process. These deviations can arise from a variety of sources, including mechanical backlash, thermal deformation, encoder inaccuracies, servo lag, tool-holder deflection, or misalignment during setup. In multi-axis Computer Numerical Control (CNC) machining environments, such errors become increasingly impactful,

particularly when the contact geometry between the tool and workpiece is sensitive to small variations in tool orientation or position. This sensitivity is magnified in large-radius spherical surfaces where the surface curvature is subtle — thus any slight misalignment of the cutting edge can result in geometrically significant protrusions or irregularities that degrade the optical quality of the component.

In modern manufacturing systems, closed-loop machine tools and precision equipment are often employed to mitigate such errors. However, even high-end imported machinery suffers from degradation over time, limited setup flexibility, and environmental influences such as thermal fluctuations, which further contribute to mispositioning. These conditions are exacerbated in complex geometries like concave or freeform surfaces where the spatial tool–surface interaction is nontrivial and continuously changing. Therefore, accurately modeling, detecting, and compensating for tool positioning errors becomes crucial for ensuring the desired surface profile and minimizing costly rework or post-processing.

Numerous studies have addressed the general problem of machining errors, surface roughness, and tool wear in freeform and precision surface manufacturing. For instance, in [1] introduced a surface-shape-based method for finishing freeform surfaces, optimizing local finishing performance but assuming ideal tool positioning. Likewise, in [2] modeled the topology of freeform surfaces generated by ball-end milling, providing critical insight into surface generation mechanics but not explicitly accounting for spatial tool mispositioning. The sensitivity of optical components to surface errors was further emphasized in [3], which demonstrated the design and fabrication of concave lens arrays on aspheric curved surfaces — highlighting the need for high geometric fidelity throughout the process.

In [4] focused on the surface roughness variations during concave surface milling of AISI 420 steel, exploring how process parameters affect surface texture, though again under the assumption of ideal tool trajectory. A more direct link to positioning control is found in [5] which proposed a double-point contact tool positioning method using a barrel cutter for concave surfaces, thereby enhancing surface conformity and reducing form error. Similarly, in [6] investigated the influence of tool path strategies and cutting parameters on tool deflection and cutting forces — factors that directly impact the spatial accuracy of tool positioning.

Advances in CNC programming and specialized tooling have also contributed to reducing form errors. Research [7, 8] introduced precision machining methods for concave-arc and cone-end milling cutters respectively, offering improved tool profiles that can reduce sensitivity to positional deviations. Additionally, in [9] explored surface micro-textures to enhance anti-adhesion properties, indirectly contributing to better tool stability and reduced deviation during dry cutting conditions. In [10] presented CNC form milling for concave-convex arc-line gears, underscoring the precision demands in gear tooth profile machining.

From a modeling perspective, mathematical approaches have been used to predict and correct machining deviations. In [11] presented a model for machining internal double-arc spiral bevel gears using finger milling cutters, integrating geometric and kinematic considerations relevant to positioning accuracy. In [12] laid out foundational knowledge in optics fabrication and precision machining methodologies, establishing key metrics for surface quality and geometric error evaluation. Furthermore, in [13] examined geometrical errors in surfaces produced using concave and convex profile tools, highlighting the need for precise tool placement during milling. Research [14] provided a comprehensive review of polishing techniques for optical parts, emphasizing the need for minimal surface error prior to finishing. More recently, in [15] investigated the influence of tool structure and cutting orientation on force generation, offering insight into the positional stability of double-arc milling cutters under dynamic load conditions.

Despite these advances, there remains a significant gap in the literature regarding the explicit modeling and compensation of tool positioning errors as a source of localized protrusions or profile distortions, especially in the context of large-radius spherical surface machining. Traditional error compensation methods tend to focus on general tool wear, thermal effects, or vibration-related deviations, and often assume a predictable, uniform error distribution. However, positioning errors are frequently non-uniform and manifest as localized surface features which are difficult to detect in-line and are even harder to remove through polishing or re-machining without compromising the global geometry.

To bridge this gap, the present study proposes a geometric modeling framework and compensation strategy specifically targeting surface deviations caused by cutting tool positioning errors. The focus is on the manufacturing of optical parts with large-radius spherical surfaces, where the tool-surface interaction is highly susceptible to positional misalignments. A detailed mathematical model is developed to analyze how deviations in tool position translate into profile errors, particularly protrusions that are often observed in post-process inspection of high-precision optical components.

The model incorporates parameters, such as tool geometry, intended tool path, surface curvature, and local tool orientation. Furthermore, a compensation strategy is proposed that can either adjust the tool path in the Computer-Aided Manufacturing software prior to machining or guide post-process corrective actions to eliminate the resulting surface defects.

Experimental observation of protrusion formation

Precision machining of large radius spherical surfaces requires strict control over cutting tool positioning, as even minor deviations can lead to significant geometric defects. A critical issue observed in the milling of concave optical components, particularly on the LOH300 machine, is the formation of a central protrusion, a defect directly linked to tool alignment errors and path inaccuracies. This protrusion degrades both the geometric accuracy and optical functionality of the machined surfaces.

In the standard setup, the workpiece rotates on a table while the inclined milling cutter feeds vertically,

with longitudinal adjustments refining the cutting path. This configuration enables high precision spherical concavities, but its success hinges on exact tool positioning. Deviations in the cutter inclination angle, vertical descent, or longitudinal displacement disrupt the intended tool workpiece interaction, often resulting in a residual central protrusion. Such errors are exacerbated in large radius machining where slight misalignments produce amplified surface distortion.

To visualize the setup, consider the schematic shown in Fig. 1, which illustrates the spatial arrangement of the milling cutter and the rotating workpiece. As shown, the inclined cutter rotating edge intersects with the workpiece surface to remove material, thereby generating the concave spherical profile.

To better understand the material removal process, a cross-sectional analysis of the cutter and workpiece interaction is presented in Fig. 2. The cutting edge of the milling tool traces a spatial path that forms a segment of a toroidal surface. This toroidal surface intersects with the spherical volume of the rotating workpiece. Material removal occurs at the intersection point, generating a concave surface whose curvature is a function of both cutter geometry and motion parameters.

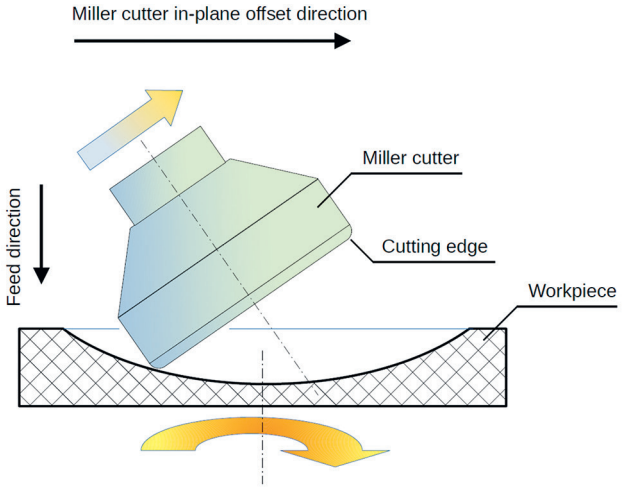

Fig. 1. Schematic of the milling setup

The cutter is capable of longitudinal and angular movement relative to the rotation axis of the workpiece. Errors in these movements can result in undercutting, manifesting as a protrusion at the center of the concave surface.

The physical manifestation of this defect is shown in Fig. 3. A clear central protrusion appears at the center of the machined spherical surface, resulting from misalignment in the cutter path relative to the workpiece axis. This geometrical error distorts the intended concave profile and is especially problematic in precision applications like optical components.

After identifying the source of the error, corrective measures were applied to adjust the cutter inclination angle and ensure proper longitudinal feed displacement. The result of these adjustments is presented in Fig. 4.

As seen in Fig. 4, the protrusion has been completely removed. The surface now conforms to the desired spherical geometry, confirming that precise tool alignment and feed optimization are essential for defect-free machining.

This real-world observation provides context for the geometric behavior illustrated in the following schematic, which conceptually models the cutter misalignment and its consequence on surface geometry.

Geometric modeling of cutter misalignment effects

The phenomenon of central protrusion is further analyzed through geometric modeling. Fig. 5 depicts how cutter misalignment during the milling process results in a protrusion forming at the center of the machined surface.

To mathematically describe the protrusion formation, the motion of the cutter edge relative to the workpiece is modeled using hypocycloidal geometry which can be described by the following parametric equation:

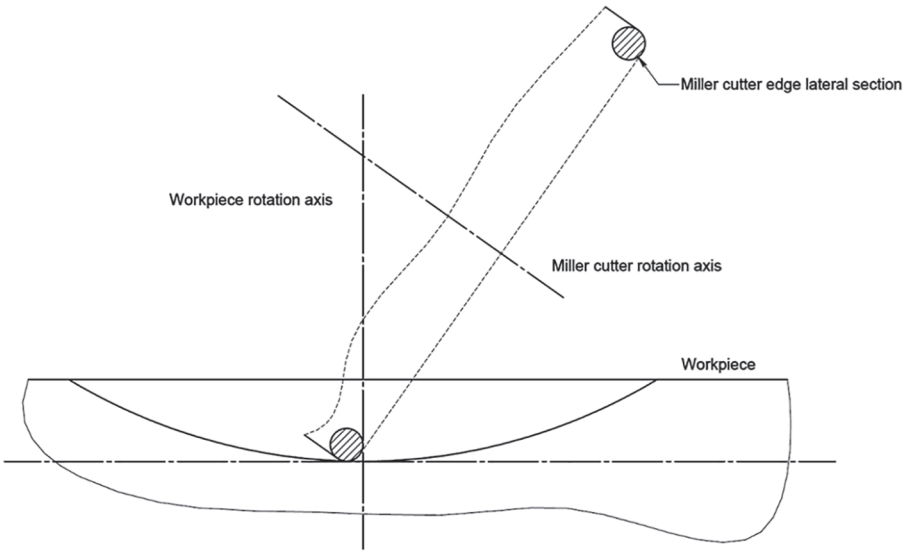

Fig. 2. Cross-section of the milling cutter and workpiece during processing

$$\begin{cases} x = (R_L - R_S)\cos\beta + R_S\cos\dfrac{R_L - R_S}{R_S}\beta \\ y = (R_L - R_S)\sin\beta - R_S\sin\dfrac{R_L - R_S}{R_S}\beta \end{cases}, \quad (1)$$

where R_L is the radius of the larger circle, R_S is the radius of the smaller circle, and β is the rotation angle of the smaller circle.

When a circle with radius R_S rolls along the inner side of a circle with radius R_L by an angle β, point A, located on the circumference of the smaller circle, moves to point A', according to the equation (1) the coordinates change as the following equation:

$$A(\beta) = \left((R_L - R_S)\cos\beta + R_S\cos\frac{R_L - R_S}{R_S}\beta, \right.$$
$$\left. (R_L - R_S)\sin\beta - R_S\sin\frac{R_L - R_S}{R_S}\beta \right). \quad (2)$$

Fig. 3. Machined part before correction — central protrusion clearly visible

Fig. 4. Machined part after proper tool alignment — protrusion successfully removed

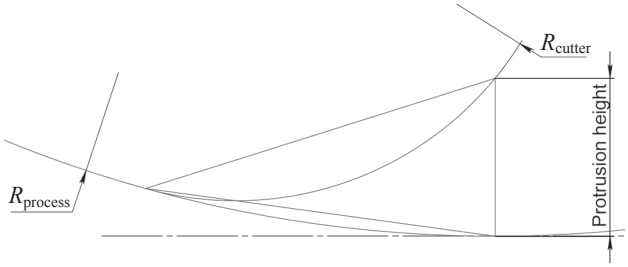

Fig. 5. The protrusion formed during the milling process; $R_{process}$ is desired radius of the machined spherical surface; R_{cutter} is the radius of the milling cutter tip used for machining

This motion is conceptually depicted in Fig. 6, where the smaller circle rolls inside the larger circle and traces a path via a fixed point on its circumference.

The central protrusion can also be analyzed geometrically as shown in Fig. 7. A larger circle with R_L and center at point O represents the reference spherical surface. The smaller circle rolls along the inner circumference from left to right. A vertical line OB intersects the lower point of the larger circle. Point A marked on the smaller circle rolls and eventually coincides with point B on line OB. The straight-line AO connecting point A to the center O change its length during this motion, ranging from $R_L - R_S$ to R_L.

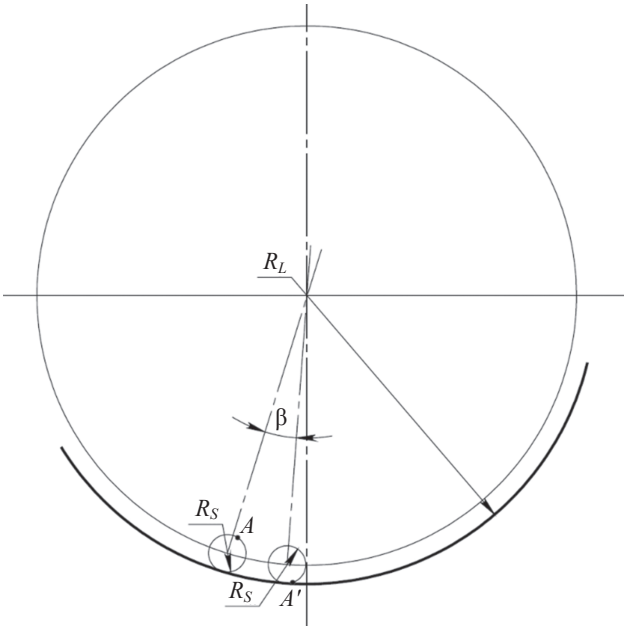

Fig. 6. Rolling of a smaller-radius circle along the inner circumference of a larger-radius circle

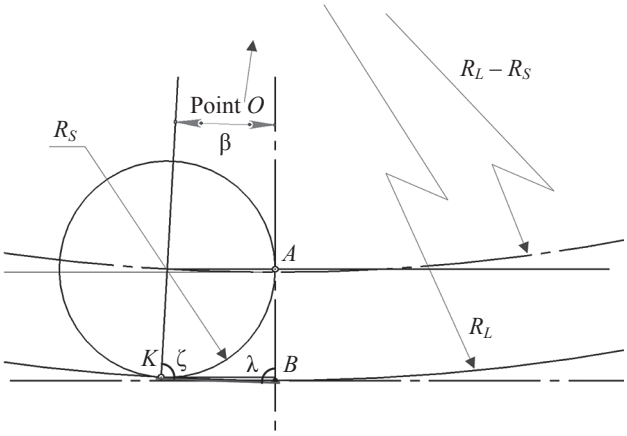

Fig. 7. The rolling of the milling cutter onto the workpiece (point O is outside the drawing); K is a point of tangency between the milling cutter and the machined surface profile; ζ is the angle formed at point K between the milling cutter radius and the machined surface; λ is angle at point B between the machined surface profile and the horizontal reference (axis)

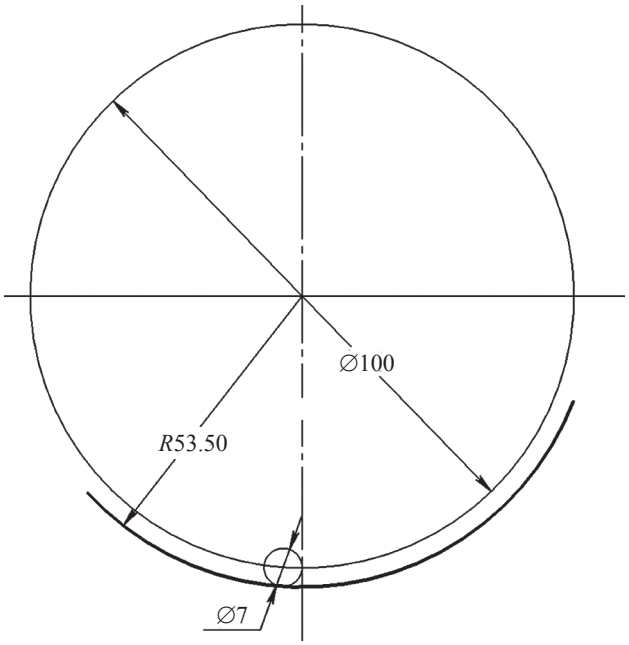

Fig. 8. Cross-section of the milling cutter and the workpiece (diameter 100 mm refers to the cutting radius center point)

The length of the line segment AO can be calculated as a function of angle β:

$$AO(\beta) = \sqrt{(x_A - x_O)^2 + (y_A - y_O)^2}, \quad (3)$$

where (x_O, y_O) is the coordinate of point O, which are both zero; (x_A, y_A) is the coordinate of point, which were determined in equation (2). By substituting the values from equation (2) into equation (3), the full expression becomes:

$$AO(\beta) = \sqrt{\left((R_L - R_S)\cos\beta + R_S\cos\frac{R_L - R_S}{R_S}\beta\right)^2 + \left((R_L - R_S)\sin\beta - R_S\sin\frac{R_L - R_S}{R_S}\beta\right)^2}. \quad (4)$$

Geometric analysis of tool-surface interaction in spherical machining

To validate the analytical model and gain visual insight into the tool movement and material engagement, a computation study was conducted using numerical modeling. The objective of the model was to replicate the real-world conditions of the machining process and examine the interaction between the cutter and the workpiece, especially in the region near the axis of rotation.

The analysis modeled a two-dimensional cross-section of the cutter and the workpiece, capturing the exact geometry of the toroidal tool and the spherical cavity. Parameters used in the model included:

— Cutter radius R_S = 3.5 mm;
— Workpiece radius R_L = 53.5 mm.

This setup is visualized in Fig. 8.

A key part of the modeling involved determining how far the cutter needed to move to ensure full coverage of the spherical surface. This was analyzed by constructing an arc that represented the necessary rolling distance of the cutter. Fig. 9 shows this construction which defines the chord length and the corresponding arc required for the tool edge to move from the periphery to the central region.

Using this setup, the angle β corresponding to the arc length was measured to be 6°09' (for substitution into the hypocycloid model 0.1073 rad). Substituting angle value into equation (4), precise distance from the cutter edge to the workpiece center during operation is computed

Evaluation of protrusion formation based on tool path error analysis

The computational analysis results validated the theoretical prediction that a hypocycloidal cutter trajectory may produce a central protrusion if the longitudinal displacement is not properly aligned. Specifically, the computed value of $AO(\beta)$, using equation (4) with β = 0.1073 rad, yielded a distance of 50.871 mm. This result closely approximates the target spherical radius R_L = 53.5 mm, thereby demonstrating the high accuracy and practical relevance of the analytical model in the context of precision machining.

The close agreement between model-based results and analytical predictions confirms the suitability of hypocycloidal trajectory modeling for analyzing and optimizing spherical surface machining. In particular, the visual outputs of the computational model clearly illustrated that insufficient longitudinal feed, tool inclination error, or tool positioning inaccuracies can result in undercutting or surface protrusions at the center of the machined region. Such deviations highlight the sensitivity of the process to even minor misalignments in the cutter path.

Beyond validating the analytical formulation, the computational modeling process also provided valuable insights into machining process optimization. It was shown that increasing the longitudinal feed displacement or adjusting the cutter inclination angle can ensure full engagement between the tool and the spherical surface. Moreover, minimizing tool positioning errors — both in the feed direction and angular alignment — is essential

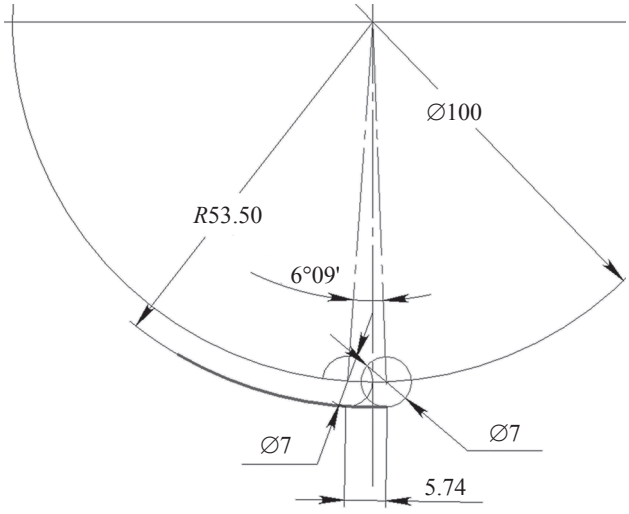

Fig. 9. Construction of the arc of the required length

to maintain geometric accuracy and prevent surface defects. These measures help improve surface finish, reduce dimensional deviations, and minimize the need for corrective rework.

These findings have direct implications for high-precision machining applications, particularly in areas such as optical component manufacturing where sub-micron accuracy and flawless surface integrity are critical. Incorporating the hypocycloidal trajectory model into machining strategies enables accurate prediction and compensation for potential deviations caused by tool misalignment or positioning errors. The strong correlation between the analytical model and computed results confirms that this approach can significantly enhance machining reliability and accuracy, especially for components with large-radius spherical geometries.

Conclusion

This study presented a comprehensive investigation into the geometric modeling and compensation of cutting tool positioning errors in the machining of large-radius spherical surfaces. Starting from experimental observations of central protrusion defects, the work identified that minor misalignments in the tool inclination or longitudinal feed can lead to significant surface deviations, particularly in applications requiring sub-micron geometric fidelity such as optical components. A mathematical framework based on hypocycloidal geometry was developed to describe the cutter motion relative to the spherical surface. This model accurately predicted the formation and characteristics of protrusions and was supported through geometric constructions reflecting real-world tool-surface interaction. The consistency between analytical predictions and modeled tool paths confirmed the effectiveness of the proposed approach. Beyond theoretical formulation, the study demonstrated that adjusting tool inclination and feed displacement, guided by the geometric model, effectively eliminates protrusions and restores the intended surface geometry. These findings were reinforced by corrected machining outcomes, where the surface profile was restored after applying model-based adjustments. By addressing a specific and underexplored class of machining errors and localized defects caused by the cutter mispositioning, this research provides a robust predictive model and compensation strategy. The method is both practical and adaptable, offering a valuable approach for improving surface integrity, reducing post-processing and enhancing reliability in ultra-precision manufacturing environments.

References

1. Bey M., Bendifallah M., Kader S., Boukhalfa K. A new approach for finishing free-form surfaces based on local shapes. *International Journal of Computer Integrated Manufacturing*, 2014, vol. 27, no. 9, pp. 840–857. https://doi.org/10.1080/0951192x.2013.838323
2. Vyboishchik A.V. Modelling topology of freeform surfaces with ball-end milling. *Procedia Engineering*, 2016, vol. 150, pp. 761–767. https://doi.org/10.1016/j.proeng.2016.07.103
3. Mo J., Chang X., Renqing D., Zhang J., Liao L., Luo S. Design, fabrication, and performance evaluation of a concave lens array on an aspheric curved surface. *Optics Express*, 2022, vol. 30, no. 18, pp. 33241–33258. https://doi.org/10.1364/oe.471055
4. Juiña L.C., Dávalos E.J., Landazurí D.S., Guaño S.E., Moreno N.V. Roughness analysis of a concave surface as a function of machining parameters and strategies for AISI 420 steel. *Materials Today: Proceedings*, 2022, vol. 49, part 1. https://doi.org/10.1016/j.matpr.2021.07.477
5. Yu Z., Zhi-Tong C., Yun Z., Tao N. Tool positioning method for achieving double-point contact in flank milling of a concave surface with a barrel cutter. *International Journal of Advanced Manufacturing Technology*, 2017, vol. 93, no. 5-8. P. 1791–1807. https://doi.org/10.1007/s00170-017-0472-1
6. Gok A., Gologlu C., Demirci H.I. Cutting parameter and tool path style effects on cutting force and tool deflection in machining of convex and concave inclined surfaces. *International Journal of Advanced Manufacturing Technology*, 2013, vol. 69, no. 5-8, pp. 1063–1078. https://doi.org/10.1007/s00170-013-5075-x
7. Chen W.F., Lai H.Y., Chen C.K. Design and NC machining of concave-arc ball-end milling cutters. *International Journal of Advanced Manufacturing Technology*, 2002, vol. 20, no. 3, pp. 169–179. https://doi.org/10.1007/s001700200140
8. Chen W.F., Lai H.Y., Chen C.K. A precision tool model for concave cone-end milling cutters. *International Journal of Advanced Manufacturing Technology*, 2001, vol. 18, no. 8, pp. 567–578. https://doi.org/10.1007/s001700170033
9. Kang Z., Fu Y., Chen Y., Ji J., Fu H., Wang S., Li R. Experimental investigation of concave and convex micro-textures for improving anti-adhesion property of cutting tool in dry finish cutting. *International Journal of Precision Engineering and Manufacturing - Green Technology*, 2018, vol. 5, no. 5, pp. 583–591. https://doi.org/10.1007/s40684-018-0060-3

Литература

1. Bey M., Bendifallah M., Kader S., Boukhalfa K. A new approach for finishing free-form surfaces based on local shapes // International Journal of Computer Integrated Manufacturing. 2014. V. 27. N 9. P. 840–857. https://doi.org/10.1080/0951192x.2013.838323
2. Vyboishchik A.V. Modelling topology of freeform surfaces with ball-end milling // Procedia Engineering. 2016. V. 150. P. 761–767. https://doi.org/10.1016/j.proeng.2016.07.103
3. Mo J., Chang X., Renqing D., Zhang J., Liao L., Luo S. Design, fabrication, and performance evaluation of a concave lens array on an aspheric curved surface // Optics Express. 2022. V. 30. N 18. P. 33241–33258. https://doi.org/10.1364/oe.471055
4. Juiña L.C., Dávalos E.J., Landazurí D.S., Guaño S.E., Moreno N.V. Roughness analysis of a concave surface as a function of machining parameters and strategies for AISI 420 steel // Materials Today: Proceedings. 2022. V. 49. Part 1. https://doi.org/10.1016/j.matpr.2021.07.477
5. Yu Z., Zhi-Tong C., Yun Z., Tao N. Tool positioning method for achieving double-point contact in flank milling of a concave surface with a barrel cutter // International Journal of Advanced Manufacturing Technology. 2017. V. 93. N 5-8. P. 1791–1807. https://doi.org/10.1007/s00170-017-0472-1
6. Gok A., Gologlu C., Demirci H.I. Cutting parameter and tool path style effects on cutting force and tool deflection in machining of convex and concave inclined surfaces // International Journal of Advanced Manufacturing Technology. 2013. V. 69. N 5-8. P. 1063–1078. https://doi.org/10.1007/s00170-013-5075-x
7. Chen W.F., Lai H.Y., Chen C.K. Design and NC machining of concave-arc ball-end milling cutters // International Journal of Advanced Manufacturing Technology. 2002. V. 20. N 3. P. 169–179. https://doi.org/10.1007/s001700200140
8. Chen W.F., Lai H.Y., Chen C.K. A precision tool model for concave cone-end milling cutters // International Journal of Advanced Manufacturing Technology. 2001. V. 18. N 8. P. 567–578. https://doi.org/10.1007/s001700170033
9. Kang Z., Fu Y., Chen Y., Ji J., Fu H., Wang S., Li R. Experimental investigation of concave and convex micro-textures for improving anti-adhesion property of cutting tool in dry finish cutting // International Journal of Precision Engineering and Manufacturing - Green Technology. 2018. V. 5. N 5. P. 583–591. https://doi.org/10.1007/s40684-018-0060-3

10. Chen Y. Yao L. Study on a method of CNC form milling for the concave convex arc line gear. *International Journal of Advanced Manufacturing Technology*, 2018, vol. 99, no. 9-12. pp. 2327–2339. https://doi.org/10.1007/s00170-018-2566-9
11. Zhang D., Wang Z., Yao L., Xie D. Mathematical modeling and machining of the internal double-arc spiral bevel gear by finger milling cutters for the nutation drive mechanism. *Machines*, 2022, vol. 10, no. 8, pp. 663. https://doi.org/10.3390/machines10080663
12. Schwertz K. An introduction to the optics manufacturing process. OptoMechanics (OPTI 521) Report, 2008.
13. Fomin A.A., Gusev V.G., Sattarova Z.G. Geometrical errors of surfaces milled with convex and concave profile tools. *Solid State Phenomena*, 2018, vol. 284, pp. 281–288. https://doi.org/10.4028/www.scientific.net/SSP.284.281
14. Xie M., Pan Y., An Z., Huang S., Dong M. Review on surface polishing methods of optical parts. *Advances in Materials Science and Engineering*, 2022, vol. 2022, pp. 8723269. https://doi.org/10.1155/2022/8723269
15. Fan M., Bi C., Liu X., Yue C., Hu D. Effects of tool structure factor, cutting orientations, and cutting parameters of double-arc milling cutter on cutting force. *International Journal of Advanced Manufacturing Technology*, 2024, vol. 134, no. 9-10, pp. 4701–4716. https://doi.org/10.1007/s00170-024-14375-0

10. Chen Y. Yao L. Study on a method of CNC form milling for the concave convex arc line gear // International Journal of Advanced Manufacturing Technology. 2018. V. 99. N 9-12. P. 2327–2339. https://doi.org/10.1007/s00170-018-2566-9
11. Zhang D., Wang Z., Yao L., Xie D. Mathematical modeling and machining of the internal double-arc spiral bevel gear by finger milling cutters for the nutation drive mechanism // Machines. 2022. V. 10. N 8. P. 663. https://doi.org/10.3390/machines10080663
12. Schwertz K. An introduction to the optics manufacturing process // OptoMechanics (OPTI 521) Report. 2008.
13. Fomin A.A., Gusev V.G., Sattarova Z.G. Geometrical errors of surfaces milled with convex and concave profile tools // Solid State Phenomena. 2018. V. 284. P. 281–288. https://doi.org/10.4028/www.scientific.net/SSP.284.281
14. Xie M., Pan Y., An Z., Huang S., Dong M. Review on surface polishing methods of optical parts // Advances in Materials Science and Engineering. 2022. V. 2022. P. 8723269. https://doi.org/10.1155/2022/8723269
15. Fan M., Bi C., Liu X., Yue C., Hu D. Effects of tool structure factor, cutting orientations, and cutting parameters of double-arc milling cutter on cutting force // International Journal of Advanced Manufacturing Technology. 2024. V. 134. N 9-10. P. 4701–4716. https://doi.org/10.1007/s00170-024-14375-0

Authors

Muhamad Albani Rizki — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 58038476200, https://orcid.org/0000-0001-7502-1699, muhamadalbanirizki@gmail.com
Yuri V. Fedosov — PhD, Head of Laboratory, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57194080548, https://orcid.org/0000-0003-1869-0081, Yf01@yandex.ru

Maxim Y. Afanasiev — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57194081345, https://orcid.org/0000-0003-4061-1407, amax@niuitmo.ru

Anastasia A. Krylova — PhD, Lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-5822-6702, ananasn94@gmail.com

Авторы

Ризки Мухамад Албани — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 58038476200, https://orcid.org/0000-0001-7502-1699, muhamadalbanirizki@gmail.com
Федосов Юрий Валерьевич — кандидат технических наук, заведующий лабораторией, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57194080548, https://orcid.org/0000-0003-1869-0081, Yf01@yandex.ru
Афанасьев Максим Яковлевич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57194081345, https://orcid.org/0000-0003-4061-1407, amax@niuitmo.ru
Крылова Анастасия Андреевна — кандидат технических наук, преподаватель, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-5822-6702, ananasn94@gmail.com

Received 25.02.2025
Approved after reviewing 27.08.2025
Accepted 15.09.2025

Статья поступила в редакцию 25.02.2025
Одобрена после рецензирования 27.08.2025
Принята к печати 15.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-825-832
УДК 543.421/.424; 66.914

Контроль состава и определение дозировки ингибиторов гидратообразования по их инфракрасным спектрам

Юлия Сергеевна Кожевина[1]✉, Татьяна Николаевна Носенко[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] leta-x@mail.ru✉, https://orcid.org/0009-0006-1359-1235
[2] tata-nostra@yandex.ru, tnnosenko@itmo.ru, https://orcid.org/0000-0003-4159-133X

Аннотация

Введение. Исследована возможность повышения точности и оперативности применения инфракрасных спектров термодинамических ингибиторов для контроля их состава и расчета дозировки, необходимой для предотвращения гидратообразования в нефтяной и газовой промышленности. Предложенный метод заключается в определении количества ингибитора для исследуемой системы «газ-вода» и величины снижения температуры начала гидратообразования. Актуальность работы и ее новизна в сравнении с традиционным экспериментальным подходом состоит в появлении возможности качественной и количественной идентификаций до девяти компонентов в составе термодинамического ингибитора, сокращении временных затрат на процессы расчетов. **Метод.** Для решения задачи определения концентрации веществ используется метод инфракрасной спектрометрии с преобразованием Фурье. Инфракрасные спектры растворов измерялись в режиме нарушенного полного внутреннего отражения. Для повышения точности измерений концентрации веществ по инфракрасному спектру в условиях многокомпонентности и схожести компонентов по химическому строению предложено применение регрессионной нейронной сети. В обучающую выборку были включены инфракрасные спектры чистых веществ — каждого отдельного компонента, двух- и трехкомпонентные смесевые водные растворы (вода + спирт + гликоль), а также ряд четырехкомпонентных растворов (гликоли + вода). Полученные данные о составе ингибитора использовались при расчете его дозировки для предотвращения гидратообразования в заданных условиях. **Основные результаты.** Продемонстрирована возможность обученной нейронной сети определять концентрации до девяти схожих по своим свойствам веществ в составе термодинамических ингибиторов гидратообразования: метанол, этанол, пропанол, моноэтиленгликоль, диэтиленгликоль, триэтиленгликоль пропиленгликоль, глицерин. Показано, что применение нейронной сети обеспечивает точность определения концентраций до 2 % об. Апробация предложенного метода обработки результатов контроля состава и определения дозировки термодинамического ингибитора для подавления процесса образования гидратов показала хорошее соответствие результатам традиционно применяемого метода. **Обсуждение.** Предложенный подход позволяет повысить оперативность подбора дозировки ингибиторов. Результаты работы могут найти применение в нефтепромысловой химии для входного контроля и прогнозирования эффективности применения ингибиторов гидратообразования термодинамического типа действия при добыче, подготовке или транспортировке углеводородного сырья.

Ключевые слова

ингибиторы гидратообразования, газогидраты, инфракрасная спектроскопия, хемометрические методы анализа, нейронные сети

Благодарности

Работа была выполнена при поддержке Центра химической инженерии Университета ИТМО.

Ссылка для цитирования: Кожевина Ю.С., Носенко Т.Н. Контроль состава и определение дозировки ингибиторов гидратообразования по их инфракрасным спектрам // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 825–832. doi: 10.17586/2226-1494-2025-25-5-825-832

Control of composition and determination of dosage of hydrate formation inhibitors by their infrared spectra

Iuliia S. Kozhevina[1]✉, Tatiana N. Nosenko[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] leta-x@mail.ru✉, https://orcid.org/0009-0006-1359-1235

[2] tata-nostra@yandex.ru, tnnosenko@itmo.ru, https://orcid.org/0000-0003-4159-133X

Abstract

The possibility of increasing the accuracy and efficiency of using infrared spectra of thermodynamic inhibitors to control their composition and calculate the dosage required for preventing hydrate formation in the oil and gas industry has been studied. The proposed method consists of determining the amount of inhibitor for the studied "gas-water" system and the magnitude of the decrease in the temperature of the onset of hydrate formation. The relevance of the work and its novelty in comparison with the traditional experimental approach consists in the emergence of the possibility of qualitative and quantitative identification of up to nine components in the composition of the thermodynamic inhibitor, reducing the time costs for calculation processes. To solve the problem of determining the concentration of substances, the method of infrared spectrometry with Fourier transformation is used. The infrared spectra of the solutions were measured in the mode of attenuated total internal reflection. To improve the accuracy of measuring the concentration of substances by the infrared spectrum in conditions of multicomponentity and similarity of components by chemical structure, the use of a regression neural network is proposed. The training sample included infrared spectra of pure substances, two-component and three-component mixed aqueous solutions (water + alcohol + glycol), as well as a number of four-component solutions (glycols + water). The obtained data on the composition of the inhibitor were then used to calculate its dosage to prevent hydrate formation under specified conditions. The capability of the trained neural network to determine the concentrations of up to nine substances similar in their properties in the composition of thermodynamic hydrate formation inhibitors has been demonstrated: methanol, ethanol, propanol, monoethylene glycol, diethylene glycol, triethylene glycol, propylene glycol, glycerol. It has been shown that the use of the neural network ensures the accuracy of concentration determination up to 2 % vol. Testing of the proposed method for processing the results of composition control and determining the dosage of the thermodynamic inhibitor for suppressing the hydrate formation process has shown good agreement with the results of the traditionally used method. The proposed approach allows increasing the efficiency of inhibitor dosage selection. The results of the work can be used in oilfield chemistry for incoming control and forecasting the efficiency of using thermodynamic type hydrate inhibitors during the extraction, preparation or transportation of hydrocarbon raw materials.

Keywords

hydrate formation inhibitors, gas hydrates, Fourier-transform infrared spectroscopy, projection on latent structures, chemometric methods of analysis, neural networks

Acknowledgements

The study was facilitated through the provision of materials and resources by the Center for Chemical Engineering of the ITMO University.

For citation: Kozhevina Iu.S., Nosenko T.N. Control of composition and determination of dosage of hydrate formation inhibitors by their infrared spectra. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 825–832 (in Russian). doi: 10.17586/2226-1494-2025-25-5-825-832

Введение

Добыча нефти и газа в основном сопровождается образованием газовых гидратов, которые, несмотря на свою необходимость в вопросах хранения и транспортировки значительных объемов газа [1–6], являются причинами закупорки технологического оборудования [7]. Для предупреждения и борьбы с гидратными отложениями, кроме различных технических и технологических решений, выделяют ингибиторы гидратообразования — химические составы (реагенты), обладающие способностью растворять уже образовавшиеся агломерации кристаллов гидратов или подавлять их рост и скопление [8]. Ингибиторы различают по механизму воздействия на процесс образования гидратов. Разные типы ингибиторов отличаются по составу действующих веществ [9, 10].

В настоящей работе исследуются ингибиторы гидратообразования термодинамического типа действия, которые за счет изменения термодинамических свойств системы сокращают области температур и давлений, при которых в системе может начаться гидратообразование. Способность ингибитора предотвращать гидратообразование оценивается по величине снижения температуры его начала и количеству (дозировке) ингибитора, которое это снижение обеспечивает. При этом поддерживается условие неизменности состава газа и термобарического режима в исследуемой системе без ингибитора и с ингибитором [11]. Одним из наиболее достоверных способов подбора эффективной дозировки ингибитора является экспериментальный [12, 13]. Однако для его реализации требуются значительные временные (время проведения эксперимента может занимать 12–24 ч) и материальные (специализированное оборудование) затраты. Таким образом, актуально привлечение экспрессных способов определения дозировки ингибитора для обеспечения безгидратного режима работы исследуемой системы.

Как показывают последние исследования, снижение температуры начала гидратообразования и дозировка ингибитора при фиксированном давлении зависят главным образом от качественного и количественного составов ингибитора [14, 15]. Паспортные данные коммерческих ингибиторов не могут дать точные представ-

ления о его составе, концентрации компонентов ингибитора представлены в широком диапазоне, что не дает возможности оценить его необходимое количество.

Для контроля состава нефтепромысловых химических реагентов, в том числе ингибиторов гидратообразования, применяются методы рамановской спектроскопии [16], инфракрасной (ИК) спектроскопии [17], масс-спектроскопии [18] и тонкослойной хроматографии [19]. В известных работах представлены в основном ингибиторы коррозии, деэмульгаторы. Чаще решается задача установления качества реагента (подлинность, постоянство состава), но не определения качественного и количественного составов. Последнее актуально для ингибиторов гидратообразования, когда необходимо установление его дозировки. Детализация компонентного состава нефтепромысловых реагентов определяется как практическая значимая техническая задача [20].

В настоящей работе представлен новый алгоритм оперативного определения эффективной дозировки термодинамических ингибиторов гидратообразования.

Материалы и методы

Исследуемые материалы (пробоподготовка). Среди веществ, обладающих термодинамически ингибирующим действием на газовые гидраты, чаще всего применяются одно-, двух- и трехатомные спирты, которые составляют активную основу компонентов ингибиторов. Для проведения исследования приготовлены модельные растворы, в состав которых вошли 9 типичных для них веществ: вода (в качестве растворителя) и действующие вещества: метанол (химически чистый (хч), АО «Вектон»), этанол (хч, АО «Вектон»), пропанол (хч, ООО «КОМПОНЕНТ-РЕАКТИВ»), моноэтиленгликоль (хч, АО «ЭКОС-1»), диэтиленгликоль (чистый для анализа (чда), АО «РоссПолимер»), триэтиленгликоль (хч, АО «РоссПолимер»), пропиленгликоль (хч, АО «ЭКОС-1»), глицерин (чда, АО «ЭКОС-1»). В смесевых модельных растворах содержание компонентов было равномерно распределено в области от 0 до 100 % по объему (% об.). Двухкомпонентные растворы — водные растворы каждого из 8 перечисленных спиртов с шагом концентрации в 20 % по объему. Трехкомпонентные растворы: растворитель — вода, первый действующий компонент — одноатомный спирт (метанол/этанол/пропанол), второй — многоатомный спирт (моноэтиленгликоль/диэтиленгликоль/триэтиленгликоль/пропиленгликоль/глицерин).

Подобная схема приготовления применялась для 15 возможных из указанных веществ пар: одноатомный спирт — многоатомный спирт (например, вода-метанол-этиленгликоль, вода-метанол-глицерин, вода-этанол-этиленгликоль и т. д.). Четырехкомпонентные растворы: вода и смесь многоатомных спиртов.

Оборудование. Регистрация спектров чистых веществ — каждого отдельного компонента и приготовленных смесевых модельных растворов проводилась на ИК-Фурье-спектрометре BRUKER Tenzor 37, оснащенном приставкой нарушенного полного внутреннего отражения (НПВО) с кристаллом KBr с алмазным напылением. Инструментальные параметры регистрации спектров: диапазон волновых чисел от 600 до 4000 $см^{-1}$ с разрешением 2 $см^{-1}$, усреднение по 16 измерениям.

В процессе регистрации спектра обеспечивалось его непрерывное добавление на поверхность кристалла приставки НПВО для поддержания постоянства концентрации летучих соединений в составе образца.

Обработка результатов измерений. Полученные ИК спектры проходили предобработку (корректировка базовой линии) с использованием программного обеспечения OPUS 7.5. Методика выбора метода предобработки ИК спектров для создания обучающей выборки описана в работе [21].

Несмотря на то, что количественное и качественное определения состава вещества по его ИК спектру типично для современной химии, с ростом компонентности исследуемых образцов (более трех составляющих) задача определения количественного состава смесей усложняется и сопровождается потерей точности измерений. Это характерно для растворов и смесей, содержащих вещества, сходные по химическому строению, например, растворов органических соединений. Проблема наличия перекрывающихся областей колебаний характеристических групп в случае определения количественного состава частично решается хемометрическими методами. В настоящей работе предложено применить искусственную нейронную сеть (ИНС). Такой подход, наряду с другими методами многомерного моделирования, все чаще используется для качественного и количественного регрессионного спектрального анализа [22].

В работе выбрана регрессионная нейронная сеть, реализованная в виде многослойного перцептрона. Его обучение осуществляется с использованием метода обратного распространения ошибки.

Определение минимально необходимой дозировки, обеспечивающей снижение температуры гидратообразования для достижения безгидратного режима работы в заданных условиях, на основании величин концентраций активных компонентов ингибитора выполнялось с использованием зависимости Н.А. Шостака [23].

Программный код предлагаемого алгоритма реализован на языке Python[1].

Результаты и обсуждение

Описание блок-схемы (модулей) обработки результатов измерений. На вход предлагаемой схемы (рис. 1) поступают предобработанные ИК спектры ингибитора: полного состава («первого» спектра $\mathbf{X_f}$) и после испарения летучей части («последнего» спектра $\mathbf{X_l}$). Каждый спектр представляет собой вектор из 3525 значений амплитуд при различных значениях волновых чисел.

В первом модуле по последнему спектру ($\mathbf{X_l}$) ингибитора определяется его тип действия **type** [10].

[1] [Электронный ресурс]. Режим доступа: https://fips.ru/EGD/46e9ef46-f15b-4d0c-b81d-6bbdc16cb314 (дата обращения: 04.09.2025).

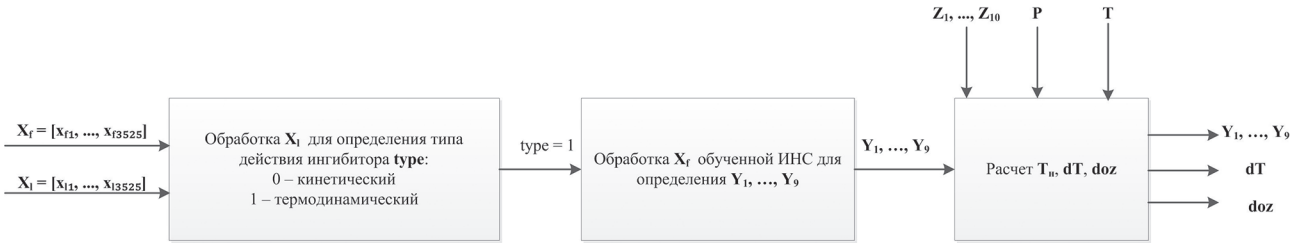

Рис. 1. Блок-схема обработки результатов измерений состава и расчета дозировки ингибиторов гидратообразования по их инфракрасным спектрам

Fig. 1. Block diagram for processing the results of composition measurements and calculating the dosage of hydrate formation inhibitors based on their infrared spectra

Во втором модуле производится определение количественного и качественного составов ($\mathbf{Y_1, ..., Y_9}$) термодинамического ингибитора по его «первому» спектру ($\mathbf{X_f}$), заключающееся в использовании обученной на модельных образцах регрессионной нейронной сети.

В третьем модуле производится расчет минимально необходимой дозировки термодинамического ингибитора **doz**, обеспечивающей снижение температуры гидратообразования **dT** для достижения безгидратного режима работы в заданных условиях. Расчет выполнен на основании значения температуры начала гидратообразования в системе без ингибитора $\mathbf{T_н}$ (вычисляется по зависимостям [23]), а также определенных в предыдущем модуле значений концентраций компонентов и введенных пользователем состава газа $\mathbf{Z_1, ..., Z_{10}}$ (мол.%) и рабочих давлении **P** и температуре **T**.

Результатом обработки исходных параметров являются значения объемной концентрации девяти компонентов $\mathbf{Y_1, ..., Y_9}$ (вода (растворитель), метанол, этанол, пропанол, моноэтиленгликоль, диэтиленгликоль, триэтиленгликоль, пропиленгликоль, глицерин), объемной дозировки ингибитора **doz** и расчетного значения снижения температуры гидратообразования **dT**, которое обеспечивает расчетная дозировка.

Обучение искусственной нейронной сети

На основании известных спектров (160 спектров) модельных смесей составлена обучающая выборка для регрессионной нейронной сети. В обучающую выборку включены ИК спектры чистых веществ — каждого отдельного компонента, двух- и трехкомпонентные (вода + спирт + гликоль) смесевые водные растворы, а также ряд четырехкомпонентных (гликоли + вода) для учета взаимного влияния гликолей в присутствии воды. Пример спектральных данных модельных растворов, использующихся в обучающей выборке, приведен на рис. 2.

Видно, что в спектрах приготовленных модельных растворов с увеличением концентрации спирта наблю-

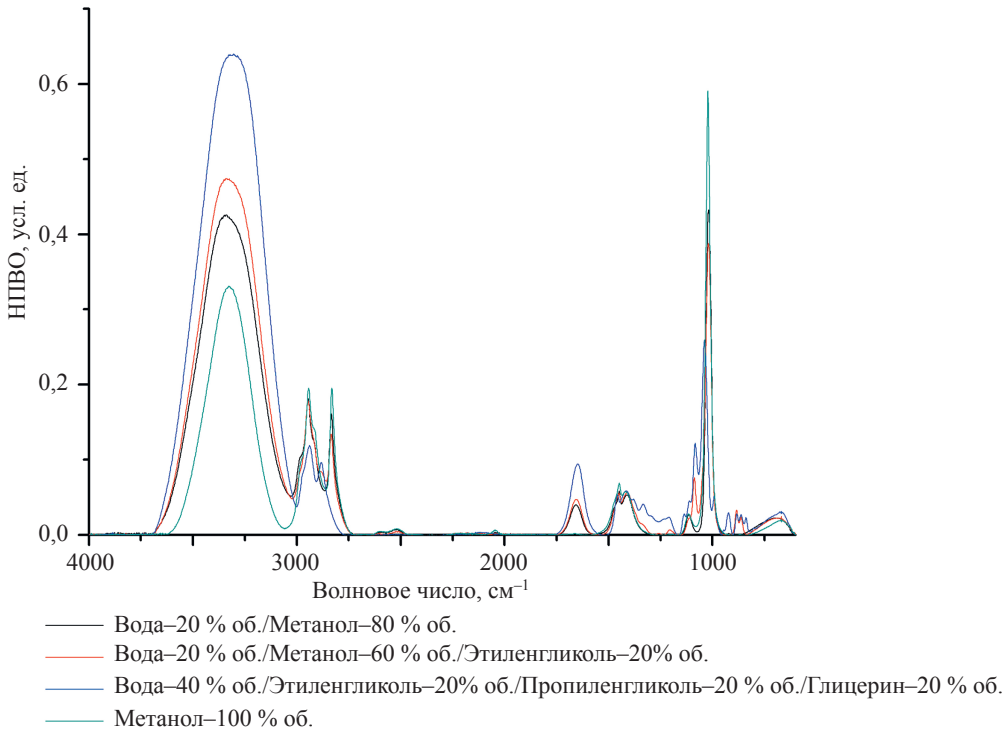

Рис. 2. Инфракрасные спектры модельных растворов в составе обучающей выборки

Fig. 2. Infrared spectra of model solutions in the training set

дается уменьшение интенсивности: высокочастотной полосы 3300 $см^{-1}$ валентных колебаний гидроксильных ОН-групп; деформационных колебаний связанных ОН-групп молекул воды в области 1650–1637 $см^{-1}$.

При этом, увеличивается интенсивность: для группы пиков в интервале 3000–2800 $см^{-1}$, которая соответствует валентным колебаниям связей С–Н; колебаний группы $–CH_2$ в области спектра около 2920 $см^{-1}$ (кроме водно-метанольного раствора); в области деформационных колебаний связей С–Н (1450–1400 $см^{-1}$); пиков при волновом числе 1110–1080 $см^{-1}$ деформационных колебаний гидроксильной группы спиртов; при волновом числе 1040–1020 $см^{-1}$, соответствующей полосе поглощения валентных колебаний связи С–О; деформационных колебаний связи С–О (880 $см^{-1}$, 860 $см^{-1}$) (кроме водно-метанольного раствора).

Спектральные данные модельных растворов представляют собой вектора (массивы) из 3525 значений амплитуд при разных значениях волновых чисел. Эти данные использованы для обучения регрессионной нейронной сети, способной определять концентрации компонентов в составе исследуемых растворов на основании ИК спектра. ИНС автоматически выявляет в этом спектре сложные, неочевидные закономерности, например, наличие определенных функциональных групп или связей, преобразует их через последовательность математических операций и выдает на выходе результат — концентрацию. Набор нейронов на входе — спектральные данные (полученные в средней ИК области в режиме НПВО); модельного раствора — набор интенсивностей (НПВО уд. ед.), каждая из которых соответствует волновому числу из диапазона измерения (3525 значений); на выходе — концентрация вещества — 9 значений, что соответствует числу определяемых компонентов. По результатам настройки гиперпараметров многослойного перцептрона определяются оптимальные, с точки зрения прогностической способности, алгоритмы оптимизации и функции активации по каждому определяемому компоненту в составе смеси. На основании [22] определено, что для достижения точности в многослойном перцептроне необходимо использовать два скрытых слоя, содержащих 128 и 64 нейрона. Определено оптимальное количество максимальных итераций. Для моделей по определению содержания воды, диэтиленгликоля и триэтиленгликоля оказалось достаточным 200 максимальных итераций, для остальных — 500. Значения остальных параметров использовались по умолчанию, которые при изменении не оказали существенного воздействия на улучшение прогностической способности нейронной сети. Вследствие различия оптимальных параметров для нейронной сети в зависимости от определяемого вещества, было обучено 9 моделей для определения каждого вещества в отдельности. Параметры качества отражены в табл. 1. Оценка прогностической способности нейронной сети проведена на тестовой выборке, состоящей из 30 проверочных образцов. Для каждого вещества по результатам определения его концентрации в каждом из проверочных образцов определено значение средней абсолютной погрешности (Mean Absolute Error, MAE) (табл. 1). Из данных табл. 1 следует, что максимальное значение MAE определения концентрации вещества с учетом округления до целых не превышает значения 2 % об.

Экспериментальные данные. Апробация предлагаемого алгоритма выполнена на примере коммерческого ингибитора, состоящего из 95 % об. метанола и 5 % об. воды.

ИК спектры — «первый» (полного состава) и «последний» (после испарения летучей части) исследуемого коммерческого ингибитора представлены на рис. 3.

По «последнему» спектру коммерческого ингибитора определен тип действия — термодинамический.

Результат работы нейронной сети по определению качественного и количественного составов исследуемого коммерческого ингибитора по его «первому» спектру отображен в табл. 2.

Расчет дозировки исследуемого коммерческого ингибитора для предотвращения гидратообразования проводился для составов газов: метан — 82,48 мол. %, этан — 4,23 мол. %, азот — 2,78 мол. %, углекислый газ — 1,95 мол. %, пропан — 1,44 мол. %, бутан — 1,22 мол. %; при рабочих давлении и температуре — 3 МПа и 262,5 К. Для расчета дозировки ингибитора при заданных условиях предварительно вычислены температура начала гидратообразования при задан-

Таблица 1. Метрики качества обученной нейронной сети
Table 1. Quality metrics of constructed neural networks

Вещество	RMSE, % об.	R-Square	MAE, % об.
Метанол	0,801	0,978	0,404
Этанол	0,820	0,978	0,610
Пропанол	0,805	0,998	0,401
Вода	1,815	0,979	0,812
Моноэтиленгликоль	1,102	0,998	0,702
Пропиленгликоль	1,606	0,989	0,800
Глицерин	1,910	0,979	1,181
Диэтиленгликоль	0,500	0,979	0,405
Триэтиленгликоль	1,408	0,989	0,400

Примечание. RMSE — среднеквадратичное отклонение прогноза; R-Square — доля дисперсии зависимых переменных.

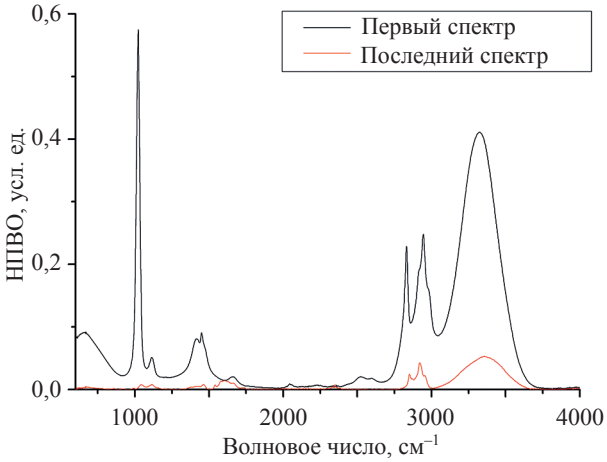

Рис. 3. Полученные инфракрасные спектры коммерческого ингибитора («первый» — спектр полного состава, «последний» — спектр после испарения летучей части)

Fig. 3. The obtained IR spectra of the commercial inhibitor (the first spectrum is the spectrum of the complete composition, the second spectrum is the spectrum after evaporation of the volatile part)

Таблица 2. Состав тестируемого ингибитора, определенный при реализации предлагаемого алгоритма

Table 2. The composition of the tested inhibitor, determined during the implementation of the proposed algorithm

Компонент	Прогнозируемое значение, % об.
Метанол	**95,2**
Этанол	0,3
Пропанол	0,0
Вода	4,0
Моноэтиленгликоль	0,8
Пропиленгликоль	0,0
Глицерин	0,3
Диэтиленгликоль	0,0
Триэтиленгликоль	0,0

ном давлении в системе без ингибитора, значение которой составило 279 К, и минимально необходимое для заданных условий снижение температуры начала гидратообразования — 16,5 К. Расчетное значение дозировки ингибитора, необходимое для достижения определенного снижения температуры гидратообразования и обеспечивающей безгидратный режим работы при заданных условиях составило 19 % об.

Сравнение результатов работы предлагаемого алгоритма с результатами, которые получаются при реализации традиционного экспериментального подхода [24, 25] для пяти коммерческих ингибиторов, показало расхождение в определении дозировки, необходимой для предотвращения гидратообразования, не более 2 % об. При этом получено снижение температуры начала гидратообразования максимум на 0,5 К, что в производственных условиях технологически приемлемо. Длительность традиционного экспериментального подбора дозировки составляет в среднем 12–24 ч [24, 25], использование предложенного алгоритма сокращает эту процедуру до 5 мин, повышая оперативность процесса подбора дозировки и обеспечивая значительную экономию времени и ресурсов.

Заключение

В работе предложено использование инфракрасных спектров термодинамических ингибиторов гидратообразования для контроля их состава и расчета дозировки, необходимой для предотвращения гидратообразования. Представлен способ обработки спектров, позволяющий повысить оперативность (длительность сокращается до 5 мин) определения предотвращающей гидратообразование дозировки термодинамических ингибиторов в сравнении с традиционным экспериментальным подходом за счет использования инфракрасной спектроскопии. Предложена нейронная сеть, обученная на инфракрасных спектрах модельных составов термодинамических ингибиторов. Применение созданной нейронной сети обеспечивает качественную (до 9 компонентов) и количественную оценки состава термодинамических ингибиторов с точностью, согласно значению средней абсолютной погрешности, до 2 % об. по инфракрасным спектрам, что повышает информативность контроля состава ингибиторов. Результаты работы могут найти применение в нефтепромысловой химии для входного контроля и прогнозирования эффективности применения ингибиторов гидратообразования термодинамического типа действия, используемых для предотвращения образования газогидратов при добыче, подготовке или транспортировке углеводородного сырья.

Литература

1. Истомин В.А., Якушев В.С., Квон В.Г., Долгаев С.И., Чувилин Е.М. Направления современных исследований газовых гидратов // Газохимия. 2009. № 5. С. 56–63.
2. Макогон Ю.Ф. Природные газовые гидраты: распространение, модели образования, ресурсы // Российский химический журнал. 2003. Т. 47. № 3. С. 70–79.
3. Соловьёв В.А. Природные газовые гидраты как потенциальное полезное ископаемое // Россйкий химический журнал. 2003. Т. 47. № 3. С. 59–69.
4. Makogon Y.F. Natural gas hydrates — A promising source of energy // Journal of Natural Gas Science and Engineering. 2010. V. 2. N 1. P. 49–59. https://doi.org/10.1016/j.jngse.2009.12.004

References

1. Istomin V.A., Iakushev B.C., Kvon V.G., Dolgaev S.I., Chuvilin E.M. State-of-the-art research of the gas hydrates. *Gazohimiya*, 2009, no. 5, pp. 56–63. (in Russian)
2. Makogon Iu.F. Natural gas hydrates: distribution, formation models, resources. *Rossijskij Himicheskij Zhurnal*, 2003, vol. 47, no. 3, pp. 70–79. (in Russian)
3. Solovev V.A. Natural gas hydrates as a potential mineral resource. *Rossijskij Himicheskij Zhurnal*, 2003, vol. 48, no. 3, pp. 59–69. (in Russian)
4. Makogon Y.F. Natural gas hydrates — A promising source of energy. *Journal of Natural Gas Science and Engineering*, 2010, vol. 2, no. 1, pp. 49–59. https://doi.org/10.1016/j.jngse.2009.12.004

5. Hongsheng D., Wang J., Xie Z., Wang B., Zhang L., Shi Q. Potential applications based on the formation and dissociation of gas hydrates // Renewable and Sustainable Energy Reviews. 2021. V. 143. P. 110928. https://doi.org/10.1016/j.rser.2021.110928
6. Thakre N., Jana A.K. Physical and molecular insights to Clathrate hydrate thermodynamics // Renewable and Sustainable Energy Reviews. 2021. V. 135. P. 110150. https://doi.org/10.1016/j.rser.2020.110150
7. Sloan E.D., Koh C.A., Sum A. Natural Gas Hydrates in Flow Assurance. Gulf Professional Publishing, 2010. 224 p.
8. Грицишин Д.Н., Квон В.Г., Истомин В.А., Минигулов Р.М. Технологии предупреждения гидратообразования в промысловых системах: проблемы и перспективы // Газохимия. 2009. № 10. С. 32–40.
9. Келланд М.А. Промысловая химия в нефтегазовой отрасли. СПб.: Профессия, 2015. 607 с.
10. Ворожцова Ю.С., Носенко Т.Н., Успенская М.В. Определение типа действия ингибиторов гидратообразования по их инфракрасным спектрам // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 4. С. 669–675. https://doi.org/10.17586/2226-1494-2023-23-4-669-675
11. Semenov A.P., Medvedev V.I., Gushchin P.A., Vinokurov V.A. Kinetic inhibition of hydrate formation by polymeric reagents: effect of pressure and structure of gas hydrates // Chemistry and Technology of Fuels and Oils. 2016. V. 51. N 6. P. 679–687. https://doi.org/10.1007/s10553-016-0658-5
12. Gjertsen L.H., Fadnes F.H. Measurements and predictions of hydrate equilibrium conditions // Annals of the New York Academy of Sciences. 2006. V. 912. N 1. P. 722–734. https://doi.org/10.1111/j.1749-6632.2000.tb06828.x
13. Tohidi B., Burgass R.W., Danesh A., Ostergaard K.K., Todd A.C. Improving the accuracy of gas hydrates dissociation point measurements // Annals of the New York Academy of Sciences. 2000. V. 912. N 1. P. 924–931. https://doi.org/10.1111/j.1749-6632.2000.tb06846.x
14. Zaporozhets E.P., Shostak N.A. Efficiency estimation of the singleand multicomponent anti-hydrate reagents // Journal of Mining Institute. 2019. V. 238. P. 423–429. https://doi.org/10.31897/PMI.2019.4.423
15. Муратова Э.Ж., Крапивин В.Б., Истомин В.А, Федулов Д.М., Квон В.Г., Герасимов Ю.А., Сергеева Д.В., Тройникова А.А., Семенов А.П. Ингибитор гидратообразования на основе смесей моноэтиленгликоля и метанола // Вести газовой науки. 2023. № 4 (56). С. 145–154.
16. Гусаков В.Н., Катермин А.В., Михайлова Л.Р., Горбунов В.В., Невядовский Е.Ю. Разработка методологии оперативного контроля качества нефтепромысловых химических реагентов // Нефтегазовое дело. 2021. Т. 19. № 4. С. 81–89. https://doi.org/10.17122/ngdelo-2021-4-81-89
17. Ракитин А.Р., Боженкова Г.С., Киселев С.А. Стеванович Е., Кильмаматов А.А. Инфракрасная спектроскопия для контроля качества ингибиторов коррозии // Нефтепромысловое дело. 2022. № 11 (647). С. 69–76. https://doi.org/10.33285/0207-2351-2022-11(647)-69-76
18. Суховерхов С.В., Задорожный П.А., Полякова Н.В. Применение инструментальных методов для анализа объектов нефтепромысловой химии // Вестник Дальневосточного отделения Российской академии наук. 2021. № 5 (219). С. 134–143. https://doi.org/10.37102/0869-7698_2021_219_05_11
19. Саранцева В.Д., Бадамшин А.Г., Каштанова Л.Е. Оценка возможности применения методов тонкослойной хроматографии и ИК-спектроскопии в лабораторных исследованиях по подбору химических реагентов // Практические аспекты нефтепромысловой химии. 2023. С. 146–147.
20. Ишмияров Э.Р., Прокудина В.Д. Формирование инструментального лабораторного подхода по контролю качества нефтепромысловых химических реагентов // Экспозиция Нефть Газ. 2024. № 8 (109). С. 134–143. https://doi.org/10.24412/2076-6785-2024-8-134-143
21. Кожевина Ю.С., Носенко Т.Н., Успенская М.В. Оценка количественного состава ингибиторов гидратообразования по их инфракрасным спектрам // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 3. С. 366–374. https://doi.org/10.17586/2226-1494-2024-24-3-366-374
22. Лаптинский К.А., Буриков С.А., Сарманова О.Э., Вервальд А.М., Утегенова Л.С., Пластинин И.В., Доленко Т.А. Диагностика вредных примесей в водных средах с помощью спектроскопиче-

5. Hongsheng D., Wang J., Xie Z., Wang B., Zhang L., Shi Q. Potential applications based on the formation and dissociation of gas hydrates. *Renewable and Sustainable Energy Reviews*, 2021, vol. 143, pp. 110928. https://doi.org/10.1016/j.rser.2021.110928
6. Thakre N., Jana A.K. Physical and molecular insights to Clathrate hydrate thermodynamics. *Renewable and Sustainable Energy Reviews*, 2021, vol. 135, pp. 110150. https://doi.org/10.1016/j.rser.2020.110150
7. Sloan E.D., Koh C.A., Sum A. *Natural Gas Hydrates in Flow Assurance*. Gulf Professional Publishing, 2010, 224 p.
8. Gritcishin D.N., Kvon V.G., Istomin V.A., Minigulov R.M. Technologies for preventing hydrate formation in production systems: challenges and prospects. *Gazohimiya*, 2009, no. 10, pp. 32–40. (in Russian)
9. Kelland M.A. *Production Chemicals for the Oil and Gas Industry*. CRC Press, 2009, 456 p.
10. Vorozhtsova Iu.S., Nosenko T.N., Uspenskaya M.V. Determination of the action type of hydrate formation inhibitors by their infrared spectra. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 669–675. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-4-669-675
11. Semenov A.P., Medvedev V.I., Gushchin P.A., Vinokurov V.A. Kinetic inhibition of hydrate formation by polymeric reagents: effect of pressure and structure of gas hydrates. *Chemistry and Technology of Fuels and Oils*, 2016, vol. 51, no. 6, pp. 679–687. https://doi.org/10.1007/s10553-016-0658-5
12. Gjertsen L.H., Fadnes F.H. Measurements and predictions of hydrate equilibrium conditions. *Annals of the New York Academy of Sciences*, 2006, vol. 912, no. 1, pp. 722–734. https://doi.org/10.1111/j.1749-6632.2000.tb06828.x
13. Tohidi B., Burgass R.W., Danesh A., Ostergaard K.K., Todd A.C. Improving the accuracy of gas hydrates dissociation point measurements. *Annals of the New York Academy of Sciences*, 2000, vol. 912, no. 1, pp. 924–931. https://doi.org/10.1111/j.1749-6632.2000.tb06846.x
14. Zaporozhets E.P., Shostak N.A. Efficiency estimation of the singleand multicomponent anti-hydrate reagents. *Journal of Mining Institute*, 2019, vol. 238, pp. 423–429. https://doi.org/10.31897/PMI.2019.4.423
15. Muratova E.Zh., Krapivin V.B., Istomin V.A., Fedulov D.M., Kvon V.G., Gerasimov Iu.A., Sergeeva D.V., Troinikova A.A., Semenov A.P. Hydrate inhibitor based on mixtures of monoethylene glycol and methanol. *Vesti Gazovoj Nauki*, 2023, no. 4 (56). pp. 145–154. (in Russian)
16. Gusakov V.N., Katermin A.V., Mikhailova L.R., Gorbunov V.V., Neviadovskii E.Iu. Methodology for quality control and application of oilfield chemicals. *Petroleum Engineering*, 2021, vol. 19, no. 4, pp. 81–89. (in Russian). https://doi.org/10.17122/ngdelo-2021-4-81-89
17. Rakitin A.R., Bozhenkova G.S., Kiselev S.A. Stevanovich E., Kilmamatov A.A. Infrared spectroscopy for quality control of corrosion inhibitors. *Oilfield Engineering*, 2022, no. 11 (647), pp. 69–76. (in Russian). https://doi.org/10.33285/0207-2351-2022-11(647)-69-76
18. Sukhoverkhov S.V., Zadorozhnyi P.A., Poliakova N.V. Application of instrumental methods for oilfield chemistry objects analysis. *Vestnik of the Far East Branch of the Russian Academy of Sciences*, 2021, no. 5 (219), pp. 134–143. (in Russian). https://doi.org/10.37102/0869-7698_2021_219_05_11
19. Sarantceva V.D., Badamshin A.G., Kashtanova L.E. Evaluation of the possibility of using thin-layer chromatography and IR spectroscopy methods in laboratory studies concerning the selection of chemical reagents. *Proc. of the Practical Aspects of Oilfield Chemistry*, 2023, pp. 146–147. (in Russian)
20. Ishmiiarov E.R., Prokudina V.D. Ormation of an instrumental laboratory approach to quality control of oilfield chemical reagents. *Exposition Oil Gas*, 2024, no. 8 (109), pp. 134–143. (in Russian). https://doi.org/10.24412/2076-6785-2024-8-134-143
21. Kozhevina Iu.S., Nosenko T.N., Uspenskaya M.V. Assessment of the quantitative composition of hydrate formation inhibitors by their infrared spectra. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 3, pp. 366–374. (in Russian). https://doi.org/10.17586/2226-1494-2024-24-3-366-374
22. Laptinskiy K. A., Burikov S.A., Sarmanova O.E., Vervald A.M., Utegenova L.S., Plastinin I.V., Dolenko T. A. Diagnostics of harmful impurities in aqueous media using spectroscopic methods and

ских методов и алгоритмов машинного обучения // Оптика и спектроскопия. 2023. Т. 131 № 6. С. 810–816. https://doi.org/10.21883/OS.2023.06.55915.106-23
23. Запорожец Е.П., Шостак Н.А. Расчет эффективности одно- и многокомпонентных антигидратных реагентов, Записки Горного института. 2019. Т. 238. С. 423–429. https://doi.org/10.31897/PMI.2019.4.423
24. Kunakova A.M., Usmanova F.G., Vorozhtsova I.S., Lanchuk I.V. Approaches to the selection of effective inhibitors of gas hydrate formation // Proc. of the SPE Russian Petroleum Technology Conference. 2019. P. 1–23. https://doi.org/10.2118/196781-MS
25. Кунакова А.М., Усманова Ф.Г., Ворожцова Ю.С., Гоголева А.Д. Оценка эффективности ингибиторов гидратообразования изотермическим методом // PROнефть. Профессионально о нефти. 2019. № 1 (11). С. 18–21. https://doi.org/10.24887/2587-7399-2019-1-18-21

machine learning algorithms. *Optics and Spectroscopy*, 2023, vol. 131, no. 6, pp. 765–771. https://doi.org/10.61011/EOS.2023.06.56664.106-23
23. Shostak N.A., Zaporozhets E.P. Efficiency estimation of the single- and multicomponent anti-hydrate reagents. *Journal of Mining Institute*, 2019, vol. 238, pp. 423-429. (in Russian). https://doi.org/10.31897/PMI.2019.4.423
24. Kunakova A.M., Usmanova F.G., Vorozhtsova I.S., Lanchuk I.V. Approaches to the selection of effective inhibitors of gas hydrate formation. *Proc. of the SPE Russian Petroleum Technology Conference*. 2019. pp. 1–23. https://doi.org/10.2118/196781-MS
25. Kunakova A.M., Usmanova F.G., Vorozhtcova Iu.S., Gogoleva A.D. Evaluation of the effectiveness of hydrate-formation inhibitors by the isothermal method. *PROneft. Professionals about Oil*, 2019, no. 1 (11), pp. 18–21. (in Russian). https://doi.org/10.24887/2587-7399-2019-1-18-21

Авторы

Кожевина Юлия Сергеевна — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57215118092, https://orcid.org/0009-0006-1359-1235, leta-x@mail.ru
Носенко Татьяна Николаевна — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57190940294, https://orcid.org/0000-0003-4159-133X, tata-nostra@yandex.ru, tnnosenko@itmo.ru

Authors

Iuliia S. Kozhevina — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57215118092, https://orcid.org/0009-0006-1359-1235, leta-x@mail.ru
Tatiana N. Nosenko — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57190940294, https://orcid.org/0000-0003-4159-133X, tata-nostra@yandex.ru, tnnosenko@itmo.ru

Статья поступила в редакцию 12.05.2025
Одобрена после рецензирования 05.09.2025
Принята к печати 25.09.2025

Received 12.05.2025
Approved after reviewing 05.09.2025
Accepted 25.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-833-843
УДК 004.89

Классификация двигательной активности человека на основе анализа мультисенсорных данных

Артём Дмитриевич Обухов✉

Тамбовский государственный технический университет, Тамбов, 392000, Российская Федерация
obuhov.art@gmail.com✉, https://orcid.org/0000-0002-3450-5213

Аннотация

Введение. Выполнен анализ мультисенсорных данных, полученных от электромиографа, инерциальных измерительных устройств, системы компьютерного зрения и трекеров виртуальной реальности, для решения задачи классификации двигательной активности человека. Актуальность решения данной задачи обусловлена необходимостью анализа и распознавания двигательной активности человека при использовании различных программно-аппаратных комплексов, например, реабилитационных и тренажерных систем. Для оптимального решения задачи распознавания типа движений рук с наибольшей точностью оценивается вклад каждого источника сигналов, а также выполняется сравнение различных моделей машинного обучения. **Метод.** Подход к обработке мультисенсорных данных включает синхронизированный сбор потоков от различных источников, разметку исходных данных, фильтрацию сигналов; двойное выравнивание временных рядов по частоте и длительности с аппроксимацией до общей константы, формирование общего набора данных, обучение и выбор модели машинного обучения для распознавания двигательной активности рук. Рассматриваются девять моделей машинного обучения: логистическая регрессия, *k*-ближайших соседей, наивный байесовский классификатор, дерево решений и ансамбли на их основе (случайный лес, AdaBoost, Extreme Gradient Boosting, Voting и Stacking Classifier). Разработанный подход синхронизации, фильтрации и двойного выравнивания потоков данных позволяет сформировать унифицированный набор данных мультисенсорных данных для обучения моделей. **Основные результаты.** Проведен эксперимент по классификации девяти категорий движений рук на основе анализа мультисенсорных данных (собрано 629 записей от 15 участников). Обучение выполнялось на 80 % собранных данных с пятикратной перекрестной проверкой. Показано, что ансамбль AdaBoost обеспечивает точность классификации 98,8 % на наборе данных из объединенных от четырех различных источников информации. В ходе абляционного анализа для сравнения источников данных, наибольшее влияние на итоговую точность классификации оказывает информация от трекеров виртуальной реальности (до 98,73 ± 1,78 % точности на модели AdaBoost), данные о мышечной активности от электромиографа являются наименее информативными. Определено, что высокая точность классификации двигательной активности может быть получена с использованием инерциальных измерительных устройств. **Обсуждение.** Исследование формализует воспроизводимый подход к обработке мультисенсорных данных и позволяет объективно сравнить вклад различных источников информации и моделей машинного обучения при решении задачи классификации двигательной активности рук пользователя в рамках реабилитационных и виртуальных тренажерных систем. Показано, что при ограничениях по ресурсам возможно отказаться от части источников данных без существенной потери точности классификации, упростив аппаратную конфигурацию систем отслеживания, перейти от закрытых коммерческих систем (трекеров виртуальной реальности) к более доступным и компактным инерциальным измерительным устройствам.

Ключевые слова

классификация движений человека, двигательная активность, машинное обучение, анализ мультисенсорных данных

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках проекта «Разработка иммерсивной системы взаимодействия с виртуальной реальностью для профессиональной подготовки на основе всенаправленной платформы» (124102100628-3).

Ссылка для цитирования: Обухов А.Д. Классификация двигательной активности человека на основе анализа мультисенсорных данных // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 833–843. doi: 10.17586/2226-1494-2025-25-5-833-843

Classification of human motor activity based on multisensory data analysis

Artem D. Obukhov✉

Tambov State Technical University, Tambov, 392000, Russian Federation

obuhov.art@gmail.com✉, https://orcid.org/0000-0002-3450-5213

Abstract

An analysis of multisensor data obtained from an electromyograph, inertial measurement devices, a computer-vision system, and virtual-reality trackers was performed in order to solve the problem of classifying human motor activity. The relevance of solving this problem is determined by the necessity of analyzing and recognizing human motor activity when using various hardware and software complexes, for example, rehabilitation and training systems. For the optimal solution of the task of recognizing the type of hand movements with the highest accuracy, the contribution of each signal source is evaluated, and a comparison of various machine-learning models is performed. The approach to processing multisensor data includes: synchronized acquisition of streams from different sources; labeling of the initial data; signal filtering; dual alignment of time series by frequency and duration with approximation to a common constant; formation of a common dataset; training and selection of a machine-learning model for recognizing motor activity of the hands. Nine machine-learning models are considered: logistic regression, *k*-nearest neighbors, naïve Bayes classifier, decision tree, and ensembles based on them (Random Forest, AdaBoost, Extreme Gradient Boosting, Voting, and Stacking Classifier). The developed approach of synchronization, filtering, and dual alignment of data streams makes it possible to form a unified dataset of multisensor data for model training. An experiment was carried out on the classification of nine categories of hand movements based on the analysis of multisensor data (629 recordings collected from 15 participants). Training was performed on 80 % of the collected data with five-fold cross-validation. The AdaBoost ensemble provides a classification accuracy of 98.8 % on the dataset composed of the combined information from four different sources. In the course of ablation analysis for comparing the data sources, the greatest influence on the final classification accuracy is exerted by information from virtual-reality trackers (up to 98.73 % ± 1.78 % accuracy on the AdaBoost model), while data on muscle activity from the electromyograph turned out to be the least informative. It was determined that high classification accuracy of motor activity can be obtained using inertial measurement devices. The considered study formalizes a reproducible approach to processing multisensor data and makes it possible to objectively compare the contribution of different sources of information and machine-learning models in solving the problem of classifying the motor activity of the user's hands within rehabilitation and virtual training systems. It is shown that under resource limitations it is possible to refuse part of the data sources without significant loss of classification accuracy, simplifying the hardware configuration of tracking systems and making it possible to move from closed commercial systems (virtual-reality trackers) to more accessible and compact inertial measurement devices.

Keywords

classification of human movements, motor activity, machine learning, multisensory data analysis

Acknowledgements

The work was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation within the framework of the project "Development of an immersive virtual reality interaction system for vocational training based on an omnidirectional platform" (124102100628-3).

For citation: Obukhov A.D. Classification of human motor activity based on multisensory data analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 833–843 (in Russian). doi: 10.17586/2226-1494-2025-25-5-833-843

Введение

Создание передовых реабилитационных и тренажерных комплексов, объединяющих технологии виртуальной реальности и дополнительное имитационное или нагрузочное оборудование, невозможно без комплексного анализа данных о процессе двигательной активности пользователя. Это, с одной стороны, может использоваться для анализа паттернов его поведения и качества выполнения упражнений [1], с другой стороны применяться для последующего формирования его цифровой копии (аватара) с высокой степенью реалистичности [2].

Для осуществления комплексного анализа двигательной активности пользователя подобного рода программно-аппаратных комплексов необходимо ориентироваться на мультисенсорный подход, т. е. объединять данные от нескольких независимых источников. Среди основных типов систем отслеживания двигательной активности стоит выделить электромиографические (Electromyography, EMG) датчики, системы инерциальной навигации (Inertial Measurement Unit, IMU) и системы компьютерного зрения (Computer Vision, CV) [3]. Необходимость комплексного подхода обусловлена тем, что каждая из систем отслеживания имеет свои сильные и слабые стороны, а также ограничения [1].

Специализированные трекеры для виртуальной реальности (Virtual Reality, VR) для абсолютного позиционирования требуют дополнительного оборудования в виде базовых станций. С другой стороны, IMU, к которым относятся различные устройства на основе акселерометров и гироскопов, включая коммерческие (Kat Loco S), позволяют осуществлять относительное позиционирование и поворот в пространстве [4]. IMU

также достаточно легко реализуются на базе широко распространенных датчиков, например, MPU-9250 [5]. Их интеграция в ESP32 позволяет создавать легкие, беспроводные носимые устройства без ограничений по количеству датчиков, их расположению, а также с возможностью самостоятельно реализовывать программную логику обработки данных [6]. Ключевой проблемой данного типа систем отслеживания является накопление ошибки интегрирования при расчете скорости движения акселерометра и, тем более, перемещения датчика. Тем не менее, IMU могут достаточно точно определять углы поворотов и направление движения, что может быть использовано при анализе двигательной активности человека.

При анализе двигательной активности большое значение имеет информация о состоянии мышечной системы человека, что можно получить путем обработки данных от датчиков EMG. EMG — широко используемый метод измерения мышечной активности путем обнаружения электрических сигналов, генерируемых мышечными сокращениями [7]. Датчики EMG могут быть прикреплены к поверхности кожи или внедрены в мышечную ткань. Поверхностные датчики EMG неинвазивны и просты в использовании, что делает их популярным выбором для отслеживания мышечной активности в различных областях применения. При использовании EMG необходимо учитывать, что данный тип датчиков очень чувствителен и подвержен помехам, в том числе, при значительной двигательной активности.

С развитием технологий CV значительное внимание уделяется методам Human Pose Estimation, позволяющим с высокой точностью определять положения ключевых точек скелета человека [8]. Современные решения, такие как MediaPipe Pose, YOLO, Movenet, OpenPose и другие, используют глубокие нейронные сети для построения скелетных моделей, что обеспечивает возможность построения трехмерных цифровых копий тела для взаимодействия с виртуальным пространством. Однако эффективность данных методов существенно зависит от качества изображений, разрешения камеры и условий освещения, что ограничивает их применение в сценариях с интенсивной динамикой или при слабом освещении. Кроме того, нужно учитывать, что активное перемещение или повороты могут сделать часть сегментов тела недоступными для камеры, что приводит к искажениям при реконструкции модели тела. Наконец, CV испытывает проблемы при определении дальности до объектов (координата Z), что частично решается стереокамерами и камерами с датчиками глубины, но точность и стоимость данных решений далеки от оптимальных.

Учитывая перечисленные проблемы, для отслеживания двигательной активности человека и ее последующего анализа (например, с использованием алгоритмов машинного обучения) необходимо объединение и обработка больших объемов данных от различных систем отслеживания. Интеграция различных методов отслеживания и мониторинга физиологических параметров позволит получить комплексный набор информации о движениях пользователя.

Эффективность применения технологий машинного обучения для решения задачи анализа и классификации двигательной активности подтверждается многочисленными исследованиями [9]. Технологии машинного обучения позволяют достаточно точно распознавать различные паттерны движений, например, при анализе походки человека на основе датчиков IMU [10], классификации различных движений [11], в том числе, при комбинации EMG и IMU [12]. В работе [13] рассмотрена задача классификации движений рук для определения статических и динамических жестов, в качестве исходных данных также выступают IMU и EMG. Итоговые результаты по точности классификации варьировались от 93 до 99 %, точность распознавания жеста от 56–70 % (для IMU) до 88–90 % (для EMG). Результаты, полученные в [13], подтвердили высокие перспективы использования машинного обучения в задачах классификации движений рук и оправданность мультисенсорного подхода для решения подобных задач. Проведенный анализ показывает возможность решения задачи классификации двигательной активности человека с использованием технологий машинного обучения. В настоящей работе, кроме анализа данных EMG и IMU, выполнено исследование эффективности мультисенсорного подхода и добавлена в качестве дополнительного источника система CV и трекеры виртуальной реальности для расширения исходного набора данных.

Предмет исследования

В представленной работе решается задача автоматической классификации двигательной активности человека на основе совмещенного анализа мультисенсорных данных. Цель исследования — разработать методику распознавания типов движений рук с использованием совокупности разнородных сигналов (EMG, IMU, CV и VR-трекеры) и оценить вклад каждого из этих источников в точность классификации. Для достижения этой цели будет сформирован набор данных, включающий синхронизированные записи от выбранных четырех типов сенсоров, собранные при выполнении пользователями различных движений верхних конечностей. Проведено обучение и сравнение нескольких моделей машинного обучения (как базовых алгоритмов, так и ансамблевых методов) по распознаванию девяти типовых движений рук на данном объединенном наборе данных. Основными задачами работы являются: количественно оценить эффективность каждого типа системы отслеживания в составе мультисенсорных данных (т. е. определить информативность сигналов EMG, IMU, CV и VR для решения задачи классификации); выявить модель машинного обучения, обеспечивающую наибольшую точность распознавания движений.

Впервые предложен и применен оригинального подход к предварительной обработке и анализу мультисенсорных временных рядов для классификации движений, где данные различных источников имеют разную частоту дискретизации и размерность. Выполнено двойное выравнивание временных рядов по частоте

дискретизации и по длительности сигналов, что обеспечивает точную синхронизацию разнородных потоков данных без потери информативности. Кроме того, впервые комплексно объединены данные сразу четырех различных типов датчиков (IMU, ЭМГ, CV и VR-трекеры) для решения задачи распознавания движений. Проведен абляционный анализ влияния каждого из каналов сигналов на качество классификации, что позволяет оценить наиболее значимые источники данных для распознавания двигательной активности.

Подход к обработке мультисенсорных данных для решения задачи классификации двигательной активности

Рассмотрим основные этапы обработки мультисенсорных данных (рисунок).

Этап 1. Процедуры по теоретической подготовке (формализация процессов обработки данных, разработка архитектур моделей машинного обучения, выбор метрик оценки моделей).

Этап 2. Эксперименты, направленные на сравнение моделей машинного обучения и достижение поставленной цели в виде решения задачи классификации двигательной активности на основе обработанных мультисенсорных данных.

На этапе 1 осуществляется формализация процесса сбора мультисенсорных данных. К объектам относятся компоненты мультисенсорной системы, отслеживающей состояние человека: EMG, IMU, система CV и VR. В совокупности они формируют входные и выходные данные для последующего обучения моделей машинного обучения. Однако перед формированием набора данных должна быть выполнена предварительная подготовка данных, которая включает следующие процедуры.

Процедура 1. Синхронизация потоков данных из различных источников в единое множество с учетом временных меток.

Процедура 2. Разделение множества информации на временные отрезки (диапазоны) путем автоматической или ручной разметки (посредством анализа собранных видеоданных о выполняемых действиях и выбором интервала, где выполняется определенное действие).

Процедура 3. Предобработка данных с использованием соответствующих фильтров.

Этап 1 наиболее актуален для данных от IMU, где выполняются дополнительные преобразования для удаления высокочастотного шума и дрейфа базовой линии.

После завершения этапа 1 формируется итоговое множество входных и выходных данных. Предполагается, что в качестве входных данных выступают мультисенсорные данные от IMU, EMG, CV и VR-трекеров. Выходными данными являются типы (категории) движений пользователя.

Далее осуществляется решение задачи классификации двигательной активности рук пользователя по нескольким категориям на основе анализа входных данных. Наибольший интерес представляет исследование влияния выбранных источников на точность классификации движений.

На этапе 2 выполняется проведение экспериментальных исследований. Для их успешного завершения необходима разработка нескольких альтернативных архитектур моделей машинного обучения, что позволит осуществить объективное сравнение различных подходов и их эффективность в контексте решаемых задач.

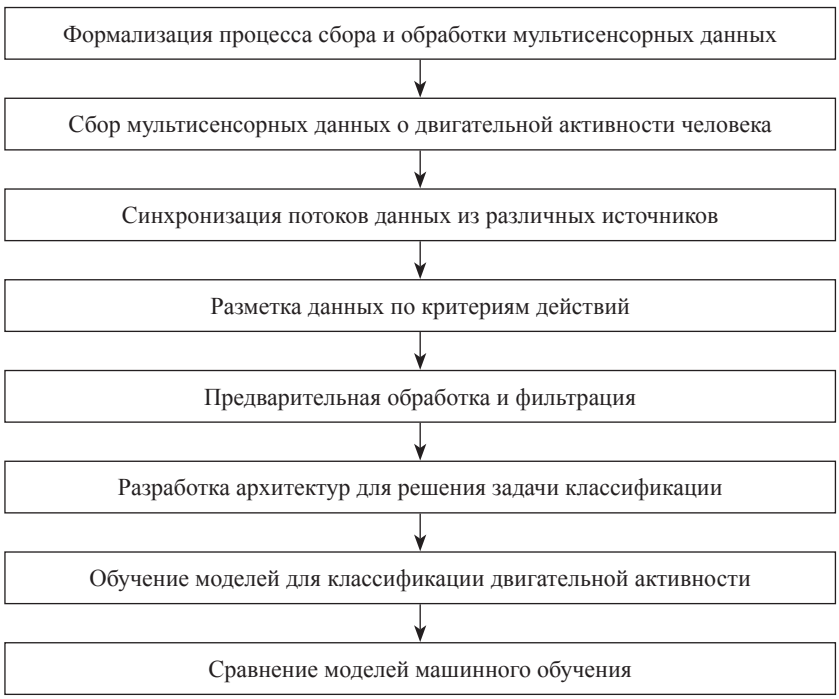

Рисунок. Структура обработки мультисенсорных данных
Figure. Multisensory data processing framework

В результате, после необходимой подготовки, осуществляется сбор данных, его подготовка в соответствии с процедурами 1–3, формирование набора данных, обучение и сравнение моделей. Модели с наилучшими показателями будут использованы в различных программно-аппаратных комплексах для мониторинга и классификации двигательной активности пользователя.

В соответствии с предложенным подходом проведем формализацию процесса сбора и обработки мультисенсорных данных.

Пусть исходные данные поступают от некоторой мультисенсорной системы сбора информации, объединяющей совокупность сенсоров:

$$S = \{s_{EMG}, s_{IMU}, s_{CV}, s_{VR}\},$$

где s_{EMG} — поверхностные EMG-датчики для регистрации электрической активности мышц; s_{IMU} — инерциальные датчики (акселерометры и гироскопы) для измерения ускорений и угловых скоростей; s_{CV} — метод CV для регистрации положения пальцев и ладони; s_{VR} — VR-трекеры для высокоточного отслеживания положения руки.

Пусть P — множество пользователей, где $p \in P$ — конкретный пользователь. Данные от всех сенсоров обозначим как B. Между множествами S и B задано соответствие, так, что есть $s \in S$, соответствующий определенному $b \in B$.

Зададим некоторую функцию φ: $P \times S \to B$, которая описывает процесс сбора данных от пользователя с помощью сенсоров. Для каждого $p \in P$ и набора датчиков S сбор данных будет иметь вид: $B_p = \{b_{p,i} = \varphi(p, s_i)\}$, $\forall s_i \in S$. После формирования сигнал $b_{p,i}(t)$ передается датчиком s_i в модуль обработки для получения исходных («сырых») данных: $r_{p,i} = \varphi(b_{p,i}, s_i)$. Функция передачи данных φ осуществляет следующее отображение φ: $B \times S \to R$, где R — множество «сырых» данных в модуле обработки.

Выполним обработку полученных данных R. Обозначим через γ: $R \times A \to E$ функцию, осуществляющую преобразование исходных данных R в обработанные E с использованием множества алгоритмов A. Для конкретных данных получим: $e_{p,i} = \gamma(r_{p,i}, a)$. Форма функции γ и используемый алгоритм $a \in A$ зависят от характеристик исходного сигнала и, следовательно, источника данных (сенсора).

В процессе обработки выполняется несколько важных процедур.

Процедура 4. Синхронизация данных для устранения $\tau_i(\forall s_i \in S)$ для всех сенсоров путем синхронизации начальных временных меток в выборках от разных сенсоров в рамках каждой сессии записи (перед началом сессии проводится калибровка положения датчиков и фиксация единого времени отсчета).

Процедура 5. Применение фильтров для снижения шумов (для IMU используется комбинация из фильтра Калмана и ориентационного фильтра Маджевика, показавшего свою высокую эффективность, для остальных источников фильтры или процедуры преобразования применяются при необходимости).

В результате произведенных преобразований получим множество обработанных данных E, среди которых заданы соответствующие выбранным сенсорам подмножества: $E = \{E_{EMG}, E_{IMU}, E_{CV}, E_{VR}\}$.

Для каждого источника данных задана постоянная частота записи, которую можно обозначить через множество $FPS = \{\nu_{EMG} = 250, \nu_{IMU} = 80, \nu_{CV} = 30, \nu_{VR} = 65\}$ (Frames Per Second, FPS), где каждый элемент принимает целые значения в герцах.

Выполним несколько преобразований для подготовки исходных данных для обучения моделей. Необходимо сохранить исходные данные после их разметки по категориям действий. Очевидно, что количество записей в каждом замере в каждом источнике значительно варьируется и напрямую зависит от частоты записи данных и длины размеченного интервала. Все это может затруднить подготовку и обучение моделей машинного обучения. Для решения данной проблемы предлагается выравнивание размерности входных данных под единую длину с учетом наибольшей частоты среди всех источников путем аппроксимации. Данный подход приведет к тому, что все данные будут представлены в виде матриц с одинаковой частотой v_{all} (эта частота будет соответствовать наибольшей и равна v_{EMG}). Недостаток данного подхода заключается в том, что объем данных будет значительно увеличен (например, для CV — до 8,3 раз). С другой стороны имеются следующие преимущества подхода: все данные синхронизированы между собой, что позволяет сравнивать равные по индексам замеры полученные от разных источников (это может использоваться в решении задачи прогнозирования уточненных значений); все данные текущего замера могут быть объединены в единую матрицу размерностью $N_{all,j} \times M_{all}$, где количество записей $N_{all,j}$ соответствует наибольшему количеству записей от источника в текущем замере, а количество столбцов M_{all} — суммарному количеству признаков от всех источников S.

Однако выполнение выравнивания в рамках замера недостаточно, так как каждый замер имеет свое значение $N_{all,j}$, которое может значительно отличаться. Данный аспект затруднит использование таких временных рядов различной длины при обучении модели, поэтому предлагается дополнительно осуществить общее выравнивание всех данных по следующей процедуре. Введем константу времени T_{all}, одинаковую для оценки всех действий. Предлагается руководствоваться правилом, что значение T_{all} превышает 95 % времени всех замеренных упражнений. Тогда осуществим выравнивание всех замеров к данной константе: $T_j \to T_{all}$, что приведет к изменению длины замера от $N_{all,j}$ к $N_{all} = v_{all}T_{all}$, которое будет одинаковым для всех замеров. Соответствующие данные будут аппроксимированы в сторону растяжения или сжатия длины временного ряда.

Таким образом, входные данные для обучения моделей получены на основе исходных массивов данных от каждого источника $E = \{E_{EMG}, E_{IMU}, E_{CV}, E_{VR}\}$ и представляют собой массив объединенных данных X от всех источников после второго выравнивающего преобразования (в рамках всего набора данных).

В качестве выходных данных выступают $Y = \{Y_{Kj} \in \mathbb{N}\}_{j=1\ldots J}$ — идентификаторы категории действий пользователя, заданные для каждого замера.

Тогда для решения поставленной задачи необходимо найти такую модель Machine Learning (ML), обеспечивающую наибольшую точность классификации:

$$ML: X \rightarrow Y.$$

Выбор моделей машинного обучения для решения задачи классификации

С учетом проведенного обзора существующих исследований в области применения различных моделей машинного обучения для классификации двигательной активности, а также опыта коллектива в применении различных моделей при решении задач классификации был выполнен отбор набора моделей [14, 15]. Выбор гиперпараметров моделей производился путем либо аналитически на основе предыдущего опыта, либо с использованием метода GridSearchCV для поиска оптимальной глубины деревьев (для тех моделей, где они используются). Для линейных и прочих моделей использовались параметры по умолчанию, если не указано иного. Получен следующий перечень моделей машинного обучения.

1. Logistic Regression. Линейная модель, использующая логистическую функцию для расчета вероятности принадлежности объекта к заданному классу. Используется с параметрами по умолчанию.
2. Nearest Neighbors Classification. Метод классификации, основанный на поиске ближайших соседей в пространстве признаков. В качестве количества соседей выбрано значение 5.
3. Decision Tree Classifier. Дерево решений, используемое для классификации. В качестве параметра глубины дерева выбрано значение 10.
4. Random Forest Classifier. Ансамбль деревьев решений с ограниченной глубиной деревьев (модель 3) и количеством деревьев равным 10. Объединяя прогнозы нескольких слабых моделей, метод снижает дисперсию и повышает устойчивость к шуму.
5. AdaBoost Classifier. Метод бустинга, который итеративно обучает 50 слабых классификаторов с использованием алгоритма SAMME. Каждый новый классификатор фокусируется на ошибках предыдущих, а итоговое решение получается посредством взвешенного голосования.
6. Gaussian Naive Bayes. Наивный байесовский классификатор, предполагающий независимость признаков и использующий нормальное распределение для оценки вероятностей. Используются параметры по умолчанию.
7. XGBClassifier. Градиентный бустинг, применяемый для классификации. В качестве основных параметров выбраны n_estimators=50 и max_depth=5.
8. Stacking Classifier. Ансамблевая модель, объединяющая прогнозы нескольких базовых классификаторов (Logistic Regression, ближайших соседей, двух вариантов Decision Tree с глубиной 5 и 10). Финальный мета-классификатор (Logistic Regression) объединяет прогнозы классификаторов для получения итогового решения.
9. Voting Classifier. Ансамблевая модель, которая объединяет прогнозы базовых моделей посредством «soft voting». Итоговая вероятность для каждого класса вычисляется как средневзвешенное значение вероятностей, предсказанных базовыми моделями. Используется перечень моделей, аналогичный представленному в п. 8.

Процесс распознавания типа движения является задачей многоклассовой классификации на конечном множестве движений. Определим множество классов возможных движений как C, $\{Y_{Kj} \in C\}_{j=1\ldots J}$. Мерой качества решения задачи многоклассовой классификации могут выступать следующие оценки [16, 17]:

— кроссэнтропийная потеря, используемая в качестве функции потерь при обучении нейронных сетей:

$$H(y, \hat{y}) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} \ln(\hat{y}_{i,j}),$$

где N — количество примеров в тестовой выборке; C — количество классов; $y_{i,j}$ — истинная метка класса j для примера i; $\hat{y}_{i,j}$ — предсказанная вероятность принадлежности примера i к классу j;

— точность классификации (доля правильно предсказанных классов):

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = y_i)}{N},$$

где $1(\hat{y}_i = y_i)$ — индикаторная функция (равна 1, если предсказание совпадает с истинным классом, иначе 0);

— средняя Precision:

$$\text{Precision} = \frac{1}{C}\sum_{k=1}^{C} \text{Precision}_k,$$

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k},$$

где TP_k (True Positives) — число объектов, правильно определенных в класс k; FP_k (False Positives) — число объектов, ошибочно отнесенных к классу k;

— средняя Recall:

$$\text{Recall} = \frac{1}{C}\sum_{k=1}^{C} \text{Recall}_k,$$

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k},$$

где FN_k (False Negatives) — число объектов класса k, ошибочно отнесенных к другим классам;

— F1-score (гармоническое среднее Precision и Recall):

$$\text{F1} = \frac{1}{C}\sum_{k=1}^{C} \text{F1}_k,$$

$$\text{F1}_k = \frac{2\text{Precision}_k\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}.$$

Соответственно, в качестве элементов тестовых выборок будут выступать элементы множества Y, не

участвовавшие в процессе обучения моделей. Также в ходе оценки моделей машинного обучения при решении задачи классификации большой интерес имеет важность признаков с целью выявления наиболее информативных составляющих мультисенсорных данных. Это может быть реализовано в процессе абляционного анализа: модели сравниваются на наборах данных с отдельными видами сенсоров (EMG, IMU, CV, VR). Такой анализ позволит количественно оценить, насколько наличие каждого типа сигнала влияет на итоговую точность классификации.

Результаты экспериментальных исследований

В рамках эксперимента были собраны данные о двигательной активности рук с использованием мультисенсорной системы, включающей EMG-датчики, IMU, VR-трекеры, систему CV. Эксперимент состоял из следующих этапов.

Этап 1. Подготовка программного обеспечения для синхронной записи данных.

Этап 2. Подготовка оборудования и калибровка датчиков.

Этап 3. Сбор данных. Участники выполняли набор заданных движений, представленный в табл. 1, включающий сгибание/разгибание локтя, круговые движения кисти, а также перемещения вдоль различных осей. Каждое движение повторялось несколько раз для получения достаточного объема данных.

Этап 4. Разметка данных. Для разметки данных использовалось разработанное программное обеспечение, которое позволяет создавать сессии для записи отдельных упражнений, визуально отслеживать по видео записанные движения и записывать время начала и окончания движения для последующего извлечения данных. При разметке осуществляется синхронное извлечение данных ото всех источников с учетом времени начала и окончания, выбранного сотрудником, отвечающим за разметку. Каждый фрагмент сохраняется отдельным файлом в формате csv.

Этап 5. Обработка сигналов. Проводится необходимая фильтрация шумов с применением полосовых фильтров, после чего формируется единая, синхронизированная выборка из источников, приходящих с разной частотой и в разных форматах. Этап 5 включает совмещение всех данных на единой временной шкале, которая затем растягивается до фиксированного времени в 5 с (наиболее распространенная длина движения), что позволяет получить единообразную размерность для всех записей в 1200 строк, которые затем поступают на вход моделей.

Этап 6. Обучение моделей. На основе выделенных признаков и синхронизированных данных были обучены девять моделей для решения задачи классификации. Этап 6 включает исследование абляции моделей классификации с целью выделения влияния отдельных источников данных на точность решения задачи.

В результате проведенного эксперимента с привлечением 15 частников собрано 629 записей, распределение упражнений представлено в табл. 1, размерность каждой записи составляет 1200 строк с 78 значениями. Каждой записи соответствует размеченная категория. Обучение моделей машинного обучения проводилось на 80 % полученных данных, оставшиеся 20 % использовались для перекрестной пятикратной проверки.

На этапе 1 проведено обучение моделей на полном наборе данных. Результаты обучения представлены в табл. 2. Сравнение моделей проведено по метрикам точности (Accuracy), Precision, Recall, F1-score по результатам перекрестной пятикратной проверки, также отражено время одного прогноза моделью. Результаты упорядочены по убыванию средней точности.

Выполнено исследование влияния используемого источника на точность классификации. Полученные результаты представлены в табл. 3. Лучшие результаты по каждому источнику выделены полужирным. Значения Accuracy получены по результатам перекрестной пятикратной проверки.

Таблица 1. Описание категорий упражнений

Table 1. Description of exercise categories

Описание упражнения		Количество собранных записей
Сгибание и разгибание руки в локте		70
Сгибание и разгибание запястья		70
Протягивание руки перед собой		70
Разведение руки вбок и приведение обратно к груди		67
Круговые движения рукой вдоль тела		71
Движение ладони перед телом вдоль оси	*X*	64
	Y	79
	Z	75
Имитация захвата предмета вытянутой рукой и перемещение к груди		63

Таблица 2. Сравнение моделей при классификации, %
Table 2. Comparison of models in classification

Модель	Accuracy EMG	Precision	Recall	F1-score	Время, мс
AdaBoost	98,89 ± 0,39	99,01 ± 0,33	98,86 ± 0,37	98,89 ± 0,37	4,4
XGBClassifier	97,94 ± 0,81	98,06 ± 0,76	97,93 ± 0,77	97,93 ± 0,79	26,7
Voting Classifier	97,14 ± 1,29	97,36 ± 1,29	97,15 ± 1,23	97,11 ± 1,31	48,6
Stacking Classifier	96,98 ± 1,84	97,34 ± 1,60	96,99 ± 1,81	97,00 ± 1,83	48,4
Decision Tree	94,12 ± 1,28	94,54 ± 1,22	94,12 ± 1,25	94,07 ± 1,24	0,1
Gaussian Naive Bayes	89,98 ± 1,96	91,48 ± 1,18	90,04 ± 1,89	90,06 ± 1,97	4,4
Nearest Neighbors	89,83 ± 3,26	91,38 ± 2,39	90,00 ± 3,19	89,90 ± 3,33	49,1
Logistic Regression	89,04 ± 3,06	89,75 ± 2,69	89,20 ± 3,06	89,03 ± 2,97	0,1
Random Forest	79,02 ± 7,23	77,99 ± 8,06	79,10 ± 7,18	75,70 ± 8,32	15,9

Таблица 3. Исследование абляции при классификации, %
Table 3. The study of ablation in classification

Модель	EMG	IMU	CV	VR
Logistic Regression	60,89 ± 1,35	88,72 ± 3,59	93,80 ± 1,83	96,98 ± 1,27
Nearest Neighbors	**83,47 ± 4,04**	89,83 ± 3,26	94,92 ± 2,33	96,03 ± 2,19
Decision Tree	50,71 ± 3,96	89,83 ± 2,85	88,40 ± 4,20	92,69 ± 1,15
Random Forest	44,99 ± 1,90	67,42 ± 6,00	76,65 ± 8,65	77,75 ± 4,28
AdaBoost	72,49 ± 3,58	**97,14 ± 0,39**	**96,19 ± 1,62**	**98,73 ± 1,78**
Gaussian Naive Bayes	51,04 ± 4,80	76,93 ± 6,69	83,62 ± 2,10	80,76 ± 4,41
XGBClassifier	72,33 ± 3,66	96,66 ± 1,29	93,48 ± 2,26	97,14 ± 0,81
Stacking Classifier	83,15 ± 2,72	92,37 ± 2,22	95,71 ± 2,22	97,62 ± 2,07
Voting Classifier	71,54 ± 1,70	94,60 ± 2,53	95,23 ± 1,94	96,67 ± 1,69

По результатам, полученным в табл. 2 и 3, можно сделать следующие выводы. Проведенный анализ всех данных показал, что наилучшую точность демонстрирует метод AdaBoost со значением более 98,8 %. Близкие результаты показали модели XGBClassifier и Voting Classifier, однако имели более низкую производительность и точность. При исследовании абляции моделей выявлены следующие закономерности: VR-трекеры являются наиболее информативным и содержательным источником данных (за счет объединения значений абсолютных координат и поворотов), что, в итоге, позволяет получить точность до 100 % для модели AdaBoost на некоторых выборках (в среднем 98,73 ± 1,78 %). Учитывая специфичность данного типа источника данных (современные шлемы переходят на системы CV и отказываются от использования базовых станций), стоит обратить внимание на остальные составляющие мультисенсорного сигнала. CV и IMU показывают достаточно высокие результаты, приемлемые для многих сценариев использования. Результаты EMG в данной ситуации являются самым ненадежным и сложным источником данных, неприменимым для большинства моделей.

В результате проведенного исследования необходимо отметить, что для решения поставленной задачи классификации двигательной активности можно использовать как всю совокупность мультисенсорных данных, так и отдельные ее составляющие (за исключением EMG). В ходе сравнения моделей предпочтение отдается ансамблям, наилучшие показатели имели такие модели как AdaBoost, XGBClassifier, Voting Classifier и Stacking Classifier. Подчеркнем, что для EMG модель Nearest Neighbors показывала высокие результаты.

Дополнительно был выполнен анализ полученных результатов с существующими исследованиями в данной области, чтобы оценить такие факторы, как используемая архитектура и источник данных. Полученные результаты сравнения представлены в табл. 4.

Проведенное сравнение показывает, что нейросетевые модели на базе LSTM, GRU, Transformer и нейросетевых архитектур могут обеспечивать высокую точность, превышающую 99 %, благодаря эффективному извлечению пространственно-временных признаков. С другой стороны, такие модели, судя по существующему опыту и работе [18], не обладают хорошей производительностью и не могут использоваться в режиме реального времени или с большой частотой вызова. Небольшое увеличение точности полностью не компенсирует значительного снижения быстродействия.

Также по сравнению с существующими исследованиями в рамках данной работы рассмотрены не два или один источник данных, а четыре, проведен абля-

Таблица 4. Сравнение точности классификации со сторонними исследованиями
Table 4. Comparison of classification accuracy with third-party studies

Источник	Источник данных	Тип модели	Точность классификации, %	Время расчета, мс
[18]	EMG + IMU	LSTM-Res	99,67	3220
[18]	EMG + IMU	GRU-Res	99,49	2820
[18]	EMG + IMU	Transformer-CNN	98,96	2450
[13]	EMG + IMU	DQN (глубокая Q-сеть)	97,50 ± 1,13	33
[19]	CV (MediaPipe)	Swin Transformer	99,7	Нет данных
[20]	VR-трекеры	CNN-Transformer	100	Нет данных
Настоящая работа	EMG + IMU + CV(MediaPipe) + VR-трекеры	AdaBoost	98,89 ± 0,39	4,4

ционный анализ их вклада в итоговую точность классификации. Тем не менее, рассмотренное сравнение показывает перспективность дальнейших исследований по реализации нейросетевых архитектур для решения данной задачи с учетом необходимости оптимизации данных архитектур для их высокой производительности без ущерба для точности.

Заключение

Специализированные трекеры для виртуальной реальности (Virtual Reality, VR) для абсолютного позиционирования требуют дополнительного оборудования в виде базовых

Рассмотрена задача комплексного анализа двигательной активности человека для реабилитационных и тренировочных VR-комплексов на основе мультисенсорных данных (Inertial Measurement Unit (IMU), Electromyography (EMG), компьютерное зрение и VR-трекеры). Предложен подход к обработке мультисенсорных данных, включающий двойное выравнивание временных рядов для синхронизации разнородных потоков и формирования унифицированного набора данных без потери информативности.

Эксперимент с 15 участниками и 629 размеченными записями подтвердил эффективность применения ансамблевых моделей машинного обучения. Метод AdaBoost продемонстрировал наибольшую точность и стабильность (98,89 ± 0,39 %) при полном наборе сенсоров. В ходе исследования абляции наибольший вклад внесли VR-трекеры (до 98,73 ± 1,78 % точности при использовании AdaBoost). IMU и система компьютерного зрения обеспечили достаточную точность (96–97 %), EMG-канал показал самую низкую информативность (83 %).

Полученные результаты количественно доказали преимущества мультисенсорного подхода, так как собранные данные в дальнейшем могут быть обработаны для получения наибольшей точности. Однако необходимо учитывать, что некоторые источники данных могут снижать общую точность (как в случае информации о мышечной активности, полученной в ходе EMG). Практическая значимость работы заключается в том, что для массовых и ресурсо-ограниченных решений достаточно комбинации IMU и компьютерного зрения, так как в ряде случаев добавление VR-трекеров невозможно. Использование только одного EMG целесообразно в узкоспециализированных сценариях, ориентированных только на оценку мышечной активности.

Таким образом, разработанный подход и обученные модели могут быть использованы в рамках реабилитационных и профессиональных виртуальных тренажеров и комплексов для распознавания типов двигательной активности пользователя, а в дальнейшем и оценки качества выполнения действий. Более того, результаты абляционного анализа позволяют оптимизировать конфигурацию датчиков под конкретные приложения и требования предметной области. В частности показано, что для массовых решений достаточно комбинации IMU и/или компьютерного зрения, тогда как добавление VR-трекеров может быть опциональным. Направлением дальнейших исследований является исследование применимости более сложных архитектур, например, сверточных нейронных сетей, Transformer-ов или их сочетаний для решения задачи классификации.

Литература

1. Obukhov A., Volkov A., Pchelintsev A., Nazarova A., Teselkin D., Surkova E., Fedorchuk I. Examination of the accuracy of movement tracking systems for monitoring exercise for musculoskeletal rehabilitation // Sensors. 2023. V. 23. N 19. P. 8058. https://doi.org/10.3390/s23198058
2. Obukhov A., Dedov D., Volkov A., Teselkin D. Modeling of nonlinear dynamic processes of human movement in virtual reality based on digital shadows // Computation. 2023. V. 11. N 5. P. 85. https://doi.org/10.3390/computation11050085

References

1. Obukhov A., Volkov A., Pchelintsev A., Nazarova A., Teselkin D., Surkova E., Fedorchuk I. Examination of the accuracy of movement tracking systems for monitoring exercise for musculoskeletal rehabilitation. *Sensors*, 2023, vol. 23, no. 19, pp. 8058. https://doi.org/10.3390/s23198058
2. Obukhov A., Dedov D., Volkov A., Teselkin D. Modeling of nonlinear dynamic processes of human movement in virtual reality based on digital shadows. *Computation*, 2023, vol. 11, no. 5, pp. 85. https://doi.org/10.3390/computation11050085

3. Islam M.M., Nooruddin S., Karray F., Muhammad G. Human activity recognition using tools of convolutional neural networks: a state of the art review, data sets, challenges, and future prospects // Computers in Biology and Medicine. 2022. V. 149. P. 106060. https://doi.org/10.1016/j.compbiomed.2022.106060
4. Ergun B.G., Şahiner R. Embodiment in virtual reality and augmented reality games: an investigation on user interface haptic controllers // Journal of Soft Computing and Artificial Intelligence. 2023. V. 4. N 2. P. 80–92. https://doi.org/10.55195/jscai.1409156
5. Franček P., Jambrosic K., Horvat M., Planinec V. The performance of inertial measurement unit sensors on various hardware platforms for binaural head-tracking applications // Sensors. 2023. V. 23. N 2. P. 872. https://doi.org/10.3390/s23020872
6. Ghorbani F., Ahmadi A., Kia M., Rahman Q., Delrobaei M. A decision-aware ambient assisted living system with IoT embedded device for in-home monitoring of older adults // Sensors. 2023. V. 23. N 5. P. 2673. https://doi.org/10.3390/s23052673
7. Eliseichev E.A., Mikhailov V.V., Borovitskiy I.V., Zhilin R.M., Senatorova E.O. A review of devices for detection of muscle activity by surface electromyography // Biomedical Engineering. 2022. V. 56. N 1. P. 69–74. https://doi.org/10.1007/s10527-022-10169-4
8. Chung J.L., Ong L.Y., Leow M.C. Comparative analysis of skeleton-based human pose estimation // Future Internet. 2022. V. 14. N 12. P. 380. https://doi.org/10.3390/fi14120380
9. Zhang S., Li Y., Zhang S., Shahabi F., Xia S., Deng Y., Alshurafa N. Deep learning in human activity recognition with wearable sensors: a review on advances // Sensors. 2022. V. 22. N 4. P. 1476. https://doi.org/10.3390/s22041476
10. Lin J.J., Hsu C.K., Hsu W.L., Tsao T.C., Wang F.C., Yen J.Y. Machine learning for human motion intention detection // Sensors. 2023. V. 23. N 16. P. 7203. https://doi.org/10.3390/s23167203
11. Mazon D.M., Groefsema M., Schomaker L.R.B., Carloni R. IMU-based classification of locomotion modes, transitions, and gait phases with convolutional recurrent neural networks // Sensors. 2022. V. 22. N 22. P. 8871. https://doi.org/10.3390/s22228871
12. Gonzales-Huisa O.A., Oshiro G., Abarca V.E., Chavez-Echajaya J.G., Elias D.A. EMG and IMU data fusion for locomotion mode classification in transtibial amputees // Prosthesis. 2023. V. 5. N 4. P. 1232–1256. https://doi.org/10.3390/prosthesis5040085
13. Vásconez J.P., López L.I.B., Caraguay A.L.V., Benalcázar M.E. Hand gesture recognition using EMG-IMU signals and deep q-networks // Sensors. 2022. V. 22. N 24. P. 9613. https://doi.org/10.3390/s22249613
14. Sulla-Torres J., Gamboa A.C., Llanque C.A., Osorio J.A., Carnero M.Z. Classification of motor competence in schoolchildren using wearable technology and machine learning with hyperparameter optimization // Applied Sciences. 2024. V. 14. N 2. P. 707. https://doi.org/10.3390/app14020707
15. Stančić I., Music J., Grujic T., Vasic M.K., Bonkovic M. Comparison and evaluation of machine learning-based classification of hand gestures captured by inertial sensors // Computation. 2022. V. 10. N 9. P. 159. https://doi.org/10.3390/computation10090159
16. Ogundokun R.O., Maskeliunas R., Misra S., Damasevicius R. Hybrid inceptionv3-svm-based approach for human posture detection in health monitoring systems // Algorithms. 2022. V. 15. N 11. P. 410. https://doi.org/10.3390/a15110410
17. Farhadpour S., Warner T.A., Maxwell A.E. Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices // Remote Sensing. 2024. V. 16. N 3. P. 533. https://doi.org/10.3390/rs16030533
18. Jiang Y., Song L., Zhang J., Song Y., Yan M. Multi-category gesture recognition modeling based on sEMG and IMU signals // Sensors. 2022. V. 22. N 15. P. 5855. https://doi.org/10.3390/s22155855
19. Lin W.C., Tu Y.C., Lin H.Y., Tseng M.H. A comparison of deep learning techniques for pose recognition in Up-and-Go pole walking exercises using skeleton images and feature data // Electronics. 2025. V. 14. N 6. P. 1075. https://doi.org/10.3390/electronics14061075
20. Mohammadzadeh A.K., Alinezhad E., Masoud S. Neural-Network-Driven intention recognition for enhanced Human–Robot Interaction: a Virtual-Reality-Driven approach // Machines. 2025. V. 13. N 5. P. 414. https://doi.org/10.3390/machines13050414

3. Islam M.M., Nooruddin S., Karray F., Muhammad G. Human activity recognition using tools of convolutional neural networks: a state of the art review, data sets, challenges, and future prospects. *Computers in Biology and Medicine*, 2022, vol. 149, pp. 106060. https://doi.org/10.1016/j.compbiomed.2022.106060
4. Ergun B.G., Şahiner R. Embodiment in virtual reality and augmented reality games: an investigation on user interface haptic controllers. *Journal of Soft Computing and Artificial Intelligence*, 2023, vol. 4, no. 2, pp. 80–92. https://doi.org/10.55195/jscai.1409156
5. Franček P., Jambrosic K., Horvat M., Planinec V. The performance of inertial measurement unit sensors on various hardware platforms for binaural head-tracking applications. *Sensors*, 2023, vol. 23, no. 2, pp. 872. https://doi.org/10.3390/s23020872
6. Ghorbani F., Ahmadi A., Kia M., Rahman Q., Delrobaei M. A decision-aware ambient assisted living system with IoT embedded device for in-home monitoring of older adults. *Sensors*, 2023, vol. 23, no. 5, pp. 2673. https://doi.org/10.3390/s23052673
7. Eliseichev E.A., Mikhailov V.V., Borovitskiy I.V., Zhilin R.M., Senatorova E.O. A review of devices for detection of muscle activity by surface electromyography. *Biomedical Engineering*, 2022, vol. 56, no. 1, pp. 69–74. https://doi.org/10.1007/s10527-022-10169-4
8. Chung J.L., Ong L.Y., Leow M.C. Comparative analysis of skeleton-based human pose estimation. *Future Internet*, 2022, vol. 14, no. 12, pp. 380. https://doi.org/10.3390/fi14120380
9. Zhang S., Li Y., Zhang S., Shahabi F., Xia S., Deng Y., Alshurafa N. Deep learning in human activity recognition with wearable sensors: a review on advances. *Sensors*, 2022, vol. 22, no. 4, pp. 1476. https://doi.org/10.3390/s22041476
10. Lin J.J., Hsu C.K., Hsu W.L., Tsao T.C., Wang F.C., Yen J.Y. Machine learning for human motion intention detection. *Sensors*, 2023, vol. 23, no. 16, pp. 7203. https://doi.org/10.3390/s23167203
11. Mazon D.M., Groefsema M., Schomaker L.R.B., Carloni R. IMU-based classification of locomotion modes, transitions, and gait phases with convolutional recurrent neural networks. *Sensors*, 2022, vol. 22, no. 22, pp. 8871. https://doi.org/10.3390/s22228871
12. Gonzales-Huisa O.A., Oshiro G., Abarca V.E., Chavez-Echajaya J.G., Elias D.A. EMG and IMU data fusion for locomotion mode classification in transtibial amputees. *Prosthesis*, 2023, vol. 5, no. 4, pp. 1232–1256. https://doi.org/10.3390/prosthesis5040085
13. Vásconez J.P., López L.I.B., Caraguay A.L.V., Benalcázar M.E. Hand gesture recognition using EMG-IMU signals and deep q-networks. *Sensors*, 2022, vol. 22, no. 24, pp. 9613. https://doi.org/10.3390/s22249613
14. Sulla-Torres J., Gamboa A.C., Llanque C.A., Osorio J.A., Carnero M.Z. Classification of motor competence in schoolchildren using wearable technology and machine learning with hyperparameter optimization. *Applied Sciences*, 2024, vol. 14, no. 2, pp. 707. https://doi.org/10.3390/app14020707
15. Stančić I., Music J., Grujic T., Vasic M.K., Bonkovic M. Comparison and evaluation of machine learning-based classification of hand gestures captured by inertial sensors. *Computation*, 2022, vol. 10, no. 9, pp. 159. https://doi.org/10.3390/computation10090159
16. Ogundokun R.O., Maskeliunas R., Misra S., Damasevicius R. Hybrid inceptionv3-svm-based approach for human posture detection in health monitoring systems. *Algorithms*, 2022, vol. 15, no. 11, pp. 410. https://doi.org/10.3390/a15110410
17. Farhadpour S., Warner T.A., Maxwell A.E. Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing*, 2024, vol. 16, no. 3, pp. 533. https://doi.org/10.3390/rs16030533
18. Jiang Y., Song L., Zhang J., Song Y., Yan M. Multi-category gesture recognition modeling based on sEMG and IMU signals. *Sensors*, 2022, vol. 22, no. 15, pp. 5855. https://doi.org/10.3390/s22155855
19. Lin W.C., Tu Y.C., Lin H.Y., Tseng M.H. A comparison of deep learning techniques for pose recognition in Up-and-Go pole walking exercises using skeleton images and feature data. *Electronics*, 2025, vol. 14, no. 6, pp. 1075. https://doi.org/10.3390/electronics14061075
20. Mohammadzadeh A.K., Alinezhad E., Masoud S. Neural-Network-Driven intention recognition for enhanced Human–Robot Interaction: a Virtual-Reality-Driven approach. *Machines*, 2025, vol. 13, no. 5, pp. 414. https://doi.org/10.3390/machines13050414

Автор

Обухов Артём Дмитриевич — доктор технических наук, профессор, ведущий научный сотрудник, Тамбовский государственный технический университет, Тамбов, 392000, Российская Федерация, SC 56104232400, https://orcid.org/0000-0002-3450-5213, obuhov.art@gmail.com

Author

Artem D. Obukhov — D.Sc., Professor, Leading Researcher, Tambov State Technical University, Tambov, 392000, Russian Federation, SC 56104232400, https://orcid.org/0000-0002-3450-5213, obuhov.art@gmail.com

Статья поступила в редакцию 21.05.2025
Одобрена после рецензирования 24.08.2025
Принята к печати 21.09.2025

Received 21.05.2025
Approved after reviewing 24.08.2025
Accepted 21.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-844-855
УДК 004.05:616

Универсальная модель архитектуры краудсорсинговой системы разметки и подготовки медицинских данных

Лев Алексеевич Коваленко[1]✉, Иван Станиславович Блеканов[2], Федор Валерьевич Ежов[3], Евгений Сергеевич Ларин[4], Глеб Ирламович Ким[5]

[1,2,3,4,5] Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация
[1] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[5] Санкт-Петербургский государственный университет, Клиника высоких медицинских технологий им. Н.И. Пирогова, Санкт-Петербург, 190020, Российская Федерация

[1] levozavr1@gmail.com✉, https://orcid.org/0009-0007-8233-4387
[2] i.blekanov@spbu.ru, https://orcid.org/0000-0002-7305-1429
[3] st056053@student.spbu.ru, https://orcid.org/0009-0007-1468-0042
[4] evgeny.larin@spbu.ru, https://orcid.org/0009-0007-6199-3607
[5] gikim.cor@gmail.com, https://orcid.org/0000-0002-9344-5724

Аннотация

Введение. Методы машинного обучения и искусственного интеллекта в целом все больше используются для обработки и интеллектуального анализа медицинских данных. Для применения этих методов требуется наличие специализированных наборов размеченных данных больших размеров. Организация процесса качественной разметки медицинских данных требует привлечения большого числа асессоров и экспертов в конкретной области медицины, а также наличие специализированных инструментов, упрощающих данный процесс и учитывающих специфику обработки медицинских данных. **Метод.** В работе предложена универсальная архитектурная модель краудсорсинговой системы, специализированной для разметки медицинских данных. Модель поддерживает обработку различных медицинских форматов данных, имеет механизмы анонимизации и многоуровневого контроля качества, позволяет организовать распределенный процесс разметки с привлечением экспертного сообщества. **Основные результаты.** Приведена классификация актуальных проблем процесса сбора и разметки медицинских данных, сформулированы критерии качества и безопасности для сравнительного анализа систем разметки медицинских данных. Предложена схема обобщенного сценария взаимодействия групп пользователей с краудсорсинговой системой в контексте решения задач искусственного интеллекта в области медицины. Построена универсальная модель архитектуры такой системы. На ее основе реализована специализированная краудсорсинговая система разметки медицинских данных на базе Computer Vision Annotation Tool. Проведено тестирование и апробация реализованной системы кардиохирургами Клиники высоких медицинских технологий им. Н.И. Пирогова Санкт-Петербургского государственного университета. **Обсуждение.** Предложенная модель архитектуры краудсорсинговой системы может быть использована для повышения эффективности и безопасности организации и построения процесса разметки медицинских данных пациентов при решении различных прикладных задач машинного обучения/искусственного интеллекта, таких как семантическая сегментация внутренних органов и их патологий, детекция и классификация заболеваний по медицинским снимкам (например, компьютерной томографии). Разработанное решение может использоваться врачами различной специализации, исследователями и разработчиками, направленными на развитие и создание методов и технологий искусственного интеллекта в области медицины.

Ключевые слова

краудсорсинг, разметка медицинских данных, модель программной архитектуры, критерии качества систем краудсорсинга, обработка медицинских данных, пользовательский сценарий взаимодействия

Ссылка для цитирования: Коваленко Л.А., Блеканов И.С., Ежов Ф.В., Ларин Е.С., Ким Г.И. Универсальная модель архитектуры краудсорсинговой системы разметки и подготовки медицинских данных // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 844–855. doi: 10.17586/2226-1494-2025-25-5-844-855

A universal architecture model of a crowdsourcing medical data labeling system designed

Lev A. Kovalenko[1]✉, Ivan S. Blekanov[2], Fedor V. Ezhov[3], Evgenii S. Larin[4], Gleb I. Kim[5]

[1,2,3,4,5] St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation
[1] ITMO University, Saint Petersburg, 197101, Russian Federation
[5] Saint Petersburg State University Hospital, Saint Petersburg, 190020, Russain Federation

[1] levozavr1@gmail.com✉, https://orcid.org/0009-0007-8233-4387
[2] i.blekanov@spbu.ru, https://orcid.org/0000-0002-7305-1429
[3] st056053@student.spbu.ru, https://orcid.org/0009-0007-1468-0042
[4] evgeny.larin@spbu.ru, https://orcid.org/0009-0007-6199-3607
[5] gikim.cor@gmail.com, https://orcid.org/0000-0002-9344-5724

Abstract

Machine Learning (ML) and Artificial Intelligence (AI) methods are used to process and intelligently analyze medical data. The application of ML/AI methods requires specialized sets of labeled medical data of large dimensions. Process organization of quality medical data labeling requires the involvement of a large number assessors and specialists in a particular field of medicine as well as the availability of specialized tools for labeling process optimization considering the specifics of medical data processing. In this paper a universal architectural model of a crowdsourcing system specifically designed for medical data labeling was proposed. The model supports processing of diverse medical data formats, incorporates data anonymization mechanisms and multi-level quality control, while enabling a distributed annotation process with expert community involvement. As a result, classification of actual problems of the process of medical data labeling and data collection, and a quality and safety criteria for comparative analysis of medical data labeling systems was detected and formulated. The scheme of generalized scenario of users' groups interaction with crowdsourcing system in the context of solving AI problems in the field of medicine was proposed. A universal model of such system architecture was designed and a specialized crowdsourcing system of medical data labeling based on Computer Vision Annotation Tool was implemented on its basis. Testing and approbation of the realized system was carried out at the Pirogov Clinic of High Medical Technologies. The proposed universal model of crowdsourcing system architecture can be used to improve the efficiency and safety of organization and construction of the process of labeling patients' medical data in the context of solving various applied ML/AI tasks, such as semantic segmentation of internal organs and their pathologies, detection and classification of diseases based on medical images (e.g. computed tomography scans). The developed solution can be used by doctors of various specializations, researchers and developers aimed at the development and creation of methods and technologies of AI in the field of medicine.

Keywords

crowdsourcing, medical data annotation, software architecture model, quality criteria for crowdsourcing systems, medical data processing, use case

For citation: Kovalenko L.A., Blekanov I.S., Ezhov F.V., Larin E.S., Kim G.I. A universal architecture model of a crowdsourcing medical data labeling system designed. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 844–855 (in Russian). doi: 10.17586/2226-1494-2025-25-5-844-855

Введение

Сегодня искусственный интеллект (Artificial Intelligence, AI) и машинное обучение (Machine Learning, ML) широко применяются в здравоохранении для диагностики, прогнозирования и персонализированной терапии [1–4]. Эффективность ML/AI-моделей критически зависит от наличия больших объемов качественно размеченных данных, требующих участия экспертов [4–7]. Однако традиционные методы разметки не справляются с растущими объемами данных, что обуславливает необходимость специализированных краудсорсинговых платформ [8–11].

Подготовка медицинских наборов данных связана с рядом уникальных сложностей, обусловленных спецификой медицинских данных. В отличие от стандартных форматов (текст, изображения, видео), медицинские данные (Digital Imaging and Communications in Medicine (DICOM), Neuroimaging Informatics Technology Initiative (NifTI) и другие форматы) требуют специальной предобработки перед разметкой, а отсутствие единых стандартов их представления дополнительно усложняет этот процесс [12–15]. Особую проблему представляет обработка конфиденциальных медицинских сведений, требующая строгого соблюдения нормативных требований (ФЗ-152[1], ФЗ-323[2], GDPR[3], HIPAA[4]), включая обязательную анонимизацию и защиту от утечек, что существенно ограничивает возможности применения краудсорсинга [15, 16]. Дополнительные сложности возникают при контроле качества разметки, поскольку различия в квалификации врачей-экспертов, асессоров и неоднозначность медицинской информации повышают риск ошибок и требуют разработки специальных механизмов валидации [17–21].

Настоящая работа посвящена решению перечисленных проблем посредством проектирования и разработки системы краудсорсинга, адаптированной под специфику медицинских данных и взаимодей-

[1] Федеральный закон от 27.07.2006 №152-ФЗ «О персональных данных» (ред. от 24.02.2021).

[2] Федеральный закон от 21.11.2011 №323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации» (ред. от 01.07.2021).

[3] Regulation (EU) 2016/679 (General Data Protection Regulation) от 27.04.2016.

[4] Public Law 104-191 (Health Insurance Portability and Accountability Act) от 21.08.1996.

ствие с врачами-экспертами в удаленном формате. Рассматривается методология разметки медицинских данных врачами-экспертами, основанная на современных методах и инструментах краудсорсинга.

Актуальные проблемы краудсорсинга в медицине

Краудсорсинговая разметка медицинских данных — ключевой инструмент для создания наборов данных, необходимых в машинном обучении, однако его внедрение в процесс разработки AI-решений в медицине сталкивается с рядом проблем. Прежде всего, обработка медицинской информации требует строгого соблюдения норм конфиденциальности (ФЗ-152[1], ФЗ-323[2], GDPR[3], HIPAA[4]) и применения надежных методов анонимизации для защиты персональных медицинских данных пациентов [15, 16, 22–24].

Еще одной важной проблемой является обеспечение качества разметки медицинских данных. Различия в квалификации асессоров и субъективность интерпретации медицинских терминов и концепций могут приводить к ошибкам, что негативно влияет на надежность данных [17–19]. Для минимизации этих рисков предлагаются такие подходы, как обучение асессоров [25–27], кросс-аннотация с привлечением врачей-экспертов [28, 29], а также автоматизация валидации с использованием специализированных метрик и методов машинного обучения [18, 30–32].

Сложности также связаны с разнообразием форматов медицинских данных (DICOM, NIfTI и др.) и отсутствием единых стандартов их представления [12–14, 33]. Это требует разработки инструментов для преобразования медицинских данных в удобный для разметки вид без потери метаинформации, что особенно важно для обеспечения согласованности наборов данных [34, 35].

Большинство существующих платформ разметки не адаптированы для работы с медицинскими данными. Они часто не поддерживают специализированные форматы, не обеспечивают должного уровня безопасности и не предназначены для использования в закрытых контурах медицинских учреждений [13, 14, 36]. Эти ограничения делают актуальной разработку специализированных решений, учитывающих особенности медицинской отрасли.

В настоящей работе предлагаются классификация проблем и актуальные направления развития краудсорсинговых систем разметки медицинских данных. В табл. 1 выделено три класса проблем: обеспечение безопасности персональных данных, качества данных и качества получаемой разметки.

Учитывая обозначение в табл. 1 проблемы использования инструментов краудсорсинга для разметки медицинских данных, можно составить перечень критериев для оценки и сравнения существующих инструментов и платформ (табл. 2).

Существующие решения

Современные инструменты для разметки медицинских данных существенно различаются по функциональности — от простых аннотаторов изображений до комплексных систем с интеграцией машинного обучения и AI-технологий [37, 38]. По специализации их можно разделить на три категории: специализированные медицинские решения, универсальные платформы

Таблица 1. Классификация актуальных проблем краудсорсинговой разметки медицинских данных
Table 1. Classification of current challenges in crowdsourced medical data annotation

Класс проблем	Название проблемы	Описание проблемы
Безопасность данных пациентов	Соблюдение правовых норм и законов	При обработке персональных медицинских данных необходимо соблюдать законы, например ФЗ-152[1] и ФЗ-323[2]
	Анонимизация данных	Не анонимизированные медицинские данные содержат персональную информацию
	Утечка данных	Риски утечки персональных данных пациентов при передаче данных по сети
Обеспечение качества данных	Разнообразие форматов	Существует большое количество специфичных форматов медицинских данных
	Отсутствие единой стандартизации	Различные медучреждения используют свои стандарты и оборудование
Обеспечение качества разметки	Разный уровень квалификации	Разметка асессоров с различным уровнем квалификации может различаться
	Потребность во врачах-экспертах	Оценка качества полученной разметки требует привлечения врачей-экспертов
	Отсутствие методов валидации разметки	Из-за сложных форматов данных и задач разметки медицинских данных не применимы общеизвестные способы валидации разметки

[1] Федеральный закон от 27.07.2006 №152-ФЗ «О персональных данных» (ред. от 24.02.2021).

[2] Федеральный закон от 21.11.2011 №323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации» (ред. от 01.07.2021).

Таблица 2. Критерии качества существующих решений с учетом класса проблем
Table 2. Quality assessment criteria for existing solutions by problem category

Критерий качества	Класс проблем	Описание
Соответствие стандартам защиты медицинских данных	Безопасность данных пациентов	Наличие инструментов анонимизации и защиты персональных данных
Режим локального запуска		Возможность запуска платформы разметки в закрытом контуре
Импортозамещенность		Доступность системы в Российской Федерации и соответствие законам
Поддержка разных типов данных для разметки	Обеспечение качества данных	Возможность загружать для разметки данные в специализированных медицинских форматах
Наличие Application Programming Interface		Возможность автоматизировать поставку и предварительную подготовку данных
Удобство и гибкость интерфейса		Возможность настроить способы представления данных для разметки
Поддержка типов аннотаций	Обеспечение качества разметки	Поддержка аннотации текстов, изображений и видео для различных задач
Система автоматической разметки		Возможность подключения машинного обучения для проведения автоматической разметки данных
Управление процессом разметки		Функционал для отслеживания прогресса и статуса задач разметки данных
Ролевая модель		Встроенная ролевая модель для распределения прав пользователей
Система контроля качества разметки		Возможность вводить метрики или инструменты для контроля качества

для разметки стандартных форматов данных и гибридные системы. Вариативность наблюдается и в способах развертывания: open-source решения, коробочные продукты и облачные сервисы. Характеристики ключевых современных платформ представлены в табл. 3, где отражены их тип, направленность и основные функциональные возможности.

Для выбранного набора решений разметки медицинских данных можно провести сравнительный анализ, согласно критериям оценки в табл. 2. Результаты такого анализа приведены в табл. 4.

Выполненный анализ выявил ключевые недостатки рассмотренных решений: отсутствие механизмов анонимизации медицинских данных, слабую поддержку управления процессом разметки (включая ролевые модели и контроль качества), а также ограниченную доступность на территории Российской Федерации. Наиболее перспективными оказались решения CVAT и LabelStudio, которые, несмотря на отсутствие поддержки специализированных медицинских форматов (например, DICOM), в наибольшей степени соответствуют предъявляемым требованиям. На основе сравнитель-

Таблица 3. Современные решения для разметки медицинских данных
Table 3. Existing medical data labeling solutions

Решение	Тип решения	Направленность	Описание	Ссылка
3D Slicer	Коробочное	Медицинские данные	Позволяет сегментировать медицинские снимки	slicer.org
Supervisely	Облачное	Общее решение с поддержкой медицинских данных	Платная платформа разметки изображений, в том числе медицинских	supervisely.com
LabsV7			Платная система разметки данных с поддержкой медицинских снимков	v7labs.com
Encord			Платная платформа разметки с поддержкой DICOM и NifTi снимков	encord.com
AmazonMTurk		Медицинские данные	Платная платформа для создания аннотированных наборов медицинских данных	mturk.com
LabelStudio	Open-source	Общее решение	Открытое решение разметки общеизвестных форматов	labelstud.io
Computer Vision Annotation Tool (CVAT)			Открытое решение разметки общеизвестных форматов	cvat.ai

Таблица 4. Оценка существующих решений согласно критериям (табл. 2)
Table 4. Comparative analysis of existing solutions based on Table 2 criteria

Название критерия	Решения						
	3D Slicer	Super visely	Labs V7	Encord	Label Studio	CVAT	Amazon MTurk
Соответствие стандартам защиты медицинских данных	–	–	–	–	–	–	–
Локальный режим работы	+	–	–	–	+	+	–
Импортозамещенность	+	–	–	–	+	+	–
Поддержка разных типов данных для разметки	+	+	+	+	–	–	–
Наличие Application Programming Interface	–	–	–	–	+	+	+
Удобство и гибкость интерфейса	+	+	+	+	+	+	+
Поддержка различных типов аннотаций	–	+	+	+	+	+	+
Система автоматической разметки	–	+	+	+	+	+	+
Управление процессом разметки	–	+	–	–	+	+	+
Ролевая модель	–	–	–	–	–	+	–
Система контроля качества разметки	–	–	–	–	+	+	–

Примечание. «+» означает соответствие критерию, а «–» — не соответствие. Жирным шрифтом выделены наиболее перспективные решения.

ного анализа была разработана универсальная архитектура краудсорсинговой системы с использованием CVAT в качестве базовой платформы, дополненной функционалом для работы с медицинскими данными.

Проектирование модели и сценария пользовательского взаимодействия с системой краудсорсинга разметки медицинских данных

Разработка ML/AI-решений представляет собой сложный итеративный процесс [39–41], где краудсорсинговая система разметки играет ключевую роль в создании аннотированных наборов данных для медицинских задач, включая семантическую сегментацию анатомических структур [42–44] и диагностику по медицинским изображениям [45–47]. Общая схема процесса представлена на рис. 1.

Краудсорсинговая система разметки может применяться на различных стадиях процесса разработки AI-решения (рис. 1): сбор и подготовка данных, интерпретация и анализ работы моделей, а также для демонстрации результатов работы ML/AI-решения и сбора обратной связи от врачей-экспертов.

На рис. 2 отображен сценарий взаимодействия ключевых пользователей с краудсорсинговой системой разметки медицинских данных в процессе разработки ML/AI-решения. К ключевым пользователям системы разметки данных можно отнести четыре основные роли: врач, врач-эксперт, асессор и разработчик ML/AI. Обобщенный сценарий их взаимодействия с системой описан в алгоритме.

Алгоритм. Алгоритм обобщенного сценария взаимодействия различных групп пользователей с краудсорсинговой системой разметки медицинских данных состоит из следующих этапов.

Этап 1. Врач собирает и загружает медицинские данные пациентов в систему разметки.

Этап 2. Врач-эксперт создает задачи разметки на основе собранных данных и распределяет их на асессоров.

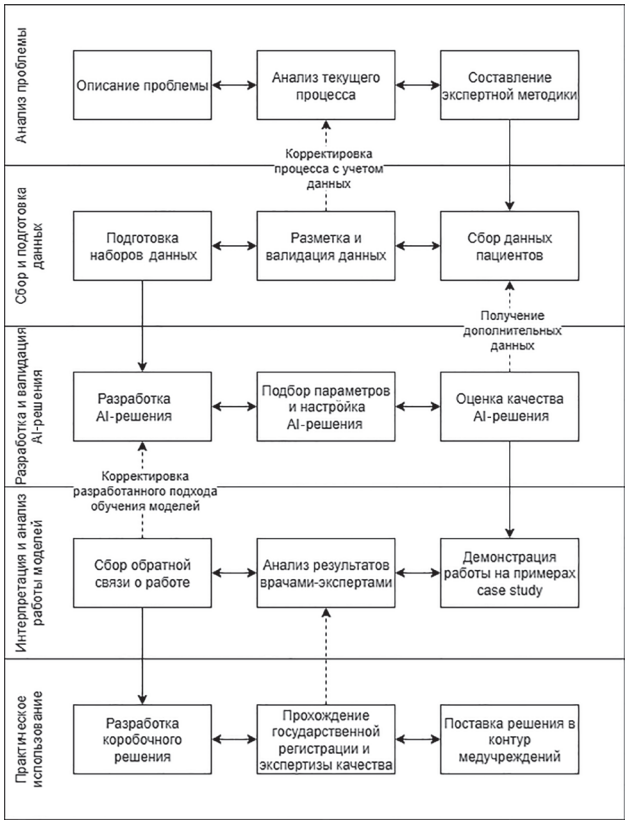

Рис. 1. Схема разработки AI-решения в области медицины
Fig. 1. AI solution development framework in medical domain

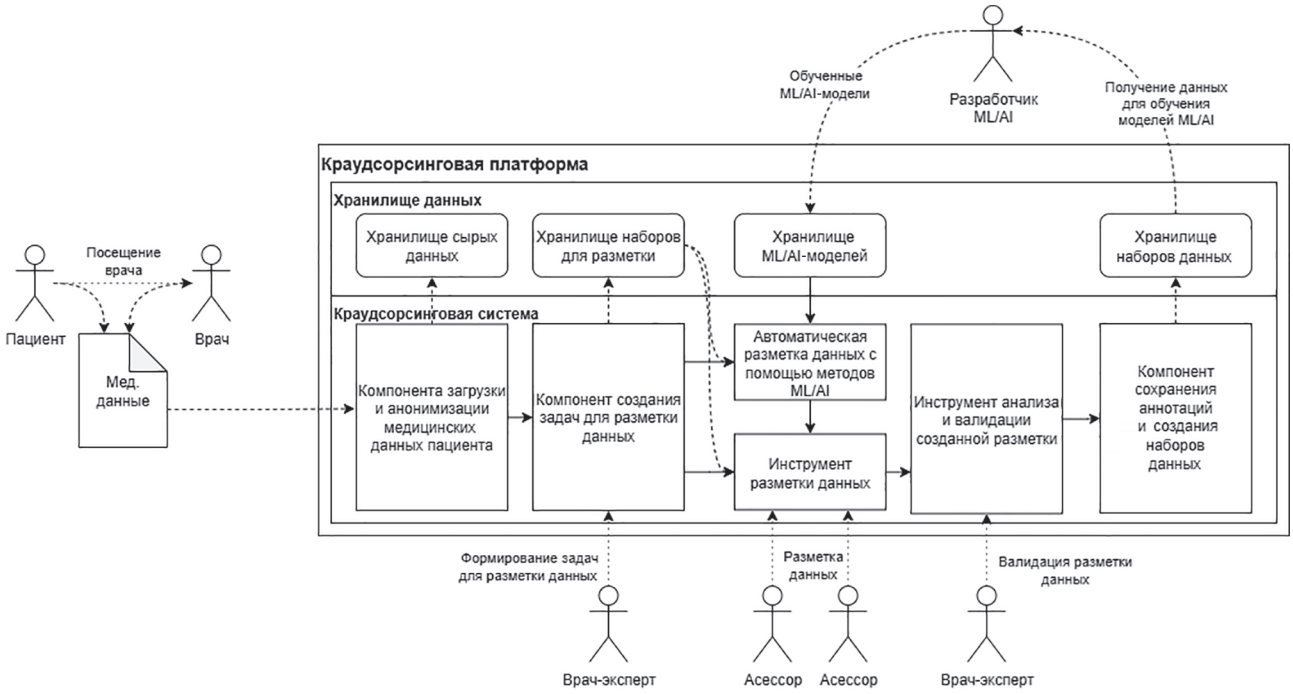

Рис. 2. Модель пользовательского взаимодействия с краудсорсинговой системой разметки медицинских данных в контексте разработки ML/AI-решения по алгоритму

Fig. 2. User interaction model with the crowdsourcing medical data annotation system within ML/AI solution development (Algorithm)

Этап 3. Асессоры размечают данные самостоятельно или корректируют разметку ML/AI-решения в назначенных им задачах.

Этап 4. Врач-эксперт проводит валидацию разметки асессоров и ее финальное уточнение.

Этап 5. Разработчик ML/AI получает подготовленные и размеченные наборы медицинских данных.

Этап 6. Разработчик ML/AI поставляет обученную ML/AI-модель для автоматизации разметки.

Результатом данного взаимодействия с системой является итеративное развитие и улучшение разрабатываемого ML/AI-решения. Для работы алгоритма необходимо обеспечить не только сбор данных для последующего увеличения качества и обобщающей способности ML/AI-моделей, но и получение экспертной обратной связи по качеству работы моделей при разметке данных, что позволит корректировать процесс обучения моделей для учета текущих ошибок и недостатков их работы.

Проектирование универсальной модели архитектуры краудсорсинговой системы разметки медицинских данных

Предлагаемая универсальная архитектура краудсорсинговой системы (рис. 3) интегрирует четыре взаимосвязанных компонента, обеспечивающих комплексное решение задач медицинской разметки. Основу системы составляют пользовательские сервисы на базе CVAT, предназначенные для взаимодействия с асессорами и проведения разметки. Архитектура дополнена специализированными функциональными модулями, выполняющими пред- и постобработку данных, включая операции анонимизации, автоматическую разметку с использованием ML-моделей и формирование итоговых наборов данных. Для хранения информации предусмотрено надежное хранилище, поддерживающее консистентность данных и сохранение связей между исходными медицинскими снимками и их PNG-представлениями. Система оркестрации гарантирует стабильность работы, горизонтальную масштабируемость и отказоустойчивость за счет централизованного управления всеми этапами обработки данных.

Сервис разметки CVAT (рис. 3) представляет собой гибкий компонент архитектуры, допускающий замену на другие системы (LabelStudio или собственные решения). Ключевые преимущества предложенной архитектуры включают: стандартизацию медицинских данных различных форматов, механизмы анонимизации и защиты информации, отказоустойчивость за счет системы оркестрации, а также поддержку распределенной обработки для повышения производительности. Реализация выполнена на Python 3.11 с использованием инструмента Celery для оркестрации процессов и системы MinIO в качестве объектного хранилища медицинских данных.

Апробация использования краудсорсинговой системы разметки медицинских данных

Экспериментальная апробация системы проводилась кардиохирургами Клиники высоких медицинских технологий им. Н.И. Пирогова Санкт-Петербургского государственного университета для разметки снимков компьютерной томографии (КТ) грудной аорты. Процесс включал:

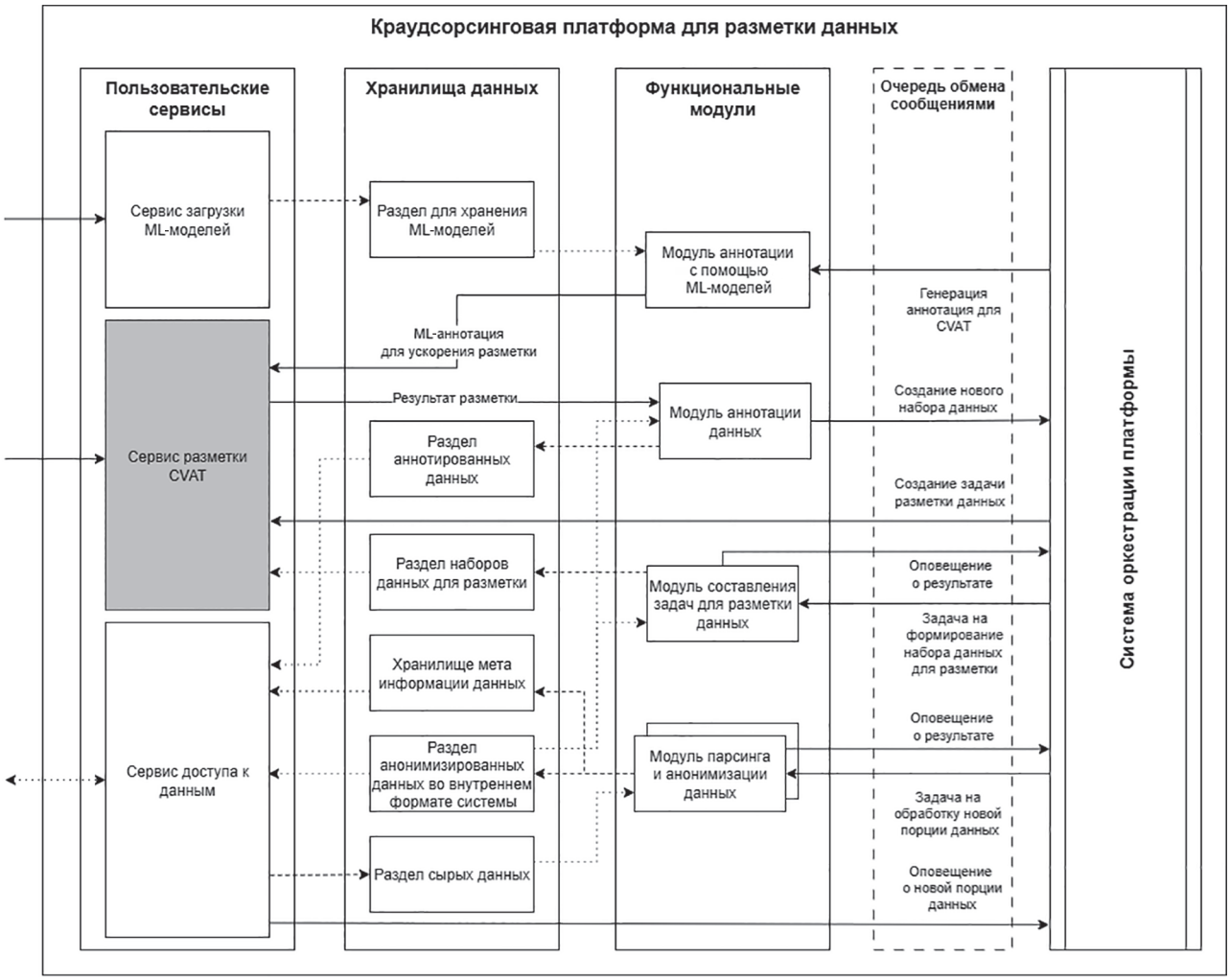

Рис. 3. Универсальная модель архитектуры краудсорсинговой системы разметки медицинских данных

Fig. 3. Universal architecture model of the crowdsourcing medical data annotation platform

— загрузку и настройку отображения DICOM-данных (рис. 4);

— анонимизацию через MD5-шифрование персональных данных;

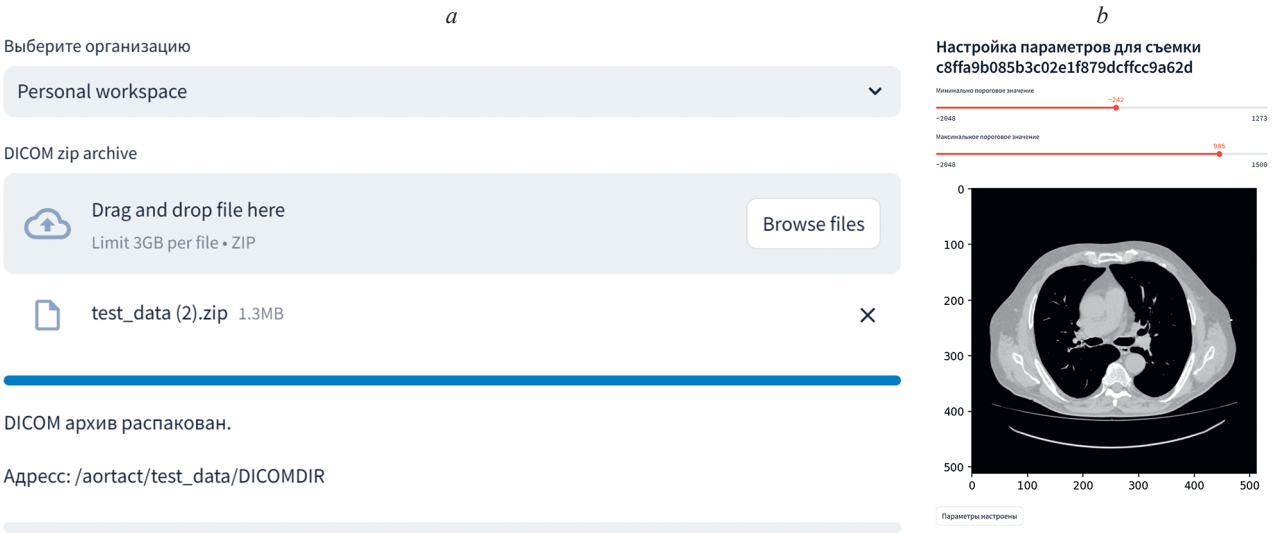

Рис. 4. Примеры интерфейсов сервиса доступа к данным: загрузки DICOM снимка (*a*); настройки параметров отображения КТ-изображений (*b*)

Fig. 4. Data access service interface example: DICOM image upload interface (*a*); CT image display settings interface (*b*)

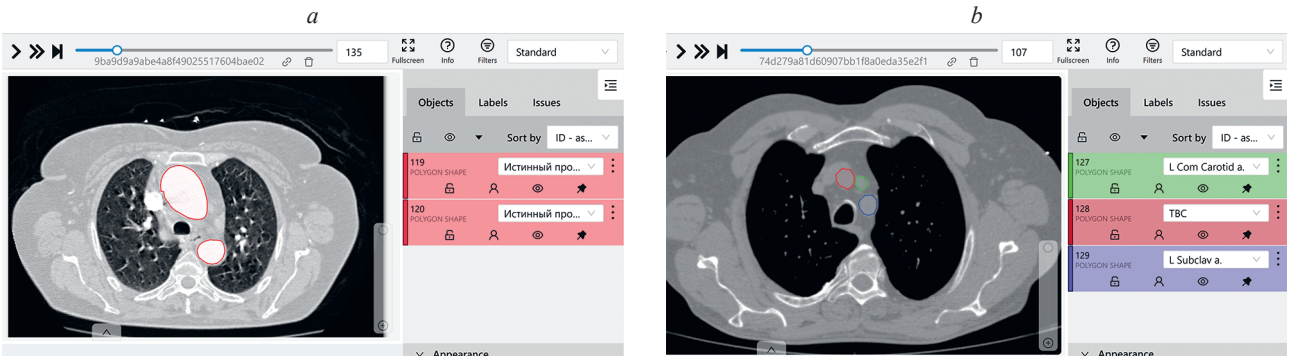

Рис. 5. Интерфейсы разметок КТ-снимков: грудной аорты (*a*); коронарных артерий (*b*)

Fig. 5. Interface for marking CT images: thoracic aorta (*a*), coronary arteries (*b*)

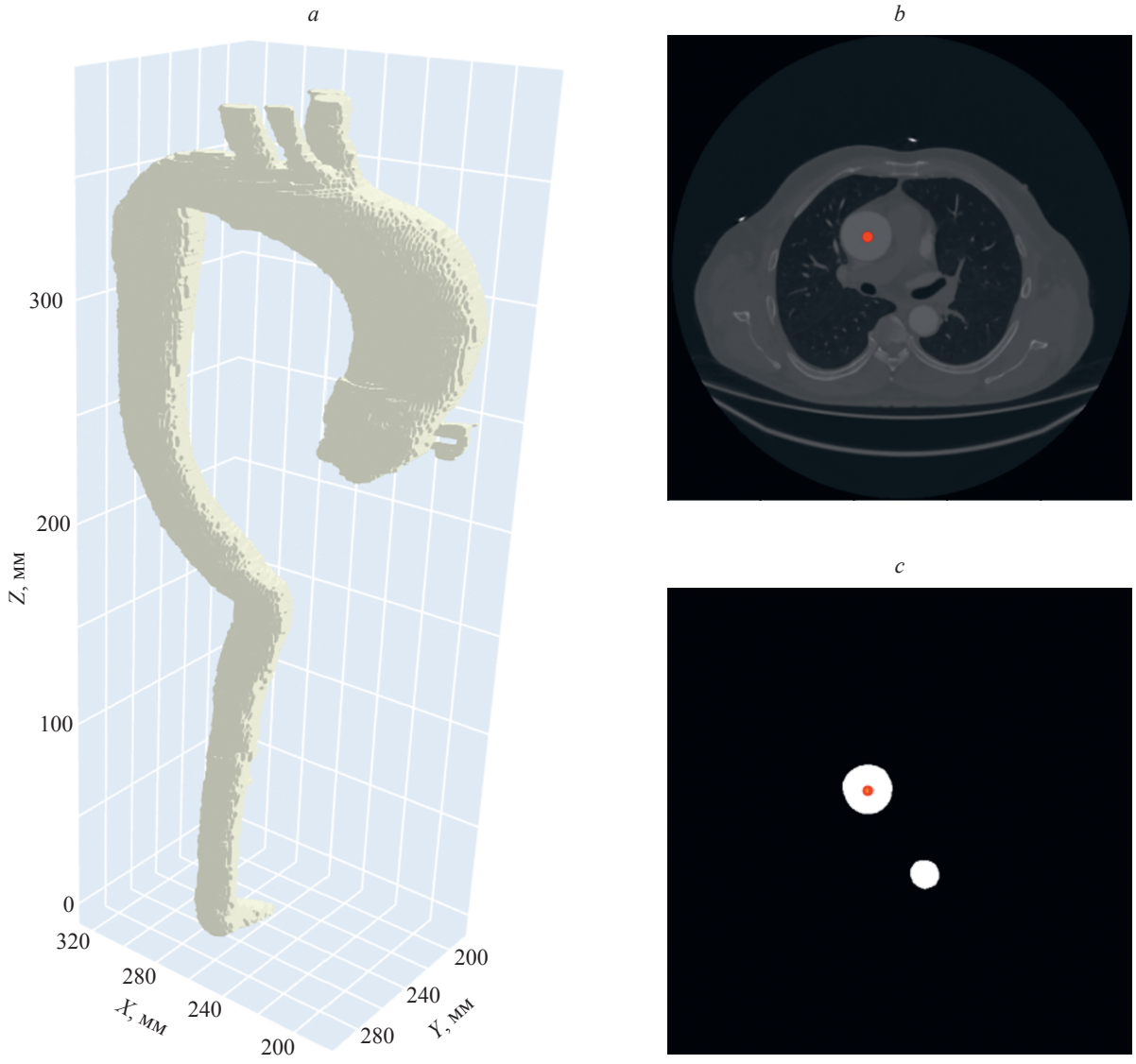

Рис. 6. Примеры размеченных данных аорты: 3D-модель аорты на основе полученной разметки (*a*); поперечный срез КТ-снимка аорты (*b*); разметка среза аорты в виде маски (*c*)

Fig. 6. Examples of annotated aorta data: 3D aorta model based on annotations (*a*); transverse CT slice of aorta (*b*); aorta slice annotation mask (*c*)

— автоматическое преобразование медицинских снимков в PNG-формат и создание задач разметки в CVAT.

В ходе апробации системы 30 асессоров и два врача-эксперта выполнили разметку 50 КТ-снимков аорты сердца (в среднем по 500 срезов на снимок) с исполь-

Таблица 5. Временные характеристики, затраченные на обработку одной аорты при создании набора данных из 50 аорт, с
Table 5. Time metrics for single aorta processing when creating a dataset of 50 aortas, s

Операция	Минимальное время	Среднее время	Максимальное время
Анонимизация входных данных одного КТ-снимка	не более 0,01	Менее 0,01	Менее 0,01
Формирование задачи разметки аорты в CVAT	32,6	71,3	117,4
Формирование связей между исходными данными КТ-снимка и изображениями для разметки в CVAT	не более 0,01	0,02	0,04
Проведение автоматической разметки загруженных снимков	16,7	17,4	21,7
Построение набора данных на основе полученной разметки	9,0	12,3	17,2
Разметка одной аорты асессорами вручную	7200	18 000	28 800

зованием 20 уникальных классов. Система автоматически сформировала набор данных для обучения ML-моделей, содержащий 25 655 размеченных срезов. Как показано на рис. 5, процесс разметки включал выделение областей аорты и сосудов. Результаты работы предложенной системы (рис. 6) демонстрируют возможность преобразования размеченных данных (рис. 6, *b*) в наборы масок (рис. 6, *c*) и 3D-модели (рис. 6, *a*) сердечно-сосудистых структур.

В процессе использования разработанной системы краудсорсинговой разметки медицинских данных были проведены замеры скорости автоматического построения набора данных на основе разметки пользователей в CVAT и исходных данных DICOM-изображений. Для обучения моделей требуется использовать не преобразованные к PNG-формату данные, а исходные данные медицинского снимка в более широком диапазоне, поэтому необходимо автоматически сопоставлять разметку с данными КТ-снимка, связь между которыми храниться в системе. В реализованном решении на базе универсальной модели архитектуры такие связи хранятся в формате JSON. В табл. 5 приведены замеры значений временных характеристик выполнения различных операций в процессе построения размеченного набора данных, состоящего из 50 аорт (25 655 срезов), для задач машинного и глубокого обучений.

Заключение

В работе спроектирована универсальная модель архитектуры краудсорсинговых систем для разметки медицинских данных (на примере семантической сегментации компьютерной томографии грудной аорты) и построения специализированных наборов данных для решения задач машинного и глубокого обучений в области медицины и кардиохирургии, в частности. Предложенная универсальная модель удовлетворяет критериям для оценки эффективности и безопасности существующих решений в этой области (табл. 2), а также позволяет решать основные актуальные проблемы организации процесса разметки медицинских данных в соответствии с обобщенной в настоящей работе классификацией проблем краудсорсинговой разметки данных (табл. 1). По сравнению с существующими решениями предложенная модель ориентирована на специфику медицинских данных и может быть адаптирована к любому формату данных в зависимости от решаемой задачи по интеллектуальному анализу таких данных. По результатам сравнительного анализа современные системы разметки показали низкую адаптивность к разметке медицинских данных, в частности, выявлены проблемы с безопасностью медицинских данных пациентов, организацией процесса и контроля качества разметки.

Предложена модель взаимодействия различных групп пользователей (как отдельных врачей, так и разработчиков машинного обучения) при разработке систем искусственного интеллекта в области медицины. Описан в виде алгоритма обобщенный сценарий взаимодействия различных групп пользователей с системой краудсорсинга на основе представленной универсальной модели архитектуры.

Разработанная краудсорсинговая система на основе предложенной модели архитектуры была успешно протестирована и апробирована в задаче сбора данных для семантической сегментации аорты сердца врачами-экспертами Клиники высоких медицинских технологий им. Н.И. Пирогова Санкт-Петербургского государственного университета. Временные характеристики работы новой функциональности, замеренные в ходе тестирования системы, незначительны в сравнении со временем, затраченным на саму разметку аорты.

Литература

1. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books, 2019. 341 p.
2. Obermeyer Z., Emanuel E.J. Predicting the future — big data, machine learning, and clinical medicine // New England Journal of Medicine. 2016. V. 375. N 13. P. 1216–1219. https://doi.org/10.1056/nejmp1606181
3. Jiang F., Jiang Y., Zhi H., Dong Y., Li H., Ma S., et al. Artificial intelligence in healthcare: past, present and future // Stroke and

References

1. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019, 341 p.
2. Obermeyer Z., Emanuel E.J. Predicting the future — big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 2016, vol. 375, no. 13, pp. 1216–1219. https://doi.org/10.1056/nejmp1606181
3. Jiang F., Jiang Y., Zhi H., Dong Y., Li H., Ma S., et al. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular*

Vascular Neurology. 2017. V. 2. N 4. P. 230–243. https://doi.org/10.1136/svn-2017-000101
4. Secinaro S., Calandra D., Secinaro A., Muthurangu V., Biancone P. The role of artificial intelligence in healthcare: a structured literature review // BMC Medical Informatics and Decision Making. 2021. V. 21. N 1. P. 125. https://doi.org/10.1186/s12911-021-01488-9
5. Roh Y., Heo G., Whang S.E. A survey on data collection for machine learning: a big data – AI Integration perspective // IEEE Transactions on Knowledge and Data Engineering. 2021. V. 33. N 4. P. 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162
6. Апанасович К.С., Махныткина О.В., Кабаров В.И., Далевская О.П. RuPersonaChat: корпус диалогов для персонификации разговорных агентов // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 2. С. 214–221. https://doi.org/10.17586/2226-1494-2024-24-2-214-221
7. Shaheen Z., Mouromtsev D.I., Postny I. RuLegalNER: a new dataset for Russian legal named entities recognition. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2023. V. 23. N 4. P. 854–857. https://doi.org/10.17586/2226-1494-2023-23-4-854-857
8. Sayin B., Krivosheev E., Yang J., Passerini A., Casati F. A review and experimental analysis of active learning over crowdsourced data // Artificial Intelligence Review. 2021. V. 54. N 7. P. 5283–5305. https://doi.org/10.1007/s10462-021-10021-3
9. Xintong G., Hongzhi W., Song Y., Hong G. Brief survey of crowdsourcing for data mining // Expert Systems With Application. 2014. V. 41. N 17. P. 7987–7994. https://doi.org/10.1016/j.eswa.2014.06.044
10. Hecht R., Kalla M., Krüger T. Crowd-sourced data collection to support automatic classification of building footprint data // Proc. of the ICA. 2018. V. 1. P. 54. https://doi.org/10.5194/ica-proc-1-54-2018
11. Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., et al. Human-level control through deep reinforcement learning // Nature. 2015. V. 518. N 7540. P. 529–533. https://doi.org/10.1038/nature14236
12. Rahmani A.M., Yousefpoor E., Yousefpoor M.S., Mehmood Z., Haider A., Hosseinzadeh M., Naqvi R.A. Machine learning (ML) in medicine: review, applications, and challenges // Mathematics. 2021. V. 9. N 22. P. 2970. https://doi.org/10.3390/math9222970
13. Wang C., Han L., Stein G., Day S., Bien-Gund C, Mathews A., et al. Crowdsourcing in health and medical research: a systematic review // Infectious Diseases of Poverty. 2020. V. 9. N 1. P. 8. https://doi.org/10.1186/s40249-020-0622-9
14. Ellis R.J., Sander R.M., Limon A. Twelve key challenges in medical machine learning and solutions // Intelligence-Based Medicine. 2022. V. 6. P. 100068. https://doi.org/10.1016/j.ibmed.2022.100068
15. Xia H., McKernan B. Privacy in crowdsourcing: a review of the threats and challenges // Computer Supported Cooperative Work (CSCW). 2020. V. 29. N 3. P. 263–301. https://doi.org/10.1007/s10606-020-09374-0
16. Rother A., Niemann U., Hielscher T., Völzke H., Ittermann T., Spiliopoulou M. Assessing the difficulty of annotating medical data in crowdworking with help of experiments // PLOS ONE. 2021. V. 16. N 7. P. e0254764. https://doi.org/10.1371/journal.pone.0254764
17. Ye C., Coco J., Epishova A., Hajaj C., Bogardus H., Novak L., et al. A crowdsourcing framework for medical data sets // AMIA Joint Summits on Translational Science proceedings. 2018. P. 273–280.
18. Kittur A., Nickerson J., Bernstein M., Gerber E., Shaw A., Zimmerman J., et al. The future of crowd work // Proc. of the Conference on Computer Supported Cooperative Work. 2013. P. 1301–1318. https://doi.org/10.1145/2441776.2441923
19. Ørting S.N., Doyle A., van Hilten A., Hirth M., Inel O., Madan C.R., et al. A survey of crowdsourcing in medical image analysis // Human Computation. 2020. V. 7. N 1. P. 1–26. https://doi.org/10.15346/hc.v7i1.1
20. Lu J., Li W., Wang Q., Zhang Y. Research on data quality control of crowdsourcing annotation: a survey // Proc. of the IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). 2020. P. 201–208. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00044
21. Lu X., Ratcliffe D., Kao T.-T., Tikhonov A., Litchfield L., Rodger C., Wang K. Rethinking quality assurance for crowdsourced multi-ROI image segmentation // Proc. of the 11th AAAI Conference on Human

Neurology, 2017, vol. 2, no. 4, pp. 230–243. https://doi.org/10.1136/svn-2017-000101
4. Secinaro S., Calandra D., Secinaro A., Muthurangu V., Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, 2021, vol. 21, no. 1, pp. 125. https://doi.org/10.1186/s12911-021-01488-9
5. Roh Y., Heo G., Whang S.E. A survey on data collection for machine learning: a big data — Al Integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 2021, vol. 33, no. 4, pp. 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162
6. Apanasovich K.S., Makhnytkina O.V., Kabarov V.I., Dalevskaya O.P. RuPersonaChat: a dialog corpus for personalizing conversational agents. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 2, pp. 214–221. (in Russian). https://doi.org/10.17586/2226-1494-2024-24-2-214-221
7. Shaheen Z., Mouromtsev D.I., Postny I. RuLegalNER: a new dataset for Russian legal named entities recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 854–857. https://doi.org/10.17586/2226-1494-2023-23-4-854-857
8. Sayin B., Krivosheev E., Yang J., Passerini A., Casati F. A review and experimental analysis of active learning over crowdsourced data. *Artificial Intelligence Review*, 2021, vol. 54, no. 7, pp. 5283–5305. https://doi.org/10.1007/s10462-021-10021-3
9. Xintong G., Hongzhi W., Song Y., Hong G. Brief survey of crowdsourcing for data mining. *Expert Systems With Application*, 2014, vol. 41, no. 17, pp. 7987–7994. https://doi.org/10.1016/j.eswa.2014.06.044
10. Hecht R., Kalla M., Krüger T. Crowd-sourced data collection to support automatic classification of building footprint data. *Proc. of the ICA*, 2018, vol. 1, pp. 54. https://doi.org/10.5194/ica-proc-1-54-2018
11. Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., et al. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, no. 7540, pp. 529–533. https://doi.org/10.1038/nature14236
12. Rahmani A.M., Yousefpoor E., Yousefpoor M.S., Mehmood Z., Haider A., Hosseinzadeh M., Naqvi R.A. Machine learning (ML) in medicine: review, applications, and challenges. *Mathematics*, 2021, vol. 9, no. 22, pp. 2970. https://doi.org/10.3390/math9222970
13. Wang C., Han L., Stein G., Day S., Bien-Gund C, Mathews A., et al. Crowdsourcing in health and medical research: a systematic review. *Infectious Diseases of Poverty*, 2020, vol. 9, no. 1, pp. 8. https://doi.org/10.1186/s40249-020-0622-9
14. Ellis R.J., Sander R.M., Limon A. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, 2022, vol. 6, pp. 100068. https://doi.org/10.1016/j.ibmed.2022.100068
15. Xia H., McKernan B. Privacy in crowdsourcing: a review of the threats and challenges. *Computer Supported Cooperative Work (CSCW)*, 2020, vol. 29, no. 3, pp. 263–301. https://doi.org/10.1007/s10606-020-09374-0
16. Rother A., Niemann U., Hielscher T., Völzke H., Ittermann T., Spiliopoulou M. Assessing the difficulty of annotating medical data in crowdworking with help of experiments. *PLOS ONE*, 2021, vol. 16, no. 7, pp. e0254764. https://doi.org/10.1371/journal.pone.0254764
17. Ye C., Coco J., Epishova A., Hajaj C., Bogardus H., Novak L., et al. A crowdsourcing framework for medical data sets. *AMIA Joint Summits on Translational Science proceedings*, 2018, pp. 273–280.
18. Kittur A., Nickerson J., Bernstein M., Gerber E., Shaw A., Zimmerman J., et al. The future of crowd work. *Proc. of the Conference on Computer Supported Cooperative Work*, 2013, pp. 1301–1318. https://doi.org/10.1145/2441776.2441923
19. Ørting S.N., Doyle A., van Hilten A., Hirth M., Inel O., Madan C.R., et al. A survey of crowdsourcing in medical image analysis. *Human Computation*, 2020, vol. 7, no. 1, pp. 1–26. https://doi.org/10.15346/hc.v7i1.1
20. Lu J., Li W., Wang Q., Zhang Y. Research on data quality control of crowdsourcing annotation: a survey. *Proc. of the IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2020, pp. 201–208. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00044
21. Lu X., Ratcliffe D., Kao T.-T., Tikhonov A., Litchfield L., Rodger C., Wang K. Rethinking quality assurance for crowdsourced multi-ROI

Computation and Crowdsourcing. 2023. V. 11. N 1. P. 103–114. https://doi.org/10.1609/hcomp.v11i1.27552
22. Тесленко Е.В. Искусственный интеллект в медицине. Правовые аспекты // Наука молодых — будущее России: сборник научных статей 8-й Международной научной конференции перспективных разработок молодых ученых. Курск: Университетская книга, 2023. С. 435–438.
23. Hulsen T. Sharing is caring — data sharing initiatives in healthcare // International Journal of Environmental Research and Public Health. 2020. V. 17. N 9. P. 3046. https://doi.org/10.3390/ijerph17093046
24. Sims M.H., Shaw M.H., Gilbertson S., Storch J., Halterman M.W. Legal and ethical issues surrounding the use of crowdsourcing among healthcare providers // Health Informatics Journal. 2019. V. 25. N 4. P. 1618–1630. https://doi.org/10.1177/1460458218796599
25. Mason W., Suri S. Conducting behavioral research on Amazon's Mechanical Turk // Behavior Research Methods. 2012. V. 44. N 1. P. 1–23. https://doi.org/10.3758/s13428-011-0124-6
26. Buecheler T., Sieg J.H., Füchslin R.M., Pfeifer R. Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework // Proc. of the 12th International Conference on the Synthesis and Simulation of Living Systems. 2010. P. 679–686.
27. Dortheimer J. Collective intelligence in design crowdsourcing // Mathematics. 2022. V. 10. N 4. P. 539. https://doi.org/10.3390/math10040539
28. Le K.H., Tran T.V., Pham H.H., Nguyen H.T., Le T.T., Nguyen H. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis // IEEE Access. 2023. V. 11. P. 14105–14114. https://doi.org/10.1109/ACCESS.2023.3243845
29. Petrović N., Moyà-Alcover G., Varona J., Jaume-i-Capó A. Crowdsourcing human-based computation for medical image analysis: a systematic literature review // Health Informatics Journal. 2020. V. 26. N 4. P. 2446–2469. https://doi.org/10.1177/1460458220907435
30. Vindas Y., Guépié B.K., Almar M., Roux E., Delachartre P. Semi-automatic data annotation based on feature-space projection and local quality metrics: An application to cerebral emboli characterization // Medical Image Analysis. 2022. V. 79. P. 102437. https://doi.org/10.1016/j.media.2022.102437
31. Philbrick K. A., Weston A.D., Akkus Z., Kline T.L., Korfiatis P., Sakinis T., et al. RIL-Contour: a medical imaging dataset annotation tool for and with deep learning // Journal of Digital Imaging. 2019. V. 32. N 4. P. 571–581. https://doi.org/10.1007/s10278-019-00232-0
32. Li H., Zhang B., Zhang Y., Liu W.W., Mao Y.J., Huang J.C., Wei L.F. A semi-automated annotation algorithm based on weakly supervised learning for medical images // Biocybernetics and Biomedical Engineering. 2020. V. 40. N 2. P. 787–802. https://doi.org/10.1016/j.bbe.2020.03.005
33. Larobina M., Murino L. Medical image file formats // Journal of Digital Imaging. 2014. V. 27. N 2. P. 200–206. https://doi.org/10.1007/s10278-013-9657-9
34. Willemink M.J., Koszek W.A., Hardell C., Wu J., Fleischmann D., Harvey H., et al. Preparing medical imaging data for machine learning // Radiology. 2020. V. 295. N 1. P. 4–15. https://doi.org/10.1148/radiol.2020192224
35. Pfob A., Lu S.-C., Sidey-Gibbons C. Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison // BMC Medical Research Methodology. 2022. V. 22. N 1. P. 282. https://doi.org/10.1186/s12874-022-01758-8
36. Кондратенко С.С., Коржук В.М. Архитектура системы обработки медицинских данных с учетом требований обеспечения целостности. Сборник тезисов докладов конгресса молодых ученых. 2023. [Электронный ресурс]. URL: https://kmu.itmo.ru/digests/article/11444
37. Васильев Ю.А., Савкина Е.Ф., Владзимирский А.В., Омелянская О.В., Арзамасов К.М. Обзор современных средств разметки цифровых диагностических изображений // Казанский медицинский журнал. 2023. Т. 104. № 5. С. 750–760. https://doi.org/10.17816/KMJ349060
38. Ежов Ф.В., Коваленко Л.А., Разумилов Е.С., Блеканов И.С. Инструменты краудсорсинга для анализа и обработки медицинских изображений в виде снимков КТ // Процессы управления и устойчивость. 2023. Т. 10. № 1. С. 291–297.
39. Saltz J.S., Krasteva I. Current approaches for executing big data science projects — a systematic literature review // PeerJ Computer Science. 2022. V. 8. P. e862. https://doi.org/10.7717/peerj-cs.862

image segmentation. *Proc. of the 11th AAAI Conference on Human Computation and Crowdsourcing*, 2023, vol. 11, no. 1, pp. 103–114. https://doi.org/10.1609/hcomp.v11i1.27552
22. Teslenko E.V. Artificial intelligence in medicine. Legal aspects. *Proc. of the Science of the young — the future of Russia*. 2023. pp. 435–438. (in Russian)
23. Hulsen T. Sharing is caring — data sharing initiatives in healthcare. *International Journal of Environmental Research and Public Health*, 2020, vol. 17, no. 9, pp. 3046. https://doi.org/10.3390/ijerph17093046
24. Sims M.H., Shaw M.H., Gilbertson S., Storch J., Halterman M.W. Legal and ethical issues surrounding the use of crowdsourcing among healthcare providers. *Health Informatics Journal*, 2019, vol. 25, no. 4, pp. 1618–1630. https://doi.org/10.1177/1460458218796599
25. Mason W., Suri S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 2012, vol. 44, no. 1, pp. 1–23. https://doi.org/10.3758/s13428-011-0124-6
26. Buecheler T., Sieg J.H., Füchslin R.M., Pfeifer R. Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. *Proc. of the 12th International Conference on the Synthesis and Simulation of Living Systems*, 2010, pp. 679–686.
27. Dortheimer J. Collective intelligence in design crowdsourcing. *Mathematics*, 2022, vol. 10, no. 4, pp. 539. https://doi.org/10.3390/math10040539
28. Le K.H., Tran T.V., Pham H.H., Nguyen H.T., Le T.T., Nguyen H. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *IEEE Access*, 2023, vol. 11, pp. 14105–14114. https://doi.org/10.1109/ACCESS.2023.3243845
29. Petrović N., Moyà-Alcover G., Varona J., Jaume-i-Capó A. Crowdsourcing human-based computation for medical image analysis: a systematic literature review. *Health Informatics Journal*, 2020, vol. 26, no. 4, pp. 2446–2469. https://doi.org/10.1177/1460458220907435
30. Vindas Y., Guépié B.K., Almar M., Roux E., Delachartre P. Semi-automatic data annotation based on feature-space projection and local quality metrics: An application to cerebral emboli characterization. *Medical Image Analysis*, 2022, vol. 79, pp. 102437. https://doi.org/10.1016/j.media.2022.102437
31. Philbrick K. A., Weston A.D., Akkus Z., Kline T.L., Korfiatis P., Sakinis T., et al. RIL-Contour: a medical imaging dataset annotation tool for and with deep learning. *Journal of Digital Imaging*, 2019, vol. 32, no. 4, pp. 571–581. https://doi.org/10.1007/s10278-019-00232-0
32. Li H., Zhang B., Zhang Y., Liu W.W., Mao Y.J., Huang J.C., Wei L.F. A semi-automated annotation algorithm based on weakly supervised learning for medical images. *Biocybernetics and Biomedical Engineering*, 2020, vol. 40, no. 2, pp. 787–802. https://doi.org/10.1016/j.bbe.2020.03.005
33. Larobina M., Murino L. Medical image file formats. *Journal of Digital Imaging*, 2014, vol. 27, no. 2, pp. 200–206. https://doi.org/10.1007/s10278-013-9657-9
34. Willemink M.J., Koszek W.A., Hardell C., Wu J., Fleischmann D., Harvey H., et al. Preparing medical imaging data for machine learning. *Radiology*, 2020, vol. 295, no. 1, pp. 4–15. https://doi.org/10.1148/radiol.2020192224
35. Pfob A., Lu S.-C., Sidey-Gibbons C. Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison. *BMC Medical Research Methodology*, 2022, vol. 22, no. 1, pp. 282. https://doi.org/10.1186/s12874-022-01758-8
36. Kondratenko S.S., Korzhuk V.M. Architecture of a medical data processing system taking into account integrity requirements. *Collection of abstracts from the Congress of Young Scientists*. 2023. Available at: https://kmu.itmo.ru/digests/article/11444 (in Russian)
37. Vasilev Y.A., Savkina E.F., Vladzimirskii A.V., Omelianskaia O.V., Arzamasov K.M. Overview of modern digital diagnostic image markup tools. *Kazan Medical Journal*, 2023, vol. 104, no. 5, pp. 750–760. (in Russian). https://doi.org/10.17816/KMJ349060
38. Ezhov F.V., Kovalenko L.A., Razumilov E.S., Blekanov I.S. Crowdsourcing tools for the analysis and processing of medical CT images. *Processy Upravlenija i Ustojchivost*', 2023, vol. 10, no. 1, pp. 291–297. (in Russian)
39. Saltz J.S., Krasteva I. Current approaches for executing big data science projects — a systematic literature review. *PeerJ Computer Science*, 2022, vol. 8, pp. e862. https://doi.org/10.7717/peerj-cs.862

40. Saltz J.S. CRISP-DM for data science: strengths, weaknesses and potential next steps // Proc. of the IEEE International Conference on Big Data. 2021. P. 2337–2344. https://doi.org/10.1109/bigdata52589.2021.9671634
41. Saltz J., Hotz N. Factors that influence the selection of a data science process management methodology: an exploratory study // Proc. of the 54th Hawaii International Conference on System Sciences. 2021. P. 949–958. https://doi.org/10.24251/hicss.2021.116
42. Zhao X., Zhang P., Song F., Fan G.D., Sun Y.Y., Wang Y.J., et al. D2A U-Net: Automatic segmentation of COVID-19 CT slices based on dual attention and hybrid dilated convolution // Computers in Biology and Medicine. 2021. V. 135. P. 104526. https://doi.org/10.1016/j.compbiomed.2021.104526
43. Xie Y., Padgett J., Biancardi A.M., Reeves A.P. Automated aorta segmentation in low-dose chest CT images // International Journal of Computer Assisted Radiology and Surgery. 2014. V. 9. N 2. P. 211–219. https://doi.org/10.1007/s11548-013-0924-5
44. Ким Г.И., Блеканов И.С., Ежов Ф.В., Коваленко Л.А., Ларин Е.С., Разумилов Е.С. [и др.] Методы искусственного интеллекта в сердечно-сосудистой хирургии и диагностика патологии аорты и аортального клапана (обзор литературы) // Сибирский журнал клинической и экспериментальной медицины. 2024. Т. 39. № 2. С. 36–45. https://doi.org/10.29001/2073-8552-2024-39-2-36-45
45. Gao R., Zhao S., Aishanjiang K., Cai H., Wei T., Zhang Y.C., et al. Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced CT and clinical data // Journal of Hematology & Oncology. 2021. V. 14. N 1. P. 154. https://doi.org/10.1186/s13045-021-01167-2
46. Chen P.-T., Wu T.H., Wang P.C., Chang D.W., Liu K.L., Wu M.S., et al. Pancreatic cancer detection on CT scans with deep learning: a nationwide population-based study // Radiology. 2023. V. 306. N 1. P. 172–182. https://doi.org/10.1148/radiol.220152
47. Zhou H., Li L., Liu Z., Zhao K., Chen X., Lu M., et al. Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images // European Radiology. 2021. V. 31. N 6. P. 3931–3940. https://doi.org/10.1007/s00330-020-07454-9

40. Saltz J.S. CRISP-DM for data science: strengths, weaknesses and potential next steps. *Proc. of the IEEE International Conference on Big Data*, 2021, pp. 2337–2344. https://doi.org/10.1109/bigdata52589.2021.9671634
41. Saltz J., Hotz N. Factors that influence the selection of a data science process management methodology: an exploratory study. *Proc. of the 54th Hawaii International Conference on System Sciences*, 2021, pp. 949–958. https://doi.org/10.24251/hicss.2021.116
42. Zhao X., Zhang P., Song F., Fan G.D., Sun Y.Y., Wang Y.J., et al. D2A U-Net: Automatic segmentation of COVID-19 CT slices based on dual attention and hybrid dilated convolution. *Computers in Biology and Medicine*, 2021, vol. 135, pp. 104526. https://doi.org/10.1016/j.compbiomed.2021.104526
43. Xie Y., Padgett J., Biancardi A.M., Reeves A.P. Automated aorta segmentation in low-dose chest CT images. *International Journal of Computer Assisted Radiology and Surgery*, 2014, vol. 9, no. 2, pp. 211–219. https://doi.org/10.1007/s11548-013-0924-5
44. Kim G.I., Blekanov I.S., Ezhov F.V., Kovalenko L.A., Larin E.S., Razumilov E.S., et al. Artificial intelligence methods in cardiovascular surgery and diagnosis of pathology of the aorta and aortic valve (literature review). *Siberian Journal of Clinical and Experimental Medicine*, 2024, vol. 39, no. 2, pp. 36–45. (in Russian). https://doi.org/10.29001/2073-8552-2024-39-2-36-45
45. Gao R., Zhao S., Aishanjiang K., Cai H., Wei T., Zhang Y.C., et al. Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced CT and clinical data. *Journal of Hematology & Oncology*, 2021, vol. 14, no. 1, pp. 154. https://doi.org/10.1186/s13045-021-01167-2
46. Chen P.-T., Wu T.H., Wang P.C., Chang D.W., Liu K.L., Wu M.S., et al. Pancreatic cancer detection on CT scans with deep learning: a nationwide population-based study. *Radiology*, 2023, vol. 306, no. 1. pp. 172–182. https://doi.org/10.1148/radiol.220152
47. Zhou H., Li L., Liu Z., Zhao K., Chen X., Lu M., et al. Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images. *European Radiology*, 2021, vol. 31, no. 6, pp. 3931–3940. https://doi.org/10.1007/s00330-020-07454-9

Авторы

Коваленко Лев Алексеевич — ведущий программист, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация; ведущий программист, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 59225183700, https://orcid.org/0009-0007-8233-4387, levozavr1@gmail.com

Блеканов Иван Станиславович — кандидат технических наук, доцент, заведующий кафедрой, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 56149559700, https://orcid.org/0000-0002-7305-1429, i.blekanov@spbu.ru

Ежов Федор Валерьевич — инженер-программист, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 59224591300, https://orcid.org/0009-0007-1468-0042, st056053@student.spbu.ru

Ларин Евгений Сергеевич — ведущий аналитик, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, https://orcid.org/0009-0007-6199-3607, evgeny.larin@spbu.ru

Ким Глеб Ирламович — кандидат медицинских наук, врач сердечно-сосудистый хирург, Санкт-Петербургский государственный университет. Клиника высоких медицинских технологий им. Н.И. Пирогова, Санкт-Петербург, 190020, Российская Федерация; доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 57704764600, https://orcid.org/0000-0002-9344-5724, gikim.cor@gmail.com

Authors

Lev A. Kovalenko — Leading Software Developer, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation; Leading Software Developer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 59225183700, https://orcid.org/0009-0007-8233-4387, levozavr1@gmail.com

Ivan. S. Blekanov — PhD, Associate Professor, Head of Department, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation, SC 56149559700, https://orcid.org/0000-0002-7305-1429, i.blekanov@spbu.ru

Fedor V. Ezhov — Software Developer - Engineer, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation, SC 59224591300, https://orcid.org/0009-0007-1468-0042, st056053@student.spbu.ru

Evgenii S. Larin — Leading Analyst, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation, https://orcid.org/0009-0007-6199-3607, evgeny.larin@spbu.ru

Gleb I. Kim — PhD (Medicine), Cardiovascular Surgeon, St. Petersburg State University (SPbSU), Saint Petersburg, 199034, Russian Federation; Associate Professor, Saint Petersburg State University Hospital, Saint Petersburg, 190020, Russain Federation, SC 57704764600, https://orcid.org/0000-0002-9344-5724, gikim.cor@gmail.com

Статья поступила в редакцию 15.01.2025
Одобрена после рецензирования 30.07.2025
Принята к печати 30.09.2025

Received 15.01.2025
Approved after reviewing 30.07.2025
Accepted 30.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-856-865
УДК 004.421.2

Методы роевой оптимизации частиц и локальных эвристик для решения мультиагентной задачи коммивояжёра

Эльдар Наилевич Мифтахов[1], Андрей Анатольевич Акимов[2], Юлия Ахнафовна Гнатенко[3]✉

[1,2] МИРЭА — Российский технологический университет, Москва, 119454, Российская Федерация
[3] Стерлитамакский филиал Уфимского университета науки и технологий, Стерлитамак, 453103, Российская Федерация

[1] promif@mail.ru, https://orcid.org/0000-0002-0471-5949
[2] andakm@yandex.ru, https://orcid.org/0000-0003-3387-2959
[3] y.a.gnatenko@struust.ru✉, https://orcid.org/0009-0009-9264-3989

Аннотация

Введение. Представлены результаты разработки и апробации метода решения мультиагентной задачи коммивояжёра (Multiple Traveling Salesman Problem, mTSP) с целью минимизации максимальной длины маршрутов («минимаксная» оптимизация). Объектом исследования является комбинаторное пространство маршрутов, возникающее при распределении городов между несколькими агентами, что обуславливает необходимость равномерного распределения нагрузки и предотвращения перегрузки отдельных маршрутов. Новизна подхода заключается в создании дискретного аналога классического алгоритма роевой оптимизации частиц (Particle Swarm Optimization, PSO), адаптированного для работы с перестановками, а также в интеграции его с локальными эвристическими процедурами и муравьиным алгоритмом (Ant Colony Optimization, ACO). **Метод.** Предложенный метод базируется на преобразовании исходной задачи mTSP в классическую задачу коммивояжёра для одного агента (TSP) посредством введения фиктивных депо, что позволяет однозначно разделить общий маршрут на отдельные части для каждого агента. Ключевым элементом является модификация PSO с использованием новых операций для дискретного пространства, таких как вычисление минимальной последовательности обменов (транспозиций) между перестановками, масштабирование скорости и применение ее к маршруту. Данный подход позволяет эффективно исследовать комбинаторное пространство решений и предотвращать преждевременную сходимость алгоритма. **Основные результаты.** Экспериментальное исследование проведено на тестовых наборах стандартной библиотеки TSPLIB (eil51.tsp, berlin52.tsp, eil76.tsp, rat99.tsp) для задачи TSP, в ходе которого сравнивались два сценария: классический PSO со случайной инициализацией и гибридный метод PSO_ACO, где метод ACO используется для формирования начальной популяции. Результаты эксперимента продемонстрировали существенное улучшение по «минимаксному» критерию по сравнению с методами CPLEX, LKH3, OR-Tools, а также современными подходами DAN, NCE и EA, что подтверждает эффективность предложенного решения. **Обсуждение.** Разработанный алгоритм может найти применение в логистике, транспортном планировании, распределении потоков в сетях связи и иных областях, где требуется оптимальное распределение ресурсов. Представленный метод будет полезен специалистам в области оптимизации, алгоритмического моделирования и практикам, занимающимся разработкой систем управления и планирования.

Ключевые слова

роевый алгоритм оптимизации частиц, мультиагентная задача коммивояжёра, минимаксная оптимизация, дискретная оптимизация, муравьиный алгоритм, локальные эвристики

Ссылка для цитирования: Мифтахов Э.Н., Акимов А.А., Гнатенко Ю.А. Методы роевой оптимизации частиц и локальных эвристик для решения мультиагентной задачи коммивояжёра // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 856–865. doi: 10.17586/2226-1494-2025-25-5-856-865

Particle swarm optimization methods and local heuristics for solving the multiple traveling salesman problem

Eldar N. Miftakhov[1], Andrey A. Akimov[2], Yuliya A. Gnatenko[3]✉

[1,2] MIREA — Russian Technological University, Moscow, 119454, Russian Federation
[3] Branch of the Ufa University of Science and Technology, Sterlitamak, 453103, Russian Federation

[1] promif@mail.ru, https://orcid.org/0000-0002-0471-5949
[2] andakm@yandex.ru, https://orcid.org/0000-0003-3387-2959
[3] y.a.gnatenko@struust.ru✉, https://orcid.org/0009-0009-9264-3989

Abstract

This paper presents the development and evaluation of a method for solving the Multiple Traveling Salesman Problem (mTSP), with the objective of minimizing the maximum route length ("minimax" optimization). The study addresses the combinatorial route-space arising from distributing cities among multiple agents, requiring balanced workload distribution to avoid overloading individual routes. The novelty of the proposed approach lies in creating a discrete analogue of the classical Particle Swarm Optimization (PSO) algorithm adapted specifically for permutation-based route representations, and integrating it with local heuristic procedures and the Ant Colony Optimization (ACO) algorithm. The proposed method transforms the original mTSP into a classical single-agent Traveling Salesman Problem (TSP) by introducing artificial (dummy) depots, thus allowing an unambiguous separation of the overall route into individual segments for each agent. A key element of the solution involves adapting the PSO algorithm through novel discrete operations, such as computing the minimal sequence of exchanges (transpositions) between permutations, scaling velocity, and applying this velocity to routes. This approach ensures efficient exploration of the combinatorial solution space and prevents premature convergence of the algorithm. The experimental study was conducted on benchmark instances from the TSPLIB library (eil51.tsp, berlin52.tsp, eil76.tsp, rat99.tsp) for the TSP, comparing two scenarios: a classical PSO with random initialization and a hybrid PSO_ACO method where the ACO algorithm is used to generate the initial population. The results demonstrate a significant improvement in the minimax criterion compared to CPLEX, LKH3, OR-Tools as well as state-of-the-art approaches DAN, NCE, and EA, confirming the effectiveness of the proposed solution. The practical importance of this research lies in potential applications of the developed algorithm in logistics, transport planning, network traffic management, and other domains where optimal resource allocation is crucial. The proposed method is valuable for specialists in optimization, algorithmic modeling, and practitioners developing planning and management systems.

Keywords

particle swarm optimization, multiple traveling salesman problem, minimax optimization, discrete optimization, ant colony optimization, local heuristics

For citation: Miftakhov E.N., Akimov A.A., Gnatenko Yu.A. Particle swarm optimization methods and local heuristics for solving the multiple traveling salesman problem. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 856–865 (in Russian). doi: 10.17586/2226-1494-2025-25-5-856-865

Введение

Мультиагентная задача коммивояжёра (Multiple Traveling Salesman Problem, mTSP) широко применяется в логистике, транспортном планировании, распределении потоков связи и производственных процессах [1]. В классической постановке требуется, чтобы m коммивояжёров (агентов), начинающих и заканчивающих маршрут из фиксированного депо, посетили n городов ровно по одному разу с оптимизацией общей стоимости маршрутов (например, суммарного или максимального расстояния — критерий «минимакс») [1, 2]. «Минимаксная» постановка обеспечивает равномерное распределение нагрузки и предотвращает перегрузку отдельных маршрутов [3, 4].

Метод оптимизации роя частиц (Particle Swarm Optimization, PSO), изначально разработанный для непрерывных задач [5], получил множество дискретных модификаций, применимых к задачам маршрутизации [6]. Основная идея PSO состоит в том, что каждая частица (агент) перемещается в пространстве решений, опираясь на свое личное лучшее решение и коллективный опыт роя — лучшее решение, обнаруженное в глобальном или локальном окружении. Для адаптации PSO к mTSP необходимо переопределить понятия «позиция» и «скорость» в терминах перестановок, где маршрут представляется как упорядоченный список городов, а «псевдоскорость» — как набор операций, изменяющих порядок обхода или распределение городов между агентами. Дополнительная интеграция алгоритма с локальными процедурами поиска (например, методом 2-opt) повышает точность решения и предотвращает преждевременную сходимость [7].

Таким образом, основная проблема стандартного PSO при решении mTSP заключается в переходе от непрерывного к дискретному представлению решений, что требует разработки корректного дискретного аналога PSO для эффективного исследования комбинаторного пространства маршрутов в условиях «минимаксной» оптимизации.

Обзор тематических научных работ

Классическая задача коммивояжёра является NP-полной, что обуславливает применение эвристических и метаэвристических методов для поиска приближенных решений [2]. Среди них широко используются генетические алгоритмы [8, 9] и алгоритмы оптимизации колонии муравьев (Ant Colony Optimization, ACO) [5], демонстрирующие высокую эффективность при решении комбинаторных задач.

Отдельную группу методов решения задачи mTSP представляют нейросетевые способы, в том числе метод Decentralized Attention Network (DAN) [10], который

сочетает графовую сеть внимания и локальный поиск, позволяя агентам кооперативно распределять города и тем самым минимизировать максимальную длину тура. Близкий по идее метод Neuro Cross-Exchange (NCE) использует графовую нейросеть для прогнозирования выгоды операций CROSS-exchange, сокращая сложность поиска с $O(n^4)$ до $O(n^2)$ без ущерба качеству решений [10].

Метод PSO вызывает значительный интерес, однако его применение осложнено дискретным характером mTSP. Поскольку стандартный PSO разработан для непрерывных пространств решений, его нельзя напрямую применять к mTSP. Для преодоления этой проблемы предложены различные методы дискретизации PSO. Например, в работе [11] разработано правило наименьшего значения позиции (Special Purpose Vehicle), позволяющее преобразовывать непрерывное представление в дискретное, а исследования в [3] показали, что включение генетических операторов (например, кроссовера и мутации) способствует улучшению поиска оптимальных решений в дискретном пространстве. Работа [12] направлена на переопределение базовых операций PSO (обновления скорости и положения частиц) для работы с перестановками, что снижает избыточность обновлений и предотвращает преждевременную сходимость.

Современные исследования все чаще фокусируются на гибридных алгоритмах, объединяющих PSO, ACO и другие эвристические подходы для решения mTSP с учетом критериев балансировки нагрузки (задача «минимакс») [4, 13]. Эффективная дискретизация PSO, позволяющая сохранять разнообразие решений и детально исследовать комбинаторное пространство маршрутов, является ключевым условием реализации таких гибридных методов.

Целью настоящей работы является разработка модифицированного алгоритма роевой оптимизации частиц, адаптированного для эффективного решения мультиагентной задачи коммивояжёра, позволяющего улучшить качество маршрутов по «минимаксному» критерию и предотвратить преждевременную сходимость за счет интеграции с локальными эвристиками и алгоритмом ACO.

Методология

Одним из методов решения мультиагентной задачи коммивояжёра mTSP является преобразование ее в классическую задачу коммивояжёра TSP с одним агентом, для которой можно применить современные эвристические алгоритмы [14]. Все города вместе с депо представляются в виде графа, где вершина 1 (депо) служит общей стартовой и конечной точкой для всех агентов. Исходную задачу с m агентами и n городами сведем к TSP с $n + m - 1$ городами путем добавления $m - 1$ фиктивных депо, номера которых задаются по схеме:

$$1, k_2^{(1)}, k_3^{(1)}, \ldots, n+1, k_i^{(2)}, k_{i+1}^{(2)}, \ldots, n+2, \ldots, n+m-1, k_j^{(m)}, k_{j+1}^{(m)}, \ldots, k_{n-1}^{(m)}, k_n^{(m)}. \quad (1)$$

Задача — минимизировать самый длинный путь из маршрутов всех коммивояжёров (задача «минимакс»).

Для решения полученной одноагентной задачи коммивояжёра в работе [8] предложен алгоритм PSO, основанный на моделировании поведения стаи птиц. Принцип алгоритма PSO состоит в том, что каждая «птица» в стае рассматривается как частица, обладающая механизмом «памяти», который помогает находить оптимальное решение путем взаимодействия с другими частицами в рое.

В стандартном PSO каждая частица рассматривается как точка без массы и объема, заданная в N-мерном пространстве. Позиции частиц являются потенциальными решениями задачи. Положение j-ой частицы представляется вектором $\mathbf{X}_j = (X_{j1}, X_{j2}, \ldots, X_{jn})$, а скорость полета частицы вектором $\mathbf{v}_j = (v_{j1}, v_{j2}, \ldots, v_{jn})$.

Алгоритм PSO является итерационным методом оптимизации. На каждой итерации частица j последовательно обновляет свою скорость и положение:

$$\mathbf{v}_j^{i+1} = \omega \mathbf{v}_j^i + c_1 r_1 \cdot (\mathbf{P}_j - \mathbf{X}_j^i) + c_2 r_2 \cdot (\mathbf{G} - \mathbf{X}_j^i), \quad (2)$$

$$\mathbf{X}_j^{i+1} = \mathbf{X}_j^i + \mathbf{v}_j^{i+1}, \quad (3)$$

где $\mathbf{v}_j^i$ и $\mathbf{X}_j^i$ — скорость и положение частицы j на i-ой итерации; $\mathbf{P}_j$ — лучшая позиция j-й частицы (личный опыт); $\mathbf{G}$ — лучшая позиция всего роя частиц (глобальный опыт); ω — коэффициент инерции скорости; c_1, c_2 — коэффициенты, определяющие влияние личного и коллективного опыта; r_1, r_2 — случайные числа из промежутка [0, 1].

Благодаря учету информации о ранее достигнутых «лучших» позициях (как индивидуальных, так и общих для всего роя), каждая частица в процессе итераций перемещается в те области пространства, где с наибольшей вероятностью находится (глобальное или локальное) оптимальное решение поставленной задачи.

Применение метода PSO для решения задачи TSP

При решении задачи TSP с помощью алгоритма PSO положение частицы в многомерном пространстве определяет возможный маршрут — последовательность посещения городов (перестановок) вида (1). Скорость должна изменять положение частицы и представляется в виде последовательности обменов (транспозиций) городов.

Формулы (2), (3) основаны на стандартных операциях над векторами. В отличие от непрерывных векторных пространств, пространство перестановок не является линейным, и прямое сложение или вычитание таких последовательностей не дает корректного результата. Исходя из этого, вводятся новые операции.

1. $\alpha \cdot v$, $0 \le \alpha \le 1$ — скалярное умножение скорости (масштабирование вектора скорости), $\alpha \cdot v$ определяется как выбор первых $k = \lceil \alpha \cdot |v| \rceil$ преобразований (обменов) из последовательности v. Здесь $|v|$ — длина (число обменов) последовательности v. Округление вверх до ближайшего целого числа $\lceil \cdot \rceil$ гарантирует при любом $\alpha > 0$ выбор по крайней мере одного элемента из последовательности v, что обеспечивает ненулевой результат.

Если $\alpha > 1$, то $k = \lceil \alpha \cdot |v| \rceil > |v|$. В этом случае требуется взять элементов больше, чем доступно в v. В дис-

кретных преобразованиях это не имеет смысла, поэтому случай $\alpha > 1$ не рассматривается.

Таким образом, $\alpha \cdot v$ — усеченная версия последовательности v, от которой берутся первые k преобразований (обменов) пропорционально коэффициенту α.

2. Операция $X \oplus v$ — приложение скорости к позиции. Пусть X — позиция (перестановка городов), а $v = [s_1, s_2, \ldots, s_m]$ — скорость, т. е. последовательность обменов (транспозиций). Тогда $X \oplus v = s_m \circ s_{m-1} \circ \ldots \circ s_1(X)$. Иными словами, оператор $\oplus$ поочередно применяет все обмены из v к перестановке X, начиная с s_1 и заканчивая s_m.

3. Операция $Y \Theta X$ — минимальная разность позиций $Y \Theta X = v_{\min}$, где $v_{\min}$ – минимальная по длине последовательность обменов, преобразующая X в Y. Другими словами, $Y \Theta X$ — скорость (цепочка преобразований), применяя которую к X, получим Y.

В формуле (2) на текущей итерации учитывается влияние как личного опыта частицы $\mathbf{P}_j$, так и коллективного опыта роя $\mathbf{G}$. Однако в дискретном пространстве, где решения представлены в виде перестановок, попытки совместить два направления обновления часто приводят к тому, что результирующая последовательность обменов оказывается неинформативной или содержит избыточные операции. Может возникнуть ситуация, когда локальные минимумы частиц начнут слишком быстро сближаться с глобальным минимумом, что, в свою очередь, уменьшит разнообразие решений в рое и будет способствовать преждевременному «застреванию» алгоритма в локальном оптимуме.

Для решения этой проблемы предлагается вектор обновления скорости определить непосредственно через минимальную последовательность обменов, которая преобразует текущее решение в выбранное целевое (личное или глобальное). Если разница между текущим и личным минимумом оказывается меньше заданного порога $\varepsilon > 0$ (частица достигает стационарного состояния, «застревает»), применяется случайное возмущение, например, эвристический метод локального поиска 2-opt: решение улучшается путем инвертирования последовательности вершин между двумя случайно выбранными вершинами маршрута. В результате получается новый маршрут, позволяющий выйти из локального минимума. Это предотвращает полное совпадение локального и глобального минимумов, сохраняя разнообразие решений и обеспечивая способность алгоритма продолжать эффективное исследование поискового пространства.

С учетом введенных операций рассмотрим этапы дискретного аналога алгоритма PSO.

Этап 1. Инициализация:

1) генерация начальных решений: случайным образом сгенерировать начальные перестановки $\{X_j^0\}_{j=1}^N$ (маршруты коммивояжёров);
2) установка минимумов: для каждой частицы установить личное лучшее решение $P_j \leftarrow X_j^0$ и выбрать глобальное решение $G = \arg\min_j f(X_j^0)$, где $f(X)$ — целевая функция задачи mTSP (задача «минимакс»).

Этап 2. Основной цикл (итерации $i = 0, 1, \ldots, I-1$). Для каждой частицы j выполняется:

1) выбор целевого решения T_j:
$$T_j = \begin{cases} G, \text{ с вероятностью } p, \\ P_j, \text{ с вероятностью } 1-p; \end{cases}$$
2) вычисление разности: рассчитывается минимальная последовательность обменов, переводящая X_j^i в T_j: $v_j = T_j \Theta X_i^j$;
3) масштабирование разности (усечение): выбирается коэффициент r в зависимости от цели:
— если $T_j = G$, то $r = r_1 \in (0, 1)$,
— если $T_j = P_j$, то $r = r_2 \in (0, 1]$,
тогда усеченная последовательность обменов:
$$v_j^{i+1} = r \cdot v_j = [s_1, s_2, \ldots, s_k],\ k = |r \cdot |v_j||;$$
4) обновление позиции частицы:
— если частица не находится в стационарном состоянии $|f(X_j^i) - f(P_j)| \geq \varepsilon$, то обновить положение по формуле $X_j^{i+1} = X_j^i \bigoplus (r \cdot (T_j \Theta X_j^i))$;
— если частица находится в стационарном состоянии $|f(X_j^i) - f(P_j)| < \varepsilon$, то применяется случайное возмущение $X_j^{i+1} = X_j^i \bigoplus R(X_j^i, k')$, где $R(X_j^i, k')$ — оператор случайного возмущения, генерирует случайную последовательность из k' обменов;
5) обновление личного минимума: если $f(X_j^{i+1}) < f(P_j)$ и $|f(X_j^i) - f(P_j)| \geq \varepsilon$, то обновляется $P_j \leftarrow X_j^{i+1}$;
6) обновление глобального решения: после обработки всех частиц глобальное решение пересматривается $G = \arg\min_j f(X_j^{i+1})$.

Этап 3. Завершение. Алгоритм выполняется до достижения заданного количества итераций I или до отсутствия улучшения решения в течение нескольких итераций. В конце возвращается глобальное решение G как оптимальное найденное решение.

Таким образом, каждая частица на каждом этапе учитывает собственное лучшее найденное решение P_j и глобальное лучшее решение G, а также вносит случайные корректировки, позволяющие «выходить» из возможных локальных минимумов в задаче минимизации максимального маршрута (метрика «минимакс»).

Вычислительный эксперимент

Алгоритм PSO реализован на языке Python. Вычисления проводились на аппаратной платформе, оснащенной процессором 11-го поколения Intel® Core™ i7-11800H с тактовой частотой 2,30 ГГц. Использовалось 16 ядер процессора, что позволило задействовать 16 параллельных потоков для повышения производительности вычислений.

Для оценки абсолютной эффективности алгоритма PSO при решении задачи mTSP применялись тестовые наборы стандартной библиотеки TSPLIB: eil51.tsp, berlin52.tsp, eil76.tsp, rat99.tsp на 51, 52, 76 и 99 вершин соответственно [15].

Все эксперименты проводились при фиксированном наборе параметров. Размер популяции частиц

$N = 8 \cdot 10^5$; заданное количество частиц обеспечивает устойчивое покрытие пространства решений для наборов до 100 городов. При обновлении скоростей на текущей итерации использованы только «когнитивная» и «социальная» составляющие с равными коэффициентами $c_1 = c_2 = 0{,}5$. В предлагаемой модификации метода PSO инерционный вес не учитывался ($\omega = 0$), так как движение «по инерции» подразумевает сохранение прежнего «направления» (т. е. повторное применение той же цепочки преобразований к последовательности городов), что может привести к циклическому возвращению к уже посещенным перестановкам. Вместо инерционного слагаемого вводится случайное возмущение — короткая цепочка из двух независимых транспозиций (эвристический метод локального поиска 2-opt).

Рассматривалось два экспериментальных сценария. В сценарии 1 (алгоритм PSO_random) на этапе 1 случайным образом генерировались 800 тыс. начальных частиц (перестановок с добавлением фиктивных депо для разбиения маршрута между коммивояжёрами), затем в течение 600 итераций каждая частица улучшала свое решение согласно механизму PSO. В сценарии 2 (алгоритм PSO_ACO) сначала работал алгоритм ACO, на каждой итерации которого отбиралось 45 лучших решений, что в итоге привело к накоплению 800 тыс. маршрутов. Далее полученные маршруты использовались в качестве начального приближения для оптимизации методом PSO в течение 300 итераций.

Для алгоритма ACO применена классическая схема с изоляцией лучших особей. После генерации всех маршрутов феромон усиливался вдоль лучших 45 маршрутов, определяемых, так называемыми «элитными» муравьями, и одновременно испарялся с коэффициентом $\rho = 0{,}85$. Веса феромона и видимости выбраны $\alpha = 1{,}2$ и $\beta = 1{,}09$ соответственно. Эти значения параметров обеспечили наилучший баланс «поиск/эксплуатация» на предварительных прогонах.

На рис. 1 изображены графики динамики оптимизации для двух, трех, 5 и 7 агентов для набора rat99.tsp (каждый график в отдельной системе координат). Для остальных наборов и при различном количестве агентов наблюдаются аналогичные тенденции: на начальном этапе происходит резкое снижение целевой метрики, за которым следует фаза постепенного уменьшения. Варианты с бóльшим количеством агентов демонстрируют более низкое итоговое значение максимальной длины маршрута (рис. 2).

Алгоритм PSO имеет вероятностный характер, что проявляется в снижении темпов улучшения решения по мере увеличения числа итераций. Начальные итерации обеспечивают значительное улучшение, однако с течением времени прирост оптимизации становится менее выраженным. При сравнении экспериментов с двумя тремя и с 5–7 агентами видно, что лучшие показатели минимаксного критерия достигаются при большем числе агентов. Это может быть связано с более равномерным распределением нагрузки между маршрутами

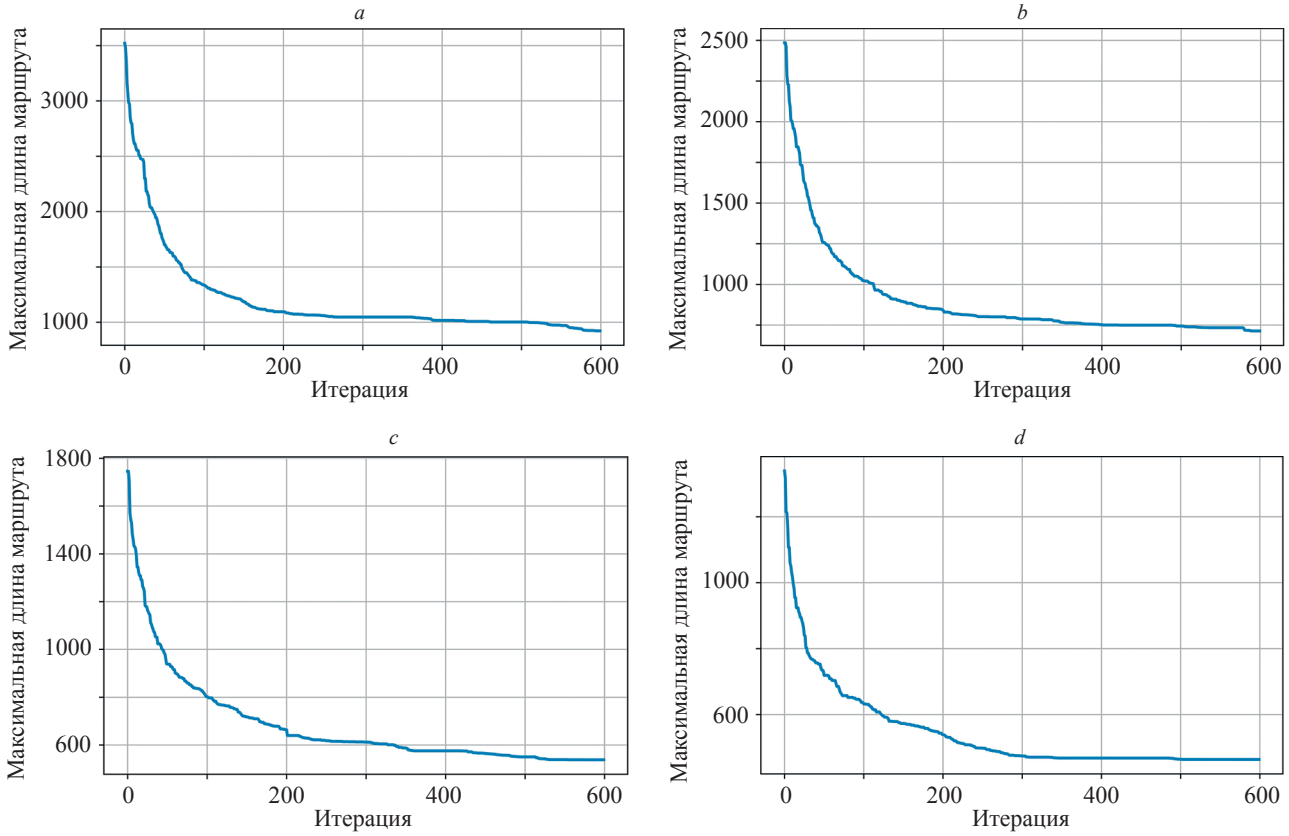

Рис. 1. Динамика оптимизации методом PSO для двух (*a*); трех (*b*); 5 (*c*) и 7 (*d*) агентов на примере набора rat99.tsp

Fig. 1. PSO optimization dynamics for the multiple traveling salesmen problem with 2 (*a*), 3 (*b*), 5 (*c*), and 7 (*d*) salesmen on the rat99.tsp benchmark instance

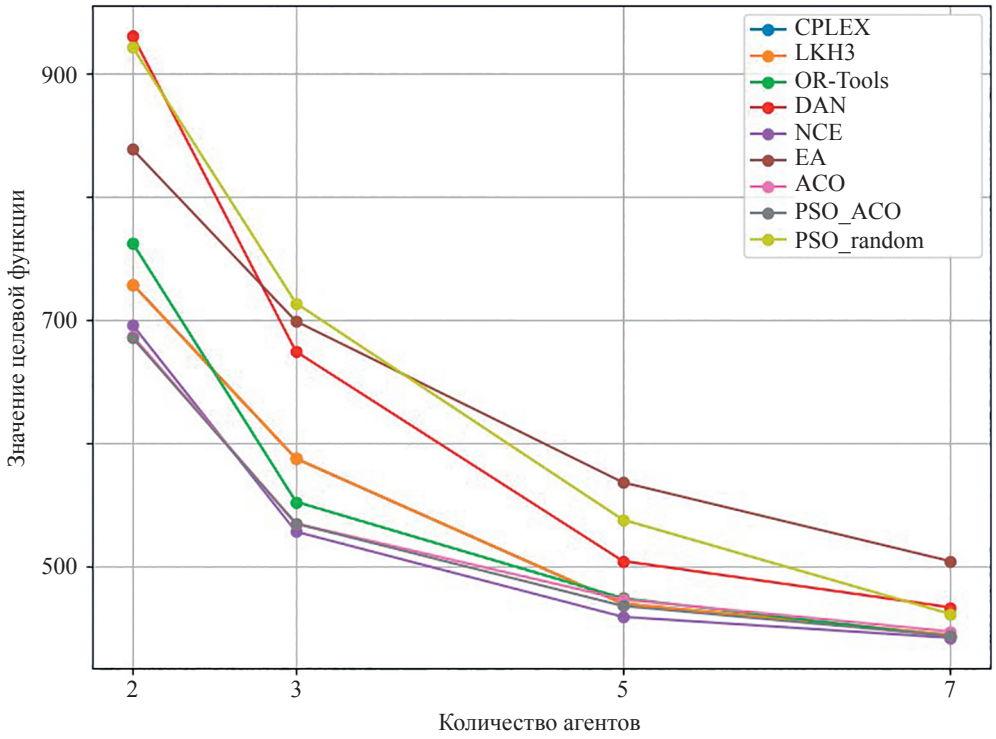

Рис. 2. Сравнение итоговых значений метрики «минимакс» для разных методов при различном числе агентов на примере rat99.tsp

Fig. 2. Comparison of final minimax metric values for different methods with varying numbers of salesmen on the rat99.tsp benchmark instance

и особыми свойствами окрестностей глобального оптимума в пространстве решений.

По ходу оптимизации PSO наблюдается поиск решений в локальной окрестности текущего глобального оптимума, которые можно определить как множество маршрутов, отличающихся от глобального оптимума одним (loc-1), двумя (loc-2) или несколькими обменами городов. По мере увеличения такого расстояния (количества обменов) плотность благоприятных решений уменьшается, и вероятность найти маршрут, более оптимальный, чем текущий глобальный, снижается. Например, для 99 вершин количество вариантов перестановок в окрестности loc-2 имеет порядок 10^8, а в окрестности loc-3 — порядок 10^{12}, при этом общее число вариантов, которые генерируются программой метода PSO при 600 итерациях и 800 тыс. частиц, не превышает $5 \cdot 10^8$, что еще раз подчеркивает эффективность и важность целенаправленного поиска в суженном пространстве решений.

Для объективного сравнения результаты работы алгоритма PSO сопоставлялись с характеристиками базовых методов (CPLEX [10], LKH–3 [15], ORTools [16]), нейросетевых методов (DAN, NCE [10]), генетического (эволюционного) метода (EA [9]), алгоритма ACO и комбинированного подхода к методам оптимизации PSO_ACO. Количественные показатели базовых методов были взяты из работ [9, 10, 15, 16]. Для корректного сопоставления использовалась метрика «минимакс», так как именно этот показатель представлен в большинстве базовых исследований. В таблице представлены сравнительные результаты работы перечисленных методов для двух, трех, 5 и 7 агентов и на разных тестовых наборах.

Сопоставление значений по «минимаксному» критерию, представленных в таблице, показывает, что комбинированная схема PSO_ACO на всех четырех наборах библиотеки TSPLIB (eil51, berlin52, eil76, rat99) и при любом из рассматриваемом числе агентов (два, три, 5, 7) неизменно возвращает маршруты меньшей длины, чем метод PSO_random. Наибольший выигрыш отмечается для более крупных наборов при двух-трех агентах, тогда как для eil51 при 7 агентах разница минимальна, но остается в пользу PSO_ACO.

На рис. 3 показаны маршруты для 5 агентов на наборе eil51 с указанием соответствующих значений «минимаксного» критерия: PSO_random (рис. 3, *a*) формирует явно неравномерную нагрузку («минимакс» равен 151,7), ACO (рис. 3, *c*) и PSO_ACO (рис. 3, *d*) демонстрируют более сбалансированные маршруты («минимакс» равен 118,2 и 118,1 соответственно), а CPLEX уступает ACO и PSO_ACO по визуальной четкости и показателю «минимакс», равному 124,0 (рис. 3, *b*).

Таким образом, PSO_ACO не только устойчиво превосходит PSO_random количественно на всех тестах, но и визуально обеспечивает более равномерное распределение нагрузки между агентами. Во всех 16 экспериментальных конфигурациях значения, полученные PSO_ACO, не превышают соответствующие результаты PSO_random, что указывает на более высокую результативность гибридного алгоритма и его стабильность при варьировании размеров задачи и числа агентов. Данное превосходство подтверждает высокий потенци-

Таблица. Значения «минимаксного» критерия для разных методов решения

Table. Minimax criterion values for the various solution methods

Метод решения	Тестовый набор													
	eil51				berlin52			eil76				rat99		
	Число агентов				Число агентов			Число агентов				Число агентов		
	2	3	5	7	2	3	5	7	2	3	5	7	2	3	5	7
CPLEX	**222,7**	**159,6**	124,0	**112,1**	**4110,2**	3244,4	**2441,4**	**2440,9**	**280,9**	197,3	150,3	139,6	728,8	587,2	469,3	443,9
LKH3	**222,7**	**159,6**	124,0	**112,1**	**4110,2**	3244,4	**2441,4**	**2440,9**	**280,9**	197,3	150,3	139,6	728,8	587,2	469,3	443,9
OR-Tools	243,3	170,5	127,5	**112,1**	4665,5	3311,3	2482,6	**2440,9**	318,0	212,4	**143,4**	128,3	762,2	552,1	473,7	442,5
DAN	274,2	178,9	158,6	118,1	5226,0	4278,0	2759,0	2697,0	361,1	251,5	170,9	148,5	930,8	674,1	504,0	466,4
NCE	235,0	170,3	121,6	**112,1**	**4110,2**	3274,0	2660,0	2441,0	285,5	211,0	144,6	**127,6**	695,8	527,8	458,6	**441,6**
EA	248,5	190,6	134,8	116,5	4472,8	3567,4	2694,6	2441,4	342,0	259,2	190,5	158,8	838,4	698,7	567,9	504,0
ACO	**222,7**	**159,6**	118,2	112,5	4115,9	3188,1	2466,8	**2440,9**	281,9	197,0	150,9	128,2	686,6	534,4	472,7	446,7
PSO_ACO	**222,7**	**159,6**	**118,1**	**112,1**	**4110,2**	**3171,6**	2455,6	**2440,9**	281,8	**196,0**	143,8	127,8	**685,2**	**534,3**	**467,5**	443,2
PSO_random	264,5	183,7	151,7	112,7	5200,3	3561,1	2758,2	2445,2	382,4	262,6	167,7	149,5	921,5	713,4	537,6	461,1

Примечание. Полужирным шрифтом выделены наименьшие значения «минимаксного» критерия для каждого теста и различного числа агентов.

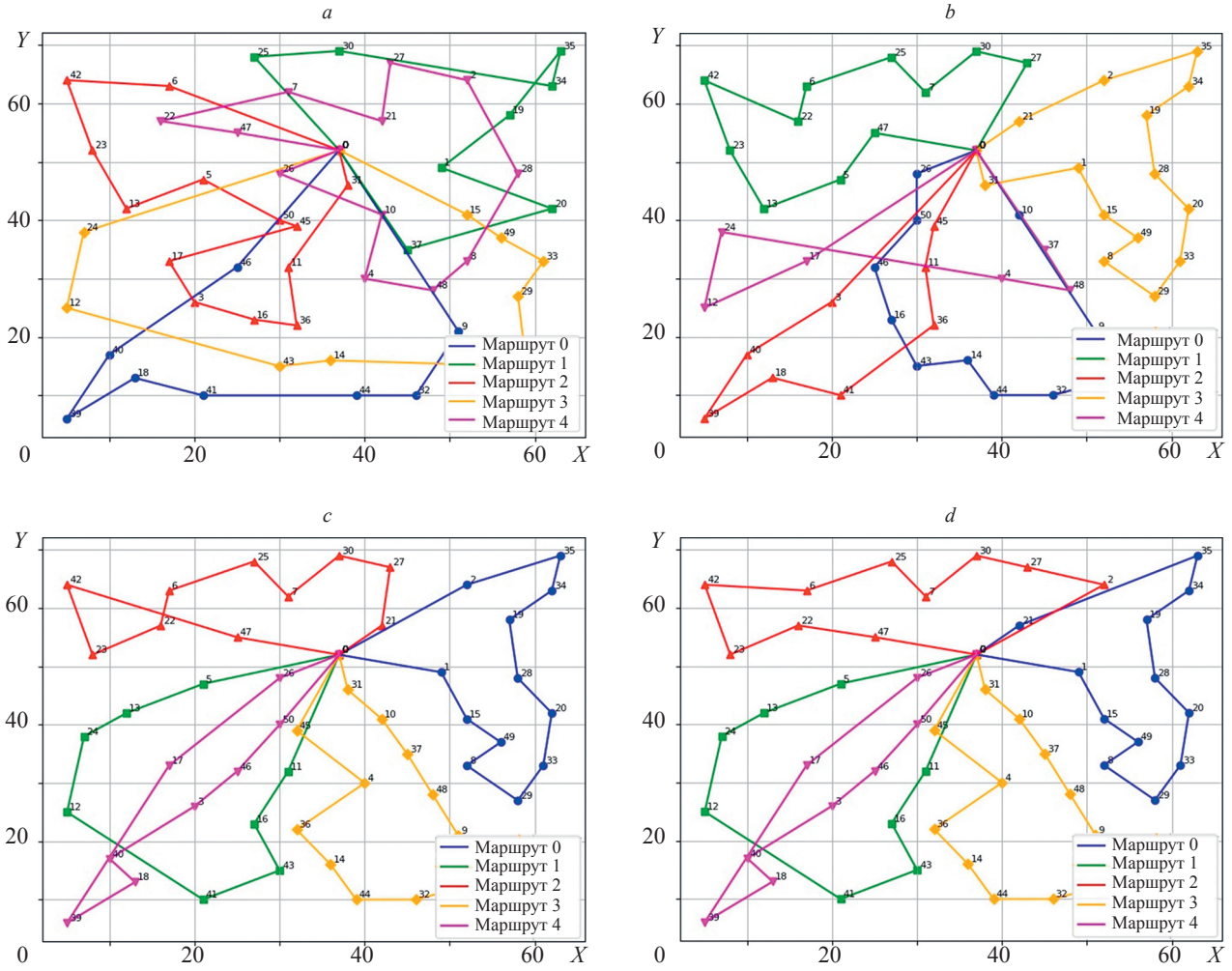

Рис. 3. Визуализация маршрутов для 5 агентов на наборе eil51 с общим депо 0 при разных методах решения: PSO_random (*a*); CPLEX (*b*); ACO (*c*); PSO_ACO (*d*)

Fig. 3. Routes of 5 agents on the eil51 benchmark (common depot 0) obtained by different methods: PSO_random (*a*), CPLEX (*b*), ACO (*c*), PSO_ACO (*d*)

ал комбинирования роевой оптимизации с алгоритмами ACO для решения задачи mTSP по «минимаксному» критерию.

Заключение

Алгоритм роя частиц (Particle Swarm Optimization, PSO) обладает рядом преимуществ, таких как способность к глобальному поиску, простота реализации и возможность параллелизации вычислений.

Можно выделить несколько направлений для дальнейшего развития метода PSO. Одним из них является интеграция дискретного PSO с локальными эвристиками, такими как 2-opt, а также с элементами алгоритмов муравьиной оптимизации для более глубокого исследования пространства решений. Важным направлением является адаптивная настройка параметров алгоритма PSO по мере эволюции роя, что позволяет избежать преждевременного «застревания» в локальных минимумах. Кроме того, представляет интерес целенаправленное исследование разреженных областей в окрестностях локальных минимумов (loc-k) с использованием дополнительных метрик расстояния между перестановками и методов кластеризации. Наконец, дальнейший теоретический анализ вычислительной сложности алгоритма, а также получение новых результатов о верхних границах длины цепочек обменов будут способствовать более глубокому пониманию и более надежному применению данного подхода.

Проведенный вычислительный эксперимент демонстрирует потенциал разработанного подхода к методу PSO для решения мультиагентной задачи коммивояжёра (Multiple Traveling Salesman Problem, mTSP) по «минимаксному» критерию, а представленные результаты позволяют визуально и количественно оценить эффективность предложенной методики на тестовых TSP-наборах.

Литература

1. Carter A.E., Ragsdale C.T. A new approach to solving the multiple traveling salesperson problem using genetic algorithms // European Journal of Operational Research. 2006. V. 175. N 1. P. 246–257. https://doi.org/10.1016/j.ejor.2005.04.027
2. Смирнов А.В. Исследование влияния степени овражности целевой функции на погрешность определения координат ее минимума // Российский технологический журнал. 2023. Т. 11. № 6. С. 57–67. https://doi.org/10.32362/2500-316X-2023-11-6-57-67
3. Boudjelaba K., Ros F., Chikouche D. Potential of particle swarm optimization and genetic algorithms for FIR filter design // Circuits, Systems, and Signal Processing. 2014. V. 33. N 10. P. 3195–3222. https://doi.org/10.1007/s00034-014-9800-y
4. Soylu B. A general variable neighborhood search heuristic for multiple traveling salesmen problem // Computers & Industrial Engineering. 2015. V. 90. P. 390–401. https://doi.org/10.1016/j.cie.2015.10.010
5. Elloumi W., Abeda H.E., Abraham A., Alimi A.M. A comparative study of the improvement of performance using a PSO modified by ACO applied to TSP // Applied Soft Computing. 2014. V. 25. P. 234–241. https://doi.org/10.1016/j.asoc.2014.09.031
6. Tang L., Liu J., Rong A., Yang Z. A multiple traveling salesman problem model for hot rolling scheduling in Shanghai Baoshan Iron & Steel Complex // European Journal of Operational Research. 2000. V. 124. N 2. P. 267–282. https://doi.org/10.1016/S0377-2217(99)00380-X
7. Lu L.C., Yue T.W. Mission-oriented ant-team ACO for min–max MTSP // Applied Soft Computing. 2019. V. 76. P. 436–444. https://doi.org/10.1016/j.asoc.2018.11.048
8. Eberhart R.C., Shi Y. Comparison between genetic algorithms and particle swarm optimization // Lecture Notes in Computer Science. 1998. V. 1447. P. 611–616. https://doi.org/10.1007/BFb0040812
9. Lupoaie V.-I., Chili I.-A., Breaban M.E., Raschip M. SOM-guided evolutionary search for solving MinMax Multiple-TSP // Proc. of the IEEE Congress on Evolutionary Computation (CEC). 2019. P. 73–80. https://doi.org/10.1109/cec.2019.8790276
10. Kim M., Park J., Park J. Learning to CROSS exchange to solve min-max vehicle routing problems // Proc. of the 11th International Conference on Learning Representations (ICLR). 2023. P. 1–12.
11. Tasgetiren M.F., Sevkli M., Liang Y.C., Gencyilmaz G. Particle swarm optimization algorithm for permutation flowshop sequencing problem // Lecture Notes in Computer Science. 2004. V. 3172. P. 382–389. https://doi.org/10.1007/978-3-540-28646-2_38
12. Liao C.J., Tseng C.T., Luarn P. A discrete version of particle swarm optimization for flowshop scheduling problems // Computers and Operations Research. 2007. V. 34. N 10. P. 3099–3111. https://doi.org/10.1016/j.cor.2005.11.017
13. Junjie P., Dingwei W. An ant colony optimization algorithm for Multiple Travelling Salesman Problem // Proc. of the First International Conference on Innovative Computing, Information and Control (ICICIC'06). 2006. V. 1. P. 210–213. https://doi.org/10.1109/icicic.2006.40
14. Necula R., Breaban M., Raschip M. Tackling the Bi-criteria facet of Multiple Traveling Salesman Problem with Ant Colony Systems // Proc. of the IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI). 2015. P. 873–880. https://doi.org/10.1109/ICTAI.2015.127
15. Helsgaun K. An Extension of the Lin-Kernighan-Helsgaun TSP solver for constrained traveling salesman and vehicle routing problems // Occasional Paper of the Roskilde University, International Development Studies. 2017. V. 12. P. 966–980.
16. Perron L., Furnon V. OR-Tools v9.6. 2019. URL: https://developers.google.com/optimization/

References

1. Carter A.E., Ragsdale C.T. A new approach to solving the multiple traveling salesperson problem using genetic algorithms. *European Journal of Operational Research*, 2006, vol. 175, no. 1, pp. 246–257. https://doi.org/10.1016/j.ejor.2005.04.027
2. Smirnov A.V. Investigation of influence of objective function valley ratio on the determination error of its minimum coordinates. *Russian Technological Journal*, vol. 11, no. 6, pp. 57–67. (in Russian). https://doi.org/10.32362/2500-316X-2023-11-6-57-67
3. Boudjelaba K., Ros F., Chikouche D. Potential of particle swarm optimization and genetic algorithms for FIR filter design. *Circuits, Systems, and Signal Processing*, 2014, vol. 33, no. 10, pp. 3195–3222. https://doi.org/10.1007/s00034-014-9800-y
4. Soylu B. A general variable neighborhood search heuristic for multiple traveling salesmen problem. *Computers & Industrial Engineering*, 2015, vol. 90, pp. 390–401. https://doi.org/10.1016/j.cie.2015.10.010
5. Elloumi W., Abeda H.E., Abraham A., Alimi A.M. A comparative study of the improvement of performance using a PSO modified by ACO applied to TSP. *Applied Soft Computing*, 2014, vol. 25, pp. 234–241. https://doi.org/10.1016/j.asoc.2014.09.031
6. Tang L., Liu J., Rong A., Yang Z. A multiple traveling salesman problem model for hot rolling scheduling in Shanghai Baoshan Iron & Steel Complex. *European Journal of Operational Research*, 2000, vol. 124, no. 2, pp. 267–282. https://doi.org/10.1016/S0377-2217(99)00380-X
7. Lu L.C., Yue T.W. Mission-oriented ant-team ACO for min–max MTSP. *Applied Soft Computing*, 2019, vol. 76, pp. 436–444. https://doi.org/10.1016/j.asoc.2018.11.048
8. Eberhart R.C., Shi Y. Comparison between genetic algorithms and particle swarm optimization. *Lecture Notes in Computer Science*, 1998, vol. 1447, pp. 611–616. https://doi.org/10.1007/BFb0040812
9. Lupoaie V.-I., Chili I.-A., Breaban M.E., Raschip M. SOM-guided evolutionary search for solving MinMax Multiple-TSP. *Proc. of the IEEE Congress on Evolutionary Computation (CEC)*, 2019, pp. 73–80. https://doi.org/10.1109/cec.2019.8790276
10. Kim M., Park J., Park J. Learning to CROSS exchange to solve min-max vehicle routing problems. *Proc. of the 11th International Conference on Learning Representations (ICLR)*, 2023, pp. 1–12.
11. Tasgetiren M.F., Sevkli M., Liang Y.C., Gencyilmaz G. Particle swarm optimization algorithm for permutation flowshop sequencing problem. *Lecture Notes in Computer Science*, 2004, vol. 3172, pp. 382–389. https://doi.org/10.1007/978-3-540-28646-2_38
12. Liao C.J., Tseng C.T., Luarn P. A discrete version of particle swarm optimization for flowshop scheduling problems. *Computers and Operations Research*, 2007, vol. 34, no. 10, pp. 3099–3111. https://doi.org/10.1016/j.cor.2005.11.017
13. Junjie P., Dingwei W. An ant colony optimization algorithm for Multiple Travelling Salesman Problem. *Proc. of the First International Conference on Innovative Computing, Information and Control (ICICIC'06)*, 2006, vol. 1, pp. 210–213. https://doi.org/10.1109/icicic.2006.40
14. Necula R., Breaban M., Raschip M. Tackling the Bi-criteria facet of Multiple Traveling Salesman Problem with Ant Colony Systems. *Proc. of the IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2015, pp. 873–880. https://doi.org/10.1109/ICTAI.2015.127
15. Helsgaun K. An Extension of the Lin-Kernighan-Helsgaun TSP solver for constrained traveling salesman and vehicle routing problems. *Occasional Paper of the Roskilde University, International Development Studies*, 2017, vol. 12, pp. 966–980.
16. Perron L., Furnon V. *OR-Tools v9.6*, 2019. Available at : https://developers.google.com/optimization/

Авторы

Мифтахов Эльдар Наилевич — доктор физико-математических наук, профессор, МИРЭА — Российский технологический университет, Москва, 119454, Российская Федерация, SC 56178153800, https://orcid.org/0000-0002-0471-5949, promif@mail.ru

Акимов Андрей Анатольевич — кандидат физико-математических наук, доцент, доцент, МИРЭА — Российский технологический университет, Москва, 119454, Российская Федерация, SC 56428598700, https://orcid.org/0000-0003-3387-2959, andakm@yandex.ru

Authors

Eldar N. Miftakhov — D.Sc. (Physics & Mathematics), Professor, MIREA — Russian Technological University, Moscow, 119454, Russian Federation, SC 56178153800, https://orcid.org/0000-0002-0471-5949, promif@mail.ru

Andrey A. Akimov — PhD (Physics & Mathematics), Associate Professor, Associate Professor, MIREA — Russian Technological University, Moscow, 119454, Russian Federation, SC 56428598700, https://orcid.org/0000-0003-3387-2959, andakm@yandex.ru

Гнатенко Юлия Ахнафовна — кандидат физико-математических наук, доцент, доцент, Стерлитамакский филиал Уфимского университета науки и технологий, Стерлитамак, 453103, Российская Федерация, sc 9234055300, https://orcid.org/0009-0009-9264-3989, y.a.gnatenko@struust.ru

Yuliya A. Gnatenko — PhD (Physics & Mathematics), Associate Professor, Associate Professor, Branch of the Ufa University of Science and Technology, Sterlitamak, 453103, Russian Federation, sc 9234055300, https://orcid.org/0009-0009-9264-3989, y.a.gnatenko@struust.ru

Статья поступила в редакцию 11.03.2025
Одобрена после рецензирования 20.08.2025
Принята к печати 24.09.2025

Received 11.03.2025
Approved after reviewing 20.08.2025
Accepted 24.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-866-875

Accelerating and analyzing performance of shortest path algorithms on GPU using CUDA platform: Bellman-Ford, Dijkstra, and Floyd-Warshall algorithms

Deep Bodra[1]✉, Sushil Khairnar[2]

[1] Harrisburg University of Science and Technology, Harrisburg, 17101, USA
[2] Virginia Tech, Virginia, 24061, USA

[1] Deepbodra97@gmail.com✉, https://orcid.org/0009-0009-4173-2447
[2] sushilk@vt.edu, https://orcid.org/0009-0006-5192-0175

Abstract
The computational demands of the shortest path algorithms on large-scale graphs with millions of vertices and edges pose significant challenges for serial implementations, often requiring hours of execution time even on powerful CPUs. This paper evaluates Graphic Processing Units implementations of three fundamental shortest path algorithms — Bellman-Ford, Dijkstra, and Floyd-Warshall using NVIDIA CUDA platform. We implemented and compared multiple variants of each algorithm, starting with basic parallel approaches and applying various optimization techniques, including grid-stride loops, shared memory utilization, memory coalescing, and algorithm-specific enhancements such as flag-based early termination for Bellman-Ford and tiled computation for Floyd-Warshall. Our study provides performance analysis comparing different optimization strategies and their effectiveness across various graph datasets.

Keywords
GPU computing, CUDA platform, shortest path algorithms, parallel algorithms, graph algorithms, Bellman-Ford, Dijkstra, Floyd-Warshall, performance optimization

For citation: Bodra D., Khairnar S. Accelerating and analyzing performance of shortest path algorithms on GPU using CUDA platform: Bellman-Ford, Dijkstra, and Floyd-Warshall algorithms. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 866–875. doi: 10.17586/2226-1494-2025-25-5-866-875

УДК 004.424.4

Ускорение и анализ производительности алгоритмов поиска кратчайшего пути на GPU с использованием платформы CUDA: алгоритмы Беллмана–Форда, Дейкстры и Флойда–Уоршелла

Дип Бодра[1]✉, Сушил Хайрнар[2]

[1] Гаррисбергский университет науки и технологий, Гаррисберг, 17101, США
[2] Технологический институт Вирджинии, Вирджиния, 24061, США

[1] Deepbodra97@gmail.com✉, https://orcid.org/0009-0009-4173-2447
[2] sushilk@vt.edu, https://orcid.org/0009-0006-5192-0175

Аннотация
Вычислительные требования к алгоритмам поиска кратчайшего пути на больших графах с миллионами вершин и ребер представляют собой значительную проблему для последовательных реализаций, часто требуя многочасового времени выполнения даже с помощью мощных процессоров. В работе выполнена оценка реализации на графических процессорах трех фундаментальных алгоритмов поиска кратчайшего пути: Беллмана–Форда, Дейкстры и Флойда–Уоршелла с использованием платформы NVIDIA CUDA. Проведено сравнение нескольких вариантов каждого алгоритма, от базовых параллельных подходов до специфических алгоритмов улучшения. Исследованы базовые методы оптимизации, включая циклы с шагом сетки, использование общей памяти, объединение памяти. Также выполнен анализ алгоритмов улучшения, таких как раннее завершение на основе флагов для алгоритма Беллмана–Форда и тайловые вычисления для алгоритма

Флойда–Уоршелла. В исследовании представлен анализ производительности, выполнено сравнение различных стратегий оптимизации и их эффективности на различных наборах графовых данных.

Ключевые слова
вычисления на GPU, платформа CUDA, алгоритмы поиска кратчайшего пути, параллельные алгоритмы, алгоритмы графов, Беллман–Форд, Дейкстра, Флойд–Уоршелл, оптимизация производительности

Ссылка для цитирования: Бодра Д., Хайрнар С. Ускорение и анализ производительности алгоритмов поиска кратчайшего пути на GPU с использованием платформы CUDA: алгоритмы Беллмана–Форда, Дейкстры и Флойда–Уоршелла // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 866–875 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-866-875

Introduction

Shortest path algorithms have applications across domains including transportation networks, communication systems, social network analysis, and Very Large-Scale Integration chip design [1]. These algorithms solve the problem of finding the minimum cost path between vertices in weighted graphs, which is required in real-world scenarios such as GPS navigation systems to find optimal routes, network protocols to determine efficient data transmission paths, and circuit designers to optimize signal routing. The computational complexity of the problem becomes challenging as graph sizes grow to millions of vertices and edges, commonly encountered in modern applications, making serial implementations impractical for time-sensitive applications [1, 2].

Modern Graphics Processing Units (GPUs) contain many cores capable of executing operations in parallel which make them suitable for algorithms that can exploit data parallelism. However, parallelizing shortest path algorithms on GPU presents unique challenges including irregular memory access patterns, varying computational loads across threads, and complex data dependencies that can limit parallel efficiency [1, 2]. The inherent nature of shortest path algorithms can affect their suitability for GPU implementation. Bellman-Ford can handle negative edge weights but requires multiple iterations, Dijkstra provides better performance for non-negative weights but has inherent sequential dependencies, and Floyd-Warshall computes all-pairs shortest paths with high computational complexity [3].

This paper presents a comprehensive study of GPU implementations for three fundamental shortest path algorithms using Compute Unified Device Architecture (CUDA). We implement and evaluate multiple optimization strategies for each algorithm, progressing from basic parallel approaches to sophisticated techniques including memory optimization, algorithmic enhancements, and architecture-specific optimizations. Our evaluation focuses on understanding the performance characteristics and trade-offs of different implementation approaches across various graph datasets.

Background of the problem

The Bellman-Ford algorithm finds the shortest paths from a source vertex to all other vertices in a weighted graph [2]. The key advantage of Bellman-Ford over Dijkstra algorithm is its ability to detect negative cycles and handle graphs containing negative cycles reachable from the source vertex [4]. The algorithm performs a series of relaxation and iteratively improves distance estimates until optimal paths are found. For a graph with $|V|$ vertices, the algorithm performs $|V| - 1$ iterations, where each of iteration relaxes all edges and updates distance estimates if a shorter path is discovered [2]. Since the order of edge relaxation within iteration does not affect correctness, parallelization can be achieved by allowing multiple edges to be processed simultaneously.

When applied to the all-pairs shortest path problem, Bellman-Ford must be executed $|V|$ times, once for each source vertex. For number of edges $|E|$, the sequential time complexity becomes $O(|V|^2 \times |E|)$, as each of the $|V|$ source vertices require $O(|V| \times |E|)$ time for single-source computation.

The space complexity for storing all-pairs distances is $O(|V|^2)$ for the distance matrix, plus $O(|V| + |E|)$ for the graph representation, yielding total space complexity $O(|V|^2 + |E|)$. The answer size is $O(|V|^2)$ for distance values only, or $O(|V|^3)$ if complete path information is stored for all vertex pairs [2].

Dijkstra algorithm solves the single-source shortest path problem by maintaining a priority queue of vertices ordered by their current shortest distance estimate and greedily selects the vertex with minimum distance for processing. The serial version uses a min-heap to extract the closest unvisited vertex, and then relaxes all outgoing edges from that vertex. This inherently sequential process of selecting the next minimum vertex poses significant challenges for parallelization [1, 3]. However, parallel versions can be implemented by running multiple instances simultaneously, computing shortest paths from different source vertices. For all-pairs shortest path computation, Dijkstra algorithm must be executed $|V|$ times, once from each source vertex.

The sequential time complexity becomes $O(|V| \times (|V| + |E|) \log_2|V|)$, representing $|V|$ executions of single-source Dijkstra with $O((|V| + |E|) \log_2|V|)$ complexity each. The space complexity requires $O(|V|^2)$ for storing the complete distance matrix, plus $O(|V|)$ for the priority queue and temporary arrays during each execution, yielding total space complexity $O(|V|^2)$. The answer size is $O(|V|^2)$ for distance information, or $O(|V|^3)$ when complete shortest path trees are maintained for all source vertices [2].

The Floyd-Warshall's algorithm computes shortest paths between all pairs of vertices in a weighted graph [2]. It can handle negative edge weights but not negative cycles. The algorithm employs dynamic programming, considering each vertex as an intermediate point in potential shortest paths. The algorithm executes $|V|$ iterations, where iteration k considers vertex k as an intermediate vertex for all vertex pairs (i, j). For each pair, it checks whether the path $i \to k \to j$ offers a shorter distance than the current best path

from i to j [3]. This structure makes Floyd-Warshall highly suitable for parallelization, as all pairwise distance updates (within each iteration) can be performed independently. Floyd-Warshall is inherently designed for the all-pairs shortest path problem. The sequential time complexity is $O(|V|^3)$, as it performs $|V|$ iterations, each examining all $|V|^2$ vertex pairs for potential improvement through the current intermediate vertex. The space complexity is $O(|V|^2)$ for storing the distance matrix which directly represents the shortest distances between all vertex pairs. The answer size is exactly $O(|V|^2)$ for distance information, or $O(|V|^3)$ when complete path reconstruction information is maintained. Unlike Bellman-Ford and Dijkstra, Floyd-Warshall complexity does not scale with the number of edges $|E|$, making it particularly efficient for dense graphs where $|E|$ approaches $|V|^2$ [5, 6].

CUDA Computing

GPUs feature a massively parallel architecture designed for high-throughput computation[1]. Unlike CPUs with few powerful cores optimized for sequential processing, modern GPUs contain thousands of smaller cores that excel at executing the same operation across multiple data elements simultaneously. This Single Instruction, Multiple Data architecture makes GPUs particularly effective for data-parallel algorithms[1].

CUDA provides a programming framework for general-purpose GPU computing developed by NVIDIA. In CUDA, parallel work is organized into kernels — functions executed simultaneously by many threads. Threads are grouped into blocks, and blocks are organized into a grid structure. This hierarchical organization enables efficient resource management and communication patterns[1].

The CUDA programming model encompasses several key concepts relevant to this work. The thread hierarchy organizes individual threads that execute kernel code into blocks that can share memory and synchronize operations. The memory hierarchy provides different levels of storage including global memory with large capacity but high latency, shared memory that offers fast access but limited capacity per block, and registers that provide the fastest access but are limited per thread[1] Grid-stride loops represent a programming pattern that allows kernels to process datasets larger than the number of available threads. Atomic operations provide hardware-supported mechanisms for thread-safe memory updates, which are crucial for avoiding race conditions in parallel algorithms [1].

Challenges in GPU Graph Algorithm Implementation

Implementing graph algorithms on GPUs presents several significant challenges [3] that have been extensively documented in the literature. Graph traversal often results in irregular memory access patterns that can substantially reduce GPU efficiency, as the hardware is optimized for coalesced memory accesses [6, 7]. Load balancing presents another critical challenge, as vertices may have vastly different degrees, leading to uneven work distribution across threads and resulting in some threads completing their work much earlier than others [6, 7]. Synchronization requirements in many graph algorithms necessitate coordination between threads, which can potentially limit the achievable parallelism and introduce performance bottlenecks [1, 3]. Memory bandwidth limitations arise when large graphs exceed GPU memory capacity or create memory bandwidth bottlenecks that constrain overall performance [3]. Understanding these fundamental challenges is essential for developing effective GPU implementations of shortest path algorithms and forms the foundation for the optimization strategies explored in this work.

Related Works

Yang et al. [8] proposed a Fast APSP algorithm that combines Floyd-Warshall with Dijkstra algorithms for large sparse graphs, achieving an average speedup of 16.97 times compared to CPU Dijkstra and 7.09 times compared to GPU Dijkstra implementations. Their work addresses graphs with over 11 million vertices using 2048 GPUs, demonstrating scalability beyond single-GPU implementations. Prihozhy and Karasik[2] conducted comprehensive comparisons of competing all-pairs shortest path algorithms for both sparse and dense graphs, providing valuable insights into algorithm selection criteria. However, their approach requires distributed computing clusters, whereas our work focuses on optimizing single-GPU performance through tiling and shared memory techniques.

Tang et al. [9] developed GPU-accelerated all-pairs shortest path algorithms specifically for stochastic road networks, reporting “thousands of times improvement” in acceleration for real-world navigation applications. While their work targets similar applications to ours, their focus on stochastic networks with uncertainty handling differs from our deterministic graph optimization approach. Recent research has explored innovative GPU-based methods for related graph problems. Spridon et al. [10] introduced novel GPU-based approaches for the generalized maximum flow problem, demonstrating the continued evolution of GPU graph algorithm development. These advances in related graph problems inform optimization strategies applicable to shortest path algorithms.

Traditional GPU implementations of Bellman-Ford have focused on basic parallelization strategies. Agarwal and Dutta [11] introduced flag-based optimization for GPU Bellman-Ford, which serves as a foundation for our strided with flag implementation. However, recent literature shows limited advancement in Bellman-Ford GPU optimization techniques beyond basic parallelization patterns, indicating a gap that our comprehensive optimization analysis addresses.

Early GPU implementations by Harish and Narayanan [1] demonstrated the potential for GPU acceleration of graph algorithms, achieving significant speedups for

[1] CUDA C++ Programming Guide. Available at: https://docs.nvidia.com/cuda/cuda-c-programming-guide/, free. English lang. (accessed: 02.06.2025).

[2] Prihozhy A.A., & Karasik O.N. (2024). Competing all-pairs shortest paths algorithms for sparse/dense graphs: implementation and comparison. Available at: http://dx.doi.org/10.21122/2309-4923-2024-4-4-12, free. English lang. (accessed: 06.06.2025).

single-source shortest path computation on graphs with millions of vertices. Recent work by Song [12] explored high-performance parallelization of Dijkstra algorithm using hybrid Message Passing Interface and CUDA approaches, demonstrating the continued relevance of GPU acceleration for shortest path problems. However, these implementations focused on single-source shortest paths, while our parallel all-pairs approach addresses the inherent limitations of parallelizing Dijkstra sequential vertex selection process.

Contemporary research has expanded GPU shortest path applications beyond traditional graph problems. Bengtsson et al. [13] applied GPU-accelerated routing to warehouse optimization problems, combining clustering and dynamic systems modeling with GPU-based shortest path computation. Kumar et al. [14] investigated Artificial Intelligent based navigation in quasi-structured environments, highlighting the growing importance of GPU-accelerated path planning in robotics and autonomous systems. These applications demonstrate the practical relevance of efficient GPU shortest path implementations across diverse domains.

The effectiveness of shared memory utilization in GPU graph algorithms has been demonstrated across multiple studies [6]. However, comprehensive analysis of the trade-offs between different memory optimization techniques for shortest path algorithms remains limited in recent literature. Our work contributes detailed performance analysis of shared memory vs. global memory approaches specifically for Floyd-Warshall tiled implementations.

Recent research has emphasized the importance of achieving performance portability across different computing architectures. Morgan et al. [15] investigated simplified approaches to achieve parallel performance and portability across CPU and GPU architectures, highlighting the challenges of maintaining efficiency across diverse hardware platforms. This work underscores the importance of architecture-specific optimizations like those explored in our study.

Harris [7] introduced grid-stride loops as a fundamental CUDA optimization pattern. While this technique has been applied to various GPU algorithms, systematic evaluation of its effectiveness for different shortest path algorithm variants has not been thoroughly investigated in recent literature.

Our Contribution

Our work advances the current state of GPU shortest path algorithm research through several distinct contributions that address gaps in the existing literature. Unlike recent studies that focus on individual algorithms or specific application domains, we provide a comprehensive systematic comparison across three fundamental shortest path algorithms using consistent experimental methodology and hardware platforms. This multi-algorithm approach enables direct performance comparisons and reveals algorithm-specific optimization opportunities that have not been thoroughly explored in recent literature.

We implement and evaluate multiple optimization strategies for each algorithm, progressing from basic parallel implementations to sophisticated techniques that combine algorithmic and architectural optimizations. This progressive optimization evaluation methodology provides detailed insights into the incremental effects of different optimization techniques which has been limited in recent publications that typically focus on single optimization approaches.

Our study contributes quantitative analysis of the effectiveness of specific optimization techniques, including flag-based early termination for Bellman-Ford, tiled computation for Floyd-Warshall, and shared memory utilization across algorithms. The systematic evaluation of these techniques provides performance trade-off analysis that has been absent from recent literature.

We demonstrate that optimal GPU implementation strategies vary significantly across different shortest path algorithms, providing practical guidance for algorithm selection and optimization in real-world applications. This algorithm-specific optimization insight fills a gap in current literature where optimization strategies are often presented as universally applicable without considering algorithm-specific characteristics.

While prior work has established the foundation for GPU shortest path algorithms, our comprehensive analysis of optimization strategies and their trade-offs provides novel insights that advance the current state of knowledge in this domain.

Experimental Setup

All experiments were conducted on an NVIDIA GeForce RTX 2080 Ti with CUDA compute capability 7.5, CUDA driver version 10.1, 4352 CUDA cores, and 11 GB GDDR6 memory. This section describes the algorithms and their variants that were used for benchmarking. For each algorithm, we implemented multiple variants with different optimization approaches, ranging from basic parallel implementations to more sophisticated techniques with memory optimizations and algorithmic enhancements. All implementation code is available at platform GitHub[1].

Experimental Setup: Bellman-Ford algorithm

For the Bellman-Ford algorithm evaluation, we used the DIMACS Road Networks Dataset[2], which provides real-world road network graphs suitable for single-source shortest path analysis. The dataset contains various road networks with different scales, allowing us to evaluate performance across graphs of varying sizes and densities. All Bellman-Ford variants use Compressed Sparse Row (CSR) format to efficiently handle large graphs in GPU global memory [2].

One thread per vertex: This implementation assigns one thread to each vertex, where each thread is responsible for relaxing all outgoing edges from its assigned vertex [1]. It uses two distance arrays: *previousDistance*

[1] CUDA Parallel Shortest Path: Implementation. Available at: https://github.com/deepbodra97/cuda-parallel-shortest-path, free. English lang. (accessed: 05.06.2025).

[2] 9th DIMACS Implementation Challenge: Shortest Paths. Available at: http://www.diag.uniroma1.it/~challenge9/download.shtml, free. English lang. (accessed: 03.06.2025).

stores costs from the previous iteration, while distance accumulates updated costs in the current iteration. This dual-array approach prevents threads from reading updated values within the same iteration to maintain algorithmic correctness. After each iteration, values are copied from distance to *previousDistance*. A challenge arises when multiple threads attempt to update the same destination vertex simultaneously, creating race conditions. We address this using CUDA *atomicMin* operation, which ensures thread-safe updates by atomically selecting the minimum value among competing writes[1]. The kernel is launched $|V| - 1$ times, with each launch handling one complete iteration of edge relaxations. The time complexity is $O((|V| - 1) \times (|E|/P + sync\_cost))$, where $P = \min(|V|, GPU\ cores)$, as each thread processes edges in parallel but synchronization is required between iterations. Space complexity remains $O(|V|^2 + |E|)$ for the all-pairs distance matrix and CSR graph representation.

Strided: The one thread per vertex implementation scalability is limited by the maximum number of threads a GPU device supports. The strided version overcomes this limitation using a grid-stride loop pattern [7], where each thread processes multiple vertices depending on the grid size and total vertex count. In this approach, a thread with ID *tid* processes vertices at positions *tid*, *tid* + *stride*, *tid* + + 2 × *stride*, and so forth, where stride equals *blockDimension* × *gridDimension*. This pattern allows the same kernel to handle graphs of arbitrary size by using fewer threads than vertices [1]. The optimal number of threads for a given graph can be determined experimentally, balancing resource utilization with memory bandwidth. Time complexity becomes $O((|V| - 1) \times (\lceil |V|/P \rceil \times \times avg\_degree + sync\_cost))$ where $avg\_degree = |E|/|V|$, as each thread now processes multiple vertices sequentially within each iteration. The load balancing factor ranges from $O(1)$ for uniform degree distribution to $O(max\_degreee/avg\_degree)$ for skewed distributions, maintaining the same $O(|V|^2 + |E|)$ space complexity.

Strided with flag [11]: Both previous implementations suffer from inefficient work distribution, as threads process vertices whose distances changed since the previous iteration. The flag-based optimization addresses this by tracking which vertices had distance updates in the previous iteration. We maintain a Boolean flag array of size $|V|$, where $flag[i]$ indicates whether vertex i--th distance changed in the previous iteration. During the distance update phase, if $previousDistance[i] > distance[i]$, we set $flag[i] = true$. In the subsequent iteration, threads only process outgoing edges from vertices with $flag[i] = true$, significantly reducing unnecessary computations [11]. The flag is reset to false when a vertex edges are processed, preparing for the next iteration. The time complexity per iteration i becomes $O(Active\_vertices(i) \times avg\_degree/P)$, where $Active\_vertices(i)$ represents vertices with updated distances. In the expected case, this yields $O(|E| \times H\_|V|/P)$ where $H\_|V|$ is the $|V|^{th}$ harmonic number. Space complexity remains $O(|V|^2 + |E|)$ plus $O(|V|)$ for the flag array.

[1] CUDA C++ Programming Guide. Available at: https://docs.nvidia.com/cuda/cuda-c-programming-guide/, free. English lang. (accessed: 02.06.2025).

Experimental Setup: Dijkstra algorithm

For Dijkstra algorithm evaluation, we used Stanford's Peer-to-peer network dataset [16], which is well-suited for evaluating all-pairs shortest path computations in moderately-sized graphs. The graph is represented using an adjacency matrix format for efficient neighbor lookups during shortest path computation.

Dijkstra algorithm presents parallelization challenges due to its inherently sequential nature of selecting the minimum unvisited vertex [1, 3, 17]. Instead of parallelizing the core algorithm logic, we implement a parallel all-pairs approach where each thread computes shortest paths from a different source vertex. Each thread runs an instance of Dijkstra algorithm using a different source vertex. Since GPU threads cannot efficiently maintain individual priority queues due to memory constraints, each thread uses a simple array-based approach to find the next minimum unvisited vertex [3]. The kernel assigns one thread per source vertex, with thread *src* computing shortest paths from vertex *src* to all other vertices. Each thread maintains its own visited array and distance array within the global distance matrix. This approach achieves parallelism across different source vertices while preserving the sequential correctness of individual Dijkstra computations. Our approach executes $|V|$ independent Dijkstra instances in parallel, yielding time complexity $O((|V|^2 + |E|) \times |V|/P)$ where each thread performs $O(|V|^2 + |E|)$ work for its assigned source. Space complexity is $O(|V|^2 + |V| \times P)$ for the distance matrix and per-thread visited arrays, with optimal performance when $P = |V|$.

Experimental Setup: Floyd-Warshall algorithm

For the Floyd-Warshall algorithm evaluation, we used Stanford's Peer-to-peer network dataset [16] and the graph is represented using an adjacency matrix format.

One thread per edge: This implementation provides maximum parallelism by assigning one thread to each edge in the graph [9]. Using a 2D thread grid, thread (i, j) handles the edge from vertex i to vertex j. In each of the iterations $|V|$, every thread checks whether using the current intermediate vertex k provides a shorter path. This approach achieves excellent memory coalescing as threads in the same warp access consecutive memory locations in the distance matrix [10]. However, it requires launching $|V|^2$ threads, limiting scalability for very large graphs due to GPU resource constraints. Time complexity is $O(|V|^3/P)$, where $P = \min(|V|^2, GPU\ cores)$, as $|V|^2$ threads perform $O(1)$ work per iteration across $|V|$ iterations. This achieves optimal parallelization when sufficient GPU cores are available. Space complexity remains $O(|V|^2)$ for the distance matrix with optimal memory coalescing patterns

One thread per edge with shared memory: This variant optimizes the one thread per edge approach by reducing global memory accesses through shared memory [6, 18] utilization. In iteration k, threads in the same row all access $distance[i][k]$, creating an opportunity for shared memory optimization. We use 1D thread blocks where the first thread in each block loads $distance[i][k]$ into shared memory, making it available to all threads in the block through shared memory broadcast [6]. This reduces global memory bandwidth requirements, though synchronization overhead can limit performance gains. Time complexity

remains $O(|V|^3/P)$ but with reduced memory access latency due to shared memory utilization. Space complexity includes additional O (*shared_memory_per_block*) for cached data, though synchronization overhead may offset some performance benefits for smaller graphs.

One thread per vertex implementation: To improve scalability, the one thread per vertex implementation reduces thread count by assigning one thread per vertex rather than per edge. Each thread handles one row of the distance matrix, iterating through all potential destinations for its assigned source vertex. Thread i processes all edges from vertex i, checking each destination j to determine if routing through intermediate vertex k offers improvement. This approach requires fewer threads while maintaining reasonable parallelism, making it suitable for larger graphs. Time complexity becomes $O(|V|^3/P)$, where $P = \min(|V|,$ *GPU cores*), but each thread now performs $O(|V|^2)$ work across all iterations. This provides better scalability for large graphs by reducing thread count requirements while maintaining the same $O(|V|^2)$ space complexity for the distance matrix.

Tiled implementation: This implementation addresses memory bandwidth limitations by dividing the adjacency matrix into 2D tiles of size *TILE_DIMENSION* × *TILE_DIMENSION*. At each iteration, tiles are processed in three phases. In phase 1, the primary tile (diagonal tile containing intermediate vertex) is processed using a single thread block, where all paths within this tile are updated using vertices from the same tile as intermediates. In phase 2, tiles sharing the same row or column as the primary tile are processed, where these tiles use vertices from the primary tile as intermediates, requiring data from both the primary tile and the current tile. In phase 3, the remaining tiles are processed using precomputed results from primary row and column tiles. This phase achieves maximum parallelism as all remaining tiles can be processed independently [5, 19, 20]. This approach ensures that each global memory location is accessed exactly once per iteration, significantly improving memory efficiency compared to the basic implementations. The time complexity becomes $O(|V|^2 + |V| \times$ *TILE_DIMENSION*) through the three-phase approach: primary tile $O(|V|)$, border tiles $O(|V|)$, and interior tiles O (*TILE_DIMENSION*) with full parallelization [5]. Space complexity remains $O(|V|^2)$ with optimal tile size *TILE_DIMENSION* = $\sqrt{}$(*GPU cores*) minimizing total execution time.

Tiled with shared memory: This implementation enhances the basic tiled approach by utilizing shared memory [6] to cache frequently accessed data within each thread block. In phase 1, the primary tile data is loaded into shared memory once and reused for all computations within the tile, eliminating redundant global memory accesses. During phase 2, each thread block loads the relevant portion of the primary tile into shared memory alongside its tile data, enabling fast access to intermediate vertex information. In phase 3, thread blocks load data from their corresponding row and column tiles into shared memory, significantly reducing global memory bandwidth requirements as multiple threads access the same cached values. Time complexity improves to $O(|V|^2 + |V| \times$ $\times$ *TILE_DIMENSION*/*memory_speedup*), where *memory_speedup* ≈ 10–100 times from shared memory utilization [6]. Space complexity includes $O(|V|^2 + \min($*num_blocks* × *TILE_DIMENSION*^{2}, *total_shared_memory*$))$ for the distance matrix plus shared memory usage, representing the most optimized implementation combining algorithmic restructuring with memory hierarchy optimization.

Results and Analysis

The Bellman-Ford algorithm experiments were conducted on the DIMACS Road Networks dataset[1], with performance measured across different graph sizes. Fig. 1 shows the execution times for the three implementation variants across various datasets. The one thread per vertex implementation serves as the baseline, providing straightforward parallelization but facing scalability limitations for larger graphs.

The strided implementation shows mixed performance results: it runs slower than the baseline for small graphs due to the overhead of grid-stride loops [7], but demonstrates better scalability for larger graphs where the naive version becomes resource-constrained.

The strided with flag optimization delivers the most significant performance improvements, achieving approximately 2.8 times faster execution compared to the one thread per vertex implementation. This optimization proves highly effective because it eliminates unnecessary work by processing only vertices whose distances changed in the previous iteration [11]. For the largest dataset (Eastern USA with 3.6 million vertices and 8.8 million edges), the strided with flag variant completed in 409 s compared to 1137 s for the baseline implementation. The performance characteristics reveal that the flag optimization becomes increasingly beneficial as graph size grows, since larger graphs tend to have more vertices with unchanged distances in later iterations of the algorithm.

Dijkstra algorithm evaluation used Stanford's Peer-to-peer network dataset [16]. Table shows the performance comparison between CPU and GPU implementations. The parallel all-pairs GPU implementation achieved significant speedup over the serial CPU version, completing the p2p-Gnutella04 dataset with 11K vertices and 40K edges in approximately 120 s compared to 24 minutes for the CPU implementation. However, the inherently sequential nature of Dijkstra algorithm limits the parallelization benefits compared to the other algorithms studied [1, 3].

The GPU implementation performance is constrained by the need for each thread to maintain its distance and visited arrays, along with the sequential process of finding minimum unvisited vertices within each thread computation. Despite these limitations, the GPU version still provides meaningful acceleration for all-pairs shortest path computation.

The Floyd-Warshall algorithm experiments demonstrate the most performance improvements among the three algorithms studied. Fig. 2 illustrates the execution times across different implementation variants and dataset sizes.

[1] 9th DIMACS Implementation Challenge: Shortest Paths. Available at: http://www.diag.uniroma1.it/~challenge9/download.shtml, free. English lang. (accessed: 03.06.2025).

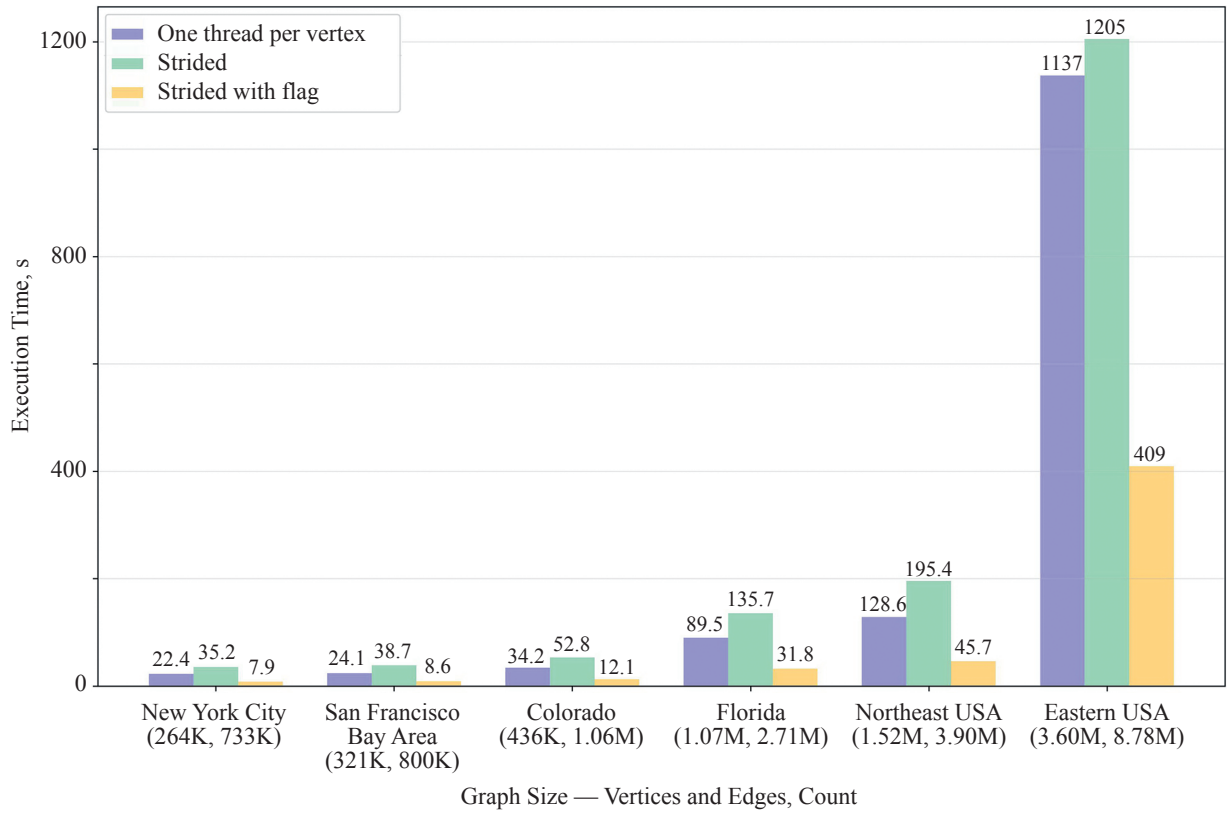

Fig. 1. Execution time comparison for Bellman-Ford variants on DIMACS Road Networks Dataset

Table. Execution time comparison for Dijkstra on Stanford's Peer-to-peer network dataset

Dataset	Number of Vertices	Number of Edges	CPU	GPU
p2p-Gnutella04	10,876	39,994	24 mins	119.941 s

The one thread per edge implementation achieves excellent performance for small to medium graphs due to maximum parallelization and good memory coalescing[1]. However, its scalability is limited by the requirement for $|V|^2$ threads. The variant of one thread per edge implementation with shared memory shows some improvement through reduced global memory accesses [6], though synchronization overhead can limit gains.

The one thread per vertex implementation provides better scalability by reducing thread count, making it suitable for larger graphs while maintaining reasonable performance. However, the tiled implementations show the most significant improvements [5].

The tiled implementation using global memory outperforms all basic variants, running approximately 4–6 times faster than the one thread per vertex version. The tiled implementation with shared memory delivers the best overall performance, achieving roughly 8 times speedup over the baseline and approximately 2 times improvement over the global memory tiled version [6].

For the p2p-Gnutella04 dataset, the tiled implementation with shared memory completed in 2.38 s compared to over 30 s for the one thread per vertex variant. This improvement demonstrates the effectiveness of combining algorithmic restructuring (tiling) with memory hierarchy optimization (shared memory) [5, 6].

When comparing across algorithms, Floyd-Warshall shows the greatest potential for GPU acceleration due to its inherently parallel structure and regular memory access patterns [17]. The algorithm $O(|V|^3)$ complexity makes GPU acceleration particularly valuable for reducing computation time.

Bellman-Ford demonstrates good parallelization potential, especially with the flag optimization that reduces unnecessary work [11]. The algorithm iterative nature and edge-based parallelism translate well to GPU architectures [1, 2].

Dijkstra algorithm shows the most limited parallelization benefits due to its inherently sequential vertex selection process [1, 3]. However, the all-pairs parallel approach still provides meaningful acceleration over serial CPU implementations.

The results highlight the importance of algorithm-specific optimizations: flag-based early termination for Bellman-Ford [11], tiled computation with shared memory for Floyd-Warshall [5, 6], and parallel source processing for Dijkstra. Memory access patterns and work distribution significantly impact GPU performance, with regular access

[1] CUDA C++ Programming Guide. Available at: https://docs.nvidia.com/cuda/cuda-c-programming-guide/, free. English lang. (accessed: 02.06.2025).

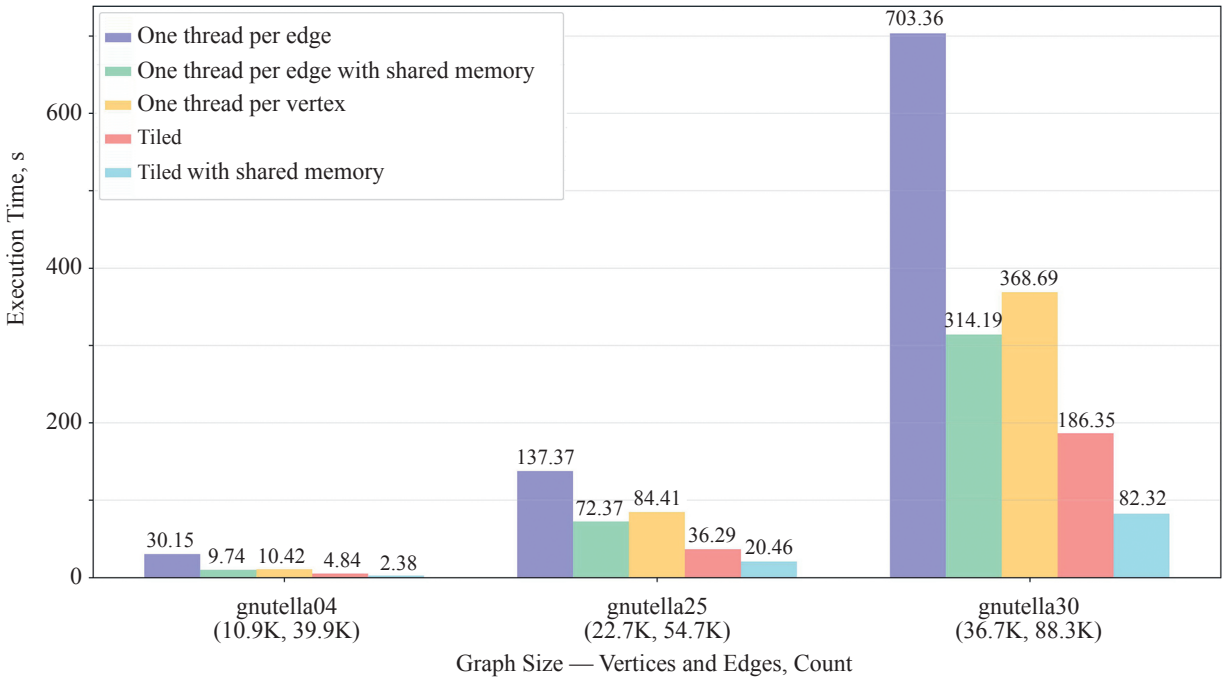

Fig. 2. Execution time comparison for Floyd-Warshall on Stanford's Peer-to-peer network dataset

patterns and balanced workloads yielding the best results [1, 2, 21, 22][1].

A common pattern is that all GPU implementations, even on the largest datasets, outperform the corresponding serial CPU versions on the smallest datasets used for experimenting. This demonstrates the computational advantage that GPU parallelization provides to enable the processing of much larger problem instances in less time than traditional approaches require for smaller problems.

Future Work

Several directions emerge for extending this research, including investigating advanced graph representations, such as ELL format to address control divergence issues [7], developing multi-GPU implementations for very large graphs [8], and adapting algorithms for dynamic graphs where edges are modified during computation. Additional opportunities include implementing memory-efficient techniques like staged loading for graphs exceeding GPU capacity, exploring hybrid CPU-GPU approaches to optimize resource utilization, and developing application-specific optimizations tailored to domains such as transportation or social networks. Comparative analysis with alternative parallel platforms such as OpenCL or distributed computing frameworks would provide broader insights into parallel shortest path computation trade-offs across different architectures.

Conclusion

This paper presented an evaluation of GPU implementations for three fundamental shortest path algorithms: Bellman-Ford, Dijkstra, and Floyd-Warshall using Compute Unified Device Architecture. Through systematic implementation and optimization of multiple variants, we demonstrated significant performance benefits of GPU acceleration for graph processing. Floyd-Warshall achieved the most dramatic improvements with the tiled shared memory implementation delivering approximately 8 times speedup, while Bellman-Ford showed substantial acceleration through flag-based optimization achieving 2.8 times performance improvement. Although Dijkstra algorithm exhibited more limited parallelization benefits due to its sequential nature, it still provided meaningful acceleration over CPU implementations. The results reveal that GPU implementations on large datasets consistently outperform CPU versions on small datasets, highlighting the transformative potential of parallel computing for shortest path problems and enabling practical processing of real-world graph sizes previously computationally prohibitive.

[1] CUDA C++ Programming Guide. Available at: https://docs.nvidia.com/cuda/cuda-c-programming-guide/, free. English lang. (accessed: 02.06.2025).

References

1. Harish P., Narayanan P.J. Accelerating large graph algorithms on the GPU using CUDA. *Lecture Notes in Computer Science*, 2007, vol. 4873, pp. 197–208. https://doi.org/10.1007/978-3-540-77220-0_21
2. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. *Introduction to Algorithms*. MIT press, 2009. 1292 p.
3. Katz G.J., Kider J.T. All-pairs shortest-paths for large graphs on the GPU. *Proc. of the 23rd ACM SIGGRAPH/EUROGRAPHICS symposium on Graphics hardware*, 2008, pp. 47–55.
4. Lebedev S.S., Novikov F.A. The Necessary and sufficient condition for Dijkstra's algorithm applicability. *Computer Tools in Education*, 2017, no. 4, pp. 5–13. (in Russian)
5. Lund B., Smith J.W. A multi-stage cuda kernel for floyd-warshall. *arXiv*, 2010, arXiv:1001.4108. https://doi.org/10.48550/arXiv.1001.4108
6. Winkler D., Meister M., Rezavand M., Rauch W. gpuSPHASE—A shared memory caching implementation for 2D SPH using CUDA. *Computer Physics Communications*, 2017, vol. 213, pp. 165–180. https://doi.org/10.1016/j.cpc.2016.11.011
7. Harris M. CUDA *Pro Tip: write flexible kernels with grid-stride loops*. Available at: https://developer.nvidia.com/blog/cuda-pro-tip-write-flexible-kernels-grid-stride-loops. (accessed: 30.05.2025)
8. Yang S., Liu X., Wang Y., He X., Tan G. Fast All-Pairs Shortest Paths algorithm in large sparse graph. *Proc. of the 37th International Conference on Supercomputing*, 2023, pp. 277–288. https://doi.org/10.1145/3577193.3593728
9. Tang W., Chen T., Armstrong M.P. GPU-accelerated parallel all-pair shortest path routing within stochastic road networks. *International Journal of Geographical Information Science*, 2025, vol. 39, no. 1, pp. 53–85. https://doi.org/10.1080/13658816.2024.2394651
10. Spridon D.E., Deaconu A.M., Tayyebi J. Novel GPU-based method for the generalized maximum flow problem. *Computation*, 2025, vol. 13, no. 2, pp. 40. https://doi.org/10.3390/computation13020040
11. Agarwal P., Dutta M. New approach of Bellman Ford algorithm on GPU using compute unified design architecture (CUDA). *International Journal of Computer Applications*, 2015, vol. 110, no. 13, pp. 1–5. https://doi.org/10.5120/19375-1027
12. Song B. High-performance parallelization of Dijkstra's algorithm using MPI and CUDA. *arXiv*, 2025, arXiv:2504.03667. https://doi.org/10.48550/arXiv.2504.03667
13. Bengtsson M., Wittsten J., Waidringer J. Warehouse storage and retrieval optimization via clustering, dynamic systems modeling, and GPU-accelerated routing. *arXiv*, 2025, arXiv:2504.20655. https://doi.org/10.48550/arXiv.2504.20655
14. Kumar H.S., Singh A., Ojha M.K. Artificial intelligence based navigation in quasi structured environment. *arXiv*, 2024, arXiv:2407.17508. https://doi.org/10.48550/arXiv.2407.17508
15. Morgan N., Yenusah C., Diaz A., Dunning D., Moore J., Heilman E., et al. On a simplified approach to achieve parallel performance and portability across CPU and GPU architectures. *Information*, 2024, vol. 15, no. 11, pp. 673. https://doi.org/10.3390/info15110673
16. Leskovec J., Krevl A. *SNAP Datasets: Stanford large network dataset collection*. 2014. Available at: http://snap.stanford.edu/data
17. Buluç A., Gilbert J.R., Budak C. Solving path problems on the GPU. *Parallel Computing*, 2010, vol. 36, no. 5-6, pp. 241–253. https://doi.org/10.1016/j.parco.2009.12.002
18. Merrill D., Garland M., Grimshaw A. Scalable GPU graph traversal. *ACM SIGPLAN Notices*, 2012, vol. 47, no. 8, pp. 117–128. https://doi.org/10.1145/2370036.2145832
19. Kirk D.B., Hwu W.M.W. *Programming Massively Parallel Processors: a Hands-on Approach*. Morgan Kaufmann, 2016, 576 p.
20. Nickolls J., Buck I., Garland M., Skadron K. Scalable parallel programming with CUDA. *Queue*, 2008, vol. 6, no. 2, pp. 40–53. https://doi.org/10.1145/1365490.1365500
21. Bodra D., Khairnar S. Comparative performance analysis of modern NoSQL data technologies: Redis, Aerospike, and Dragonfly. *Journal of Research, Innovation and Technologies*, 2025, vol. 4, no. 2, pp. 193–200. https://doi.org/10.57017/jorit.v4.2(8).05
22. Khairnar S., Bodra D. Recommendation engine for Amazon magazine subscriptions. *International Journal of Advanced Computer Science and Applications*, 2025, vol. 16, no. 7, pp. 1–8. https://doi.org/10.14569/ijacsa.2025.0160796

Литература

1. Harish P., Narayanan P.J. Accelerating large graph algorithms on the GPU using CUDA // Lecture Notes in Computer Science. 2007. V. 4873. P. 197–208. https://doi.org/10.1007/978-3-540-77220-0_21
2. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. Introduction to Algorithms. MIT press, 2009. 1292 p.
3. Katz G.J., Kider J.T. All-pairs shortest-paths for large graphs on the GPU // Proc. of the 23rd ACM SIGGRAPH/EUROGRAPHICS symposium on Graphics hardware. 2008. P. 47–55.
4. Лебедев С.С., Новиков Ф.А. Необходимое и достаточное условие применимости алгоритма Дейкстры // Компьютерные инструменты в образовании. 2017. № 4. С. 5–13.
5. Lund B., Smith J.W. A multi-stage cuda kernel for floyd-warshall // arXiv. 2010. arXiv:1001.4108. https://doi.org/10.48550/arXiv.1001.4108
6. Winkler D., Meister M., Rezavand M., Rauch W. gpuSPHASE—A shared memory caching implementation for 2D SPH using CUDA // Computer Physics Communications. 2017. V. 213. P. 165–180. https://doi.org/10.1016/j.cpc.2016.11.011
7. Harris M. CUDA Pro Tip: write flexible kernels with grid-stride loops. URL: https://developer.nvidia.com/blog/cuda-pro-tip-write-flexible-kernels-grid-stride-loops. (accessed: 30.05.2025)
8. Yang S., Liu X., Wang Y., He X., Tan G. Fast All-Pairs Shortest Paths algorithm in large sparse graph // Proc. of the 37th International Conference on Supercomputing. 2023. P. 277–288. https://doi.org/10.1145/3577193.3593728
9. Tang W., Chen T., Armstrong M.P. GPU-accelerated parallel all-pair shortest path routing within stochastic road networks // International Journal of Geographical Information Science. 2025. V. 39. N 1. P. 53–85. https://doi.org/10.1080/13658816.2024.2394651
10. Spridon D.E., Deaconu A.M., Tayyebi J. Novel GPU-based method for the generalized maximum flow problem // Computation. 2025. V. 13. N 2. P. 40. https://doi.org/10.3390/computation13020040
11. Agarwal P., Dutta M. New approach of Bellman Ford algorithm on GPU using compute unified design architecture (CUDA) // International Journal of Computer Applications. 2015. V. 110. N 13. P. 1–5. https://doi.org/10.5120/19375-1027
12. Song B. High-performance parallelization of Dijkstra's algorithm using MPI and CUDA // arXiv. 2025. arXiv:2504.03667. https://doi.org/10.48550/arXiv.2504.03667
13. Bengtsson M., Wittsten J., Waidringer J. Warehouse storage and retrieval optimization via clustering, dynamic systems modeling, and GPU-accelerated routing // arXiv. 2025. arXiv:2504.20655. https://doi.org/10.48550/arXiv.2504.20655
14. Kumar H.S., Singh A., Ojha M.K. Artificial intelligence based navigation in quasi structured environment // arXiv. 2024. arXiv:2407.17508. https://doi.org/10.48550/arXiv.2407.17508
15. Morgan N., Yenusah C., Diaz A., Dunning D., Moore J., Heilman E., et al. On a simplified approach to achieve parallel performance and portability across CPU and GPU architectures // Information. 2024. V. 15. N 11. P. 673. https://doi.org/10.3390/info15110673
16. Leskovec J., Krevl A. SNAP Datasets: Stanford large network dataset collection. 2014. URL: http://snap.stanford.edu/data
17. Buluç A., Gilbert J.R., Budak C. Solving path problems on the GPU // Parallel Computing. 2010. V.36. N 5-6. P. 241–253. https://doi.org/10.1016/j.parco.2009.12.002
18. Merrill D., Garland M., Grimshaw A. Scalable GPU graph traversal // ACM SIGPLAN Notices. 2012. V. 47. N 8. P. 117–128. https://doi.org/10.1145/2370036.2145832
19. Kirk D.B., Hwu W.M.W. Programming Massively Parallel Processors: a Hands-on Approach. Morgan Kaufmann, 2016. 576 p.
20. Nickolls J., Buck I., Garland M., Skadron K. Scalable parallel programming with CUDA // Queue. 2008. V. 6. N 2. P. 40–53. https://doi.org/10.1145/1365490.1365500
21. Bodra D., Khairnar S. Comparative performance analysis of modern NoSQL data technologies: Redis, Aerospike, and Dragonfly // Journal of Research, Innovation and Technologies. 2025. V. 4. N 2. P. 193–200. https://doi.org/10.57017/jorit.v4.2(8).05
22. Khairnar S., Bodra D. Recommendation engine for Amazon magazine subscriptions // International Journal of Advanced Computer Science and Applications. 2025. V. 16. N 7. P. 1–8. https://doi.org/10.14569/ijacsa.2025.0160796

Authors

Deep Bodra — Magister, Student, Harrisburg University of Science and Technology, Harrisburg,17101, USA, SC 57216618940, https://orcid.org/0009-0009-4173-2447, Deepbodra97@gmail.com
Sushil Khairnar — Magister, Student, Virginia Tech, Virginia, 24061, USA, SC 57204777066, https://orcid.org/0009-0006-5192-0175, sushilk@vt.edu

Авторы

Бодра Дип — магистр, студент, Гаррисбергский университет науки и технологий, Гаррисберг, 17101, США, SC 57216618940, https://orcid.org/0009-0009-4173-2447, Deepbodra97@gmail.com
Хайрнар Сушил — магистр, студент, Технологический институт Вирджинии, Вирджиния, 24061, США, SC 57204777066, https://orcid.org/0009-0006-5192-0175, sushilk@vt.edu

Received 20.06.2025
Approved after reviewing 27.08.2025
Accepted 22.09.2025

Статья поступила в редакцию 20.06.2025
Одобрена после рецензирования 27.08.2025
Принята к печати 22.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-876-887

Anomaly detection for IIoT: analyzing Edge-IIoTset dataset with varied class distributions

Wafaa Ferhi[1]✉, Djilali Moussaoui[2], Mourad Hadjila[3], Al Baraa Bouidaine[4]

[1,2,3,4] University of Abu Bekr Belkaid, Tlemcen, 13000, Algeria

[1] wafaa.ferhi@univ-tlemcen.dz✉, https://orcid.org/0009-0005-7574-8368
[2] djilali.moussaoui@univ-tlemcen.dz, https://orcid.org/0000-0003-3478-263X
[3] mourad.hadjila@univ-tlemcen.dz https://orcid.org/0000-0002-6554-3925
[4] albaraa.bouidaine@univ-tlemcen.dz, https://orcid.org/0009-0005-2204-9117

Abstract

In the context of the Industrial Internet of Things (IIoT), cybersecurity refers to preventing unauthorized access, attacks, and vulnerabilities to interconnected devices, networks, and data. Given the inherent interconnectedness of IIoT devices, ensuring security is of paramount importance to mitigate potential disruptions, data breaches, and malicious activities. As IIoT systems continue to proliferate, the significance of robust security measures, effective intrusion detection, and intelligent detection techniques escalates to safeguard critical infrastructure and sensitive data from cyber threats. This work aims to contribute towards establishing a secure and resilient industrial environment through the utilization of a hybrid model: Convolutional Neural Network with Deep Neural Network, accommodating distinct class distributions. The recent "Edge IIoTset" dataset is harnessed to enhance the model efficacy. Throughout the evaluation process, diverse metrics are employed, encompassing Accuracy, Precision, Recall, and the F1-score. By applying thorough preprocessing and using various class distribution scenarios (2, 6, 9, 10, and 15 classes), the model achieved excellent classification results. Notably, the 9-class configuration reached an Accuracy of 99.13 %, while the 6-class and 10-class setups also delivered strong performance at 97.13 % and 96.11 %, respectively. Our architecture effectively combines feature extraction and deep classification layers, resulting in a robust solution adaptable to complex IIoT traffic.

Keywords

anomaly, convolutional neural network, deep neural network, Edge IIoTset dataset, Industrial Internet of Things, intelligent detection, metrics, security

For citation: Ferhi W., Moussaoui D., Hadjila M., Bouidaine A.B. Anomaly detection for IIoT: analyzing Edge-IIoTset dataset with varied class distributions. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 876–887. doi: 10.17586/2226-1494-2025-25-5-876-887

УДК 004.056.5

Обнаружение аномалий для IIoT: анализ набора данных Edge-IIoTset с различными распределениями классов

Вафаа Ферхи[1]✉, Джилали Муссауи[2], Мурад Хаджила[3], Аль Бараа Буиден[4]

[1,2,3,4] Университет Абу Бекра Белкаида, Тлемсен, 13000, Алжир

[1] wafaa.ferhi@univ-tlemcen.dz✉, https://orcid.org/0009-0005-7574-8368
[2] djilali.moussaoui@univ-tlemcen.dz, https://orcid.org/0000-0003-3478-263X
[3] mourad.hadjila@univ-tlemcen.dz https://orcid.org/0000-0002-6554-3925
[4] albaraa.bouidaine@univ-tlemcen.dz, https://orcid.org/0009-0005-2204-9117

Аннотация

Кибербезопасность промышленного интернета вещей (Industrial Internet of Things, IIoT) означает предотвращение несанкционированного доступа, атак и уязвимостей взаимосвязанных устройств, сетей и данных. Учитывая внутреннюю взаимосвязь устройств IIoT, обеспечение безопасности имеет первостепенное значение для предотвращения потенциальных сбоев, утечек данных и вредоносных действий. По мере распространения

систем IIoT возрастает важность надежных мер безопасности, эффективного обнаружения вторжений и интеллектуальных методов обнаружения для защиты критически важной инфраструктуры и конфиденциальных данных от киберугроз. В данной работе исследованы вопросы создания безопасной и устойчивой промышленной среды посредством использования гибридной модели: сверточной нейронной сети и глубокой нейронной сети, учитывающей различные распределения классов. Для повышения эффективности модели применен набор данных Edge IIoTset. В процессе оценки использованы различные метрики, включая Accuracy, Precision, Recal и F1-меру. Благодаря тщательной предварительной обработке и использованию различных сценариев распределения классов (2, 6, 9, 10 и 15 классов) модель показала хорошие результаты классификации. Конфигурация с 9 классами достигла точности 99,13 %, в то время как конфигурации с 6 и 10 классами — 97,13 % и 96,11 % соответственно. Предложенная архитектура эффективно сочетает уровни извлечения признаков и глубокой классификации, что приводит к созданию надежного решения, адаптируемого к сложному трафику IIoT.

Ключевые слова

аномалия, сверточная нейронная сеть, глубокая нейронная сеть, набор данных Edge IIoTset, промышленный интернет вещей, интеллектуальное обнаружение, метрики, безопасность

Ссылка для цитирования: Ферхи В., Муссауи Д., Хаджила М., Буиден А.Б. Обнаружение аномалий для IIoT: анализ набора данных Edge-IIoTset с различными распределениями классов // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 876–887 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-876-887

Introduction

The Internet of Things (IoT) is described as the interconnection of multiple devices that use unique identifiers to share data and other pertinent information across a network without the assistance of individuals [1]. Using sensing devices, any physical device can be simply operated, minimizing the need for human labor [2]. IoT applications have been deployed in virtually all fields, ranging from healthcare and agriculture to transportation and manufacturing. These applications have revolutionized and transformed industries by connecting devices, collecting data, and enabling intelligent decision-making processes [3]. Furthermore, as the industrial world progresses towards more advanced and complicated systems, the need for Industrial IoT (IIoT) has emerged. IIoT takes the principles of IoT and applies them specifically to industrial processes, enabling remote monitoring, intelligent analytics, and control of industrial operations. It introduces a higher level of automation, scalability, and efficiency, addressing the unique challenges and requirements of the manufacturing sector. With IIoT, industries can optimize production, improve resource utilization, and enhance overall operational performance [4]. IIoT, when integrated with Cyber-Physical Systems (CPS), brings a transformative shift to industrial operations. CPS is a system that integrates the physical and virtual worlds, fostering connectivity between them [5], which encompass a transformative realm where the physical and cyber worlds intricately interlace, imbuing the operational landscape with heightened intelligence and efficiency [6]. They use sensors and actuators to gather data from the physical world and software to analyze and act on that data, promoting seamless connectivity [7]. The integration of IIoT and CPS enhances connectivity among smart devices [8]. The rapid expansion of IIoT has led to a surge in connected devices, significantly increasing data generation [9]. This presents challenges in data security and anomaly detection, making the confidentiality, integrity, and availability of IIoT data crucial for protecting critical infrastructure [10]. Anomaly detection is essential in identifying security vulnerabilities or inefficiencies [11]. Artificial Intelligence and Deep Learning (DL) enable real-time anomaly detection in data traffic, device behavior, and system performance, allowing proactive threat mitigation [12]. These technologies strengthen cybersecurity defenses against malware, ransomware, and phishing, ensuring data integrity and system security. The main contributions of our research are outlined below:

— introduction of a novel combined DL model, integrating Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN), which demonstrates enhanced performance;
— employing a contemporary dataset known as Edge-IIoTset to facilitate the training and evaluation of the proposed MC-CNN-DNN (Multiclassification CNN-DNN) model. Diverse multiclass distributions are introduced and analyzed;
— the evaluation of our model performance incorporates several metrics, including Accuracy and Precision.

Related work

In recent studies, researchers have presented various approaches to addressing cybersecurity vulnerabilities and breaches in IIoT environments. In [13], authors present a DL-based intrusion detection model combining CNNs for spatial feature extraction and Long Short-Term Memory (LSTM) for temporal feature extraction (Network Intrusion Detection System (NIDS))-CNN-LSTM). Tested on datasets like KDD CUP99, NSL KDD, and UNSW NB15, the model showed robust Accuracy and performance in binary and multi-classification tasks. Similarly, in [14], another group of scientists propose a DL framework leveraging CNNs, Recurrent Neural Networks DNNs, and Generative Adversarial Networks for cyber threat detection in IoT-driven IIoT networks, achieving 95 %–97 % Accuracy on intrusion datasets. The study conducted by [15] examined seven Machine Learning (ML) classifiers on the CICIDS2017 dataset, with K-Nearest Neighbors outperforming others in Precision, Recall, Accuracy, and F1-score. In a related work [16], which uses the same dataset as [15], a combination of ML algorithms and Principal Component Analysis techniques for Distributed Denial of Service (DDoS) detection using the CICIDS2017 and CSE-CIC-IDS 2018 datasets, showing

superior results [17]. In this paper, a novel approach is presented in which the authors develop a DLmodel using DNN and Decision Trees to handle unbalanced ICS datasets, improving attack detection. Another study [18] introduces a Nonsymmetric Deep AutoEncoder for unsupervised feature learning in intrusion detection which is specifically designed for unsupervised feature learning. In [19] the researchers present a groundbreaking anomaly-based intrusion detection model that uses a CNN to build both binary and multi-class classification models [20, 21]. Refer to numerous interesting surveys that deal with ML and DL techniques for Intrusion Detection Systems (IDS). These surveys examine publicly available intrusion datasets used in recent IDS to reveal present-day challenges and future directions. The review published in [22] focuses on several advancements IDS datasets, specifically from CSE-CIC-IDS-2017 to CSE-CIC-IDS-2018. This update includes the addition of new attack categories. The review study discussed in [23] explores and analyses intrusion detection and prevention methods specifically aimed at mitigating DDoS attacks. The study delves into the classification of IDS and explores different anomaly detection approaches. Similarly, and in the same context, the researchers in [24] are using the Difficult Set Sampling Technique (DSSTE) algorithm. The purpose of DSSTE is to improve the learning of unbalanced network data in a classification model by increasing the number of minority samples to be learned. DSSTE aims to address the problem of unbalanced network traffic and improve the classification Accuracy for the minority class. In [25], the researchers will thoroughly analyze and provide solutions to the problems arising from dataset imbalance in both the training and inference phases.

Background Framework of the Study

DNN

DL, a subset of ML, utilizes artificial neural networks to learn complex patterns from data [26]. A neural network consists of an input layer, an output layer, and one or more hidden layers. The perceptron, the fundamental unit of neural networks, processes multiple inputs by applying weights, summing them with a bias term, and passing the result through an activation function [27]. Mathematically, this is expressed as:

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b,$$

$$\begin{cases} f_1 = w_{11}x_1 + w_{12}x_2 + b_1 \\ f_2 = w_{21}x_1 + w_{22}x_2 + b_2 \\ f_3 = w_{31}x_1 + w_{32}x_2 + b_3 \end{cases},$$

where $x_1, x_2, \ldots x_n$, are inputs; $w_1, w_2, \ldots w_n$, are weights, and b is the bias term. Early DNNs were structured as multilayer perceptrons, where each perceptron computed outputs based on weighted inputs [28, 29]. In the case of three connected perceptrons, such as illustrated in Fig. 1, *a*, the first two perceptrons receive inputs w and x_2, perform calculations based on their respective parameters, and generate outputs y_1 and y_2. These outputs are then passed to the third perceptron, which further performs calculations to produce the final output y_3. Modern DNNs use advanced training techniques like backpropagation (Fig. 1, *b*) which optimizes learning by adjusting weights and biases efficiently.

CNN

In the evolution of neural networks, significant advancements were made with the introduction of multilayer perceptron variants and CNNs. In 1989, the concept of multilayer perceptron models emerged, marking a key milestone in neural network development. However, it was Yann LeCun who revolutionized the field by inventing the first CNNs [29]. LeCun's CNNs were inspired by the organization and functionality of the visual cortex in animals [30]. These networks were specifically designed to learn and process spatial hierarchies of features in an automated and adaptive manner. CNNs are a mathematical framework that typically consists of three fundamental layer types: convolutional layers, pooling layers, and fully connected layers [31]. Convolutional layers extract important features from input data using learnable filters. Pooling layers reduce the spatial dimensions of feature maps through down-sampling, improving computational efficiency and translation invariance. Fully connected layers analyze extracted features to make final predictions,

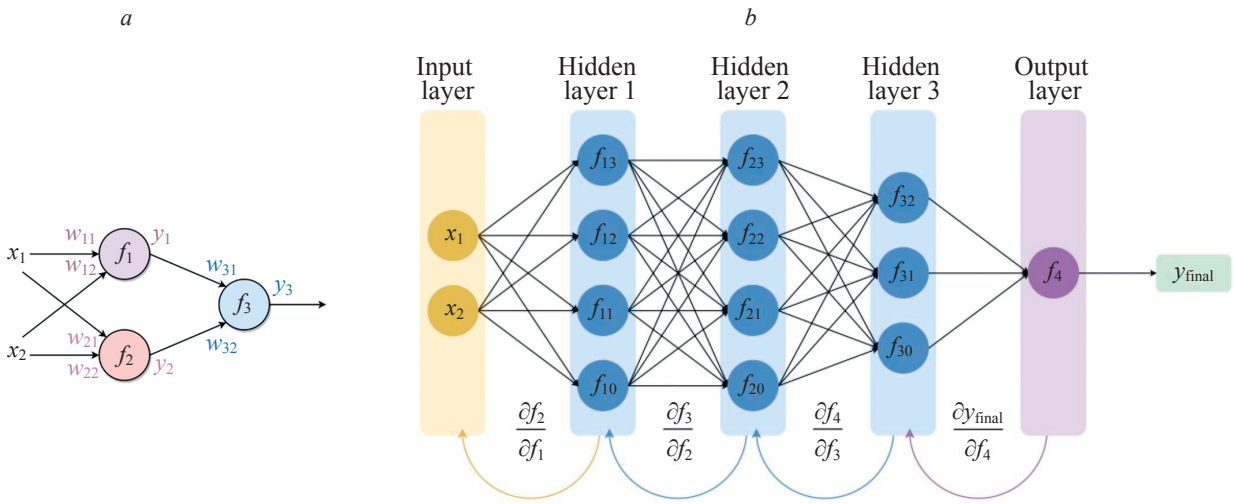

Fig. 1. DNN structure and learning mechanism: multilayer perceptron model (*a*); backpropagation process (*b*)

connecting all neurons from the previous layer to capture complex patterns. By combining these layers in a sequential manner and adjusting their parameters through processes like backpropagation and gradient descent, CNNs can learn complex patterns and make predictions on various tasks.

Evaluation Metrics

The evaluation of previous algorithms used for securing IIoT often involves employing various performance measures. These measures, including Accuracy, Precision, Recall, F1-score, true positive rate, false alarm rate, false positive rate, receiver operating characteristic curve, and area under the curve, are commonly utilized for assessing their effectiveness.

Model proposed

Dataset

The choice of dataset is critical for anomaly detection algorithms. This study makes use of the "Edge-IIoTset" dataset [32]. The dataset was created leveraging a purpose built IoT/IIoT testbed that includes a wide range of devices. The dataset contains data on 14 attacks related to IoT and IIoT connectivity protocols which are classified into five threat categories: DoS/DDoS attacks, information gathering, man-in-the-middle attacks, injection attacks, and malware assaults. It also includes features sourced from several sources, such as alarms, system resources, logs, and network traffic. The "Edge-IIoTset" dataset contains 61 features with two target variables: 'Attack label' for binary classification and 'Attack type' for multiclass classification. The 'Attack label' is explicitly designed for binary classification tasks, aiming to differentiate between two classes: "Attack" and "Normal". The target variable 'Attack label' assigns a binary label of 1 to instances representing attacks, and a label of 0 to instances representing normal traffic. On the other hand, the 'Attack type' target variable is intended for multiclass classification, enabling the categorization of instances into Different Attack (DA) types and normal traffic. Table 1 presents a summary of the instances of different IoT traffic types observed in the "Edge-IIoTset" dataset.

Table 1. Edge-IIoTset dataset Type Instances 'Attack type'

Class	Traffic Type	Instances
Normal	NORMAL	1,615,643
Attack	DDoS UDP	121,568
	DDoS ICMP	116,436
	SQL injection	51,203
	Password	50,153
	Vulnerability scanner	50,110
	DDoS TCP	50,062
	DDoS HTTP	49,911
	Uploading	37,634
	Backdoor	24,862
	Port Scanning	22,564
	XSS	15,915
	Ransomware	10,925
	MITM	1,214
	Fingerprinting	1,001

Experimental Approaches

The proposed model in this work is a composite algorithm consisting of a CNNs followed by a complex DNN. The design of the hybrid CNN-DNN model was motivated by the complementary strengths of both architectures. CNN layers are well-suited for capturing spatial and temporal patterns in sequential feature representations, while DNN layers are effective in combining those features through deep nonlinear transformations for robust classification. Initial tests with DNN-only models showed a tendency toward overfitting and limited generalization. Conversely, CNN-only models struggled to differentiate between closely related attack classes. The hybrid configuration achieved a balance between rich feature extraction and accurate classification resulting in superior performance across multiple class distributions. These results validate the architectural synergy of the CNN-DNN combination and reflect the trade-offs considered during model development.

However, before implementing the model, a preprocessing step is performed on the dataset to enhance the performance of the created model. Preprocessing the dataset is an essential step in any ML or DL task. It consists of transforming and preparing the raw data in a way that makes it suitable for training the model. The steps involved in preprocessing utilized in this study are:

— **Load the dataset:** The code loads the dataset from a *csv* file.
— **Drop unnecessary columns:** Certain columns in the dataset are not needed for the ML model, so they are dropped using the drop method of the DataFrame.
— **Drop rows with missing values:** Rows containing any missing values are removed from the dataset using the *dropna* method.
— **Shuffle the dataframe:** The rows in the DataFrame are shuffled randomly using the shuffle function from the *sklearn.utils* module. This is done to ensure that the data is not biased in any particular order.
— **Encode categorical variables:** Some columns in the dataset are categorical, meaning they represent categories rather than numerical values. To convert these categorical variables into a numerical format suitable for the model, one-hot encoding is performed using the *pd.get* dummies method.
— **Normalize the features:** The numerical features in the dataset are normalized to achieve a mean of '0' and a standard deviation of '1'. This is done by leveraging the StandardScaler from the *sklearn.preprocessing* module.
— **Encode the target variable:** The target variable, which is the 'Attack type' column representing the attack class, is encoded leveraging label encoding. Label encoding maps use the different classes to integer values.
— **Split the data:** The preprocessed data is split into training and testing sets using the train test split function from *sklearn.model* selection. The training set is utilized for model training, and the testing set is employed to evaluate its performance.

Binary classification

We build a DNN model for binary classification leveraging the given dataset. The model consists of Dense layers with Rectified Linear Unit (ReLU) activation functions and a Sigmoid activation function for the output layer. We compile and train the model using the Adaptive Moment Estimation (Adam) optimizer, binary crossentropy loss function, and Accuracy as the evaluation metric. The algorithm is depicted in Algorithm 1.

Algorithm 1 DNN Model for Binary Classification

Require: Input *x* train scaled, *y* train, *x* test scaled, *y* test

1: Perform label encoding on *y* train and *y* test to convert class labels into numerical format. build DNN model

2: Create a Sequential model

3: Add a Dense layer with 256 neurons and activation function ReLU, with input dimension equal to the number of features in *x* train scaled.

4: Add another Dense layer with 164 neurons and activation function ReLU.

5: Add another Dense layer with 82 neurons and activation function ReLU.

6: Add another Dense layer with 32 neurons and activation function ReLU.

7: Add the output Dense layer with 1 neuron and activation function Sigmoid (binary classification).

8: Call build DNN model() to build the DNN model for binary classification.

9: Compile the model using Adam optimizer and binary crossentropy loss function, with Accuracy as the evaluation metric.

10: Train the model with 25 epochs and a batch size of 32. Validate the model using *x* test scaled and *y* test encoded.

Multiclass Classification

In a well-structured dataset with efficient pre-processing, it is possible to modify the number of classes in the target according to our objectives and the use of the model in our environment. In Edget IIoTset the 'attacks-types' target is generally used to perform a multiclass classification. Manipulating the number of classes in the target variable can be beneficial in various ways:

— **Remove non-essential classes:** To simplify the model and improve focus on critical attacks, rare or less relevant attack classes were removed. The study retained the nine most common attack classes, where the model demonstrated high Accuracy. Fig. 2, *b* illustrates the distribution of these selected classes.

— **Merging similar classes:** Classes with similar attack characteristics were combined to reduce complexity while maintaining data representativeness. After analysis, 15 similar attack classes were merged into six broader categories (Fig. 2, *c*) ensuring essential attack features were preserved while enhancing model efficiency.

— **Aggregate classes:** Some classes had low Accuracy or insufficient data points, making them difficult to distinguish. Instead of removing them, they were grouped into a single new class called DA (Fig. 2, *a*). This aggregation increased representativeness and improved predictive performance.

— **Duplicate classes:** In certain cases, classes were duplicated to represent specific attack subcategories, enhancing the model ability to differentiate between various attack scenarios and improving Accuracy.

Once the data was pre-processed and the classes defined, we proceeded to model design (Fig. 2). The proposed architecture, called MC-CNN-DNN, combines a CNN for feature extraction and a DNN for classification. The CNN part includes three 1D convolutional layers with 256, 128, and 64 filters, respectively, each followed by a max-pooling layer (pool size = 2). The extracted features are flattened and passed to a DNN consisting of four fully connected layers with 256, 164, 82, and 32 neurons, all using ReLU activation. L2 regularization ($\lambda = 0.00001$) is applied to all dense layers. The final output layer has 6 (or 9, 10, 15) neurons with Softmax activation for multiclass classification. After one-hot encoding and normalization, the input data contained 96 features per sample, reshaped to match the CNN input format of (96, 1) representing 96 features and one channel. This shape is optimal for Conv1D layers. The model was compiled with the Adam optimizer (learning rate = 0.0001) and categorical cross-entropy as the loss function. It was trained over 25 epochs with a batch size of 32, using 20 % of the training data for validation. The full structure and implementation of the model are detailed in Algorithm 2.

Algorithm 2 MC-CNN-DNN Model

Require: *x* train, *y* train, *x* test, *y* test.

1: Initialize the model as Sequential()

2: Add Conv1D Layer with filters=256, kernel size=3, activation= 'relu', input shape=(xtrain.shape[1], 1)

3: Add MaxPooling1D Layer with pool size=2

4: Add Conv1D Layer with filters=128, kernel size=3, activation='relu'

5: Add MaxPooling1D Layer with pool size=2

6: Add Conv1D Layer with filters=64, kernel size=3, activation='relu'

7: Add MaxPooling1D Layer with pool size=2

8: Add Flatten Layer

9: Add Dense Layer with units=256, activation='relu', kernel regularizer=l2(0.00001)

10: Add Dense Layer with units=164, activation='relu', kernel regularizer=l2(0.00001)

11: Add Dense Layer with units=82, activation='relu', kernel regularizer=l2(0.00001)

12: Add Dense Layer with units=32, activation='relu', kernel regularizer=l2(0.00001)

13: Add Dense Layer with units=classnum, activation='softmax'

14: Set the optimizer as Adam with learning rate 0.001

15: Set the loss function as categorical crossentropy

16: Compile the model with optimizer and loss function

17: Train the model

18: Fit the model on *x* train and *y* train for 20 epochs with batch size=512 and validation split=0.2

19: Save the training history in history

20: Evaluate the model on *x* test and *y* test, and save the results in score.

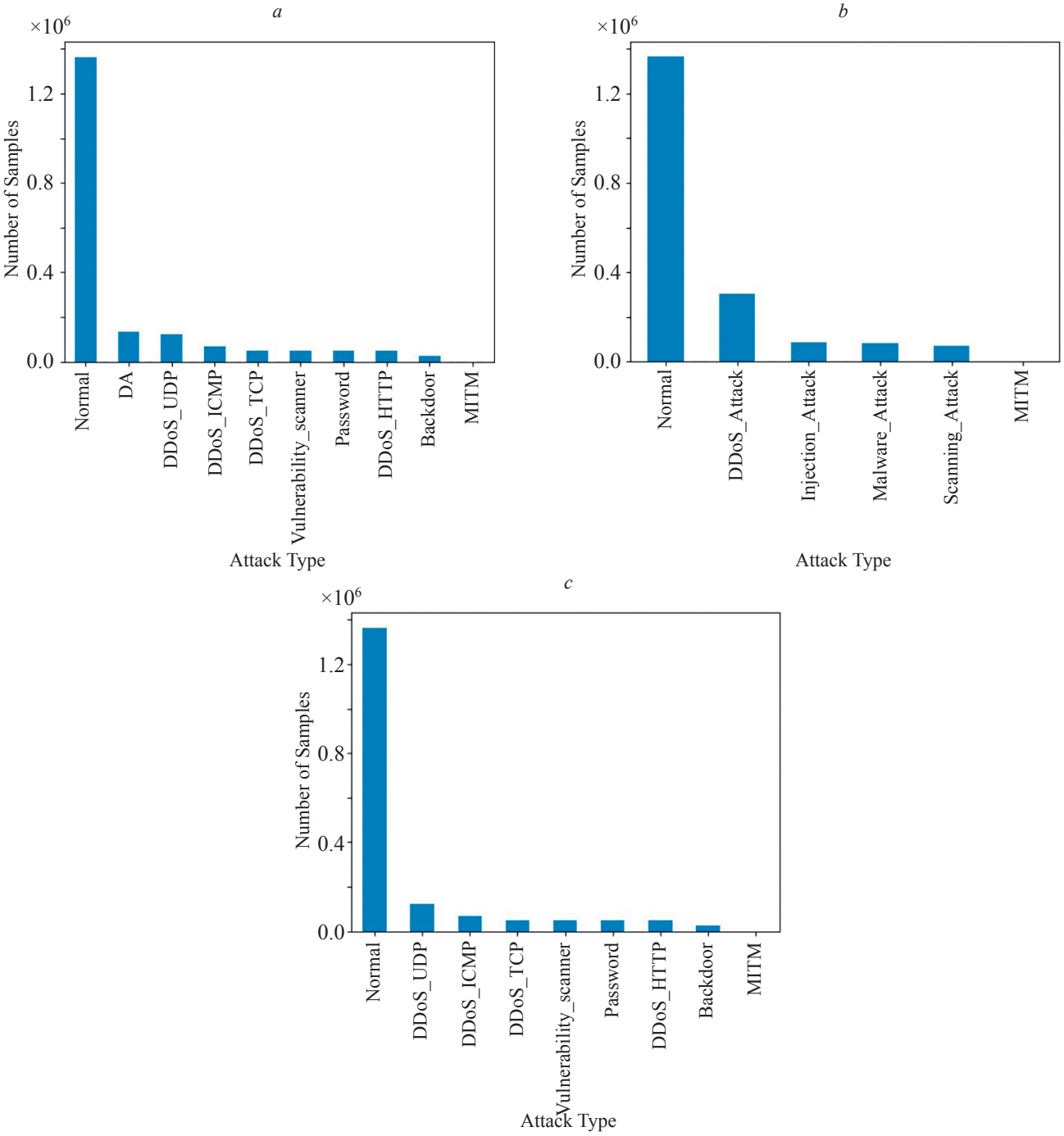

Fig. 2. Bar distribution across class settings: 10-class (*a*); 6-class (*b*); 9-class (*c*)

Results and Discussion

This paper proposes an innovative and efficient method for modern intrusion detection systems which are crucial for identifying unauthorized activity within computer networks. Despite the use of state-of-the-art algorithms to categorize a wide range of intrusion scenarios, their overall performance remains suboptimal. The experiment results show the model outstanding proficiency in distinguishing between the two classes, 'Normal' and 'Attack'. Achieving a perfect score (100 %) across all binary classification metrics — Accuracy, Precision, Recall, and F1-score, highlights its ability to classify instances with complete Accuracy while minimizing misclassifications. These results confirm the model exceptional suitability for binary classification tasks.

The Accuracy results across the testing, validation, and training sets using the "Edge IIoTset" dataset are illustrated in Fig. 4, with representing different class distribution scenarios.

Notably, our proposed method, the MC-CNN-DNN model, consistently demonstrates exceptional Accuracy across all approaches examined. Particularly, when assessing different class distributions, the 9-class distribution approach emerges as the standout performer, boasting an impressive Accuracy rate of 99.50 %. Similarly,

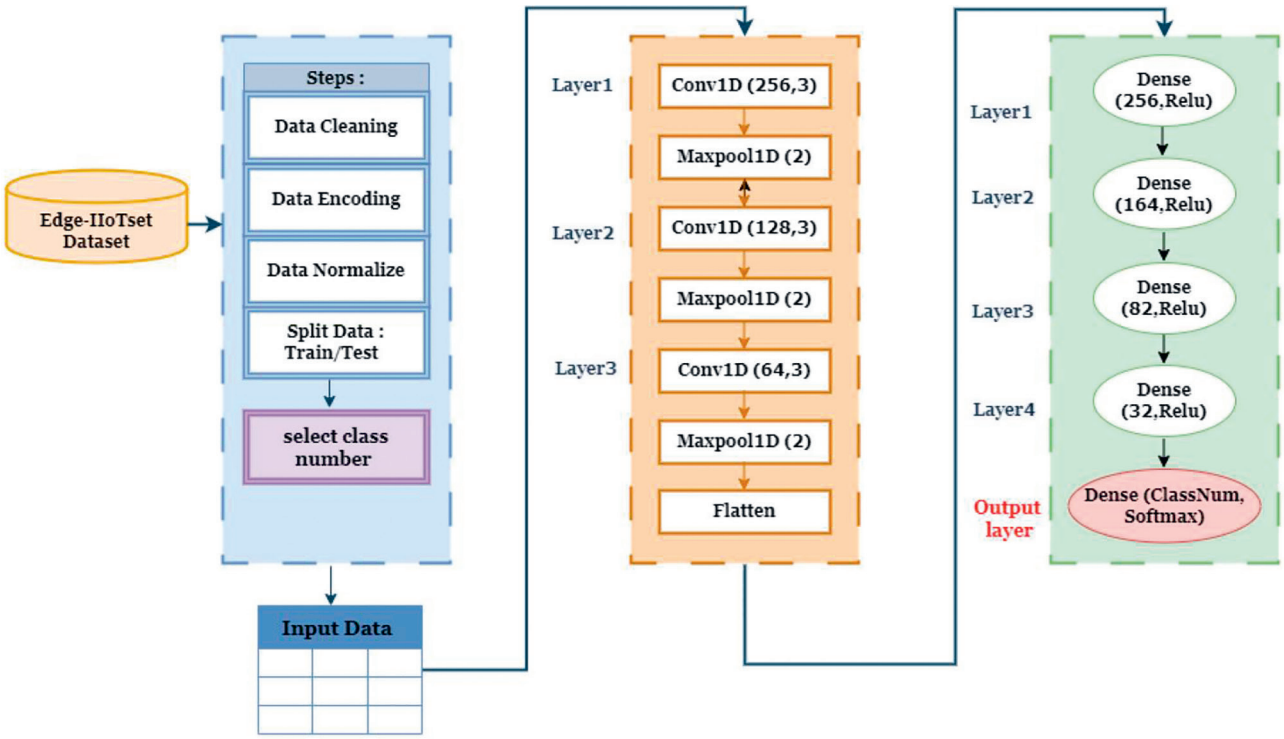

Fig. 3. Proposed methodology of the study

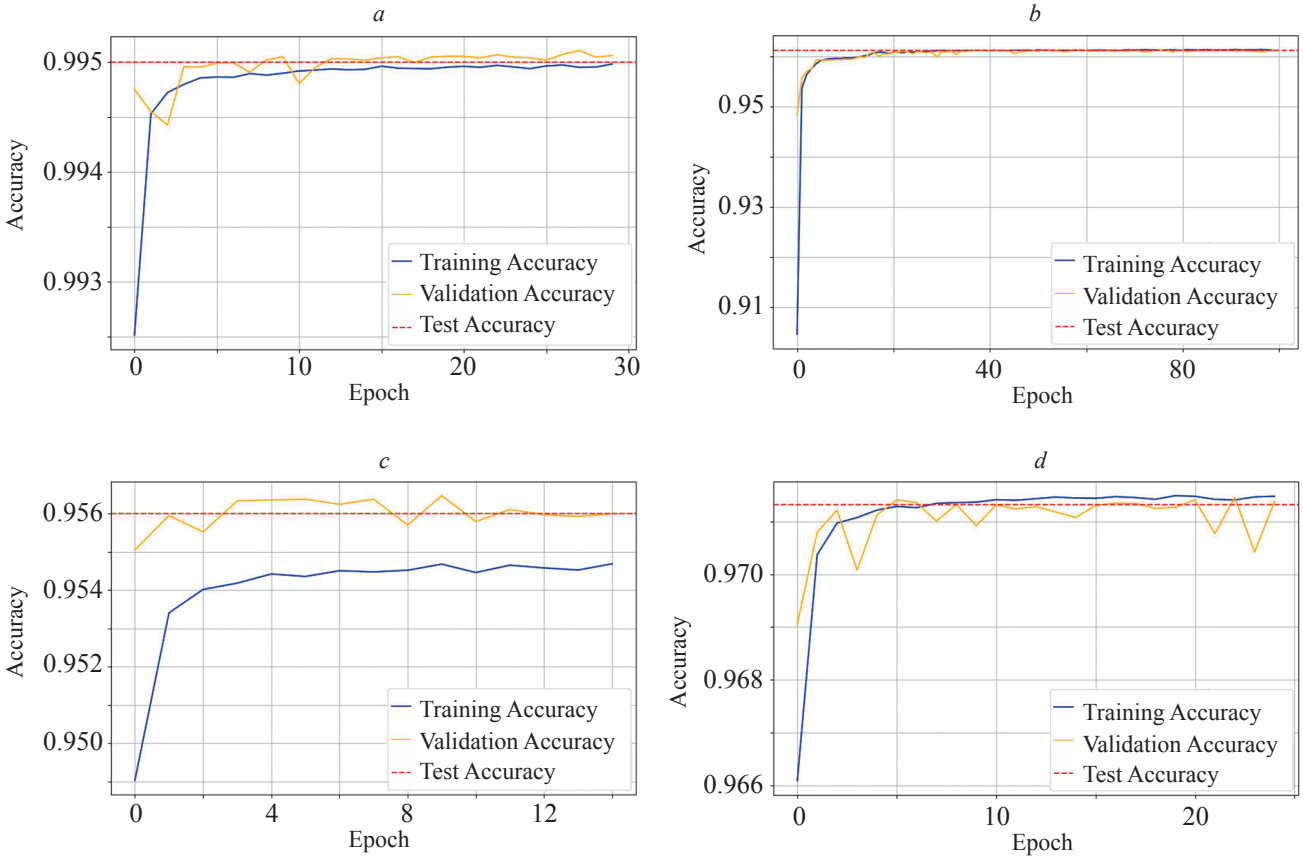

Fig. 4. Accuracy performance across various class distributions for the proposed MC-CNN-DNN model: 9-class (99.50 %) (*a*); 10-class (96.12 %) (*b*); 15-class (95.6 %) (*c*); and 6-class (97.14 %) (*d*)

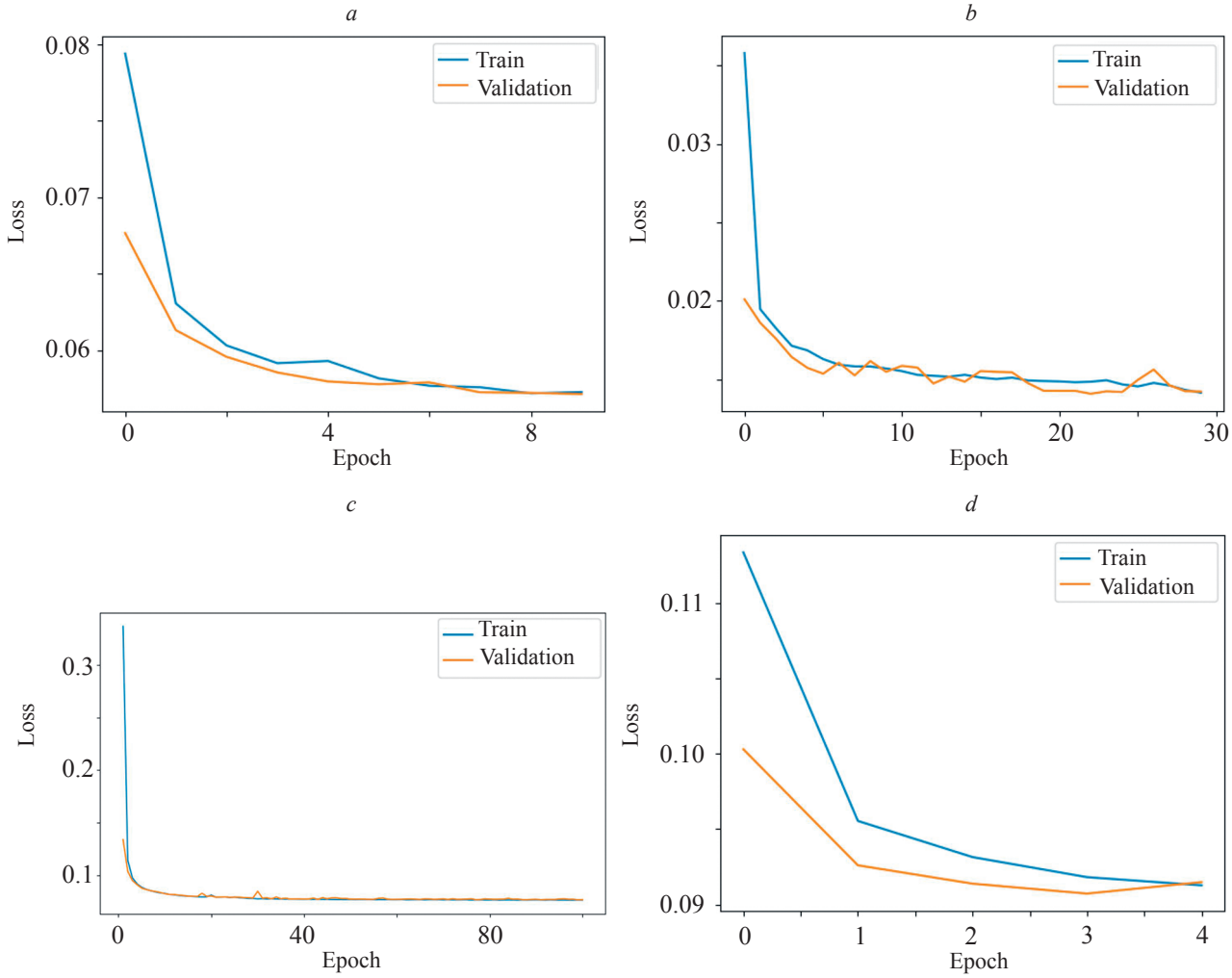

Fig. 5. Loss function across various class distributions: 6-class (*a*); 9-class (*b*); 10-class (*c*); 15-class (*d*)

the 6-class distribution approach exhibits a robust Accuracy level of 97.14 %. Meanwhile, the Accuracy for the 10-class distribution approach remains noteworthy at 96.12 %, followed closely by the 15-class distribution approach with an Accuracy of 95.6 %.

Fig. 5 subpictures reveal consistently low loss values across all scenarios, highlighting the model stability during training. Furthermore, the close alignment between training and validation loss curves indicates the absence of overfitting. These results underscore the efficacy of our proposed MC-CNN-DNN model in achieving high Accuracy across diverse class distribution scenarios, further affirming its potential for robust intrusion detection within the complex landscape of the "Edge IIoTset" dataset.

Table 2 shows the evaluation metrics in terms of Precision, Recall, and F1-score of a MC-CNN-DNN model on 15-class distribution, Classes like "Normal", "Backdoor", "DDoS HTTP", "DDoS ICMP", "DDoS TCP", "DDoS UDP", "Fingerprinting", "MITM", "Password", "Port Scanning", "Ransomware", "SQL injection", "Uploading", "Vulnerability scanner", and "XSS".

Table 3 shows the performance of a MC-CNN-DNN model on 9-class distribution. Classes like "Normal", "DDoS UDP", "DDoS ICMP", and "MITM" show perfect

Table 2. Evaluation Metrics for 15-class Distribution, %

Class	Precision	Recall	F1-score
Normal	100	100	100
Backdoor	94	97	96
DDoS HTTP	74	96	84
DDoS ICMP	100	100	100
DDoS TCP	84	100	91
DDoS UDP	100	100	100
Fingerprinting	35	46	40
MITM	100	100	100
Password	91	19	32
Port Scanning	85	57	69
Ransomware	100	75	86
SQL injection	46	91	61
Uploading	67	48	56
Vulnerability scanner	100	83	90
XSS	62	35	44

Table 3. Evaluation Metrics for 9-class Distribution, %

Class	Precision	Recall	F1-score
Normal	100	100	100
DDoS UDP	100	100	100
DDoS ICMP	100	100	100
DDoS TCP	99	100	100
Vulnerability scanner	100	100	100
Password	98	85	91
DDoS HTTP	87	98	92
Backdoor	100	98	99
MITM	100	100	100

Precision, Recall, and F1-score, indicating that the model performs exceptionally well on these classes. The "DDoS TCP" class has also high Precision 99 %, suggesting that there might be a few false positives. However, the Recall and F1-score are still high. "Vulnerability scanner", "Password", and "Backdoor" classes also show good performance, although "Password" has relatively lower Recall, impacting its F1-score. "DDoS HTTP" class has a lower Precision 87 % but a high Recall 98 %, resulting in a good F1-score.

In Table 4, the model demonstrates excellent performance in classifying various 10-class of attack subtypes. For "DDoS UDP" and "DDoS ICMP" classes, it achieves perfect Precision and Recall. In the "DDoS TCP" class, the model achieves a Precision of 82 % and a Recall of 100 %, resulting in an F1-score of 92 %. In "DA" class, the model performance is reasonable, achieving a Precision of 71 % and a Recall of 85 %, leading to an F1-score of 77 %. The model performs well on the "Vulnerability scanner" class with a Precision of 91 % and a Recall of 81 %. However, for the "Password" class, Precision is perfect at 100 %, and Recall is 84 %. In the "Backdoor" class, the model performs admirably with high Precision of 99 % and Recall of 95 %, resulting in an F1-score of 97 %.

In Table 5, for 6-class distribution the model performance remains strong. In "DDoS attack" it achieves a Precision of 68 % and a high Recall of 99 %, resulting in an F1-score of 81 %. For "Injection attack", "Scanning attack", and "MITM", the model excels with perfect Precision, Recall, and F1-score for these classes. However, in "Malware attack", the model performance is moderate, attaining a Precision of 95 % but a lower Recall of 51 %, which leads to an F1-score of 66 %. Table 6 summarizes the results obtained in terms of both Accuracy and loss function.

Table 7 offers a concise comparison of model performances within the domain of intrusion detection leveraging the "Edge IIoTset" dataset. Our MC-CNN-DNN hybrid model stands out with the highest Accuracy, indicating its robustness and potential for enhanced security measures in industrial IoT environments. This comparison sheds light on the advancements made in intrusion detection techniques, further contributing to the development of effective solutions for safeguarding IIoT systems.

Table 4. Evaluation Metrics for 10-class Distribution, %

Class	Precision	Recall	F1-score
Normal	100	100	100
DA	71	85	77
DDoS UDP	100	100	100
DDoS ICMP	100	99	100
DDoS TCP	82	100	92
Vulnerability scanner	91	81	92
Password	100	84	91
DDoS HTTP	75	94	84
Backdoor	99	95	97
MITM	100	100	100

Table 5. Evaluation Metrics for 6-class Distribution, %

Class	Precision	Recall	F1-score
Normal	95	100	97
DDoS Attack	68	99	81
Injection attack	100	100	100
Malware attack	95	51	66
Scanning attack	100	100	100
MITM	98	73	84

Table 6. Summary of results

Class Num	Accuracy, %	Loss Function
2-class	100	$5.52 \cdot 10^{-6}$
6-class	97	0.062
9-class	99	0.014
10-class	96	0.070
15-class	95	0.080

Table 7. Comparison of the results with previous studies using the 'Edge IIoTset' dataset

Authors	Model	Accuracy, %
[32]	DNN	96.0
[33]	CNN-LSTM	98.7
[34]	Inception Time	94.9
Our work	CNN-DNN	99.5

Limitations and Future Work

While our proposed Multiclassification CNN-DNN model achieved outstanding results on the Edge-IIoTset dataset, certain considerations remain for future exploration. As with most DL models, performance can vary with dataset size and class balance, suggesting that larger or more diverse datasets may further enhance generalization. The hybrid architecture, though highly effective, introduces a moderate computational cost that could be optimized

for edge deployments. Moreover, extending validation to other IIoT datasets would further confirm the model adaptability across varying industrial environments. Future work will focus on improving efficiency and portability, while exploring integration with other learning strategies such as autoencoders, Reinforcement Learning, or Graph Neural Networks.

Conclusion

This study presented a hybrid Convolutional Neural Network with Deep Neural Network model for intrusion detection, trained and tested on the Edge-IIoTset dataset. By applying thorough preprocessing and using various class distribution scenarios (2, 6, 9, 10, and 15 classes), the model achieved excellent classification results. Notably, the 9-class configuration reached an Accuracy of 99.13 %, while the 6-class and 10-class setups also delivered strong performance at 97.13 % and 96.11 %, respectively. Our architecture effectively combines feature extraction and deep classification layers, resulting in a robust solution adaptable to complex Industrial Internet of Things traffic. Future work will focus on integrating other models like Reinforcement Learning, autoencoders, and Graph Neural Networks, along with evaluating the system on new datasets and in real-time industrial environments.

References

1. Jaidka H., Sharma N., Singh R. Evolution of IoT to IIoT: applications & challenges. *Proc. of the International Conference on Innovative Computing & Communications (ICICC)*, 2020, pp. 1–6. https://doi.org/10.2139/ssrn.3603739
2. Farhan L., Kharel R., Kaiwartya O., Quiroz-Castellanos M., Alissa A., Abdulsalam M. A concise review on internet of things (IoT)-problems, challenges and opportunities. *Proc. of the 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*, 2018, pp. 1–6. https://doi.org/10.1109/CSNDSP.2018.8471762
3. Chalishazar T. *Peerbits exploring the applications of IoT in different industries*, 2023. Available at: https://www.peerbits.com/blog/iot-applications-in-different-industries.html (accessed: 24.06.2023)
4. Qiu T., Chi J., Zhou X., Ning Z., Atiquzzaman M., Wu D.O. Edge computing in Industrial Internet of Things: architecture, advances and challenges. *IEEE Communications Surveys & Tutorials*, 2020, vol. 22, no. 4, pp. 2462–2488. https://doi.org/10.1109/COMST.2020.3009103
5. Alguliyev R., Imamverdiyev Y., Sukhostat L. Cyber-physical systems and their security issues. *Computers in Industry*, 2018, vol. 100, no. 1, pp. 212–223.
6. Mohamed N., Al-Jaroodi J., Jawhar I. Cyber–physical systems forensics: today and tomorrow. *Journal of Sensor and Actuator Networks*, 2020, vol. 9, no. 3, pp. 37. https://doi.org/10.3390/jsan9030037
7. Javaid M., Haleem A., Singh R.P., Suman R., Gonzalez E.S. Understanding the adoption of industry 4.0 technologies in improving environmental sustainability. *Sustainable Operations and Computers*, 2022, vol. 3, pp. 203–217. https://doi.org/10.1016/j.susoc.2022.01.008
8. Mirani A.A., Velasco-Hernandez G., Awasthi A., Walsh J. Key challenges and emerging technologies in industrial iot architectures: A review. *Sensors*, 2022, vol. 22, no. 15, pp. 5836. https://doi.org/10.3390/s22155836
9. Younan M., Houssein E.H., Elhoseny M., Ali A.A. Challenges and recommended technologies for the industrial internet of things: A comprehensive review. *Measurement*, 2020, vol. 151, pp. 107198. https://doi.org/10.1016/j.measurement.2019.107198
10. Gebremichael T., Ledwaba L.P., Eldefrawy M.H., Hancke G.P., Pereira N., Gidlund M., Akerberg J. Security and privacy in the industrial internet of things: current standards and future challenges. *IEEE Access*, 2020, vol. 8, pp. 152351–152366. https://doi.org/10.1109/ACCESS.2020.3016937
11. Madhuri G.S., Rani M.U. Anomaly detection techniques. *Proc. of the IADS International Conference on Computing, Communications & Data Engineering (CCODE)*, 2018, pp. 1–6.
12. Munir M., Chattha M.A., Dengel A., Ahmed S. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. *Proc. of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019, pp. 561–566. https://doi.org/10.1109/icmla.2019.00105
13. Du J., Yang K., Hu Y., Jiang L. NIDS-CNNLSTM: Network intrusion detection classification model based on deep learning. *IEEE Access*, 2023, vol. 11, pp. 24808–24821. https://doi.org/10.1109/ACCESS.2023.3254915
14. Kandhro I.A., Alanazi S.M., Ali F., Kehar A., Fatima K., Uddin M. Detection of real-time malicious intrusions and attacks in IoT

Литература

1. Jaidka H., Sharma N., Singh R. Evolution of IoT to IIoT: applications & challenges // Proc. of the International Conference on Innovative Computing & Communications (ICICC). 2020. P. 1–6. https://doi.org/10.2139/ssrn.3603739
2. Farhan L., Kharel R., Kaiwartya O., Quiroz-Castellanos M., Alissa A., Abdulsalam M. A concise review on internet of things (IoT)-problems, challenges and opportunities // Proc. of the 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP). 2018. P. 1–6. https://doi.org/10.1109/CSNDSP.2018.8471762
3. Chalishazar T. Peerbits exploring the applications of IoT in different industries. 2023. URL: https://www.peerbits.com/blog/iot-applications-in-different-industries.html (accessed: 24.06.2023)
4. Qiu T., Chi J., Zhou X., Ning Z., Atiquzzaman M., Wu D.O. Edge computing in Industrial Internet of Things: architecture, advances and challenges // IEEE Communications Surveys & Tutorials. 2020. V. 22. N 4. P. 2462–2488. https://doi.org/10.1109/COMST.2020.3009103
5. Alguliyev R., Imamverdiyev Y., Sukhostat L. Cyber-physical systems and their security issues // Computers in Industry. 2018. V. 100. N 1. P. 212–223.
6. Mohamed N., Al-Jaroodi J., Jawhar I. Cyber–physical systems forensics: today and tomorrow // Journal of Sensor and Actuator Networks. 2020. V. 9. N 3. P. 37. https://doi.org/10.3390/jsan9030037
7. Javaid M., Haleem A., Singh R.P., Suman R., Gonzalez E.S. Understanding the adoption of industry 4.0 technologies in improving environmental sustainability // Sustainable Operations and Computers. 2022. V. 3. P. 203–217. https://doi.org/10.1016/j.susoc.2022.01.008
8. Mirani A.A., Velasco-Hernandez G., Awasthi A., Walsh J. Key challenges and emerging technologies in industrial iot architectures: A review // Sensors. 2022. V. 22. N 15. P. 5836. https://doi.org/10.3390/s22155836
9. Younan M., Houssein E.H., Elhoseny M., Ali A.A. Challenges and recommended technologies for the industrial internet of things: A comprehensive review // Measurement. 2020. V. 151. P. 107198. https://doi.org/10.1016/j.measurement.2019.107198
10. Gebremichael T., Ledwaba L.P., Eldefrawy M.H., Hancke G.P., Pereira N., Gidlund M., Akerberg J. Security and privacy in the industrial internet of things: current standards and future challenges // IEEE Access. 2020. V. 8. P. 152351–152366. https://doi.org/10.1109/ACCESS.2020.3016937
11. Madhuri G.S., Rani M.U. Anomaly detection techniques // Proc. of the IADS International Conference on Computing, Communications & Data Engineering (CCODE). 2018. P. 1–6.
12. Munir M., Chattha M.A., Dengel A., Ahmed S. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data // Proc. of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA). 2019. P. 561–566. https://doi.org/10.1109/icmla.2019.00105
13. Du J., Yang K., Hu Y., Jiang L. NIDS-CNNLSTM: Network intrusion detection classification model based on deep learning // IEEE Access. 2023. V. 11. P. 24808–24821. https://doi.org/10.1109/ACCESS.2023.3254915
14. Kandhro I.A., Alanazi S.M., Ali F., Kehar A., Fatima K., Uddin M. Detection of real-time malicious intrusions and attacks in IoT

empowered cybersecurity infrastructures. *IEEE Access*, 2023, vol. 11, pp. 9136– 9148. https://doi.org/10.1109/ACCESS.2023.3238664
15. Alrowaily M., Alenezi F., Lu Z. Effectiveness of machine learning based intrusion detection systems. *Lecture Notes in Computer Science*, 2019, vol. 11611, pp. 277–288. https://doi.org/10.1007/978-3-030-24907-6_21
16. Cam N.T., Trung N.G. An intelligent approach to improving the performance of threat detection in IoT. *IEEE Access*, 2023, vol. 11, pp. 44319–44334. https://doi.org/10.1109/ACCESS.2023.3273160
17. Al-Abassi A., Karimipour H., Dehghantanha A., Pariz R.M. An ensemble deep learning-based cyber-attack detection in industrial control system. *IEEE Access*, 2020, vol. 8, pp. 83965– 83973. https://doi.org/10.1109/ACCESS.2020.2992249
18. Shone N., Ngoc T.N., Phai V.D., Shi Q. A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018, vol. 2, no. 1, pp. 41–50. https://doi.org/10.1109/TETCI.2017.2772792
19. Ullah I., Mahmoud Q.H. Design and development of a deep learning-based model for anomaly detection in IoT networks. *IEEE Access*, 2021, vol. 9, pp. 103906–103926. https://doi.org/10.1109/ACCESS.2021.3094024
20. Gümüşbaş D., Yıldırım T., Genovese A., Scotti F. A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. *IEEE Systems Journal*, 2020, vol. 15, no. 2, pp. 1717–1731. https://doi.org/10.1109/JSYST.2020.2992966
21. Ashraf E., Areed N.F., Salem H., Salem H., Abdelhady E., Farouk A. IoT based intrusion detection systems from the perspective of machine and deep learning: a survey and comparative study. *Delta University Scientific Journal*, 2022, vol. 5, no. 2, pp. 367–386. https://doi.org/10.21608/dusj.2022.275552
22. Thakkar A., Lohiya R. A review of the advancement in intrusion detection datasets. *Procedia Computer Science*, 2020, vol. 167, pp. 636–645. https://doi.org/10.1016/j.procs.2020.03.330
23. Mishra N., Pandya S. Internet of Things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. *IEEE Access*, 2021, vol. 9, pp. 59353–59377. https://doi.org/10.1109/ACCESS.2021.3073408
24. Liu L., Wang P., Lin J., Liu L. Intrusion detection of imbalanced network traffic based on machine learning and deep learning. *IEEE Access*, 2020, vol. 9, pp. 7550–7563. https://doi.org/10.1109/ACCESS.2020.3048198
25. Ito A., Saito K., Ueno R., Homma N. Imbalanced data problems in deep learning-based side-channel attacks: Analysis and solution. *IEEE Transactions on Information Forensics and Security*, 2021, vol. 16, pp. 3790–3802. https://doi.org/10.1109/TIFS.2021.3092050
26. Goyal P., Pandey S., Jain K. *Deep Learning for Natural Language Processing: Creating Neural Networks with Python*. Apress, 2018, 294 p.
27. Chinnathambi R.A., Plathottam S.J., Hossen T., Nair A.S., Ranganathan P. Deep neural networks (DNN) for day-ahead electricity price markets. *Proc. of the IEEE electrical power and energy conference (EPEC)*, 2018, pp. 1–6. https://doi.org/10.1109/EPEC.2018.8598327
28. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958, vol. 65, no. 6, pp. 386–408. https://doi.org/10.1037/h0042519
29. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989, vol. 1, no. 4, pp. 541–551. https://doi.org/10.1162/neco.1989.1.4.541
30. Hubel D.H., Wiesel T.N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 1968, vol. 195, no. 1, pp. 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455
31. Yamashita R., Nishio M., Do R.K.G., Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 2018, vol. 9, pp. 611–629. https://doi.org/10.1007/s13244-018-0639-9
32. Ferrag M.A., Friha O., Hamouda D., Maglaras L., Janicke H. Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEE Access*, 2022, vol. 10, pp. 40281–40306. https://doi.org/10.1109/access.2022.3165809
33. Khacha A., Saadouni R., Harbi Y., Aliouat Z. Hybrid deep learning-based intrusion detection system for industrial Internet of Things. *Proc. of the 5th International Symposium on Informatics and its*

empowered cybersecurity infrastructures // IEEE Access. 2023. V. 11. P. 9136– 9148. https://doi.org/10.1109/ACCESS.2023.3238664
15. Alrowaily M., Alenezi F., Lu Z. Effectiveness of machine learning based intrusion detection systems // Lecture Notes in Computer Science. 2019. V. 11611. P. 277–288. https://doi.org/10.1007/978-3-030-24907-6_21
16. Cam N.T., Trung N.G. An intelligent approach to improving the performance of threat detection in IoT // IEEE Access. 2023. V. 11. P. 44319–44334. https://doi.org/10.1109/ACCESS.2023.3273160
17. Al-Abassi A., Karimipour H., Dehghantanha A., Pariz R.M. An ensemble deep learning-based cyber-attack detection in industrial control system // IEEE Access. 2020. V. 8. P. 83965– 83973. https://doi.org/10.1109/ACCESS.2020.2992249
18. Shone N., Ngoc T.N., Phai V.D., Shi Q. A deep learning approach to network intrusion detection // IEEE Transactions on Emerging Topics in Computational Intelligence. 2018. V. 2. N 1. P. 41–50. https://doi.org/10.1109/TETCI.2017.2772792
19. Ullah I., Mahmoud Q.H. Design and development of a deep learning-based model for anomaly detection in IoT networks // IEEE Access. 2021. V. 9. P. 103906–103926. https://doi.org/10.1109/ACCESS.2021.3094024
20. Gümüşbaş D., Yıldırım T., Genovese A., Scotti F. A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems // IEEE Systems Journal. 2020. V. 15. N 2. P. 1717–1731. https://doi.org/10.1109/JSYST.2020.2992966
21. Ashraf E., Areed N.F., Salem H., Salem H., Abdelhady E., Farouk A. IoT based intrusion detection systems from the perspective of machine and deep learning: a survey and comparative study // Delta University Scientific Journal. 2022. V. 5. N 2. P. 367–386. https://doi.org/10.21608/dusj.2022.275552
22. Thakkar A., Lohiya R. A review of the advancement in intrusion detection datasets // Procedia Computer Science. 2020. V. 167. P. 636–645. https://doi.org/10.1016/j.procs.2020.03.330
23. Mishra N., Pandya S. Internet of Things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review // IEEE Access. 2021. V. 9. P. 59353–59377. https://doi.org/10.1109/ACCESS.2021.3073408
24. Liu L., Wang P., Lin J., Liu L. Intrusion detection of imbalanced network traffic based on machine learning and deep learning // IEEE Access. 2020. V. 9. P. 7550–7563. https://doi.org/10.1109/ACCESS.2020.3048198
25. Ito A., Saito K., Ueno R., Homma N. Imbalanced data problems in deep learning-based side-channel attacks: Analysis and solution // IEEE Transactions on Information Forensics and Security. 2021. V. 16. P. 3790–3802. https://doi.org/10.1109/TIFS.2021.3092050
26. Goyal P., Pandey S., Jain K. Deep Learning for Natural Language Processing: Creating Neural Networks with Python. Apress, 2018. 294 p.
27. Chinnathambi R.A., Plathottam S.J., Hossen T., Nair A.S., Ranganathan P. Deep neural networks (DNN) for day-ahead electricity price markets // Proc. of the IEEE electrical power and energy conference (EPEC). 2018. P. 1–6. https://doi.org/10.1109/EPEC.2018.8598327
28. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // Psychological Review. 1958. V. 65. N 6. P. 386–408. https://doi.org/10.1037/h0042519
29. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D. Backpropagation applied to handwritten zip code recognition // Neural Computation. 1989. V. 1. N 4. P. 541–551. https://doi.org/10.1162/neco.1989.1.4.541
30. Hubel D.H., Wiesel T.N. Receptive fields and functional architecture of monkey striate cortex // The Journal of Physiology. 1968. V. 195. N 1. P. 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455
31. Yamashita R., Nishio M., Do R.K.G., Togashi K. Convolutional neural networks: an overview and application in radiology // Insights into Imaging. 2018. V. 9. P. 611–629. https://doi.org/10.1007/s13244-018-0639-9
32. Ferrag M.A., Friha O., Hamouda D., Maglaras L., Janicke H. Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning // IEEE Access. 2022. V. 10. P. 40281–40306. https://doi.org/10.1109/access.2022.3165809
33. Khacha A., Saadouni R., Harbi Y., Aliouat Z. Hybrid deep learning-based intrusion detection system for industrial Internet of Things // Proc. of the 5th International Symposium on Informatics and its

Applications (ISIA), 2022, pp. 1–6. https://doi.org/10.1109/ISIA55826.2022.9993487
34. Tareq I., Elbagoury B.M., El-Regaily S., El-Horbaty E.S.M. Analysis of ToN-IoT, UNW-NB15, and edge-IIoT datasets using dl in cybersecurity for IoT. *Applied Sciences*, 2022, vol. 12, no. 19, pp. 9572. https://doi.org/10.3390/app12199572

Applications (ISIA). 2022. P. 1–6. https://doi.org/10.1109/ISIA55826.2022.9993487
34. Tareq I., Elbagoury B.M., El-Regaily S., El-Horbaty E.S.M. Analysis of ToN-IoT, UNW-NB15, and edge-IIoT datasets using dl in cybersecurity for IoT // Applied Sciences. 2022. V. 12. N 19. P. 9572. https://doi.org/10.3390/app12199572

Authors

Wafaa Ferhi — PhD Student, Assistant, University of Abu Bekr Belkaid, Tlemcen, 13000, Algeria, sc 58480659800, https://orcid.org/0009-0005-7574-8368, wafaa.ferhi@univ-tlemcen.dz
Djilali Moussaoui — Lecturer, University of Abu Bekr Belkaid, Tlemcen, 13000, Algeria, sc 56360232600, https://orcid.org/0000-0003-3478-263X, djilali.moussaoui@univ-tlemcen.dz
Mourad Hadjila — Lecturer, University of Abu Bekr Belkaid, Tlemcen, 13000, Algeria, sc 56440246000, https://orcid.org/0000-0002-6554-3925, mourad.hadjila@univ-tlemcen.dz
Al Baraa Bouidaine — PhD Student, Assistant, University of Abu Bekr Belkaid, Tlemcen, 13000, Algeria, sc 58482050500, https://orcid.org/0009-0005-2204-9117, albaraa.bouidaine@univ-tlemcen.dz

Авторы

Ферхи Вафаа — аспирант, ассистент, Университет Абу Бекра Белкаида, Тлемсен, 13000, Алжир, sc 58480659800, https://orcid.org/0009-0005-7574-8368, wafaa.ferhi@univ-tlemcen.dz
Муссауи Джилали — преподаватель, Университет Абу Бекра Белкаида, Тлемсен, 13000, Алжир, sc 56360232600, https://orcid.org/0000-0003-3478-263X, djilali.moussaoui@univ-tlemcen.dz
Хаджила Мурад — преподаватель, Университет Абу Бекра Белкаида, Тлемсен, 13000, Алжир, sc 56440246000, https://orcid.org/0000-0002-6554-3925, mourad.hadjila@univ-tlemcen.dz
Буиден Аль Бараа — аспирант, ассистент, Университет Абу Бекра Белкаида, Тлемсен, 13000, Алжир, sc 58482050500, https://orcid.org/0009-0005-2204-9117, albaraa.bouidaine@univ-tlemcen.dz

Received 11.02.2025
Approved after reviewing 27.08.2025
Accepted 25.09.2025

Статья поступила в редакцию 11.02.2025
Одобрена после рецензирования 27.08.2025
Принята к печати 25.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-888-901

Incorporating negative examples into Hidden Markov Model-based classification of peptide sequences

Valeriia A. Polezhaeva[1]✉, Denis A. Kleverov[2]✉, Anatoly A. Shalyto[3], Maxim Artyomov[4]

[1,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation
[2,4] Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 631110, USA

[1] polezhaevalera@yandex.ru✉, https://orcid.org/0009-0001-7469-7440
[2] denklewer@gmail.com✉, https://orcid.org/0009-0002-1362-486X
[3] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077
[4] martyomov@pathology.wustl.edu, https://orcid.org/0000-0002-1133-4212

Abstract
Hidden Markov Models (HMMs) trained to identify binding regions in peptide sequences have demonstrated the ability to uncover shared amino acid patterns in peptides bound to major histocompatibility complex molecules. In this work, we present an enhanced approach for predicting peptide binding using an ensemble of HMMs. Building on a previously proposed method, we extend it to a classification setting by incorporating both binding (positive) and non-binding (negative) peptide sequences. Our strategy involves training two sets of models on these distinct datasets and selecting ensemble members based on conditional probability estimates. The method was evaluated across six alleles of major histocompatibility complex using two model architectures: simplified architecture with 9 states representing the peptide binding core region and two cycle-states for the amino acids outside this region, and extended architecture, in which each cycle state was replaced by 9 additional states. Models evaluated in comparison with the state-of-the-art MixMHC2pred predictor. Results show a statistically significant improvement in prediction accuracy. Notably, incorporating non-binding peptides during training improved performance in several cases, highlighting the importance of background sequence information in distinguishing binding-specific patterns.

Keywords
peptide binding, Hidden Markov Models, binding prediction, negative examples, epitope prediction, major histocompatibility complex, binding motifs, machine learning

For citation: Polezhaeva V.A., Kleverov D.A., Shalyto A.A., Artyomov M. Incorporating negative examples into Hidden Markov Model-based classification of peptide sequences. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 888–901. doi: 10.17586/2226-1494-2025-25-5-888-901

УДК 004.85

Классификация пептидных последовательностей с использованием скрытых марковских моделей, учитывающих отрицательные примеры

Валерия Александровна Полежаева[1]✉, Денис Анатольевич Клеверов[2]✉, Анатолий Абрамович Шалыто[3], Максим Артемов[4]

[1,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2,4] Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США

[1] polezhaevalera@yandex.ru✉, https://orcid.org/0009-0001-7469-7440
[2] denklewer@gmail.com✉, https://orcid.org/0009-0002-1362-486X
[3] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077
[4] martyomov@pathology.wustl.edu, https://orcid.org/0000-0002-1133-4212

Аннотация

Введение. Скрытые марковские модели могут применяться к задаче идентификации ядра связывания пептида с молекулами главного комплекса гистосовместимости, выявляя общие аминокислотные паттерны анализируемых последовательностей. Представлен усовершенствованный подход к решению этой задачи на основе ансамбля скрытых марковских моделей. Ранее предложенный авторами метод адаптирован к задаче классификации пептидов на два класса: связывающиеся и не связывающиеся. **Метод.** Разработанный подход включает в себя обучения двух типов моделей: первый тип — обученный с использованием связывающихся пептидов (положительных примеров данных), второй — не связывающихся пептидов (отрицательных примеров данных). Отбор моделей в ансамбль и классификация последовательностей выполнялась на основе оценки условной вероятности между полученными моделями. **Основные результаты.** Модифицированная стратегия обучения ансамбля моделей протестирована для шести различных аллелей главного комплекса гистосовместимости с использованием двух архитектур моделей. В первом случае использовалась упрощенная структура с девятью состояниями модели, соответствующими ядру связывания пептида, и двумя состояниями-циклами для аминокислот вне этого ядра. Во втором случае применялась расширенная схема, где состояния-циклы заменялись девятью дополнительными состояниями. Оценка эффективности моделей производилась в сравнении с современным методом MixMHC2pred, в ходе которой обученные модели продемонстрировали статистически значимое повышение точности предсказаний класса пептидов. **Обсуждение.** Разработанная стратегия обучения моделей учитывает как связывающиеся, так и не связывающиеся с комплексом пептиды, позволяет повысить точность предсказания класса связывания скрытыми марковскими моделями даже в условиях ограниченного объема положительных данных. Повышение предсказания в этом случае достигается за счет использования фонового распределения аминокислотных последовательностей, полученного из отрицательной выборки.

Ключевые слова

связывание пептидов, скрытые марковские модели, отрицательная выборка, предсказание связывающей способности, предсказание эпитопов, главные комплексы гистосовместимости, связывающие мотивы, машинное обучение

Ссылка для цитирования: Полежаева В.А., Клеверов Д.А., Шалыто А.А., Артемов М. Классификация пептидных последовательностей с использованием скрытых марковских моделей, учитывающих отрицательные примеры // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 888–901 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-888-901

Introduction

A key characteristic of the immune system is its ability to distinguish foreign (non-self) targets from the body own tissues (self), a process primarily mediated by T cells. T cells play a central role in recognizing antigenic peptides presented by the Major Histocompatibility Complex (MHC) of Antigen-Presenting Cells (APCs). The interaction between T Cell Receptors (TCRs) and peptide-MHC (pMHC) complexes enables T cells to identify and respond to foreign pathogens while maintaining tolerance to self-antigens [1].

Identifying peptide sequences involved in this immune response, known as neoepitopes, is a critical step in cancer vaccine development [2]. Peptides that are likely to be bound and presented by MHC molecules and subsequently recognized by immune cells can serve as components of personalized vaccination strategies. These peptides can stimulate immune responses, even in cases where natural immune activation does not occur [3]. They exhibit high binding affinity — the strength of their interaction with MHC molecules [4].

MHC molecules exhibit specificity in peptide binding, governed by the structure of their peptide-binding grooves. The strength and stability of MHC-peptide interactions significantly influence antigen presentation and T cell activation [4].

MHC class I molecules primarily bind peptides of 9 amino acids in length, though they can accommodate peptides ranging from 8 to 15 amino acids. Their binding site consists of two helices forming a closed pocket, with interactions mainly occurring at the peptide terminal residues. The central region may play a comparatively lesser role [5, 6]. In contrast, MHC class II molecules bind longer peptides, typically ranging from 12 to 25 amino acids, with preference for 15-mers. Their open-ended binding groove allows the epitope to extend beyond the binding site, although the main stabilizing interaction with a binding groove is mediated by a nine-amino-acid region, known as the binding core. Therefore, despite the variability in peptide length, the fixed length core of MHC class II molecules plays a critical role in determining the specificity and stability of MHC-peptide interactions. Moreover, each individual can express up to six MHC class I and II types (alleles), but population-level diversity is vast — over 20,000 Class I and 8,000 Class II variants exist [5]. Each allele has unique peptide-binding preferences, generating distinct binding motifs.

From a computational perspective, selecting peptide candidates for vaccination can be formulated as a classification problem in machine learning. A model processes an input protein sequence and outputs a numerical score reflecting the likelihood of a peptide successfully passing through different stages of antigen processing.

This classification distinguishes:

— Binders — peptides that successfully bind MHC and undergo antigen processing;

— Non-binders — peptides that fail to bind or be processed effectively.

In this approach, protein sequences are encoded using standard amino acid labels (ACDEFGHIKLMNPQRSTVWY, where ‘A’ represents alanine and ‘Y’ represents tyrosine).

Hidden Markov Models (HMMs) offer a flexible and interpretable approach for identifying binding cores and predicting MHC-peptide binding affinity. Unlike neural networks, which often struggle with interpretability, HMMs can effectively model peptide sequences of different lengths by capturing underlying probabilistic dependencies [7].

Recent studies [5, 8, 9] have demonstrated the effectiveness of HMMs in binding affinity analysis, particularly due to their online learning capabilities, which allow incremental updates to transition and emission probabilities without requiring full retraining. This adaptability makes HMMs well-suited for evolving peptide datasets.

Additionally, HMMs have been explored for refining binding core identification, improving affinity estimation accuracy. However, comparative studies on different HMM architectures remain limited, and existing research often lacks detailed analyses of structural variations in these models. While some studies [5] have evaluated HMM performance for specific MHC classes, a broader comparison across multiple alleles is still needed to fully understand their potential in MHC-peptide interaction modeling.

In this study, we extend the methodology of Hidden Markov-based peptide analysis models by adapting a previously proposed approach to the machine learning task of peptide classification. Our method involves the construction of an ensemble of predictors, where each predictor is built from two independently trained HMMs: one using binder peptides and the other using non-binders, which serve as a representation of the background distribution of the peptide space. Predictors are included in the final ensemble based on conditional probability estimates derived from these models. In the following sections, we first outline the models training and validation procedures followed by ensemble construction pipeline. We then present the results of our comparative experiments across multiple MHC alleles, demonstrating improved classification performance relative to state-of-the-art MixMHCpred predictor and identifying optimal model architectures and training parameters.

Development of the Hidden Markov Model classification ensemble

Training dataset generation

Experimental data. MHC class II peptide data used to train the models were obtained from the Immune Epitope Database (IEDB) [10]. Peptides were categorized as binders based on IEDB quantitative binding data, specifically those with IC50 ≤ 500 nM or binding quality annotations such as *positive*, *positive-high*, or *positive-intermediate*. Peptides not meeting these criteria were categorized as non-binders. All protein sequences were encoded using standard 20-letter amino acid notation (ACDEFGHIKLMNPQRSTVWY), allowing direct computational processing. To ensure robust evaluation we applied four-fold cross validation [11], splitting the dataset into four equal parts. In each fold, three subsets were used for training and one for validation, resulting in four independent training and evaluation cycles per experiment.

Random sequence generation. Not all MHC alleles are equally well characterized. For some alleles, the amount of available binding data is limited [12], and existing datasets are often biased toward artificially mutated peptides derived from known binders [13]. As a result, experimental peptide datasets may not accurately represent the true background distribution of natural human peptides.

To address this, we generated an additional dataset consisting of random peptide sequences sampled from the human proteome, using data from the GENCODE database [14]. These peptides represent the natural background distribution of human sequences that could realistically occur in vivo. Given that MHC molecules bind fewer than 1 % of possible peptides [1], this random peptide set not only offers a biologically plausible background but also serves as a reasonable approximation of a negative dataset for model training and evaluation.

HMM Models training procedure

Following the training procedure outlined in a previous study [5], we train HMM by iteratively updating its three key components:

1) transition probability matrix, which defines the probabilities of transitions between hidden states;
2) emission probability matrix, which describes the probabilities of amino acid occurrences in each state;
3) initial state probability vector, which determines the probability of a sequence starting from a specific state.

We consider two standard algorithms for HMM training and optimization: the Viterbi [15] and Baum–Welch [16] algorithms. The Viterbi algorithm updates model parameters based on the single most probable state path for each sequence, preserving the interpretability and uniqueness of individual states. In contrast, the Baum–Welch algorithm averages over all possible state paths, which can dilute motif specificity. To prioritize interpretability and enable detailed analysis of binding core alignments and sequence motifs, we employ the Viterbi algorithm in this work.

A key contribution of our approach is the introduction of dedicated non-binder models, trained using negative peptide examples. To our knowledge, this represents a novel direction in MHC-peptide modeling. For each MHC allele, a single predictor consists of two separate HMMs: one trained on confirmed binder peptides, and another on non-binders. The non-binder models can either be trained using experimentally validated negative data from IEDB or initialized using amino acid frequencies derived from randomly sampled peptides from the human proteome.

Training parameters

Model architecture. We evaluated two architectures of models in this study. The first, referred to as the base model, consists of two loop states (representing peptide start and end fragments) and a central body of 9 states corresponding to the peptide binding core (Fig. 1, *a*). The second architecture, referred to as the **extended model**, introduces an additional set of 9 states which replace loop states (Fig. 1, *b*). Both include 9 states for the binding core (C_1, …, C_9). However, in the base model, the peptide terminal amino acids transition into loop regions (T_1, T_2), whereas the extended model includes extended set of states representing terminal amino acids (T_1, …, T_{18}). This design enables the model to capture not only the amino

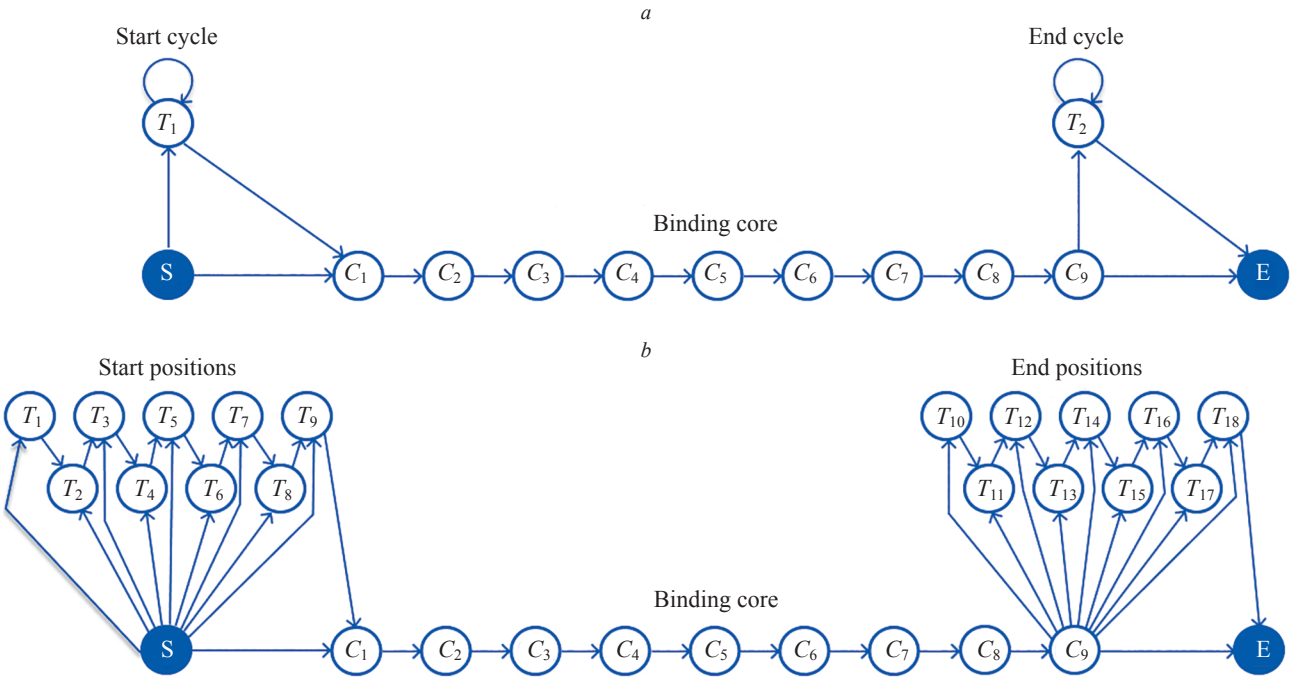

Fig. 1. Architecture of HMM: "Base model" with 2 loop states (T_1 and T_2) (*a*); Extended architecture with 9 start and end states (T_1 … T_{18}) (*b*)

acid distributions within the binding core, but also the properties of the flanking regions (known as N-terminal and C-terminal) of the peptide, which do not directly participate in binding. Each emission probability in the initial states was initialized with a random value drawn from a uniform distribution over the standard 20-letter amino acid alphabet.

Anchor state parameter. Anchor states are positions within the model that play a key role in binding prediction, typically corresponding to conserved residues within the binding core. In these states, high emission probabilities are assigned to a limited set of amino acids that occur most frequently during training. To improve model specificity, we restricted evaluated probabilities to the top four to 9 most common amino acids, as suggested in previous work [17]. This selective focus allows the model to emphasize biologically relevant sequence features while minimizing noise. To evaluate the impact of this parameter, we trained models using both architectures with anchor state values ranging from four to 9 and assessed their effect on classification performance.

Building an ensemble of models

To avoid the risk of suboptimal convergence — a known limitation of the Viterbi algorithm commonly used in HMM training — we implemented an ensemble-based strategy to improve the prediction stability. For each fold in the cross-validation procedure, 50 independent training runs were performed with different random initializations. We selected the top 70 % based on their validation performance. For each peptide, we then computed the final log-probability score by taking the median of its scores across this selected ensemble. This was done separately for both the binder and non-binder models, ensuring that the resulting scores are robust to the randomness of individual training runs.

Using these median log-probabilities, we computed the final classification score — the estimated probability that a peptide is a binder — using the following conditional probability formula:

$$Score = \frac{\exp\left(\frac{\ln P_b}{L}\right)}{\left(\exp\left(\frac{\ln P_b}{L}\right) + \exp\left(\frac{\ln P_{nb}}{L}\right)\right)},$$

where P_b — probability to generate peptide for binder model; P_{nb} — probability to generate peptide for non-binder model; L — the length of peptide.

Statistical evaluation of model performance

To assess performance differences during hyperparameter optimization of our model, we used DeLong's nonparametric test for comparing correlated Receiver Operating Characteristic (ROC) curves [18]. The test was implemented via the delong_test function from the MLstatkit Python package [19]. We evaluated significance between different configurations by comparing their predicted probabilities for each peptide against ground truth labels from testing datasets.

It was applied in the following comparisons:

— negative examples dataset choice — all combinations of anchor state parameters and architectures were compared between models trained on confirmed non-binders; randomly initialized negative examples;

— model architecture comparisons — all combinations of anchor state parameters and negative datasets were tested between the base and extended architectures.

For benchmarking against the state-of-the-art MixMHX2pred tool, we performed a paired t-test on ROC Curve (ROC AUC) scores across all data splits, comparing

our best-performing model with MixMHX2pred. The test was implemented using SciPy's ttest_rel function [20].

Results

Experiments Design

To evaluate the effectiveness of our method, we conducted the following experiments.

1. The anchor state parameter tuning. We evaluated the performance of base models using different values of the anchor state parameters to identify the optimal setting for each MHC allele.
2. Effect of non-binder dataset choice. We assessed the impact of incorporating experimentally confirmed non-binders into training by comparing ensembles built with two types of non-binder models. This allowed us to determine the most suitable negative dataset for each allele.
3. Architecture comparison. We compared ensemble performance between the base and extended model architectures to assess how model design affects prediction accuracy.
4. Overall, method evaluation. We benchmarked our HMM-based classification approach against the state-of-the-art MixMHC2pred tool, using identical validation data to ensure a fair performance comparison.

Anchor state parameter tuning

Fig. 2 presents boxplots of ROC AUC scores for the base architecture across different alleles, with varying values of the anchor state parameter. For most of the alleles we indicated non-decreasing trend. As the value of the parameter increased, the model performance remained the same or increased too. However, for HLA-DRB4*01:01 allele there was an upward trend in ROC AUC as the anchor state increases, although variability remains high. This suggests that a universal optimal value cannot be determined, and instead, the anchor state parameter must be fine-tuned individually for each allele to achieve the best predictive performance.

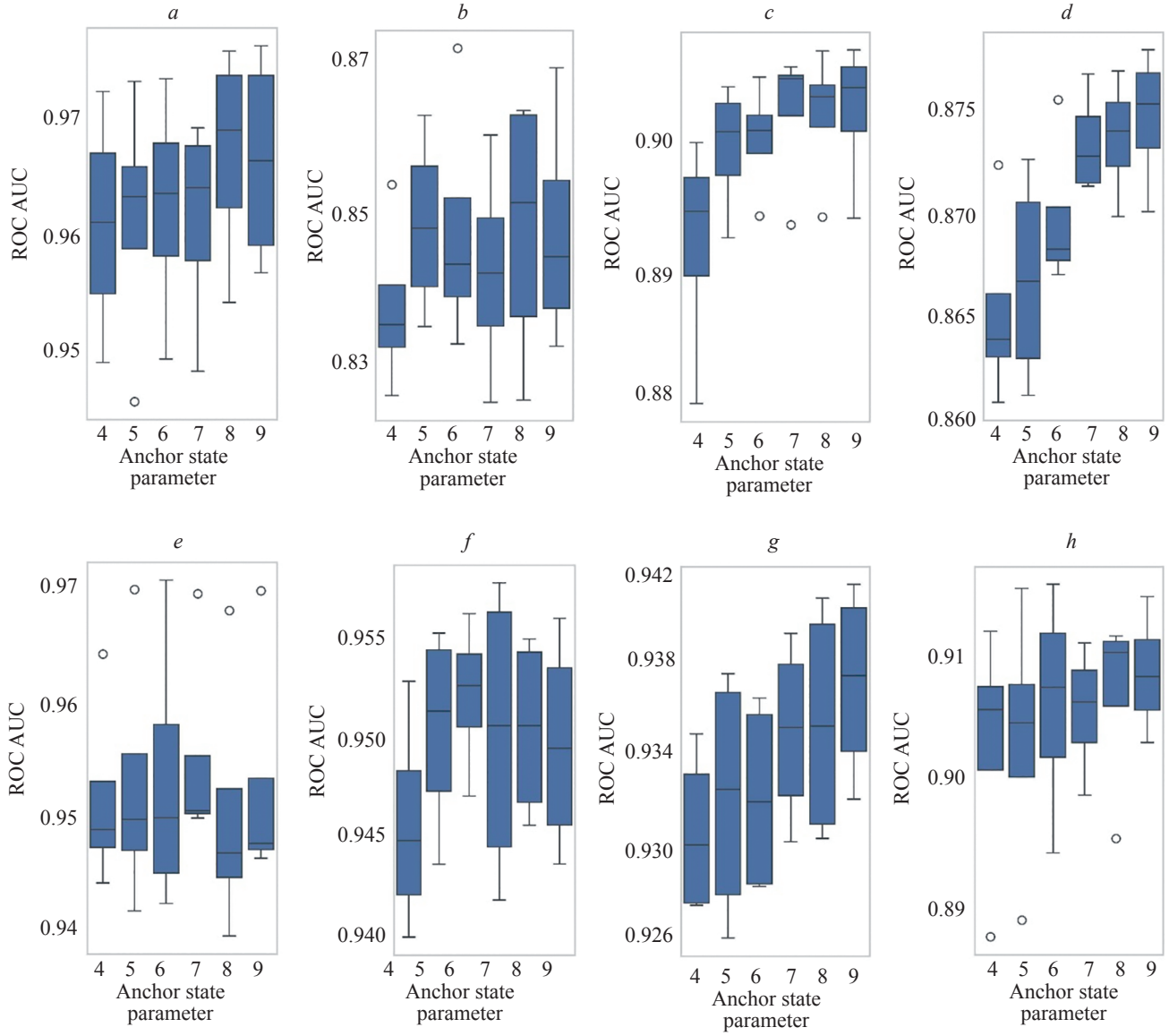

Fig. 2. ROC AUC score for base architecture for different values of anchor state parameter: HLA-DRB3*02:02 (*a*); HLA-DRB4*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB1*03:01 (*g*); HLA-DRB1*11:01 (*h*)

Negative examples dataset choice

In this experiment, we compared the impact of training and scoring with different types of negative datasets on the performance of different models. Specifically, five out of eight models using the base architecture and six out of eight models with the extended architecture performed better when using non-binders as the background distribution. Incorporating non-binders as negative dataset during training led to an improvement of ROC AUC scores (Fig. 3). Demonstrated performance enhancement appears to depend on the characteristics of the non-binders used in training. When the non-binders exhibited a more consistent and normal pattern — particularly with peptides of length 15 — the models achieved better performance. However, when the non-binders resembled random sequences, as observed in the HLA-DRB1*15:01 and HLA-DRB1*03:01 alleles, performance decreased, and random scoring yielded better results. This suggests that carefully selecting non-binders, rather than using completely random sequences, helps the HMMs learn more relevant distinctions between binders and non-binders, ultimately leading to improved predictive performance.

The length distribution of non-binder peptides (Fig. 5) further supports this observation, as it highlights variations in non-binder quality. A more uniform length distribution (e.g., predominantly 15-mers) correlates with improved model performance, whereas fragmented or highly variable distributions — indicative of less reliable non-binders — may contribute to reduced predictive accuracy.

Comparing model architectures

Among the eight evaluated models, the extended architecture exhibited significantly improved performance in two cases (HLA-DRB1*15:01 and HLA-DRB1*03:01 alleles) (Fig. 6). Interpretability of the model can be accessed via graph alignments for both base and extended models (Fig. 7, *a*, *b*). In this graph edges represent transitions between states of the models (with respective probabilities depicted as weights for edges) and nodes represent a single state of the models, which correspond to a single position within a peptide. To highlight the importance of the given amino acid we transformed the probability to information value of the amino acid given the emission probabilities of the state (the bigger the letter — the higher importance) (Fig. 7, *c*). Both the baseline and extended models accurately identified the binding core (Fig. 7, *a*, *b*); however, the extended architecture provided enhanced interpretability due to its more detailed terminal states. The optimal parameter combination — comprising anchor state configuration, scoring function (comparison to non-binders versus random peptides), and model architecture — favored the second approach in more cases (Fig. 6). This suggests that the selected configuration may exhibit greater sensitivity to variations in other parameters.

Comparing the best HMM models to the state-of-the-art MixMHX2pred tool

Best-performing HMM models consistently outperform the state-of-the-art MixMHC2pred tool [21] in terms of ROC AUC score across multiple alleles. In particular, significant improvements are observed for HLA-DRB1*03:01, HLA-DRB1*07:01, HLA-DRB1*12:01, HLA-DRB1*15:01, and HLA-DRB3*02:02 (Fig. 8). Even in cases where the improvement is less pronounced, such as HLA-DRB1*11:01 and HLA-DRB3*01:01, our models perform comparably to MixMHC2pred.

Discussion

Building upon the methodology outlined in a previous study [5], we developed a rigorous HMMbased training and validation pipeline with a strong emphasis on model evaluation using the area under the receiver operating characteristic curve (ROC AUC) as the primary performance metric. A central enhancement in our approach was the incorporation of non-binding peptides as a dataset of negative examples during training. Specifically, we trained two separate HMMs: one on experimentally validated binding peptides and another on non-binding peptides, which represent a background distribution of the possible peptides space.

The inclusion of negative examples enabled a more nuanced estimation of conditional probabilities, allowing us to rank test peptides based on their relative likelihood of being true binders, and solve the machine learning classification problem. Strikingly the HMM-based approach led to a clear and consistent improvement in ROC AUC scores across multiple alleles, demonstrating the value of this dual-model approach for enhancing predictive power as seen from the comparison with the state-of-the art MixMHC2pred predictor.

These findings highlight the efficacy of HMMs in MHC-peptide binding prediction and underscore the importance of incorporating negative sampling and specialized anchor states to enhance discriminative power. Notably, we observed that the quality and representativeness of non-binding peptides significantly influence the performance of models trained on non-binders, suggesting that careful selection of negative examples is critical.

Future work could explore further refinement of models architectures and training procedures, particularly improving the quality and handling of non-binder datasets, as models tend to be dependent on the quality of this data.

Additionally, one open challenge is the ability of such models to generalize to alleles exhibiting multiple peptide-binding motifs. For these alleles, models considering a single motif show reduced predictive power [22]. Addressing this limitation may require adapting the current approach to a multiclass classification framework or incorporating unsupervised motif clustering strategies.

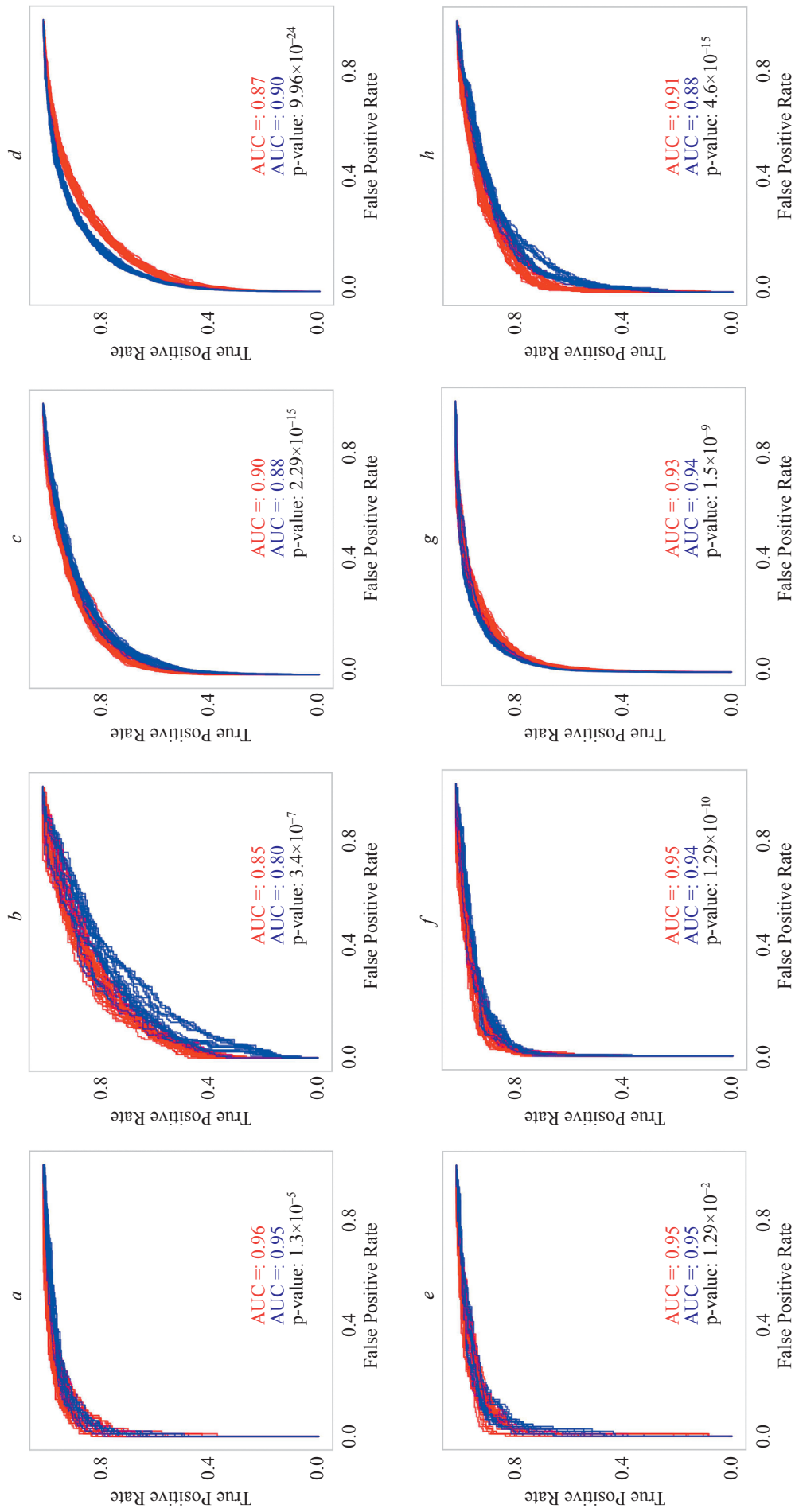

Fig. 3. ROC curves comparing non-binders (blue) vs. random (red) peptides scoring in base architecture models. P-values are used to assess the statistical significance of the difference in ROC AUC scores between the two types of test datasets: Non-binders (red) and Randoms (blue) calculated using T-test: HLA-DRB3*02:02 (*a*); HLA-DRB4*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB1*03:01 (*g*); HLA-DRB1*11:01 (*h*)

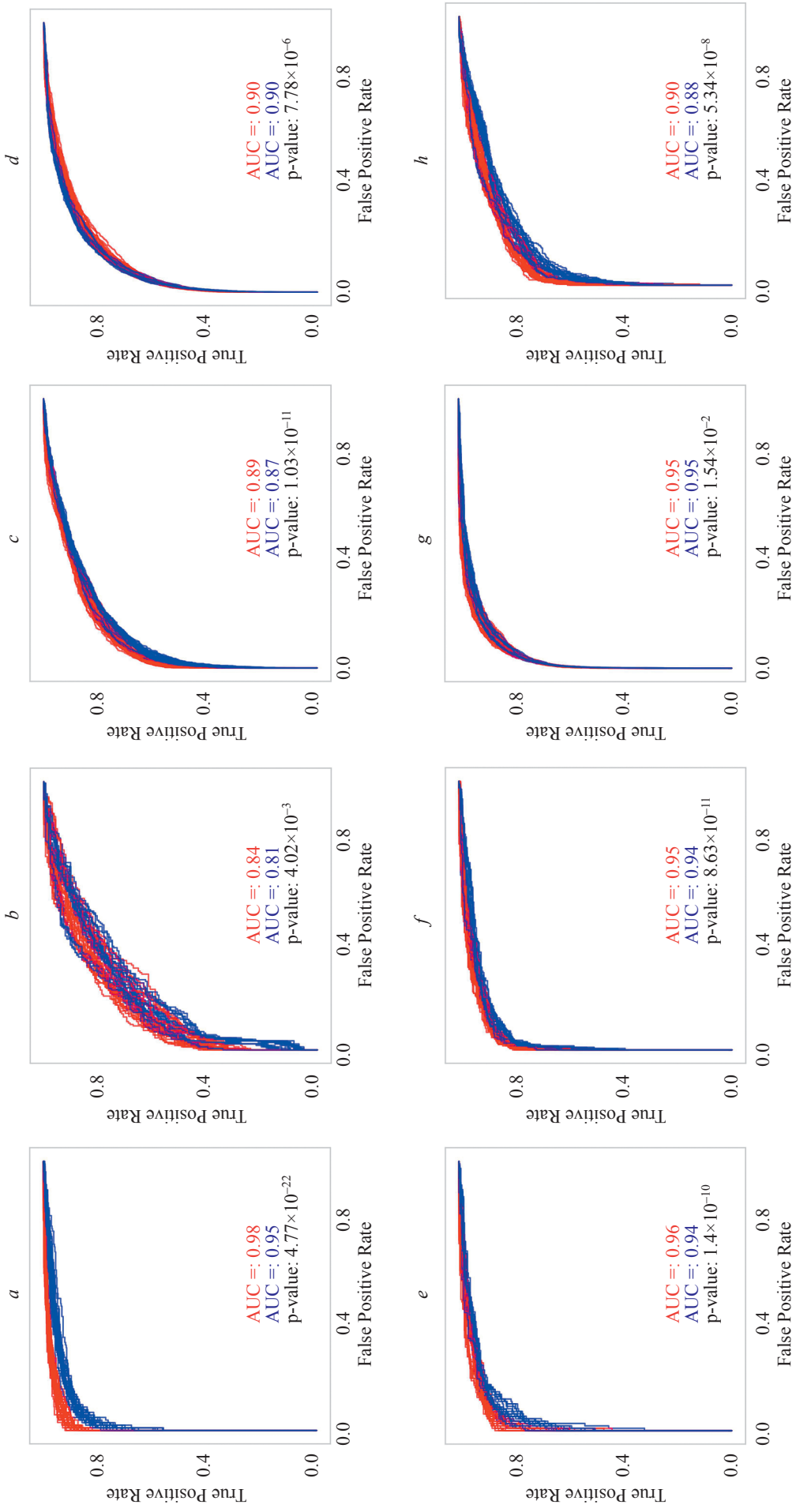

Fig. 4. ROC curves comparing non-binders (blue) vs. random (red) peptides scoring in extended architecture models. P-values are used to assess the statistical significance of the difference in ROC AUC scores between the two types of test datasets: Non-binders (red) and Randoms (blue) calculated using T-test: HLA-DRB3*02:02 (*a*); HLA-DRB4*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB1*03:01 (*g*); HLA-DRB1*11:01 (*h*)

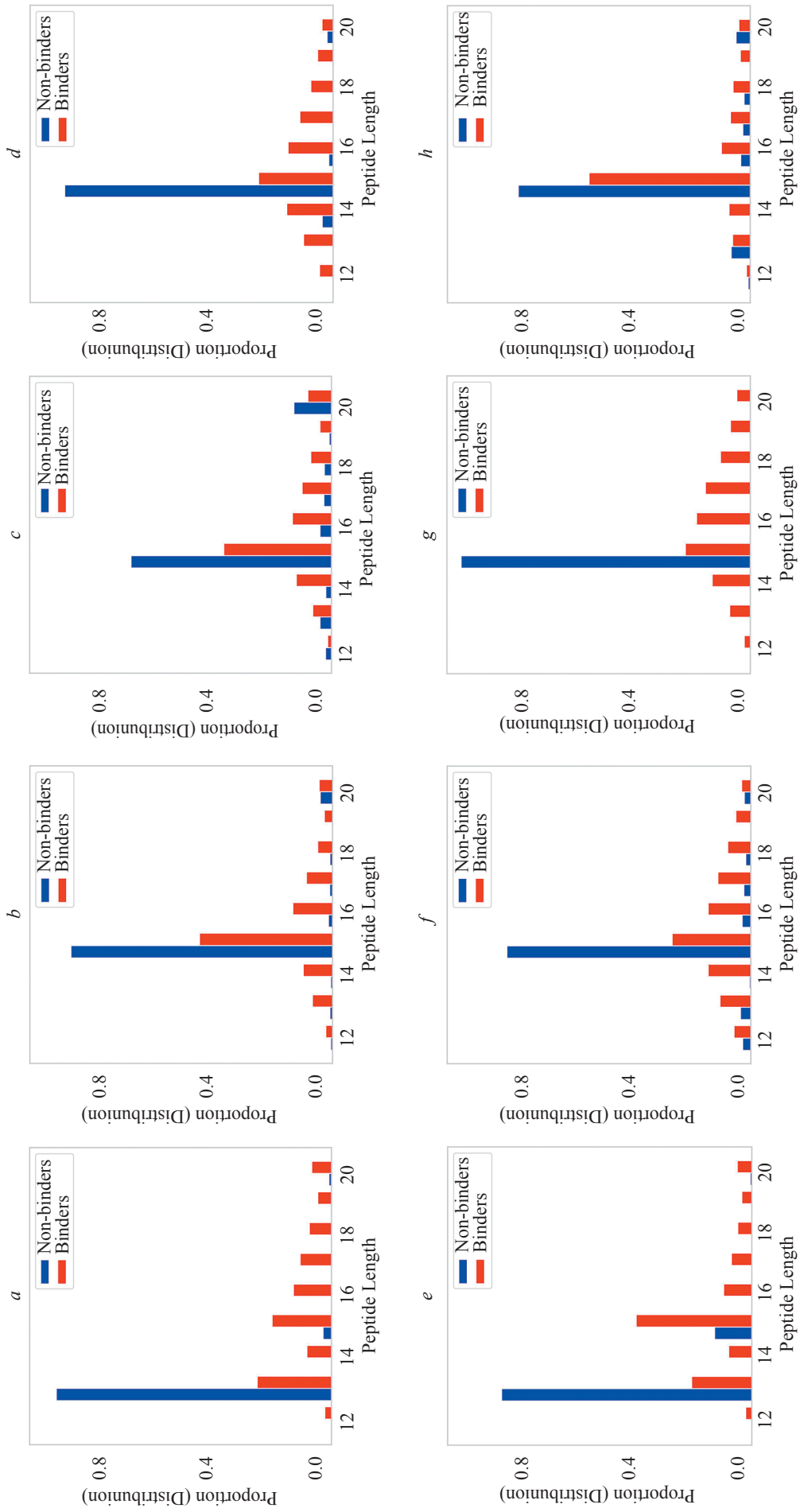

Fig. 5. Length distribution of non-binders. The predominance of 15-mer sequences suggests higher-quality non-binders, making them suitable for negative sampling when training HMMs: HLA-DRB1*03:01 (*a*); HLA-DRB1*07:01 (*b*); HLA-DRB1*11:01 (*c*); HLA-DRB1*12:01 (*d*); HLA-DRB1*15:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB3*02:02 (*g*); HLA-DRB4*01:01 (*h*)

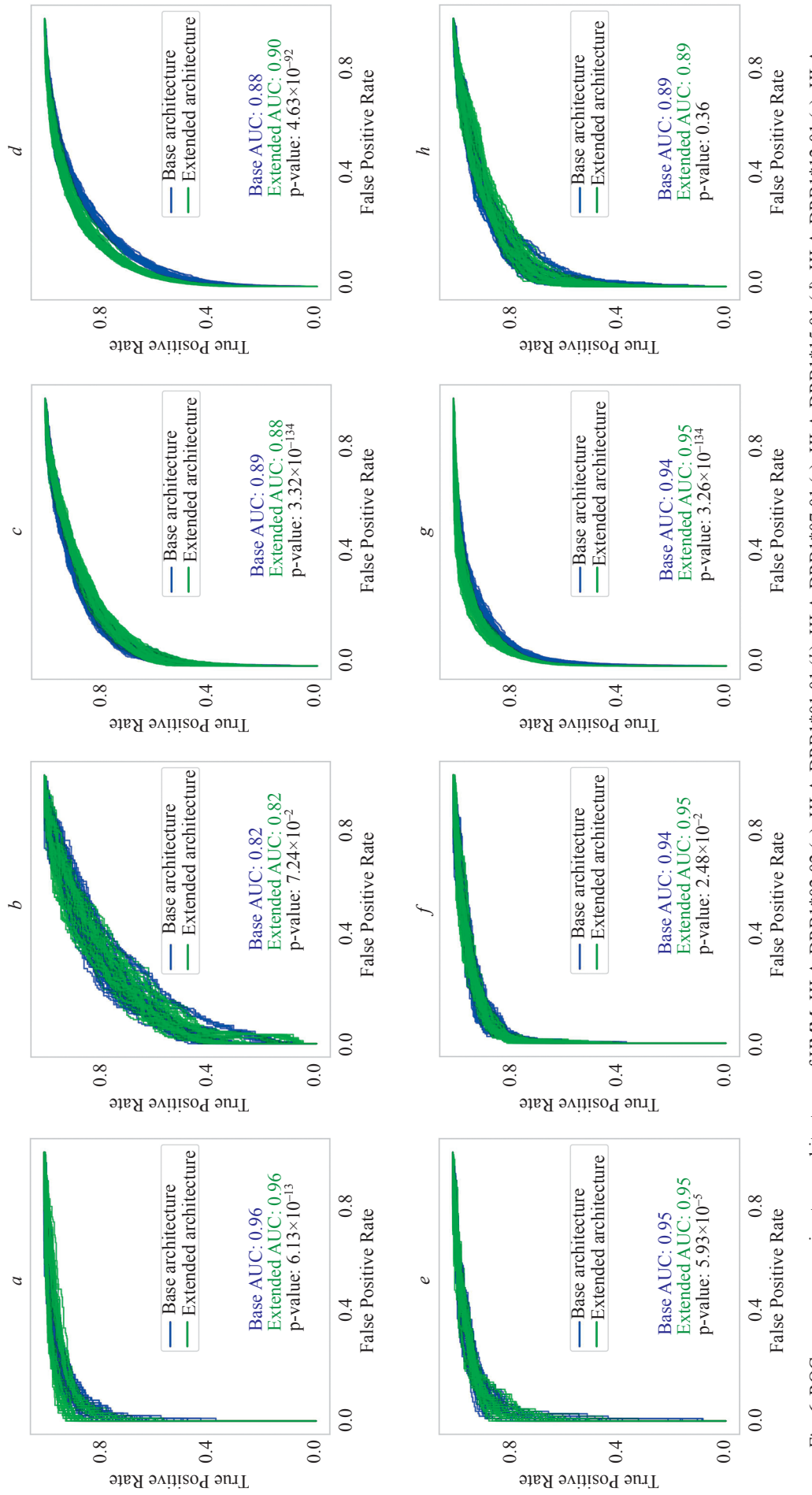

Fig. 6. ROC curves comparing two architectures of HMM: HLA-DRB1*02:02 (*a*); HLA-DRB1*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB3*03:01 (*g*); HLA-DRB4*11:01 (*h*)

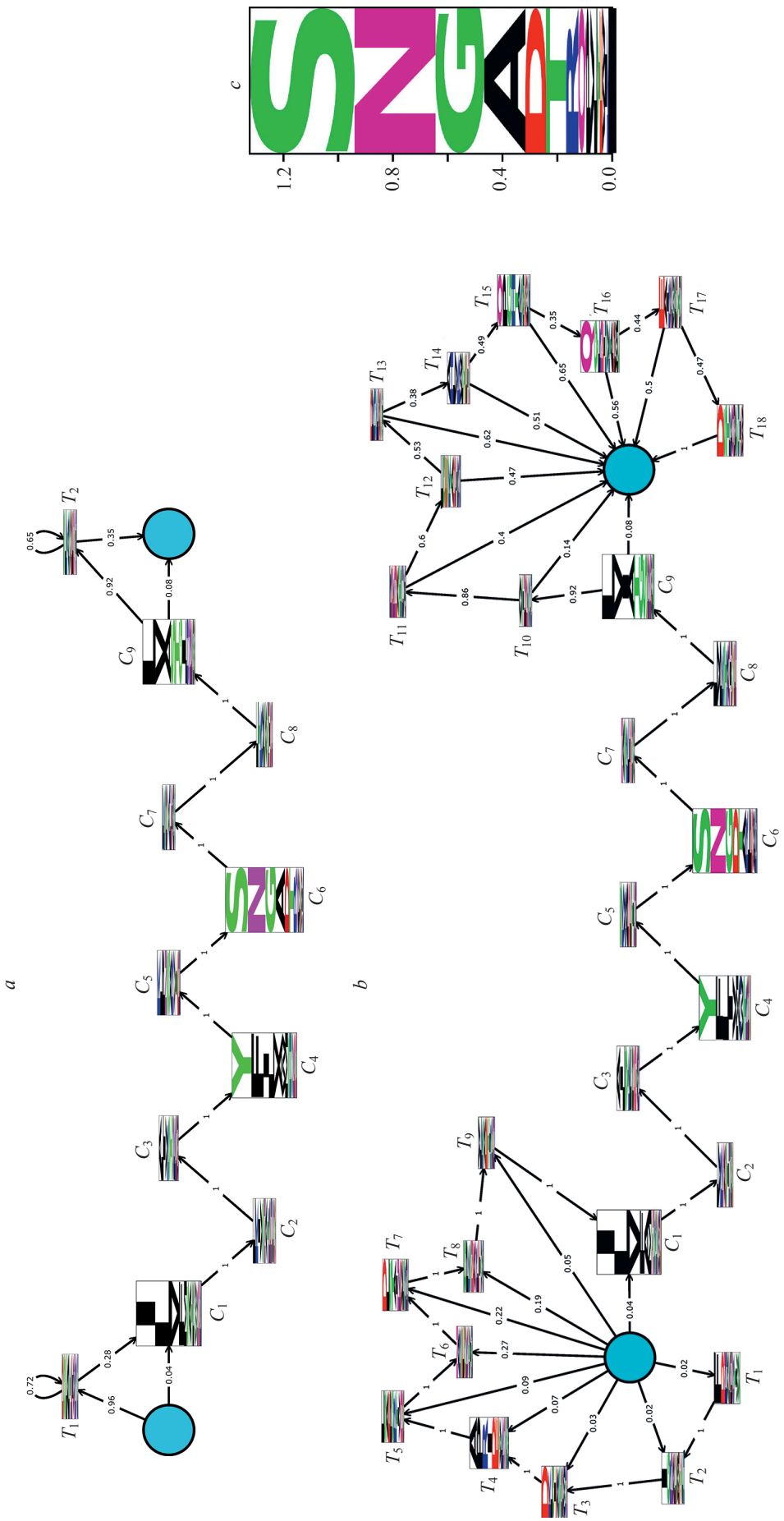

Fig. 7. The example of a state distribution plot with transition probabilities after training a model: Base architecture (*a*); Extended architecture (*b*); State distribution example represented as information content of amino acids (*c*)

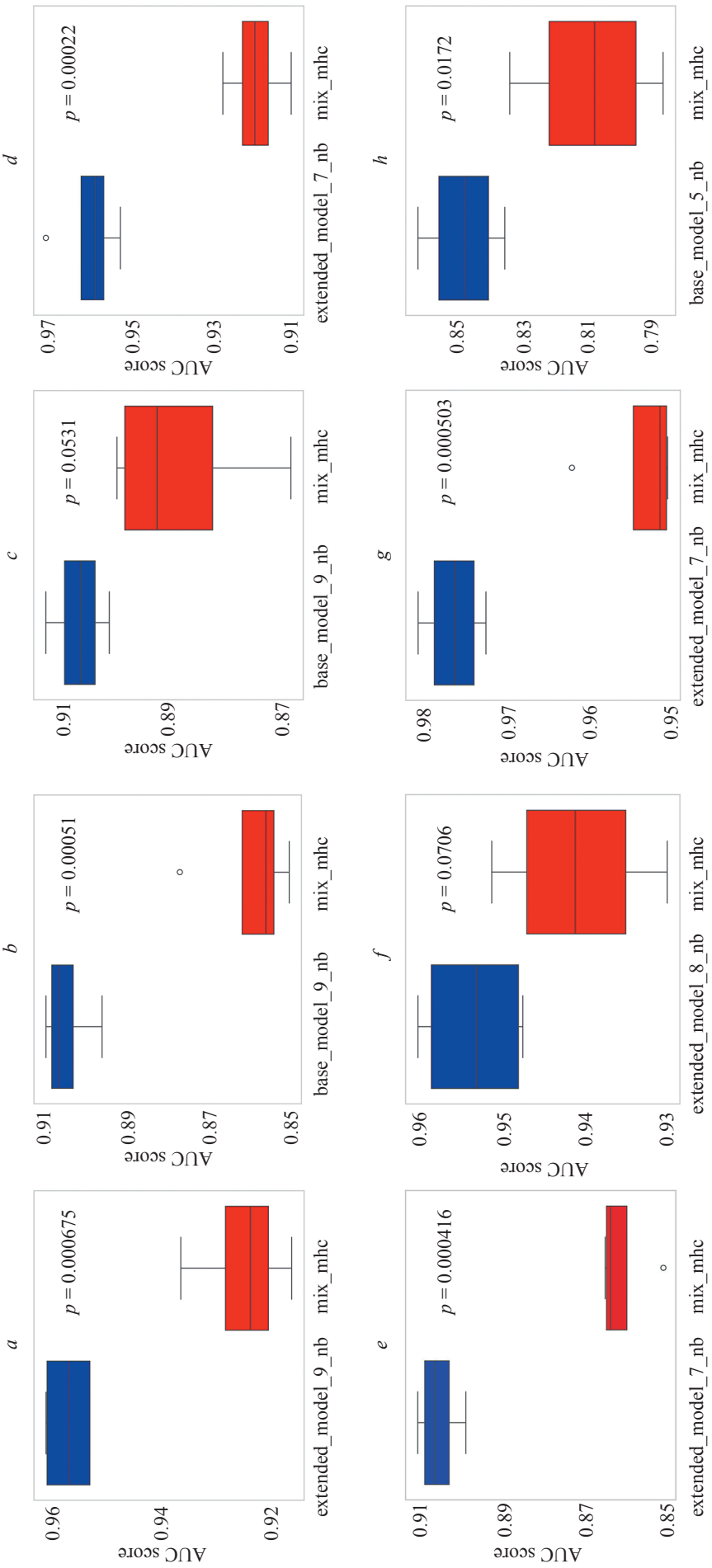

Fig. 8. Performance comparison of AUC score between fine-tuned HMM ensemble predictors with MixMHC2pred tool: HLA-DRB1*03:01 (*a*); HLA-DRB1*07:01 (*b*); HLA-DRB1*11:01 (*c*); HLA-DRB1*12:01 (*d*); HLA-DRB1*15:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB3*02:02 (*g*); HLA-DRB4*01:01 (*h*)

Conclusion

In this study, we demonstrated the potential of Hidden Markov Models (HMMs) to address the challenge of peptide to Major Histocompatibility Complex (MHC) binding affinity prediction. Our findings underscore the effectiveness of HMMs in predicting peptide to MHC binding affinity, offering a statistically grounded and interpretable alternative to existing methods. A key innovation was negative examples model introduction, in which separate HMMs were trained on binding and non-binding peptides. This dual-model strategy effectively captured the background distribution of proteomic sequences and led to statistically significant improvements in predictive performance, as measured by Receiver Operating Characteristic Curve (ROC AUC), compared to the current state-of-the-art, MixMHC2pred.

Our work also involved systematic optimization of anchor state parameters within the HMM architecture, guided by a conditional probability scoring framework. Performance was evaluated through extensive cross-validation and comparative analysis against models trained on either experimentally confirmed non-binders or naturally occurring peptide sequences.

References

1. Corradin G. Antigen processing and presentation. *Immunology Letters*, 1990, vol. 25, no. 1–3, pp. 11–13. https://doi.org/10.1016/0165-2478(90)90082-2
2. Abualrous E.T., Sticht J., Freund C. Major histocompatibility complex (MHC) class I and class II proteins: impact of polymorphism on antigen presentation. *Current Opinion in Immunology*, 2021, vol. 70, pp. 95–104. https://doi.org/10.1016/j.coi.2021.04.009
3. Waldman A.D., Fritz J.M., Lenardo M.J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology*, 2020, vol. 20, no. 11, pp. 651–668. https://doi.org/10.1038/s41577-020-0306-5
4. Wieczorek M., Abualrous E.T., Sticht J., Alvaro-Benito M., Stolzenberg S., Noé F., Freund C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Frontiers in Immunology*, 2017, vol. 8, pp. 292. https://doi.org/10.3389/fimmu.2017.00292
5. Kleverov D.A., Shalyto A.A., Artyomov M.N. A method for constructing interpretable hidden Markov models for the task of identifying binding cores in sequences. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 5, pp. 989–1000. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-5-989-1000
6. Gutiérrez S.E., Esteban E.N., Lützelschwab C.M., Juliarena M.A. Major histocompatibility complex-associated resistance to infectious diseases: the case of bovine leukemia virus infection. *Trends and Advances in Veterinary Genetics*, 2017, pp. 101–126. https://doi.org/10.5772/intechopen.68416
7. Eddy S.R. Profile hidden Markov models. *Bioinformatics*, 1998, vol. 14, no. 9, pp. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755
8. Alspach E., Lussier D.M., Miceli A.P., Kizhvatov I., DuPage M., Luoma A.M., et al. MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature*, 2019, vol. 574, no. 7780, pp. 696–701. https://doi.org/10.1038/s41586-019-1671-8
9. Kim M.W., Gao W., Lichti C.F., Gu X., Dykstra T., Cao J., et al. Endogenous self-peptides guard immune privilege of the central nervous system. *Nature*, 2025, vol. 637, no. 8044, pp. 176–183. https://doi.org/10.1038/s41586-024-08279-y
10. Vita R., Blazeska N., Marrama D., Duesing S., Bennett J., Greenbaum J., et al. The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Research*, 2025, vol. 53, no. D1, pp. D436–D443. https://doi.org/10.1093/nar/gkae1092
11. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009, 767 p. https://doi.org/10.1007/978-0-387-84858-7
12. Capietto A.H., Jhunjhunwala S., Pollock S.B., Lupardus P., Wong J., Hänsch L., et al. Mutation position is an important determinant for predicting cancer neoantigens. *Journal of Experimental Medicine*, 2020, vol. 217, no. 4, pp. e20190179. https://doi.org/10.1084/jem.20190179
13. Rahman K.S., Chowdhury E.U., Sachse K., Kaltenboeck B. Inadequate reference datasets biased toward short non-epitopes confound B-cell epitope prediction. *The Journal of Biological Chemistry*, 2016, vol. 291, no. 28, pp. 14585–14599. https://doi.org/10.1074/jbc.M116.729020

Литература

1. Corradin G. Antigen processing and presentation // Immunology Letters. 1990. V. 25. N 1–3. P. 11–13. https://doi.org/10.1016/0165-2478(90)90082-2
2. Abualrous E.T., Sticht J., Freund C. Major histocompatibility complex (MHC) class I and class II proteins: impact of polymorphism on antigen presentation // Current Opinion in Immunology. 2021. V. 70. P. 95–104. https://doi.org/10.1016/j.coi.2021.04.009
3. Waldman A.D., Fritz J.M., Lenardo M.J. A guide to cancer immunotherapy: from T cell basic science to clinical practice // Nature Reviews Immunology. 2020. V. 20. N 11. P. 651–668. https://doi.org/10.1038/s41577-020-0306-5
4. Wieczorek M., Abualrous E.T., Sticht J., Alvaro-Benito M., Stolzenberg S., Noé F., Freund C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation // Frontiers in Immunology. 2017. V. 8. P. 292. https://doi.org/10.3389/fimmu.2017.00292
5. Клеверов Д.А., Шалыто А.А., Артемов М. Метод построения интерпретируемых скрытых марковских моделей для задачи поиска связываемых участков пептидов в последовательностях белков // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 5. С. 989–1000. https://doi.org/10.17586/2226-1494-2023-23-5-989-1000
6. Gutiérrez S.E., Esteban E.N., Lützelschwab C.M., Juliarena M.A. Major histocompatibility complex-associated resistance to infectious diseases: the case of bovine leukemia virus infection // Trends and Advances in Veterinary Genetics. 2017. P. 101–126. https://doi.org/10.5772/intechopen.68416
7. Eddy S.R. Profile hidden Markov models // Bioinformatics. 1998. V. 14. N 9. P. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755
8. Alspach E., Lussier D.M., Miceli A.P., Kizhvatov I., DuPage M., Luoma A.M., et al. MHC-II neoantigens shape tumour immunity and response to immunotherapy // Nature. 2019. V. 574. N 7780. P. 696–701. https://doi.org/10.1038/s41586-019-1671-8
9. Kim M.W., Gao W., Lichti C.F., Gu X., Dykstra T., Cao J., et al. Endogenous self-peptides guard immune privilege of the central nervous system // Nature. 2025. V. 637. N 8044. P. 176–183. https://doi.org/10.1038/s41586-024-08279-y
10. Vita R., Blazeska N., Marrama D., Duesing S., Bennett J., Greenbaum J., et al. The Immune Epitope Database (IEDB): 2024 update // Nucleic Acids Research. 2025. V. 53. N D1. P. D436–D443. https://doi.org/10.1093/nar/gkae1092
11. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. 767 p. https://doi.org/10.1007/978-0-387-84858-7
12. Capietto A.H., Jhunjhunwala S., Pollock S.B., Lupardus P., Wong J., Hänsch L., et al. Mutation position is an important determinant for predicting cancer neoantigens // Journal of Experimental Medicine. 2020. V. 217. N 4. P. e20190179. https://doi.org/10.1084/jem.20190179
13. Rahman K.S., Chowdhury E.U., Sachse K., Kaltenboeck B. Inadequate reference datasets biased toward short non-epitopes confound B-cell epitope prediction // The Journal of Biological Chemistry. 2016. V. 291. N 28. P. 14585–14599. https://doi.org/10.1074/jbc.M116.729020

14. Mudge J.M., Carbonell-Sala S., Diekhans M., Martinez J.G., Hunt T., Jungreis I., et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research*, 2025, vol. 53, no. D1, pp. D966–D975. https://doi.org/10.1093/nar/gkae1078
15. Forney G.D. The viterbi algorithm. *Proceedings of the IEEE*, 1973, vol. 61, no. 3, pp. 268–278. https://doi.org/10.1109/proc.1973.9030
16. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, vol. 77, no. 2, pp. 257–286. https://doi.org/10.1109/5.18626
17. Nielsen M., Lundegaard C., Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 2007, vol. 8, pp. 238. https://doi.org/10.1186/1471-2105-8-238
18. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 1988, vol. 44, no. 3, pp. 837–845. https://doi.org/10.2307/2531595
19. Sun X., Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 2014, vol. 21, no. 11, pp. 1389–1393. https://doi.org/10.1109/LSP.2014.2337313
20. Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 2020, vol. 17, no. 3, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2
21. Racle J., Michaux J., Rockinger G.A., Arnaud M., Bobisse S., Chong C., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nature Biotechnology*, 2019, vol. 37, no. 11, pp. 1283–1286. https://doi.org/10.1038/s41587-019-0289-6
22. Koşaloğlu-Yalçın Z., Sidney J., Chronister W., Peters B., Sette A. Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs recognized by HLA-DQ2.5. *Immunology*, 2021, vol. 162, no. 2, pp. 235–247. https://doi.org/10.1111/imm.13279

14. Mudge J.M., Carbonell-Sala S., Diekhans M., Martinez J.G., Hunt T., Jungreis I., et al. GENCODE 2025: reference gene annotation for human and mouse // Nucleic Acids Research. 2025. V. 53. N D1. P. D966–D975. https://doi.org/10.1093/nar/gkae1078
15. Forney G.D. The viterbi algorithm // Proceedings of the IEEE. 1973. V. 61. N 3. P. 268–278. https://doi.org/10.1109/proc.1973.9030
16. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. 1989. V. 77. N 2. P. 257–286. https://doi.org/10.1109/5.18626
17. Nielsen M., Lundegaard C., Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method // BMC Bioinformatics. 2007. V. 8. P. 238. https://doi.org/10.1186/1471-2105-8-238
18. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach // Biometrics. 1988. V. 44. N 3. P. 837–845. https://doi.org/10.2307/2531595
19. Sun X., Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves // IEEE Signal Processing Letters. 2014. V. 21. N 11. P. 1389–1393. https://doi.org/10.1109/LSP.2014.2337313
20. Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python // Nature Methods. 2020. V. 17. N 3. P. 261–272. https://doi.org/10.1038/s41592-019-0686-2
21. Racle J., Michaux J., Rockinger G.A., Arnaud M., Bobisse S., Chong C., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes // Nature Biotechnology. 2019. V. 37. N 11. P. 1283–1286. https://doi.org/10.1038/s41587-019-0289-6
22. Koşaloğlu-Yalçın Z., Sidney J., Chronister W., Peters B., Sette A. Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs recognized by HLA-DQ2.5 // Immunology. 2021. V. 162. N 2. P. 235–247. https://doi.org/10.1111/imm.13279

Authors

Valeriia A. Polezhaeva — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0001-7469-7440, polezhaevalera@yandex.ru
Denis A. Kleverov — Visiting Researcher, Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 631110, USA, sc 58741254400, https://orcid.org/0009-0002-1362-486X, denklewer@gmail.com

Anatoly A. Shalyto — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru

Maxim Artyomov — PhD (Chemistry), Full Professor, Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 631110, USA; Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 9242717500, https://orcid.org/0000-0002-1133-4212, martyomov@pathology.wustl.edu

Авторы

Полежаева Валерия Александровна — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0001-7469-7440, polezhaevalera@yandex.ru
Клеверов Денис Анатольевич — приглашенный научный сотрудник, Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США, sc 58741254400, https://orcid.org/0009-0002-1362-486X, denklewer@gmail.com
Шалыто Анатолий Абрамович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru
Артемов Максим — PhD, химические науки, профессор (исследователь), профессор, Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США; профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 9242717500, https://orcid.org/0000-0002-1133-4212, martyomov@pathology.wustl.edu

Received 06.05.2025
Approved after reviewing 29.08.2025
Accepted 20.09.2025

Статья поступила в редакцию 06.05.2025
Одобрена после рецензирования 29.08.2025
Принята к печати 20.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-902-909

Vector embeddings compression using clustering with the ensemble of oblivious decision trees and separate centroids storage

Nikita A. Tomilov✉

ITMO University, Saint Petersburg, 197101, Russian Federation

programmer174@icloud.com✉, https://orcid.org/0000-0001-9325-0356

Abstract

The modern approach to search textual and multimodal data in large collections involves the transformation of the documents into vector embeddings. To store these embeddings efficiently different approaches could be used, such as quantization, which results in loss of precision and reduction of search accuracy. Previously, a method was proposed that reduces the loss of precision during quantization. In that method, clustering of the embeddings with *k*-Means algorithm is performed, then a bias, or delta, being the difference between the cluster centroid and vector embedding, is computed, and then only this delta is quantized. In this article a modification of that method is proposed, with a different clustering algorithm, the ensemble of Oblivious Decision Trees. The essence of the method lies in training an ensemble of binary Oblivious Decision Trees. This ensemble is used to compute a hash for each of the original vectors, and the vectors with the same hash are considered as belonging to the same cluster. In case when the resulting cluster count is too big or too small for the dataset, a reclustering process is also performed. Each cluster is then stored using two different files: the first file contains the per-vector biases, or deltas, and the second file contains identifiers and the positions of the data in the first file. The data in the first file is quantized and then compressed with a general-purpose compression algorithm. The usage of Oblivious Decision Trees allows us to reduce the size of the storage compared to the same storage organization with *k*-Means clustering. The proposed clustering method was tested on Fashion-MNIST-784-Euclidean and NYT-256-angular dataset against the *k*-Means clustering. The proposed method demonstrates a better compression quality compared to clustering via *k*-Means, demonstrating up to 4.7 % less storage size for NF4 quantization for Brotli compression algorithm. For other compression algorithms the storage size reduction is less noticeable. However, the proposed clustering algorithm provides a bigger error value compared to *k*-Means, up to 16 % in the worst-case scenario. Compared to Parquet, the proposed clustering method demonstrates a lesser error value for the Fashion-MNIST-784-Euclidean dataset when using quantizations FP8 and NF4. For the NYT-256-angular dataset, compared to Parquet, the proposed method allows better compression for all tested quantization types. These results suggest that the proposed clustering method can be utilized not only for the nearest neighbor search applications, but also for compression tasks, when the increase in the quantization error can be ignored.

Keywords

vector representations, embeddings, oblivious decision tree, clustering, compression of vector embeddings

For citation: Tomilov N.A. Vector embeddings compression using clustering with the ensemble of oblivious decision trees and separate centroids storage. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 902–909. doi: 10.17586/2226-1494-2025-25-5-902-909

УДК 004.021

Сжатие векторных представлений с использованием кластеризации с помощью ансамбля небрежных решающих деревьев и раздельного хранения центроидов

Никита Андреевич Томилов✉

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

programmer174@icloud.com✉, https://orcid.org/0000-0001-9325-0356

Аннотация

Введение. Современные подходы к поиску текстовых и мультимодальных данных в больших коллекциях предполагают преобразование документов в векторные представления. Для эффективного хранения этих векторов применяется квантизация, которая снижает точность представления и, как следствие, ухудшает качество поиска. Ранее был предложен метод, уменьшающий потери точности при квантизации, в котором векторы кластеризуются с помощью алгоритма *k*-средних, вычисляется смещение, или дельта, являющееся разницей между центроидом кластера и исходным вектором, после чего квантуется только смещение. В настоящей работе предлагается модификация этого метода, использующая другой алгоритм кластеризации — ансамбль небрежных решающих деревьев. **Метод.** Разработанный метод основывается на обучении ансамбля бинарных небрежных решающих деревьев. Этот ансамбль используется для вычисления хэша каждого исходного векторного представления, после чего векторные представления с одинаковым хэшем рассматриваются как принадлежащие одному кластеру. В случае, если число результирующих кластеров слишком большое или слишком маленькое для используемого набора данных, производится процесс перекластеризации. Каждый кластер сохраняется в двух отдельных файлах: первый содержит смещения для каждого вектора, второй — идентификаторы и позиции данных в первом файле. Данные в первом файле подвергаются квантизации и затем сжимаются с помощью универсального алгоритма сжатия. **Основные результаты.** Предложенный метод кластеризации протестирован на наборах данных fashion-mnist-784-euclidean и NYT-256-angular и сравнивался с кластеризацией методом *k*-средних. Метод показал лучшее качество сжатия, демонстрируя до 4,7 % меньшее занимаемое дисковое пространство при использовании квантизации NF4 и алгоритма сжатия Brotli. Для других алгоритмов сжатия увеличение пространства оказалось менее значительным. Однако представленный алгоритм кластеризации демонстрирует большее значение показателя ошибки квантизации по сравнению с методом *k*-средних, минимум до 16 %. По сравнению с форматом Parquet разработанный метод кластеризации продемонстрировал меньший показатель ошибки для набора данных fashion-mnist-784-euclidean при использовании квантизаций FP8 и NF4. Для набора данных NYT-256-angular по сравнению с Parquet предложенный метод при использовании алгоритма сжатия Brotli позволяет добиться лучшего качества сжатия для всех протестированных типов квантизации. **Обсуждение.** Полученные результаты свидетельствуют о том, что разработанный метод кластеризации может быть использован не только в задачах поиска ближайших соседей, но и в задачах сжатия данных, когда увеличением ошибки квантизации можно пренебречь.

Ключевые слова

векторные представления, эмбеддинг, небрежное решающее дерево, кластеризация, сжатие векторных представлений

Ссылка для цитирования: Томилов Н.А. Сжатие векторных представлений с использованием кластеризации с помощью ансамбля небрежных решающих деревьев и раздельного хранения центроидов // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 902–909 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-902-909

Introduction

One of the effective strategies for organizing search in large collections of textual and multimodal data involves transforming documents into vector embeddings, being sequences of floating-point numbers that encode semantic content [1]. In this framework, machine learning techniques are used to construct such embeddings for both the documents and the search queries. Efficient search is then performed via nearest-neighbor retrieval in the resulting high-dimensional space, using precomputed indices [2]. Storing uncompressed embeddings in high-precision formats such as FP32 requires significant memory and disk space. To address this, quantization methods are often employed to convert values from FP32 to lower-precision formats such as FP16 or FP8 [3], enabling substantial space savings. However, this comes at the cost of reduced accuracy, which may negatively affect downstream applications, including further machine learning tasks [4].

In [5] authors explored vector embeddings compression with the help of *k*-Means clustering, where we cluster embeddings and separately store the centroids and per-vector biases (deltas). Each embedding can thus be represented as a pair of the index of its closest centroid and the delta vector between the embedding and the centroid. These components can be compressed with different levels of precision, for instance, storing centroids in FP32 to preserve accuracy while applying quantization to the deltas using FP16 or FP8. In [6] authors explored the clustering algorithm via the ensemble of Oblivious Trees, which proved to be beneficial for improving the approximate nearest neighbor search.

In scope of this research, we combine these previous works and present the usage of the clustering algorithm using the ensemble of Oblivious Trees for compressing various kinds of vector embeddings via the method of storing centroids and deltas separately [5]. Our hypothesis suggests that usage of this clustering algorithm will improve the compression compared to *k*-Means [7] clustering.

Methods

Description of the clustering method used

In our previous study [6] a clustering algorithm was proposed using an ensemble of binary Oblivious Decision Trees (ODTs) [8]. To manage computational costs, each tree is trained on a random subsample, having size N, taken from the original dataset of size N_{dataset}, such that $N = N_{\text{dataset}} \times TrainRatio$, where $TrainRatio$ is between 0.1 and 1. We transform each original vector embedding into multiple low-dimensional variants by randomly selecting subsets of its components. This results in M so-called subvectors, having dimension d_m, per each of the original vector embedding. These subvectors are essentially different views of the data, each used to train a separate tree. A mapping table is maintained to associate the indices of the projected components with their positions in the original vector, allowing reconstruction during ensemble creation.

Each ODT partitions the input space hierarchically. At each tree level j, we select a component K_j and a threshold X_j splitting the data into two groups depending on whether the value of the selected component is less than or equal to X_j, or bigger than X_j. The rule by which the component and threshold is selected is called a partitioning rule and is a hyper-parameter of the algorithm, and in scope of our research we find the values that perform the partition to minimize the variance of distances between each vector embedding and the average vector embedding for that group.

The same splitting rule is applied across all branches at a given level, which is a defining property of oblivious trees. This process repeats until the tree reaches the desired depth *depth*. Due to the nature of the splits, it is possible for some groups to be empty at deeper levels, which is expected and does not affect the structure.

After training, each tree encodes a vector as a binary string of zeros and ones, depending on the comparison outcome on each tree level with respect to the corresponding threshold X_j. This binary string is effectively a hash for each vector. This hash is called *SH* and has length of *depth*. With M trees, each vector receives M such hashes which are then concatenated into a single composite hash string *TH*. This final hash effectively clusters vectors, as those with the same composite hash are placed into the same group, or cluster.

It is worth mentioning that this method does not have precise control over the resulting number of clusters and can lead to a huge number of them, in the worst-case scenario being comparable to the amount of vector embeddings in the dataset. Since the maximum resulting cluster number is $2^{M\,depth}$, having 5 trees of depth 4 can result in 1,048,576 clusters. This is inappropriate for datasets having comparable or lower numbers of vectors. To mitigate this problem, we introduce so-called reclustering with the parameter F and threshold T. If the maximum resulting number of clusters R_1 is less than T, we perform the clustering as normal, and then split each of the resulting clusters up to F sub-clusters with the different clustering algorithm, such as *k*-Means, so that the maximum number of resulting sub-clusters is $R_1 \times F$. We call this process reclustering up. If the maximum resulting number of clusters R_2 is more than T, we pick the centroids of the resulting clusters, and then cluster those centroids using different clustering algorithms, such as *k*-Means, with such parameters, that the maximum number of resulting sub-clusters is R_2/F. We then use these new centroids to cluster the original dataset via *k*-Nearest Neighbors algorithm. We call this process reclustering down.

The example resulting clusters count for $T = 4{,}000$ are presented in Table 1.

As a result of this reclustering we gain more precise control over the maximum number of clusters. It is worth mentioning that such reclustering is only necessary for relatively small datasets and is not necessary for such combinations of M and *depth* so that the resulting maximum number of clusters is significantly less than the number of vectors in the source dataset.

This research has four important differences compared to [6]. Firstly, in this research we focus on data compression instead of approximate nearest neighbor search. Secondly, we were able to optimize the training phase of the ODTs, so that now we can train the ODTs on more embeddings, specifically 15 % of the original dataset ($TrainRatio = 0.15$), instead of the first five thousand vectors of the dataset as it was in the previous research. Also, we focus only on binary trees, since ternary trees proved to create too many clusters for compression purposes, and introduce reclustering to mitigate the issue with having too many clusters.

Description of the storing method used

To validate the compression quality of the Oblivious Tree clustering algorithm compared to *k*-Means clustering algorithm, we use the storage method [5] with some key differences. The essence of the method is to store

Table 1. Reclustering for $T = 4{,}000$

Number of trees M	Depth of each tree (*depth*)	Maximum number of clusters without reclustering	Maximum number of clusters after reclustering up with $F = 10$	Maximum number of clusters after reclustering down with $F = 30$
4	2	256	2,560	—
4	3	4,096	—	137
4	4	65,536	—	2,185
5	2	1,024	10,240	—
5	3	32,768	—	1,092
5	4	1,048,576	—	34,953

clustered embeddings in binary files, so that each cluster is represented by two files. The first binary file, called page file, contains entries for a given document index, where an entry contains the metadata, such as segment index, followed by the serialized floating-point values array. This array is generated by subtracting the original vector embedding value from the centroid of the respective cluster. During serialization, this floating-point values array can be quantized from their original FP32 representation via any scalar quantization, such as FP8 [9], to reduce space.

The second binary file, called the page index file, serves as an index, mapping each primary identifier to its byte offset in the page file, indicating where the entry for a given document index is. All offsets in this index file, as well as indexes in the page file, are encoded using variable length encoding to reduce overhead. After the page is finalized, meaning there are no more vector embeddings for this page, the page file is compressed using any general-purpose compression algorithm, to further reduce space.

To convert the dataset to the proposed storage, it is necessary to cluster the vector embeddings and generate the page files and page index files as described above. To retrieve the vector embedding for a given document index, it is necessary to traverse page index files to find the precise page that contains the necessary embedding, then decompress the necessary page file, and fetch the entry for the retrieved offset. After that, it is necessary to deserialize the floating-point array, containing the vector embedding delta, and sum it with the respective vector centroid to restore the original vector embedding.

It is worth mentioning that this storage organization method can be used with different clustering algorithms, which is why we can use it to compare Oblivious Tree clustering against *k*-Means clustering in the context of vector embeddings compression while having the storage layer similar for both clustering algorithms.

The key difference between the method presented in this research and what was explored in [5] is that in this research the whole page file is compressed, while in [5] the compression was used for each vector embedding individually. Working with the full page will improve compression as there will be more data to find repetitive patterns to utilize during compression. However, it also increases the vector retrieval time, which is out of the scope of the research. On top of that, NF4 [10] is also considered as another quantization method that can pack a floating-point value to just 4 bits.

Storage implementation

To perform the experiments, we implemented the proposed storage solution in Kotlin programming language. To serialize the entries, we used an Apache Avro library. Each of the entries contains a byte array that was constructed by converting the delta value directly from the FP32 representation to specified quantization format to one of the target quantizations, namely FP16, FP8 and NF4, reducing the storage space from four bytes per the component of delta representation to two, one or half of a byte respectively. The centroids of the clusters are stored separately, serialized via the built-in Java serialization mechanism, in their original FP32 representation. As an additional metadata, the entry contains a segment index that is always zero for the datasets used to conduct the experiment.

To convert the datasets to store them using the proposed storage implementation, we perform clustering of the dataset via the *k*-Means algorithm, and via the ensemble of the Oblivious Trees. Then, vectors belonging to each of the clusters get converted into the page files. After that, the page files are compressed using the following compression algorithms: LZMA, Zstandard, Brotli [11]. To use these algorithms in Kotlin, the Apache Commons Compress[1], zstd-jni[2] and brotli4j[3] libraries were used. For Zstandard and Brotli, the compression level value was set to 22 and 11, respectively. For LZMA, the default parameters provided by the library were used.

Results

Experiment setup

To verify the compression quality when using the Oblivious Tree clustering algorithm, we selected the same two datasets as before, NYT-256-angular[4] and Fashion-MNIST-784-Euclidean [12], taken from the popular ANN-Benchmarks [13] collection. These datasets, as well as others from the same collection, are often used to benchmark average nearest neighbor vector search algorithms. They differ from each other in the way the data was originally created.

These datasets were clustered using the following clustering algorithms:

1. *k*-Means clustering, with the maximum number of clusters K being an arbitrary number between 1,000 and 15,000.
2. ODT ensemble with the following parameters:
   a. $M = 5$;
   b. $d_m = 196$ for the Fashion-MNIST dataset, 32 for the NYT dataset (meaning, 25 % from the source dataset dimensionality for Fashion-MNIST and 12.5 % for NYT);
   c. $depth = 2, 4$;
   d. $TrainRatio = 0.15$ %;
   e. $F = 10$ for $depth = 2$ and 30 for $depth = 4$;
   f. $T = 4{,}000$.

For both clustering algorithms we then calculate the delta values and quantize them to their FP16, FP8 and NF4 representations, after which the data is compressed using LZMA, ZStd and Brotli algorithms. The centroids are stored in their original FP32 representation and are not compressed. The exact FP8 quantization type used is E3M4, meaning 3 bits for the exponent, 4 bits for the mantissa, and one sign bit.

[1] Commons Compress — Overview. URL: https://commons.apache.org/proper/commons-compress/index.html (accessed: 11.05.2025).

[2] luben/zstd-jni: JNI binding for Zstd. URL: https://github.com/luben/zstd-jni (accessed: 11.05.2025).

[3] hyperxpro/Brotli4j: Brotli4j provides Brotli compression and decompression for Java. URL: https://github.com/hyperxpro/Brotli4j (accessed: 11.05.2025).

[4] Newman D. Bag of Words // UCI Machine Learning Repository. 2008. URL: https://archive.ics.uci.edu/dataset/164/bag+of+words (accessed: 11.05.2025). DOI: 10.24432/C5ZG6P

Upon creating the storage files, we measure the storage size represented as the sum of the sizes of each of the individual files generated for each of these clustering methods, against the error value, represented as Euclidean or Angular distance, depending on the dataset used, between the retrieved vector embedding and the original vector embedding. Assuming the vector embedding of size n, retrieved from the storage and having the component values (vc_0, …, vc_n), as vc, and the original vector embedding of the same size n, having component values (vs_0, …, vs_n), as vs, the error value e, being the distance metric, can be calculated using the following equations:

— Euclidean distance

$$e = \sqrt{\sum_{k=0}^{n} (vc_k - vs_k)^2},$$

— Angular distance

$$e = \frac{\sum_{k=0}^{n} vc_k vs_k}{\sqrt{\sum_{k=0}^{n} vc_k^2} \sqrt{\sum_{k=0}^{n} vs_k^2}}.$$

The lower the metric, the closer the retrieved vector embedding to the original one, meaning the lesser the error caused by lossy compression via quantization.

We then select only the best two parameter combinations for the k-Means and Oblivious Tree clustering, judging by the lowest possible storage size, and provide the values for all combinations of compression algorithm and quantization type for these value combinations. To provide a baseline, we also store the clustered values in Parquet [14] format in both lossless (float array) and lossy (byte arrays of quantized float arrays) for the specified compression algorithms supported by Parquet with compression levels configured similarly to the proposed solution.

The test setup has the following specifications: AMD Ryzen 7 7700X 8C16T; 64 GB RAM; NVMe WD SN850X 2TB; OS Ubuntu 22.04; OpenJDK 22; a tool for comparing vector search algorithms, developed in previous research [15], using Java Microbenchmark Harness [16].

Experiment results

The results for the Fashion-MNIST-784-Euclidean dataset are presented in Table 2. The vector embeddings in this dataset are grayscale 28 by 28-pixel images of clothing having the values from 0 to 255, with a lot of values being zero indicating the black pixels, making this a sparse dataset. The size of the original dataset, excluding the data that was not part of the experiment like precomputed closest neighbors, in HDF5 file without any compression and in the original FP32 format is 179.44 MB. This value is also present in the table as the Original HDF5 file row. The notation for the ODT clustering corresponds to $M/d_m/depth/F/\{up, down\}$, where M, d_m, $depth$, F are the parameters specified above, and *up* or *down* indicate reclustering direction.

Table 2. Storage size and error value (as Euclidean distance) for Fashion-MNIST-784-Euclidean dataset, for different compression algorithms and quantizations

Storage	Quantization	Storage size for compression algorithm, MB				Error value
		Without compression	LZMA	ZStd	Brotli	
Original HDF5 file	FP32	179.44	—	—	—	—
Parquet	FP32	34.81	—	26.78	27.42	—
	FP16	91.43	—	32.88	34.11	0.02 ± 0.01
	FP8	46.56	—	18.17	18.65	85.23 ± 29.23
	NF4	24.12	—	9.53	10.02	380.86 ± 211.83
k-Means, K = 1,000	FP32	182.48	29.18	32.59	30.83	—
	FP16	93.58	44.42	47.17	44.24	0.25 ± 0.20
	FP8	48.73	20.44	21.29	20.39	55.04 ± 13.29
	NF4	26.25	12.21	12.63	12.24	155.90 ± 50.78
k-Means, K = 10,000	FP32	182.49	29.44	33.14	30.80	—
	FP16	94.79	45.54	48.10	45.17	0.24 ± 0.18
	FP8	49.93	21.66	22.39	21.57	53.36 ± 14.26
	NF4	27.50	13.53	13.78	13.47	152.80 ± 47.02
ODT, 5/192/4/30/down	FP32	182.55	29.18	32.22	30.91	—
	FP16	92.88	43.48	47.20	43.36	0.29 ± 0.13
	FP8	48.02	19.90	21.12	19.80	60.91 ± 8.99
	NF4	25.58	11.63	12.14	11.66	169.57 ± 32.90
ODT, 5/192/2/10/up	FP32	183.48	31.10	34.84	31.90	—
	FP16	93.85	45.00	49.25	44.47	0.24 ± 0.13
	FP8	48.98	20.77	21.98	20.69	54.41 ± 12.80
	NF4	26.54	12.74	13.11	12.57	151.71 ± 37.53

For this dataset, the Oblivious Tree clustering method demonstrates worse compression compared to *k*-Means for original FP32 vector embedding representations, but better compression when using delta storage with quantization. This can be seen for both selected parameter combinations for both LZMA and Brotli compression algorithms, and for the first parameter combination for ZStd compression algorithm. For example, for FP16 quantization and Brotli compression, the Oblivious Tree clustering requires 43.36 MB of size in the best-case scenario, compared to 44.24 MB for *k*-Means clustering. It means that Oblivious Tree clustering can use 0.88 MB less compared to *k*-Means clustering. Relative to the storage size of *k*-Means clustering this difference equals to around 2 % less storage size. For NF4 quantization and Brotli compression, the values are 12.24 and 11.66 MB respectively, which is 0.58 MB less storage size, or 4.7 % less relative to *k*-Means clustering.

This advantage is offset by the fact that Oblivious Tree clustering demonstrates bigger error value, showing bigger loss during quantization. For the combinations mentioned above, the average error value for ODTs is 0.04 and 14 more than for *k*-Means clustering, or 16 % and 9 % more relative to *k*-Means clustering, respectively. On top of that, for this dataset, when using original FP32 vector representations and without delta storage both clustering methods cannot demonstrate better compression compared to Parquet, which is able to effectively encode repeating zero values in the vector embeddings. This encoding is so effective that its lossless compression is better than the lossy compression for FP16 encoding for the proposed method and both clustering algorithms. Overall, using ODT clustering combined with delta storage cannot offer better compression compared to quantized Parquet storage, however it can demonstrate less error value for some of the quantizations. Compared to Parquet, using ODTs can lower the error value from around 85 and 380 for FP8 and NF4 quantizations, respectively, to around 61 and 170 for the parameter combinations demonstrating the least storage size.

The results for the NYT-256-angular dataset are presented in Table 3. The vector embeddings in this dataset are text embeddings of the NYT articles, having dimensionality 256 and all the values between –0.342 and 0.346, making it a dense dataset. As the angular distance is usually smaller, in the table the error value is multiplied by 1,000. The size of the original dataset, excluding the data that was not part of the experiment like precomputed closest neighbors, in HDF5 file without any compression and in the original FP32 format is 283.21 MB.

For this dataset, the results are comparable to the results of the previous dataset, represented in Table 2. The Oblivious Tree clustering method demonstrates better compression when using delta storage with quantization. For example, for FP8 quantization and Brotli compression,

Table 3. Storage size and error value (as angular distance) for NYT-256-angular dataset, for different compression algorithms and quantizations

Storage	Quantization	Storage size for compression algorithm, MB				Error value, multiplied by 1.000
		Without compression	LZMA	ZStd	Brotli	
Original HDF5 file	FP32	283.21	—	—	—	—
Parquet	FP32	289.82	—	244.85	239.05	—
	FP16	156.95	—	133.18	133.53	0.00 ± 0.00
	FP8	85.14	—	47.89	46.12	6.94 ± 0.89
	NF4	49.73	—	28.58	27.98	74.70 ± 12.58
k-Means, K = 1,000	FP32	298.27	239.04	239.20	237.14	—
	FP16	157.65	124.43	126.95	121.89	0.00 ± 0.00
	FP8	86.80	45.71	45.01	44.88	6.93 ± 1.49
	NF4	51.37	28.27	27.06	26.92	70.12 ± 14.35
k-Means, K = 10,000	FP32	298.44	240.24	239.73	238.29	—
	FP16	166.48	134.46	136.47	131.74	0.00 ± 0.00
	FP8	95.74	55.30	53.96	53.64	7.27 ± 4.24
	NF4	60.28	39.14	36.24	35.93	63.90 ± 34.65
ODT, 5/32/4/30/down	FP32	298.62	239.76	239.67	237.90	—
	FP16	157.18	124.55	125.03	121.82	0.00 ± 0.00
	FP8	86.32	45.19	46.94	44.00	7.44 ± 1.31
	NF4	50.87	28.68	28.32	26.18	73.27 ± 8.88
ODT, 5/32/2/10/up	FP32	303.23	244.36	244.34	242.29	—
	FP16	161.83	128.82	131.26	126.21	0.00 ± 0.00
	FP8	90.97	50.60	49.80	49.62	6.53 ± 1.13
	NF4	55.48	33.13	31.68	31.47	71.18 ± 15.58

the Oblivious Tree clustering requires 44 MB of size in the best-case scenario, compared to 44.88 MB for *k*-Means clustering, which is 0.88 MB less storage size or almost 2 % less relative to *k*-Means clustering. For NF4 quantization and Brotli compression, the values are 26.18 and 26.92 MB respectively, which is 0.74 MB less or is around 2.7 % less size relative to *k*-Means clustering. For FP16 compression the difference between compression algorithms is negligible. However, for this dataset both clustering methods demonstrate better compression than Parquet for the same compression algorithm.

The error value for the FP16 encoding for this dataset is almost zero, making FP16 a good choice for this dataset, allowing the storage size to significantly decrease without compromising the precision. As for more lossy quantizations, the results are also the same as for the previous dataset. The Oblivious Tree quantization also has a bigger error value. For the combinations mentioned above, the average error value for ODTs is 0.00051 and 0.00315 more than for *k*-Means clustering, or 7.4 % and 4.5 % more relative to *k*-Means clustering, respectively. Compared to Parquet, for this dataset ODT clustering demonstrates better compression with comparable error value. For NF4 quantization, ODT clustering achieves an error value of up to about 73, while for Parquet the value is around 75. For this quantization and Brotli compression, Parquet requires 27.98 MB of storage, which is 1.8 MB more than for ODT clustering, or 6.8 % more relative to ODT clustering. Overall, using ODT clustering in combination with delta storage can offer better compression compared to quantized Parquet storage with a comparable error value.

Conclusion

In this work, we proposed a storage solution for vector embeddings that combines oblivious decision tree clustering with delta encoding and quantization. Our experiments demonstrated that this method reduces storage requirements compared to *k*-Means clustering, with savings of up to 4.7 % for FP8 and NF4 quantization. We also showed its loss in precision compared to *k*-Means, which indicates that the method is most effective when storage efficiency is prioritized over absolute precision.

The main advantages of the method are the improved compression through delta quantization, the retention of centroids in FP32 precision, and the possibility of fine-tuning cluster counts via reclustering. Its limitations, namely greater retrieval time and significant training cost, are inherited from the underlying clustering algorithm and storage implementation. It is also worth mentioning that the reclustering step does not allow determining the target cluster solely based on the hash value of the embedding, negating the sharding benefits mentioned in [6]. It is possible to achieve sharding when reclustering up is being performed, however, in this case ODT can help determine only the target node, rather than target cluster of vector embeddings on the node, which would be possible without reclustering. We consider these trade-offs acceptable for applications where compact storage is critical such as large-scale embedding databases.

The experiment results highlight that while oblivious tree clustering is beneficial for compression tasks where an increase in quantization error is acceptable, for sparse datasets like Fashion-MNIST-784-Euclidean, Parquet format with lossless compression can outperform both ODT clustering and *k*-Means clustering methods when using original FP32 representations, due to its effectiveness in encoding repeating zero values, and shows better compression for lossy quantizations. For dense datasets like NYT-256-angular, both clustering methods demonstrate better compression than Parquet for the same compression algorithm and same quantization. Additionally, for the NYT-256-angular dataset, FP16 encoding yields an almost zero error value, making it a suitable choice for significant space reduction without compromising precision.

The introduced reclustering step allows for better control of the resulting number of clusters, improving upon one of the drawbacks of ODT clustering, at the cost of limiting the sharding possibilities. Other advantages and disadvantages of both the delta storage and ODT clustering remain unchanged.

These results suggest that the proposed clustering method can be utilized not only for the nearest neighbor search applications, as already explored in [6], but also for storage and compression of the vector embeddings for later use, such as machine learning tasks over the previously gathered embeddings, when the quantization error can be ignored. Further advancements in this area include experimenting with other partitioning rules to better group the vector embeddings, achieving better compression.

References

1. Grbovic M., Cheng H. Real-time personalization using embeddings for search ranking at airbnb. *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 311–320. https://doi.org/10.1145/3219819.3219885
2. Korn F., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z. Fast nearest neighbor search in medical image databases. *Proc. of the 22th International Conference on Very Large Data Bases*, 1996, pp. 215–226. https://doi.org/10.5555/645922.673493
3. Zhou W., Lu Y., Li H., Tian Q. Scalar quantization for large scale image search. *Proc. of the 20th ACM International Conference on Multimedia*, 2012, pp. 169–178. https://doi.org/10.1145/2393347.2393377
4. Zhang J., Yang J., Yuen H. Training with low-precision embedding tables. *Systems for Machine Learning Workshop at NeurIPS*, 2018, vol. 2018.

Литература

1. Grbovic M., Cheng H. Real-time personalization using embeddings for search ranking at airbnb // Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018. P. 311–320. https://doi.org/10.1145/3219819.3219885
2. Korn F., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z. Fast nearest neighbor search in medical image databases // Proc. of the 22th International Conference on Very Large Data Bases. 1996. P. 215–226. https://doi.org/10.5555/645922.673493
3. Zhou W., Lu Y., Li H., Tian Q. Scalar quantization for large scale image search // Proc. of the 20th ACM International Conference on Multimedia. 2012. P. 169–178. https://doi.org/10.1145/2393347.2393377
4. Zhang J., Yang J., Yuen H. Training with low-precision embedding tables // Systems for Machine Learning Workshop at NeurIPS. 2018. V. 2018.

5. *Tomilov N.A., Turov V.P., Babayants A.A., Platonov A.V.* A method of storing vector data in compressed form using clustering. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 1, pp. 112–117. (in Russian). https://doi.org/10.17586/2226-1494-2024-24-1-112-117
6. Tomilov N.A., Turov V.P., Babayants A.A., Platonov A.V. Vector search using method of clustering using ensemble of oblivious trees. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 2. pp. 339–344. https://doi.org/10.17586/2226-1494-2025-25-2-339-344
7. Kanungo T., Mount D., Netanyahu N., Piatko Ch., Silverman R., Wu A. The analysis of a simple k-means clustering algorithm. *Proc. of the 16th Annual Symposium on Computational Geometry*, 2000, pp. 100–109. https://doi.org/10.1145/336154.336189
8. Lou Y., Obukhov M. BDT: gradient boosted decision tables for high accuracy and scoring efficiency. *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1893–1901. https://doi.org/10.1145/3097983.3098175
9. Kuzmin A., van Baalen M., Ren Y., Nagel M., Peters J., Blankevoort T. Fp8 quantization: The power of the exponent. *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 14651–14662.
10. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QloRA: efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 1–5.
11. Alakuijala J., Farruggia A., Ferragina P., Kliuchnikov E., Obryk R., Szabadka Z., Vandevenne L. Brotli: A general-purpose data compressor. *ACM Transactions on Information Systems (TOIS)*, 2018, vol. 37, no. 1, pp. 1–30. https://doi.org/10.1145/3231935
12. Xiao H., Rasul K., Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, 2017, arXiv:1708.07747. https://doi.org/10.48550/arXiv.1708.07747
13. Aumüller M., Bernhardsson E., Faithfull A. ANN-Benchmarks: a benchmarking tool for approximate nearest neighbor algorithms. *Lecture Notes in Computer Science*, 2017, vol. 10609, pp. 34–49. https://doi.org/10.1007/978-3-319-68474-1_3
14. Zeng X., Hui Y., Shen J., Pavlo A., McKinney W., Zhang H. An empirical evaluation of columnar storage formats. *Proc. of the VLDB Endowment*, 2023, vol. 17, no. 2, pp. 148–161. https://doi.org/10.14778/3626292.3626298
15. Turov V.P., Tomilov N.A., Babayants A.A. Developing a tool to compare vector search algorithms. *Proc. of the 11th Congress of Young Scientists*, 2022, vol. 1, pp. 446–450. (in Russian)
16. Laaber C., Leitner P. An evaluation of open-source software microbenchmark suites for continuous performance assessment. *Proc. of the 15th International Conference on Mining Software Repositories (MSR '18)*, 2018, pp. 119–130. https://doi.org/10.1145/3196398.3196407

5. Томилов Н.А., Туров В.П., Бабаянц А.А., Платонов А.В. Метод хранения векторных представлений в сжатом виде с применением кластеризации // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 1. С. 112–117. https://doi.org/10.17586/2226-1494-2024-24-1-112-117
6. Tomilov N.A., Turov V.P., Babayants A.A., Platonov A.V. Vector search using method of clustering using ensemble of oblivious trees // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2025. V. 25. N 2. P. 339–344. https://doi.org/10.17586/2226-1494-2025-25-2-339-344
7. Kanungo T., Mount D., Netanyahu N., Piatko Ch., Silverman R., Wu A. The analysis of a simple k-means clustering algorithm // Proc. of the 16th Annual Symposium on Computational Geometry. 2000. P. 100–109. https://doi.org/10.1145/336154.336189
8. Lou Y., Obukhov M. BDT: gradient boosted decision tables for high accuracy and scoring efficiency // Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017. P. 1893–1901. https://doi.org/10.1145/3097983.3098175
9. Kuzmin A., van Baalen M., Ren Y., Nagel M., Peters J., Blankevoort T. Fp8 quantization: The power of the exponent // Advances in Neural Information Processing Systems. 2022. V. 35. P. 1–10.
10. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QloRA: efficient finetuning of quantized LLMs // Advances in Neural Information Processing Systems. 2023. V. 36. P. 1–5.
11. Alakuijala J., Farruggia A., Ferragina P., Kliuchnikov E., Obryk R., Szabadka Z., Vandevenne L. Brotli: A general-purpose data compressor // ACM Transactions on Information Systems (TOIS). 2018. V. 37. N 1. P. 1–30. https://doi.org/10.1145/3231935
12. Xiao H., Rasul K., Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms // arXiv. 2017. arXiv:1708.07747. https://doi.org/10.48550/arXiv.1708.07747
13. Aumüller M., Bernhardsson E., Faithfull A. ANN-Benchmarks: a benchmarking tool for approximate nearest neighbor algorithms // Lecture Notes in Computer Science. 2017. V. 10609. P. 34–49. https://doi.org/10.1007/978-3-319-68474-1_3
14. Zeng X., Hui Y., Shen J., Pavlo A., McKinney W., Zhang H. An empirical evaluation of columnar storage formats // Proc. of the VLDB Endowment. 2023. V. 17. N 2. P. 148–161. https://doi.org/10.14778/3626292.3626298
15. Туров В.П., Томилов Н.А., Бабаянц А.А. Разработка инструмента сравнения алгоритмов векторного поиска // XI Конгресс молодых учёных: сборник научных трудов. СПб: федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО». 2022. Т. 1. С. 446–450.
16. Laaber C., Leitner P. An evaluation of open-source software microbenchmark suites for continuous performance assessment // Proc. of the 15th International Conference on Mining Software Repositories (MSR '18). 2018. P. 119–130. https://doi.org/10.1145/3196398.3196407

Author

Nikita A. Tomilov — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57225127284, https://orcid.org/0000-0001-9325-0356, programmer174@icloud.com

Автор

Томилов Никита Андреевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57225127284, https://orcid.org/0000-0001-9325-0356, programmer174@icloud.com

Received 11.05.2025
Approved after reviewing 29.08.2025
Accepted 21.09.2025

Статья поступила в редакцию 11.05.2025
Одобрена после рецензирования 29.08.2025
Принята к печати 21.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-910-922

Enhanced detection of denial-of-service attacks in Kubernetes: a multi-framework machine learning approach integrating node and application metrics

Ghadeer Darwesh[1], Jaafar Hammoud[2], Alisa A. Vorobeva[3]✉

[1,2,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] ghadeerdarwesh32@gmail.com, https://orcid.org/0000-0003-1116-9410
[2] hammoudgj@gmail.com, https://orcid.org/0000-0002-2033-0838
[3] vorobeva@itmo.ru✉, https://orcid.org/0000-0001-6691-6167

Abstract

The widespread adoption of Kubernetes as a platform for orchestrating containerized applications has heightened the need for effective security mechanisms, particularly to counter Denial-of-Service (DoS) attacks. This article proposes an approach to DoS attack detection based on two key components the use of comprehensive metrics and the application of ensemble Machine Learning models. The approach involves the collection and analysis of comprehensive metrics from node-level (CPU, memory) and application-level (network activity, file descriptors) data from containers running on various frameworks (Flask, Django, FastAPI, Node.js, Golang). To implement this approach, a dataset containing 49,990 instances of network activity, characterized by 28 features (comprehensive metrics), was created. Statistical analysis (Student's t-test, Pearson correlation) identified the metrics most relevant for attack detection, including total CPU time (cpu_sec_total) and resident memory usage (resident_memory_total). A comparison of nine Machine Learning models for attack detection was conducted, including ensemble methods (Random Forest, XGBoost, LightGBM) which demonstrated the highest effectiveness, achieving 100 % accuracy (F1-score equals 1.0) and perfect class separation (AUC equals 1.0). The XGBoost model also eliminated false positives (precision equals 1.0). Feature importance analysis revealed the most significant metrics for classification: CPU usage (cpu_sec_total, cpu_sec_idle), network packet transmission (transmit_packets), system load average, and memory usage (virtual_memory_total, resident_memory_total). The work emphasizes the importance of integrating multi-level metrics for building resilient anomaly detection systems. The proposed approach is scalable and independent of specific frameworks, making it applicable for protecting containerized environments. The research results serve as a foundation for developing proactive Kubernetes security systems capable of countering sophisticated attack vectors.

Keywords

Kubernetes, DoS attack detection, machine learning, node-level metrics, application-level metrics, anomaly detection, ensemble models

For citation: Darwesh G., Hammoud J., Vorobeva A.A. Enhanced detection of denial-of-service attacks in Kubernetes: a multi-framework machine learning approach integrating node and application metric. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 910–922. doi: 10.17586/2226-1494-2025-25-5-910-922

УДК 004.056

Повышение эффективности обнаружения DoS-атак в Kubernetes: подход на основе машинного обучения с интеграцией метрик уровня узлов и приложений для мультифреймворковых сред

Гадир Дарвиш[1], Жаафар Хаммуд[2], Алиса Андреевна Воробьева[3]✉

[1,2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] ghadeerdarwesh32@gmail.com, https://orcid.org/0000-0003-1116-9410
[2] hammoudgj@gmail.com, https://orcid.org/0000-0002-2033-0838
[3] vorobeva@itmo.ru✉, https://orcid.org/0000-0001-6691-6167

Аннотация
Широкое распространение Kubernetes как платформы для оркестрации контейнеризированных приложений атакам типа «Отказ в обслуживании» (Denial-of-Service, DoS). В работе предложен подход к обнаружению DoS-атак, основанный на двух ключевых компонентах: использование комплексных метрик и применение ансамблевых моделей машинного обучения. Подход предполагает сбор и анализ комплексных метрик: уровня узлов (Central Processing Unit (CPU), память) и уровня приложений (сетевая активность, файловые дескрипторы) из контейнеров, работающих на различных фреймворках (Flask, Django, FastAPI, Node.js, Golang). Для реализации подхода создан набор данных, содержащий 49 990 экземпляров сетевой активности, охарактеризованных 28 признаками (комплексными метриками). Статистический анализ (t-критерий Стьюдента, корреляция Пирсона) выявил наиболее релевантные для детектирования атак метрики, включая общее время использования CPU (cpu_sec_total) и объем задействованной оперативной памяти (resident_memory_total). Сравнение девяти моделей машинного обучения для детектирования атак, включая ансамблевые методы (Random Forest, XGBoost, LightGBM), показало наивысшую эффективность (F1-мера равна 1,0) и полное разделение классов (AUC равна 1,0). Применение модели XGBoost позволило исключить ложноположительные срабатывания (precision равна 1.0). Анализ важности признаков выявил наиболее значимые для классификации метрики, связанные с использованием CPU (cpu_sec_total, cpu_sec_idle), передачей сетевых пакетов (transmit_packets), средней загрузкой системы и использованием памяти (virtual_memory_total, resident_memory_total). Проведенное исследование показало важность интеграции разноуровневых метрик для создания устойчивых систем обнаружения аномалий. Предложенный подход является масштабируемым и независимым от конкретных фреймворков, что делает его применимым для защиты контейнеризированных сред. Результаты исследования служат основой для разработки проактивных систем безопасности Kubernetes, способных противостоять сложным векторам атак.

Ключевые слова
Kubernetes, обнаружение DoS-атак, машинное обучение, метрики уровня узлов, метрики уровня приложений, обнаружение аномалий, ансамблевые модели

Ссылка для цитирования: Дарвиш Г., Хаммуд Ж., Воробьева А.А. Повышение эффективности обнаружения DoS-атак в Kubernetes: подход на основе машинного обучения с интеграцией метрик уровня узлов и приложений для мультифреймворковых сред // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 910–922 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-910-922

Introduction

Kubernetes, the de facto standard for container orchestration, has revolutionized cloud-native architectures by automating the deployment, scaling, and management of containerized applications. However, the complexity and dynamic nature of Kubernetes clusters expose them to a wide range of security threats, with Denial-of-Service (DoS) attacks standing out as a prominent risk. These attacks exploit Kubernetes resource scaling mechanisms to inundate cluster resources, potentially leading to service disruptions and substantial financial repercussions [1, 2]. The containerized workloads managed by Kubernetes pose unique challenges in detecting and mitigating DoS attacks. Containers often demonstrate dynamic, ephemeral, and unpredictable resource utilization patterns which blur the line between legitimate traffic bursts and malicious activity. Conventional DoS detection methods, such as static thresholds and signature-based techniques, fall short in such settings due to their inability to adapt to Kubernetes highly dynamic operational environments. These limitations often result in high false-positive rates, rendering traditional approaches unsuitable for production environments [1, 2]. Machine Learning (ML) has emerged as a promising paradigm for tackling these challenges. By leveraging anomaly detection techniques, ML-based systems can dynamically identify deviations in traffic and resource utilization patterns without relying on predefined rules or static thresholds. However, many existing studies focus on specific application frameworks, such as Flask, or narrow workloads, thereby limiting their generalizability across diverse operational contexts [1, 2]. Recent advancements in Kubernetes security emphasize the importance of hybrid approaches that combine ML with runtime monitoring and rule-based mechanisms. Tools such as extended Berkeley Packet Filter (eBPF) and Express Data Path (XDP) offer lightweight, high-performance anomaly detection at the kernel level, enabling real-time security insights. Despite these innovations, significant gaps remain in evaluating the effectiveness of such approaches across diverse frameworks and workloads, leaving room for improvement in generalizability and robustness [3, 4]. Building upon our prior work [5], where we developed an ML-based DoS detection framework tailored to the Flask framework, this study extends the scope to encompass multiple application frameworks, including Django, FastAPI, Flask, Golang, and Node.js. This broader scope addresses the generalizability concerns raised in our earlier work and ensures applicability to a wider range of Kubernetes environments.

This study makes three key contributions: it provides a comparative analysis of ML-based DoS detection across multiple frameworks to improve generalizability and robustness; it integrates lightweight runtime monitoring tools to enhance detection efficiency; and it evaluates advanced classifiers for distinguishing between natural workload variations and attack-induced anomalies in Kubernetes environments [5–7].

Literature Review and Previous Work

DoS attacks are among the most prevalent security threats in containerized environments. Kubernetes, with its dynamic orchestration and auto-scaling capabilities, provides an efficient platform for managing modern workloads but also presents a large attack surface for

adversaries to exploit. DoS attacks in Kubernetes often target resource management mechanisms, overwhelming cluster components like Central Processing Unit (CPU), memory, and network bandwidth to render applications unresponsive [1, 2]. Studies emphasize the difficulty in distinguishing between natural workload variations and attack-induced overloads, as both may manifest as anomalies in resource usage [1, 4].

Conventional DoS detection techniques rely on predefined thresholds or static rules to identify anomalous behaviors. While computationally inexpensive, these methods are rigid and fail to adapt to dynamic environments like Kubernetes [1, 2, 4].

Recent advancements have introduced more sophisticated approaches, such as Host-Based Intrusion Detection Systems (HIDS). Researchers in [8] developed a real-time HIDS for Linux containers that monitors system calls from the host kernel to detect anomalies. Their method achieved a 100 % detection rate with a false positive rate of just 2 %.

Several studies focus on Docker containers, the most widely used container runtime. For example: Researchers in [9] proposed an online anomaly detection system using an optimized isolation forest algorithm. By assigning weights to resource metrics and incorporating weighted feature selection, this approach improved accuracy while maintaining minimal performance overhead, crucial for differentiating between attack-induced and natural resource overloads. In [10], a probabilistic real-time IDS was proposed using *n*-grams of system calls and probabilistic models like Maximum Likelihood Estimator. This system achieved detection accuracy between 87 % and 97 % across datasets. Dynamic approaches using anomaly-based methods have demonstrated significant improvements over static techniques. In [11], researchers evaluated dynamic schemes on 28 real-world container vulnerability exploits, with results indicating that dynamic methods detected 22 out of 28 exploits compared to only three detected by static methods.

ML-based anomaly detection systems offer significant advantages in identifying DoS attacks. Techniques like Random Forest, Gradient Boosting, and Neural Networks have been widely adopted for detecting anomalies in Kubernetes. The introduction of eBPF and XDP has enabled lightweight, kernel-level anomaly detection for real-time insights [1, 2]. Studies such as [12] have demonstrated the potential of ML classifiers like Decision Trees and Random Forests in distinguishing between legitimate and malicious activities at the container level. These approaches achieved F-measures of 99.8 % for attack detection while maintaining low resource overheads.

Most existing research evaluates DoS detection methods on specific frameworks, such as Flask or FastAPI, without accounting for their generalizability to other workloads [1, 2, 4]. Researchers in [13] highlighted the need for cross-framework evaluations by analyzing security mechanisms in Docker containers across multiple deployment scenarios. Their results emphasized that framework-agnostic detection systems are critical for robust Kubernetes security.

Studies are often constrained to single frameworks, making their findings less applicable to diverse Kubernetes workloads [1, 2, 4].

Despite these advancements, the current research landscape reveals several persistent and interconnected limitations that hinder the development of robust, practical detection systems. A primary constraint is the lack of generalizability across technological stacks. The majority of studies evaluate proposed methods within the context of a single application framework, neglecting validation across diverse environments [1, 2, 4, 13]. This significantly limits the applicability of such solutions in real-world Kubernetes clusters which are inherently heterogeneous and host applications built with different languages and frameworks.

Furthermore, a significant efficiency-effectiveness trade-off remains unresolved. While static methods are computationally efficient yet inflexible, the more effective dynamic and ML approaches typically demand large volumes of training data and introduce substantial computational overhead, making them costly to deploy in resource-sensitive environments [10, 14].

Finally, the critical challenge of integration is still largely unaddressed. Only a limited number of studies have explored the combination of real-time monitoring techniques (e.g., eBPF) with ML systems to achieve the dual objectives of high accuracy and low-latency detection in Kubernetes production settings [1].

Thus, the identified gaps — limited generalizability, an unoptimized accuracy-overhead trade-off, and a lack of integrated real-time solutions — form the core research challenge addressed by this work.

This study directly addresses these limitations by proposing a comprehensive detection framework validated across multiple application frameworks and programming languages, including Flask, Django, FastAPI, Golang, and Node.js. Our approach leverages lightweight, eBPF-based runtime monitoring to minimize performance impact and detection latency. We conduct an extensive comparative evaluation of ML classifiers to identify optimal strategies for DoS detection in Kubernetes. Through these contributions, this research provides a foundation for developing scalable, framework-agnostic security solutions capable of protecting complex Kubernetes deployments against evolving DoS threats.

Data and Statistical Study

These contributions aim to provide a framework-agnostic, efficient, and accurate solution to securing Kubernetes environments against evolving threats. The dataset[1] consists of 49,990 instances with 28 features, encompassing both node-level and app-level metrics collected from multiple frameworks deployed in Kubernetes by using a collector developed in [15]. The target variable, *attack*, is binary, indicating the presence (1) or absence (0) of DoS attacks. Table 1 presents the node-level metrics gathered from the frameworks, while Table 2 details the app-level metrics.

We conducted a comprehensive statistical analysis to demonstrate the robustness and reliability of the dataset. This analysis provides insights into the distribution, central

[1] Available at: https://github.com/ghadeerda/Kubernetes-model-agent (accessed: 29.08.2025).

Table 1. This table lists and describes the metrics collected at the node level, including CPU, memory, disk, and network-related features

Column Name	Description
id	Unique identifier for the observation
time	Timestamp of the data record
cpu_sec_idle	Percentage of CPU idle time during the interval
disk_av_per	Available disk space as a percentage
disk_read	Amount of data read from the disk (in bytes)
disk_write	Amount of data written to the disk (in bytes)
net_receive	Network data received (in bytes)
mem_pressure	Memory pressure indicator (value reflects memory load)
mem_av_per	Available memory as a percentage
forks_total	Total number of process forks
intr	Number of interrupts handled by the CPU
load1, load5, load15	CPU load average over 1, 5, and 15 minutes respectively
receive_drop	Number of network packets dropped during reception
receive_errs	Number of network reception errors
transmit_packets	Number of packets transmitted over the network
ipv4_sock_inuse	Number of IPv4 sockets currently in use
est_conn	Number of established connections
lis_conn	Number of listening connections
open_fds	Number of open file descriptors
attack	Indicator for attack presence (1 for attack, 0 for no attack)

Table 2. This table provides details of the application-level metrics, including resource utilization features specific to the application frameworks

Column Name	Description
id	Unique identifier for the observation
time	Timestamp of the data record
cpu_sec_total	Total CPU usage in seconds
virtual_memory_total	Total virtual memory usage (in bytes)
resident_memory_total	Total resident memory usage (in bytes)
open_fds	Number of open file descriptors
attack	Indicator for attack presence (1 for attack, 0 for no attack)

tendencies, and variability of both node-level and app-level metrics, highlighting the dataset ability to capture diverse system behaviors under different conditions. By examining the relationships between features and their correlations with the target variable (*attack*), we confirmed that the dataset effectively encapsulates the characteristics required for detecting DoS attacks. This statistical study not only validates the dataset integrity but also underscores its potential for developing and benchmarking robust anomaly detection models in containerized environments.

To thoroughly understand the dataset characteristics and assess its suitability for detecting DoS attacks, we conducted an in-depth statistical analysis. This analysis aimed to explore key metrics at both node-level and app-level, evaluate their relationships, and identify patterns that distinguish between attack and non-attack states. By combining descriptive statistics, hypothesis testing, and correlation analysis, we established a solid foundation for understanding the dataset structure and its potential for training robust ML models. The following sections detail the findings from these analyses.

Descriptive Statistics

During attacks, CPU usage increases significantly while idle time decreases, indicating elevated system load. Virtual and resident memory also spike, reflecting stress on memory resources. Open file descriptors rise, suggesting heavier application activity. Disk and network I/O metrics show moderate changes but still contribute to overall anomaly detection.

Hypothesis Testing

T-tests revealed statistically significant differences between attack and non-attack states. Features such as

cpu_sec_total, resident_memory_total, and open_fds were highly significant, while disk_read, disk_write, and virtual_memory_total showed moderate significance. These findings confirm that attacks consistently alter resource usage patterns.

Correlation Analysis

Correlation analysis showed that cpu_sec_total, resident_memory_total, and open_fds are positively correlated with attacks, while cpu_sec_idle and disk_av_per are negatively correlated. Some memory-related features showed redundancy, while others like forks_total and disk_av_per contributed unique information.

Key Insights

The most predictive indicators of DoS attacks include increased CPU and memory usage, along with higher counts of open file descriptors. These resource usage patterns correspond to the system stress and resource exhaustion typically induced by attacks. Fig. 1 presents histograms comparing the distributions of key. Fig. 2 shows boxplots of metrics. These visualizations were generated by the authors based on the labeled dataset of 49,990 instances, which includes node- and application-level metrics collected from real Kubernetes workloads using five frameworks (Flask, Django, FastAPI, Node.js, and Golang). The patterns observed in these figures are derived from the statistical analysis discussed earlier (t-tests and correlation), and form the empirical basis for the ML models used in this study.

Sample Representativeness and Realism of Workload Simulation

To ensure a realistic and representative dataset, we designed the data collection process to emulate real

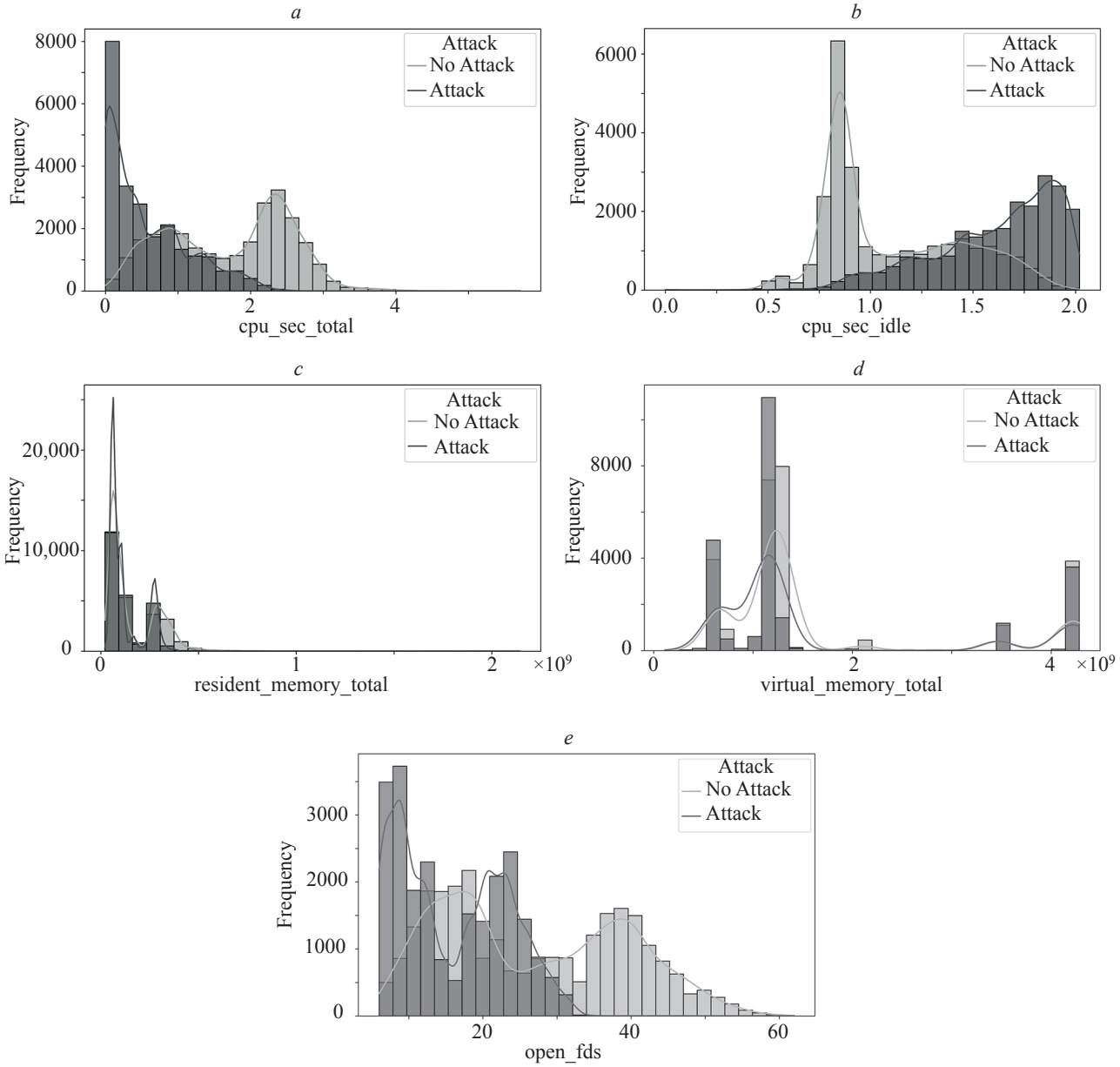

Fig. 1. Histograms of Frequency Distribution of CPU and memory usage metrics (cpu_sec_total, resident_memory_total) under attack and non-attack conditions, based on the collected dataset: cpu_usage_total (*a*); cpu_sec_idle (*b*); resident_memory_total (*c*); virtual_memory_total (*d*); open_fds (*e*)

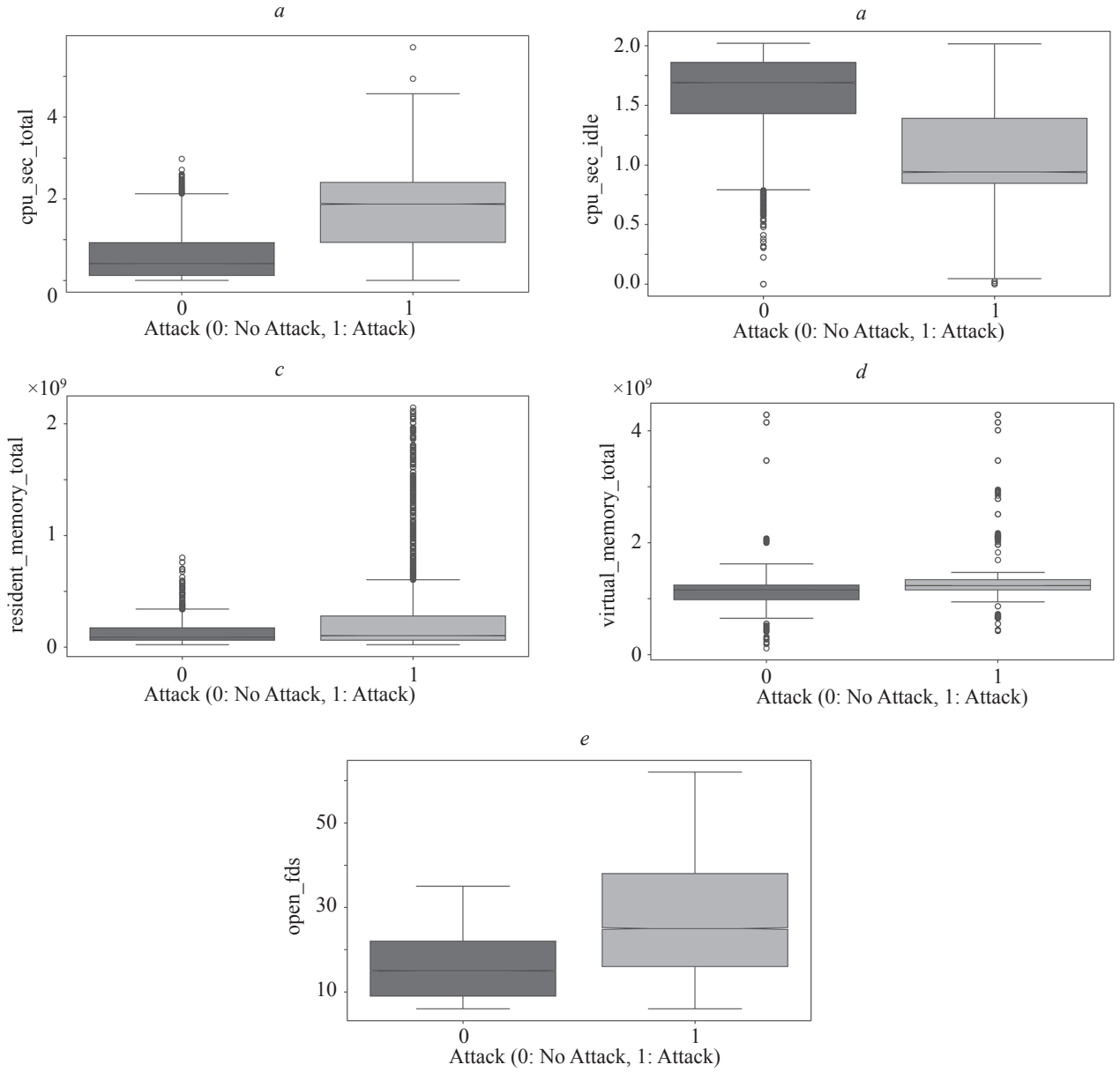

Fig. 2. Distribution shifts for key metrics. Feature Boxplots comparing CPU idle time (cpu_sec_idle) and open file descriptors (open_fds) between normal and attack states in Kubernetes workloads: cpu_usage_total (*a*); cpu_sec_idle (*b*); resident_memory_total (*c*); virtual_memory_total (*d*); open_fds (*e*)

Kubernetes environments. The dataset includes both node- and application-level metrics from applications built with five frameworks — Flask, Django, FastAPI, Golang, and Node.js — covering diverse architectures and performance patterns. Benign traffic was generated using synthetic tools and real user-like behavior, simulating varying request rates, CPU/memory loads, and I/O operations to reflect real-world workload dynamics such as traffic spikes and batch processing. Attack data was produced using standard DoS tools to stress both network and application layers, mimicking real adversarial scenarios like volumetric floods and resource exhaustion. This blend of framework diversity, metric richness, and controlled anomaly injection results in a dataset that reflects real operational conditions and provides a solid foundation for training effective, generalizable ML models.

ML Models and Detection Approach

We evaluated multiple supervised learning algorithms for detecting DoS attacks, focusing on their ability to generalize across different frameworks and languages (Flask, Django, FastAPI, Nodejs, Golang). The classifiers include: Logistic Regression: A simple linear model for baseline comparisons [16]. Random Forest: A tree-based ensemble model known for its robustness to overfitting [17]. Gradient Boosting: An iterative boosting model suitable for handling imbalanced datasets [18]. Support Vector Machine (SVM): Effective for high-dimensional feature spaces [19]. Decision Tree: A fast and interpretable model [20]. Naive Bayes: Suitable for datasets where feature independence can be assumed [21]. K-Nearest Neighbors: A distance-based approach for capturing non-

linear patterns [22]. XGBoost: A high-performance gradient boosting model [23]. LightGBM: A scalable and efficient tree-based model optimized for large datasets [24].

All features were standardized using StandardScaler to ensure uniform scaling across models. Any missing values were imputed based on the mean or median of the respective feature.

A stratified 5-fold cross-validation strategy was employed to ensure robust evaluation across diverse data splits. For tree-based models, feature importance scores were computed to interpret the contribution of individual metrics.

Accuracy, precision, recall, and F1-score were computed for each classifier. Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores were used to assess model performance. In addition to the Confusion Matrix which provided a detailed breakdown of true positives, true negatives, false positives, and false negatives for each classifier.

The detection pipeline was implemented using Python, leveraging libraries, such as Scikit-learn, XGBoost, and LightGBM. The training and evaluation processes were automated to facilitate reproducibility and ensure consistent results across multiple frameworks.

Results and Discussion

This section presents a detailed analysis of the results derived from the machine learning classifiers applied to the combined dataset, integrating both application and node-level metrics from multiple frameworks. The discussion includes model performance metrics, feature importance analysis, and insights into classifier behavior for detecting anomalies.

The cross-validation accuracy of the classifiers was evaluated to assess their ability to generalize across different data subsets. The accuracy boxplot shows that ensemble models like Random Forest, Gradient Boosting, XGBoost, and LightGBM consistently outperformed other models, achieving nearly perfect performance with minimal variance. Among all classifiers: Random Forest and Decision Tree achieved perfect classification accuracy with no false predictions. XGBoost achieved an accuracy of 100 % with robust predictive power. Linear classifiers like Logistic Regression and simpler models like Naive Bayes showed relatively lower but acceptable accuracies. The classifier cross-validation accuracy plot highlights the stability of ensemble models compared to others (Fig. 3).

The confusion matrices provide detailed insights into the classification behavior of each ML model. Fig. 4 illustrates these matrices for all evaluated classifiers, showing the distribution of true positives, true negatives, false positives, and false negatives. Notably, XGBoost and Random Forest achieved perfect classification, with zero misclassifications of either attack or non-attack instances. In contrast, models like Naive Bayes and SVM showed some weaknesses. For example, Naive Bayes falsely identified 4,967 normal cases as attacks, reflecting its limitation in environments with overlapping feature distributions. SVM exhibited a slightly elevated false positive rate due to its sensitivity to kernel parameterization. These confusion matrices were generated by the authors using the labeled dataset of 49,990 samples, collected from a realistic Kubernetes cluster and detailed in the dataset description. The results confirm the relative strengths of ensemble models and provide a comparative evaluation of precision and recall across all classifiers.

Feature Importance

Feature importance was analyzed using: tree-based native importance for ensemble models and permutation importance for other classifiers where native importance is unavailable.

The following trends were observed: For tree-based models like Random Forest, LightGBM, and XGBoost, the most critical features were: CPU utilization metrics (cpu_sec_total, cpu_sec_idle), Packet transmission metrics (transmit_packets), System load averages (load1, load5, load15), and Memory metrics (virtual_memory_total, resident_memory_total). XGBoost and Gradient Boosting

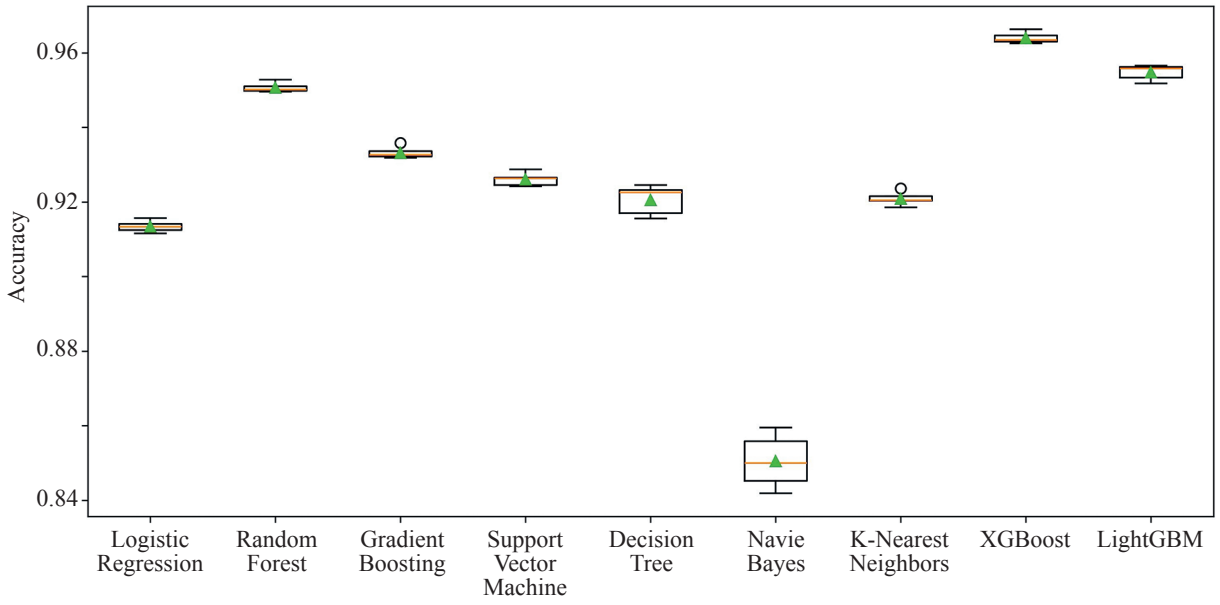

Fig. 3. Boxplot comparing cross-validation accuracy across all evaluated machine learning models

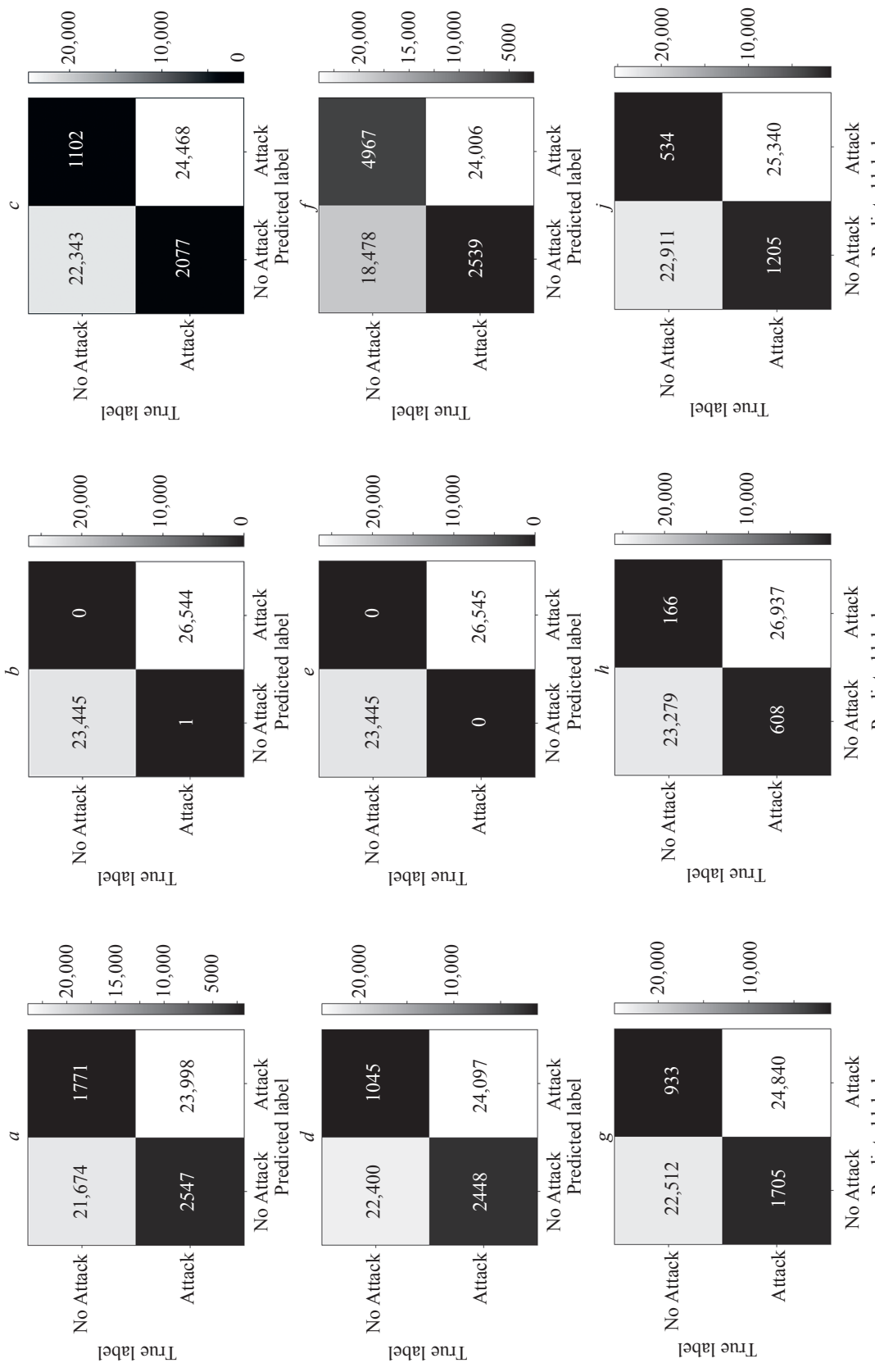

Fig. 4. Individual confusion matrices for each classifier, highlighting true positives, false positives, true negatives, and false negatives: Logistic Regression (*a*) Random Forest (*b*); Gradient Boosting (*c*); Support Vector Machine (*d*); Decision Tree (*e*); Naive Bayes (*f*); K-Nearest Neighbors (*g*); XGBoost (*h*); LigthGBM (*j*)

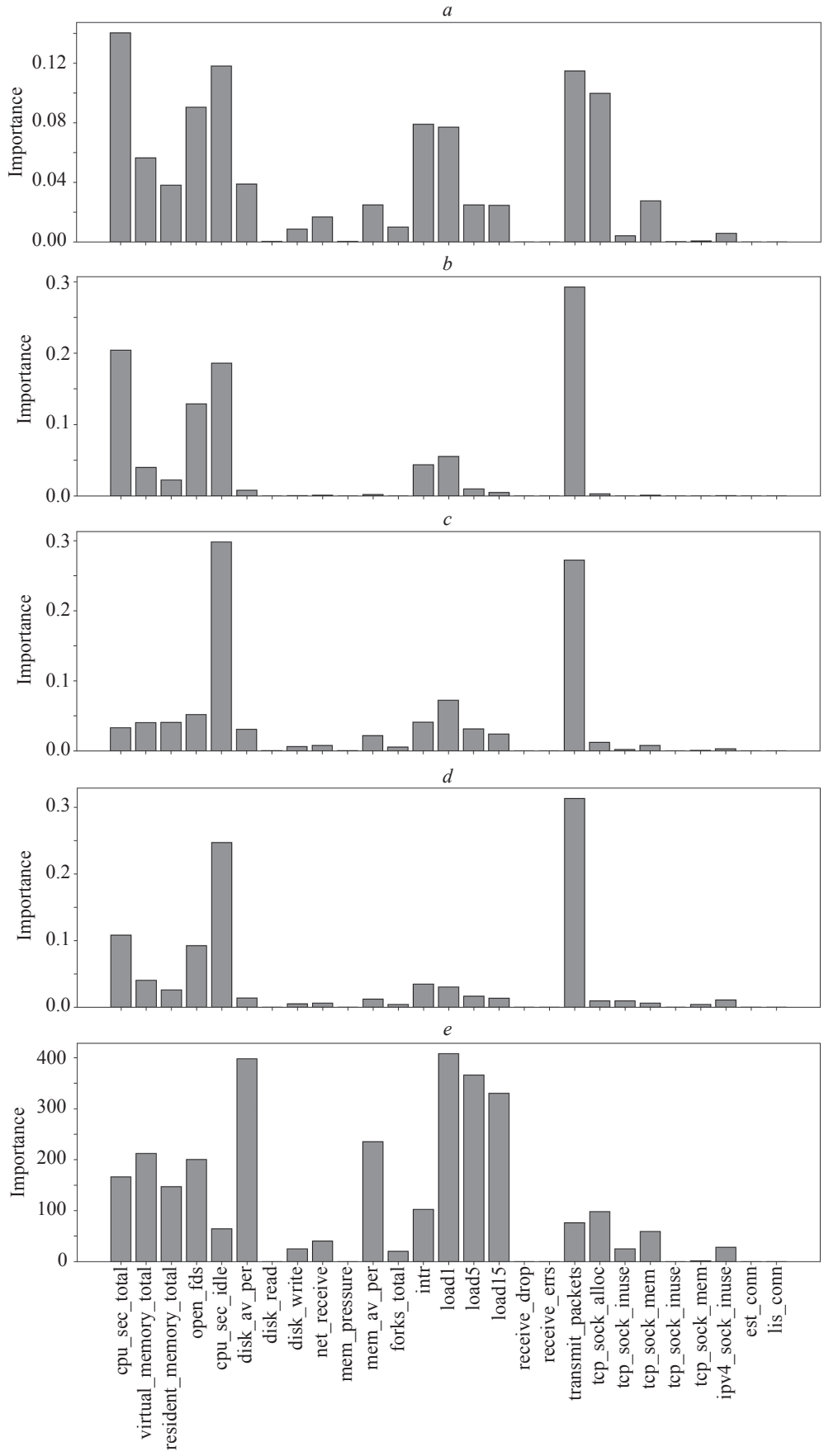

Fig. 5. Bar plots showing feature importance rankings for models that provide native importance measures (e.g., Random Forest, XGBoost, LightGBM): Random Forest (*a*); Gradient Boosting (*b*); Decision Tree (*c*); XGBoost (*d*); LigthGBM (*e*)

highlighted the dominance of network-level metrics such as transmit_packets and tcp_sock_alloc. Decision Tree models underscored similar features, with additional emphasis on resource allocation metrics (open_fds, mem_av_per). The variation across feature importance profiles indicates that different models prioritize features differently, depending on their intrinsic algorithms and data processing mechanics.

Separate plots for feature importances across classifiers and a combined feature importance comparison provide detailed insights (Fig. 5).

Receiver Operating Characteristic Analysis

The ROC curves demonstrated high AUC values for all classifiers: XGBoost, Random Forest, and Decision Tree exhibited an AUC of 1.0, confirming perfect separability of the classes. Models like Naive Bayes and Logistic Regression achieved slightly lower AUCs (about 0.94 and 0.97, respectively), indicating lower sensitivity to certain features. The ROC curves for all classifiers are included for a holistic view of performance across false-positive and true-positive rates (Fig. 6).

Quantitative Evaluation of Early Detection Capabilities

In addition to accuracy, a key strength of a detection system is how early it can identify anomalies before traditional alerts are triggered. We analyzed the time gap between model predictions and system-level resource exhaustion warnings. XGBoost and Random Forest detected DoS attack signatures from 3 to 12 seconds before critical signs like CPU saturation, memory exhaustion, or application failures occurred. This early warning enables proactive actions, such as auto-scaling, traffic throttling, or container isolation — reducing the risk of service disruption. Detection lead time was measured by aligning attack labels with resource usage traces and locating inflection points in key metrics (cpu_sec_total, resident_memory_total, open_fds). Fig. 7 illustrates this timeline, showing that our models consistently flag anomalies before threshold breaches, confirming their effectiveness in real-time Kubernetes defense systems.

Key Insights and Implications

Ensemble models like Random Forest, XGBoost, LightGBM, and Gradient Boosting consistently outperformed other classifiers in accuracy, robustness, and feature prioritization, proving highly effective for high-dimensional, multi-source datasets. In contrast, simpler models such as Naive Bayes provided baseline performance but lacked the sophistication needed for complex environments. Resource and network metrics — including CPU utilization, packet transmission, and system load — were key indicators for anomaly detection. Integrating node- and application-level data significantly improved detection accuracy and scalability, making ensemble approaches well-suited for real-time deployment in Kubernetes and edge environments. These results establish a strong foundation for developing generalizable frameworks that can distinguish between attack-induced and natural workload fluctuations in containerized systems.

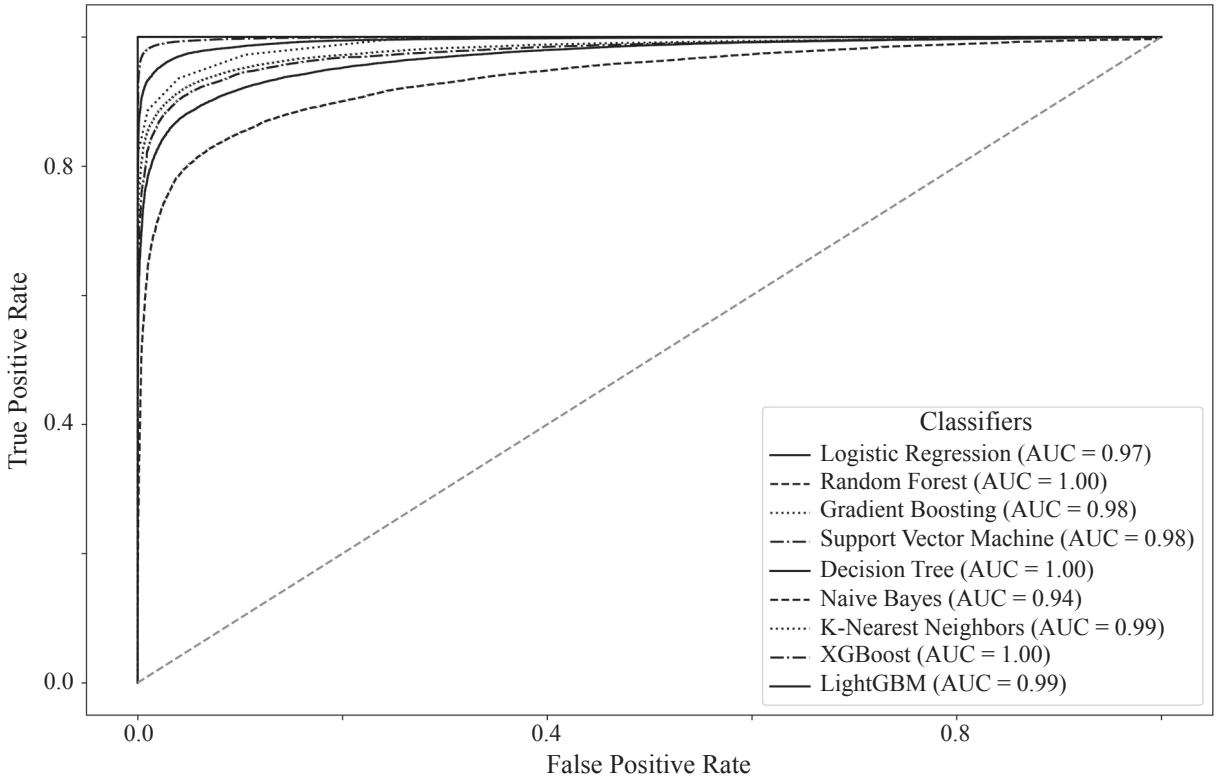

Fig. 6. ROC curves for all classifiers, illustrating their ability to differentiate between attack and non-attack states

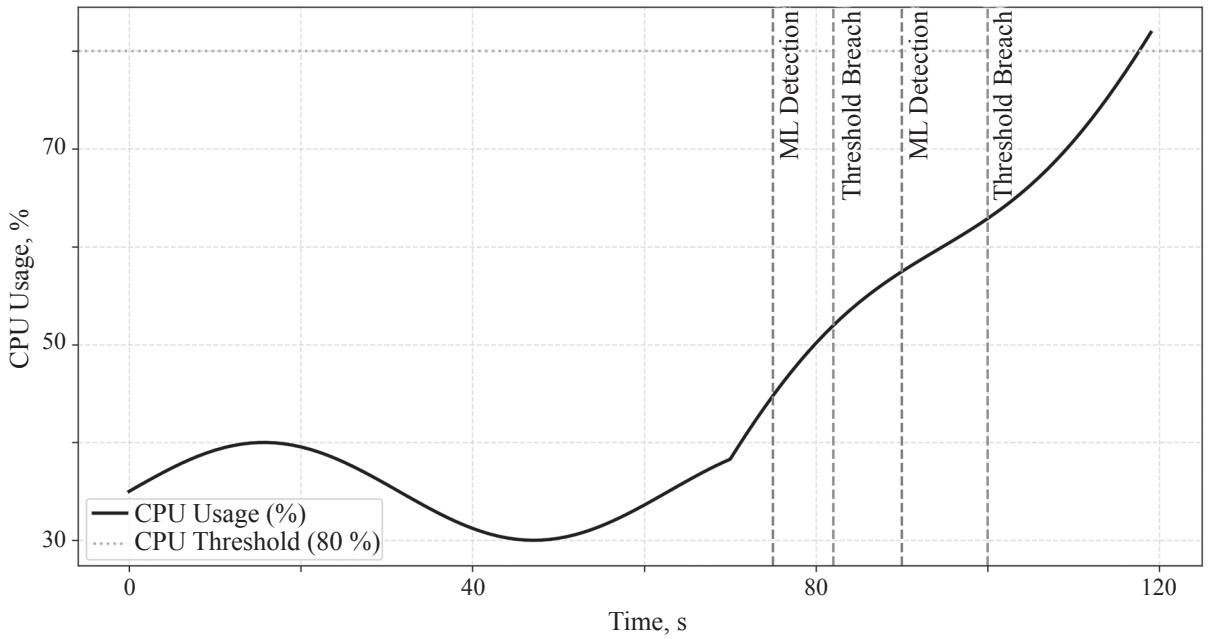

Fig. 7. Simulated comparative timeline of CPU usage during Denial-of-Service (DoS) attack scenarios. The figure illustrates two events where ML models (XGBoost, Random Forest) detect anomalous activity several seconds before the system reaches critical CPU utilization thresholds (80 %). This early detection window — ranging from 3 to 12 seconds — enables proactive mitigation before visible service degradation occurs, demonstrating the practical value of the proposed approach in real-time Kubernetes environments

Conclusion

This study evaluated machine learning methods for detecting DoS attacks in Kubernetes using metrics from five frameworks — Flask, Django, FastAPI, Node.js, and Golang. By combining node- and application-level metrics, we built a robust dataset that captures diverse workload behaviors. Statistical and correlation analyses highlighted the predictive power of CPU usage, memory consumption, and open file descriptors. Ensemble classifiers, especially XGBoost, Random Forest, and LightGBM, achieved the highest performance in identifying attacks. The approach proved scalable and adaptable across frameworks, advancing beyond static or framework-specific methods. Future work will focus on integrating lightweight monitoring tools and testing in resource-constrained edge environments to further improve real-time detection in cloud-native systems.

References

1. Sadiq A., Syed H.J., Ansari A.A., Ibrahim A.O., Alohaly M., Elsadig M. Detection of denial of service attack in cloud based kubernetes using eBPF. *Applied Sciences*, 2023, vol. 13, no. 8, p. 4700. https://doi.org/10.3390/app13084700
2. Cao C., Blaise A., Verwer S., Rebecchi F. Learning state machines to monitor and detect anomalies on a kubernetes cluster. *Proc. of the 17th International Conference on Availability, Reliability and Security*, 2022, pp. 1–9. https://doi.org/10.1145/3538969.3543810
3. Koksal S., Catak F. O., Dalveren Y. Flexible and lightweight mitigation framework for distributed denial-of-service attacks in container-based edge networks using Kubernetes. *IEEE Access*, 2024, vol. 12, pp. 172980–172991. https://doi.org/10.1109/ACCESS.2024.3501192
4. Tripathi A.A. *Attacking and Defending Kubernetes*. PhD thesis. Dublin Business School, 2024. Available at: https://esource.dbs.ie/items/eda4ea15-cedf-456b-93f9-6ce67e25c4bb (accessed: 02.12.2024).
5. Darwesh G., Hammoud J., Vorobeva A.A. Enhancing Kubernetes security with machine learning: a proactive approach to anomaly detection. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 1007–1015. https://doi.org/10.17586/2226-1494-2024-24-6-1007-1015
6. Ghadeer D., Jaafar H., Vorobeva A.A. Security in Kubernetes: best practices and security analysis. *Journal of the Ural Federal District. Information Security*, 2022, no. 2 (44), pp. 63–69. https://doi.org/10.14529/secur220209

Литература

1. Sadiq A., Syed H.J., Ansari A.A., Ibrahim A.O., Alohaly M., Elsadig M. Detection of denial of service attack in cloud based kubernetes using eBPF // Applied Sciences. 2023. V. 13. N 8. P. 4700. https://doi.org/10.3390/app13084700
2. Cao C., Blaise A., Verwer S., Rebecchi F. Learning state machines to monitor and detect anomalies on a kubernetes cluster // Proc. of the 17th International Conference on Availability, Reliability and Security. 2022. P. 1–9. https://doi.org/10.1145/3538969.3543810
3. Koksal S., Catak F. O., Dalveren Y. Flexible and lightweight mitigation framework for distributed denial-of-service attacks in container-based edge networks using Kubernetes // IEEE Access. 2024. V. 12. P. 172980–172991. https://doi.org/10.1109/ACCESS.2024.3501192
4. Tripathi A.A. Attacking and Defending Kubernetes. PhD thesis. Dublin Business School, 2024. [Online]. URL: https://esource.dbs.ie/items/eda4ea15-cedf-456b-93f9-6ce67e25c4bb (accessed: 02.12.2024).
5. Darwesh G., Hammoud J., Vorobeva A.A. Enhancing Kubernetes security with machine learning: a proactive approach to anomaly detection // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2024. V. 24. N 6. P. 1007–1015. https://doi.org/10.17586/2226-1494-2024-24-6-1007-1015
6. Ghadeer D., Jaafar H., Vorobeva A.A. Security in Kubernetes: best practices and security analysis // Journal of the Ural Federal District. Information Security. 2022. N. 2 (44). P. 63–69. https://doi.org/10.14529/secur220209

7. Darwesh G., Hammoud J., Vorobeva A.A. Enhancing kubernetes security: the crucial role of DevSecOps. *Proc. of the Institute for Systems Analysis Russian Academy of Sciences*, 2024, vol. 74, no. 3, pp. 78-88. https://doi.org/10.14357/20790279240309
8. Abed A.S., Clancy C., Levy D.S. Intrusion detection system for applications using linux containers. *Lecture Notes in Computer Science*, 2024, vol. 9331, pp. 123–135. https://doi.org/10.1007/978-3-319-24858-5_8
9. Zou Z., Xie Y., Huang K., Xu G., Feng D., Long D. A docker container anomaly monitoring system based on optimized isolation forest. *IEEE Transactions on Cloud Computing*, 2022, vol. 10, no. 1, pp. 134–145. https://doi.org/10.1109/TCC.2019.2935724
10. Srinivasan S., Kumar A., Mahajan M., Sitaram D., Gupta S. Probabilistic real-time intrusion detection system for docker containers. *Communications in Computer and Information Science*, 2019, vol. 969, pp. 336–347. https://doi.org/10.1007/978-981-13-5826-5_26
11. Tunde-Onadele O., He J., Dai T., Gu X. A study on container vulnerability exploit detection. *Proc. of the IEEE International Conference on Cloud Engineering (IC2E)*, 2019, pp. 121–127. https://doi.org/10.1109/IC2E.2019.00026
12. Flora J., Gonçalves P., Antunes N. Using attack injection to evaluate intrusion detection effectiveness in container-based systems. *Proc. of the IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2020, pp. 60–69. https://doi.org/10.1109/PRDC50213.2020.00017
13. Haq M.S., Nguyen T.D., Tosun A.S., Vollmer F., Korkmaz T., Sadeghi A.-R. SoK: a comprehensive analysis and evaluation of docker container attack and defense mechanisms. *Proc. of the IEEE Symposium on Security and Privacy (SP)*, 2024, pp. 4573–4590. https://doi.org/10.1109/sp54263.2024.00268
14. Lin Y., Tunde-Onadele O., Gu X. Cdl: Classified distributed learning for detecting security attacks in containerized applications. *Proc. of the 36th Annual Computer Security Applications Conference*, 2020, pp. 179–188. https://doi.org/10.1145/3427228.3427236
15. Darwesh G., Hammoud J., Vorobeva A.A. A novel approach to feature collection for anomaly detection in Kubernetes environment and agent for metrics collection from Kubernetes nodes. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 538–546. https://doi.org/10.17586/2226-1494-2023-23-3-538-546
16. LaValley M.P. Logistic regression. *Circulation*, 2008, vol. 117, no. 18, pp. 2395–2399. https://doi.org/10.1161/circulationaha.106.682658
17. Rigatti S.J. Random Forest. *Journal of Insurance Medicine*, 2017, vol. 47, no. 1, pp. 31–39. https://doi.org/10.17849/insm-47-01-31-39.1
18. Natekin A., Knoll A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 2013, vol. 7, pp. 21. https://doi.org/10.3389/fnbot.2013.00021
19. Suthaharan S. Support vector machine. *Integrated Series in Information Systems*, 2016, vol. 36, pp. 207–235. https://doi.org/10.1007/978-1-4899-7641-3_9
20. Song Y., Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 2015, vol. 27, no. 2, pp. 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
21. Rish I. An empirical study of the naive Bayes classifier. *Proc. of the IJCAI-2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001, pp. 41–46.
22. Kramer O. K-Nearest neighbors. *Intelligent Systems Reference Library*, 2013, vol. 51, pp. 13–23. https://doi.org/10.1007/978-3-642-38652-7_2
23. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785
24. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: a highly efficient gradient boosting decision tree. *Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3149–3157.

7. Darwesh G., Hammoud J., Vorobeva A.A. Enhancing kubernetes security: the crucial role of DevSecOps // Proc. of the Institute for Systems Analysis Russian Academy of Sciences. 2024. V. 74. N 3. P. 78-88. https://doi.org/10.14357/20790279240309
8. Abed A.S., Clancy C., Levy D.S. Intrusion detection system for applications using linux containers // Lecture Notes in Computer Science. 2024. V. 9331. P. 123–135. https://doi.org/10.1007/978-3-319-24858-5_8
9. Zou Z., Xie Y., Huang K., Xu G., Feng D., Long D. A docker container anomaly monitoring system based on optimized isolation forest // IEEE Transactions on Cloud Computing. 2022. V. 10. N 1. P. 134–145. https://doi.org/10.1109/TCC.2019.2935724
10. Srinivasan S., Kumar A., Mahajan M., Sitaram D., Gupta S. Probabilistic real-time intrusion detection system for docker containers // Communications in Computer and Information Science. 2019. V. 969. P. 336–347. https://doi.org/10.1007/978-981-13-5826-5_26
11. Tunde-Onadele O., He J., Dai T., Gu X. A study on container vulnerability exploit detection // Proc. of the IEEE International Conference on Cloud Engineering (IC2E). 2019. P. 121–127. https://doi.org/10.1109/IC2E.2019.00026
12. Flora J., Gonçalves P., Antunes N. Using attack injection to evaluate intrusion detection effectiveness in container-based systems // Proc. of the IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC). 2020. P. 60–69. https://doi.org/10.1109/PRDC50213.2020.00017
13. Haq M.S., Nguyen T.D., Tosun A.S., Vollmer F., Korkmaz T., Sadeghi A.-R. SoK: a comprehensive analysis and evaluation of docker container attack and defense mechanisms // Proc. of the IEEE Symposium on Security and Privacy (SP). 2024. P. 4573–4590. https://doi.org/10.1109/sp54263.2024.00268
14. Lin Y., Tunde-Onadele O., Gu X. Cdl: Classified distributed learning for detecting security attacks in containerized applications // Proc. of the 36th Annual Computer Security Applications Conference. 2020. P. 179–188. https://doi.org/10.1145/3427228.3427236
15. Darwesh G., Hammoud J., Vorobeva A.A. A novel approach to feature collection for anomaly detection in Kubernetes environment and agent for metrics collection from Kubernetes nodes // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2023. V. 23. N 3. P. 538–546. https://doi.org/10.17586/2226-1494-2023-23-3-538-546
16. LaValley M.P. Logistic regression // Circulation. 2008. V. 117. N 18. P. 2395–2399. https://doi.org/10.1161/circulationaha.106.682658
17. Rigatti S.J. Random Forest // Journal of Insurance Medicine. 2017. V. 47. N 1. P. 31–39. https://doi.org/10.17849/insm-47-01-31-39.1
18. Natekin A., Knoll A. Gradient boosting machines, a tutorial // Frontiers in Neurorobotics. 2013. V. 7. P. 21. https://doi.org/10.3389/fnbot.2013.00021
19. Suthaharan S. Support vector machine // Integrated Series in Information Systems. 2016. V. 36. P. 207–235. https://doi.org/10.1007/978-1-4899-7641-3_9
20. Song Y., Lu Y. Decision tree methods: applications for classification and prediction // Shanghai Archives of Psychiatry. 2015. V. 27. N 2. P. 130–135, https://doi.org/10.11919/j.issn.1002-0829.215044
21. Rish I. An empirical study of the naive Bayes classifier // Proc. of the IJCAI-2001 Workshop on Empirical Methods in Artificial Intelligence. 2001. P. 41–46.
22. Kramer O. K-Nearest neighbors // Intelligent Systems Reference Library. 2013. V. 51. P. 13–23. https://doi.org/10.1007/978-3-642-38652-7_2
23. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794. https://doi.org/10.1145/2939672.2939785
24. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: a highly efficient gradient boosting decision tree // Proc. of the 31st International Conference on Neural Information Processing Systems. 2017. P. 3149–3157.

Authors

Ghadeer Darwesh — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57226287648, https://orcid.org/0000-0003-1116-9410, ghadeerdarwesh32@gmail.com
Jaafar Hammoud — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57222044000, https://orcid.org/0000-0002-2033-0838, hammoudgj@gmail.com
Alisa A. Vorobeva — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57191359167, https://orcid.org/0000-0001-6691-6167, vorobeva@itmo.ru

Авторы

Дарвиш Гадир — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57226287648, https://orcid.org/0000-0003-1116-9410, ghadeerdarwesh32@gmail.com
Хаммуд Жаафар — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57222044000, https://orcid.org/0000-0002-2033-0838, hammoudgj@gmail.com
Воробьева Алиса Андреевна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57191359167, https://orcid.org/0000-0001-6691-6167, vorobeva@itmo.ru

Received 26.03.2025
Approved after reviewing 02.09.2025
Accepted 30.09.2025

Статья поступила в редакцию 26.03.2025
Одобрена после рецензирования 02.09.2025
Принята к печати 30.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-923-932

Experimental results of using AES-128 in LoRaWAN

Abdelouahab Nouar[1], Mounir Tahar Abbes[2]✉, Selma Boumerdassi[3], Mostefa Chaib[4]

[1,4] Hassiba Ben Bouali University (UHBC), LMA Laboratory, Chlef, 02010, Algeria
[2] Hassiba Ben Bouali University (UHBC), Chlef, 02010, Algeria
[3] Conservatoire National des Arts et Metiers (CNAM), Paris, 75141, France

[1] a.nouar@univ-chlef.dz, https://orcid.org/0009-0001-3355-1912
[2] m.taharabbes@univ-chlef.dz✉, https://orcid.org/0000-0001-5132-2366
[3] selma.boumerdassi@inria.fr, https://orcid.org/0000-0003-2603-2433
[4] m.chaib@univ-chlef.dz, https://orcid.org/0000-0001-9137-9527

Abstract
In the Internet of Things (IoT), Low Power Wide Area Networks (LPWAN) technologies have been obtaining considerable attention. Long-Range Wide-Area Networks (LoRaWAN) was created by the Long Range (LoRa) Alliance as an open standard operating over the unlicensed band. Its advantages include a large coverage area, low power consumption, and inexpensive transceiver chips. The standard of LoRaWAN encryption uses a 128-bit symmetric algorithm called Advanced Encryption Standard (AES). This standard secures communication and entities which are beneficial for resource-constrained devices on the IoT for efficient communication and security. The security problems with LoRa networks and devices remain an important challenge considering the technology large deployment for numerous applications. Even though LoRaWAN network architecture and security have been enhanced by the LoRa Alliance, the most recent version still has some weaknesses such as its susceptibility to attacks. Many studies and researchers have indicated that LoRaWAN versions 1.0 and 1.1 have security risks and vulnerabilities. This research proposes a method to construct and integrate cryptographic algorithms (AES-128) within widely utilized wireless Network Server Simulators NS-3. This module aims to increase the security of data in LoRa networks by protecting critical information from unauthorized access. Consequently, implementing the AES-128 encryption algorithm within the NS-3 simulator will benefit the scientific community greatly. This will enable an examination of the impact of various security measures on network performance metrics, including latency, overhead, energy consumption, throughput, and packet size.

Keywords
LoRaWAN, cryptography, LoRa, AES-128, security, NS-3, IoT

For citation: Nouar A., Tahar Abbes M., Boumerdassi S., Chaib M. Experimental results of using AES-128 in LoRaWAN. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 923–932. doi: 10.17586/2226-1494-2025-25-5-923-932

УДК 004.056.55

Экспериментальные результаты использования AES-128 в LoRaWAN

Абделуахаб Нуар[1], Мунир Тахар Аббес[2]✉, Сельма Бумердасси[4], Мостефа Хаиб[4]

[1,4] Университет Асиба Бенбуали Лаборатория ЛМА, Шлеф, 02010, Алжир
[2] Университет Асиба Бенбуали, Шлеф, 02010, Алжир
[3] Национальная консерватория искусств и ремесел, Париж, 75141, Франция

[1] a.nouar@univ-chlef.dz, https://orcid.org/0009-0001-3355-1912
[2] m.taharabbes@univ-chlef.dz✉, https://orcid.org/0000-0001-5132-2366
[3] selma.boumerdassi@inria.fr, https://orcid.org/0000-0003-2603-2433
[4] m.chaib@univ-chlef.dz, https://orcid.org/0000-0001-9137-9527

Аннотация
Технология Low Power Wide Area Networks (LPWAN) привлекает значительное внимание в Интернете вещей (IoT). Long-Range Wide-Area Networks (LoRaWAN) создан компанией Long Range (LoRa) как открытый

нелицензионный стандарт. Его преимущества включают большую зону покрытия, низкое энергопотребление и недорогие чипы приемопередатчиков. Стандарт шифрования LoRaWAN использует 128-битный симметричный алгоритм Advanced Encryption Standard (AES). Этот стандарт защищает связь и объекты, что выгодно для устройств с ограниченными ресурсами в IoT для эффективной связи и безопасности. Проблемы безопасности сетей и устройств LoRa остаются важной задачей, учитывая широкое распространение этой технологии в многочисленных приложениях. Несмотря на то, что создатели LoRa улучшили архитектуру и безопасность сети LoRaWAN, последняя версия все еще имеет некоторые недостатки, такие как уязвимость к атакам. Многочисленные исследования показали, что версии LoRaWAN 1.0 и 1.1 содержат угрозы безопасности и уязвимости. В работе предлагается метод построения и интеграции криптографических алгоритмов (AES-128) в широко используемых симуляторах беспроводных сетей NS-3. Целью данного средства является повышение безопасности данных в сетях LoRaWAN путем защиты критически важной информации от несанкционированного доступа. Внедрение алгоритма шифрования AES-128 в симулятор NS-3 позволит изучить влияние различных мер безопасности на показатели производительности сети, включая задержку, накладные расходы, энергопотребление, пропускную способность и размер пакета.

Ключевые слова
LoRaWAN, криптография, LoRa, AES-128, security, NS-3, IoT

Ссылка для цитирования: Нуар А., Тахар Аббес М., Бумердасси С., Хаиб М. Экспериментальные результаты использования AES-128 в LoRaWAN // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 923–932 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-923-932

Introduction

Long-Range Wide-Area Networks (LoRaWAN) is a Low-Power Wide Area Network (LPWAN) protocol that uses low-power algorithm to send data over long distances. LoRaWAN utilizes the unlicensed wireless spectrum, meaning anyone can use it without government permission. Multiple End Devices (EDs) communicate with a central gateway using a star topology [1]. The gateway then transmits the data packets to a Network Server (NS) that manages all devices and processes the data.

LoRaWAN is widely used in smart city applications [1], with many cities around the world, using the technology for various use cases, such as intelligent lighting, waste management, and air quality monitoring. Overall, these statistics highlight the growing popularity and adoption of LoRaWAN and its suitability for a wide range of Internet of Things (IoT) applications [2].

In LoRaWAN, cryptography plays a critical role in securing data transmission. One of the cryptographic methods used in LoRaWAN is Advanced Encryption Standard (AES) 128 bits. The latest version of LoRaWAN v.1.1 has provided a security framework that includes data privacy protection, data integrity control, device authentication, and key management [3]. The LoRaWAN protocol uses AES-128 algorithm in the core encryption mechanism to guarantee confidentiality, integrity, and authentication through two layers:

— Network Layer Encryption: This layer uses a Network Session Key (NwkSKey) to assure end-to-end device connection with the network server. It protects the integrity of the message to ensure that it comes from a legitimate device.
— Application Layer: This layer assures the confidentiality of data from the ED to the Application Server (AS), by introducing an Application Session Key called (AppSKey). It provides encryption to protect the payload, ensuring that only the AS can decrypt the actual message content.

In light of the latest development in resource-constrained IoT devices, various versions of LoRaWAN have been published to improve its performance in terms of security, scalability, and real-time long-range communication. The following summarizes the evolution of the LoRaWAN specifications, including the year of release and major changes:

— LoRaWAN version 1.0 (January 2015)[1]: first approval of LoRaWAN 1.0;
— LoRaWAN version 1.0.1 (February 2016) [4]: This update added a new frequency plan, modified some MAC-layer instructions;
— LoRaWAN version 1.0.2 (July 2016)[2]: Many problems were resolved, and others MAC commands were created in this version;
— LoRaWAN version 1.1 (October 2017)[3]: With the addition of a new server named Join Server, this significant revision introduced a new architecture. It also brought many improvements to the security mechanism, including support for roaming handover and the use of two root keys rather than one to derive the session security keys. Finally, it added numerous countermeasures to mitigate some of the vulnerabilities that had been reported in earlier versions;
— LoRaWAN version 1.0.3 (July 2018)[4]: In this revision a little number of MAC commands for class A devices are added;
— LoRaWAN version 1.0.4 (October 2020) [5]: This small update for v1.0.3 clarified various issues on Adaptive Data Rate (ADR) behavior, FCnt usage and behaviors, joining channel selection process, and retransmission backoff. This release includes two significant security-related changes that observed: first, DevNonce

[1] L. Specification, “LoRaWAN specification v1. 0”, San Francisco, CA, USA, 2015. Available at: https://lora-alliance.org/resource_hub/lorawan-specification-v1-0 (accessed: 17.04.2024).

[2] L. Alliance, “LoRaWAN specification v1. 0.2”, Date of retrieval, 2016. Available at: https://resources.lora-alliance.org/technical-specifications-v1-0-2/ (accessed: 11.11.2024).

[3] LoRaWAN specification v1.1. Available at: https://resources.lora-alliance.org/technical-specifications/ (accessed: 29.11.2023).

[4] L. Specification, “LoRaWAN specification v1.0.3”, San Francisco, CA, USA, 2018. Available at: https://lora-alliance.org/resource_hub/lorawan-specification-v1-0-3 (accessed: 12.03.2024).

generation is now incremental rather than random, and second, JoinEUI and AppNonce have been substituted for AppEUI and AppNonce.

This research focuses on the implementation of AES-128-bit cryptography under the NS-3 simulator; by doing so, it will improve the realism and security aspects of simulations, especially when the work will be on IoT or wireless network research. The main idea is to simulate secure communication within networks and to study the effects of encryption on network performance.

The rest of the paper is organized as follows: first, similar works are presented, then the activation methods and key derivation are discussed. Some basic ideas and an AES overview are presented in the following section. Next section illustrates the implementation of the AES-128 algorithm using NS-3. The results and discussion of the impact of cryptography on LoRaWAN performance and the concluding remarks are presented in the final sections.

Similar Works

AES-128 is considered to be the block cipher of choice for many applications in the future. However, that does not mean that the communication protocol is secure. Butun et al. [6] conclude that there are multiple attack vectors to LoRaWAN and that the security is dependent on the implementation. The authors state that there are a few critical mechanisms in the implementation that need to be considered.

There has been ample work on LoRaWAN. The literature shows that LoRaWAN version 1.0 has some security vulnerabilities. Many of these vulnerabilities have since been fixed in version 1.1 and have improved the security of LoRaWAN.

In [7], the authors proposed the use of PHYSEC-based key management which is based on physical layer security in LoRaWAN. The authors research showed that it can be a good solution to current key management solutions while having low energy consumption costs when compared to other key management methods.

The work presented in [8] elaborated an experimental performance analysis of the Over-the-Air-Activation (OTAA) procedure using a real LoRaWAN deployment in the field, with the objective of analyzing the delay in activation and energy consumption on a large-scale LoRaWAN. The authors came to the conclusion that high network traffic is a big problem in OTAA activation. Long activation delays occurred (50 % of the devices took more than 2 hours to activate). There were also a high number of packet retransmissions. Three main factors affect the performance of the OTAA procedure: collisions; retransmissions; and the communication request work cycle.

Another proposed secure LoRaWAN backend [9] Server Session Key Generation (S2KG), which uses it to generate network session keys.

As an example, the vulnerability of missing beacon authentication in Class B mechanism and the ADR spoofing attack are controlled using ChirpOTLE [10] by updating the LoRaWAN protocol.

Another research [11] presents a solution based on hybridization between GNU Radio and software-defined radio; this architecture is without LoRaWAN transceivers.

The authors in [12] propose Low-Power AES Data Encryption Architecture (LPADA) for LoRaWAN in physical layer based on different hardware construction. The core of this solution is composed from a low-energy lookup table to complete AES substitution and to optimize the energy consumption in several rounds.

Another interesting contribution [13] illustrates the high-level use with various AES key sizes alongside differing payload dimensions. The findings indicate that the costs associated with delay and energy consumption are moderate, and employing longer key sizes is a viable approach to enhance security.

Naoui et al. [14] assessed the security of the LoRaWAN 1.0 protocol. The authors concluded that the LoRaWAN protocol is susceptible to two potential assaults. The first one is the parameter DevNonce, this is a 16-bit counter that is increased by one with each join request, starting at 0 when the ED is first switched on. The attacker can use replay attacks when the DevNonce is not encrypted. Also, AppNonce is generated when the server receives join-request message from the EDs. After that the AppNonce is passed to both the ED and AS for authentication. In the next message, an attacker can send the ED the relevant join acceptance message which it initiated. The authors designed a trusted third-party computer which is utilized to dispatch the session key for NSs and ASs. The trusted third-party computer creates a timeline, and the NS stores the timeline when it receives a join-request message so as to prevent a replay attack.

Jakub et al. [15] aimed to integrate the fog computing concept into LoRaWAN. The basic tenet of this paradigm is to increase efficiency for massive volumes of data by putting data processing and storage closer to the EDs. In this regard, the authors presented three fog computing-based IoT network architecture. To determine the best architecture, each of the suggested architectures was simulated and compared in terms of service time. By reducing latency, bandwidth, and efficiency, fog computing offers several advantages to IoT sectors. But security concerns must not be overlooked.

According to Qadir et al. [16], EDs that are located on the network edge is a major target for cyber-attackers. In light of safe key management, they therefore provide a remedy known as the Key Generation and Distribution (KGD) method which lessens cyber attacks. There are three steps involved in the KGD algorithm. Initially, it uses a cryptographically safe deterministic random bit generator approach to produce the secret keys. The Elliptic-Curve Diffie-Hellman technique is then used to exchange the produced keys between the ED and Join Server. The Elliptic Curve Digital Signature Algorithm, a key authentication procedure, is subsequently taken into consideration to confirm if the keys were transferred to the authorized parties. The results demonstrate that their suggested KGD has authentication, integrity, and transmission secrecy.

The authors in [17] propose a novel security protocol that reduces the total time required for key creation and renewal. The technique initiates with random pairing locations and utilizes Lagrange interpolation, effectively decreasing the message count while generating a group key. The chain of hashes concept facilitates the renewal

of a group certificate through a single message, thereby negating the necessity for additional message exchanges. The evaluation results indicate that this strategy significantly decreases both the volume of messages and the configuration time relative to prior methods. This enhancement increases the efficiency of secure communication and fortifies overall security by reducing potential vulnerabilities.

The authors in [18] introduce FLoRa, a technique for key generation at the physical layer. The initial key is generated using an adaptive multibit quantization method which enhances the initiation process influences to the rate of bit generation. This minimizes key reconciliation duration and enhances the recovery rate. The method utilizes a robust algorithm to assess channel conditions for the optimization of the key generation process.

To the best of our knowledge, there is a lack of existing research regarding the implementation of the AES-128 standard under network simulators. Furthermore, not much research has been done on how these security paradigms affect ED energy usage.

Activation Methods and Key Derivation

LoRaWAN supports two distinct methods of activating devices: OTAA and Activation by Personalization [19]. Both of these methods are interchangeable. Independently, the activation mechanism for LoRaWAN is explained in the two following Fig. 1, and Fig. 2.

Step 1: The ED consistently initiates the joining procedure in all instances. A join-request message is sent to the network by the final device intending to join. This message includes critical information regarding the device identity and capabilities. This mechanism preserves network integrity by ensuring that each join request is novel and not a repetition of previous attempts (referred to as a replay attack [20]).

The *AppKey* used to calculate the Message Integrity Code (MIC) by using all the fields in the join-request message.

The join-request message is then updated with the computed MIC and is not encrypted, nor is the AppKey transmitted, as illustrated in Fig. 2.

Step 2: The message requesting to join the network is processed and generated by the server (NwkSKey and AppSKey).

Step 3: As part of the normal down-link mode, the NS gives the encrypted join-accept data to the ED, and the NS does not accept the Join-request message, the ED will not receive any response from the server.

Step 4: The role of NS is to maintain the NwkSKey and also distribute AppSKey to the AS.

Step 5: The join-accept information is deciphered by means of the ED via the AES encryption method. Each of the two keys are produced by the ED, the AppSKey and the NwkSKey, using the AppKey and AppNonce [21].

AES Overview

AES employs a symmetric block cipher scheme and offers key lengths of 128, 192, and 256 bits for encryption and decryption [22]; these key lengths determine the number of rounds in the encryption process to meet different environmental needs, as is illustrated in Table 1.

Implementation of AES-128 Algorithm under NS-3

Despite the enormous research and the various works carried out by researchers and labs, in all the literature, according to our knowledge, we do not find the deployment

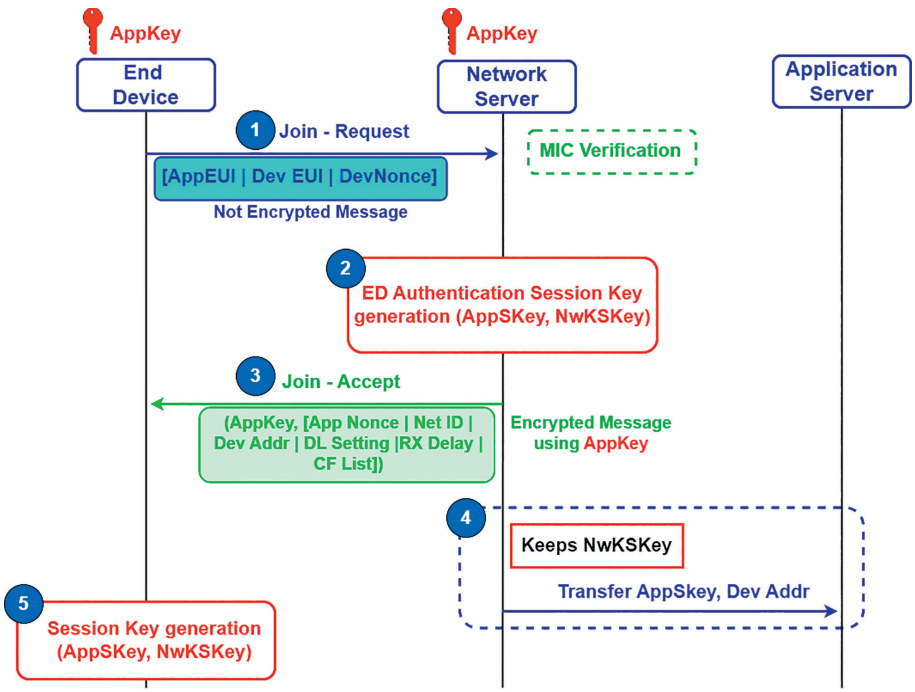

Fig. 1. OTAA message flow in LoRaWAN Network

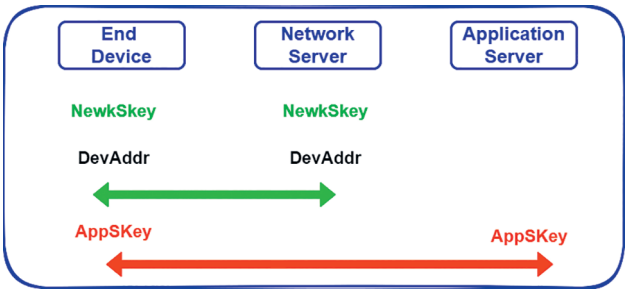

Fig. 2. Pre-sharing DevAddr and session keys for Activation by Personalization

of cryptography using the NS-3 simulator. So, the objective is to implement the AES algorithm under the NS-3 simulator; for this, it will work as follows:

— Modify the application layer code to add encryption;
— Encrypt data before calling the *Send()* function, and decrypt it after receiving the packet.

As illustrated in Fig. 3, at the physical layer, the packet will be split and take just the payload in plaintext, then encrypt only the payload according to the specification [9], using the AES-128 encryption algorithm based on the library accessible via the following link[1], once the ciphertext is successfully received by the receiver (by the NS), it will be decrypted with the same encryption key, and finally the decrypted uplink message will be displayed in plaintext.

The MAC layer is responsible for transmitting the packets, and the helpers are charged with initializing the configuration parameters of each scenario, including the kind of encryption (AES-128, AES-192, and AES-256 [23]). The encryption keys are configured in the *PeriodicSenderHelper*.

[1] Tiny AES in C. Available at: https://github.com/kokke/tiny-AES-c (accessed: 08.06.2024).

Table 1. Key sizes in AES [22]

Parameter	AES-128	AES-192	AES-256
Rounds, numbers	10	12	14
Key sizes, bit	128	192	256
Data block lengths, bit	128	128	128

This operation calls the *setAesKey()* function and instantiates an object of the *PeriodicSender class* which is responsible for encryption. Through the *encrypt()* function, the encryption is done before the packet will be sent.

The *PeriodicSender* class even implements the *Send Packet()* function to call an object of the *LorawanMac class* which implements another *Send()* function whose role is to send messages as shown in Fig. 4.

The *encrypt()* function is between timespec begin and timespec end to calculate CPU time; the time needed by the CPU to execute the encryption as shown in Fig. 5.

On the other side, in the NS side, the same steps are done for the decryption once the message is received. The *NetworkServerHelper* class initializes the shared parameters (Symmetric Encryption) by the *SetDecrypt()* function and calls an object of the NetworkServer model class to do the necessary, as shown in the Fig. 6:

— Receive the message by the *Receive()* function;
— Remove the headers;
— Decrypt the message via the *Decrypt()* function.

The *decrypt()* function aims to decrypt the message received using the same pre-shared encryption key, as illustrated in Fig. 7.

Results and Discussion

To measure the energy consumption, Packet Delivery Rate (PDR) and Time on Air (ToA) induced by the cryptographic primitives used in the LoRaWAN stack according to various packet sizes based on the parameters

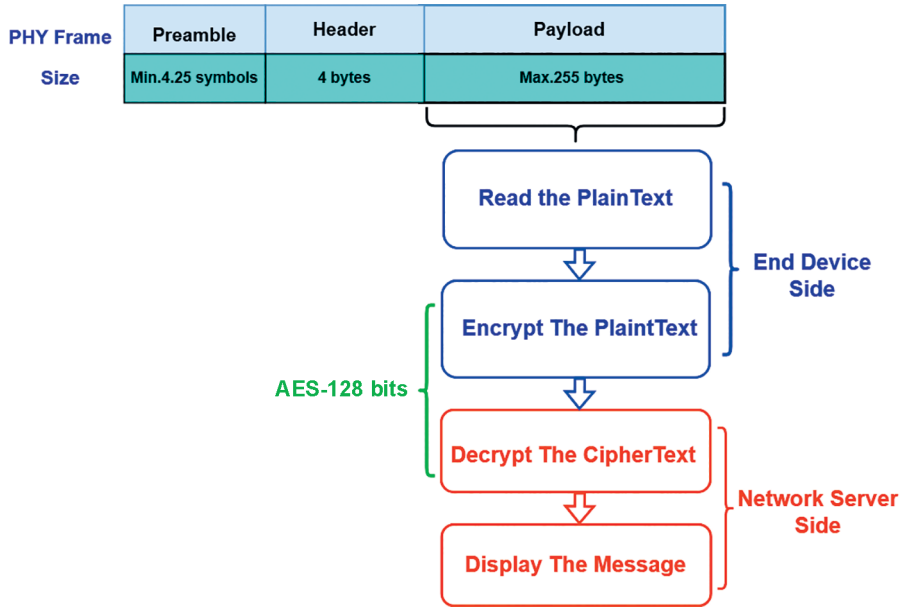

Fig. 3. AES-128 Flow Implementation

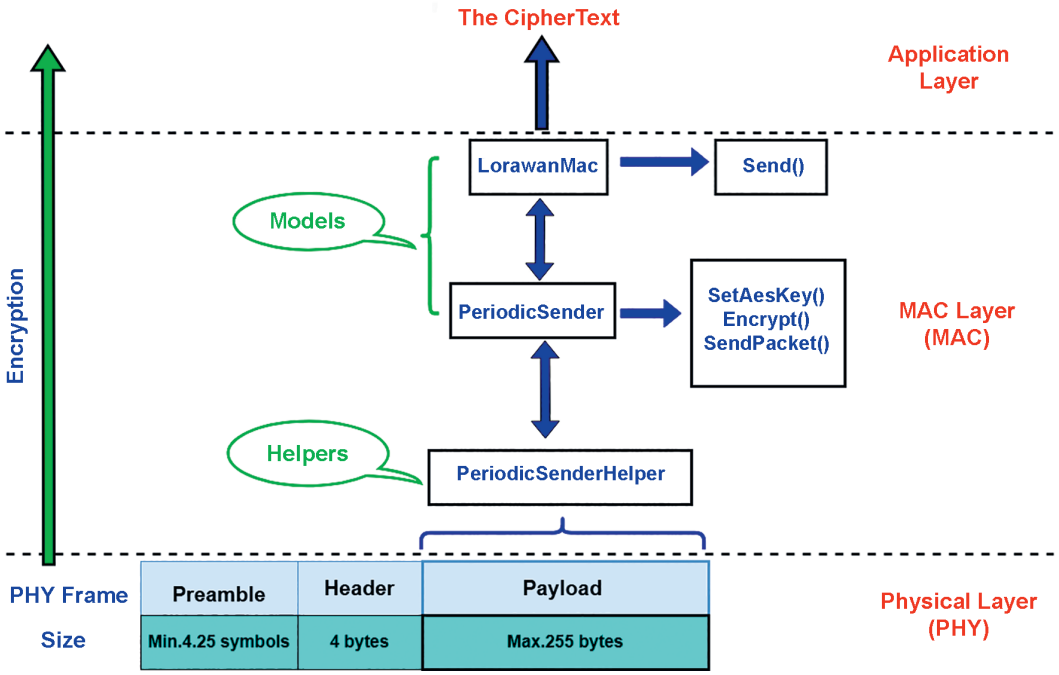

Fig. 4. Encryption process

```
90  void
91  PeriodicSender::SetAesKey (int encryption, unsigned char* msg,size_t input)
92  {
93      m_encryption = encryption;
94      unsigned char* cipher = encrypt(msg,input);
95      m_encryptedMsg = std::string((char*)cipher);
96
97  }
98
99  unsigned char*
100 PeriodicSender::encrypt(unsigned char* shellcode,size_t input)
101 {
102     // beginning timestamp  ++++++++++++++++++++++++++++++
103     struct timespec begin;    timespec_get(&begin, TIME_UTC);
104     struct timespec begin2;    clock_gettime(CLOCK_PROCESS_CPUTIME_ID, &begin2);
105     // end beginning timestamp ---------------------------
```

Fig. 5. *encrypt()* function

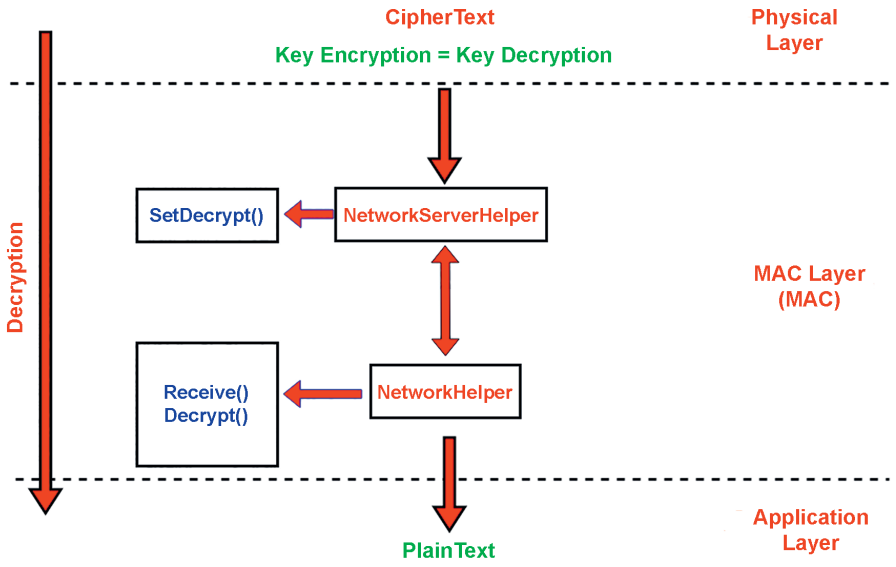

Fig. 6. Decryption process

```
std::string
NetworkServer::decrypt(uint32_t len,  std::string text, int s)
{
  char cc[2];
   unsigned char shellcode [len];
   for (uint32_t i = 0; i < len-1; ++i)
   {
        sprintf(cc,"%c",text[i]);
        shellcode[i] = reinterpret_cast<unsigned char&>(cc);
   }

    unsigned char key[] = "2b7e151628aed2a6abf7158809cf4f3c";
    unsigned char iv[] = "\x9d\x02\x35\x3b\xa3\x4b\xec\x26\x13\x88\x58\x51\x11\x47\xa5\x98";    printf("\n");
```

Fig. 7. decrypt() function

Table. 2. Simulation Parameters

Parameter	Value	Unit
N of Nodes	50	—
Radius	20	m
Period	5	s
Packet Size	12, 24, 32, 64, 128, 192, 216	byte
GateWay (GW)	1	—
Simulation Time	3,600	s
Energy Initial	200	mAh
Battery PD2032	2,664	J
	3.7	V
Simulator	NS-3 (Version 3.35)	—
Operating System	Ubuntu 24.4 64 bit	—

used in Table 2, successfully gathering data from multiple experiments. This information allowed us to analyze the performance metrics comprehensively, providing insights into the efficiency of the cryptographic methods in relation to varying packet sizes and their impact on overall network reliability.

Consider a scenario that sends a packet of 12-byte every 5 s with the maximum transmission power (+14 dBm). For a simulation time of 3,600 s, the PD2032 battery model has a capacity of 4,000 mAh and 2,664 J, with the EDs randomly distributed in a radius of 20 m around a single (01) GW, as the same parameters used in experience [24]. In order to demonstrate the influence of AES-128 cryptography, the difference with and without AES-128 in value is shown in Table 3.

Fig. 8 shows the ToA for different payload lengths (in bytes), using a radius equal to 20 m and a single GW. In LoRaWAN, ToA defines the elapsed time for a LoRaWAN packet between the ED and GW. ToA for different configurations for each packet can be calculated using a formula provided in LoRaWAN specifications [9]. As expected, payload length plays an important role for ToA.

The ToA increases with increasing packet size with and without AES-128 encryption; there is a slight difference between the two histograms.

Since ToA is directly related to the amount of energy, a node needs to spend to transmit the data packet, it is important to determine the battery life of a node. Simulate different scenarios by increasing the packet size ∈ {12, 32, 64, 128, 192, 216}. Fig. 9 examines the energy remaining of nodes in a network during 1 hour of simulation, focusing on packet sizes, with energy measured in joules. A significant decrease in the remaining energy occurs with increasing the packet size from 64 to 216 byte, highlighting the impact of encryption AES-128, packet size, and communication frequency on energy usage.

In Fig. 10, the PDR is calculated based on the packet size where 7 different packet size values are plotted with and without AES-128. The Table 4 shows the difference in value.

As expected, with the packet size increasing, the PDR also decreases for both histograms, either with or

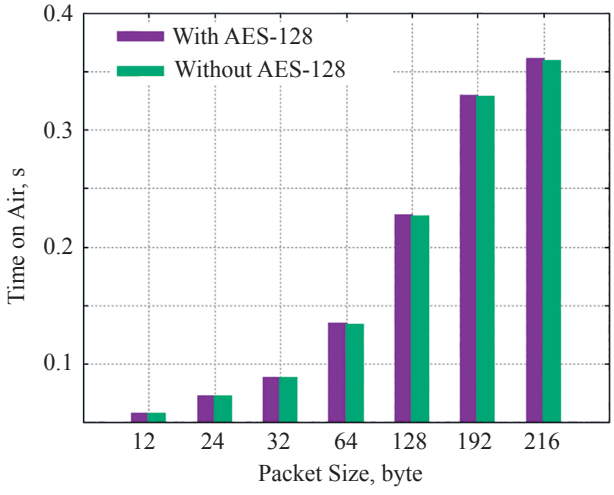

Fig. 8. Time on Air vs. Packet Size

Table 3. Time difference in ToA with and without AES-128, bits

Parameter	12 byte	24 byte	32 byte	64 byte	128 byte	192 byte	216 byte
ToA Without.AES-128	0.0564760	0.0718360	0.0871960	0.133376	0.225536	0.327936	0.358656
ToA With.AES-128	0.0567319	0.0721809	0.0876003	0.133918	0.226560	0.329472	0.360356
Difference, s	0.0002559	0.0003449	0.0004043	0.000542	0.001024	0.001536	0.001700

Table 4. PDR with and without AES-128, bits

Parameter	12 byte	24 byte	32 byte	64 byte	128 byte	192 byte	216 byte
Packet sent with AES-128	30,862	25,300	20,769	13,600	8,050	5,550	5,050
Packet received with AES-128	25,697	19,027	14,724	8,154	3,166	1,665	1,395
Packet sent without AES-128	30,873	25,180	20,840	13,650	8,010	5,750	5,150
Packet received without AES-128	25,697	19,027	14,792	8,192	3,182	1,782	1,442
PDR with AES-128, %	83.26	75.50	70.89	59.95	39.32	30.00	27.62
PDR without AES-128, %	83.23	75.56	70.97	60.01	39.72	30.99	28.00

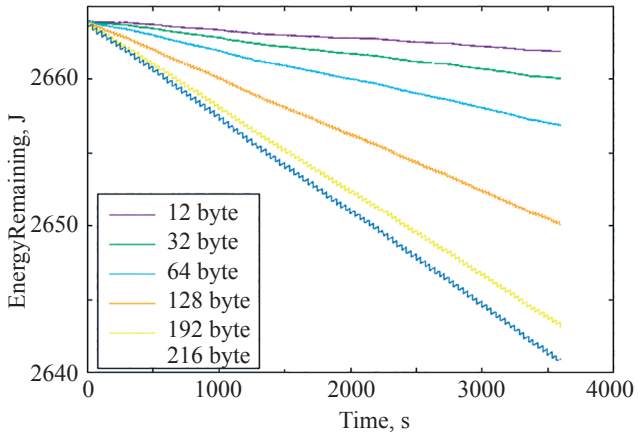

Fig. 9. Energy remaining vs. Time

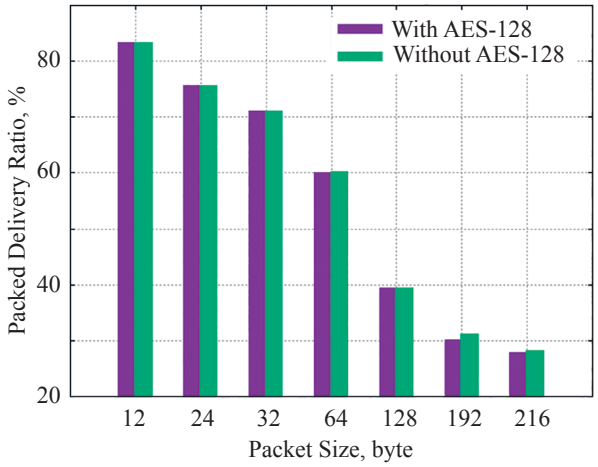

Fig. 10. Packet Delivery Ratio vs. Packet Size

without AES-128 encryption. Notice that for the smallest size of 12 byte with the lowest airtime, almost 83 % PDR is achieved. While for the bigger size of 216 byte, the PDR value drops to only less than 28 % due to the longer packet transmission time, packets are more vulnerable to collisions. In addition, it should be noted that the transmission of packets encrypted with AES-128 impacts the transmission time which is longer than without using AES-128 cryptography and which increases the propensity to collisions and therefore directly impacts the PDR. This

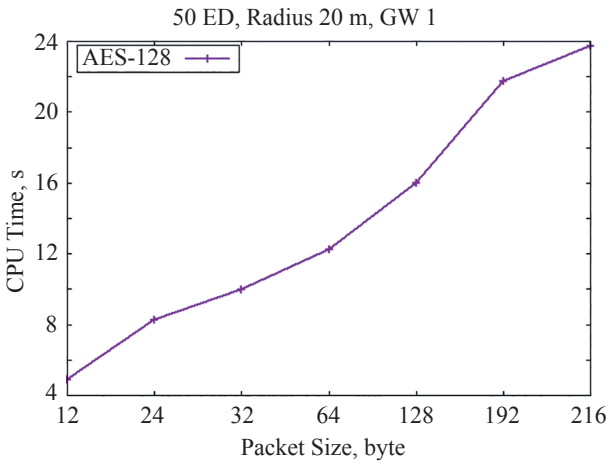

Fig. 11. CPU Time vs. Packet size with AES-128

confirms the results obtained in Fig. 8 when the packet size increases.

Fig. 11 shows the execution time in seconds compared with the packet size using the AES-128 encryption. There is a linear increase up to the maximum size which took 24 s for a packet of 216 bytes.

Conclusion

In this research, we have introduced and extensively evaluated the implementation of the AES-128 encryption standard for Long-Range Wide-Area Networks (LoRaWAN) using the NS-3 simulator. When simulating such systems, it is important to include encryption to make the simulation results more realistic. This contribution lies in the implementation of AES-128. The results show that when transmitting ciphertext, AES-128 suffers an average delay of about 0.7867 ms. A significant decrease in the remaining energy occurs when the packet size increases from 64 to 216 byte. Even if security has some negative effects on network performance, the trade-off is necessary.

Currently, while AES-128 is the standard in LoRaWAN devices, despite its relatively small key size (128 bits), AES-128 is considered secure and is widely used across various industries. It offers sufficient protection for most LoRaWAN use cases. This work offers the way for the integration of other security modes such as AES-192 and AES-256, expanding the application scope of LoRaWAN.

References

1. Mostefa C., Mounir T.A., Abdelmadjid A.M., Nouar A. Ft-CSMA: A fine-tuned CSMA protocol for LoRa-based networks. *Journal of Communications*, 2024, vol. 19, no. 2, pp. 65–77. https://doi.org/10.12720/jcm.19.2.65-77
2. Umbreen S., Shehzad D., Shafi N., Khan B., Habib U. An energy-efficient mobility-based cluster head selection for lifetime enhancement of wireless sensor networks. *IEEE Access*, 2020, vol. 8, pp. 207779–207793. https://doi.org/10.1109/access.2020.3038031
3. Mostefa C., Abdelouahab N., Mounir T.A., Boumerdassi S., Femmam S., Amel Z.A. Formal validation of ADR protocol in LoRaWAN network using Event-b. *Proc. of the 7th International Conference on Computer, Software and Modeling (ICCSM)*, 2023, pp. 11–15. https://doi.org/10.1109/ICCSM60247.2023.00011
4. Sornin N., Luis M., Eirich T., Kramp T., Hersent O. *LoRaWAN Specification*. V. 1. LoRa Alliance Inc., 2015, 82 p.
5. LoRaWAN® L2 1.0.4 Specification (TS001-1.0.4). *Lora Alliance Technical Committee*, 2020, 90 p.
6. Butun I., Pereira N., Gidlund M. Analysis of LoRaWAN v1.1 security: research paper. *Proc. of the 4th ACM MobiHoc Workshop on Experiences with the Design and Implementation of Smart Objects*, 2018, pp. 1–6. https://doi.org/10.1145/3213299.3213304
7. Andreas W., de la Fuente A.G., Christoph L., Michael K. Physical layer security based key management for LoRaWAN. *arXiv*, 2021, arXiv:2101.02975. https://doi.org/10.48550/arXiv.2101.02975
8. El Fehri C., Baccour N., Berthou P., Kammoun I. Experimental analysis of the over-the-air activation procedure in LoRaWAN. *Proc. of the 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2021, pp. 30–35. https://doi.org/10.1109/wimob52687.2021.9606301
9. Tsai K.-L., Leu F.-Y., Hung L.-L., Ko C.-Y. Secure session key generation method for LoRaWAN servers. *IEEE Access*, 2020, vol. 8, pp. 54631–54640. https://doi.org/10.1109/ACCESS.2020.2978100
10. Hessel F., Almon L., Alvarez F. ChirpOTLE: a framework for practical LoRaWAN security evaluation. *Proc. of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2020, pp. 306–316. https://doi.org/10.1145/3395351.3399423
11. Pospisil O., Fujdiak R., Mikhaylov K., Ruotsalainen H., Misurec J. Testbed for LoRaWAN security: design and validation through man-in-the-middle attacks study. *Applied Sciences*, 2021, vol. 11, no. 16, pp. 7642. https://doi.org/10.3390/app11167642
12. Tsai K.-L., Leu F.-Y., You I., Chang S.-W., Hu S.-J., Park H. Low-power AES data encryption architecture for a LoRaWAN. *IEEE Access*, 2019, vol. 7, pp. 146348–146357. https://doi.org/10.1109/access.2019.2941972
13. Thaenkaew P., Quoitin B., Meddahi A. Evaluating the cost of beyond AES-128 LoRaWAN security. *Proc. of the International Symposium on Networks, Computers and Communications (ISNCC)*, 2022, pp. 1–6. https://doi.org/10.1109/isncc55209.2022.9851811
14. Naoui S., Elhdhili M.E., Saidane L.A. Trusted third party based key management for enhancing LoRaWAN security. *Proc. of the IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 2017, pp. 1306–1313. https://doi.org/10.1109/AICCSA.2017.73
15. Jalowiczor J., Rozhon J., Voznak M. Study of the efficiency of fog computing in an optimized LoRaWAN cloud architecture. *Sensors*, 2021, vol. 21, no. 9, pp. 3159. https://doi.org/10.3390/s21093159
16. Qadir J., Butun I., Gastaldo P., Aiello O., Caviglia D.D. Mitigating cyber attacks in LoRaWAN via lightweight secure key management scheme. *IEEE Access*, 2023, vol. 11, pp. 68301–68315. https://doi.org/10.1109/ACCESS.2023.3291420
17. Hanna Y., Cebe M., Leon J., Akkaya K. Efficient group key management for resilient operation of LoRaWAN-based smart grid applications. *IEEE Transactions on Control Systems Technology*, 2024, vol. 32, no. 5, pp. 1706–1717. https://doi.org/10.1109/tcst.2024.3378988
18. Han B., Li Y., Wang X., Li H., Huang J. FLoRa: Sequential fuzzy extractor based physical layer key generation for LPWAN. *Future Generation Computer Systems*, 2023, vol. 140, pp. 253–265. https://doi.org/10.1016/j.future.2022.10.018
19. Islam M., Jamil H.M.M., Pranto S.A., Das R.K., Amin A., Khan A. Future industrial applications: exploring LPWAN-driven IoT protocols. *Sensors*, 2024, vol. 24, no. 8, pp. 2509. https://doi.org/10.3390/s24082509

Литература

1. Mostefa C., Mounir T.A., Abdelmadjid A.M., Nouar A. Ft-CSMA: A fine-tuned CSMA protocol for LoRa-based networks // Journal of Communications. 2024. V. 19. N 2. P. 65–77. https://doi.org/10.12720/jcm.19.2.65-77
2. Umbreen S., Shehzad D., Shafi N., Khan B., Habib U. An energy-efficient mobility-based cluster head selection for lifetime enhancement of wireless sensor networks // IEEE Access. 2020. V. 8. P. 207779–207793. https://doi.org/10.1109/access.2020.3038031
3. Mostefa C., Abdelouahab N., Mounir T.A., Boumerdassi S., Femmam S., Amel Z.A. Formal validation of ADR protocol in LoRaWAN network using Event-b // Proc. of the 7th International Conference on Computer, Software and Modeling (ICCSM). 2023. P. 11–15. https://doi.org/10.1109/ICCSM60247.2023.00011
4. Sornin N., Luis M., Eirich T., Kramp T., Hersent O. LoRaWAN Specification. V. 1. LoRa Alliance, Inc. 2015. 82 p.
5. LoRaWAN® L2 1.0.4 Specification (TS001-1.0.4) // Lora Alliance Technical Committee, 2020. 90 p.
6. Butun I., Pereira N., Gidlund M. Analysis of LoRaWAN v1.1 security: research paper // Proc. of the 4th ACM MobiHoc Workshop on Experiences with the Design and Implementation of Smart Objects. 2018. P. 1–6. https://doi.org/10.1145/3213299.3213304
7. Andreas W., de la Fuente A.G., Christoph L., Michael K. Physical layer security based key management for LoRaWAN // arXiv. 2021. arXiv:2101.02975. https://doi.org/10.48550/arXiv.2101.02975
8. El Fehri C., Baccour N., Berthou P., Kammoun I. Experimental analysis of the over-the-air activation procedure in LoRaWAN // Proc. of the 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). 2021. P. 30–35. https://doi.org/10.1109/wimob52687.2021.9606301
9. Tsai K.-L., Leu F.-Y., Hung L.-L., Ko C.-Y. Secure session key generation method for LoRaWAN servers // IEEE Access. 2020. V. 8. P. 54631–54640. https://doi.org/10.1109/ACCESS.2020.2978100
10. Hessel F., Almon L., Alvarez F. ChirpOTLE: a framework for practical LoRaWAN security evaluation // Proc. of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks. 2020. P. 306–316. https://doi.org/10.1145/3395351.3399423
11. Pospisil O., Fujdiak R., Mikhaylov K., Ruotsalainen H., Misurec J. Testbed for LoRaWAN security: design and validation through man-in-the-middle attacks study // Applied Sciences. 2021. V. 11. N 16. P. 7642. https://doi.org/10.3390/app11167642
12. Tsai K.-L., Leu F.-Y., You I., Chang S.-W., Hu S.-J., Park H. Low-power AES data encryption architecture for a LoRaWAN // IEEE Access. 2019. V. 7. P. 146348–146357. https://doi.org/10.1109/access.2019.2941972
13. Thaenkaew P., Quoitin B., Meddahi A. Evaluating the cost of beyond AES-128 LoRaWAN security // Proc. of the International Symposium on Networks, Computers and Communications (ISNCC). 2022. P. 1–6. https://doi.org/10.1109/isncc55209.2022.9851811
14. Naoui S., Elhdhili M.E., Saidane L.A. Trusted third party based key management for enhancing LoRaWAN security // Proc. of the IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). 2017. P. 1306–1313. https://doi.org/10.1109/AICCSA.2017.73
15. Jalowiczor J., Rozhon J., Voznak M. Study of the efficiency of fog computing in an optimized LoRaWAN cloud architecture // Sensors. 2021. V. 21. N 9. P. 3159. https://doi.org/10.3390/s21093159
16. Qadir J., Butun I., Gastaldo P., Aiello O., Caviglia D.D. Mitigating cyber attacks in LoRaWAN via lightweight secure key management scheme // IEEE Access. 2023. V. 11. P. 68301–68315. https://doi.org/10.1109/ACCESS.2023.3291420
17. Hanna Y., Cebe M., Leon J., Akkaya K. Efficient group key management for resilient operation of LoRaWAN-based smart grid applications // IEEE Transactions on Control Systems Technology. 2024. V. 32. N 5. P. 1706–1717. https://doi.org/10.1109/tcst.2024.3378988
18. Han B., Li Y., Wang X., Li H., Huang J. FLoRa: Sequential fuzzy extractor based physical layer key generation for LPWAN // Future Generation Computer Systems. 2023. V. 140. P. 253–265. https://doi.org/10.1016/j.future.2022.10.018
19. Islam M., Jamil H.M.M., Pranto S.A., Das R.K., Amin A., Khan A. Future industrial applications: exploring LPWAN-driven IoT protocols // Sensors. 2024. V. 24. N 8. P. 2509. https://doi.org/10.3390/s24082509

20. Na S., Hwang D., Shin W., Kim K.-H. Scenario and countermeasure for replay attack using join request messages in LoRaWAN. *Proc. of the International Conference on Information Networking (ICOIN)*, 2017, pp. 718–720. https://doi.org/10.1109/ICOIN.2017.7899580
21. Kang J.-M., Lim D.-W. On the quasi-orthogonality of LoRa modulation. *IEEE Internet of Things Journal*, 2023, vol. 10, no. 14, pp. 12366–12378. https://doi.org/10.1109/jiot.2023.3245885
22. Tsai K.-L., Huang Y.-L., Leu F.-Y., You I., Huang Y.-L., Tsai C.-H. AES-128 based secure low power communication for LoRaWAN IoT environment. *IEEE Access*, 2018, vol. 6, pp. 45325–45334. https://doi.org/10.1109/access.2018.2852563
23. Abboud S., Abdoun N. Enhancing LoRaWAN security: an advanced AES-based cryptographic approach. *IEEE Access*, 2024, vol. 12, P. 2589–2606. https://doi.org/10.1109/ACCESS.2023.3348416
24. Nouar A., Abbes M.T., Boumerdassi S., Chaib M. Impact of mobility model on LoRaWAN performance. *Journal of Communications*, 2024, vol. 19, no. 1. pp. 7–18. https://doi.org/10.12720/jcm.19.1.7-18

20. Na S., Hwang D., Shin W., Kim K.-H. Scenario and countermeasure for replay attack using join request messages in LoRaWAN // Proc. of the International Conference on Information Networking (ICOIN). 2017. P. 718–720. https://doi.org/10.1109/ICOIN.2017.7899580
21. Kang J.-M., Lim D.-W. On the quasi-orthogonality of LoRa modulation // IEEE Internet of Things Journal. 2023. V. 10. N 14. P. 12366–12378. https://doi.org/10.1109/jiot.2023.3245885
22. Tsai K.-L., Huang Y.-L., Leu F.-Y., You I., Huang Y.-L., Tsai C.-H. AES-128 based secure low power communication for LoRaWAN IoT environment // IEEE Access. 2018. V. 6. P. 45325–45334. https://doi.org/10.1109/access.2018.2852563
23. Abboud S., Abdoun N. Enhancing LoRaWAN security: an advanced AES-based cryptographic approach // IEEE Access. 2024. V. 12. P. 2589–2606, https://doi.org/10.1109/ACCESS.2023.3348416
24. Nouar A., Abbes M.T., Boumerdassi S., Chaib M. Impact of mobility model on LoRaWAN performance // Journal of Communications. 2024. V. 19. N 1. P. 7–18. https://doi.org/10.12720/jcm.19.1.7-18

Authors

Abdelouahab Nouar — PhD Student, Hassiba Ben Bouali University (UHBC), LMA Laboratory, Chlef, 02010, Algeria, SC 58865584200, https://orcid.org/0009-0001-3355-1912, a.nouar@univ-chlef.dz
Mounir Tahar Abbes — Professor, Hassiba Ben Bouali University (UHBC), Chlef, 02010, Algeria, SC 57212811077, https://orcid.org/0000-0001-5132-2366, m.taharabbes@univ-chlef.d
Selma Boumerdassi — Professor, Conservatoire National des Arts et Metiers (CNAM), Paris, 75141, France, SC 6602291128, https://orcid.org/0000-0003-2603-2433, selma.boumerdassi@inria.fr
Mostefa Chaib — PhD, Researcher, Hassiba Ben Bouali University (UHBC), LMA Laboratory, Chlef, 02010, Algeria, SC 58835296600, https://orcid.org/0000-0001-9137-9527, m.chaib@univ-chlef.dz

Авторы

Нуар Абделуахаб — аспирант, Университет Асиба Бенбуали Лаборатория ЛМА, Шлеф, 02010, Алжир, SC 58865584200, https://orcid.org/0009-0001-3355-1912, a.nouar@univ-chlef.dz
Тахар Аббес Мунир — профессор, Университет Асиба Бенбуали, Шлеф, 02010, Алжир, SC 57212811077, https://orcid.org/0000-0001-5132-2366, m.taharabbes@univ-chlef.dz
Бумердасси Сельма — профессор, Национальная консерватория искусств и ремесел, Париж, 75141, Франция, SC 6602291128, https://orcid.org/0000-0003-2603-2433, selma.boumerdassi@inria.fr
Хаиб Мостефа — PhD, исследователь, Университет Асиба Бенбуали Лаборатория ЛМЕ, Шлеф, 02010, Алжир, SC 58835296600, https://orcid.org/0000-0001-9137-9527, m.chaib@univ-chlef.dz

Received 13.12.2024
Approved after reviewing 02.09.2025
Accepted 25.09.2025

Статья поступила в редакцию 13.12.2024
Одобрена после рецензирования 02.09.2025
Принята к печати 25.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-933-942
УДК 527.62, 629.05

Решение задачи автономной навигации беспилотного летательного аппарата на основе интеграции инерциальной и оптической систем измерения

Сергей Викторович Соколов[1], Елена Григорьевна Чуб[2]✉

[1] Московский технический университет связи и информатики, Москва, 123423, Российская Федерация
[1,2] Ростовский государственный экономический университет, Ростов-на-Дону, 344002, Российская Федерация

[1] s.v.s.888@yandex.ru, https://orcid.org/0000-0002-5246-841X
[2] elenachub111@gmail.com✉, https://orcid.org/0000-0002-3012-4181

Аннотация

Введение. При построении навигационных систем беспилотных летательных аппаратов основными требованиями к ним являются автономность, точность и миниатюрность исполнения. Автономность навигации беспилотных летательных аппаратов может быть обеспечена с помощью бесплатформенной инерциальной навигационной системы, но ее недостатком является ухудшение точности решения навигационной задачи с течением времени. Для коррекции ошибок бесплатформенной инерциальной навигационной системы используется ее интеграция с различными неинерциальными навигационными системами, среди которых одной из наиболее перспективных с точки зрения выполнения перечисленных требований является система навигации по измерениям оптического потока. Но при традиционном использовании такой системы определяются только составляющие линейной и угловой скоростей беспилотных летательных аппаратов. Подобное определение скоростей является лишь частью общей задачи навигации и не позволяет решить ее в целом. Для поиска решения представлен подход, позволяющий объединить возможности бесплатформенной инерциальной навигационной системы, обеспечивающей решение задачи навигации в целом, и системы навигации по оптическому потоку, позволяющей осуществлять автономное наблюдение параметров линейного и углового движений с минимальными аппаратными затратами. **Метод.** Предложенное решение задачи автономной навигации беспилотных летательных аппаратов получено на основе сильносвязанной интеграции бесплатформенной инерциальной навигационной системы и системы навигации по оптическому потоку с использованием методов стохастической нелинейной фильтрации. Синтез навигационного алгоритма построен на формировании уравнений оцениваемого вектора навигационных параметров по инерциальным измерениям, а уравнений его наблюдателя — по измерениям оптического потока, с последующей реализацией на их основе единого навигационного фильтра, учитывающего дискретный характер используемых измерений. Для оценки полного вектора параметров движения беспилотных летательных аппаратов по измерениям интегрированной инерциально-оптической навигационной системы применен модифицированный расширенный дискретный фильтр Калмана для коррелированных шумов объекта и наблюдателя. **Основные результаты.** Апробация предложенного подхода выполнена на основе численного эксперимента, в ходе которого смоделировано пространственно-угловое движение среднескоростного беспилотного летательного аппарата с одновременным формированием зашумленных измерений параметров его движения. Уровень помех измерения выбран соответствующим уровню помех среднеточных инерциальных и оптических измерителей. Алгоритм оценивания вектора навигационных параметров беспилотного летательного аппарата реализован на основе предложенного модифицированного расширенного дискретного фильтра Калмана. Полученные значения погрешностей оценки всех параметров движения беспилотного летательного аппарата показали возможность выполнения требований к точности не только современных, но и перспективных автономных навигационных систем. **Обсуждение.** Сильносвязанная интеграция инерциальной и оптической навигационных систем по вычислительным затратам и по точности оценки параметров движения оказывается более эффективной в сравнении с традиционным методом определения только составляющих линейной и угловой скоростей объекта по параметрам оптического потока. Основными преимуществами предложенной инерциально-оптической навигационной системы являются автономность и возможность наблюдения всех параметров движения беспилотного летательного аппарата. Устойчивость и точность оценки, простота технической реализации позволяют использовать предложенное

решение для автономной помехоустойчивой навигации беспилотных летательных аппаратов самого различного назначения.

Ключевые слова
бесплатформенная инерциальная навигационная система, параметры оптического потока, инерциально-оптическая навигационная система, стохастическая фильтрация

Благодарности
Работа выполнена в рамках Госзадания № 1023080200012-3-2.3.4.

Ссылка для цитирования: Соколов С.В., Чуб Е.Г. Решение задачи автономной навигации беспилотного летательного аппарата на основе интеграции инерциальной и оптической систем измерения // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 933–942. doi: 10.17586/2226-1494-2025-25-5-933-942

Solving the problem of autonomous drone navigation based on the integration of inertial and optical measurement systems

Sergey V. Sokolov[1], Elena G. Chub[2]✉

[1] Moscow Technical University of Communications and Informatics, Moscow, 123423, Russian Federation
[1,2] Rostov State University of Economics, Rostov-on-Don, 344002, Russian Federation

[1] s.v.s.888@yandex.ru, https://orcid.org/0000-0002-5246-841X
[2] elenachub111@gmail.com✉, https://orcid.org/0000-0002-3012-4181

Abstract
When building drone navigation systems, the main requirements for them are autonomy, accuracy and miniaturization of execution. Drone navigation autonomy can be achieved using strapdown, but its disadvantage is that the accuracy of solving the navigation problem deteriorates over time. To correct strapdown errors, its integration with various non-inertial navigation systems is used, among which one of the most promising in terms of meeting the above requirements is the optical flow measurement navigation system. However, in its traditional use, only the components of the linear and angular velocities of unmanned aerial vehicles are determined. Such a determination of speeds is only part of the overall navigation task and does not allow us to solve it as a whole. In this regard, the article considers an approach that allows combining the capabilities of a free-form inertial navigation system that provides a solution to the navigation problem as a whole, and an optical flow navigation system that allows for autonomous monitoring of linear and angular motion parameters with minimal hardware costs. The proposed solution to the drones autonomous navigation problem is based on the strongly coupled integration of strapdown and an optical flow navigation system using stochastic nonlinear filtering methods. The synthesis of the navigation algorithm is based on the formation of equations of the estimated vector of navigation parameters based on inertial measurements, and the equations of its observer based on optical flow measurements, followed by the implementation of a single navigation filter based on them, taking into account the discrete nature of the measurements used. To estimate the full vector of motion parameters of drones based on measurements of the integrated inertial optical navigation system, a modified extended discrete Kalman filter was used for correlated object and observer noise. The proposed approach was tested on the basis of a numerical experiment during which the spatial and angular motion of a medium-speed drone was modeled with the simultaneous formation of noisy measurements of its motion parameters. The measurement interference level is selected according to the interference level of the medium-range inertial and optical meters. The algorithm for estimating the vector of navigation parameters of the drone is implemented based on the proposed modified extended discrete filter Kalman. The obtained error values for estimating all drone motion parameters have shown that it is possible to meet the accuracy requirements of not only modern, but also promising autonomous navigation systems. The highly coupled integration of inertial and optical navigation systems in terms of computational costs and accuracy of estimating motion parameters turns out to be more effective than the traditional method of determining only the components of the linear and angular velocities of an object based on the parameters of the optical flow. The main advantages of the proposed inertial optical navigation system are autonomy and the ability to monitor all motion parameters of an unmanned aerial vehicle. The stability and accuracy of the assessment, the simplicity of the technical implementation make it possible to use the proposed solution for autonomous noise-resistant navigation of drones for various purposes.

Keywords
strapdown, optical flow parameters, inertial-optical navigation system, stochastic filtering

Acknowledgements
The work was carried out within the framework of the State Assignment No. 1023080200012-3-2.3.4.

For citation: Sokolov S.V., Chub E.G. Solving the problem of autonomous drone navigation based on the integration of inertial and optical measurement systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 933–942 (in Russian). doi: 10.17586/2226-1494-2025-25-5-933-942

Введение

Одной из основных проблем разработки навигационных систем (НС) современных беспилотных летательных аппаратов (БПЛА) является одновременное выполнение таких требований к ним как автономность, точность и миниатюрность исполнения. Традиционное использование спутниковых НС, обеспечивающих выполнение последних двух требований [1–5], оказывается, как правило, невозможным при воздействии

естественных или искусственных радиопомех повышенной интенсивности. В этом случае автономность навигации БПЛА может быть обеспечена с помощью бесплатформенной инерциальной НС (БИНС), основным недостатком которой является ухудшение точности решения навигационной задачи с течением времени. Неизбежные ошибки БИНС, в свою очередь, корректируются различными неинерциальными НС, интегрируемыми с БИНС [5–8], среди которых одной из наиболее перспективных с точки зрения перечисленных требований является НС, использующая измерения оптического потока.

Исследованию данной НС посвящено значительное количество работ [9–23]. Основной идеей всех публикаций на эту тему является поиск наименее затратных в вычислительном отношении и более точных алгоритмов определения составляющих линейной и угловой скоростей объекта по изменению интенсивности оптического потока, снимаемого видеокамерой во время движения. В большинстве работ данная задача решается путем определения соответствия характерных точек на последовательности изображений с использованием различных вероятностных [9], геометрических [10–12] и алгебраических [13, 14] методов. Более универсальным подходом является применение основного уравнения оптического потока [15–18], позволяющего определять так называемые параметры оптического потока, аналитически связанные с проекциями линейной и угловой скоростей объекта [19]. Но имеющая место некорректность задачи идентификации параметров оптического потока привела к многочисленным разработкам, привлекающим дополнительную информацию о регистрируемых изображениях, что неизбежно увеличивает и без того значительный объем вычислительных затрат [20–23]. При этом следует отметить, что определение линейной и угловой скоростей объекта (в рассматриваемом случае — БПЛА) является лишь промежуточным этапом решения общей задачи навигации — текущего позиционирования БПЛА и определения его угловой ориентации, и не позволяет решить ее в целом. В связи с этим рассмотрим подход, обеспечивающий непосредственное определение не промежуточных навигационных параметров БПЛА (проекций линейной и угловой скоростей), а искомых — координат его местоположения и углов ориентации, причем, в условиях неизбежных помех измерения.

Постановка задачи

Для решения задачи автономной навигации БПЛА актуальна разработка подходов, объединяющих как совместные возможности БИНС, обеспечивающих решение задачи навигации в целом, и систем навигации по оптическому потоку (СНОП), позволяющих наблюдать параметры линейного и углового движений с минимальными аппаратными затратами, так и методов, учитывающих влияние помех при оценке навигационных параметров БПЛА, т. е. методов современной теории стохастической фильтрации [1, 2, 6, 8, 24]. Поставленную проблему будем решать на основе интеграции традиционных БИНС и СНОП. Для возможности последующего решения используем следующие системы координат (СК) [2, 6, 8, 25]:

— приборную СК (ПСК) $J\,Oxyz$ с началом в центре масс БПЛА, оси которой направлены по взаимно ортогональным осям чувствительности чувствительных элементов, образующих измерительный комплекс БИНС;
— невращающуюся инерциальную СК (ИСК) $I\,O\xi\eta\zeta$ с началом в центре сферы Земли, осью $O\eta$, совпадающей с вектором угловой скорости вращения Земли Ω; осью $O\zeta$, лежащей в начальный момент времени в плоскости нулевого меридиана, и осью $O\xi$, дополняющей СК до правой;
— сопровождающую (ССК) $S\,OXYZ$ с началом в центре масс БПЛА, ось OY которой лежит в плоскости местного меридиана, ось OZ направлена от центра Земли по местной вертикали, а ось OX дополняет СК до правой;
— СК, связанную с видеокамерой (ВСК) $OX^*Y^*Z^*$ (рис. 1), где $x^* = f\dfrac{X^*}{Z^*}$, $y^* = f\dfrac{Y^*}{Z^*}$ — известные координаты проекции точки сканирования P на плоскость изображения, f — фокусное расстояние; Z^* — глубина изображения (проекция точки сканирования P); V_{X^*}, V_{Y^*}, V_{Z^*} — проекции вектора линейной скорости центра масс БПЛА на оси ВСК; ω_{X^*}, ω_{Y^*}, ω_{Z^*} — проекции вектора угловой скорости БПЛА на оси ВСК. При этом оси ориентированы следующим образом: OZ^* совпадает с линией визирования видеокамеры; начало ВСК — с началом ПСК; в начальный момент времени OX^* ВСК — с направлением оси Oy ПСК; OY^* ВСК — с направлением оси Ox ПСК; OZ^* ВСК направлена противоположно оси Oz ПСК.

Состав измерительного комплекса БИНС традиционный — содержит три акселерометра и три датчика угловой скорости (ДУС), оси чувствительности которых направлены по осям ПСК.

Описание предлагаемого метода

Дискретная модель БИНС. В качестве модели БИНС используем уравнения БИНС в общей нелиней-

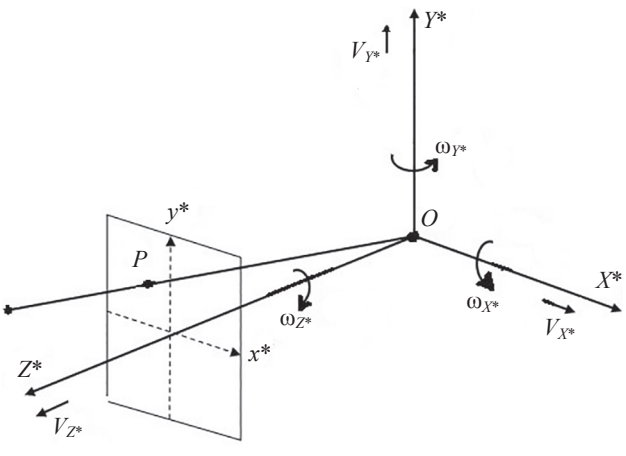

Рис. 1. Система координат, связанная с видеокамерой
Fig. 1. The coordinate system associated with the video camera

ной форме [8, 25]. Так как обработка информации в бортовых навигационных процессорах происходит в цифровом виде, рекуррентные уравнения БИНС представим в дискретной форме:

$$\begin{vmatrix}\alpha_k\\ \beta_k\\ \gamma_k\end{vmatrix} = \begin{vmatrix}\alpha_{k-1}\\ \beta_{k-1}\\ \gamma_{k-1}\end{vmatrix} + \tau \begin{vmatrix} \dfrac{\sin\gamma_{k-1}}{\cos\beta_{k-1}} & \dfrac{\cos\gamma_{k-1}}{\cos\beta_{k-1}} & 0\\ \cos\gamma_{k-1} & -\sin\gamma_{k-1} & 0\\ \sin\gamma_{k-1}\operatorname{tg}\beta_{k-1} & \cos\gamma_{k-1}\operatorname{tg}\beta_{k-1} & 1\end{vmatrix} \times$$
$$\times (\mathbf{Z}_{d_k} - \mathbf{W}_{d_k}),$$

$$\begin{vmatrix}\lambda_k\\ \varphi_k\end{vmatrix} = \begin{vmatrix}\lambda_{k-1}\\ \varphi_{k-1}\end{vmatrix} + \tau \begin{vmatrix}(\cos\varphi_{k-1})^{-1} & 0\\ 0 & 1\end{vmatrix} \begin{vmatrix}V_{X_{k-1}}\\ V_{Y_{k-1}}\end{vmatrix} (r + h_{k-1})^{-1},$$
$$h_k = h_{k-1} + \tau V_{Z_{k-1}},$$

$$\begin{vmatrix}V_{X_k}\\ V_{Y_k}\\ V_{Z_k}\end{vmatrix} = \begin{vmatrix}V_{X_{k-1}}\\ V_{Y_{k-1}}\\ V_{Z_{k-1}}\end{vmatrix} + \tau \left\{ \mathbf{C}^T(\alpha_{k-1}, \beta_{k-1}, \gamma_{k-1}, \lambda_{k-1}, \varphi_{k-1}) \times \right. \quad (1)$$

$$\times \mathbf{Z}_{a_k} + \left(\left(2 \begin{vmatrix}0\\ \Omega\cos\varphi_{k-1}\\ \Omega\sin\varphi_{k-1}\end{vmatrix} + (r + h_{k-1})^{-1} \begin{vmatrix}-V_{Y_{k-1}}\\ V_{X_{k-1}}\\ V_{X_{k-1}}\operatorname{tg}\varphi_{k-1}\end{vmatrix} \right) \times \right.$$
$$\left. \times \begin{vmatrix}V_{X_{k-1}}\\ V_{Y_{k-1}}\\ V_{Z_{k-1}}\end{vmatrix} \right) - \begin{vmatrix}0\\ -\Omega^2 (r + h_{k-1})\cos\varphi_{k-1}\sin\varphi_{k-1}\\ \Omega^2 (r + h_{k-1})\cos^2\varphi_{k-1} + g\end{vmatrix} -$$
$$\left. - \mathbf{C}^T(\alpha_{k-1}, \beta_{k-1}, \gamma_{k-1}, \lambda_{k-1}, \varphi_{k-1})\, \mathbf{W}_{a_k} \right\},$$

где τ — временной шаг; α_k, β_k, γ_k — углы Эйлера–Крылова, определяющие ориентацию ПСК относительно ИСК на k-ом шаге; $\mathbf{Z}_{d_k}$ — вектор измерений трех ортогональных ДУС на k-ом шаге; $\mathbf{W}_{d_k}$ — вектор помех измерения ДУС на k-ом шаге, аппроксимируемый в дальнейшем центрированной гауссовской последовательностью с матрицей дисперсий $\mathbf{D}_{d_k}$; λ_k — долгота; φ_k — широта центра масс БПЛА на k-ом шаге; $|V_{X_k}\ V_{Y_k}\ V_{Z_k}|^T$ — вектор скорости БПЛА относительно Земли на k-ом шаге; r — радиус Земли; h_k — высота БПЛА на k-ом шаге; Ω — скорость вращения Земли; g — гравитационное ускорение; $\mathbf{Z}_{a_k}$ — вектор выходных сигналов акселерометров на k-ом шаге; $\mathbf{W}_{a_k}$ — вектор помех акселерометров на k-ом шаге, аппроксимируемый центрированной гауссовской последовательностью с матрицей дисперсий $\mathbf{D}_{a_k}$; $\mathbf{C}(\alpha_{k-1}, \beta_{k-1}, \gamma_{k-1}, \lambda_{k-1}, \varphi_{k-1}) = \mathbf{D}(\alpha_{k-1}, \beta_{k-1}, \gamma_{k-1}) \times \mathbf{B}^T(\lambda_{k-1}, \varphi_{k-1})$ — матрица поворота 2-го рода [25] ПСК относительно ССК, матрицы поворота: $\mathbf{D}(\alpha_{k-1}, \beta_{k-1}, \gamma_{k-1})$ — ПСК относительно ИСК, $\mathbf{B}(\lambda_{k-1}, \varphi_{k-1})$ — ССК относительно ИСК [8].

В векторной форме уравнения (1), полученные при самых общих предположениях о траектории движения БПЛА и стохастическом характере помех чувствительных элементов БИНС, можно представить как:

$$\mathbf{Y}_k = \mathbf{F}(\mathbf{Y}_{k-1}, k-1) + \mathbf{F}_1(\mathbf{Y}_{k-1}, k-1)\mathbf{W}_{\mathbf{Y}_k}, \quad (2)$$

где $\mathbf{Y}_k = |\alpha_k\ \beta_k\ \gamma_k\ \lambda_k\ \varphi_k\ V_{X_k}\ V_{Y_k}\ V_{Z_k}\ h_k|^T$, $\mathbf{Y}(0) = \mathbf{Y}_0$, $\mathbf{W}_{\mathbf{Y}_k} = |W_{d_k}^T\ W_{a_k}^T|^T$; $\mathbf{F}(\mathbf{Y}_{k-1}, k-1)$ — вектор-функция, определяющая регулярную составляющую вектора состояния БИНС; $\mathbf{F}_1(\mathbf{Y}_{k-1}, k-1)$ — матрица, определяющая влияние помех измерения на динамику вектора состояния БИНС.

Полученная модель (2) является полной дискретной стохастической моделью БИНС, обеспечивающей возможность ее использования при построении интегрированной НС в качестве модели вектора навигационных параметров. Необходимость интеграции обусловлена тем обстоятельством, что измерения *всех* чувствительных элементов БИНС уже использованы при описании динамики полного вектора навигационных параметров в уравнении (2). А для оценки этих параметров с применением методов стохастической фильтрации обязательно наличие наблюдателя вектора состояния системы, т. е. дополнительных измерений, несущих информацию обо всех оцениваемых переменных модели (2). В качестве такого наблюдателя параметров движения БПЛА рассмотрим СНОП, обладающую такими преимуществами как конструктивная простота и автономность.

Модель СНОП. В современных СНОП используются различные методы оценки параметров движения объекта путем обработки информации, содержащейся в изменении интенсивности, цветовой гаммы, контраста отраженного светового излучения [15, 17, 20]. Одним из основных подходов к оценке параметров движения является вычисление поля скоростей оптического потока по монокулярным изображениям [10–23]. На сегодняшний день эта задача решается в детерминированной постановке [16–23]: непосредственным вычислением вектора оптического потока из так называемого основного уравнения оптического потока с последующим определением параметров движения путем решения соответствующей системы уравнений.

Под оптическим потоком здесь понимается его классическое определение как изображения видимого движения объектов, поверхностей или краев сцены, получаемого в результате перемещения наблюдателя (в частности, камеры) относительно сцены. Или, формализуя данное определение: если $I_0(x, y)$ — первый кадр видео (функция интенсивности от координаты точки на изображении), а $I_1(x, y)$ — второй кадр, содержащий те же самые точки, только смещенные (интенсивность сохранена, края изображения не рассматриваются), то векторное поле $V(x, y) = (u(x, y), v(x, y))$: $I_1(x, y) = I_0(x + u, y + v)\ \forall\ (x, y)$ и есть оптический поток. Главная особенность компонентов $u(x, y)$, $v(x, y)$ векторного поля $V(x, y)$ («параметров оптического потока») — наличие аналитической зависимости от линейной и угловой скоростей движения объекта, вычисляемых по этим параметрам, в связи с чем задача их оценки с высокой точностью, причем, в условиях неизбежных помех видеосъемки, является одной из основных в теории и практике технического зрения.

Исходным уравнением для определения параметров оптического потока $u(x, y)$, $v(x, y)$ является основное уравнение оптического потока [15–19]:

$$\frac{\partial I}{\partial t} + u\frac{\partial I}{\partial x} + v\frac{\partial I}{\partial y} = 0,$$

где u, v — компоненты скорости яркостной картины соответственно в направлениях x и y, взятые в точке сканирования P; $\frac{\partial I(x, y, t)}{\partial t}$, $\frac{\partial I(x, y, t)}{\partial x}$, $\frac{\partial I(x, y, t)}{\partial y}$ — частные производные функции яркости наблюдаемой поверхности, которые оцениваются непосредственно из изображения.

Здесь важно отметить, что использование детерминированного подхода в описании движения объекта при приближенном решении уравнения оптического потока позволяет вычислить лишь *масштабированные* линейную и угловую скорости, а также, что более существенно, не обеспечивает принципиальной возможности формирования точного решения в связи с неизбежной зашумленностью реальных изображений и методической погрешностью подхода [10–23]. В связи с этим рассмотрим иной подход к оценке навигационных параметров БПЛА — на основе интегрирования СНОП с БИНС и использования измерений СНОП в качестве вектора наблюдений навигационных параметров, описываемых уравнением БИНС (2), с последующим применением аппарата теории стохастической фильтрации.

Для построения модели наблюдателя вектора навигационных параметров, формируемых СНОП, применим уравнения, связывающие значения вектора оптического потока с параметрами движения БПЛА [16–23], которые в ВСК имеют вид:

$$\begin{vmatrix} u \\ v \end{vmatrix} = \begin{vmatrix} f\left(-\frac{V_{X^*}}{Z^*} - \omega_{Y^*} + \omega_{Z^*} y^*\right) - fx^*\left(-\frac{V_{Z^*}}{Z^*} - \omega_{X^*} y^* + \omega_{Y^*} x^*\right) \\ f\left(-\frac{V_{Y^*}}{Z^*} - \omega_{Z^*} x^* + \omega_{X^*}\right) - fy^*\left(-\frac{V_{Z^*}}{Z^*} - \omega_{X^*} y^* + \omega_{Y^*} x^*\right) \end{vmatrix} =$$

$$= \frac{f}{Z^*}\begin{vmatrix} -1 & 0 & x^* \\ 0 & -1 & y^* \end{vmatrix}\begin{vmatrix} V_{X^*} \\ V_{Y^*} \\ V_{Z^*} \end{vmatrix} + f\begin{vmatrix} y^* x^* & -(x^{*2}+1) & y^* \\ (y^{*2}+1) & -y^* x^* & -x^* \end{vmatrix}\begin{vmatrix} \omega_{X^*} \\ \omega_{Y^*} \\ \omega_{Z^*} \end{vmatrix},$$

где u, v — параметры оптического потока.

Предположим, что при использовании СНОП оптическая информация снимается с поверхности Земли (линия визирования видеокамеры в невозмущенном состоянии направлена к центру Земли).

В этом случае в проекциях на оси ПСК уравнения вектора оптического потока примут вид:

$$\begin{vmatrix} u \\ v \end{vmatrix} = \begin{vmatrix} f\left(\frac{V_y}{Z} - \omega_x + \omega_z x\right) + fy\left(-\frac{V_z}{Z} + \omega_y x - \omega_x y\right) \\ f\left(\frac{V_x}{Z} - \omega_z y + \omega_y\right) + fx\left(-\frac{V_z}{Z} + \omega_y x - \omega_x y\right) \end{vmatrix} =$$

$$= \frac{f}{Z}\begin{vmatrix} 0 & 1 & -y \\ 1 & 0 & -x \end{vmatrix}\begin{vmatrix} V_x \\ V_y \\ V_z \end{vmatrix} + f\begin{vmatrix} -(y^2+1) & yx & x \\ -yx & x^2+1 & -y \end{vmatrix}\begin{vmatrix} \omega_x \\ \omega_y \\ \omega_z \end{vmatrix},$$

где x, y — известные координаты проекции точки сканирования P на плоскость изображения; $Z = Z(x, y)$ — глубина точки сканирования P; V_x, V_y, V_z — проекции вектора линейной скорости центра масс БПЛА на оси ПСК; ω_x, ω_y, ω_z — проекции вектора угловой скорости БПЛА на оси ПСК.

Для возможности использования полученных уравнений в качестве уравнений наблюдателя вектора $\mathbf{Y}_k$ представим входящие в них переменные в дискретной форме в ССК.

В случае невозмущенного движения видеокамеры при сканировании подстилающей поверхности глубина изображения точки сканирования P определяется как $Z_k(x, y) = \sqrt{x^2 + y^2 + h_k^2}$, следовательно, в общем случае — при вращении приборного трехгранника (т. е. видеокамеры), глубина изображения трансформируется следующим образом: $Z_k(x, y) = \sqrt{x^2 + y^2 + \frac{h_k^2}{c_{33}^2(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k)}}$, где $c_{33}(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k)$ — соответствующий элемент матрицы $\mathbf{C}(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k)$. В свою очередь, векторы линейной и угловой скоростей (с учетом угловой скорости вращения Земли, а также вращения сопровождающего трехгранника за счет движения центра масс объекта) в ПСК можно представить как:

$$|V_{x_k}\ V_{y_k}\ V_{z_k}|^T = \mathbf{C}(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k)|V_{x_k}, V_{y_k}, V_{z_k}|^T,$$

$$|\omega_{x_k}\ \omega_{y_k}\ \omega_{z_k}|^T = \mathbf{Z}_{d_k} - \mathbf{C}(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k) \times$$

$$\times\left\{\begin{vmatrix} 0 \\ \Omega\cos\varphi_k \\ \Omega\sin\varphi_k \end{vmatrix} + \tau(r + h_k)^{-1}\begin{vmatrix} V_{Y_k} \\ V_{X_k}(\cos\varphi_k)^{-1} \\ 0 \end{vmatrix}\right\} - \mathbf{W}_{d_k},$$

при этом вектор $(\mathbf{Z}_{d_k} - \mathbf{W}_{d_k})$ (вектор угловой скорости в ИСК) для улучшения наблюдаемости параметров углового движения может быть выражен из уравнений (1) через углы Эйлера–Крылова следующим образом:

$$\mathbf{Z}_{d_k} - \mathbf{W}_{d_k} = \tau^{-1}\mathbf{\Phi}^{-1}(\beta_{k-1}, \gamma_{k-1})\left(\begin{vmatrix} \alpha_k \\ \beta_k \\ \gamma_k \end{vmatrix} - \begin{vmatrix} \alpha_{k-1} \\ \beta_{k-1} \\ \gamma_{k-1} \end{vmatrix}\right),$$

$$\mathbf{\Phi}^{-1}(\beta_{k-1}, \gamma_{k-1}) = \begin{vmatrix} \sin\gamma_{k-1}\cos\beta_{k-1} & \cos\gamma_{k-1} & 0 \\ \cos\gamma_{k-1}\cos\beta_{k-1} & -\sin\gamma_{k-1} & 0 \\ -\sin\beta_{k-1} & 0 & 1 \end{vmatrix},$$

где $\mathbf{\Phi}^{-1}$ — матрица, обратная матрице $\mathbf{\Phi}$.

Тогда окончательно уравнение вектора оптического потока в функции оцениваемого навигационного вектора $\mathbf{Y}_k$ примет вид:

$$\begin{vmatrix} u_k \\ v_k \end{vmatrix} = f\left(x^2 + y^2 + \frac{h_k^2}{c_{33}^2(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k)}\right)^{-1/2}\begin{vmatrix} 0 & 1 & -y \\ 1 & 0 & -x \end{vmatrix} \times$$

$$\times\ \mathbf{C}(\alpha_k, \beta_k, \gamma_k, \lambda_k, \varphi_k)\begin{vmatrix} V_{X_k} \\ V_{Y_k} \\ V_{Z_k} \end{vmatrix} +$$

$$+ f\begin{vmatrix} -(y^2+1) & yx & x \\ -yx & x^2+1 & -y \end{vmatrix}\left\{\tau^{-1}\mathbf{\Phi}^{-1}(\beta_{k-1}, \gamma_{k-1}) \times\right.$$

$$\times\left(\begin{vmatrix}\alpha_k\\ \beta_k\\ \gamma_k\end{vmatrix}-\begin{vmatrix}\alpha_{k-1}\\ \beta_{k-1}\\ \gamma_{k-1}\end{vmatrix}\right)-\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\times$$

$$\times\left(\begin{vmatrix}0\\ \Omega\cos\varphi_k\\ \Omega\sin\varphi_k\end{vmatrix}+\tau(r+h_k)^{-1}\begin{vmatrix}V_{Y_k}\\ V_{X_k}(\cos\varphi_k)^{-1}\\ 0\end{vmatrix}\right)\Bigg\}.$$

В работах [10–15, 20–23] отмечено, что в реальной СНОП компоненты вектора оптического потока u_k, v_k в каждой точке сканирования измеряются с неизбежными погрешностями W_{u_k}, W_{v_k}. В этом случае измеряемый в точке P на k-ом шаге вектор оптического потока $\mathbf{Z}_{CHOП_k}$ имеет вид:

$$\mathbf{Z}_{CHOП_k}=\begin{vmatrix}u_k\\ v_k\end{vmatrix}=f\left(x^2+y^2+\frac{h_k^2}{c_{33}^2(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)}\right)^{-1/2}\times$$

$$\times\begin{vmatrix}0&1&-y\\ 1&0&-x\end{vmatrix}\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\begin{vmatrix}V_{X_k}\\ V_{Y_k}\\ V_{Z_k}\end{vmatrix}+$$

$$+f\begin{vmatrix}-(y^2+1)&yx&x\\ -yx&x^2+1&-y\end{vmatrix}\Bigg\{\tau^{-1}\mathbf{\Phi}^{-1}(\beta_{k-1},\gamma_{k-1})\times \tag{3}$$

$$\times\left(\begin{vmatrix}\alpha_k\\ \beta_k\\ \gamma_k\end{vmatrix}-\begin{vmatrix}\alpha_{k-1}\\ \beta_{k-1}\\ \gamma_{k-1}\end{vmatrix}\right)-\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\times$$

$$\times\left(\begin{vmatrix}0\\ \Omega\cos\varphi_k\\ \Omega\sin\varphi_k\end{vmatrix}+\tau(r+h_k)^{-1}\begin{vmatrix}V_{Y_k}\\ V_{X_k}(\cos\varphi_k)^{-1}\\ 0\end{vmatrix}\right)\Bigg\}+\begin{vmatrix}W_{u_k}\\ W_{v_k}\end{vmatrix}.$$

Так как в общем случае вероятностные распределения погрешностей W_{u_k}, W_{v_k} не определены, то для возможности применения методов нелинейной фильтрации используем усреднение сигналов наблюдения $\mathbf{Z}_{CHOП_{ik}}$ (3), полученных во всех i-х точках (x_i, y_i) сканирования:

$$\mathbf{Z}_{H_k}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_{CHOПi_k}=\frac{f}{N}\times$$

$$\times\sum_{i=1}^{N}\left(x_i^2+y_i^2+\frac{h_k^2}{c_{33}^2(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)}\right)^{-1/2}\begin{vmatrix}0&1&-y_i\\ 1&0&-x_i\end{vmatrix}\times$$

$$\times\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\begin{vmatrix}V_{X_k}\\ V_{Y_k}\\ V_{Z_k}\end{vmatrix}+$$

$$+\begin{vmatrix}-(y^2+1)&yx_i&x_i\\ -yx_i&x_i^2+1&-y_i\end{vmatrix}\Bigg\{\tau^{-1}\mathbf{\Phi}^{-1}(\beta_{k-1},\gamma_{k-1})\times \tag{4}$$

$$\times\left(\begin{vmatrix}\alpha_k\\ \beta_k\\ \gamma_k\end{vmatrix}-\begin{vmatrix}\alpha_{k-1}\\ \beta_{k-1}\\ \gamma_{k-1}\end{vmatrix}\right)-\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\times$$

$$\times\left(\begin{vmatrix}0\\ \Omega\cos\varphi_k\\ \Omega\sin\varphi_k\end{vmatrix}+\tau(r+h_k)^{-1}\begin{vmatrix}V_{Y_k}\\ V_{X_k}(\cos\varphi_k)^{-1}\\ 0\end{vmatrix}\right)\Bigg\}+$$

$$+\begin{vmatrix}\frac{1}{N}\sum_{i=1}^{N}W_{ui_k}\\ \frac{1}{N}\sum_{i=1}^{N}W_{vi_k}\end{vmatrix}.$$

Так как помехи параметров оптического потока равномощны для каждой точки сканирования, то в силу центральной предельной теоремы распределение шумов $W_{u^*_k}=\frac{1}{N}\sum_{i=1}^{N}W_{ui_k}$, $W_{v^*_k}=\frac{1}{N}\sum_{i=1}^{N}W_{vi_k}$ в новом наблюдателе (4) уже при $N\geq 3$ будет близко к гауссовскому, постоянно приближаясь к нему еще более с ростом N [24]. Это дает основание аппроксимировать помехи $W_{u^*_k}$, $W_{v^*_k}$ центрированными гауссовскими последовательностями с дисперсиями D_{u^*k}, D_{v^*k} соответственно. Помимо этого, при окончательном формировании наблюдателя текущих параметров движения (на k-ом шаге) произведем аппроксимацию углов α_{k-1}, β_{k-1}, γ_{k-1} их оценкой: $\hat{\alpha}_{k-1}$, $\hat{\beta}_{k-1}$, $\hat{\gamma}_{k-1}$. В итоге искомый наблюдатель вектора навигационных параметров, обеспечивающий возможность применения существующих методов нелинейной фильтрации [24], запишем в виде:

$$\mathbf{Z}_{\mathbf{H}_k}=\mathbf{A}_1(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k,h_k)\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\times$$

$$\times\begin{vmatrix}V_{X_k}\\ V_{Y_k}\\ V_{Z_k}\end{vmatrix}+\mathbf{A}_2\Bigg\{\tau^{-1}\mathbf{\Phi}^{-1}(\hat{\beta}_{k-1},\hat{\gamma}_{k-1})\times$$

$$\times\left(\begin{vmatrix}\alpha_k\\ \beta_k\\ \gamma_k\end{vmatrix}-\begin{vmatrix}\hat{\alpha}_{k-1}\\ \hat{\beta}_{k-1}\\ \hat{\gamma}_{k-1}\end{vmatrix}\right)-\mathbf{C}(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)\times$$

$$\times\left(\begin{vmatrix}0\\ \Omega\cos\varphi_k\\ \Omega\sin\varphi_k\end{vmatrix}+\tau(r+h_k)^{-1}\begin{vmatrix}V_{Y_k}\\ V_{X_k}(\cos\varphi_k)^{-1}\\ 0\end{vmatrix}\right)\Bigg\}+\begin{vmatrix}W_{u^*_k}\\ W_{v^*_k}\end{vmatrix},$$

$$\mathbf{A}_1(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k,h_k)=\frac{f}{N}\times$$

$$\times\sum_{i=1}^{N}\left(x_i^2+y_i^2+\frac{h_k^2}{c_{33}^2(\alpha_k,\beta_k,\gamma_k,\lambda_k,\varphi_k)}\right)^{-1/2}\begin{vmatrix}0&1&-y_i\\ 1&0&-x_i\end{vmatrix},$$

$$\mathbf{A}_2=\frac{f}{N}\sum_{i=1}^{N}\begin{vmatrix}-(y_i^2+1)&y_ix_i&x_i\\ -y_ix_i&x_i^2+1&-y_i\end{vmatrix},$$

или в общем виде:

$$\mathbf{Z}_{\mathbf{H}_k}=\mathbf{H}_k(\mathbf{Y}_k,k-1)+\mathbf{W}_{\mathbf{H}_k},\ \mathbf{W}_{\mathbf{H}_k}=|W_{u_k}\ \ W_{v_k}|^T. \tag{5}$$

Важной особенностью полученного наблюдателя (5) является явная зависимость его информационной части от *всех* компонентов вектора $\mathbf{Y}_k$, что положительно влияет на сходимость и точность процесса фильтрации.

Уравнения оценки навигационных параметров в интегрированной НС. Полученные уравнения (2) и (5) в классическом виде «объект-наблюдатель» позволяют

окончательно решить поставленную задачу нелинейной фильтрации вектора $\mathbf{Y}_k$ на основе формирования расширенного дискретного фильтра Калмана, использование которого, во-первых, потенциально обеспечивает минимум среднеквадратической ошибки оценки вектора навигационных параметров, а во-вторых, позволяет достичь необходимого компромисса между требуемой точностью и объемом вычислительных затрат, реализуемым в бортовых вычислителях [24]:

$$\hat{\mathbf{Y}}_k = \mathbf{F}(\hat{\mathbf{Y}}_{k-1}, k-1) + \mathbf{K}_k(\mathbf{Z}_{\mathbf{H}_k} - \mathbf{H}_k[\mathbf{F}(\hat{\mathbf{Y}}_{k-1}, k-1)]), \quad (6)$$

где $\hat{\mathbf{Y}}_k$ — оценка вектора состояния системы в k-й момент времени; $\mathbf{F}(\hat{\mathbf{Y}}_{k-1}, k-1)$ и $\mathbf{H}_k[\mathbf{F}(\hat{\mathbf{Y}}_{k-1}, k-1)$ — экстраполированные оценки векторов состояния и наблюдений; $\mathbf{K}_k$ — коэффициент усиления фильтра:

$$\mathbf{K}_k = \mathbf{P}_{k/k-1}\mathbf{h}_k^T(\mathbf{h}_k\mathbf{P}_{k/k-1}\mathbf{h}_k^T + \mathbf{D}_{\mathbf{H}_k})^{-1},$$

$$\mathbf{P}_{k/k-1} = \mathbf{\Phi}_k\mathbf{P}_{k-1}\mathbf{\Phi}_k^T + \mathbf{D}_{\mathbf{Y}_k},\ \mathbf{P}_k = (\mathbf{E} - \mathbf{K}_k\mathbf{h}_k)\mathbf{P}_{k/k-1},$$

где $\mathbf{h}_k = \left.\frac{\partial \mathbf{H}_k(\mathbf{X})}{\partial \mathbf{X}}\right|_{\mathbf{X}=\mathbf{F}(\hat{\mathbf{Y}}_{k-1},k-1)}$, $\mathbf{\Phi}_k = \left.\frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{X}}\right|_{\mathbf{X}=\hat{\mathbf{Y}}_{k-1}}$; $\mathbf{P}_{k/k-1}$ — экстраполированная ковариационная матрица; $\mathbf{P}_k$ — ковариационная матрица в k-й момент времени; $\mathbf{E}$ — единичная матрица,

$$\mathbf{D}_{\mathbf{Y}_k} = \mathbf{F}_1(\hat{\mathbf{Y}}_{k-1}, k-1)\begin{vmatrix} D_d & 0 \\ 0 & D_a \end{vmatrix}\mathbf{F}_1^T(\hat{\mathbf{Y}}_{k-1}, k-1),\ \mathbf{D}_{\mathbf{H}_k} = \begin{vmatrix} D_{uk} & 0 \\ 0 & D_{vk} \end{vmatrix}.$$

Как правило, применение расширенного дискретного фильтра Калмана (6) обеспечивает требуемый компромисс по критерию «точность — вычислительные затраты» при реализации как в вычислителях общего назначения, так и в специализированных бортовых вычислителях. Для иллюстрации эффективности предложенного подхода был проведен следующий численный эксперимент.

Численный эксперимент

Рассматривалось движение БПЛА со следующими параметрами линейного и углового движений в ПСК:

$$V_x = 23 + 2{,}7\cos(0{,}18t),\ V_y = 14 + \sin(0{,}2t),$$
$$V_z = 0{,}12\cos(10^{-2}t)\ (\text{м/с}),$$

$$\omega_x = 0{,}2\cos(0{,}02t),\ \omega_y = 0{,}23\sin(0{,}04t),$$
$$\omega_z = 0{,}02\cos(0{,}01t)\ (^\circ/\text{с})$$

на временном интервале [0, 1000] с из точки с координатами $\lambda_0 = \frac{\pi}{6}$ рад, $\varphi_0 = \frac{\pi}{4}$ рад, $h_0 = 80$ м.

Измерения ДУС моделировались путем аддитивного наложения векторной гауссовской последовательности $\mathbf{W}_d$ с нулевым средним и матрицей дисперсий $\mathbf{D}_d = 10^{-8}\mathbf{E}_3$ (рад/с)^2, $\mathbf{E}_3$ — единичная матрица размерности 3, на вектор проекций угловой скорости в ИСК, определяемый как $\mathbf{D}^T(\alpha_{k-1}, \beta_{k-1}, \gamma_{k-1})|\omega_{x_k}\omega_{y_k}\omega_{z_k}|^T$. Моделирование измерений СНОП осуществлялось в соответствии с алгоритмом усреднения (7) при $f = 25$ мм для 100 пикселов (x_i, y_i) в координатной сетке $\{x{:}[0; 5], y{:}[0; 5]\}$ с шагом 0,5 и использованием в качестве по-

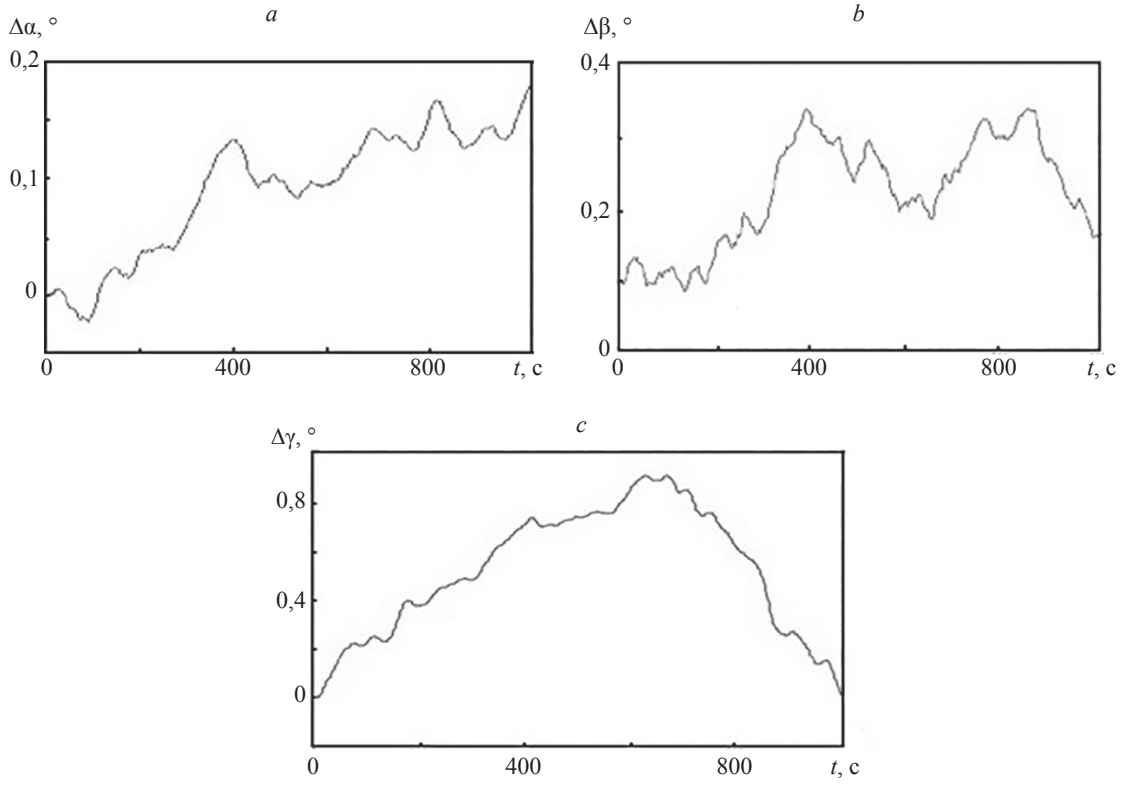

Рис. 2. Усредненные оценки погрешностей углов α (*a*), β (*b*) и γ (*c*)
Fig. 2. Average estimates of angles errors α (*a*), β (*b*), and γ (*c*)

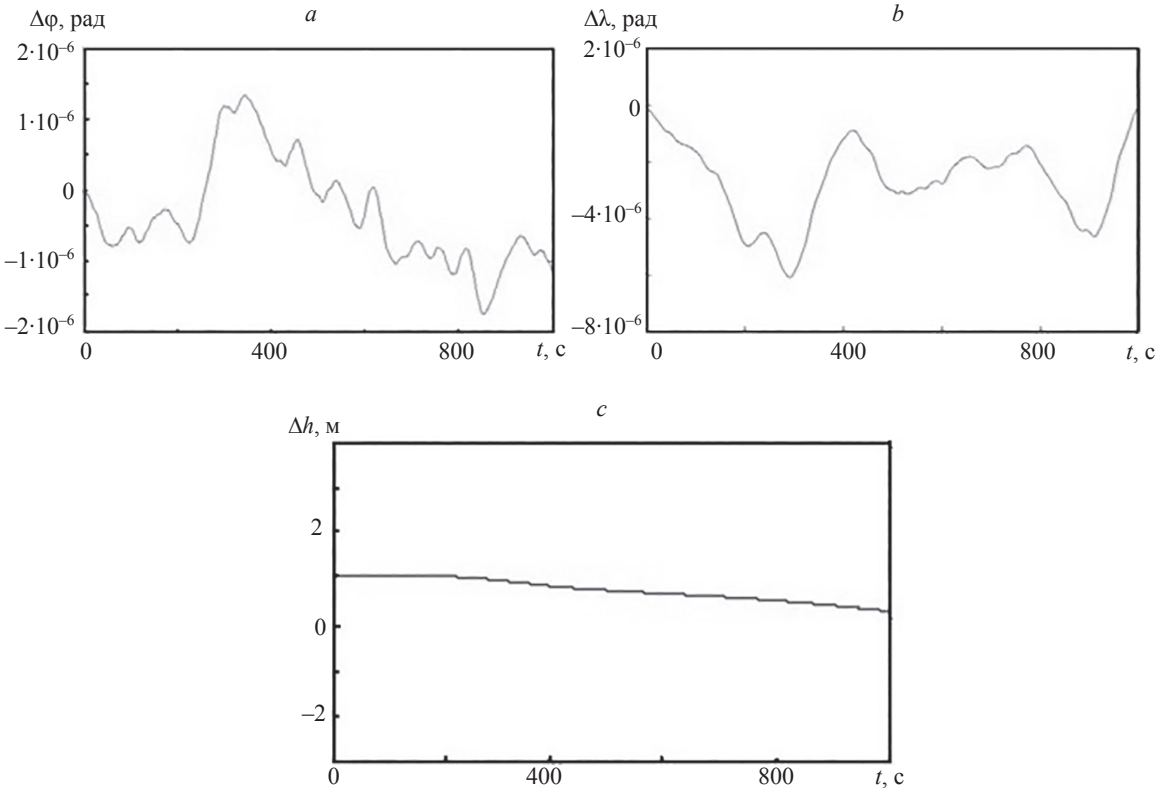

Рис. 3. Усредненные оценки погрешностей широты (*a*), долготы (*b*) и высоты (*c*) беспилотного летательного аппарата

Fig. 3. Average estimates of drone latitude (*a*), longitude (*b*), and height (*c*) errors

мех измерения W_u, W_v гауссовских последовательностей с нулевыми средними и дисперсиями $(D_u, D_v) = (2{,}2\cdot10^{-4})^2$ (м/с)^2. По полученным измерениям СНОП был реализован алгоритм расширенного дискретного фильтра Калмана (9) с шагом 0,01 с.

Навигационные параметры БПЛА оценивались для 15 траекторий с последующим усреднением погрешностей оценивания всех параметров по множеству.

На рис. 2 представлены графики усредненных погрешностей оценок углов α, β, γ, определяющих ориентацию ПСК относительно ИСК, на рис. 3 — графики усредненных погрешностей оценок широты, долготы и высоты БПЛА.

Анализ приведенных значений погрешностей оценивания углов ориентации и координат БПЛА позволяет сделать выводы об устойчивости процесса их оценивания и удовлетворении диапазонов изменения значений погрешностей оценок требованиям, предъявляемым не только к современным, но и перспективным автономным НС БПЛА.

В результате получено, что среднее значение погрешности оценки угла α не превысило на *всем* интервале моделирования 0,2°, а в конце интервала моделирования — 0,18°, для угла β — 0,35° (в конце — 0,2°), для угла γ — 0,9° (в конце — 0,1°); по широте средняя абсолютная погрешность оценки не превысила $1{,}8\cdot10^{-6}$ рад (12 м), а в конце интервала моделирования — $1{,}1\cdot10^{-6}$ рад (7 м)), по долготе — $6\cdot10^{-6}$ рад (в конце — 18 м) ($0{,}5\cdot10^{-6}$ рад (1,5 м)), по высоте — 1 м (в конце — 0,5 м).

Заключение

Полученные результаты позволяют сделать вывод о том, что использование системы навигации по оптическому потоку в качестве наблюдателя параметров движения беспилотного летательного аппарата при комплексировании с бесплатформенной инерциальной навигационной системой оказывается более эффективным по сравнению с непосредственным определением по параметрам оптического потока только составляющих линейной и угловой скоростей — как по вычислительным затратам, так и по точности оценки вектора навигационных параметров беспилотного летательного аппарата [9–23]. В целом, преимущественными особенностями рассмотренной инерциально-оптической навигационной системы являются ее автономность, невысокая размерность наблюдателя (равная 2), а также возможность наблюдения всех параметров движения беспилотного летательного аппарата, что обеспечивает устойчивость и точность оценки его навигационных параметров.

Простота технической реализации данной системы позволяет использовать ее для навигации беспилотных летательных аппаратов самого различного назначения.

Литература

1. Веремеенко К.К., Желтов С.Ю., Ким Н.В. Современные информационные технологии в задачах навигации и наведения беспилотных маневренных летательных аппаратов. М.: Физматлит, 2009. 552 с.
2. ГЛОНАСС. Принципы построения и функционирования / Под ред. А.И. Перова, В.Н. Харисова. М.: Радиотехника, 2010. 800 с.
3. Shaheen E.M. Mathematical analysis for the GPS carrier tracking loop phase jitter in presence of different types of interference signals // Gyroscopy and Navigation. 2018. V. 9. N 4. P. 267–276. https://doi.org/10.1134/s2075108718040077
4. Bhatti J., Humphreys T.E. Hostile control of ships via false GPS signals: demonstration and detection // Navigation. 2017. V. 64. N 1. P. 51–66. https://doi.org/10.1002/navi.183
5. Синютин С.А., Соколов С.В. Решение задачи тесной интеграции инерциально-спутниковых навигационных систем, комплексируемых с одометром // Инженерный вестник Дона. 2014. №4-1 (31). С 74.
6. Емельянцев Г.И., Степанов А.П. Интегрированные инерциально-спутниковые системы ориентации и навигации. СПб.: Концерн «ЦНИИ «Электроприбор», 2016. 394 с.
7. Анучин Н.О., Емельянцев Г.И. Интегрированные системы ориентации и навигации для морских подвижных объектов. СПб.: ГНЦ РФ — ЦНИИ «Электроприбор», 1999. 356 с.
8. Розенберг И.Н., Соколов С.В., Уманский В.И., Погорелов В.А. Теоретические основы тесной интеграции инерциально-спутниковых навигационных систем. М.: Физматлит, 2018. 305 с.
9. Степовой А.В. Методы оценивания вероятности наведения ЛА с пассивным оптико-электронным прибором // Известия высших учебных заведений. Приборостроение. 1999. Т. 42. № 2. С. 40–44.
10. Kitt B., Geiger A., Lategahn H. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme // Proc. of the IEEE Intelligent Vehicles Symposium. 2010. P. 486–492. https://doi.org/10.1109/ivs.2010.5548123
11. Bruss A.R., Horn B.K.P. Passive navigation // Computer Vision, Graphics, and Image Processing. 1983. V. 21. N 1. P. 3–20. https://doi.org/10.1016/s0734-189x(83)80026-7
12. Geiger A., Lenz P., Stiller C., Urtasun R. Vision meets robotics: The KITTI Dataset // The International Journal of Robotics Research. 2013. V. 32. N 11. P. 1231–1237. https://doi.org/10.1177/0278364913491297
13. Raudies F., Neumann H. A review and evaluation of methods estimating ego-motion // Computer Vision and Image Understanding. 2012. V. 116. N 5. P. 606–633. https://doi.org/10.1016/j.cviu.2011.04.004
14. Zhang T., Tomasi C. On the consistency of instantaneous rigid motion estimation // International Journal of Computer Vision. 2002. V. 46. N 1. P. 51–79. https://doi.org/10.1023/a:1013248231976
15. Хорн Б.К.П. Зрение роботов. М.: Мир, 1989. 487 с.
16. Пономарев Е.С., Григорьев А.С. Алгоритмы вычисления оптического потока в задаче определения собственного движения. // Сборник трудов 39-й междисциплинарной школы-конференции ИППИ РАН «Информационные технологии и системы 2015». Сочи: Институт проблем передачи информации им. А.А. Харкевича РАН. 2015. С. 457–470.
17. Baker S., Roth S., Scharstein D., Black M.J., Lewis J.P., Szeliski R. A database and evaluation methodology for optical flow // Proc. of the IEEE 11th International Conference on Computer Vision. 2007. P. 1–8. https://doi.org/10.1109/iccv.2007.4408903
18. Fleet D.J., Weiss Y. Optical flow estimation // Handbook of Mathematical Models in Computer Vision. 2006. P. 237–257. https://doi.org/10.1007/0-387-28831-7_15
19. Sokolov S.V., Shvidchenko S.A., Reshetnikova I.V., Vavilova E.V. Effective estimation of motion parameters of mobile robotic complexes based on information processing of technical vision systems // Proc. of the Systems of signals generating and processing in the field of on board communications. 2025. P. 1–4. https://doi.org/10.1109/ieeeconf64229.2025.10948069
20. Kanatani K. 3-D Interpretation of optical flow by renormalization // International Journal of Computer Vision. 1993. V. 11. N 3. P. 267–282. https://doi.org/10.1007/BF01469345
21. Mirabdollah H., Mertsching B. On the Second Order Statistics of Essential Matrix Elements // Lecture Notes in Computer Science. 2014. V. 8753. P. 547–557. https://doi.org/10.1007/978-3-319-11752-2_45

References

1. Veremeenko K.K., Zheltov S.Iu., Kim N.V. *Modern Information Technologies in the Tasks of Navigation and Guidance of Unmanned Aerial Vehicles*. Moscow, Fizmatlit Publ., 2009, 552 p. (in Russian)
2. Perov A.I., Kharisov V.N. (ed.) GLONASS: Design Concepts and Principles of Operation. Moscow, Radiotehnika Publ., 2010, 800 p. (in Russian)
3. Shaheen E.M. Mathematical analysis for the GPS carrier tracking loop phase jitter in presence of different types of interference signals. *Gyroscopy and Navigation*, 2018, vol. 9, no. 4, pp. 267–276. https://doi.org/10.1134/s2075108718040077
4. Bhatti J., Humphreys T.E. Hostile control of ships via false GPS signals: demonstration and detection. *Navigation*, 2017, vol. 64, no. 1, pp. 51–66. https://doi.org/10.1002/navi.183
5. Siniutin S.A., Sokolov S.V. Solution of the problem of close integration of inertial-satellite navigation systems, complexed with odometer. *Ingineering Journal of Don*, 2014, no. 4-1 (31), pp 74. (in Russian)
6. Emeliantcev G.I., Stepanov A.P *Integrated Inertial-Satellite Systems of Orientation and Navigation*. St. Petersburg, Concern CSRI Elektropribor, 2016, 394 p. (in Russian)
7. Anuchin N.O., Emeliantcev G.I. *Integrated Systems of Orientation and Navigation for Marine Moving Objects*. St. Petersburg, Concern CSRI Elektropribor, 1999, 356 p. (in Russian)
8. Rozenberg I.N., Sokolov S.V., Umanskii V.I., Pogorelov V.A. *Theoretical Framework of the Deep Integration of Inertial-Satellite Navigation Systems*. Moscow, Fizmatlit Publ., 2018, 305 p. (in Russian)
9. Stepovoi A.V. Methods for estimating the probability of an aircraft guidance with a passive optoelectronic device. *Journal of Instrument Engineering*, 1999, vol. 42, no. 2, pp. 40–44. (in Russian)
10. Kitt B., Geiger A., Lategahn H. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. *Proc. of the IEEE Intelligent Vehicles Symposium*, 2010, pp. 486–492. https://doi.org/10.1109/ivs.2010.5548123
11. Bruss A.R., Horn B.K.P. Passive navigation. *Computer Vision, Graphics, and Image Processing*, 1983, vol. 21, no. 1, pp. 3–20. https://doi.org/10.1016/s0734-189x(83)80026-7
12. Geiger A., Lenz P., Stiller C., Urtasun R. Vision meets robotics: The KITTI Dataset. *The International Journal of Robotics Research*, 2013, vol. 32, no. 11, pp. 1231–1237. https://doi.org/10.1177/0278364913491297
13. Raudies F., Neumann H. A review and evaluation of methods estimating ego-motion. *Computer Vision and Image Understanding*, 2012, vol. 116, no. 5, pp. 606–633. https://doi.org/10.1016/j.cviu.2011.04.004
14. Zhang T., Tomasi C. On the consistency of instantaneous rigid motion estimation. *International Journal of Computer Vision*, 2002, vol. 46, no. 1, pp. 51–79. https://doi.org/10.1023/a:1013248231976
15. Horn B.K.P. *Robot Vision*. Mit Pr, 1986, 480 p.
16. Ponomarev E.S., Grigorev A.S. Algorithms for optical flow calculations in the problem of the proper motion determining. *Proc. of the 39th Information Technologies and Systems*, 2015, pp. 457–470. (in Russian)
17. Baker S., Roth S., Scharstein D., Black M.J., Lewis J.P., Szeliski R. A database and evaluation methodology for optical flow. *Proc. of the IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8. https://doi.org/10.1109/iccv.2007.4408903
18. Fleet D.J., Weiss Y. Optical flow estimation. *Handbook of Mathematical Models in Computer Vision*, 2006, pp. 237–257. https://doi.org/10.1007/0-387-28831-7_15
19. Sokolov S.V., Shvidchenko S.A., Reshetnikova I.V., Vavilova E.V. Effective estimation of motion parameters of mobile robotic complexes based on information processing of technical vision systems. *Proc. of the Systems of signals generating and processing in the field of on board communications*, 2025, pp. 1–4. https://doi.org/10.1109/ieeeconf64229.2025.10948069
20. Kanatani K. 3-D Interpretation of optical flow by renormalization. *International Journal of Computer Vision*, 1993, vol. 11, no. 3, pp. 267–282. https://doi.org/10.1007/BF01469345
21. Mirabdollah H., Mertsching B. On the Second Order Statistics of Essential Matrix Elements. *Lecture Notes in Computer Science*, 2014, vol. 8753, pp. 547–557. https://doi.org/10.1007/978-3-319-11752-2_45
22. Xu L., Jia J., Matsushita Y. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine*

22. Xu L., Jia J., Matsushita Y. Motion detail preserving optical flow estimation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012. V. 34. N9. P. 1744–1757. https://doi.org/10.1109/TPAMI.2011.236
23. Zach C., Pock T., Bischof H. A duality based approach for realtime TV-L 1 optical flow // Lecture Notes in Computer Science. 2007. V. 4713. P. 214–223. https://doi.org/10.1007/978-3-540-74936-3_22
24. Тихонов В.И., Харисов В.Н. Статистический анализ и синтез радиотехнических устройств и систем. М.: Радио и связь, 2004. 608 с.
25. Ишлинский А.Ю. Ориентация, гироскопы и инерциальная навигация. М.: Наука, 1976. 670 с.

Intelligence, 2012, vol. 34, no.9, pp. 1744–1757. https://doi.org/10.1109/TPAMI.2011.236
23. Zach C., Pock T., Bischof H. A duality based approach for realtime TV-L 1 optical flow. *Lecture Notes in Computer Science*, 2007, vol. 4713, pp. 214–223. https://doi.org/10.1007/978-3-540-74936-3_22
24. Tikhonov V.I., Kharisov V.N. *Statistical Analysis and Synthesis of Radio Engineering Devices and Systems*. Moscow, Radio i svjaz' Publ., 2004, 608 p. (in Russian)
25. Ishlinskii A.Iu. *Orientation, Gyroscopes, and Inertial Navigation*. Moscow, Nauka Publ., 1976. 670 p. (in Russian)

Авторы

Соколов Сергей Викторович — доктор технических наук, профессор, заведующий кафедрой, Московский технический университет связи и информатики, Москва, 123423, Российская Федерация; профессор, Ростовский государственный экономический университет, Ростов-на-Дону, 344002, Российская Федерация, sc 35235181200, https://orcid.org/0000-0002-5246-841X, s.v.s.888@yandex.ru

Чуб Елена Григорьевна — кандидат технических наук, старший научный сотрудник, Ростовский государственный экономический университет, Ростов-на-Дону, 344002, Российская Федерация, sc 55611768900, https://orcid.org/0000-0002-3012-4181, elenachub111@gmail.com

Authors

Sergey V. Sokolov — D.Sc., Professor, Head of Department, Moscow Technical University of Communications and Informatics, Moscow, 123423, Russian Federation; Professor, Rostov State University of Economics, Rostov-on-Don, 344002, Russian Federation, sc 35235181200, https://orcid.org/0000-0002-5246-841X, s.v.s.888@yandex.ru

Elena G. Chub — PhD, Senior Researcher, Rostov State University of Economics, Rostov-on-Don, 344002, Russian Federation, sc 55611768900, https://orcid.org/0000-0002-3012-4181, elenachub111@gmail.com

Статья поступила в редакцию 26.05.2025
Одобрена после рецензирования 20.08.2025
Принята к печати 30.09.2025

Received 26.05.2025
Approved after reviewing 20.08.2025
Accepted 30.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-943-951
УДК 621.923.74+004.942

Математическая модель движения сферического ротора в процессе доводки чашечными притирами со свободным абразивом

Сергей Николаевич Федорович✉

АО «Концерн «ЦНИИ «Электроприбор», Санкт-Петербург, 197046, Российская Федерация
fedorovichsn@gmial.com✉, https://orcid.org/0009-0001-3147-9910

Аннотация

Введение. В связи с повышением требований к эксплуатационным характеристикам гироскопов с электростатическим неконтактным подвесом ротора возникает необходимость улучшения технологии производства деталей и сборки приборов. Важнейшим компонентом чувствительного элемента электростатического гироскопа является сферический ротор из бериллия. Возмущающие моменты от сил подвеса пропорциональны напряжению, подаваемому на электроды, и отклонению поверхности ротора от сферической формы. По этой причине технология финишной обработки поверхности ротора должна обеспечивать выполнение высоких требований к сферичности ротора. При изготовлении роторов всех известных типов электростатических гироскопов применяется технология бесцентровой доводки чашечными притирами со свободным абразивом. Одним из ключевых факторов, влияющих на получаемую сферичность, являются параметры движения ротора в доводочном станке. Представлена математическая модель, позволяющая определить параметры движения ротора в станке доводки под действием сил трения от вращения чашечных притиров. **Метод.** Процесс бесцентровой доводки чашечными притирами рассматривается как разновидность фрикционного привода. Движение ротора рассматривается как движение абсолютно твердого тела. Для определения параметров движения используются дифференциальные уравнения Эйлера для вращательного движения, решение которых осуществляется численно с использованием программного пакета MATLAB. Распределение давлений в парах притир-ротор рассматривается по аналогии взаимодействий в шаровом шарнире. **Основные результаты.** Показано, что математическая модель движения ротора в процессе доводки чашечными притирами помогает обнаружить основные закономерности движения ротора при бесцентровой доводке и определить граничные условия, при которых должна осуществляться обработка. Предложенная модель позволяет выявить влияние разности в моментах инерции ротора на его движение при обработке, в частности при полировке. **Обсуждение.** Разработанная модель движения ротора может быть использована при проектировании алгоритмов и систем управления станками бесцентровой доводки сфер свободным абразивом, а также в качестве составной части математических и физических моделей, описывающих обработку поверхности ротора методом доводки чашечными притирами.

Ключевые слова

электростатический гироскоп, сферическая доводка, сферический ротор, чашечные притиры, бесцентровая доводка, движение твердого тела

Ссылка для цитирования: Федорович С.Н. Математическая модель движения сферического ротора в процессе доводки чашечными притирами со свободным абразивом // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 943–951. doi: 10.17586/2226-1494-2025-25-5-943-951

Mathematical model of the motion of a spherical rotor during finishing with cup laps and free abrasive

Sergei N. Fedorovich✉

JSC Concern CSRI Elektropribor, Saint Petersburg, 197046, Russian Federation

fedorovichsn@gmail.com✉, https://orcid.org/0009-0001-3147-9910

Abstract

Due to the increasing requirements for the performance characteristics of gyroscopes with electrostatic non-contact rotor suspension, there is a need to improve the technology for manufacturing parts and assembling devices. The most important component of the sensitive element of an electrostatic gyroscope is a spherical beryllium rotor. The disturbing moments from the suspension forces are proportional to the voltage supplied to the electrodes and the deviation of the rotor surface from the spherical shape. For this reason, the technology of finishing the rotor surface must ensure that high requirements for the sphericity of the rotor are met. In the manufacture of rotors of all known types of electrostatic gyroscopes, the technology of centerless finishing with cup laps with free abrasive is used. One of the key factors influencing the resulting sphericity is the parameters of the rotor motion in the finishing machine. The article presents a mathematical model that allows one to determine the parameters of the rotor motion in the finishing machine under the action of friction forces from the rotation of the cup laps. The method of mathematical modeling was used in the work. The process of centerless finishing with cup laps is considered as a type of friction drive. The rotor motion is considered as the motion of an absolutely rigid body. To determine the motion parameters, the Euler differential equations for rotational motion are used the solution of which is carried out numerically using the MATLAB software package. The pressure distribution in the lap-rotor pairs is considered by analogy with the expression of effects in a ball joint. The result of the work is a mathematical model of the rotor motion during finishing with cup laps, which made it possible to identify the main patterns of rotor motion during centerless finishing. The model made it possible to reveal that the difference in the moments of inertia of the rotor can have a significant effect on the rotor motion during processing, in particular, during polishing. Boundary conditions were determined under which the rotor motion can be permissibly considered as the motion of a ball with equal moments of inertia. The proposed model of rotor motion can be used in designing algorithms and control systems for machines for centerless finishing of spheres with free abrasive as well as a component of mathematical and physical models describing the processing of the rotor surface by finishing with cup laps.

Keywords

electrostatic gyroscope, centerless finishing, spherical rotor, lap polishing, motion dynamics

For citation: Fedorovich S.N. Mathematical model of the motion of a spherical rotor during finishing with cup laps and free abrasive. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 943–951 (in Russian). doi: 10.17586/2226-1494-2025-25-5-943-951

Введение

Важнейшим компонентом чувствительного элемента электростатического гироскопа (ЭСГ) является ротор из бериллия [1]. Форма ротора в эксплуатационном состоянии должна быть максимально приближена к сферической, поскольку возмущающие моменты от сил электростатического подвеса ротора пропорциональны напряжению, подаваемому на электроды подвеса и отклонению поверхности ротора от идеальной сферической формы [2]. По этой причине технология финишной обработки поверхности ротора должна обеспечивать выполнение высоких требований сферичности ротора. Для финишной обработки роторов используется технология сферической доводки чашечными притирами со свободным абразивом [3]. Данная технология применяется при изготовлении роторов всех известных типов ЭСГ [4, 5] и в иных проектах, где требуется достижение высокой степени сферичности [6]. При таком виде обработки ротор не имеет жесткой фиксации. Ротор устанавливается в точку пересечения осей и поджимается чашами притиров, тем самым ограничиваются только поступательные степени свободы. При вращении притиров ротор начинает вращаться под действием момента трения, вызванного совместным действием всех притиров. Направление вращения притиров циклично изменяется по заданному алгоритму, обеспечивая изменение положения оси вращения ротора для исключения образования повторяющихся паттернов выработки на роторе и его более равномерной обработки.

Ключевой особенностью процесса сферической доводки является механизм движения обрабатываемой детали — сферического ротора. Силовое взаимодействие между поверхностями притиров и ротора, которое возникает при их относительном движении, порождают силу сопротивления, которая является движущей силой для ротора [7]. Трение притиров о ротор одновременно выступает в роли механизма движения ротора и его абразивной обработки. Таким образом, скорость, ускорение и направление движения, а также и формообразование ротора зависят от параметров движения притиров и их давления на ротор. Существующие современные теоретические концепции абразивной доводки неприменимы для решения конкретных практических задач, таких как доводка роторов ЭСГ, из-за уникальных особенностей этого процесса [8, 9].

Целью исследования является создание математической модели движения ротора в станке сферической доводки, для последующего применения при проектировании алгоритмов и систем управления станками бесцентровой доводки сфер свободным абразивом. Разработка модели может облегчить исследование процесса сферической доводки и сократить количество проводимых экспериментов.

Математическая модель движения ротора при сферической доводке

Механизм движения ротора в притирах можно представить как одну из разновидностей фрикционного привода. В данном случае притиры выполняют роль фрикционов, а сферический ротор является ведомым элементом. Рассмотрим движение ротора на примере доводочного станка, представленного на рис. 1, *а*.

Доводочный станок оснащен четырьмя серводвигателями, расположенными в вершинах тетраэдра — *A*, *B*, *C*, *D*.

На валу каждого серводвигателя находится аксиально подвижные выдвижные штоки, на которых расположены муфты *2*, в которые установлены притиры *3*. Оси валов пересекаются в центре тетраэдра. Выходные валы трех нижних сервоприводов размещены через 120° друг от друга при просмотре сверху вниз и наклонены под углом 19°28′12″ ниже плоскости горизонта. Вал верхнего серводвигателя направлен вниз в точку пересечения трех нижних валов. Чтобы ротор находился в центре тетраэдра, аксиальные силы от притиров должны быть сбалансированы. Это обеспечивается пружинами *1*, которыми притиры *3* поджимаются к ротору. Подвижные муфты *2* позволяют минимизировать поперечные силы, вызванные погрешностями при изготовлении элементов станка и неидеальной сбалансированностью аксиальных сил притиров.

В силу симметрии векторная сумма равных по модулю сил F, действующих по оси любого из притиров, будет уравновешиваться аксиальными силами оставшихся трех притиров. Весом бериллиевого ротора можно пренебречь, поскольку на практике он существенно меньше сил F — вес самого тяжелого ротора ЭСГ составляет не более 15 г.

Допустим, что ротор находится в центре тетраэдра и в процессе обработки центр ротора остается неподвижным. На практике обеспечение этого условия определяется качеством изготовления деталей станка и его конструкцией. Предположим, что притиры и ротор идеально приработаны и поверхность контакта идеализированно представима как поверхность шарового слоя.

Введем условно неподвижную декартову систему координат $Oxyz$, связанную со станиной станка. Начало неподвижной системы координат поместим в точку O пересечения осей притиров. Оси x, y, z направим между осей притиров A, B, C, D (рис. 1, *b*). Введем подвижные координаты, связанные с каждым из притиров $Ol_{k1}l_{k2}l_{k3}$, с началом в точке O, индекс k — обозначение притира (A, B, C или D). Оси l_{k3} направим по оси вращения каждого из притиров.

Определим связь систем координат $Oxyz$ и $Ol_{k1}l_{k2}l_{k3}$ с помощью углов Эйлера [10], используя последовательность поворотов z, y_1, z_2 (рис. 2). В соответствии с кинематикой станка, притиры имеют поступательную и вращательную степени свободы. В рабочем положении, когда притиры подведены к ротору, поступательная степень свободы теряется — силы поджатия компенсируют друг друга, поскольку векторная сумма аксиальных сил по условию равна нулю. Таким образом, для каждого притира остается одна вращательная степень свободы вокруг оси l_{k3}.

Используя компактную запись $c_\alpha = \cos\alpha$, $s_\alpha = \sin\alpha$ и т. д., определим матрицу перехода:

$$\mathbf{R}(\alpha, \beta, \gamma) = \mathbf{R}_{z2}(\gamma)\mathbf{R}_{y1}(\beta)\mathbf{R}_{z}(\alpha) = \begin{bmatrix} c_\alpha c_\beta c_\gamma - s_\alpha s_\gamma & c_\gamma s_\alpha + c_\alpha c_\beta s_\gamma & -c_\alpha s_\beta \\ -c_\alpha s_\gamma - c_\beta c_\gamma s_\alpha & c_\alpha c_\gamma - c_\beta s_\alpha s_\gamma & s_\alpha s_\beta \\ c_\gamma s_\beta & s_\beta s_\gamma & c_\beta \end{bmatrix}, \quad (1)$$

где α, β и γ — углы прецессии, нутации и собственного вращения притира.

В системе координат $Oxyz$ углы α и β для каждого притира определены конструкцией станка (табл. 1), которые устанавливают положение оси вращения притира, а угол γ — разворот притира вокруг его собственной оси. Подставляя значения углов в уравнение (1), полу-

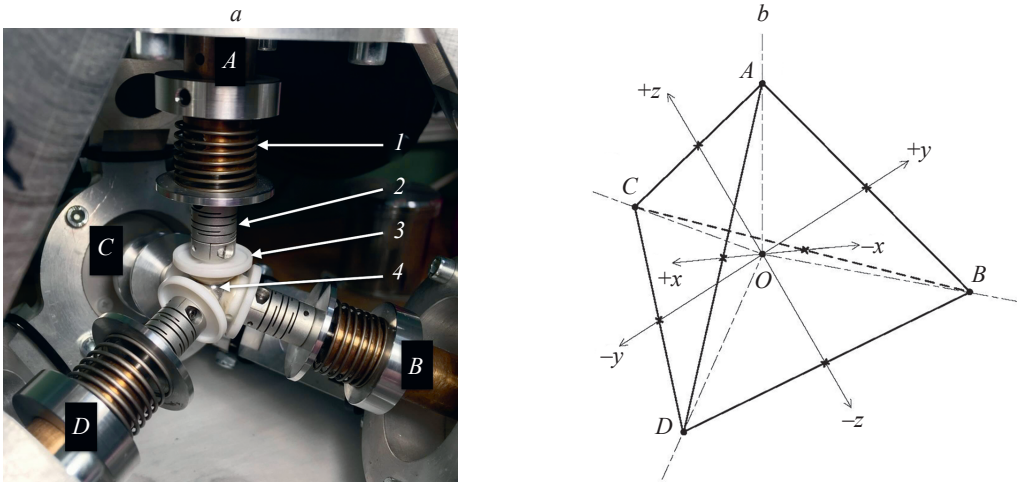

Рис. 1. Доводочный станок (*a*) и схема расположения осей притиров (*b*).

A, *B*, *C*, *D* — вершины тетраэдра с четырьмя серводвигателями; *1* — пружина; *2* — муфта; *3* — притиры; *4* — бериллиевый ротор

Fig. 1. Finishing machine (*a*) and lapping axes arrangement diagram (*b*).

A, *B*, *C*, *D* — vertices of a tetrahedron with four servomotors; *1* — spring; *2* — clutch; *3* — lapping; *4* — beryllium rotor

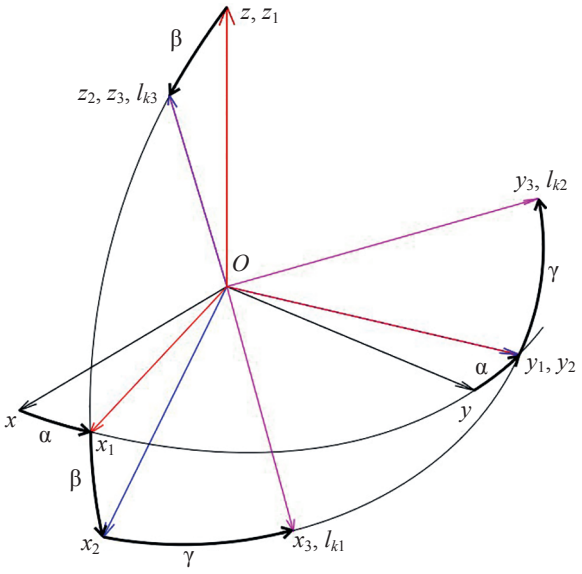

Рис. 2. Последовательность поворотов z, y_1, z_2

Fig. 2. Sequence of rotations z, y_1, z_2

чим матрицу перехода из системы координат притира $Ol_{k1}l_{k2}l_{k3}$ в систему координат станка $Oxyz$.

Рассмотрим вектор $\mathbf{v}$ в неподвижных координатах $Oxyz$. Если $\mathbf{v}'$ является одним и тем же вектором в подвижных координатах $Ol_{k1}l_{k2}l_{k3}$ то переход осуществляется как:

$$\mathbf{v}' = \mathbf{R}(\alpha, \beta, \gamma)\mathbf{v}. \tag{2}$$

Обратный переход выполняется следующим образом:

$$\mathbf{v} = \mathbf{R}(\alpha, \beta, \gamma)^T\mathbf{v}'. \tag{3}$$

Выражения (2) и (3) определяют связь систем координат притиров $Ol_{k1}l_{k2}l_{k3}$ с системой координат станка $Oxyz$. Заметим, что кинематическая связь в станке подразумевает лишь изменение угла γ при вращении притиров, а углы α, β определяют расположение осей притиров для конкретного станка. При этом положительное приращение угла γ в (2) определяет вращение притира по часовой стрелке при наблюдении из точки O в неподвижных координатах.

Перейдем к рассмотрению кинематической связи станка с ротором. Введем подвижную декартову систему координат $O\xi\eta\zeta$, связанную с ротором, начало которой совпадает с его центром. Поскольку центр ротора расположен в точке пересечения осей притиров, то начало подвижной и неподвижной систем координат также совпадают. Ось ζ направим по оси с наибольшим моментом инерции ротора, а оси ξ, η произвольным образом, поскольку ротор симметричен относительно оси с наибольшим моментом инерции. Определим связь систем координат $Oxyz$ и $O\xi\eta\zeta$ используя параметры Гамильтона и матрицу поворота ось-угол [11]. По условию, ротор обладает тремя вращательными степенями свободы и приводится в движение силами трения притиров, вращаясь при этом с мгновенной угловой скоростью ω. Для задания оси поворота используем единичный направляющий вектор $\mathbf{n}$ угловой скорости ротора, который определим в виде:

$$\mathbf{n} = \frac{\boldsymbol{\omega}}{|\boldsymbol{\omega}|}.$$

Угол, на который поворачивается ротор за малое время Δt вращаясь с мгновенной угловой скоростью ω, имеет вид:

$$\lambda = \int_t^{t+\Delta t} \omega(t)dt. \tag{4}$$

Кватернион $\mathbf{q}$, при вращении ротора вокруг $\mathbf{n}$, рассчитаем по формуле

$$\mathbf{q}(\mathbf{n}, \lambda) = \begin{bmatrix} \cos(\lambda/2) \\ \mathbf{n}\sin(\lambda/2) \end{bmatrix} = \begin{bmatrix} \cos(\lambda/2) \\ n_x \sin(\lambda/2) \\ n_y \sin(\lambda/2) \\ n_z \sin(\lambda/2) \end{bmatrix}. \tag{5}$$

Соответствующую матрицу поворота ось-угол запишем так

$$\mathbf{R}_q(\mathbf{q}) = \tag{6}$$
$$= \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix},$$

где q_0, q_1, q_2, q_3 — соответствующие элементы кватерниона (6).

Вычисляя угол поворота по уравнению (4) и подставляя (5) в (6), получим матрицу преобразования координат. Рассмотрим вектор $\mathbf{v}$ в неподвижных координатах $Oxyz$. Если $\mathbf{v}'$ является одним и тем же вектором в подвижных координатах $O\xi\eta\zeta$, то переход между координатами при повороте ротора на угол λ выражается как:

$$\mathbf{v}' = \mathbf{R}_q(\mathbf{q}(\mathbf{n}, \lambda))\mathbf{v}; \tag{7}$$

$$\mathbf{v} = \mathbf{R}_q(\mathbf{q}(\mathbf{n}, \lambda))^T\mathbf{v}'. \tag{8}$$

Выражения (7) и (8) определяют связь подвижной и неподвижной систем координат. Заметим, что выражения (4)–(8) применимы для численного решения уравнений движения ротора, в котором поворот ротора на малый угол λ осуществляется для каждого интервала времени Δt.

После того, как кинематические связи станка определены, перейдем к определению сил трения — дви-

Таблица 1. Углы α и β расположения осей притиров в системе координат $Oxyz$

Table 1. Angles α and β location of the lapping axes in the coordinate system $Oxyz$

Притир	α	β
A	45°	54°42′0″
B	135°	125°18′0″
C	225°	54°42′0″
D	315°	124°18′0″

жущих сил в рассматриваемой системе. Рассмотрим моменты сил трения, которые действуют в области контакта между притиром и ротором. На рис. 3 изображено давление притира на ротор с силой F в системе координат $Ol_1l_2l_3$ и связанной с ней сферической системой r, θ, φ, в которой r — радиус-вектор до рассматриваемой точки, θ — зенитный угол (угол между осью l_3 и радиус-вектором), φ — азимутальный угол (угол между осью l_1 и проекцией радиуса-вектора на плоскость Ol_1l_2).

Разделим поверхность ротора на малые элементы с площадью dS. Исходя из обозначений (рис. 3) элемент площади поверхности ротора с радиусом r может быть записан в виде:

$$dS = r^2 \sin(\theta)\, d\theta d\varphi.$$

В области контакта на каждый элемент площади приходится давление p со стороны притира. Сила трения f, при относительном перемещении притира и ротора, на каждой площадке dS может быть вычислена как:

$$f(\theta) = \mu p(\theta) dS, \qquad (9)$$

где μ — коэффициент трения скольжения; $p(\theta)$ — давление, приходящееся на элемент площади ротора.

Рассмотрим силу трения, в первом приближении, независимо от скорости относительного движения [12]. В действительности трение является многофакторным процессом и, безусловно, коэффициент трения нелинейно зависит от скорости [13]. Распределение давления $p(\theta)$ в контактной паре рассмотрим по аналогии выражения воздействий в шаровом шарнире — по синусоидальному закону [14]. Поверхности при этом

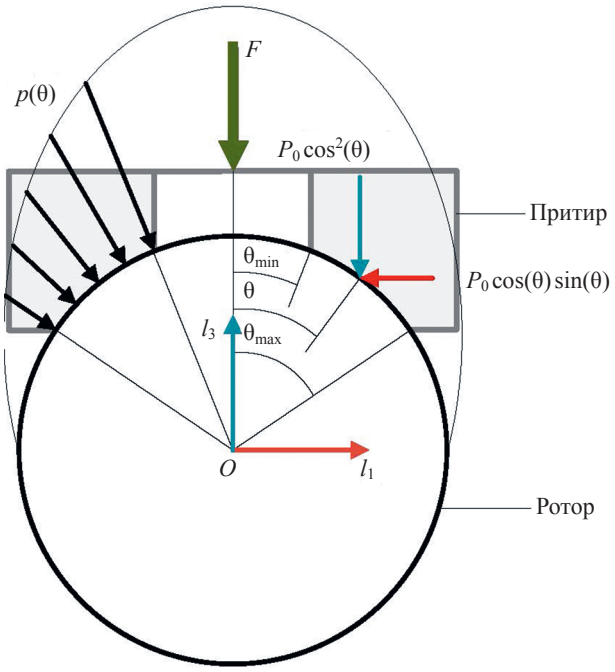

Рис. 3. Распределение давления притира на поверхности ротора

Fig. 3. Distribution of lapping pressure on the rotor surface

предполагаются идеальными, с положительным радиальным зазором во внутренней сфере притира, поскольку в паре имеется прослойка абразивной суспензии. В соответствии с синусоидальным распределением давления имеем:

$$p(\theta) = P_0 \cos(\theta), \qquad (10)$$

где P_0 — предельно возможное давление.

По условию, приложенная к притиру сила F уравновешена, а значит, равна интегральной сумме сил реакции, действующих вдоль оси приложения силы. Разделим давление на две составляющие — продольную $P_0 \cos^2\theta$ и поперечную $P_0 \cos\theta \sin\theta$ оси притира. Условие равновесия сил для области контакта может быть записано как:

$$F = \int_0^{2\pi} \int_{\theta_{min}}^{\theta_{max}} P_0 \cos^2(\theta) r^2 \sin(\theta) d\theta d\varphi. \qquad (11)$$

Интегрируя (11), получим:

$$F = 2\pi r^2 P_0 \left(\frac{\cos^3(\theta_{min}) - \cos^3(\theta_{max})}{3} \right). \qquad (12)$$

Выражая P_0 из уравнения (12) и подставляя в (10), и затем в (9), получим выражение для силы трения, действующей на каждую элементарную площадку dS:

$$f(\theta) = \frac{3\mu F \cos(\theta)}{2\pi r^2} \frac{dS}{\cos^3(\theta_{min}) - \cos^3(\theta_{max})}. \qquad (13)$$

Уравнение (13) определяет величину силы трения, при этом ее направление соответствует направлению скорости относительного движения $\boldsymbol{\omega}_{rel}$ поверхностей притира и ротора. Поверхность контакта $\mathbf{r}$, в соответствии с принятыми допущениями, можно представить как сегмент сферы. В соответствии с обозначениями (рис. 3) в векторном представлении имеем:

$$\mathbf{r} = \mathbf{r}(\theta, \varphi) = r \sin(\theta) \cos(\varphi)\, \mathbf{i} + r \sin(\theta) \sin(\varphi)\, \mathbf{j} + r \cos(\theta)\, \mathbf{k},$$
$$\varphi \in [0, 2\pi], \quad \theta \in [\theta_{min}, \theta_{max}],$$

где $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ — базисные векторы системы координат $Ol_1l_2l_3$, связанной с притиром.

Линейные скорости относительного движения на каждом элементе поверхности определяются векторным произведением относительной угловой скорости на радиус-вектор:

$$\mathbf{v} = \boldsymbol{\omega}_{rel} \times \mathbf{r}.$$

Поставив в соответствие каждому элементу dS вектор относительной скорости, получим векторное поле скоростей относительного движения поверхностей. Относительную скорость движения поверхностей можно приближенно считать скоростью движения абразива, которым обрабатывается ротор. Векторы относительной скорости, коллинеарные силе трения, действующей на ротор, могут быть использованы как направляющие векторы для сил трения:

$$\mathbf{f} = \frac{\mathbf{v}}{|\mathbf{v}|} f(\theta).$$

Момент, создаваемый силами **f**, вычисляется относительно неподвижной точки O по формуле

$$\mathbf{M}_k = \iint_S \left(\mathbf{r} \times \frac{\mathbf{f}}{dS} \right) dS. \quad (14)$$

Интегрируя векторное произведение (14) по поверхности контакта, получим момент силы, действующий со стороны притира с индексом k в системе координат $Ol_{k1}l_{k2}l_{k3}$. Определим суммарный крутящий момент действующий на ротор. Для этого осуществим переход из $Ol_{k1}l_{k2}l_{k3}$ в $Oxyz$ и просуммируем векторы крутящего момента каждого из четырех притиров:

$$\mathbf{M} = \sum_{k=1}^{4} \mathbf{R}(\alpha_k, \beta_k, \gamma_k)^T \mathbf{M}_k. \quad (15)$$

Запишем систему динамических уравнений Эйлера применительно к ротору, на который действует внешний крутящий момент $\mathbf{M}'$ в системе координат, связанной с ротором $O\xi\eta\zeta$:

$$\begin{cases} I_\xi \dot{\omega}_\xi + (I_\zeta - I_\eta)\omega_\eta \omega_\zeta = \mathbf{M}'_\xi \\ I_\eta \dot{\omega}_\eta + (I_\xi - I_\zeta)\omega_\xi \omega_\zeta = \mathbf{M}'_\eta \\ I_\zeta \dot{\omega}_\zeta + (I_\eta - I_\xi)\omega_\xi \omega_\eta = \mathbf{M}'_\zeta \end{cases}, \quad (16)$$

где I_ξ, I_η, I_ζ — осевые моменты инерции ротора.

Преобразование координат для вектора внешнего крутящего момента определим как:

$$\mathbf{M}' = \mathbf{R}_q \left(\mathbf{q} \left(\frac{\boldsymbol{\omega}_i}{|\boldsymbol{\omega}_i|}, \lambda \right) \right) \mathbf{M}. \quad (17)$$

Угол поворота ротора λ найдем из выражения (4) на каждом шаге Δt. Таким образом, уравнения (16) и (17) полностью определяют движение ротора при обработке чашечными притирами.

Раскрытие выражений (15)–(17) дает громоздкие системы уравнений, решение которых в аналитическом виде затруднено. По этой причине решение уравнений движения выполним численным методом простой подстановки Эйлера.

Результаты моделирования

Рассмотрим результаты решений предложенной модели на примере полировки полого ротора из бериллия с наружным диаметром 50 мм притирами из фторопласта. Параметры модели выбирались исходя из режимов доводки и полировки роторов, применяемых на практике. Геометрические размеры ротора соответствуют действительным размерам ротора ЭСГ, представленным в работах [15, 16]. Ось подвижной системы координат ζ совпадает с осью симметрии ротора. Осевые моменты инерции были вычислены в программе SolidWorks.

Осуществим проверку модели движения в частном случае, для которого положение оси вращения ротора является тривиальным. Выберем режим вращения притиров таким, в котором притиры A, D вращаются по часовой стрелке, а притиры B, C в противоположном направлении. При этом притиры вращаются с одинаковыми угловыми скоростями. В силу симметрии станка, в системе $Oxyz$ (рис. 1) результирующий момент сил трения действует по оси x. Соответственно, ротор должен раскручиваться относительно оси x. Зададим начальные условия, представление в табл. 2.

Рассмотрим зависимость момента сил трения от времени в координатах ротора $O\xi\eta\zeta$, (рис. 4).

В течение первых 10 с система не переходит в состояние динамического равновесия, поскольку приведенный момент не равен нулю. Вместо этого наблюдается затухающий переходный процесс, вызванный неравенством главных моментов инерции ротора. Составляющие вектора угловой скорости ротора, представлены на рис. 5. Из рис. 5, *b* видно, что ротор, раскручиваясь вокруг оси близкой к ξ, с течением времени стремится занять положение, в котором ось вращения будет совпадать с осью наибольшего момента инерции I_ζ. Длительность переходного процесса составляет 12 с,

Таблица 2. Параметры и начальные условия модели движения ротора

Table 2. Parameters and initial conditions of the rotor motion model

Параметр	Значение
I_ξ, кг·м^2	$6{,}11 \times 10^{-6}$
I_η, кг·м^2	$6{,}11 \times 10^{-6}$
I_ζ, кг·м^2	$7{,}06 \times 10^{-6}$
θ_{min}, °	45
θ_{max}, °	54
μ	0,05
F, Н	3
h, с	1×10^{-3}
ω_A, ω_D, рад/с	31,4
ω_B, ω_C, рад/с	–31,4
α, °	1

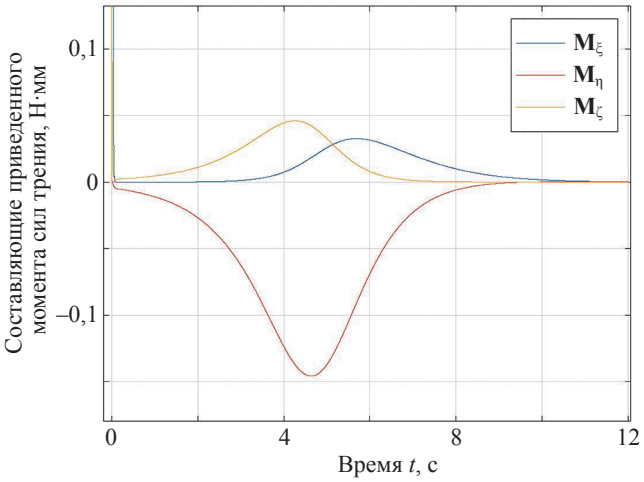

Рис. 4. Составляющие приведенного момента сил трения в системе $O\xi\eta\zeta$

Fig. 4. Components of the reduced moment of friction forces in the $O\xi\eta\zeta$ system

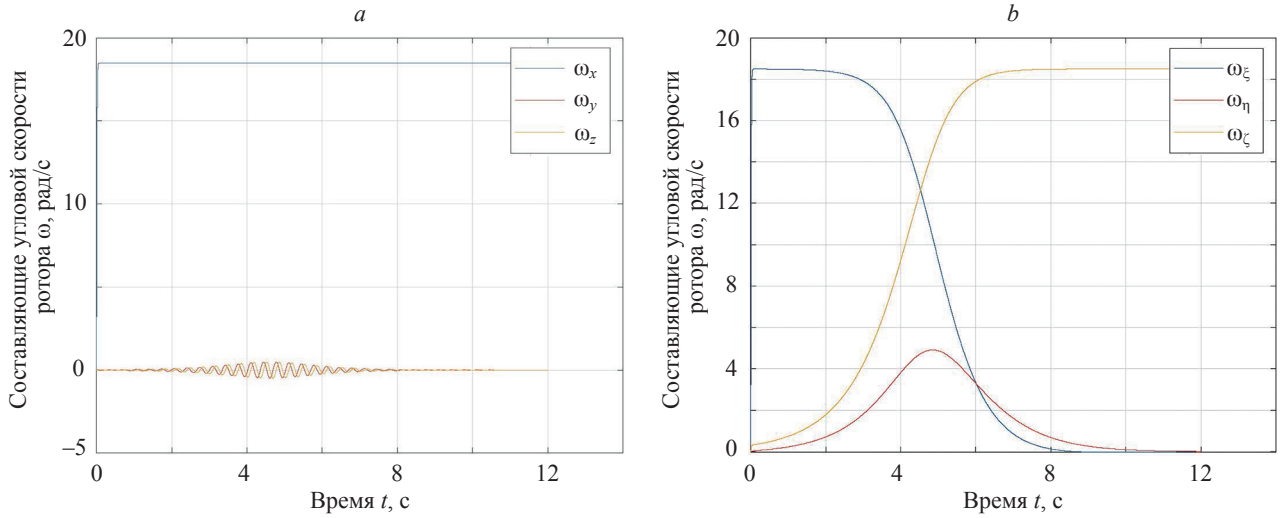

Рис. 5. Проекции угловой скорости ротора в системах: *Oxyz* (*a*) и *O*ξηζ (*b*)
Fig. 5. Projections of the angular velocity of the rotor in the systems: *Oxyz* (*a*) and *O*ξηζ (*b*)

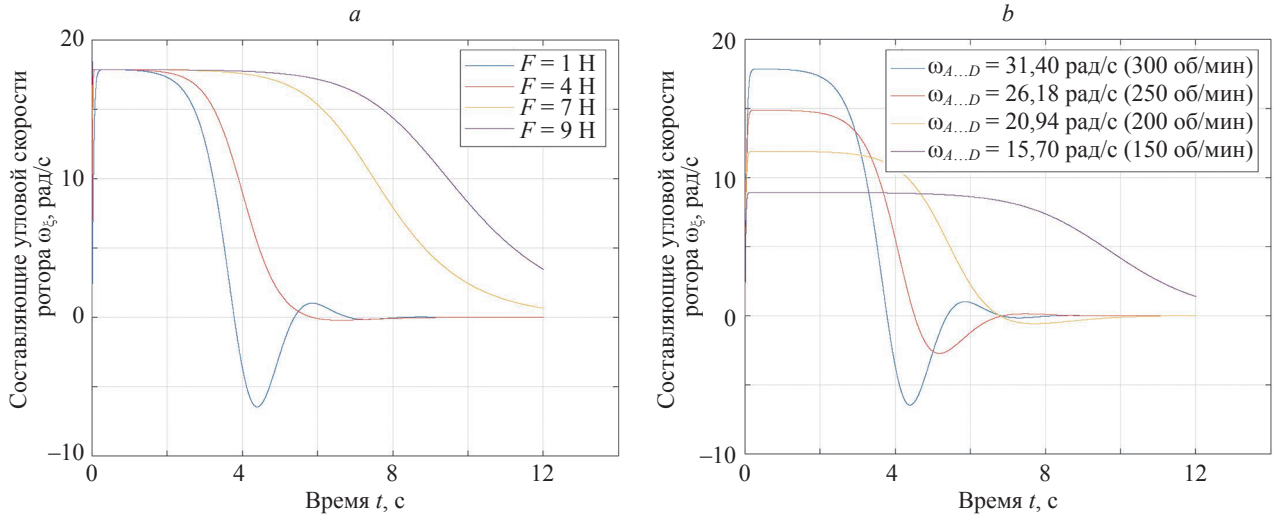

Рис. 6. Переходный процесс для различных значений: $F(\omega_{A\ldots D} = 31{,}4, \mu = 0{,}05)$ (*a*); $\omega_{A\ldots D}$ $(F = 1, \mu = 0{,}05)$ (*b*)
Fig. 6. Transient process for different values of: $F(\omega_{A\ldots D} = 31{,}4, \mu = 0{,}05)$ (*a*); $\omega_{A\ldots D}$ $(F = 1, \mu = 0{,}05)$ (*b*)

по истечении которых ротор вращается вокруг оси ζ с постоянной угловой скоростью, и система переходит в состояние динамического равновесия.

Исследование влияния начальных условий показывает, что длительность переходного процесса зависит, в основном, от четырех параметров: силы прижатия притиров F, коэффициента трения μ, скорости вращения притиров $\omega_{A\ldots D}$ и в меньшей степени от площади рабочей поверхности притиров, которая определяется углами θ_{min}, θ_{max}. Увеличение F или μ приводит к повышению длительности переходного процесса и степени затухания (рис. 6), на котором представлено изменение только одной составляющей ω_ξ для разных значений приложенной силы. Как видно из (рис. 6, *a*) увеличение силы прижатия приводит к увеличению длительности переходного процесса.

В наибольшей степени влияние разности моментов инерции ротора на его движение в станке зависит от угловой скорости чаш притиров и, соответственно, самого ротора. Пример такой зависимости представлен на рис. 6, *b*.

Заключение

Представленные зависимости переходного процесса от параметров системы позволяют определить степень влияния отличия главных моментов инерции на результирующее движение ротора. Так, для процесса полирования, в котором силы F находятся в пределах 1–9 Н и притиры вращаются с относительно высокими скоростями — в пределах 15,7–31,4 рад/с результаты моделирования показали, что для полого ротора электростатического гироскопа, отличие в главных моментах инерции оказывает существенное влияние на параметры движения ротора во время обработки — проявляется гироскопический эффект. Это приводит к тому, что в цикле обработки, при любом переключении направлений вращения притиров, в координатах

ротора ось вращения стремится выровняться с осью максимального момента инерции, что приводит к неравномерности обработки.

Для процессов доводки, в которых сила F более 9 Н или коэффициент трения более 0,3 и угловые скорости притиров достаточно низкие — до 15,7 рад/с, различие моментов инерции в первом приближении можно не учитывать, поскольку длительность переходного процесса составляет десятки секунд и более.

На основе рассмотренных зависимостей получена следующая практическая рекомендация. При полировании ротора электростатического гироскопа следует использовать скорости вращения притиров — 15 рад/с (143 об/мин) и менее, а также применять интервал между переключением направления вращения притиров не более 6 с. Указанные граничные условия позволяют избежать появления повторяющихся паттернов выработки на поверхности ротора и повысить сферичность.

Разработанная математическая модель движения показала примечательный факт, что ни сила прижатия, ни коэффициент трения скольжения не влияют на угловую скорость ротора при наступлении динамического равновесия системы в диапазоне используемых на практике режимов обработки.

Важнейшим практическим применением представленной модели является возможность определения параметров движения ротора в сферодоводочном устройстве для конкретных рабочих условий, на основе которых появляется возможность управлять положением ротора и реализовывать систему управления станком доводки.

Литература

1. Ландау Б.Е., Белаш А.А., Гуревич С.С., Левин С.Л., Романенко С.Г., Цветков В.Н. Электростатический гироскоп в системах ориентации космических аппаратов // Гироскопия и навигация. 2021. Т. 29. № 3 (114). С. 69–79. https://doi.org/10.17285/0869-7035.0071
2. Мартыненко Ю.Г. Движение твердого тела в электрических и магнитных полях. М.: Наука, 1988. 368 с.
3. Федорович С.Н. Современное состояние и перспективы развития технологии сферодоводки прецизионных узлов // Металлообработка. 2018. № 1 (103). С. 27–32.
4. Angele W. Finishing high precision quartz balls // Precision Engineering. 1980. V. 2. N 3. P. 119–122. https://doi.org/10.1016/0141-6359(80)90025-2
5. Marcelja F., DeBra D.B., Keiser G.M., Turneaure J.P. Precision spheres for the Gravity Probe B experiment // Classical and Quantum Gravity. 2015. V. 32. N 22. P. 224007. https://doi.org/10.1088/0264-9381/32/22/224007
6. Becker P., Schiel D. The Avogadro constant and a new definition of the kilogram // International Journal of Mass Spectrometry. 2013. V. 349-350. P. 219–226. https://doi.org/10.1016/j.ijms.2013.03.015
7. Федорович С.Н. Моделирование процесса доводки сферического ротора шарового гироскопа // Известия высших учебных заведений. Приборостроение. 2021. Т. 64. № 4. С. 307–315. https://doi.org/10.17586/0021-3454-2021-64-4-307-315
8. Орлов П.Н. Технологическое обеспечение качества деталей методами доводки. М. Машиностроение, 1988. 383 с.
9. Бабаев С.Г., Садыгов П.Г. Притирка и доводка поверхностей деталей машин. М.: Машиностроение, 1976. С. 6–15.
10. Маркеев А.П. Теоретическая механика. М.: Наука, 1990. 414 с.
11. Амелькин Н.И. Кинематика и динамика твердого тела. М.: МФТИ, 2000. 63 с.
12. Фаркаш З., Бартельс Г., Вольф Д.Э., Унгер Т. О силе трения при поступательном и вращательном движении плоского тела // Нелинейная динамика. 2011. Т. 7, № 1. С. 139–146. https://doi.org/10.20537/nd1101007
13. Khala M.J., Hare C., Wu C., Martin M.J., Venugopal N., Freeman T. The importance of a velocity-dependent friction coefficient in representing the flow behaviour of a blade-driven powder bed // Powder Technology. 2021. V. 385. P. 264–272. https://doi.org/10.1016/j.powtec.2021.02.060
14. Aublin M. Systèmes Mécaniques: Théorie et Dimensionnement. Dunod, 1993. 662 p.
15. Анфиногенов А.С., Парфенов О.И. Способ уменьшения деформаций внешней поверхности тонкостенных сферических роторов гироскопов // Морское приборостроение. 1969. № 1. С. 114–119.
16. Юльметова О.С. Ионно-плазменные и лазерные технологии в гироскопическом приборостроении: диссертация на соискание ученой степени доктора технических наук. СПб., 2019. 220 с.

References

1. Landau B.E., Belash A.A., Gurevich S.S., Levin S.L., Romanenko S.G., Tsvetkov V.N. Electrostatic gyroscope in spacecraft attitude reference systems. *Gyroscopy and Navigation*, 2021, vol. 12, no. 3, pp. 247–253. https://doi.org/10.1134/s2075108721030056
2. Martynenko Y.G. *Motion of a Rigid Body in Electric and Magnetic Fields*. Moscow, Nauka Publ., 1988, 368 p. (in Russian)
3. Fedorovich S.N. Current state and perspectives for development of the technology of lapping of precision spherical system elements. *Metalworking*, 2018, no. 1 (103), pp. 27–32. (in Russian)
4. Angele W. Finishing high precision quartz balls. *Precision Engineering*, 1980, vol. 2, no. 3, pp. 119–122. https://doi.org/10.1016/0141-6359(80)90025-2
5. Marcelja F., DeBra D.B., Keiser G.M., Turneaure J.P. Precision spheres for the Gravity Probe B experiment. *Classical and Quantum Gravity*, 2015, vol. 32, no. 22, pp. 224007. https://doi.org/10.1088/0264-9381/32/22/224007
6. Becker P., Schiel D. The Avogadro constant and a new definition of the kilogram. *International Journal of Mass Spectrometry*, 2013, vol. 349-350, pp. 219–226. https://doi.org/10.1016/j.ijms.2013.03.015
7. Fedorovich S. N. Modeling the process of finishing the spherical rotor of a ball gyroscope. *Journal of Instrument Engineering*, 2021, vol. 64, no. 4, pp. 307–315. (in Russian). https://doi.org/10.17586/0021-3454-2021-64-4-307-315
8. Orlov P.N. *Technological Quality Assurance of Parts by Finishing Methods*. Moscow, Mashinostroenie Publ., 1988, 383 p. (in Russian)
9. Babaev S.G., Sadygov P.G. *Lapping and Finishing of Machine Parts Surfaces*. Moscow, Mashinostroenie Publ., 1976, pp. 6–15. (in Russian)
10. Markeev A.P. *Theoretical Mechanics*. Moscow, Nauka Publ., 1990. 414 p. (in Russian)
11. Amelkin N.I. *Kinematics and Dynamics of a Rigid Body*. Moscow, MIPT Publ., 2000, 63 p. (in Russian)
12. Farkas Z., Bartels G., Unger T., Wolf D. Frictional coupling between sliding and spinning motion. *Russian Journal of Nonlinear Dynamics*, 2011, vol. 7, no. 1, pp. 139–146. (in Russian). https://doi.org/10.20537/nd1101007
13. Khala M.J., Hare C., Wu C., Martin M.J., Venugopal N., Freeman T. The importance of a velocity-dependent friction coefficient in representing the flow behaviour of a blade-driven powder bed. *Powder Technology*, 2021, vol. 385, pp. 264–272. https://doi.org/10.1016/j.powtec.2021.02.060
14. Aublin M. *Systèmes Mécaniques: Théorie et Dimensionnement. Dunod*, 1993, 662 p. (in French)
15. Anfinogenov A.S., Parfenov O.I., Method to reduce the deformations of the outer surface of thin-walled spherical rotors in gyroscopes. *Morskoe priborostroenie*, 1969, no. 1, pp. 114–119. (in Russian)
16. Yulmetova, O.S. *Ion-plasma and laser technologies in gyroscopic instrumentation*. Dissertation for the degree of doctor of technical sciences. St. Petersburg, 2019, 220 p. (in Russian)

Автор

Федорович Сергей Николаевич — начальник лаборатории, АО «Концерн «ЦНИИ «Электроприбор», Санкт-Петербург, 197046, Российская Федерация, https://orcid.org/0009-0001-3147-9910, fedorovichsn@gmail.com

Author

Sergei N. Fedorovich — Chief of Laboratory, JSC Concern CSRI Elektropribor, Saint Petersburg, 197046, Russian Federation, https://orcid.org/0009-0001-3147-9910, fedorovichsn@gmail.com

Статья поступила в редакцию 12.05.2025
Одобрена после рецензирования 16.07.2025
Принята к печати 22.09.2025

Received 12.05.2025
Approved after reviewing 16.07.2025
Accepted 22.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-952-960

Experimental study of the optically transparent gas flow and temperature field using the background oriented Schlieren method

Pavel A. Bryzgunov[1]✉, Dmitry S. Pisarev[2], Olga V. Zlyvko[3], Andrey N. Rogalev[4], Nikolay D. Rogalev[5]

[1,2,3,4,5] National Research University "Moscow Power Engineering Institute", Moscow, 111250, Russian Federation

[1] bryzgunovpa@mpei.ru✉, https://orcid.org/0000-0003-3710-5116
[2] pisarevds@mpei.ru, https://orcid.org/0009-0006-3091-4884
[3] zlyvkoov@mpei.ru, https://orcid.org/0000-0003-0554-4026
[4] rogalevan@mpei.ru, https://orcid.org/0000-0001-7256-0144
[5] rogalevnd@mpei.ru, https://orcid.org/0000-0002-6458-2869

Abstract

The article presents the results of an experimental study of the flow structure and temperature field in a plume formed above a low-power burner flame. The pulsation and spectral characteristics of the flow at key sampling points were analyzed, which allowed us to draw a conclusion about the nature of the flow at the main points of the jet. It is proposed to use time series of changes in the point displacement field to analyze the spectral characteristics of the flow. In this work, the Background Oriented Schlieren method was used to visualize the flow and determine temperatures followed by post-processing in the program developed during the study. The advantage of this approach compared to the traditional optical Schlieren method is that there is no need for parabolic mirrors as well as the ability to obtain results in digital form convenient for further processing. During the experiment, a special background with randomly located bright dots was placed behind the object of study which was filmed by a video camera. Fluctuations in the medium density caused changes in the refractive indices of the medium, as a result of which the points on the background of the video frames displaced, and the displacements of the points was proportional to the change in the refractive index which in turn is proportional to the density gradient and, accordingly, to the temperature gradient of the medium. The displacement of the points was determined by cross-correlation analysis of each frame in comparison with the frames in the absence of disturbances. Then the displacement field was filtered by a median filter in order to minimize noise and statistical outliers. The filtered displacement field was used to calculate the temperature field, while solving the Cauchy problem for temperature with a known derivative at a point and specified boundary conditions. A set of instantaneous point displacement fields, instantaneous and time-averaged temperature fields was obtained, which allowed us to draw conclusions about the flow structure. At characteristic points of the jet, oscillograms of the displacement value were obtained as well as pulsation spectra with an inertial interval corresponding to the "–5/3" law. The approach proposed in the work allows, in addition to contactless study of the temperature field, also studying turbulent flow pulsations in the case of close to two-dimensional or axisymmetric flows.

Keywords

background oriented Schlieren method, temperature field, spectral characteristics of flow, optical studies of flow, flow structure

Acknowledgements

This study conducted by the Moscow Power Engineering Institute was financially supported by the Ministry of Science and Higher Education of the Russian Federation (State Assignment No. FSWF-2023-0014, contract No. 075-03-2023-383, 18.01.2023).

For citation: Bryzgunov P.A., Pisarev D.S., Zlyvko O.V., Rogalev A.N., Rogalev N.D. Experimental study of the optically transparent gas flow and temperature field using the background oriented Schlieren method. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 952–960. doi: 10.17586/2226-1494-2025-25-5-952-960

УДК 532.5-1/-9

Экспериментальное исследование структуры течения и поля температур оптически прозрачной среды посредством фоново-ориентированного шлирен-метода

Павел Александрович Брызгунов[1]✉, Дмитрий Сергеевич Писарев[2], Ольга Владимировна Злывко[3], Андрей Николаевич Рогалев[4], Николай Дмитриевич Рогалев[5]

[1,2,3,4,5] Национальный исследовательский университет «МЭИ», Москва, 111250, Российская Федерация

[1] bryzgunovpa@mpei.ru✉, https://orcid.org/0000-0003-3710-5116
[2] pisarevds@mpei.ru, https://orcid.org/0009-0006-3091-4884
[3] zlyvkoov@mpei.ru, https://orcid.org/0000-0003-0554-4026
[4] rogalevan@mpei.ru, https://orcid.org/0000-0001-7256-0144
[5] rogalevnd@mpei.ru, https://orcid.org/0000-0002-6458-2869

Аннотация

Введение. Представлены результаты экспериментального исследования структуры течения и поля температур в конвективной струе воздуха и продуктов сгорания природного газа, формирующейся над пламенем горелки малой мощности. Проанализированы пульсационные и спектральные характеристики потока в ключевых точках отбора, что позволило сделать вывод о характере течения в основных точках струи. Предложено для анализа спектральных характеристик потока использовать временные ряды изменения поля смещений точек. **Метод.** В работе для визуализации течения и определения температур использован фоново-ориентированный шлирен-метод с последующей постобработкой в разработанной в ходе исследования программе. Преимуществом данного подхода в сравнении с традиционным оптическим шлирен-методом является отсутствие необходимости в параболических зеркалах, а также возможность получения результатов в цифровом виде, удобном для дальнейшей обработки. В ходе эксперимента за объектом исследования, который снимался видеокамерой, помещался фон со случайно расположенными черными точками. Колебания плотности среды вызывали изменения коэффициентов преломления среды, вследствие чего точки на фоне на видеокадрах смещались, причем смещение точек пропорционально изменению коэффициента преломления, который в свою очередь пропорционален градиенту плотности и, соответственно, градиенту температуры среды. Смещение точек определялось с применением кросс-корреляционного анализа каждого кадра в сравнении с кадрами при отсутствии возмущений. Далее поле смещений подвергалось фильтрации посредством медианного фильтра с целью минимизации шумов и статистических выбросов. Отфильтрованное поле смещений использовалось для вычисления поля температур, при этом решалась задача Коши относительно температуры с известной производной в точке и заданных граничных условиях. **Основные результаты.** Получена совокупность мгновенных полей смещений точек, мгновенных и осредненного полей температуры, позволивших сделать выводы о структуре течения. В характерных точках струи получены осциллограммы величины смещения, а также спектры пульсаций, имеющие инерционный интервал, соответствующий закону «–5/3». **Обсуждение.** Предложенный в работе подход позволяет в дополнение к бесконтактному исследованию поля температур также исследовать турбулентные пульсации течения в случае квазидвухмерных или осесимметричных потоков.

Ключевые слова

фоново-ориентированный шлирен-метод, поле температур, спектральные характеристики потока, оптические исследования потока, структура течения

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания № FSWF-2023-0014 (Соглашение № 075-03-2023-383 от 18 января 2023 г.) в сфере научной деятельности на 2023–2025 гг.

Ссылка для цитирования: Брызгунов П.А., Писарев Д.С., Злывко О.В., Рогалев А.Н., Рогалев Н.Д. Экспериментальное исследование структуры течения и поля температур оптически прозрачной среды посредством фоново-ориентированного шлирен-метода // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 952–960. doi: 10.17586/2226-1494-2025-25-5-952-960

Introduction

The Schlieren technique is a non-invasive method for measuring the density gradient of an optically transparent fluid. The principle of this technique is based on the fact that changes in the density of the fluid or gas also result in changes in the refractive index of the medium. There is a direct correlation between the density gradient and the change in refractive index gradient, which implies that in regions where there is a density gradient, such as those caused by temperature or pressure changes, light rays may deviate from their initial trajectory and form shadow patterns.

Currently, there are two primary classes of experimental approaches based on this phenomenon: traditional optical Schlieren and Background Oriented Schlieren (BOS) techniques.

The traditional Schlieren method [1] utilizes optical systems composed of a point light source, a subject under examination, a focusing element (either a lens or a parabolic mirror), and a “light knife” (a thin, opaque plate). As light passes through areas with varying optical properties, rays from the point source diverge from their original path, while the remaining rays maintain their course. Next, the light beam either travels through a lens or reflects off a parabolic mirror at a focal point where a

light barrier is positioned. This barrier serves to block the majority of the light (known as the illumination), while rays that deviate from the primary trajectory bypass it and strike a screen, resulting in a shadow pattern that reveals the density distribution gradient.

A notable advantage of the conventional Schlieren technique is the exceptional quality of the generated images. The primary drawbacks of the traditional Schlieren technique are: the requirement for the installation of expensive (in the case of larger diameters) optical equipment, and the complexity associated with subsequent data analysis due to its analog nature.

The BOS method, first proposed by Dalziel in 2000 [2], involves the installation of a background behind the subject under study, with points applied in a regular or irregular pattern [3]. When light rays reflected from the background travel through it, similar to the traditional Schlieren technique, they deviate from their initial trajectory. A digital video camera captures the object on the background with dots. In this instance, the deviation of the light rays from their trajectory is manifested as a visual displacement of the background dots in regions of the gradient in the refractive index compared to their position in the absence of perturbations. By using cross-correlation analysis of the acquired video frames, a field of displacement points can be generated that is directly proportional to the gradient in refractive index and, consequently, the density gradient. The main disadvantage of this approach is that the quality of the images obtained is inferior to that of the traditional method, and there is also the presence of noise. Additionally, cross-correlation image analysis is a relatively resource-intensive process.

On the other hand, the benefits of BOS include ease of installation and the ability to obtain digital data after processing in two- or three-dimensional displacement fields (in the case of multiple cameras), which can be utilized to reconstruct the density gradient field [4].

Thus, if the density of the medium within the boundaries of the area being studied is known, it is then possible to reconstruct the density field using a gradient field in the presence of [5]. This, in turn, allows for the determination of temperature, pressure, and species concentration fields under known boundary conditions and an equation of state [6–8]. With further processing, it becomes possible to measure velocity fields [9, 10].

When imaging from multiple angles, it is possible to generate a three-dimensional mass distribution [11, 12]. The visualization of mass gradient fields, known as "numerical Schlieren", can be employed to validate the results of computational modeling when compared to experimental data acquired using the Schlieren technique and shadow method [13, 14].

A substantial portion of the work centered on the implementation of the BOS approach focuses on the investigation of various flames and the convective plumes that form above them. In [15], the BOS method was used to achieve instantaneous three-dimensional refractive index characterization of unstable natural gas flames from Bunsen burners using a 23-chamber setup.

In [16], the candle flame was investigated in a three-dimensional domain using a setup with 11 cameras. In order to reduce image noise, a background pattern printed on a transparent film illuminated by LEDs, and aspherical and Fresnel lenses were employed to ensure uniform lighting of the background. This resulted in high-quality images of the refractive index and temperature distributions at various time points. Similar findings regarding the visualization of density and temperature fields can be found in [17–19]. Additionally, [14–18] demonstrate the turbulent nature of the flow, which is similar to that of currents in free jets.

However, these studies have not analyzed the spectral characteristics of the currents, specifically the frequencies and amplitudes of pulsations that occur. Nevertheless, since the field of point displacements obtained during image processing when implementing the BOS is linearly related to the density gradient and, consequently, temperature, this field can be considered a passive scalar whose pulsations are associated with those of the velocity field.

As shown in [20], using the example of freely convective flows in a cubic cavity, temperature pulsations are directly linked to velocity pulsations and their spectral characteristics (in terms of fundamental frequencies) coincide qualitatively and quantitatively. Therefore, by employing the BOS technique, it is feasible to acquire not only variable fields but also a spectral analysis of the flows.

The significant advantage of this technique in comparison with analyzing time series of temperature data obtained through contact sensor measurements (the conventional approach) is that it eliminates thermal inertia from the primary transducer (such as a thermocouple), and it is a non-contact method of measurement, as the temperature sensor may cause additional flow disturbances.

The goal of this research is to assess the feasibility of the BOS technique for examining the spectral features of the flow, using a plume formed above a low-intensity burner flame as an example as well as to investigate the structure of this jet.

Research method

The BOS technique was employed to investigate the flow field. The setup is illustrated in Fig. 1. A convective jet of heated gas above a burner flame was captured as the subject of study using an Evercam 1000-16-M high speed camera with a backdrop of a disordered dot pattern. A low-power burner with a nozzle diameter D = 5 mm and gas consumption of 0.1 gram per hour was used. The distances from the background, Z_1, to the subject and from the subject, Z_2, to the camera were 150 and 200 mm, respectively, for a total distance, Z_3, of 350 mm, X_1 width was 90 mm, and height Y_1 was 160 mm. The Cartesian coordinate system was utilized. The resulting images were processed using a Python-based program. OpenPIV was used for cross-correlation analysis. Given a fixed distance between the camera and background and a constant focal length, derivatives of the refractive index were determined as follows:

$$\frac{\partial n}{\partial x} = -K\Delta x, \qquad (1)$$

where n is the refractive index at a point; Δx is the displacement along the x axis; K is the empirical coefficient, $K > 0$.

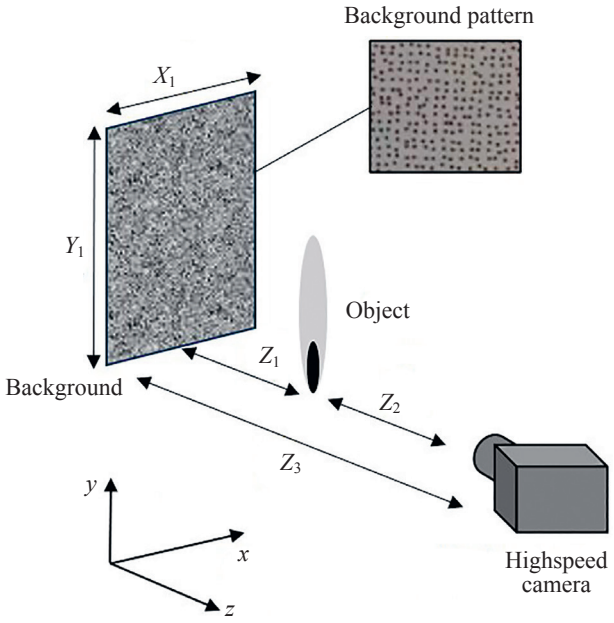

Fig. 1. Experimental setup diagram

The correlation between the refractive index and the density of the medium is linear according to the Gladstone-Dale equation:

$$n - 1 = k\rho, \tag{2}$$

where k is the Gladstone-Dale coefficient and ρ is the density of the medium.

The Gladstone-Dale coefficient for visible light and constant wavelength is dependent on the composition of the medium, but it is independent of pressure and temperature at range below 4000 K [21, 22]. Assuming combustion in air composed of 80 % nitrogen and 20 % oxygen, the volume composition of combustion products from natural gas can be estimated to be approximately 70 % nitrogen, 20 % water vapor, and 10 % carbon dioxide. Under these conditions, the Gladstone-Dale coefficient will vary between approximately $2.3 \cdot 10^{-4}$ m^3/kg for clean air and $2.5 \cdot 10^{-4}$ m^3/kg for combustion products, with a maximum possible change of less than 9 %. Given that combustion, occurs in an open environment, where the combustion products mix with surrounding air, the actual value of the Gladstone-Dale coefficient is likely to be closer to that of clean air. Therefore, a constant value of $2.3 \cdot 10^{-4}$ m^3/kg was chosen for use in this study. Considering equations (1) and (2), the derivative of the density can be calculated using the following equation:

$$\frac{\partial \rho}{\partial x} = \frac{1}{k}\frac{\partial n}{\partial x} = -\frac{1}{k}K\Delta x. \tag{3}$$

The density and temperature (T) of a gas are related through the equation of state:

$$T = \frac{P\mu}{R\rho}, \tag{4}$$

where P is the absolute pressure; R is the universal gas constant (R = 8.3 J/(mol·K)); μ is the molecular weight of the gas.

Since the tests took place in an open volume, the pressure can be assumed to be constant and equal to atmospheric pressure. The molecular weight of clean air is 29 g/mol, and the combustion products of natural gas, taking into account the composition described above, are 27.6 g/mol. The difference is less than 5 %, so μ was assumed to be constant and equal to 29 g/mol. Taking into account these and previously accepted assumptions, as well as equations (1)–(4), the expression for the temperature gradient can be written as follows:

$$\frac{\partial T}{\partial x} = -T^2\frac{R}{\mu P}\frac{\partial \rho}{\partial x} = T^2\frac{R}{\mu P}\frac{K}{k}\Delta x. \tag{5}$$

Thus, in accordance with equation (5), it is possible to determine the temperature field in the presence of boundary conditions by solving the Cauchy problem numerically in a two-dimensional computational domain.

In this study, the equation is solved with respect to the excess temperature; therefore, a zero value is assigned at the lateral boundaries. Due to the fact that the derivative is only taken with respect to x, there is an automatic boundary condition of the second kind, namely, $\partial T/\partial y = 0$, at the upper and lower limits. The empirical coefficient K, which is dependent on the optical properties of the setup, can be determined based on the measurement of flow temperature at a reference point using a thermocouple.

Fig. 2 presents examples of fields obtained during the processing. As illustrated in Fig. 2, *a*, the initial displacement field exhibits substantial noise, as well as artifacts originating from an optically opaque flame. This noise was mitigated through the application of a median filter, as depicted in Fig. 2, *b*. Based on the processed displacement field, the excess temperature field was calculated using equation (3).

It is crucial to note that owing to the inherent opacity of the flame, the temperature gradients within the flame region are assumed to be zero. Consequently, the temperatures displayed in the flame region in the treatment results approximate the closest values of the temperature in the transparent medium.

In order to investigate the oscillatory properties of the jet, we analyzed the displacement values at points *1–4* (Fig. 2, *b*). To eliminate noise during the analysis of fluctuations, we employed a Savitsky-Golay filter [23].

The spectral features of fluctuations with an amplitude Δx were determined using a fast Fourier transform. Given that the acquisition was performed at a constant frame rate, time-averaged fields were obtained by taking arithmetic means of the total instantaneous fields.

Results and Discussion

Fig. 3 illustrates the measurement results for the instantaneous temperature distributions (Fig. 3, *a–c*) and the average field (Fig. 3, *d*). As can be observed in Fig. 3, *a*, *b*, there is a periodic disturbance of vortical structures above the 120-unit mark on the vertical axis. The average time interval between frames indicating the beginning of vortex formation and the end of vortex disruption is approximately 0.1 s, allowing us to estimate a frequency of approximately 10 Hz. The average temperature field clearly reveals a slight

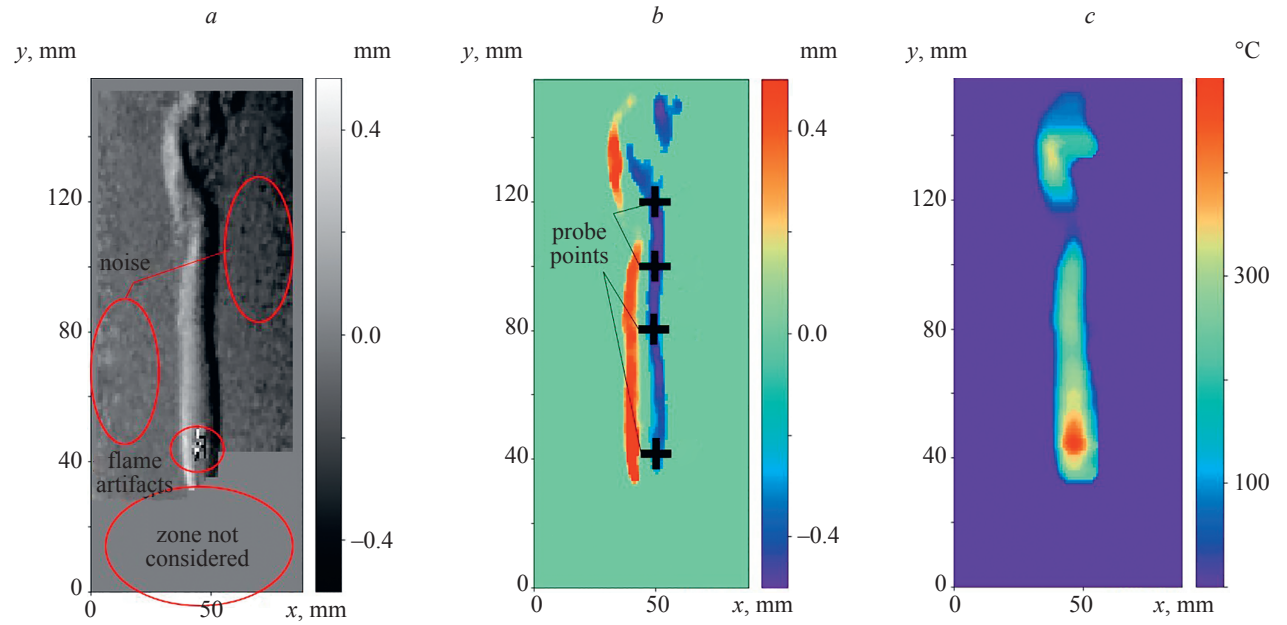

Fig. 2. Examples to the description of the data processing methodology: original displacement field (*a*); filtered offset field (*b*); instantaneous temperature field (*c*)

rightward deviation of the plume which is attributed to the fact that the plume periodically deviates rightward and then leftward in the analyzed dataset, with one additional half cycle occurring in the rightward direction compared to the leftward direction.

As shown by comparing the displacement waveforms at points *1–4* (Fig. 4) with the obtained flow patterns, plume oscillations occur every 2.5 s, corresponding to a frequency of 0.4 Hz. The *y* axis coordinate decreases as the sampling points move from *1* to *4*, and the amplitude of the oscillations becomes smaller. The frequencies of these oscillations approximately coincide, indicating a relatively stable, organized flow at point *4*. However, as the height increases, the oscillations become more disordered and chaotic, suggesting a turbulent flow.

An analysis of the average squares of the pulsations at different points (Fig. 5) depending on the *y* axis coordinate allows us to conclude that, on average, the plume decays at height *l* about 12*D*. At the same time, the average pulsation energy increases by more than four times. Thus, an analysis of the results from the perspective of pulsations in the displacement values unambiguously corresponds to the shadow pattern of the currents (Fig. 2, *a*). This allows us to confirm that the proposed approach can be used to analyze turbulence characteristics.

The results of the spectral analysis are particularly interesting (Fig. 6). The pulsation spectra at points *1* and *2* (Fig. 6, *a*, *b*) show typical turbulent spectra, with inertial regions where the amplitude of the pulsations decreases as the frequency increases to the power of "–5/3", in

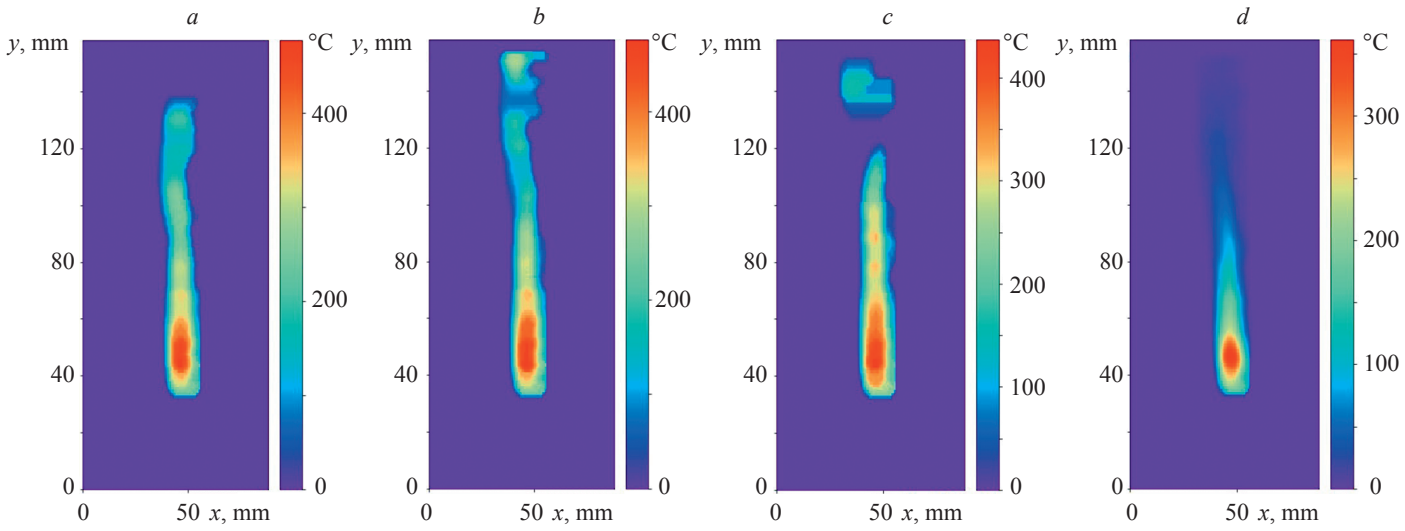

Fig. 3. Instantaneous temperature fields: the vortex formation beginning (*a*); the vortex formation (*b*); the vortex detachment (*c*); and time-averaged temperature field (*d*)

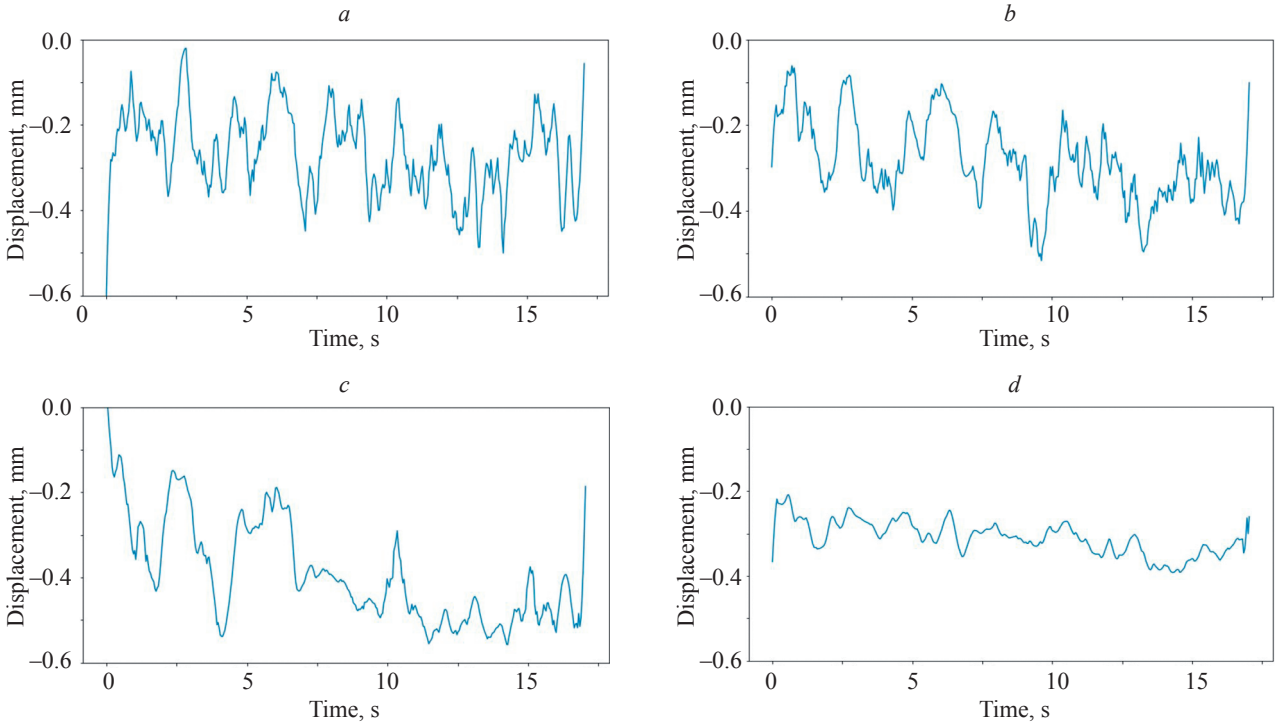

Fig. 4. Oscillograms of instantaneous displacement values at control points: *1* (*a*), *2* (*b*), *3* (*c*), and *4* (*d*)

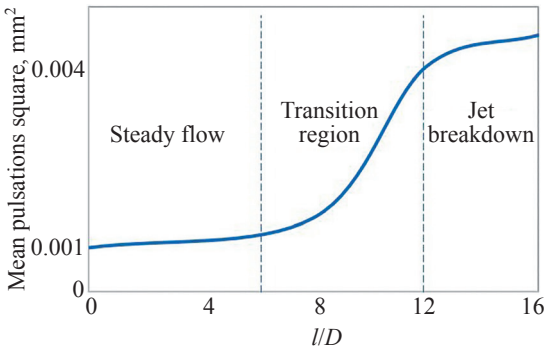

Fig. 5. Dependence of pulsation intensity vs. altitude

accordance with Kolmogorov's "–5/3" law. This confirms the conclusion from the waveform analysis that the flow becomes turbulent at a certain plume height.

At point *4* (Fig. 6, *d*), the spectrum shows a much less turbulent flow, with several peaks corresponding to the frequencies of the most significant pulsations, at approximately 0.4 Hz, 1 Hz, 2.5 Hz, and 10 Hz. The spectrum at point *3* lies between the spectrum at point *4* and those at points *1* and *2*, indicating a transition between these two types of flow. However, both in the case of this spectrum and in the spectra at points *1*, *2*, and *4*, we can easily see the presence of amplitude peaks near the frequencies of 0.4 Hz and 10 Hz. This indicates that these frequencies correspond to the main pulsations of the current.

As mentioned above, the plume oscillates from side to side at a frequency of 0.4 Hz. With a frequency of approximately 10 Hz, vortices form and break up, as was established by analyzing frames of the shadow flow pattern.

Therefore, the spectral analysis of the pulsations of displacement magnitude at characteristic points yields results that are well confirmed by the shadow patterns of the currents. This supports the hypothesis that the field Δx can act as a passive scalar and can be used for non-contact measurement of spectral flow characteristics.

The analysis of the spectral characteristics of the flow can also be performed based on the results obtained from the analysis of instantaneous temperature fields using the BOS method. However, due to the fact that the temperature field is calculated as a numerical solution to a Cauchy problem, there is an inevitable error in the calculation due to the discretization of the grid and the numerical scheme used. This can lead to the loss of a significant portion of the pulsation signals in the analysis.

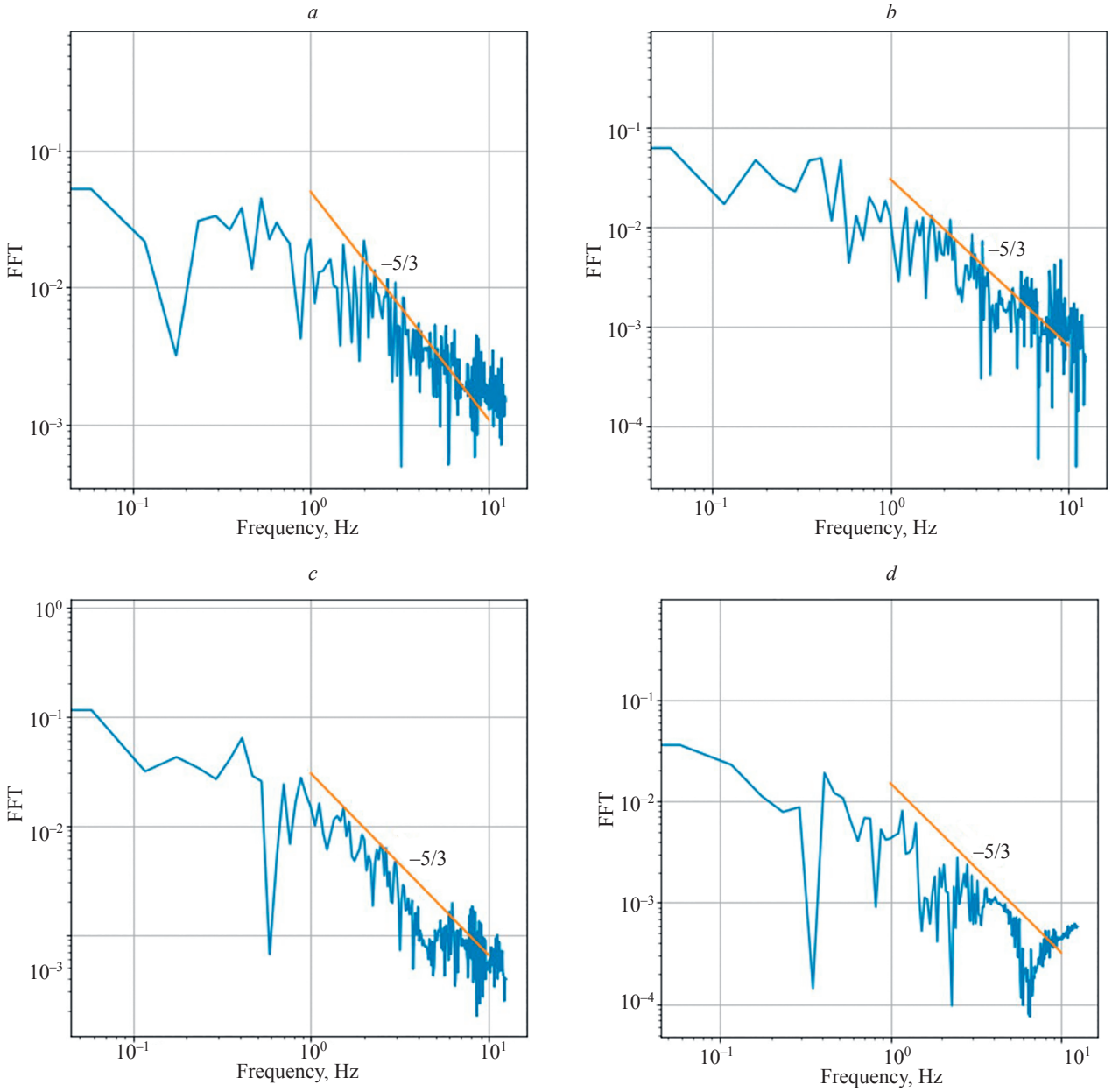

Fig. 6. Results of spectral analysis of pulsations: at points *1* (*a*), *2* (*b*), *3* (*c*), *4* (*d*)

Conclusion

Using the background oriented Schlieren method, shadow flow patterns of a convective plume formed over a laminar diffusion flame of a compact burner were obtained as well as instantaneous and averaged temperature fields.

The flow structure was analyzed and it was found that the plume has a steady flow zone below $l/D = 5$, with a transition region, and the plume disintegrates into separate vortices at $l/D = 12$.

Pulsation and spectral characteristics of the flow at control points were investigated. Spectral analysis revealed the main frequencies (0.4 and 10 Hz) of the flow corresponding to the phenomena observed in the shadow picture frames: the slow oscillation of the plume by x axis and the disruption of vortices in its upper part.

The similarity between the obtained spectral characteristics and the results of analyzing shadow patterns leads us to conclude that the instantaneous displacement fields of points obtained through visualization using the Background Oriented Schlieren method can be used as passive scalars to study flow pulsations non-invasively.

References

1. Braeuer A. Shadowgraph and schlieren techniques. *Supercritical Fluid Science and Technology*, 2015, vol. 7, pp. 283–312. https://doi.org/10.1016/B978-0-444-63422-1.00004-3
2. Dalziel S.B., Hughes G.O., Sutherland B.R. Whole-field density measurements by ‘synthetic schlieren’. *Experiments in Fluids*, 2000, vol. 28, no. 4, pp. 322–335. https://doi.org/10.1007/s003480050391
3. Shimazaki T., Ichihara S., Tagawa Y. Background oriented schlieren technique with fast Fourier demodulation for measuring large density-gradient fields of fluids. *Experimental Thermal and Fluid Science*, 2022, vol. 134, pp. 110598. https://doi.org/10.1016/j.expthermflusci.2022.110598
4. Li X., Gong S., Zhang F., Ma Z., Xun G. Three-dimensional tomographic reconstruction for gaseous fuel jets based on background oriented schlieren technique. *Journal of the Energy Institute*, 2025, vol. 120, pp. 102118. https://doi.org/10.1016/j.joei.2025.102118
5. Davami J., Juliano T.J., Moreto J.R., Liu X. Density measurements via background-oriented schlieren and parallel-ray omnidirectional integration. *Experiments in Fluids*, 2025, vol. 66, no. 4, pp. 78. https://doi.org/10.1007/s00348-025-04012-1
6. Martínez-González A., Moreno-Hernández D., Guerrero-Viramontes J.A., León-Rodríguez M., Zamarripa-Ramírez J.C.I., Carrillo-Delgado C. Temperature measurement of fluid flows by using a focusing schlieren method. *Sensors*, 2019, vol. 19, no. 1, pp. 12. https://doi.org/10.3390/s19010012
7. Ichihara S., Shimazaki T., Tagawa Y. Background-oriented schlieren technique with vector tomography for measurement of axisymmetric pressure fields of laser-induced underwater shock waves. *Experiments in Fluids*, 2022, vol. 63, no. 11, pp. 182. https://doi.org/10.1007/s00348-022-03524-4
8. Miao Y., Jia C., Hua Y., Sun L., Xu J., Wu D., Huang G., Liu H. Measurement of the concentration distribution of hydrogen jets using adaptive stream stripe- background oriented schlieren (ASS-BOS). *International Journal of Hydrogen Energy*, 2024, vol. 77, pp. 281–290. https://doi.org/10.1016/j.ijhydene.2024.06.099
9. Wang Q., Mei X.H., Wu Y., Zhao C.Y. An optimization and parametric study of a schlieren motion estimation method. *Flow, Turbulence and Combustion*, 2021, vol. 107, no. 3, pp. 609–630. https://doi.org/10.1007/s10494-021-00246-1
10. Yang S., Zhao L., Wang H., Li M., Xu W. Fluid motion prediction from schlieren for ethanol plume velocity measurement. *International Journal of Heat and Fluid Flow*, 2025, vol. 115, pp. 109889. https://doi.org/10.1016/j.ijheatfluidflow.2025.109889
11. Li J., Xiong Y., Tang Y., Han W., Pan C., Wang J. Three-dimensional diagnosis of lean premixed turbulent swirl flames using tomographic background oriented Schlieren. *Physics of Fluids*, 2024, vol. 36, no. 5, pp. 055159. https://doi.org/10.1063/5.0209235
12. Akamine M., Teramoto S., Okamoto K. Formulation and demonstrations of three-dimensional background-oriented schlieren using a mirror for near-wall density measurements. *Experiments in Fluids*, 2023, vol. 64, no. 7, pp. 134. https://doi.org/10.1007/s00348-023-03672-1
13. Bulat P.V., Volkov K.N. Numerical simulation of shock wave refraction on inclined contact discontinuity. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 3, pp. 550–558. (in Russian). https://doi.org/10.17586/2226-1494-2016-16-3-550-558
14. Bulat P.V., Volkov K.N. Numerical simulation of shock wave diffraction over right angle on unstructured meshes. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 2, pp. 354–362. (in Russian). https://doi.org/10.17586/2226-1494-2016-16-2-354-362
15. Grauer S.J., Unterberger A., Rittler A., Daun K.J., Kempf A.M., Mohri K. Instantaneous 3D flame imaging by background-oriented schlieren tomography. *Combustion and Flame*, 2018, vol. 196, pp. 284–299. https://doi.org/10.1016/j.combustflame.2018.06.022
16. Cowles R.A.P., Molnar J.P., Singh A.K., Grauer S.J. Tomographic background-oriented schlieren facility for buoyancy-driven flows and flames. *Proc. of the AIAA Science and Technology Forum and Exposition, AIAA*. SciTech Forum, 2025, https://doi.org/10.2514/6.2025-1058
17. Liu Y., Xing F., Su L., Tan H., Wang D. A mini-review of recent developments in plenoptic background-oriented schlieren technology for flow dynamics measurement. *Aerospace*, 2024, vol. 11, no. 4, pp. 303. https://doi.org/10.3390/aerospace11040303

Литература

1. Braeuer A. Shadowgraph and schlieren techniques // Supercritical Fluid Science and Technology. 2015. V. 7. P. 283–312. https://doi.org/10.1016/B978-0-444-63422-1.00004-3
2. Dalziel S.B., Hughes G.O., Sutherland B.R. Whole-field density measurements by ‘synthetic schlieren’ // Experiments in Fluids. 2000. V. 28. N 4. P. 322–335. https://doi.org/10.1007/s003480050391
3. Shimazaki T., Ichihara S., Tagawa Y. Background oriented schlieren technique with fast Fourier demodulation for measuring large density-gradient fields of fluids // Experimental Thermal and Fluid Science. 2022. V. 134. P. 110598. https://doi.org/10.1016/j.expthermflusci.2022.110598
4. Li X., Gong S., Zhang F., Ma Z., Xun G. Three-dimensional tomographic reconstruction for gaseous fuel jets based on background oriented schlieren technique // Journal of the Energy Institute. 2025. V. 120. P. 102118. https://doi.org/10.1016/j.joei.2025.102118
5. Davami J., Juliano T.J., Moreto J.R., Liu X. Density measurements via background-oriented schlieren and parallel-ray omnidirectional integration // Experiments in Fluids. 2025. V. 66. N 4. P. 78. https://doi.org/10.1007/s00348-025-04012-1
6. Martínez-González A., Moreno-Hernández D., Guerrero-Viramontes J.A., León-Rodríguez M., Zamarripa-Ramírez J.C.I., Carrillo-Delgado C. Temperature measurement of fluid flows by using a focusing schlieren method // Sensors. 2019. V. 19. N 1. P. 12. https://doi.org/10.3390/s19010012
7. Ichihara S., Shimazaki T., Tagawa Y. Background-oriented schlieren technique with vector tomography for measurement of axisymmetric pressure fields of laser-induced underwater shock waves // Experiments in Fluids. 2022. V. 63. N 11. P. 182. https://doi.org/10.1007/s00348-022-03524-4
8. Miao Y., Jia C., Hua Y., Sun L., Xu J., Wu D., Huang G., Liu H. Measurement of the concentration distribution of hydrogen jets using adaptive stream stripe- background oriented schlieren (ASS-BOS) // International Journal of Hydrogen Energy. 2024. V. 77. P. 281–290. https://doi.org/10.1016/j.ijhydene.2024.06.099
9. Wang Q., Mei X.H., Wu Y., Zhao C.Y. An optimization and parametric study of a schlieren motion estimation method // Flow, Turbulence and Combustion. 2021. V. 107. N. 3. P. 609–630. https://doi.org/10.1007/s10494-021-00246-1
10. Yang S., Zhao L., Wang H., Li M., Xu W. Fluid motion prediction from schlieren for ethanol plume velocity measurement // International Journal of Heat and Fluid Flow. 2025. V. 115. P. 109889. https://doi.org/10.1016/j.ijheatfluidflow.2025.109889
11. Li J., Xiong Y., Tang Y., Han W., Pan C., Wang J. Three-dimensional diagnosis of lean premixed turbulent swirl flames using tomographic background oriented Schlieren // Physics of Fluids. 2024. V. 36. N 5. P. 055159. https://doi.org/10.1063/5.0209235
12. Akamine M., Teramoto S., Okamoto K. Formulation and demonstrations of three-dimensional background-oriented schlieren using a mirror for near-wall density measurements // Experiments in Fluids. 2023. V. 64. N 7. P. 134. https://doi.org/10.1007/s00348-023-03672-1
13. Булат П.В., Волков К.Н. Численное моделирование рефракции ударной волны на наклонном контактном разрыве // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 3. С. 550–558. https://doi.org/10.17586/2226-1494-2016-16-3-550-558
14. Булат П.В., Волков К.Н. Численное моделирование дифракции ударной волны на прямом угле на неструктурированных сетках // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 2. С. 354–362. https://doi.org/10.17586/2226-1494-2016-16-2-354-362
15. Grauer S.J., Unterberger A., Rittler A., Daun K.J., Kempf A.M., Mohri K. Instantaneous 3D flame imaging by background-oriented schlieren tomography // Combustion and Flame. 2018. V. 196. P. 284–299. https://doi.org/10.1016/j.combustflame.2018.06.022
16. Cowles R.A.P., Molnar J.P., Singh A.K., Grauer S.J. Tomographic background-oriented schlieren facility for buoyancy-driven flows and flames // Proc. of the AIAA Science and Technology Forum and Exposition, AIAA. SciTech Forum. 2025. https://doi.org/10.2514/6.2025-1058
17. Liu Y., Xing F., Su L., Tan H., Wang D. A mini-review of recent developments in plenoptic background-oriented schlieren technology for flow dynamics measurement // Aerospace. 2024. V. 11. N 4. P. 303. https://doi.org/10.3390/aerospace11040303

18. Sasono M., Sakti S.P., Noor J.E., Soetedjo H. Application of checkerboard-based Background-Oriented Schlieren technique for invisible visualization of thermal plumes. *AIP Conference Proceedings*, 2023, vol. 2720, no. 1, pp. 040035. https://doi.org/10.1063/5.0136943
19. Gao P., Zhang Y., Yu X., Dong S., Chen Q., Yuan Y. Reconstruction method of 3D turbulent flames by background-oriented schlieren tomography and analysis of time asynchrony. *Fire*, 2023, vol. 6, no. 11, pp. 417. https://doi.org/10.3390/fire6110417
20. Vasiliev A., Sukhanovskii A., Frick P., Budnikov A., Fomichev V., Bolshukhin M., Romanov R. High Rayleigh number convection in a cubic cell with adiabatic sidewalls. *International Journal of Heat and Mass Transfer*, 2016, vol. 102, pp. 201–212. https://doi.org/10.1016/j.ijheatmasstransfer.2016.06.015
21. Liu H.C., Huang J.Q., Li L. Cai W.W. Volumetric imaging of flame refractive index, density, and temperature using background-oriented Schlieren tomography. *Science China Technological Sciences*, 2021, vol. 64, no. 1, pp. 98–110. https://doi.org/10.1007/s11431-020-1663-5
22. Wang G.T., Daniel K.A., Lynch K.P., Guildenbecher D.R., Mazumdar Y.C. High temperature and pressure Gladstone–Dale coefficient measurements in air behind reflected shock waves. *Physics of Fluids*, 2023, vol. 35, no. 8, pp. 086121. https://doi.org/10.1063/5.0162017
23. Chen Y., Cao R., Chen J., Liu L., Matsushita B. A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, vol. 180, pp. 174–190. https://doi.org/10.1016/j.isprsjprs.2021.08.015

18. Sasono M., Sakti S.P., Noor J.E., Soetedjo H. Application of checkerboard-based Background-Oriented Schlieren technique for invisible visualization of thermal plumes // AIP Conference Proceedings. 2023. V. 2720. N 1. P. 040035. https://doi.org/10.1063/5.0136943
19. Gao P., Zhang Y., Yu X., Dong S., Chen Q., Yuan Y. Reconstruction method of 3D turbulent flames by background-oriented schlieren tomography and analysis of time asynchrony // Fire. 2023. V. 6. N 11. P. 417. https://doi.org/10.3390/fire6110417
20. Vasiliev A., Sukhanovskii A., Frick P., Budnikov A., Fomichev V., Bolshukhin M., Romanov R. High Rayleigh number convection in a cubic cell with adiabatic sidewalls // International Journal of Heat and Mass Transfer. 2016. V. 102. P. 201–212. https://doi.org/10.1016/j.ijheatmasstransfer.2016.06.015
21. Liu H.C., Huang J.Q., Li L. Cai W.W. Volumetric imaging of flame refractive index, density, and temperature using background-oriented Schlieren tomography // Science China Technological Sciences. 2021. V. 64. N 1. P. 98–110. https://doi.org/10.1007/s11431-020-1663-5
22. Wang G.T., Daniel K.A., Lynch K.P., Guildenbecher D.R., Mazumdar Y.C. High temperature and pressure Gladstone–Dale coefficient measurements in air behind reflected shock waves // Physics of Fluids. 2023. V. 35. N 8. P. 086121. https://doi.org/10.1063/5.0162017
23. Chen Y., Cao R., Chen J., Liu L., Matsushita B. A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter // ISPRS Journal of Photogrammetry and Remote Sensing. 2021. V. 180. P. 174–190. https://doi.org/10.1016/j.isprsjprs.2021.08.015

Authors

Pavel A. Bryzgunov — PhD, Assistant, National Research University "Moscow Power Engineering Institute", Moscow, 111250, Russian Federation, sc 57844836600, https://orcid.org/0000-0003-3710-5116, bryzgunovpa@mpei.ru
Dmitry S. Pisarev — Senior Lecturer, National Research University "Moscow Power Engineering Institute", Moscow, 111250, Russian Federation, sc 16239539100, https://orcid.org/0009-0006-3091-4884, pisarevds@mpei.ru
Olga V. Zlyvko — PhD (Economy), Associate Professor, Associate Professor, National Research University "Moscow Power Engineering Institute", Moscow, 111250, Russian Federation, sc 57060525900, https://orcid.org/0000-0003-0554-4026, zlyvkoov@mpei.ru
Andrey N. Rogalev — D.Sc., Associate Professor, Head of Department, National Research University "Moscow Power Engineering Institute", Moscow, 111250, Russian Federation, sc 34980078500, https://orcid.org/0000-0001-7256-0144, rogalevan@mpei.ru
Nikolay D. Rogalev — D.Sc., Professor, Rector, National Research University "Moscow Power Engineering Institute", Moscow, 111250, Russian Federation, sc 6507029432, https://orcid.org/0000-0002-6458-2869, rogalevnd@mpei.ru

Авторы

Брызгунов Павел Александрович — кандидат технических наук, ассистент, Национальный исследовательский университет «МЭИ», Москва, 111250, Российская Федерация, sc 57844836600, https://orcid.org/0000-0003-3710-5116, bryzgunovpa@mpei.ru
Писарев Дмитрий Сергеевич — старший преподаватель, Национальный исследовательский университет «МЭИ», Москва, 111250, Российская Федерация, sc 16239539100, https://orcid.org/0009-0006-3091-4884, pisarevds@mpei.ru
Злывко Ольга Владимировна — кандидат экономических наук, доцент, доцент, Национальный исследовательский университет «МЭИ», Москва, 111250, Российская Федерация, sc 57060525900, https://orcid.org/0000-0003-0554-4026, zlyvkoov@mpei.ru
Рогалев Андрей Николаевич — доктор технических наук, доцент, заведующий кафедрой, Национальный исследовательский университет «МЭИ», Москва, 111250, Российская Федерация, sc 34980078500, https://orcid.org/0000-0001-7256-0144, rogalevan@mpei.ru
Рогалев Николай Дмитриевич — доктор технических наук, профессор, ректор, Национальный исследовательский университет «МЭИ», Москва, 111250, Российская Федерация, sc 6507029432, https://orcid.org/0000-0002-6458-2869, rogalevnd@mpei.ru

Received 04.07.2025
Approved after reviewing 31.07.2025
Accepted 15.09.2025

Статья поступила в редакцию 04.07.2025
Одобрена после рецензирования 31.07.2025
Принята к печати 15.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-961-970
УДК 004.896

Разработка и исследование метода обучения с подкреплением для акустической диагностики промышленного оборудования

Наталья Аркадьевна Верзун[1]✉, Михаил Олегович Колбанёв[2], Аделина Рустамовна Салиева[3]

[1,2] Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация
[1,2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В. И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация
[3] Лига цифровой экономики, Москва, 127015, Российская Федерация

[1] Verzun.n@unecon.ru✉, http://orcid.org/0000-0002-0126-2358
[2] mokolbanev@mail.ru, http://orcid.org/0000-0003-4825-6972
[3] Rustamovna.a3@gmail.com, https://orcid.org/0009-0001-9519-5773

Аннотация

Введение. Исследована актуальная задача акустической диагностики автономно работающего промышленного оборудования. Обзор существующих подходов к акустической диагностике, включая методы на основе сверточных нейронных сетей и алгоритмы обучения с учителем, показал их ограничения, такие как необходимость использования для обучения больших объемов размеченных данных, слабая адаптация к изменяющимся условиям и отсутствие механизма принятия решений в реальном времени. Предложен новый подход к акустической диагностике на основе методов обучения с подкреплением, отличающийся способностью к адаптации, высокой устойчивостью к шуму и возможностью непрерывного обучения в динамической среде. **Метод.** Представленный метод определения состояния работоспособности оборудования использует подход, основанный на исследовании акустических сигналов, издаваемых работающим оборудованием. Метод включает построение нейронной сети, выбор аудиозаписей из открытых библиотек аудиофайлов и обучение сети при помощи алгоритма с подкреплением. Процесс акустической диагностики состояния исправности/неисправности промышленного оборудования предполагает четыре этапа: фиксацию в режиме реального времени акустических данных работающего оборудования, извлечение признаков состояния оборудования, обучение с подкреплением нейронной сети и принятие решения о исправности/неисправности оборудования. **Основные результаты.** На основе размеченных аудиофайлов из открытых баз данных проведен эксперимент по идентификации различных состояний оборудования: нормальное состояние, начальная стадия дефекта, критическая неисправность. Результаты показали точность классификации от 89,7 % до 98,5 % и среднее время отклика от 0,5 до 0,7 с при низкой вычислительной нагрузке (в среднем загрузка центрального процессора 36,5 % и объем потребляемой оперативной памяти 509 МБ). **Обсуждение.** В отличие от известных систем акустической диагностики, основанных на алгоритмах обучения с учителем нейронных и сверточных нейронных сетей на предварительно размеченных базах данных, содержащих акустические сигналы, издаваемые работающим оборудованием, в предлагаемом подходе реализуется декомпозиция исходных акустических сигналов на спектральные составляющие. Каждая из этих составляющих анализируется и снабжается признаками, отражающими состояние исправности/неисправности оборудования. Такой подход позволяет: использовать алгоритмы обучения с подкреплением для принятия решений на основе стратегии; сократить время обучения модели за счет предварительного выделения значимых признаков; повысить точность диагностики; снизить вычислительную нагрузку и требования к аппаратным ресурсам. Разработанный алгоритм может применяться для непрерывного мониторинга состояния оборудования и предиктивного обслуживания в автономно функционирующих промышленных системах. Его использование позволит надежно и своевременно выявлять, и классифицировать неисправности промышленного оборудования. Алгоритм возможно доработать с учетом требований к интеграции с инфраструктурой интернета вещей, повышения устойчивости к внешним шумам и внедрения более продвинутых алгоритмов обучения с подкреплением, таких как Proximal Policy Optimization или Asynchronous Advantage Actor-Critic.

Ключевые слова
акустическая диагностика, промышленное оборудование, обучение с подкреплением, классификация состояний, RL-агент, спектральный анализ

Ссылка для цитирования: Верзун Н.А., Колбанёв М.О., Салиева А.Р. Разработка и исследование метода обучения с подкреплением для акустической диагностики промышленного оборудования // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 961–970. doi: 10.17586/2226-1494-2025-25-5-961-970

Development and research of a reinforcement learning method for acoustic diagnostics of industrial equipment

Natalya A. Verzun[1]✉, Mikhail O. Kolbanev[2], Adelina R. Salieva[3]

[1,2] Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation
[1,2] Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197376, Russian Federation
[3] Digital Economy League, Moscow, 127015, Russian Federation

[1] Verzun.n@unecon.ru✉, http://orcid.org/0000-0002-0126-2358
[2] mokolbanev@mail.ru, http://orcid.org/0000-0003-4825-6972
[3] Rustamovna.a3@gmail.com, https://orcid.org/0009-0001-9519-5773

Abstract
The actual problem of acoustic diagnostics of autonomously operating industrial equipment is investigated. An overview of existing approaches to acoustic diagnostics, including methods based on convolutional neural networks and learning algorithms with a teacher, is provided. Their limitations have been identified, such as the need to use large amounts of labeled data for training, poor adaptation to changing conditions, and the lack of a real-time decision-making mechanism. A new approach to acoustic diagnostics based on reinforcement learning methods is proposed, characterized by adaptability, high resistance to noise and the possibility of continuous learning in a dynamic environment. The proposed method for determining the state of equipment operability uses an approach based on the study of acoustic signals emitted by operating equipment. The method includes building a neural network, selecting audio recordings from open audio file libraries, and training the network using a reinforcement learning algorithm. The process of acoustic diagnostics of the state of serviceability/ malfunction of industrial equipment involves four stages: real-time recording of acoustic data of working equipment, extraction of signs of equipment condition, training with reinforcement of a neural network and making a decision on the serviceability / malfunction of the equipment. Based on tagged WAV audio files from open databases, an experiment was conducted to identify various states of the equipment: normal condition, initial stage of the defect, critical malfunction. The results showed classification accuracy from 89.7 % to 98.5 % and average response time from 0.5 to 0.7 seconds with low computing load (on average 36.5 % CPU and 509 MB RAM). Unlike the well-known acoustic diagnostic systems based on teacher-learning algorithms for neural and convolutional neural networks on pre-marked datasets containing acoustic signals emitted by running equipment, the proposed approach implements the decomposition of the initial acoustic signals into spectral components. Each of these components is analyzed and provided with signs reflecting the state of serviceability or malfunction of the equipment. This approach allows you to: use reinforcement learning algorithms for strategic decision-making; reduce model training time by pre-selecting significant features; improve diagnostic accuracy; reduce computational load and hardware resource requirements. The developed algorithm can be used for continuous monitoring of equipment condition and predictive maintenance in autonomously functioning industrial systems. Its use will allow reliable and timely detection and classification of industrial equipment malfunctions. It is possible to refine the algorithm to meet the requirements for integration with the IoT infrastructure, increase resistance to external noise, and implement more advanced RL algorithms such as PPO.

Keywords
acoustic diagnostics, industrial equipment, reinforcement learning, classification of states, RL agent, spectral analysis

For citation: Verzun N.A., Kolbanev M.O., Salieva A.R. Development and research of a reinforcement learning method for acoustic diagnostics of industrial equipment. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 961–970 (in Russian). doi: 10.17586/2226-1494-2025-25-5-961-970

Введение

С появлением большого числа автономно работающих технических систем возникает необходимость в разработке новых мер по обеспечению их надежности и отказоустойчивости [1–4]. Традиционные методы диагностики, такие как периодический осмотр и плановое обслуживание, не всегда способны предотвратить внезапные отказы оборудования [5]. Альтернативой традиционным методам могут выступать интеллектуальные системы непрерывного мониторинга, которые анализируют акустические сигналы, издаваемые оборудованием, классифицируют их и оперативно выявляют потенциальные неисправности [6, 7].

Развитие методов машинного обучения: обучение с учителем (Supervised Learning) и сверточные нейронные сети (Convolutional Neural Networks, CNN), значительно повысило эффективность решения задач классификации акустических сигналов [8]. Так, например, в области акустической диагностики использование алгоритма обучение с подкреплением (Reinforcement Learning, RL), основанного на применении глубоких нейронных сетей (Deep Q-Learning) уже продемонстрировало свою эффективность для распознавания

неисправностей в промышленных станках, улучшив точность классификации до 94,5 % [9]. Алгоритм Policy Gradient позволяет адаптироваться к изменяющимся условиям работы оборудования, например, для диагностики подшипников качения [10].

Эти алгоритмы показали хорошие результаты в распознавании звуков, однако их применение на практике сталкивается с рядом ограничений. Такие методы требуют огромных объемов размеченных данных для обучения, что затруднительно в реальных условиях. Они плохо адаптируются к динамически меняющимся условиям функционирования промышленного оборудования и не способны эффективно классифицировать неисправности. Алгоритмы обучались на размеченных «чистых» данных [11], которые представляли звуки работающего оборудования (состояние исправного/несправного оборудования), а реальные производственные условия зачастую характеризуются высоким уровнем внешних шумов [12]. Альтернативой этим методам может быть обучение с подкреплением, которое представляет собой более гибкий подход. Методы RL адаптируются к изменениям в среде, а также могут эффективно справляться с задачами, требующими последовательных решений в динамических условиях.

Сочетание глубокого обучения с методами RL предоставляет возможность объединить лучшие качества обеих технологий. Гибридные архитектуры, включающие CNN для извлечения признаков и RL-агентов для принятия решений, создают системы, способные эффективно решать задачи классификации и диагностики, даже в условиях изменяющихся характеристик оборудования. Такие подходы уже активно применяются для диагностики дефектов подшипников, мониторинга вибраций турбин и предсказания оставшегося ресурса оборудования [13, 14].

Целью исследования является разработка метода акустической диагностики промышленного оборудования на основе обучения с подкреплением и спектрального анализа сигналов, обеспечивающего высокую точность, адаптивность и устойчивость к шумам при работе в реальном времени. В отличие от известных систем акустической диагностики, с жестко заданной логикой принятия решений, разработанная система формирует оптимальную стратегию классификации за счет взаимодействия RL-агента с окружающей средой. Метод сочетает извлечение информативных спектральных признаков (амплитудный и частотный спектры, Mel-Frequency Cepstral Coefficients, MFCC) с обучением агента на основе наблюдаемых состояний, что позволяет учитывать стохастическую природу акустических сигналов и реализовать механизм обратной связи для поэтапного обновления диагностической политики в реальном времени [15, 16].

Подход к акустической диагностике промышленного оборудования

Система акустической диагностики, использующая методы RL, в общем случае состоит из: микрофонов и датчиков для сбора акустических данных; модуля предварительной обработки сигналов; модуля классификации звуковых паттернов; модуля принятия решений о техническом состоянии оборудования. Взаимодействие RL-агента и окружающей среды для предлагаемого подхода к акустической диагностике показано на рис. 1.

RL-агент — основной компонент интеллектуального анализа акустических сигналов. При обучении с подкреплением RL-агент обучается взаимодействовать с окружающей средой, выполняя действия на основе текущего состояния и получая за них вознаграждения [17]. В настоящей работе: состояния — спектральные характеристики акустических сигналов, представляющие признаки состояния оборудования; действия — принимаемые решения о классификации состояния оборудования (например, выявление нормальной работы или определение неисправности); награды — система поощрений, которая учитывает такие параметры, как точность диагностики и своевременность обнаружения неисправностей.

RL-агент обучается на основе поступающих данных и обратной связи, анализирует акустические характеристики сигналов, распознает закономерности и принимает решения о состоянии оборудования. На каждом временном шаге RL-агент оценивает текущее состояние системы, определяя, работает ли оборудование в нормальном режиме или имеются признаки неисправно-

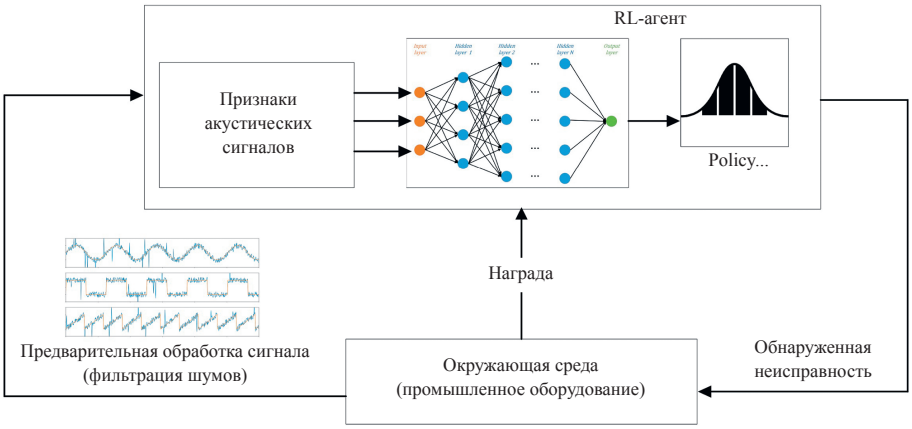

Рис. 1. Модель взаимодействия RL-агента и среды в задаче акустической диагностики

Fig. 1. The model of agent-environment interaction in the acoustic diagnostics task

сти. Его эффективность определяется системой наград, которая оценивает, насколько точно и своевременно он выявляет отклонения. Для того чтобы RL-агент мог принимать обоснованные решения, в процессе анализа используются различные признаки состояния.

Признаки состояния представляют собой набор признаков, извлеченных из акустических сигналов, при этом состояние на каждом временном шаге представлено вектором признаков, извлеченных из акустического сигнала. Такими признаками являются: амплитудный спектр (позволяет анализировать интенсивность сигналов в различных диапазонах частот); частотный спектр (отображает распределение частот в сигнале, выявляя потенциальные аномалии); MFCC.

Выделение и учет этих признаков позволяет проводить более детальный анализ сигналов, что особенно важно при диагностике сложных акустических процессов и благодаря чему достигается: устойчивость к шуму, возможность фильтрации случайных помех и выделение ключевых акустических характеристик работающего оборудования.

Математическая модель

Признаки акустических сигналов образуют многомерный вектор данных, который помогает RL-агенту распознавать изменения в звуках и повышать точность диагностики. На временном шаге t состояние $\mathbf{s}_t$ представляет собой многомерный вектор, включающий признаки [18–21] перечисленные в разделе «Подход к акустической диагностике промышленного оборудования».

Амплитудный спектр извлекается модулем предварительной обработки сигналов с использованием быстрого преобразования Фурье:

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi fn/N}, \quad (1)$$

где $x(n)$ — временной сигнал; N — количество отсчетов; f — частота; j — мнимая единица.

Частотный спектр определяется как набор значений f_i, соответствующих дискретным частотам:

$$f_i = \frac{if_s}{N_{FFT}}, \; i = 0, 1, \ldots, N_{FFT}, \quad (2)$$

где f_s — частота дискретизации; N_{FFT} — размер быстрого преобразования Фурье.

MFCC имеют вид

$$MFCC(k) = \sum_{m=1}^{M} \log(|F_m|) \cos\left[k\,(m - 0{,}5)\frac{\pi}{M}\right], \quad (3)$$

где M — количество фильтров в мел-шкале; F_m — спектральная энергия в фильтре m; k — индекс коэффициента MFCC.

Состояние $\mathbf{s}_t$ представляется как вектор

$$\mathbf{s}_t = \{f_1, f_2, \ldots, f_i\}, \; t = 1, 2, \ldots, \quad (4)$$

где f_i — спектральные признаки, извлеченные из акустического сигнала на момент времени t.

Действия $\alpha \in A$ представляют собой выбор одного из возможных состояний оборудования. Например: α_1 — оборудование в нормальном состоянии, α_2 — обнаружен незначительный дефект, α_3 — критическая неисправность. Для дискретного пространства действий используется политика $\pi(\alpha|s; \theta)$, которая определяет вероятность выбора действия α при наблюдаемом состоянии s

$$\pi(\alpha|s; \theta) = P(A = \alpha|S = s; \theta), \quad (5)$$

где θ — параметры модели (нейронной сети); $\pi(\alpha|s; \theta)$ — вероятность выбора действия α в состоянии s; s — текущее состояние (вектор признаков); α — конкретное действие.

Награда τ вычисляется как комбинация точности классификации и своевременности:

$$\tau_e = \omega_1 Accuracy(a_t, \hat{a}) - \omega_2 Delay(a_t), \quad (6)$$

где $Accuracy(a_t, \hat{a})$ — бинарный показатель точности классификации (1 — правильная классификация, 0 — ошибочная); $Delay(a_t)$ — временной штраф за позднее обнаружение дефекта; ω_1 и ω_2 — веса, задающие баланс между точностью и своевременностью.

Обучение RL-агента направлено на максимизацию ожидаемой совокупной награды:

$$J(\theta) = E_{\pi\theta}\left[\sum_{t=0}^{T} y^t \tau_t\right], \quad (7)$$

где y — коэффициент дисконтирования, задающий приоритет краткосрочным наградам; θ — параметры нейронной сети, представляющей политику. Обновление параметров будет производиться с помощью градиентного метода.

После извлечения признаков из акустического сигнала RL-агент выполняет классификацию состояния оборудования, определяя возможные неисправности. Однако для эффективного функционирования системы недостаточно просто классифицировать сигналы — необходимо также учитывать неопределенность, адаптироваться к изменениям в данных и минимизировать ошибки диагностики. Эти задачи решает модуль принятия решений, который включает компоненты: политику выбора действий — $\pi(\alpha|s)$; функцию ценности — $\pi(s; \theta)$; систему оценки уверенности — $Confidence(\alpha|s)$; механизм обучения с использованием критерия ошибок, т. е. функцию потерь — L; градиентное обновление параметров нейронной сети — θ.

Политика выбора действий $\pi(\alpha|s)$ определяет, какое решение RL-агент примет на основе наблюдаемого состояния оборудования. Задается вероятностной моделью:

$$\pi(s, \theta) = \frac{\exp(Q(s, a))}{\sum_{a' \in A} \exp(Q(s, a'))}, \quad (8)$$

где $\pi(s, \theta)$ — оценка качества действия a в состоянии s (функция ценности); $Q(s, a)$ — оценка качества действия a в состоянии s; a' — произвольное действие из множества всех возможных действий A в состоянии

s; a — параметр, который регулирует баланс между исследованием и эксплуатацией.

RL-агент выбирает действие на основе вероятности $\pi(\alpha|s)$:

$$a_t = \arg\max \pi(\alpha|\mathbf{s}_t), \quad (9)$$

где a_t — выбранное действие в момент t; arg max — операция, возвращающая аргумент (в данном случае действие), при котором функция достигает максимума; $\pi(\alpha|\mathbf{s}_t)$ — вероятность выбора действия a в состоянии $\mathbf{s}_t$. Для оценки $Q(s, a)$ используется алгоритм Deep Q-Learning.

Система оценки уверенности позволяет учитывать степень надежности каждого принятого решения:

$$Confidence(\alpha|s) = \pi(\alpha|s), \quad (10)$$

если уверенность ниже порогового значения δ, то решение откладывается, а сигнал передается для повторного анализа.

Механизм обратной связи корректирует модель RL-агента на основе предыдущего опыта, повышая ее точность и может быть организован через механизм обучения с использованием критерия ошибок (функции потерь):

$$L = \frac{1}{B}\sum_{i=1}^{B}[Q(s_i, a_i) - \hat{r}_i]^2, \quad (11)$$

где B — размер обучающей выборки; $Q(s_i, a_i)$ — предсказанное Q-значение для пары (состояние, действие); r_i — фактическое значение награды; $\sum_{i=1}^{N}[Q(s_i, a_i) - \hat{r}_i]^2$ — квадрат ошибки.

Градиентное обновление:

$$\theta \leftarrow \theta - \eta \times \nabla_\theta L, \quad (12)$$

где θ — параметры модели; η — скорость обучения; $\nabla_\theta L$ — градиент функции потерь по параметрам.

Алгоритм обучения

Используя представленную математическую модель, была создана программа [22], реализующая метод акустической диагностики (рис. 2). В ее основе лежат алгоритмы обработки сигналов, механизм обучения RL-агента и модуль принятия решений, что обеспечивает автоматическое выявление неисправностей и адаптацию к изменяющимся условиям эксплуатации оборудования. Реализация включает в себя обработку акустических сигналов и динамическое обновление модели посредством обратной связи, что повышает надежность и эффективность диагностики.

Приведенные в разделе «Математическая модель» соотношения лежат в основе процесса функционирования системы акустической диагностики, которая реализована поэтапно в соответствии с блок-схемой на рис. 2.

Приведем сопоставление математических выражений с этапами реализации предложенного алгоритма.

Этап 1. Инициализация системы. Выполняется начальная настройка системы, которая включает установку параметров обучения RL-агента, инициализацию структуры модели, логгера и директории хранения. На этапе 1 происходит задание параметров и скорости

Рис. 2. Этапы акустической диагностики

Fig. 2. Stages of acoustic diagnostics

обучения, которые далее участвуют в формуле градиентного обновления (12).

Этап 2. Загрузка данных. Акустический сигнал подвергается сегментации и нормализации, подготавливается для последующего анализа.

Этап 3. Аугментация. Применяются преобразования, имитирующие вариации внешних условий. Отметим, что на этапе 3 не используется отдельная формула, но он влияет на корректность построения вектора признаков (4).

Этап 4. Извлечение признаков. Извлечение реализуется с помощью быстрого преобразования Фурье формула (1), вычисления частотных составляющих (2) и извлечения MFCC (3).

Этап 5. Обучение модели. На основе извлеченных признаков формируется вектор состояния агента (4). Процесс обучения включает оптимизацию параметров модели θ, обновляемых по методу градиентного спуска на основе функций потерь (11) и (12).

Этап 6. Буфер воспроизведения. Осуществляется выборка обучающих примеров с приоритетами, что позволяет реализовать эффективное обучение по методу Deep Q-Learning (8)–(10).

Этап 7. Обновление модели. Производится вычисление целевой функции награды (6), а также обновление целевой сети Q-функции, что соответствует максимизации ожидаемой совокупной награды (7).

Этап 8. Оценка. Анализируются Accuracy, Loss и Reward — метрики, основанные на функции потерь, точности классификации и награде (формулы см., например, в [12]).

Этап 9. Интеграция. На данном этапе система отправляет уведомления при критических состояниях, определенных на основе вероятностной политики (5).

Этап 10. Сохранение. Выполняется сохранение обученной модели с параметрами θ для последующего применения и дообучения.

Описание эксперимента

Целью эксперимента являлась оценка эффективности разработанного алгоритма RL при диагностике различных состояний промышленного оборудования на основе анализа акустических сигналов. В рамках исследования основное внимание уделялось способности модели точно и своевременно классифицировать техническое состояние оборудования, а также устойчивости алгоритма к фоновым шумам и изменяющимся условиям эксплуатации.

Исходные данные. Для обеспечения воспроизводимости эксперимента и репрезентативности обучающей выборки были использованы открытые общедоступные базы данных, содержащие акустические и вибрационные сигналы промышленного оборудования в различных состояниях:

— Case Western Reserve University Bearing Data Center — база данных, включающая записи вибрационных и акустических сигналов подшипников с контролируемыми дефектами различной степени [23];
— MIMII Dataset (Malfunctioning Industrial Machine Investigation and Inspection) — содержит акустические записи насосов, вентиляторов, клапанов и компрессоров в нормальном и аномальном состояниях [24];
— DCASE Challenge 2020 Dataset — акустические данные для задач мониторинга технического состояния машин в условиях фоновых шумов и неопределенности среды [25].

Для обучения и тестирования модели применялись аудиофайлы в формате Waveform Audio File Format, снабженные метками состояний оборудования. Предварительная обработка включала нормализацию сигналов, выделение спектральных признаков (амплитудный и частотный спектр, MFCC) и формирование входных векторов признаков.

Оценка алгоритма акустического распознавания. Тестирование алгоритма проводилось на различных типах оборудования для оценки его эффективности в распознавании аномалий. Для оценки качества классификации были использованы общепринятые метрики (Accuracy, Precision, Recall, F1-мера). Важность их применения в данной работе обусловлена не только общепринятостью, но и необходимостью комплексной количественной оценки работы RL-агента [23]: Accuracy, Precision и Recall позволяют понять компромисс между количеством ложных срабатываний и пропущенных дефектов, что критически важно в задачах промышленной диагностики; F1-мера отражает обобщенную эффективность диагностики при наличии несбалансированных классов.

Также в рамках оценки рассчитывался показатель ложных срабатываний, что особенно важно в контексте промышленной эксплуатации, где ложные тревоги ведут к экономическим потерям.

Целевые классы и постановка задачи. Эксперимент охватывает три целевых состояния оборудования: нормальное состояние — сигнал характеризуется стабильной спектральной структурой без аномалий; начальная стадия дефекта — слабовыраженные отклонения в сигнале, указывающие на поверхностные повреждения (трещины, износ); критическая неисправность — ярко выраженные частотные всплески и высокоамплитудные аномалии, отражающие значительные дефекты (глубокие трещины, разрушение элемента).

RL-агент обучался классифицировать текущие состояния на основе анализа извлеченных признаков, а также адаптироваться к изменениям акустической картины за счет механизма обратной связи и функции награды, описанной в математической модели.

Все метрики рассчитывались на тестовой выборке, сформированной по принципу стратифицированного разбиения (80/20), с последующим усреднением по результатам 5 независимых запусков модели.

Результаты. Для оценки эффективности разработанного алгоритма обучения было проведено тестирование. Основное внимание уделялось точности классификации различных состояний подшипника, а также способности алгоритма идентифицировать неисправности в различных типах промышленного оборудования.

Результаты тестирования алгоритма обучения представлены в табл. 1 и 2. Сравнение метрик по типам оборудования и неисправностей приведено на рис. 3.

Таблица 1. Точность классификации по состояниям подшипника
Table 1. Accuracy of classification according to bearing conditions

Состояние подшипника	Точность классификации, %	Время отклика, с	Описание
Нормальное состояние	98,5	0,5	Сигналы подшипника без признаков повреждений
Начальная стадия дефекта	94,2	0,6	Легкие повреждения, такие как трещины или износ
Критическая неисправность	89,7	0,7	Серьезные повреждения: глубокие трещины и сильный износ

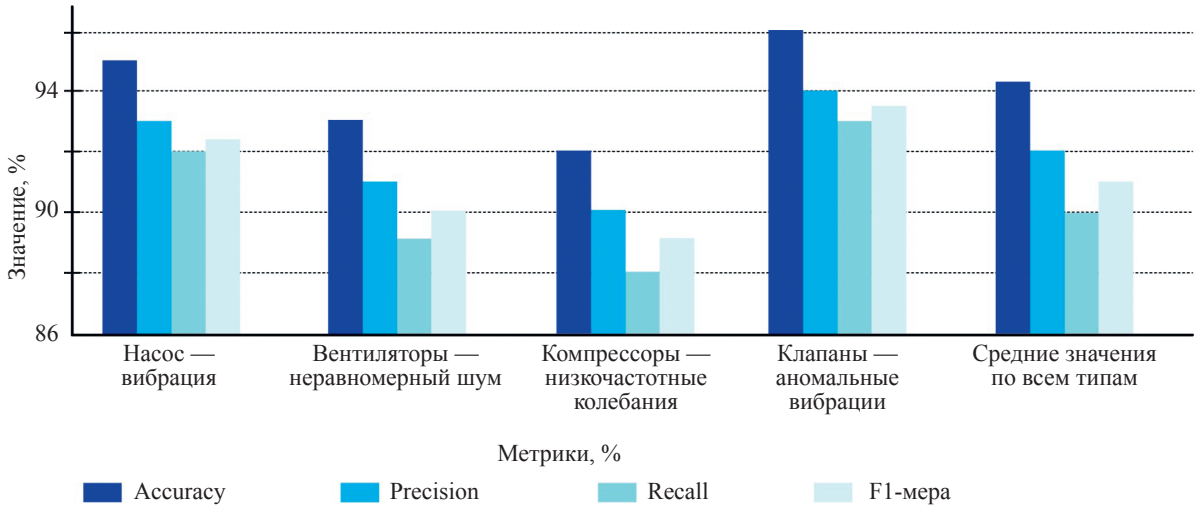

Рис. 3. Сравнение метрик по типам оборудования и неисправностей
Fig. 3. Comparison of metrics by type of equipment and malfunction

Таблица 2. Сравнение загрузки CPU, RAM и времени выполнения по типу оборудования
Table 2. Comparison of CPU, RAM, and execution time by hardware type

Тип оборудования	Тип нарушения	Загрузка CPU, %	Загрузка RAM, МБ	Время отклика, с
Насос	Вибрация	35	512	0,21
Вентиляторы	Неравномерный шум	38	490	0,26
Компрессоры	Низкочастотные колебания	40	530	0,27
Клапаны	Аномальные вибрации	33	505	0,25

Для оценки применимости разработанного алгоритма в условиях реального времени особое внимание уделялось его вычислительной эффективности. В рамках экспериментов измерялись ключевые показатели нагрузки на систему: средняя загрузка центрального процессора (CPU), объем потребляемой оперативной памяти (RAM) и время отклика алгоритма. В табл. 2 показаны показатели нагрузки на компьютер и время отклика алгоритма.

Обсуждение

Проведенные исследования подтвердили эффективность предложенного RL-алгоритма в задачах акустической диагностики неисправностей промышленного оборудования. Средние значения Accuracy, Recall и F1-меры варьируются в пределах 90–95 %, что свидетельствует о надежности и устойчивости алгоритма при обработке сложных акустических сигналов.

Наибольшие показатели достигнуты при анализе вибраций насосов и клапанов, где F1-мера составила 92,5 % и 93,5 %. Это указывает на способность модели точно классифицировать как нормальные, так и аномальные состояния. В случае вентиляторов и компрессоров, несмотря на более сложные акустические профили, алгоритм продемонстрировал стабильные результаты с F1-мерой 90 % и 89 % соответственно.

С точки зрения вычислительной эффективности, алгоритм показал низкую нагрузку на систему: средняя загрузка CPU составила 36,5 %, потребление RAM — около 509 МБ, а среднее время отклика — 0,25 с. Это делает его пригодным для использования в системах реального времени.

Низкая загрузка CPU и оперативной памяти позволяет интегрировать предложенный метод в промышленные контроллеры и edge-устройства, обеспечивая непрерывную диагностику без существенной нагрузки на вычислительные ресурсы.

Однако, несмотря на положительные результаты, метод имеет ряд ограничений:

— необходима калибровка параметров награды и архитектуры RL-агента под конкретный тип оборудования;
— использование алгоритма требует предварительного извлечения спектральных признаков, что повышает сложность подготовки данных;
— в текущей реализации не учитываются мультисенсорные источники (вибрация, температура), которые могли бы улучшить устойчивость диагностики.

Заключение

Проведен анализ существующих подходов к акустической диагностике промышленного оборудования, включая методы машинного обучения с учителем и сверточные нейронные сети. Применение данных подходов на практике позволяет достигать высокой точности распознавания неисправностей (до 94,5 %) и успешно выявлять различные аномалии в звуках работающего оборудования. Однако они обладают рядом ограничений: требуют большого количества размеченных данных, плохо адаптируются к изменяющимся условиям эксплуатации и не обеспечивают автономного принятия решений в реальном времени.

Предложен новый подход, основанный на методах обучения с подкреплением, отличающийся высокой адаптивностью, наличием механизма обратной связи и способностью работать в условиях стохастической и динамически изменяющейся среды. Разработана блок-схема алгоритма, включающая извлечение спектральных признаков, обучение агента и принятие решений на основе оптимальной стратегии.

Проведен эксперимент на размеченных аудиофайлах из открытой базы данных Case Western Reserve University. В ходе экспериментов модель успешно идентифицировала состояния оборудования даже при наличии фоновых индустриальных шумов, что подтверждает ее пригодность для работы в реальных производственных условиях.

Благодаря низкой вычислительной нагрузке: в среднем загрузка центрального процессора составляла 36,5 % и объем потребляемой оперативной памяти 509 МБ, а также быстрому времени отклика 0,5–0,7 с, предложенный алгоритм может быть эффективно реализован на edge-устройствах и промышленных контроллерах. Это открывает возможности его интеграции в системы предиктивного обслуживания и мониторинга оборудования в реальном времени, обеспечивая своевременное выявление неисправностей, снижение внеплановых простоев и повышение общей надежности технологических процессов.

Практическая значимость работы заключается в повышении точности и скорости диагностики, что способствует снижению рисков дорогостоящих отказов и оптимизации процессов технического обслуживания. В перспективе предложенный подход может быть усовершенствован за счет интеграции с технологиями интернета вещей, повышения устойчивости к шуму, внедрения более сложных алгоритмов обучения с подкреплением, таких как Proximal Policy Optimization или Asynchronous Advantage Actor-Critic, а также испытания на реальных промышленных объектах.

Литература

1. Виноградeнко А.М., Будко Н.П. Адаптивный контроль технического состояния автономных сложных технических объектов на основе интеллектуальных технологий // T-Comm: Телекоммуникации и Транспорт. 2020. Т. 14. № 1. С. 25–35. https://doi.org/10.36724/2072-8735-2020-14-1-25-35
2. Богатырев В.А., Богатырев С.В., Богатырев А.В. Оценка готовности компьютерной системы к своевременному обслуживанию запросов при его совмещении с информационным восстановлением памяти после отказов // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 3. С. 608–617. https://doi.org/10.17586/2226-1494-2023-23-3-608-617
3. Bogatyrev V., Vinokurova M. Control and safety of operation of duplicated computer systems // Communications in Computer and Information Science. 2017. V. 700. P. 331–342. https://doi.org/10.1007/978-3-319-66836-9_28
4. Bogatyrev V.A. Exchange of duplicated computing complexes in fault-tolerant systems // Automatic Control and Computer Sciences. 2011. V. 45. N 5. P. 268–276. https://doi.org/10.3103/s014641161105004x
5. Мартюгов А.С., Ершов Е.В., Виноградова Л.Н., Варфоломеев И.А. Диагностика промышленного оборудования методом акустического контроля // Оптико- электронные приборы и устройства в системах распознавания образов и обработки изображений: Материалы XVI Международной научно-технической конференции. Курск: Юго-Западный государственный университет, 2021. С. 172–174.
6. Верзун Н.А., Колбанёв М.О., Салиева А.Р. Многоагентный ансамблевый алгоритм акустического распознавания нарушений работоспособности автономного технологического оборудования // Информационно-управляющие системы. 2025. № 3 (136). С. 14–24. https://doi.org/10.31799/1684-8853-2025-3-14-24

References

1. Vinogradenko A.M., Budko N.P. Adaptive control of technical condition of autonomous complex technical objects on the basis of intelligent technologies. Electrostatic gyroscope in spacecraft attitude reference systems. *T-Comm*, 2020, vol. 14, no. 1, pp. 25–35. (in Russian). https://doi.org/10.36724/2072-8735-2020-14-1-25-35
2. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Assessment of the readiness of a computer system for timely servicing of requests when combined with information recovery of memory after failures. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 608–617. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-3-608-617
3. Bogatyrev V., Vinokurova M. Control and safety of operation of duplicated computer systems. *Communications in Computer and Information Science*, 2017, vol. 700, pp. 331–342. https://doi.org/10.1007/978-3-319-66836-9_28
4. Bogatyrev V.A. Exchange of duplicated computing complexes in fault-tolerant systems. *Automatic Control and Computer Sciences*, 2011, vol. 45, no. 5, pp. 268–276. https://doi.org/10.3103/s014641161105004x
5. Martiugov A.S., Ershov E.V., Vinogradova L.N., Varfolomeev I.A. Diagnostics of industrial equipment using acoustic testing. *Proc. of the Optical-Electronic Devices and Instruments in Image Recognition and Processing Systems*, 2021, pp. 172–174. (in Russian)
6. Verzun N.A., Kolbanev M.O., Salieva A.R. Multi-agent ensemble algorithm for acoustic recognition of malfunctions of autonomous technological equipment. *Information and Control Systems*, 2025, no. 3, pp. 14–24. (in Russian). https://doi.org/10.31799/1684-8853-2025-3-14-24
7. Shchegolkov M.V., Zinkin S.A. Overview of the main model-free reinforcement learning approaches. *Proc. of the Energy and Automation in Modern Society*, 2024, pp. 91–95. (in Russian)

7. Щегольков М.В., Зинкин С.А. Обзор основных подходов обучения с подкреплением на основе обучения без знания модели // Энергетика и автоматизация в современном обществе: Материалы VII Всероссийской научно-практической конференции обучающихся и преподавателей. СПб: Санкт-Петербургский государственный университете промышленных технологий и дизайна, 2024. С. 91–95.
8. Ye L., Ma X., Wen C. Rotating machinery fault diagnosis method by combining time-frequency domain features and CNN knowledge transfer // Sensors. 2021. V. 21. N 24. P. 8168. https://doi.org/10.3390/s21248168
9. Shao S., McAleer S., Yan R., Baldi P. Highly accurate machine fault diagnosis using deep transfer learning // IEEE Transactions on Industrial Informatics. 2019. V. 15. N 4. P. 2446–2455. https://doi.org/10.1109/tii.2018.2864759
10. Souza R.M., Nascimento E.G.S., Miranda U.A., Silva W.J.D., Lepikson H.A.. Deep learning for diagnosis and classification of faults in industrial rotating machinery // Computers and Industrial Engineering. 2021. V. 153. P. 107060. https://doi.org/10.1016/j.cie.2020.107060
11. Lyu P., Zhang K., Yu W., Wang B., Liu C. A novel RSG-based intelligent bearing fault diagnosis method for motors in high-noise industrial environment // Advanced Engineering Informatics. 2022. V. 52. P. 101564. https://doi.org/10.1016/j.aei.2022.101564
12. Zhang J., Koppel A., Bedi A.S., Szepesvari C., Wang M., Variational policy gradient method for reinforcement learning with general utilities // arXiv. 2020. arXiv:2007.02151. https://doi.org/10.48550/arXiv.2007.02151
13. Chen D., Peng P., Huang T., Tian Y. Deep reinforcement learning with spiking Q-learning // arXiv. 2022. arXiv:2201.09754. https://doi.org/10.48550/arXiv.2201.09754
14. Верзун Н.А., Колбанёв М.О., Салиева А.Р. Анализ перспектив обучения умных автономных логистических систем на основе оптимизации функции ценности // Известия СПбГЭТУ ЛЭТИ. 2024. Т. 17. № 10. С. 28–39. https://doi.org/10.32603/2071-8985-2024-17-10-28-39
15. Tama B.A., Vania M., Lee S., Lim S. Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals // Artificial Intelligence Review. 2023. V. 56. N 5. P. 4667–4709. https://doi.org/10.1007/s10462-022-10293-3
16. Wang R., Zhan X., Bai H., Dong E., Cheng Z., Jia X. A review of fault diagnosis methods for rotating machinery using infrared thermography // Micromachines. 2022. V. 13. N 10. P. 1644. https://doi.org/10.3390/mi13101644
17. Ramaswamy A., Hüllermeier E. Deep Q-Learning: theoretical insights from an asymptotic analysis // arXiv. 2020. arXiv:2008.10870. https://doi.org/10.48550/arXiv.2008.10870
18. Hansen N., Su H., Wang X. Stabilizing Deep Q-Learning with ConvNets and vision transformers under data augmentation // Proc. of the 35th International Conference on Neural Information Processing Systems. 2021. P. 3680–3693.
19. Haq A.S., Nasrun M., Setianingsih C., Murti M.A. Speech recognition implementation using MFCC and DTW algorithm for home automation // Proc. of the International Conference on Electrical Engineering Computer Science and Informatics. 2020. V. 7. P. 78–85. https://doi.org/10.11591/eecsi.v7.2041
20. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. Bradford Books, 2018. 552 p.
21. Das O., Das D.B., Birant D. Machine learning for fault analysis in rotating machinery: A comprehensive review // Heliyon. 2023. V. 9. N 6. P. e17584. https://doi.org/10.1016/j.heliyon.2023.e17584
22. Свидетельство о государственной регистрации программы для ЭВМ № 2025619237. Акустическая система диагностики неисправностей промышленного оборудования на основе обучения с подкреплением (АСД-ОП). Номер и дата поступления заявки: 2025617991 10.04.2025. Опубликовано 14.04.2025 Бюл. № 4 / Бердникова А.А., Колбанёв М.О, Верзун Н.А., Салиева А.Р. Правообладатель: Государственное бюджетное образовательное учреждение высшего образования «Нижегородский государственный инженерно-экономический университет».
23. Moharam M.H., Hany O., Hany A., Mahmoud A., Mohamed M., Saeed S. Anomaly detection using machine learning and adopted digital twin concepts in radio environments // Scientific Reports. 2025. V. 15. P. 18352. https://doi.org/10.1038/s41598-025-02759-5
24. Purohit H.P., Tanabe R., Ichige K., Endo T., Nikaido Y., Suefusa K., Kawaguchi Y. MIMII Dataset: sound dataset for malfunctioning

8. Ye L., Ma X., Wen C. Rotating machinery fault diagnosis method by combining time-frequency domain features and CNN knowledge transfer. *Sensors*, 2021, vol. 21, no. 24, pp. 8168. https://doi.org/10.3390/s21248168
9. Shao S., McAleer S., Yan R., Baldi P. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 2019, vol. 15, no. 4, pp. 2446–2455. https://doi.org/10.1109/tii.2018.2864759
10. Souza R.M., Nascimento E.G.S., Miranda U.A., Silva W.J.D., Lepikson H.A.. Deep learning for diagnosis and classification of faults in industrial rotating machinery. *Computers and Industrial Engineering*, 2021, vol. 153, pp. 107060. https://doi.org/10.1016/j.cie.2020.107060
11. Lyu P., Zhang K., Yu W., Wang B., Liu C. A novel RSG-based intelligent bearing fault diagnosis method for motors in high-noise industrial environment. *Advanced Engineering Informatics*, 2022, vol. 52, pp. 101564. https://doi.org/10.1016/j.aei.2022.101564
12. Zhang J., Koppel A., Bedi A.S., Szepesvari C., Wang M., Variational policy gradient method for reinforcement learning with general utilities. *arXiv*, 2020, arXiv:2007.02151. https://doi.org/10.48550/arXiv.2007.02151
13. Chen D., Peng P., Huang T., Tian Y. Deep reinforcement learning with spiking Q-learning. *arXiv*, 2022, arXiv:2201.09754. https://doi.org/10.48550/arXiv.2201.09754
14. Verzun N.A., Kolbanev M.O., Salieva A.R. Analysis learning prospects of smart autonomous logistics systems based on value function optimization. *LETI Transactions on Electrical Engineering & Computer Science*, 2024, vol. 17, no. 10, pp. 28–39. (in Russian). https://doi.org/10.32603/2071-8985-2024-17-10-28-39
15. Tama B.A., Vania M., Lee S., Lim S. Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals. *Artificial Intelligence Review*, 2023, vol. 56, no. 5, pp. 4667–4709. https://doi.org/10.1007/s10462-022-10293-3
16. Wang R., Zhan X., Bai H., Dong E., Cheng Z., Jia X. A review of fault diagnosis methods for rotating machinery using infrared thermography. *Micromachines*, 2022, vol. 13, no. 10, pp. 1644. https://doi.org/10.3390/mi13101644
17. Ramaswamy A., Hüllermeier E. Deep Q-Learning: theoretical insights from an asymptotic analysis. *arXiv*, 2020, arXiv:2008.10870. https://doi.org/10.48550/arXiv.2008.10870
18. Hansen N., Su H., Wang X. Stabilizing Deep Q-Learning with ConvNets and vision transformers under data augmentation. *Proc. of the 35th International Conference on Neural Information Processing Systems*, 2021, pp. 3680–3693.
19. Haq A.S., Nasrun M., Setianingsih C., Murti M.A. Speech recognition implementation using MFCC and DTW algorithm for home automation. *Proc. of the International Conference on Electrical Engineering Computer Science and Informatics*, 2020, vol. 7, pp. 78–85. https://doi.org/10.11591/eecsi.v7.2041
20. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. Bradford Books, 2018, 552 p.
21. Das O., Das D.B., Birant D. Machine learning for fault analysis in rotating machinery: A comprehensive review. *Heliyon*, 2023, vol. 9, no. 6, pp. e17584. https://doi.org/10.1016/j.heliyon.2023.e17584
22. Berdnikova A.A., Kolbanev M.O, Verzun N.A., Salieva A.R. *Acoustic system for diagnostics of industrial equipment faults based on reinforcement learning (ASD-OP)*. Certificate of state registration of the computer program RU 2025619237. 2025. (in Russian)
23. Moharam M.H., Hany O., Hany A., Mahmoud A., Mohamed M., Saeed S. Anomaly detection using machine learning and adopted digital twin concepts in radio environments. *Scientific Reports*, 2025, vol. 15, pp. 18352. https://doi.org/10.1038/s41598-025-02759-5
24. Purohit H.P., Tanabe R., Ichige K., Endo T., Nikaido Y., Suefusa K., Kawaguchi Y. MIMII Dataset: sound dataset for malfunctioning industrial machine investigation and inspection. *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, pp. 209–213. https://doi.org/10.33682/m76f-d618
25. Koizumi Y., Kawaguchi Y., Imoto K., Nakamura T., Nikaido Y., Tanabe R., Purohit H., Suefusa K., Endo T., Yasuda M., Harada N. Description and discussion on DCASE2020 challenge Task2: unsupervised anomalous sound detection for machine condition monitoring. *arXiv*, 2020, arXiv:2006.05822. https://doi.org/10.48550/arXiv.2006.05822

industrial machine investigation and inspection // Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). 2019. P. 209–213. https://doi.org/10.33682/m76f-d618
25. Koizumi Y., Kawaguchi Y., Imoto K., Nakamura T., Nikaido Y., Tanabe R., Purohit H., Suefusa K., Endo T., Yasuda M., Harada N. Description and discussion on DCASE2020 challenge Task2: unsupervised anomalous sound detection for machine condition monitoring // arXiv. 2020. arXiv:2006.05822. https://doi.org/10.48550/arXiv.2006.05822

Авторы

Верзун Наталья Аркадьевна — кандидат технических наук, доцент, доцент, Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация; доцент, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В. И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация, SC 57208320400, http://orcid.org/0000-0002-0126-2358, Verzun.n@unecon.ru

Колбанёв Михаил Олегович — доктор технических наук, профессор, профессор, Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация; профессор, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В. И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация, SC 6506189057, http://orcid.org/0000-0003-4825-6972, mokolbanev@mail.ru

Салиева Аделина Рустамовна — аспирант, младший аналитик, Лига цифровой экономики, Москва, 127015, Российская Федерация, https://orcid.org/0009-0001-9519-5773, Rustamovna.a3@gmail.com

Authors

Natalya A. Verzun — PhD, Associate Professor, Associate Professor, Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation; Associate Professor, Saint Petersburg Electrotechnical University “LETI”, Saint Petersburg, 197376, Russian Federation, SC 57208320400, http://orcid.org/0000-0002-0126-2358, Verzun.n@unecon.ru

Mikhail O. Kolbanev — D.Sc., Full Professor, Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation; Professor, Saint Petersburg Electrotechnical University “LETI”, Saint Petersburg, 197376, Russian Federation, SC 6506189057, http://orcid.org/0000-0003-4825-6972, mokolbanev@mail.ru

Adelina R. Salieva — PhD Student, Junior Analyst, Digital Economy League, Moscow, 127015, Russian Federation, https://orcid.org/0009-0001-9519-5773, Rustamovna.a3@gmail.com

Статья поступила в редакцию 28.06.2025
Одобрена после рецензирования 08.08.2025
Принята к печати 24.09.2025

Received 28.06.2025
Approved after reviewing 08.08.2025
Accepted 24.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-971-978
УДК 51-73

Методы моделирования аномальных режимов динамических процессов на основе энергетической оценки

Владислав Константинович Казанков[1]✉, Светлана Евгеньевна Холодова[2]

[1] ООО ТОР, Санкт-Петербург, 190013, Российская Федерация
[1] Научно-технологический университет «Сириус», Краснодарский край, поселок городского типа Сириус, 354340, Российская Федерация
[2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] v.kazankov98@gmail.com✉, https://orcid.org/0000-0002-7766-7730
[2] kholodovase@yandex.ru, https://orcid.org/0000-0002-2852-4952

Аннотация

Введение. Рассматривается задача о методологии прогнозирования особых режимов динамических процессов, в частности — нелинейного эффекта, возникающего в морской среде, называемого «волнами-убийцами». Волны-убийцы — волны, возникающие в океане, как правило, внезапно, существующие короткий временной промежуток и обладающие огромным разрушительным потенциалом. Существует множество направлений в изучении данного явления, основанных на применении компьютерного моделирования и численных методов. При этом наблюдается тенденция поиска волн-убийц не только в гидродинамике, но и в других предметных областях, в которых при построении моделей исследуемых явлений и процессов применяется аппарат решения соответствующих начально-краевых задач для систем дифференциальных уравнений. Как правило, исследования направлены на поиск решения дифференциальных уравнений, на основе которых удается продемонстрировать возникновение аномально высоких волн. Следует отметить, что поиск аналитических решений для некоторых дифференциальных уравнений является крайне сложной задачей или даже не решаемой. **Метод.** Предлагается альтернативный подход, позволяющий доказать существование возможности возникновения аномалии без необходимости решения системы дифференциальных уравнений. В результате производится построение модели динамической системы, похожей на формализм теории Купмана, отличающейся учетом асимптотической скорости роста образа динамического оператора в энергетическом пространстве, на основе которого возникает упорядоченная иерархия классов динамических операторов. Предлагается определение аномалии в формализме рассматриваемого математического аппарата, при этом, феномен волны-убийцы интерпретируется как частный случай возникновения аномального явления в гидродинамической системе при достаточно высоком среднем значении волнового фона. **Основные результаты.** Разработан метод, позволяющий рассматривать эволюцию динамической системы на основе взаимодействия выделенных объемов рассматриваемой среды и их обмена энергией. В рамках предложенного подхода удается сформулировать необходимые условия возникновения аномального явления и достаточные условия отсутствия возникновения аномалий. Предлагается метод обработки временного ряда, учитывающий гипотезу о периодичности возникновения аномальных явлений. **Обсуждение.** Демонстрируется существование аномалий в магнитогидродинамических процессах, для доказательства которого проводится построение модели инверсии магнитного поля, причем решение дисперсионного уравнения осуществляется с помощью модификации численного метода Иванисова–Полищука, состоящей в комбинировании его алгоритма и метода оптимизации Adam. Полученные результаты могут быть востребованы для дальнейшего развития изучения устройства динамических систем и для выявления большего количества междисциплинарных связей, позволяющих конструктивно перенести часть результатов из одной предметной области в другую.

Ключевые слова

математическое моделирование, волны-убийцы, динамические системы, временные ряды, доказательство существования аномалий

Ссылка для цитирования: Казанков В.К., Холодова С.Е. Методы моделирования аномальных режимов динамических процессов на основе энергетической оценки // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 971–978. doi: 10.17586/2226-1494-2025-25-5-971-978

Methods of modeling anomalous modes of dynamic processes based on energy estimation

Vladislav K. Kazankov[1]✉, Svetlana E. Kholodova[2]

[1] TOR company, Saint Petersburg, 190013, Russian Federation
[1] Sirius University of Science and Technology, Sirius Federal Territory, 354340, Russian Federation
[2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] v.kazankov98@gmail.com✉, https://orcid.org/0000-0002-7766-7730
[2] kholodovase@yandex.ru, https://orcid.org/0000-0002-2852-4952

Abstract

The problem of forecasting methodology for special modes of dynamic processes the nonlinear effect that occurs in the marine environment, called “rogue waves”, is considered. Rogue waves are waves that occur in the ocean, as a rule, suddenly, exist for a short period of time and have a huge destructive potential. There are many directions in the study of this phenomenon based on the application of computer modeling and numerical methods. At the same time, there is a tendency to search for rogue waves not only in hydrodynamics, but also in other subject areas, in which, when constructing models of the phenomena and processes under study, the apparatus for solving the corresponding initial boundary value problems for systems of differential equations is used. As a rule, the authors try to find solutions to differential equations, based on which it is possible to demonstrate the occurrence of abnormally high waves. It should be noted that the search for analytical solutions for some differential equations is an extremely difficult task or even impossible to solve. An alternative approach is proposed that makes it possible to prove the existence of the possibility of an anomaly without the need to solve the corresponding system of differential equations, and a model of a dynamic system is constructed similar to the formalism of Koopman theory which takes into account the asymptotic growth rate of the image of a dynamic operator in the energy space, on the basis of which an ordered hierarchy of classes of dynamic operators arises. The definition of an anomaly in the formalism of the mathematical apparatus under consideration is proposed, while the phenomenon of a rogue wave is interpreted as a special case of the occurrence of an anomalous phenomenon in a hydrodynamic system with a sufficiently high average value of the wave background. Within the framework of the proposed approach, it is possible to formulate the necessary conditions for the occurrence of an abnormal phenomenon and sufficient conditions for the absence of anomalies. A time series processing method is proposed that considers the hypothesis of the frequency of occurrence of anomalous phenomena. The existence of anomalies in magnetohydrodynamic processes is demonstrated, which is proved by constructing a model of magnetic field inversion, and the solution of the corresponding dispersion equation is carried out using a modification of the numerical Ivanisov-Polishchuk method consisting in combining the Ivanisov-Polishchuk algorithm and the Adam optimization method. The results obtained may be in demand for further development of the study of the structure of dynamic systems and for identifying more interdisciplinary connections that allow constructive transfer of some of the results from one subject area to another.

Keywords

mathematical modeling, rogue waves, dynamical systems, time series, proof of the existence of anomalies

For citation: Kazankov V.K., Kholodova S.E. Methods of modeling anomalous modes of dynamic processes based on energy estimation. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 971–978 (in Russian). doi: 10.17586/2226-1494-2025-25-5-971-978

Введение

Динамические процессы являются неотъемлемой частью физического мира. Их математические модели широко используются как в теоретических исследованиях, так и при внедрении передовых технологий в производство. Несмотря на особенности предметных областей, классическим способом описания динамики в системе является использование аппарата дифференциальных уравнений. Существует феномен, который до недавнего времени считался особенностью исключительно гидродинамических систем, но оказалось, что первопричина его возникновения заключена в структуре динамики самой системы. Явление называется «волна-убийца» [1, 2], которое возникает, как правило, в системах с нелинейной динамикой. Известны натурные эксперименты, позволившие впервые сгенерировать волну-убийцу в бассейне при пересечении пары волн под определенными углами [3]. К отличительным чертам волн-убийц относят: огромный размер, внезапность возникновения, короткий временной промежуток существования. Феноменологически существование волн-убийц крупного размера может быть объяснено тем, что природа таких волн связана с совокупной энергией всех волн выделенной сплошной среды. Следует отметить, что появление волн-убийц возможно не только в жидких средах, описываемых классической гидродинамикой, но и в различных квантово-механических процессах, постоянных относительно запаса энергии системы. Существует два основных направления исследования волн-убийц: исследование свойств уравнений математической физики вместе со структурой их решения и прогнозирование возникновения волн-убийц. В работах [4–6] исследованы нелинейные уравнения, соответствующие моделям квантовой механики, при решении которых возникают волны-убийцы. Так частные решения уравнений Шредингера нелинейного вида дают необходимые, с точки зрения математического моделирования, структуры для построения модели волн-убийц. Также исследование решений обобщенного уравнения Хироты показало наличие волн-убийц [7]. Возникает необходимость в формировании критериев прогнозирования возникновения аномалий с учетом периодической структуры их появления. Не меньший

интерес представляет исследование взаимосвязи статистических показателей волновых амплитуд и характера взаимодействия волн друг с другом. В настоящей работе предлагается в подтверждение применимости подхода к моделированию возникновения волн-убийц рассмотреть динамические системы, не подчиняющиеся в явном виде гидродинамическим законам. При этом требуется описать динамические системы, в которых выделяются структурные компоненты, необходимые для описания аномальных режимов. В качестве инструмента исследования разработана прогностическая модель, позволяющая при помощи ретроспективного анализа предоставить возможность появления аномалий в текущей системе. С точки зрения методов математического моделирования предлагается класс функций, имеющих ряд характеристик, достаточных для описания интересующей модели.

Метод

Математическая модель. Рассмотрим сплошную среду V. Произведем ее разбиение на непересекающиеся объемы v_i и зафиксируем их. Определим для любого объема v_i линейный функционал J_t: $V \to W \subset \mathbb{R}^+ \cup \{0\}$, где W — множество всех оценок $w_i(t)$, представляющих собой энергетические оценки объемов v_i в момент времени t; $\mathbb{R}^+ \cup \{0\}$ — множество действительных неотрицательных чисел; параметр $t \in T \subset \mathbb{R}^+ \cup \{0\}$ описывает непрерывное время. Энергетические оценки объемов v_i в момент времени t определяются по формуле: $J_t(v_i) = w_i(t)$. Тогда для любого t выполняется неравенство

$$\sum_{v_i \in V} J_t(v_i) \le \sup W < \infty.$$

Примем, что каждый объем v_i может передавать энергию, а значит, и увеличивать значение энергетической оценки объема v_j, где $v_i, v_j \in V$. Такой динамической структуре формально соответствует полный граф.

Пусть в динамической системе для любого элемента $w_i(t) \in W$ существует двухпараметрическое семейство замкнутых операторов $\mathfrak{D} = \{D_t^u : W \to W\}_{u \in G}$, таких, что при $t = 0$, $D_0^u = I$ и $\mathfrak{D} \subset C^1(T)$, где G — множество всех возможных пар объемов, а u — конкретная пара объемов, между которыми происходит энергетический обмен, I — оператор тождественного отображения, C^1 — класс непрерывных функций, имеющих по крайне мере одну производную.

Зададим норму в энергетическом пространстве $H = (V, J_t)$ для оператора D_t^u при $u = ij$ формулой:

$$\|D_t^{ij}\|_H = |J_t(v_j) - J_\tau(v_i)|,$$

где τ — момент времени получения энергетической оценки объемом v_i.

Для любых $\hat{D}_{t_2}^a, \bar{D}_\tau^b \in \mathfrak{D}$ определим композицию операторов как $\hat{D}_t^a \bar{D}_\tau^b = \hat{D}_{t+\tau}^{a \circ b}$, при $\tau \le t$ и $a \circ b = u \in G$, где «$\circ$» — оператор склейки пар объемов, такой, что в результате будет получена пара, соответствующая одному из ребер полного графа. Если зафиксировать объем v_i и рассматривать эволюцию динамической системы через изменение энергетической оценки объема v_i то можно упростить запись динамического оператора:

$$D_t^{ii} = D_t.$$

Пусть $\Delta\tau$ — промежуток времени, принимаемый за условную единицу, тогда описывает динамический процесс: $(D_{n\Delta\tau}(w))^n = J_t(v_i)$, где $n \in N$ количество временных интервалов. Величину $\Delta\tau$ также можно интерпретировать как параметр дискретизации времени t где $t = n\Delta\tau$, поэтому, если $\Delta\tau \to 0$, то и $n \to \infty$. Среди операторов $D_t \in \mathfrak{D}$ возникает упорядоченная иерархия классов операторов $[D_t] = \mathfrak{L}_p$, где p — порядок класса, образованный относительно скорости роста энергетических оценок.

Пусть $2\omega \ne 0$ — условная единица измерения энергии, которую возможно зарегистрировать, тогда точной верхней гранью класса $\mathfrak{L}_0$ будет представитель класса операторов $\mathfrak{L}_1$, таких, что, если $D_t \in \mathfrak{L}_1 \cap \mathfrak{L}_0$, то $\|D_t\|_H \le \underbrace{2\omega + \ldots + 2\omega}_{n} = 2\omega n = O(n)$.

Так как за каждый условный промежуток времени $\Delta\tau$ количество энергии увеличивается на постоянное значение 2ω, то $\sup\mathfrak{L}_0 = O(n)$.

Рассмотрим $\sup\mathfrak{L}_1$. Полагая, что в момент времени t количество изменения энергии равно 2ω, для любого $D_t \in \mathfrak{L}_1 \cap \mathfrak{L}_2$ будет справедлива оценка

$$\|D_t\|_H \le 2\omega + 2\cdot 2\omega + \ldots + n\cdot 2\omega = \\ = (2\omega + \omega(n-1))n = \omega(n+1)n = O(n^2),$$

а для любого оператора $D_t \in \mathfrak{L}_2 \cap \mathfrak{L}_3$, такого, что в каждый момент времени изменение количества энергии будет в 2ω раз больше, чем в предыдущей момент времени t, верна оценка

$$\|D_t\|_H \le 2\omega + (2\omega)^2 + \ldots + (2\omega)^n = \\ = \frac{2\omega(1-(2\omega)^n)}{1-2\omega} \le Ae^{\alpha n} = O(e^{\alpha n}),$$

где $A = \dfrac{2\omega}{2\omega - 1}$; $\alpha = \ln(2\omega)$.

Для произвольного оператора D_t из класса $\mathfrak{L}_p$ при $p \ge 2$ справедлива оценка

$$\|D_t\|_H \le \sum_{k=1}^{n} 2\omega \uparrow^{p-1} k,$$

где «$\uparrow^{p-1}$» — обозначение гипероператора в нотации Кнута [8].

Определение. Аномалией в динамической системе будем называть энергетическую оценку объема v_i, такую, что амплитудный критерий $\mu(v_i) \ge 2{,}1$ [9]. Амплитудный критерий будет иметь вид:

$$\mu(v_i) = \frac{\|D_t\|_H}{\dfrac{1}{|M|}\sum_{\tau \in M}\|D_t\|_H},$$

где $\|D_t\|_H$ — отображает энергетическую оценку объема, при этом в случае конкретной физической систе-

мы аналогом энергетической оценки может являться высота волны в момент времени t; M — множество, состоящее из трети самых больших энергетических оценок до момента t.

Пусть оператор $D_t \in \mathfrak{L}_0$. Тогда верно следующее неравенство:

$$\frac{1}{|M|}\sum_{\tau\in M}\|D_t\|_H \le \frac{\|D_{t-1}\|_H + \|D_{t-2}\|_H}{2},$$

с учетом регистрации не менее двух энергетических оценок до момента t.

Тогда для значения амплитудного критерия μ справедливо неравенство

$$\mu = \frac{2\|D_t\|_H}{\|D_{t-1}\|_H + \|D_{t-2}\|_H}. \qquad (1)$$

Теорема. Любой оператор $D_t \in \mathfrak{L}_2$ задает динамический процесс, в котором могут возникать аномалии.

Доказательство. Пусть до момента времени t было получено не менее двух энергетических оценок, а $\omega = 1$, тогда на основе неравенства (1) — следствия амплитудного критерия, получим нижнюю грань его числовой характеристики в энергетическом пространстве:

$$\frac{2\|D_t\|_H}{\|D_{t-1}\|_H + \|D_{t-2}\|_H} \ge \frac{2\|D_t\|_H}{2\max(\|D_{t-1}\|_H, \|D_{t-2}\|_H)} \ge$$
$$\ge \frac{\|D_t\|_H}{\|D_{t-1}\|_H} \ge \frac{(2\omega-1)(n+1)n}{2e^{(n-1)\ln 2\omega}}\bigg|_{n=3} = \frac{4\cdot 3}{2e^{2\ln 2}} = 1{,}5.$$

Рассмотрим верхнюю оценку:

$$\frac{2\|D_t\|_H}{\|D_{t-1}\|_H + \|D_{t-2}\|_H} \le \frac{2\|D_t\|_H}{2\min(\|D_{t-1}\|_H, \|D_{t-2}\|_H)} \le \frac{\|D_t\|_H}{\|D_{t-2}\|_H} \le$$

$$\le \frac{2e^{n\ln 2\omega}}{(2\omega-1)(n-1)(n-2)}\bigg|_{n=3} = \frac{2e^{3\ln 2}}{2} = 8.$$

Так как верхняя оценка числовой характеристики амплитудного критерия больше значения 2,1, а нижняя — меньше, то значение амплитудного критерия μ может превышать величину 2,1, следовательно, в динамической системе могут возникать аномальные явления. Из доказательства теоремы следуют два следствия.

Следствие 1. Необходимое условие существования аномалии в динамической системе. Если существует возможность возникновения аномалии в динамической системе, то динамический оператор $D_t \in \mathfrak{L}_p$ где $p \ge 2$.

Следствие 2. Достаточное условие отсутствия возникновения аномальных явлений в динамической системе. Если динамический оператор $D_t \in \mathfrak{L}_p$ где $p \in \{0, 1\}$, то в динамической системе не возникает аномальных явлений.

Рассмотрим возможность возникновения волны-убийцы в качестве частного случая появления аномалии в динамической системе при достаточно высоком волновом фоне. Возникает гипотеза: «Если в динамической системе существуют аномальные явления, то они должны происходить с некоторой периодичностью», для подтверждения которой предлагается метод обработки временных рядов.

Основные результаты и обсуждения

Применение метода в гидродинамике. В работе [10] проведены вычислительные эксперименты, состоящие в построении с помощью численных методов решений уравнений гидродинамики в слое идеальной жидкости со свободной поверхностью и бесконечно глубоким дном.

Результирующая волнограмма допускает интерпретацию в виде временного ряда X_t.

Выполним оценку амплитуды каждой волны согласно формуле

$$J_t = x_{mx} - (x_{mf} + x_{ms})/2,$$

где x_{mx} — высота гребня волны; x_{mf} и x_{ms} — высоты ближайших подошв, окружающих гребень.

Временной ряд X_t преобразуем во временной ряд Y_t, который отображает изменение высоты волны с течением времени. Согласно амплитудному критерию, во временном ряде Y_t зафиксировано 25 волн-убийц. Для временного ряда Y_t определим оценку локальной равномерности $\mu(U_n)$ как среднее квадратическое отклонение, рассчитанное для множества $U_n \subset Y_t$, где n — количество элементов в множестве U_n. Временной ряд Y_t состоит из 29 103 элементов, а волны-убийцы возникали только в моменты времени

$$t = [26\,925;\ 27\,164;\ 27\,198;\ 27\,232;\ 27\,403;\ \ldots$$
$$\ldots,\ 28\,429;\ 28\,463;\ 28\,497;\ 28\,531].$$

Предполагая, что возникновение волны-убийцы — систематическое явление, обладающее некоторой периодической структурой, подберем такое n, при котором значение $\mu(U_n)$ будет минимальным во все моменты времени возникновения волны-убийцы. Решение этой задачи сводится к построению целевой функции $\rho(n)$ и ее минимизации.

Пусть целевая функция имеет вид

$$\rho(n) = \frac{1}{25}\sum(\mu_k(U_n))^2 \to \min,$$

где μ_k — оценка локальной неравномерности k-го подмножества.

Перед началом оптимизации необходимо определить область допустимых значений.

Пусть $n \in [3{,}500]$. Под частицей будем понимать реализацию метода имитации отжига для целевой функции $\rho(n)$. Наименьшее значение целевой функции $\rho(n)$ достигается при $n = 17$. Это значение встречалось чаще всего и имело наименьшее значение из всех полученных в результате запуска роя из 60 частиц. На рис. 1 представлен график $\mu(U_n)$ для $n = 17$. Синим цветом показано изменение значения $\mu(U_n)$, обозначается как «Origin». За «Trend» обозначен результат экспоненциального сглаживания «Origin» с параметром $\alpha = 0{,}02$. Красная линия отмечает момент регистрации первой

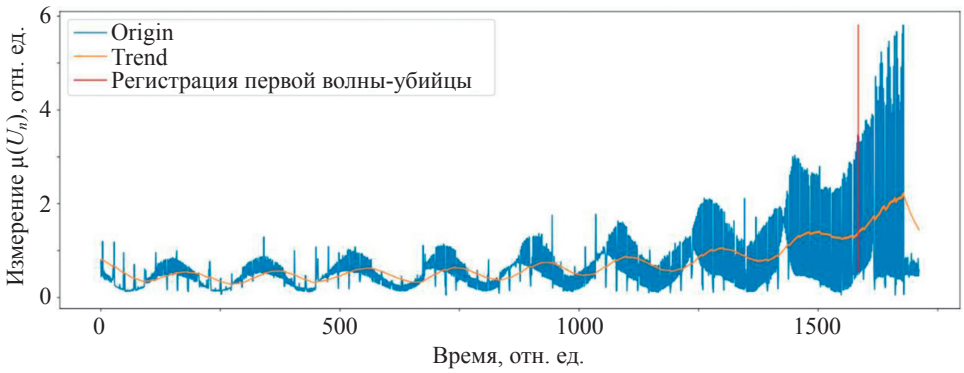

Рис. 1. Изменение μ(U_n) с течением времени при n = 17

Fig. 1. The change of μ(U_n) vs. time at n = 17

волны-убийцы, при этом она начинается со значения показателя μ(U_n) и заканчивается на максимально возможном значении μ(U_n).

На рис. 1 изображена эпоха, состоящая в явном виде из циклической составляющей в виде синусоиды и тенденции в виде экспоненты. Представленный профиль позволяет выдвинуть гипотезу о существовании некоторого периодического режима в нелинейном динамическом процессе. Применение функции локальной равномерности ρ(n) позволяет отобразить исходный временной ряд в другой, имеющий более выраженную периодическую динамику. В работе [11] проведено исследование эффекта Портевена–Ле Шателье, при котором разработана методология для обработки временных рядов с использованием гипотезы о периодичности возникновения волн-убийц. На рис. 2 представлены распределения амплитуд волн, полученные из временного ряда Y_t. По горизонтальной полуоси отображена амплитуда волн в метрах, а по вертикальной — частота возникновения. Из рисунка видно, что распределение амплитуд волн напоминает суперпозицию нормального закона с несколькими распределениями Рэлея. Благодаря использованию центральной предельной теоремы для данной системы можно сделать вывод: в динамической системе существует выраженная динамика, которая описывается законами распределения Рэлея. При этом, именно закон распределения Рэлея используют для описания распределения амплитуд волн в линейной волновой теории.

На рис. 3 изображен график распределений значений функции μ(U_{17}), где на вертикальной оси отмечены частоты значений, а на горизонтальной — относительные значения. Профиль распределения напоминает закон распределения Рэлея, из чего можно сделать вывод о том, что использование функции μ(U_n) для преобразования временного ряда Y_t позволяет перейти от описания нелинейной динамики системы к ее некоторому представлению, согласующемуся с линейной теорией волн.

Применение метода в магнитной гидродинамике. В работе [12] приведено подробное исследование системы дифференциальных уравнений в частных производных, описывающей магнитогидродинамические процессы во вращающемся слое электропроводящей жидкости переменной глубины.

Локализация динамики в точке с координатами (x_0, y_0, z_0) описывается следующим законом с точностью до аддитивной константы:

$$D_t(\mathbf{v} + \mathbf{b}) = -2\boldsymbol{\omega} \times \mathbf{v} - gz_0,$$

где **v** — вектор поля скорости; **b** — вектор магнитной индукции; **ω** — угловая скорость вращения слоя; g — величина силы тяжести.

Правая часть уравнения соответствует образу динамического оператора $D_t \in \mathfrak{Q}_2$, следовательно, согласно доказанной теореме, магнитогидродинамическая система содержит решение, описывающее аномальный режим.

В случае длинных волн малой амплитуды решение исходной краевой задачи сводится к решению одного скалярного уравнения для некоторой модифицирован-

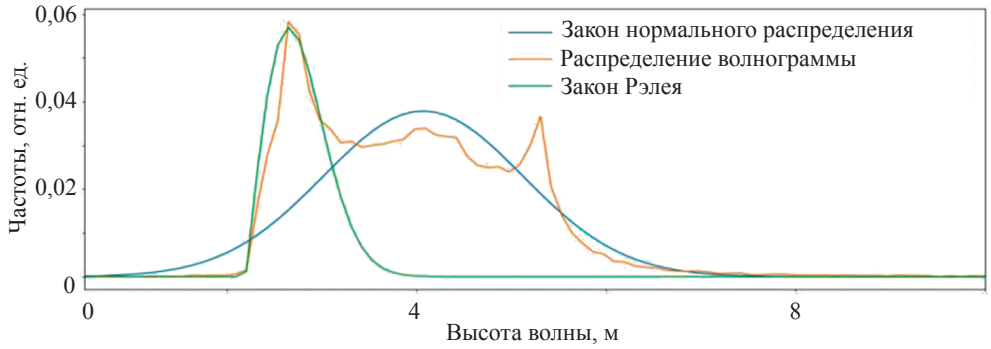

Рис. 2. Распределения амплитуд волн во временном ряду

Fig. 2. Distribution of wave amplitudes in the time series

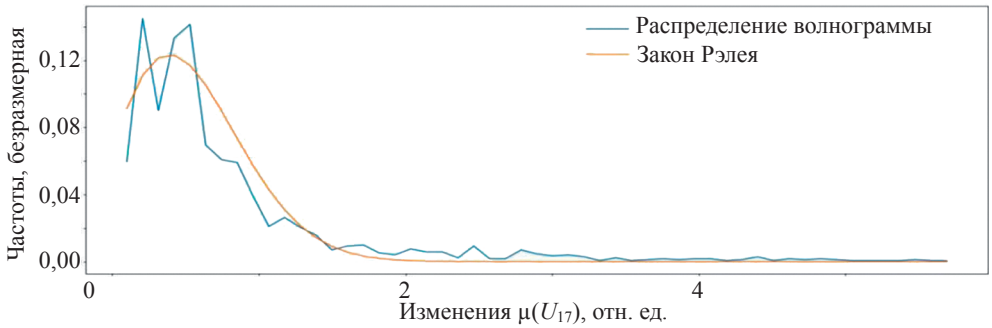

Рис. 3. Распределение функции μ(U_{17})
Fig. 3. Distribution of function μ(U_{17})

ной функции возмущения глубины слоя ξ, которая представляется в виде гармоники $\xi = \mathrm{Im}(A\exp(i(kx + ly - \sigma t))$. При этом частота σ определяется из дисперсионного уравнения 6-го порядка $P_n(\sigma) = \sum_{k=0}^{n} c_k \sigma^k$, где $c_k \in \mathbb{C}$, для решения которого предлагается использовать численный метод, являющийся результатом комбинирования алгоритмов Иванисова–Полищука [13] и Adam [14]. Численный метод состоит в использовании в качестве начального — приближения с помощью реализации алгоритма Иванисова–Полищука, и, если оно не сходится к искомому решению, то его последнее положение становится начальной точкой оптимизации целевой функции $f(\sigma) = |P_n(\sigma)|^2$, для которой применяется алгоритм Adam. В таблице представлены результаты запусков алгоритмов: итерационного метода Иванисова–Полищука (IP), алгоритма оптимизации Adam (Adam), модификации итерационного метода Иванисова–Полищука, использующей для оптимизации целевой функции алгоритм Adam (Combine). Строки таблицы *K*1 и *K*2 содержат информацию о тестировании алгоритмов на выборке полиномов, сформированной следующим образом: категория *K*1 содержит полиномы, алгебраическая кратность корней которых может соответствовать степени полинома, а *K*2 — кратность одного из корней не может быть более двух. Столбец «Total» содержит общее количество корней для каждой

Таблица. Сравнение алгоритмов
Table. Algorithm comparison

Категория полиномов	IP	Adam	Combine	Total
*K*1	109	14	149	648
*K*2	124	61	164	648

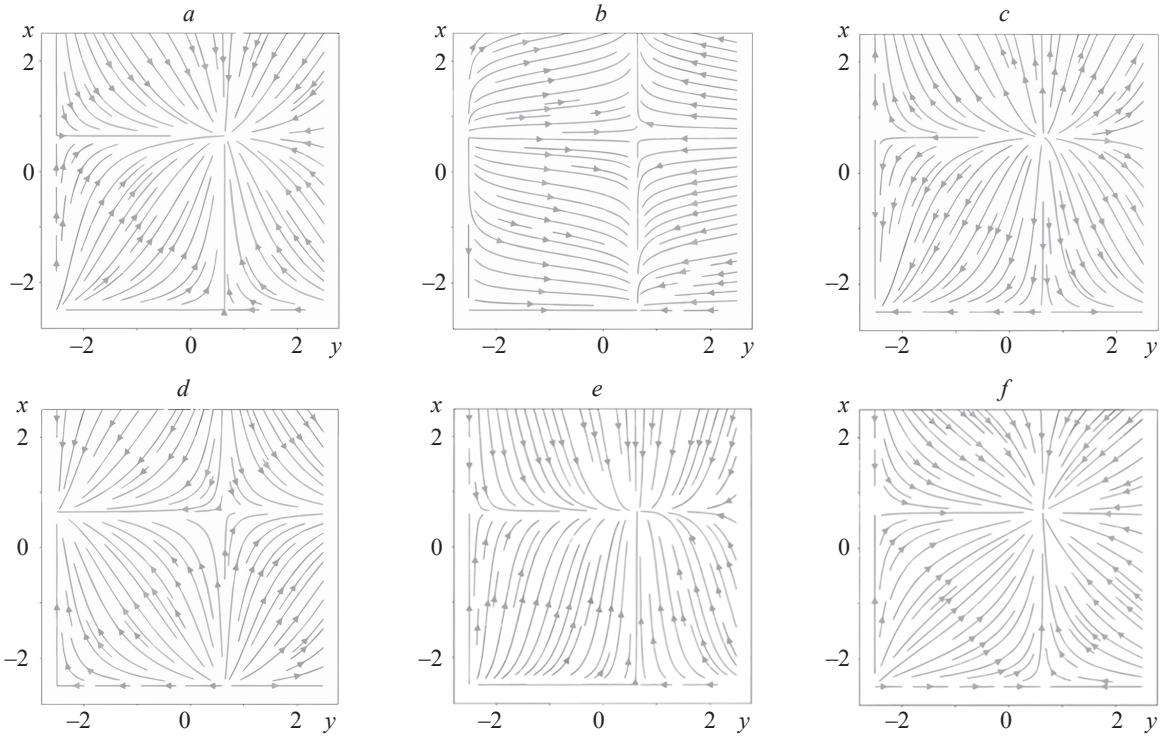

Рис. 4. Эволюция магнитного поля (*a*–*f*): начало (*a*), середина (*b*)–(*e*) и окончание (*f*) процесса
Fig. 4. Evolution of the magnetic field (*a*–*f*): at the begin (*a*), during (*b*)–(*e*) and at the end (*f*) of the process

из категории. Каждое значение в таблице описывает количество найденных коней полиномов.

Предложенная модификация алгоритмов повышает эффективность вычислений на 6,17 % по сравнению с алгоритмами Иванисова–Полищука и Adam.

В результате, найденное аналитическое решение магнитогидродинамической системы позволяет провести анализ возможных динамических режимов. На рис. 4 представлена эволюция магнитного поля в плоскости *xOy* с одинаковым временным шагом.

Видно, что процесс имеет выраженную периодическую структуру, при этом через регулярный промежуток времени происходит резкая смена направления течения силовых линий, что выглядит как аномальное явление. Такое же поведение магнитного поля описывается в теории контрастных структур. Например, в работе [15] рассматривается кинематическая модель, представляющая структуру галактического магнитного поля, в которой также возникает инверсия магнитного поля. И так, представленное исследование подтверждает гипотезу о периодическом характере возникновения аномальных явлений в динамической системе.

Заключение

Предложен метод, позволяющий рассматривать эволюцию динамической системы на основе взаимодействия выделенных объемов рассматриваемой среды и их обмена энергией. Такой подход представляет возможность доказательства существования для некоторых дифференциальных операторов особого нелинейного эффекта, возникающего в сплошной среде при достаточно интенсивном волновом фоне, называемого в случае гидромеханической системы волнами-убийцами. Особенность построения модели способствует распространению методологии обработки временных рядов на широкий класс динамических систем и прогнозированию возможности возникновения в них аномальных явлений. Дальнейшее развитие метода может состоять в применении анализа для конкретных акваторий, с целью выделения особых климатически-географических зон и временных промежутков, в которые могут возникать аномальные явления. Представленный анализ доказывает существование аномальных явлений в гидромагнитной системе и демонстрирует возникновение инверсии магнитного поля. Для нахождения соответствующих дисперсионных характеристик предложена модификация алгоритма Иванисова–Полищука. Полученные результаты могут быть востребованы для дальнейшего развития изучения устройства динамических систем и для выявления большего количества междисциплинарных связей, позволяющих конструктивно перенести часть результатов из одной предметной области в другую.

Литература

1. Куркин А.А., Пелиновский Е.Н. Волны-убийцы: факты, теория и моделирование. Москва; Берлин: Директ-Медиа, 2016. 178 с.
2. Brule S., Enoch S., Guenneau S. On the possibility of seismic rogue waves in very soft soils // arXiv. 2020. arXiv:2004.07037v1. https://doi.org/10.48550/arXiv.2004.07037
3. McAllister M.L., Draycott S., Adcock T.A.A., Taylor P.H., van den Bremer T.S. Laboratory recreation of the Draupner wave and the role of breaking in crossing seas // Journal of Fluid Mechanics. 2019. V. 860. P. 767–786. https://doi.org/10.1017/jfm.2018.886
4. Wu X.Y., Tian B., Qu Q.X., Yuan Y.Q., Du X.X. Rogue waves for a (2+1)-dimensional Gross-Pitaevskii equation with time-varying trapping potential in the Bose–Einstein condensate // Computers & Mathematics with Applications. 2020. V. 79. N 4. P. 1023–1030. https://doi.org/10.1016/j.camwa.2019.08.015
5. Yang B., Yang J. Rogue wave patterns in the nonlinear Schrödinger equation // Physica D: Nonlinear Phenomena. 2021. V. 419. P. 132850. https://doi.org/10.1016/j.physd.2021.132850
6. Li B.Q. Hybrid breather and rogue wave solution for a (2+1)-dimensional ferromagnetic spin chain system with variable coefficients // International Journal of Computer Mathematics. 2022. V. 99. N 3. P. 506–519. https://doi.org/10.1080/00207160.2021.1922678
7. Liu J.G., Zhu W.H. Multiple rogue wave solutions for (2+1)-dimensional Boussinesq equation // Chinese Journal of Physics. 2020. V. 67. P. 492–500. https://doi.org/10.1016/j.cjph.2020.08.008
8. Knuth D.E. Mathematics and computer science: coping with finiteness // Science. 1976. V. 194. N 4271. P. 1235–1242. https://doi.org/10.1126/science.194.4271.1235
9. Kharif C., Pelinovsky E., Slunyaev A. Rogue Waves in the Ocean. Springer, 2009. 216 p.
10. Шамин Р.В., Юдин А.В. Моделирование пространственно-временного распространения волн-убийц // Доклады Академии наук. 2013. Т. 448. № 5. С. 592–594. https://doi.org/10.7868/S0869565213050228
11. Kazankov V.K., Shmeleva A.G., Zaitseva E.V. Unstable plastic flow in structural materials: time series for analysis of experimental data //

References

1. Kurkin A.A., Pelinovsky E.N. *Rogue Waves: Facts, Theory and Modelling*. Moscow; Berlin, 2016, 178 p. (in Russian)
2. Brule S., Enoch S., Guenneau S. On the possibility of seismic rogue waves in very soft soils. *arXiv*, 2020, arXiv:2004.07037v1. https://doi.org/10.48550/arXiv.2004.07037
3. McAllister M.L., Draycott S., Adcock T.A.A., Taylor P.H., van den Bremer T.S. Laboratory recreation of the Draupner wave and the role of breaking in crossing seas. *Journal of Fluid Mechanics*, 2019, vol. 860, pp. 767–786. https://doi.org/10.1017/jfm.2018.886
4. Wu X.Y., Tian B., Qu Q.X., Yuan Y.Q., Du X.X. Rogue waves for a (2+1)-dimensional Gross-Pitaevskii equation with time-varying trapping potential in the Bose–Einstein condensate. *Computers & Mathematics with Applications*, 2020, vol. 79, no. 4, pp. 1023–1030. https://doi.org/10.1016/j.camwa.2019.08.015
5. Yang B., Yang J. Rogue wave patterns in the nonlinear Schrödinger equation. *Physica D: Nonlinear Phenomena*, 2021, vol. 419, pp. 132850. https://doi.org/10.1016/j.physd.2021.132850
6. Li B.Q. Hybrid breather and rogue wave solution for a (2+1)-dimensional ferromagnetic spin chain system with variable coefficients. *International Journal of Computer Mathematics*, 2022, vol. 99, no. 3, pp. 506–519. https://doi.org/10.1080/00207160.2021.1922678
7. Liu J.G., Zhu W.H. Multiple rogue wave solutions for (2+1)-dimensional Boussinesq equation. *Chinese Journal of Physics*, 2020, vol. 67, pp. 492–500. https://doi.org/10.1016/j.cjph.2020.08.008
8. Knuth D.E. Mathematics and computer science: coping with finiteness. *Science*, 1976, vol. 194, no. 4271, pp. 1235–1242. https://doi.org/10.1126/science.194.4271.1235
9. Kharif C., Pelinovsky E., Slunyaev A. *Rogue Waves in the Ocean*. Springer, 2009, 216 p.
10. Shamin R.V., Yudin A.V. Simulation of spatiotemporal spread of rogue waves. *Doklady Earth Sciences*, 2013, vol. 448, no. 2, pp. 240–242. (in Russian). https://doi.org/10.1134/S1028334X13020165
11. Kazankov V.K., Shmeleva A.G., Zaitseva E.V. Unstable plastic flow in structural materials: time series for analysis of experimental data. *Materials Physics and Mechanics*, 2022, vol. 48, no. 2, pp. 208–216. https://doi.org/10.18149/MPM.4822022_6

Materials Physics and Mechanics. 2022. V. 48. N 2. P. 208–216. https://doi.org/10.18149/MPM.4822022_6
12. Kazankov V.K., Peregudin S.I., Kholodova S.E. Mathematical modeling of geophysical processes in a layer of electrically conductive liquid of variable depth // Springer Geology. 2023. P. 331–341. https://doi.org/10.1007/978-3-031-16575-7_32
13. Ivanisov A.V., Polishchuk V.K. A method of finding the roots of polynomials which converge for any initial approximation // USSR Computational Mathematics and Mathematical Physics. 1985. V. 25. N 3. P. 1–7. https://doi.org/10.1016/0041-5553(85)90066-7
14. Kingma D.P., Ba J. Adam: a method for stochastic optimization // arXiv. 2014. arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980
15. Михайлов Е.А., Хасаева Т.Т., Тепляков И.О. Возникновение контрастных структур для галактического магнитного поля: теоретические оценки и моделирование на видеокартах // Труды Института системного программирования РАН. 2021. Т. 33. № 6. С. 253-264. https://doi.org/10.15514/ISPRAS-2021-33(6)-18

12. Kazankov V.K., Peregudin S.I., Kholodova S.E. Mathematical modeling of geophysical processes in a layer of electrically conductive liquid of variable depth. *Springer Geology*, 2023, pp. 331–341. https://doi.org/10.1007/978-3-031-16575-7_32
13. Ivanisov A.V., Polishchuk V.K. A method of finding the roots of polynomials which converge for any initial approximation. *USSR Computational Mathematics and Mathematical Physics*, 1985, vol. 25, no. 3, pp. 1–7. https://doi.org/10.1016/0041-5553(85)90066-7
14. Kingma D.P., Ba J. Adam: a method for stochastic optimization. *arXiv*, 2014, arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980
15. Mikhailov E.A., Khasaeva T.T., Teplyakov I.O. The emergence of contrast structures for galactic magnetic field: theoretical estimates and modeling on GPU. *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS)*, 2021, vol. 33, no. 6, pp. 253–264. (in Russian). https://doi.org/10.15514/ISPRAS-2021-33(6)-18

Авторы

Казанков Владислав Константинович — старший инженер-программист, ООО ТОР, Санкт-Петербург, 190013, Российская Федерация; ведущий инженер-исследователь, Научно-технологический университет «Сириус», Краснодарский край, поселок городского типа Сириус, 354340, Российская Федерация, Российская Федерация, SC 57704665700, https://orcid.org/0000-0002-7766-7730, v.kazankov98@gmail.com

Холодова Светлана Евгеньевна — доктор физико-математических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 23103155100, https://orcid.org/0000-0002-2852-4952, kholodovase@yandex.ru

Authors

Vladislav K. Kazankov — Senior Software Developer, TOR company, Saint Petersburg, 190013, Russian Federation; Leading Researching Engineer, Sirius University of Science and Technology, Sirius Federal Territory, 354340, Russian Federation, SC 57704665700, https://orcid.org/0000-0002-7766-7730, v.kazankov98@gmail.com

Svetlana E. Kholodova — D.Sc. (Physics & Mathematics), Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 23103155100, https://orcid.org/0000-0002-2852-4952, kholodovase@yandex.ru

Статья поступила в редакцию 20.07.2025
Одобрена после рецензирования 05.08.2025
Принята к печати 20.09.2025

Received 20.07.2025
Approved after reviewing 05.08.2025
Accepted 20.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-979-987
УДК 519.254

Волновая регрессия: нелинейная когнитивная эвристика

Павел Игоревич Богданов[1], Илья Алексеевич Суров[2]✉

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] pavel.bogdanov@metalab.ifmo.ru, https://orcid.org/0009-0000-1363-2451
[2] ilya.a.surov@itmo.ru✉, https://orcid.org/0000-0001-5690-7507

Аннотация

Введение. Качество регрессии определяется выбором аппроксимирующей функции, более или менее точно соответствующей процессу порождения данных. Ключевым классом таких процессов являются когнитивные процессы, часто имеющие волновой характер. Соответствующая математическая структура положена в основу метода регрессии поведенческих данных. **Метод.** Волновая регрессия строится путем обобщения коэффициентов классической линейной регрессии вещественных весов на комплекснозначные амплитуды, модули и фазы которых кодируют усиление и задержку когнитивных волн. При этом целевая величина порождается квадратом модуля суммы амплитудных влияний базисных признаков. Построенные регрессионные модели апробированы на массиве оценок успеваемости учебной группы в сравнении с линейными регрессиями с тем же числом параметров. **Основные результаты.** При большом числе базисных признаков точность волновой регрессии близка к точности линейных моделей. При уменьшении числа признаков базисных признаков ошибка линейной регрессии растет, тогда как ошибка волновой регрессии снижается. Наибольшая разница наблюдается в троичном режиме, когда целевой признак порождается парой базисных признаков. В этом случае ошибка трехпараметрической волновой регрессии на 2,5 % ниже ошибки полной линейной регрессии с 21 параметром. **Обсуждение.** Полученное преимущество обусловлено особым типом нелинейности волновой регрессии, характерной для прагматических эвристик естественного мышления. Эта нелинейность позволяет использовать смысловые корреляции признаков, не видимые другими регрессионными моделями. Представленный подход к использованию этих корреляций открывает возможности создания экономичных алгоритмов природоподобного интеллекта и анализа данных.

Ключевые слова

нелинейная регрессия, волновая логика, анализ данных, когнитивная модель, поведение, прогноз, эвристика

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 23-71-01046.

Ссылка для цитирования: Богданов П.И., Суров И.А. Волновая регрессия: нелинейная когнитивная эвристика // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 979–987. doi: 10.17586/2226-1494-2025-25-5-979-987

Wave regression: nonlinear cognitive heuristic

Pavel I. Bogdanov[1], Ilya A. Surov[2]✉

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] pavel.bogdanov@metalab.ifmo.ru, https://orcid.org/0009-0000-1363-2451
[2] ilya.a.surov@itmo.ru✉, https://orcid.org/0000-0001-5690-7507

Abstract

The quality of regression is determined by the choice of an approximation function, more or less accurately reflecting the process which generated the data. An important class of such processes is cognitive processes of largely wave nature. Here, the corresponding wave-like calculus is used in the new method of behavioral regression. We generalize classical linear regression from real weights to complex-valued amplitudes the modules and phases of which encode the amplification and delay of cognitive waves. The target feature then emerges as squared module of total amplitude

influences of all basis features. The obtained regression models are tested on the data of academic performance of the study group in comparison with linear regressions of the same number of parameters. When using all basis features, the accuracy of wave regression is close to the accuracy of linear models. With fewer basis features the quality of linear regression degrades, while the performance of wave regression improves. The largest difference is observed in triadic regime when the target feature is produced by two basis features. In this case, the error of three-parameter wave regression is 2.5 % lower than that of full linear regression with 21 parameters. This dramatic improvement is due to a special nonlinearity of wave regression, typical to pragmatic heuristics of natural thinking. This nonlinearity takes advantage of semantic correlations of features missed by classical regressions. The wave-like reduction of computational complexity opens up ways for developing more efficient and nature-like algorithms of data analysis and artificial intelligence.

Keywords
nonlinear regression, wave logic, data analysis, cognitive modeling, behavior prediction, computational complexity, heuristic

Acknowledgements
Research was funded by the Russian Science Foundation grant No. 23-71-01046.

For citation: Bogdanov P.I., Surov I.A. Wave regression: nonlinear cognitive heuristic. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 979–987 (in Russian). doi: 10.17586/2226-1494-2025-25-5-979-987

Введение

Регрессия — центральная задача анализа данных, состоящая в прогнозировании целевой величины по набору известных признаков с помощью функций определенного вида. При нелинейной связи между признаками используются степенные, гиперболические, показательные, тригонометрические, другие функции и эвристики [1, 2]. Выбор между этими вариантами обычно делается на основе априорных знаний о природе процессов, породивших рассматриваемые данные. Важнейшим таким процессом является мышление человека и других живых организмов, прямо или косвенно порождающих данные биологической природы: текстовые, звуковые, графические, поведенческие статистики социально-экономических, культурно-исторических, экологических и других процессов.

Поведение биосистем различного масштаба и сложности следует логико-вероятностной структуре, моделируемой на основе математического аппарата квантовой теории и физики волн [3–5]. При этом вещественные поведенческие вероятности в каждой ситуации-контексте порождаются комплекснозначными амплитудами «когнитивных волн» через операцию квадратного модуля [6]. Несмотря на свою универсальность [7], эта математическая структура не представлена среди методов нелинейной регрессии.

Настоящая работа устраняет этот пробел, устанавливая соответствие между *контекстами* принятия решений в квантовой когнитивистике [5, 7] и *признаками* в задаче регрессии. За основу взята волновая модель [6], связывающая вероятности двоичного выбора в трех поведенческих контекстах.

Троичная волновая регрессия

Простейшая волновая регрессия воспроизводит неотрицательную целевую величину Y по паре неотрицательных признаков X_a, X_b, которые переводятся в амплитудный вид операцией квадратного корня: $Q_a = \sqrt{X_a}$, $Q_b = \sqrt{X_b}$. Эти значения определяют стартовые амплитуды волн, распространяющихся из точек A, B в точку целевого признака Y как показано на рис. 1.

Соответствующие переходы характеризуются коэффициентами подавления/усиления волновых амплитуд r_a, r_b, а также фазовыми задержками φ_a, φ_b. В результате в точку Y приходит волна суммарной амплитуды

$$Q_y = r_a Q_a e^{i\varphi_a} + r_b Q_b e^{i\varphi_b}, \quad (1)$$

порождающая вещественную величину Y через операцию квадратного модуля аналогично тому, как мощность волны определяется ее амплитудой:

$$Y_{\text{mod}} = |Q_y|^2 = r_a^2 X_a + r_b^2 X_b + 2 r_a r_b \sqrt{X_a X_b} \cos(\varphi_b - \varphi_a). \quad (2)$$

Вещественные величины X_a, X_b, Y соответствуют наблюдаемым поведенческим данным, тогда как амплитудный вид Q_a, Q_b, Q_y представляет эти данные в когнитивной системе наблюдателя. Комплексно-амплитудное представление характерно наличием фазы «когнитивных волн», отсутствующей у вещественных чисел.

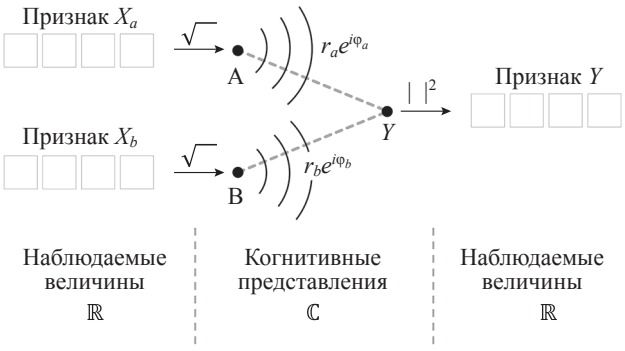

Рис. 1. Троичная волновая регрессия. Вещественные базисные признаки $X_a, X_b \geq 0$ переводятся в амплитудный вид и распространяются в виде комплекснозначных волн в когнитивном пространстве. Интерференции этих волн в точке целевого признака Y (1) порождает его значения посредством операции квадратного модуля (2)

Fig. 1. Triadic wave regression. Real-valued basis features $X_a, X_b \geq 0$ are converted to the amplitude form and propagate as complex waves in cognitive space. Interference of these waves at the target feature (1) generates its values by means of squared modulus (2)

Фазовые параметры определяют характер интерференции волновых амплитуд в мышлении субъекта: при близких значениях φ_a, φ_b амплитуды когнитивных волн Q_a, Q_b в (1) складываются конструктивно, в результате чего выходной поведенческий сигнал (2) нелинейно усиливается. В обратном случае амплитуды складываются в противофазе, подавляя поведенческий сигнал сильнее, чем это возможно в линейных моделях [6]. Поскольку суперпозиция амплитуд в выражении (1) при этом остается линейной, волновое представление, по сути, есть метод линеаризации поведенческих закономерностей, характерный для естественного мышления.

Ошибкой регрессии признака Y по признакам X_a, X_b является относительная норма разности модельных (2) и фактических данных

$$L = \frac{|Y - Y_{\text{mod}}|}{|Y|}, \tag{3}$$

минимизация которой на тренировочной выборке данных определяет величины r_a, r_b, а также разность фаз $\varphi_b - \varphi_a$. Знание этих параметров позволяет пользоваться формулой (2) для прогнозирования величины Y по величинам X_a, X_b [8, 9].

Модификации. Для практических применений суперпозицию (1) целесообразно упростить следующим образом. Во-первых, квадратный модуль в (2) не меняется при умножении амплитуды Q_y на произвольный фазовый фактор $e^{i\varphi}$. В этой связи для волновой регрессии информативна лишь разность фаз $\varphi_b - \varphi_a$, в силу чего одну из них можно приравнять нулю.

Кроме того, практика показывает, что наличие в суперпозиции (1) двух разных амплитуд перехода r_a, r_b, во многих случаях избыточно. Совместно с обнулением одной из фаз, приравнивание этих амплитуд приводит соотношение (2) к виду:

$$Y_{\text{mod}} = |rQ_a + e^{i\varphi} rQ_b|^2. \tag{4a}$$

Вещественным аналогом (4a) с тем же числом свободных параметров является линейная комбинация признаков X_a, X_b с вещественными весами w_a, w_b

$$Y_{\text{mod}} = w_a X_a + w_b X_b. \tag{4b}$$

Также оправдано упрощение, приводящее модели (4) к однопараметрическим формам

$$Y_{\text{mod}} = |Q_a + e^{i\varphi} Q_b|^2, \tag{5a}$$

$$Y_{\text{mod}} = w_a X_a + X_b, \tag{5b}$$

$$Y_{\text{mod}} = w_a X_a. \tag{5c}$$

Поскольку неотрицательные данные не могут быть центрированы на нуле, к моделям (4), (5) целесообразно добавить произвольную константу — свободный член (СЧ). В порядке возрастания числа параметров, полученные таким образом модели выстраиваются следующим образом:
один параметр: СЧ, модели (5);
два параметра: модели (5) + СЧ, модели (4);
три параметра: модели (4) + СЧ.

Свойства этих 11 моделей исследовались следующим образом.

Испытания. В качестве тестовых данных взят журнал успеваемости учебной группы из 22 человек за двухлетний период. После фильтрации пропусков и нечисловых элементов, итоговый массив данных содержит по 144 оценки каждого из 22 учащихся, нормированные на интервал от нуля до единицы. Признакам X_a, X_b, Y на схеме (1) соответствуют оценки любой тройки учащихся, один из которых является целевым, а два оставшихся — базисными. При этом для каждого целевого признака имеется $21 \times 20/2 = 210$ различных базисных пар, для каждой из которых строились перечисленные регрессии. Для тренировки каждой такой моделей из трех соответствующих строк массива многократно выбирались 75 % случайных столбцов, тогда как тестовые 25 % использовались для вычисления ошибки прогноза (3). Минимизация функции (3) выполнялась методом многомерной оптимизации L-BFGS-B [10].

Точками на рис. 2, *a* показаны значения прогнозной (тестовой) ошибки целевого признака 3 (учащийся группы под порядковым номером 3) с помощью однопараметрических моделей (5a), (5b) для всех 210 базисных пар.

Эти пары отсортированы по убыванию регрессионной информативности, т. е. по возрастанию ошибки (3) на тренировочной части данных. Линиями показано усреднение таких сортировок по всем 22 целевым признакам. Ошибка (3) волновых моделей (5a) в среднем на 7 % ниже ошибки линейных моделей (5b). При этом информативные базисные пары для линейной модели являются информативными для волновой модели и наоборот. Показанная на графике рис. 2, *b* корреляция этих ошибок составляет 78 %.

На практике из всех базисных пар целесообразно использовать те, которые показывают наилучшую прогнозную точность для целевого признака. Такие пары одновременно имеют и наилучшую тренировочную точность, о чем свидетельствует достаточно строгая монотонность линий на рис. 2, *a*. Это свойство гарантирует, что базисная пара с наименьшей ошибкой на тренировочных данных целевого признака (крайние левые точки на рис. 2, *a*) окажется наилучшей или почти наилучшей в прогнозном режиме как для линейной, так и для волновой регрессий этого признака. После усреднения по всем 22 признакам, полученные таким образом точности наилучшего прогноза для 11 вышеуказанных моделей, сгруппированных по числу параметров, представлены на рис. 3.

Среди линейных моделей наилучший результат показывают классические линейные регрессии. Как и ожидалось, ошибка этих моделей монотонно убывает с ростом числа параметров: от 0,33 у регрессии константой (нижний график) до полной линейной двухпараметрической регрессии (0,303, верхний график). Ошибка волновых регрессий с тем же числом параметров во всех случаях меньше. Наибольшая разница имеет место в однопараметрическом случае, когда волновая модель (5a) показывает ошибку $0{,}308 \pm 0{,}004$.

Среди двухпараметрических моделей линейные СЧ + $w_a X_a$ и $w_a X_a + w_b X_b$ показывают близкие ошибки

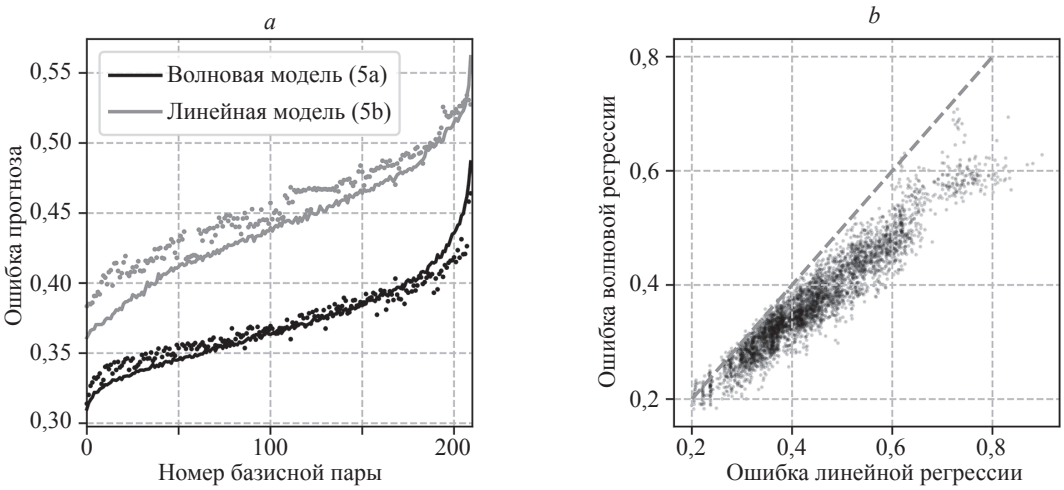

Рис. 2. Тестирование линейной и волновой однопараметрических регрессий (5a), (5b). Ошибки прогнозирования целевого признака 3 на основе всех 210 возможных пар базисных признаков (*a*); корреляция прогнозных ошибок для всех целевых признаков (*b*). Каждая точка соответствует некоторой паре базисных признаков

Fig. 2. Testing of single-parameter linear and wave regressions (5a), (5b). Prognostic errors of target feature 3 for all 210 possible basis pairs (*a*); correlation of prognostic errors for all target features (*b*). Each point stands for some pair of basis features

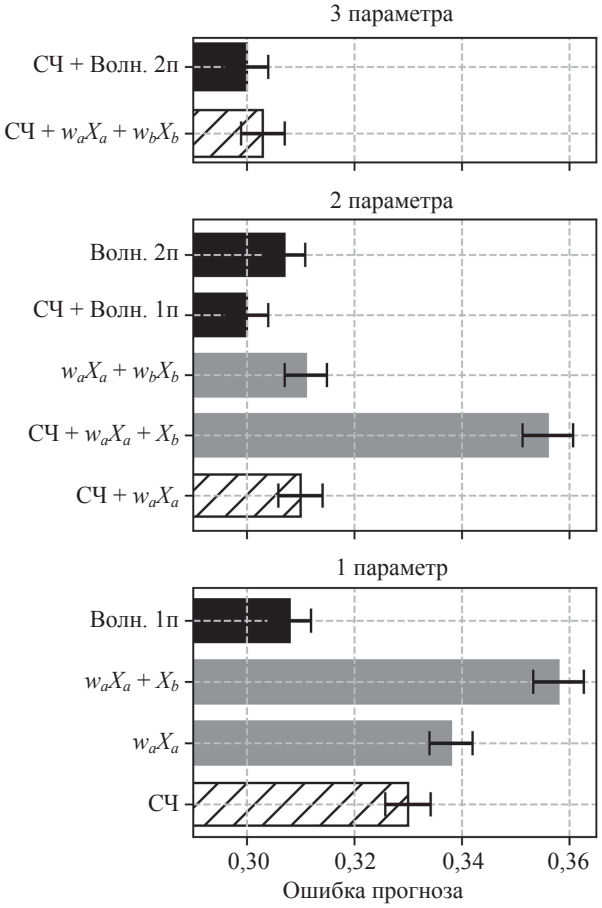

Рис. 3. Средние ошибки регрессионных моделей с одним, двумя и тремя параметрами.

Волновые модели показаны черным цветом, классические линейные регрессии — штриховкой.
Волн. 1п и волн. 2п — одно- и двухпараметрические модели; СЧ — свободный член; $w_aX_a + w_bX_b$ — линейная комбинация признаков с вещественными весами (4b); $w_aX_a + X_b$ и w_aX_a — однопараметрические модели (5b) и (5c)

Fig. 3. Average errors of regression models with one, two and three parameters. Wave models are shown in black, classical linear regressions are hatched

порядка 0,31. Волновой аналог последней (4a) дает ошибку 0,307. Если второй параметр волновой модели использовать в качестве СЧ, то ошибка снижается до 0,300 ± 0,004. Аналогичный ход для линейной модели (СЧ + $w_a X_a + X_b$) дает значительно худший результат (0,356). Точности трехпараметрических моделей отличаются незначительно (0,003) с преимуществом волнового варианта.

Многопризнаковый режим

Модель. Представленный метод обобщается на произвольное число $K \geq 2$ базисных признаков. При этом целевой признак в общем случае аппроксимируется многокомпонентным аналогом суперпозиции (1), (2). Упрощение, аналогичное троичной модели (4a), приводит это выражение к виду:

$$Y_{\text{mod}} = r^2 \left| \sum_{k=1}^{K} e^{i\varphi_k} Q_k \right|^2, \ \varphi_1 = 0, \qquad (6a)$$

в котором число регрессионных параметров равно числу базисных признаков K.

Вещественным аналогом (6) является классическая линейная регрессия

$$Y_{\text{mod}} = \sum_{k=1}^{K} w_k X_k \qquad (6b)$$

с тем же числом параметров.

Испытания. Модели (6a) и (6b) испытаны на тех же самых данных (раздел «Троичная волновая регрессия») об успеваемости учебной группы для трех разных предметов. В предельном случае единственный целевой признак моделируется на основе $K = 21$ оставшихся, причем целевым может быть любой из общего набора 22 признаков.

Полученная методом перекрестной проверки прогнозная точность, усредненная по 22 целевым признакам, представлена на гистограммах (рис. 4) первыми двумя столбцами. Для первого предмета чуть лучше сработала линейная регрессия (6b), тогда как для предметов 2 и 3 предпочтительной оказалась волновая (6a). По трем рассмотренным учебным дисциплинам, среднее преимущество волновой регрессии по сравнению с линейной составило 0,7 %.

Точности однопараметрических регрессий по лучшим базисным парам (раздел «Троичная волновая регрессия») показаны в последних двух столбцах гистограмм на рис. 4. Наибольшую ошибку во всех случаях показывает линейный вариант (5b). Волновой вариант (5a), напротив, во всех случаях дает наименьшую ошибку: на 4,4 % ниже однопараметрической линейной и на 2,5% ниже полных регрессий (6a), (6b).

Характерно поведение тренировочных ошибок, показанных на рис. 4 красным. У полных регрессий она на 4–7 % ниже прогнозной. У обеих однопараметрических моделей, напротив, тренировочная и прогнозная ошибки практически совпадают, что указывает на отсутствие переобучения.

Произвольная часть известных признаков. Полные и однопараметрические модели также апробированы в режиме, когда из $N = 22$ признаков известна произвольная часть $1 \leq K \leq 21$, тогда как остальные являются целевыми. В силу большого числа сочетаний из N по K, для каждого K рассматривались не более 200 случайных разбиений N признаков на K известных и $N - K$ неизвестных. Полные и однопараметрические регрессии каждого из неизвестных признаков строились как описано выше (разделы «Троичная волновая регрессия» и «Многопризнаковый режим»). Полученные значения в зависимости от K представлены для предмета 2 на рис. 5, *a*.

Графики на рис. 5, *a* соответствуют обозначениям на рис. 4: серый и черный пунктиры соответствуют полной регрессии в линейном (6b) и волновом (6a) вариантах. Из этих моделей линейная предпочтительнее

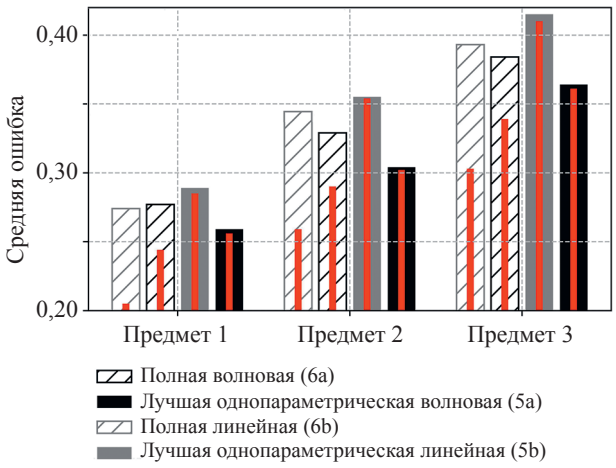

Рис. 4. Средние ошибки полных регрессий (6a), (6b) по $K = 21$ базисным признакам и однопараметрических регрессий по лучшим базисным парам для трех разных предметов. Широкими столбцами показаны прогнозные ошибки (3) согласно легенде. Красным цветом — тренировочные ошибки

Fig. 4. Average errors of full regressions (6a), (6b) with $K = 21$ basis features and single-parameter regressions with best basis pairs for three different disciplines. Wide bars show prognostic errors (3) as indicated in legend. Corresponding training errors are shown in red

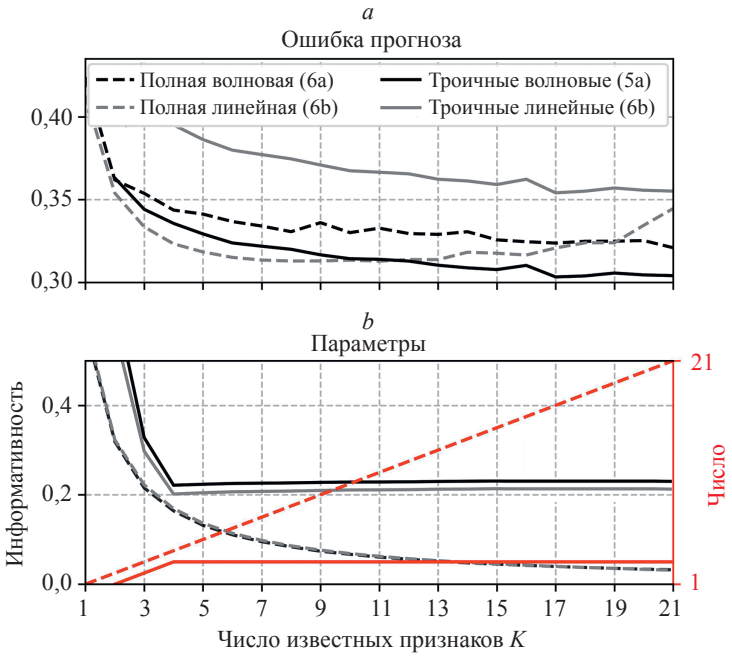

Рис. 5. Средние ошибки прогноза *N*–*K* целевых признаков с помощью полных (6a), (6b) и лучших троичных (5a), (5b) регрессий в зависимости от числа известных признаков *K* (*a*). Число и средняя информативность параметров моделей (*b*) для полных (красная штриховка) и троичных (красная линия) регрессий

Fig. 5. Average prognostic errors for *N* – *K* target features by full and best triadic regressions for different numbers *K* of known features (*a*). Number and average informativity of model parameters (*b*)

волновой всюду кроме $K > 19$, где у линейной модели имеет место переобучение, отсутствующее у волнового варианта.

Сплошные линии соответствуют линейной и волновой однопараметрическим регрессиям (5a), (5b) по лучшим базисным парам. В отличие от полных моделей, эти графики начинаются с минимального числа базисных признаков $K = 2$ (рис. 1). Как и на рис. 4, линейная однопараметрическая регрессия показывает наибольшую ошибку при всех *K*. Волновой вариант, напротив, конкурирует по точности с полной линейной регрессией, превосходя ее при большом числе известных признаков ($K \geq 12$). Крайние правые точки графиков соответствуют средней гистограмме на рис. 4.

Эти соотношения ошибок имеют место при радикальном отличии числа параметров моделей *M*, показанных красным на рис. 5, *b*.

Для полных моделей (6a), (6b) пунктир показывает число известных признаков *K*. Для однопараметрических моделей *M* складывается из одного параметра и необходимых номеров базисных признаков, лучших для данного целевого признака. При $K \geq 4$ это дает константу $M \equiv 3$.

С помощью этих значений можно вычислить среднюю информативность параметров как долю объясненной дисперсии целевых признаков 1 – ошибка (3), отнесенную к числу параметров модели *M*. Полученные таким образом величины для каждой из моделей представлены на рис. 5, *b* серыми и черными графиками. В крайних правых точках информативность параметров однопараметрических моделей примерно в 7,5 раз выше информативности параметров полных регрессий.

Обсуждение

Полученное преимущество волновых регрессий над линейными (2,5–7 %) примечательно не своей величиной, которую нетрудно превзойти более сложными методами классической регрессии. Важно то, что этот выигрыш получен переходом от вещественной «логики частиц» к комплекснозначной «логике волн» [6] в рамках одной и той же регрессионной структуры (4), (5) без увеличения числа параметров. Аналогичного эффекта можно ожидать для более сложных моделей регрессии и других алгоритмов анализа данных. В этой связи отметим следующие свойства разработанного подхода.

Троичные корреляции. При выборе регрессионной модели в первую очередь обычно рассматривается линейная функция вида (6b), учитывающая среднее значение и попарные корреляции целевого признака с данными; после этого при необходимости добавляются аналогичные слагаемые высших степеней $w_k X_k^n$, описывающие нелинейные связи целевого признака *Y* с базисными признаками X_k. Представленные результаты расширяют представления об этой последовательности, что особенно важно при наличии жестких ограничений по числу параметров и вычислительных ресурсов.

В частности, единственный доступный параметр более эффективно использовать как фазу в составе нелинейной функции (5a), чем как коэффициент в составе линейных структур (5b), (5c) или в виде константы[1]. В отличие от линейных слагаемых вида $w_k X_k^n$,

[1] Аналогичный эффект имеет место в задачах обработки изображений [11].

интерференционное соотношение (5а) описывает не попарные, а троичные корреляции, связывающие пару признаков X_a, X_b по отношению к целевому признаку Y в когнитивной системе наблюдателя. В классической регрессионной парадигме такие структуры ничем не выделяются их множества других нелинейных функций, в смеси с которыми их прогнозные качества теряются среди других компонент.

Наряду с троичными моделями имеют место множество других регрессионных структур, более сложных как по числу признаков, так и по степени нелинейности [12, 13]. Предварительные тесты показали, что обобщение выражений (1)–(5) на четыре и пять признаков ведет к сокращению преимущества волновых моделей над их линейными аналогами. Этот результат подтверждает особый статус троичных структур в естественном мышлении [14] как одну из базовых эвристик нелинейного анализа данных. В этом качестве волновые модели существенно расширяют набор средств нелинейной регрессии [15–17].

Вычислительная экономия. Большое значение имеет экономность троичной волновой регрессии, число параметров которой не зависит от числа известных признаков (рис. 5, *b*). В силу соответствующего сокращения вычислительных затрат этот выигрыш оправдал бы и некоторое снижение точности — как это имеет место для вещественных однопараметрических моделей (рис. 4, рис. 5, *a*). Волновая модификация модели, напротив, при том же выигрыше по параметрам показывает и выигрыш в точности, что свидетельствует о перспективности использованной логико-математической структуры.

Ресурсная экономичность разработанного подхода наиболее сильно проявляется при переменном разделении признаков на известные и неизвестные, которое для реальных поведенческих систем носит ситуативно-временный характер: сегодня нужно прогнозировать некоторые величины $X_1 \ldots X_3$ по признакам $X_4 \ldots X_9$, тогда как завтра значения $X_1 \ldots X_3$ могут быть известны, а целевым станет признак X_7. Комбинаторная многочисленность таких вариантов затрудняет их охват в рамках общей предобученной модели; поэтому обычное решение этой задачи состояло бы в тренировке специализированной регрессии для каждого нового разбиения. Поскольку при изменении состава известных признаков пользоваться результатам других регрессий обычно невозможно, такая тренировка выполнялась бы каждый раз с нуля. При этом в логике «чем больше данных, тем лучше» [18] использовались бы все известные признаки.

Представленный подход решает эту проблему иначе. При этом требуется лишь однократная тренировка всевозможных троичных регрессий, число которых $N(N-1)(N-2)/2$ пропорциональна третьей степени общего числа признаков N. После этого регрессия любого набора целевых признаков по любой части остальных признаков требует лишь соответствующей выборки наилучших троек из этого банка. Вычислительное преимущество этого метода при $N = 22$ можно оценить на примере рис. 5. При одинаковых средствах и методах вычислений, построение обеих полных регрессий (пунктирные графики) заняло 2 ч 47 мин; на той же вычислительной мощности, построение банка обеих троичных регрессий (рис. 2) заняло 1 мин 12 с, после чего выборка оптимальных троек (сплошные графики) заняла 6 секунд. Этот более чем 100-кратный выигрыш указывает на перспективность представленного подхода к поиску регрессионных моделей оптимальной сложности [19–21], что особенно актуально в свете энергозатратности современных моделей машинного обучения [22]. В этом отношении представленная модель соответствует качествам естественного мышления, для которого характерно использование простых, гибких и достаточно эффективных эвристик при скудности доступных данных [18, 23].

Границы применимости. В отличие от выигрыша по вычислительным ресурсам, волновое преимущество в точности имеет место не всегда. Например, такое преимущество отсутствует при моделировании средней успеваемости учебной группы, когда признаками являются средние оценки группы по различным предметам. Строгих критериев такого преимущества в настоящее время не выработано; ясно лишь, что волновая модель предпочтительнее линейной при наличии в регрессионных тройках субъективно-смысловых соотношений, которые кодируются фазовыми параметрами когнитивных волн [6, 24]. Когда признаками являются оценки отдельных людей, то такие соотношения обусловлены их межличностными отношениями, проявляющимися в учебной успеваемости. При работе со средними, напротив, фазы могли бы кодировать осмысленное отношение к различным предметам учебной группы как целого, чего в рассмотренных данных не наблюдается.

Обобщение. Возможно обобщение представленного подхода с волновых на квантово-подобные логико-математические структуры, описывающие еще более фундаментальные принципы порождения и восприятия информации благодаря специфике квантовых форматов кодирования [25, 26]. Такое обобщение позволило бы использовать инструментарий квантовой когнитивистики [5, 7] в целях регрессии и других задач анализа данных. Волновая логика, однако, предпочтительна сравнительно простой реализацией в оптике и голографии [27–30]. Представленный подход открывает новые возможности как в этом направлении, так и в рамках обычных вычислительных средств.

Заключение

В работе представлен метод нелинейной регрессии на основе алгебры волновых процессов порождения и когнитивного моделирования поведенческих данных. Полученные результаты особенно значимы потому, что модели машинного обучения, по сути, представляют собой большие регрессии, приспособленные к работе с текстовыми, графическими и другими типами данных. В этой связи достигнутое сокращение числа регрессионных параметров открывает новые подходы к проблеме вычислительно-энергетической емкости искусственного интеллекта.

Литература

1. Полежаев В.Д., Полежаева Л.Н. Нелинейные модели парной регрессии в курсе эконометрики // Современные проблемы науки и образования. 2018. № 4. С. 73
2. Колентеев Н.Я., Гончарова О.А., Гончаров В.С. Парная нелинейная регрессия и корреляция // Материалы VIII-ой Международной межвузовской научно-методической конференции. СПб.: Военная академия материально-технического обеспечения имени генерала армии А.В. Хрулёва, 2022. С. 304–312.
3. Khrennikov A. Quantum-like modeling of cognition // Frontiers in Physics. 2015. V. 3. P. 77. https://doi.org/10.3389/fphy.2015.00077
4. Asano M., Basieva I., Khrennikov A., Ohya M., Tanaka Y., Yamato I. Quantum information biology: from information interpretation of quantum mechanics to applications in molecular biology and cognitive psychology // Foundations of Physics. 2015. V. 45. N 10. P. 1362–1378. https://doi.org/10.1007/s10701-015-9929-y
5. Суров И.А., Алоджанц А.П. Модели принятия решений в квантовой когнитивистике. СПб: Университет ИТМО, 2018. 63 с.
6. Суров И.А. Логика множеств и логика волн в когнитивно-поведенческом моделировании // Информационные и математические технологии в науке и управлении. 2023. № 4(32). С. 51–66. https://doi.org/10.25729/ESI.2023.32.4.005
7. Khrennikov A.Y. Ubiquitous Quantum Structure: From Psychology to Finance. Springer, 2010. 216 p. https://doi.org/10.1007/978-3-642-05101-2
8. Surov I.A., Pilkevich S.V., Alodjants, A.P., Khmelevsky S.V. Quantum phase stability in human cognition // Frontiers in Psychology. 2019. V. 10. P. 929. https://doi.org/10.3389/fpsyg.2019.00929
9. Shan Z.H. Brainwave phase stability: predictive modeling of irrational decision // Frontiers in Psychology. 2022. V. 13. P. 617051. https://doi.org/10.3389/fpsyg.2022.617051
10. Head J.D., Zerner M.C. A Broyden—Fletcher—Goldfarb—Shanno optimization procedure for molecular geometries // Chemical Physics Letters. 1985. V. 122. N 3. P. 264–270. https://doi.org/10.1016/0009-2614(85)80574-1
11. Oppenheim A.V., Lim J.S. The importance of phase in signals // Proceedings of the IEEE. 1981. V. 69. N 5. P. 529–541. https://doi.org/10.1109/PROC.1981.12022
12. Sorkin R.D. Quantum mechanics as quantum measure theory // Modern Physics Letters A. 1994. V. 9. N 33. P. 3119–3127. https://doi.org/10.1142/s021773239400294x
13. Базилевский М.П. Критерии нелинейности квазилинейных регрессионных моделей // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 4 (23). С. 185–195. https://doi.org/10.26102/2310-6018/2018.23.4.015
14. Surov I.A. Quantum cognitive triad: semantic geometry of context representation // Foundations of Science. 2021. V. 26. N 4. P. 947–975. https://doi.org/10.1007/s10699-020-09712-x
15. Рудой Г.И. Индуктивное порождение суперпозиций в задачах нелинейной регрессии // Машинное обучение и анализ данных. 2011. Т. 1. № 2. С. 183–197.
16. Сологуб Р.А. Методы трансформации моделей в задачах нелинейной регрессии // Машинное обучение и анализ данных. 2015. Т. 1. № 14. С. 1961–1976.
17. Шестопал О.В., Черноиван Д.Н., Середина П.Б. Робастные методы построения и улучшения многомерной линейной и нелинейной регрессий // T-Comm: Телекоммуникации и транспорт. 2019. Т. 13. № 2. С. 46–51. https://doi.org/10.24411/2072-8735-2018-10235
18. Marsh B., Todd P.M., Gigerenzer G. Cognitive heuristics // Reasoning the Fast and Frugal Way. 2003. P. 273–288. https://doi.org/10.1017/cbo9780511818714.010
19. Стрижов В.В. Поиск модели оптимальной сложности в задачах нелинейной регрессии // Математические методы распознавания образов. 2005. Т. 12. № 1. С. 206–209.
20. Сологуб Р.А. Алгоритмы порождения нелинейных регрессионных моделей // Информационные технологии. 2013. № 5. С. 8–12.
21. Ширяев В.Д., Шагилова Е.В., Беляков М.Т. О выборе формы нелинейной регрессионной модели // Молодежь и системная модернизация страны: Сборник научных статей 8-й Международной научной конференции студентов и молодых ученых. В 4-х томах. Курск, ЗАО «Университетская книга», 2024. С. 137–142.

References

1. Polezhaev V.D., Polezhaeva L.N. Nonlinear paired regression models in the econometrics course. *Modern problems of science and education*, 2018, no. 4, pp. 73. (in Russian)
2. Kolenteev N.Ia., Goncharova O.A., Goncharov V.S. Paired nonlinear regression and correlation. *Proc. of the Innovative Technologies in Higher Education Pedagogy*, 2022, pp. 304–312. (in Russian)
3. Khrennikov A. Quantum-like modeling of cognition. *Frontiers in Physics*, 2015, vol. 3, pp. 77. https://doi.org/10.3389/fphy.2015.00077
4. Asano M., Basieva I., Khrennikov A., Ohya M., Tanaka Y., Yamato I. Quantum information biology: from information interpretation of quantum mechanics to applications in molecular biology and cognitive psychology. *Foundations of Physics*, 2015, vol. 45, no. 10, pp. 1362–1378. https://doi.org/10.1007/s10701-015-9929-y
5. Surov I.A., Alodzhantc A.P. *Decision Models in Quantum Cognitive Science*. St. Petersburg, ITMO University, 2018. 63 p. (in Russian)
6. Surov I.A. Logic of sets and logic of waves in cognitive-behavioral modeling. *Information and Mathematical Technologies in Science and Management*, 2023, no. 4(32), pp. 51–66. https://doi.org/10.25729/ESI.2023.32.4.005 (in Russian)
7. Khrennikov A.Y. *Ubiquitous Quantum Structure: From Psychology to Finance*. Springer, 2010, 216 p. https://doi.org/10.1007/978-3-642-05101-2
8. Surov I.A., Pilkevich S.V., Alodjants, A.P., Khmelevsky S.V. Quantum phase stability in human cognition. *Frontiers in Psychology*, 2019, vol. 10, pp. 929. https://doi.org/10.3389/fpsyg.2019.00929
9. Shan Z.H. Brainwave phase stability: predictive modeling of irrational decision. *Frontiers in Psychology*, 2022, vol. 13, pp. 617051. https://doi.org/10.3389/fpsyg.2022.617051
10. Head J.D., Zerner M.C. A Broyden—Fletcher—Goldfarb—Shanno optimization procedure for molecular geometries. *Chemical Physics Letters*, 1985, vol. 122, no. 3, pp. 264–270. https://doi.org/10.1016/0009-2614(85)80574-1
11. Oppenheim A.V., Lim J.S. The importance of phase in signals. *Proceedings of the IEEE*, 1981, vol. 69, no. 5, pp. 529–541. https://doi.org/10.1109/PROC.1981.12022
12. Sorkin R.D. Quantum mechanics as quantum measure theory. *Modern Physics Letters A*, 1994, vol. 9, no. 33, pp. 3119–3127. https://doi.org/10.1142/s021773239400294x
13. Bazilevskiy M.P. Nonlinear criteria of quasilinear regression models. *Modeling, Optimization and Information Technology*, 2018, vol. 6, no. 4 (23), pp. 185–195. (in Russian). https://doi.org/10.26102/2310-6018/2018.23.4.015
14. Surov I.A. Quantum cognitive triad: semantic geometry of context representation. *Foundations of Science*, 2021, vol. 26, no. 4, pp. 947–975. https://doi.org/10.1007/s10699-020-09712-x
15. Rudoi G.I. Inductive generation of superpositions in nonlinear regression problems. *Mashinnoe obuchenie i analiz dannyh*, 2011, vol. 1, no. 2, pp. 183–197. (in Russian)
16. Sologub R.A. Methods of the nonlinear regression model transformation. *Mashinnoe obuchenie i analiz dannyh*, 2015, vol. 1, no. 14, pp. 1961–1976. (in Russian)
17. Shestopal O.V., Chernoivan D.N., Seredina P.B. Robust methods of building and improving multidimensional linear and nonlinear regressions. *T-Comm*, 2019, vol. 13, no. 2, pp. 46–51. (in Russian). https://doi.org/10.24411/2072-8735-2018-10235
18. Marsh B., Todd P.M., Gigerenzer G. Cognitive heuristics. *Reasoning the Fast and Frugal Way*, 2003, pp. 273–288. https://doi.org/10.1017/cbo9780511818714.010
19. Strizhov V.V. Search for a model of optimal complexity in nonlinear regression problems. *MMRO*, 2005, vol. 12, no. 1, pp. 206–209. (in Russian)
20. Sologub R.A. Nonlinear model generation algorithms. *Information Technologies*, 2013, no. 5, pp. 8–12. (in Russian)
21. Shiriaev V.D., Shagilova E.V., Beliakov M.T. About the form choice of the nonlinear regression model. *Proc. of the Youth and Systemic Modernization of the Country*, 2024, pp. 137–142. (in Russian)
22. Chen S. How much energy will AI really consume? The good, the bad and the unknown. *Nature*, 2025, vol. 639, no. 8053, pp. 22–24. https://doi.org/10.1038/d41586-025-00616-z
23. Kureichik V.V., Rodzin S.I. Bio-heuristics inspired by fauna (review). *Information Technologies*, 2023, vol. 29, no. 11, pp. 559–573. (in Russian). https://doi.org/10.17587/it.29.559-573

22. Chen S. How much energy will Al really consume? The good, the bad and the unknown // Nature. 2025. V. 639. N 8053. P. 22–24. https://doi.org/10.1038/d41586-025-00616-z
23. Курейчик В.В., Родзин С.И. Биоэвристики, инспирированные фауной (обзор) // Информационные технологии. 2023. Т. 29. № 11. С. 559–573. https://doi.org/10.17587/it.29.559-573
24. Суров И.А. Процессная семантика комплексных чисел // Математические структуры и моделирование. 2023. № 4 (68). С. 71-84. https://doi.org/10.24147/2222-8772.2023.4.71-84
25. Суров И.А. Какая разница? Прагматическая формализация смысла // Искусственный интеллект и принятие решений. 2023. № 1. С. 78–89. https://doi.org/10.14357/20718594230108
26. Суров И.А. Цветовая кодировка кубитных состояний // Информатика и автоматизация. 2023. Т. 22. № 5. С. 1207–1236. https://doi.org/10.15622/ia.22.5.9
27. Павлов А.В. Реализация регрессионных моделей обработки информации методом фурье-голографии // Известия Российской академии наук. Теория и системы управления. 2005. № 2. С. 29–36.
28. Lin X., Rivenson Y., Yardimei N.T., Veli M., Luo Y., Jarrahi M., Ozcan A. All-optical machine learning using diffractive deep neural networks // Science. 2018. V. 361. N 6406. P. 1004–1008. https://doi.org/10.1126/science.aat8084
29. Краснов А.Е., Головкин М.Е., Никольский Д.Н., Благовещенский В.Г. Волновая сеть для распознавания изображений // Автоматизация в промышленности. 2022. № 10. С. 28–33. https://doi.org/10.25728/avtprom.2022.10.06
30. Райков А.Н. Полностью аналоговый фотонный искусственный интеллект // Информационное общество. 2024. № 6. С. 168–179.
24. Surov I.A. Process semantics of complex numbers. *Mathematical Structures and Modeling*, 2023, vol. 68, no. 4, pp. 71-84. (in Russian). https://doi.org/10.24147/2222-8772.2023.4.71-84
25. Surov I.A. What is the difference? Pragmatic formalization of meaning. *Artificial Intelligence and Decision Making*, 2023, no. 1, pp. 78–89. (in Russian). https://doi.org/10.14357/20718594230108
26. Surov I.A. Color coding of qubit states. *Informatics and Automation*, 2023, vol. 22, no. 5, pp. 1207–1236. (in Russian). https://doi.org/10.15622/ia.22.5.9
27. Pavlov A.V. Implementation of regression models of data processing by the fourier-holography method. *Journal of Computer and Systems Sciences International*, 2005, vol. 44, no. 2, pp. 184–191.
28. Lin X., Rivenson Y., Yardimei N.T., Veli M., Luo Y., Jarrahi M., Ozcan A. All-optical machine learning using diffractive deep neural networks. *Science*, 2018, vol. 361, no. 6406, pp. 1004–1008. https://doi.org/10.1126/science.aat8084
29. Krasnov A.E., Golovkin M.E., Nikolskii D.N., Blagoveshchenskii V.G. Wave network for image recognition. *Avtomatizacija v Promyshlennosti*, 2022, no. 10, pp. 28–33. (in Russian). https://doi.org/10.25728/avtprom.2022.10.06
30. Raikov A.N. All-analogue photonic artificial intelligence. *Information Society*, 2024, no. 6, pp. 168–179. (in Russian)

Авторы

Богданов Павел Игоревич — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0000-1363-2451, pavel.bogdanov@metalab.ifmo.ru

Суров Илья Алексеевич — кандидат физико-математических наук, доцент, научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57219761715, https://orcid.org/0000-0001-5690-7507, ilya.a.surov@itmo.ru

Authors

Pavel I. Bogdanov — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0000-1363-2451, pavel.bogdanov@metalab.ifmo.ru

Ilya A. Surov — PhD (Physics & Mathematics), Associate Professor, Scientific Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57219761715, https://orcid.org/0000-0001-5690-7507, ilya.a.surov@itmo.ru

Статья поступила в редакцию 24.06.2025
Одобрена после рецензирования 24.08.2025
Принята к печати 30.09.2025

Received 24.06.2025
Approved after reviewing 24.08.2025
Accepted 30.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-988-995
УДК 004. 942

Оценка надежности восстанавливаемого кластера контейнерной виртуализацией

Владимир Анатольевич Богатырев[1]✉, Ван Кю Фунг[2]

[1] Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация
[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] vladimir.bogatyrev@gmail.com✉, https://orcid.org/0000-0003-0213-0223
[2] phungvanquy97@gmail.com, https://orcid.org/0009-0006-3278-1106

Аннотация

Введение. При построении отказоустойчивых кластеров высокой готовности с малыми задержками обслуживания потоков запросов все большее применение находит технология контейнерной виртуализации. При проектировании высоконадежных кластеров важной задачей является их структурно-параметрический модельно-ориентированный синтез с учетом влияния числа развертываемых контейнеров на производительность, задержки обслуживания потоков запросов и надежность системы. **Метод.** Обоснование решений по обеспечению высокой готовности кластера основывается на разработке моделей восстанавливаемого кластера при реконфигурации с учетом миграции виртуальных контейнеров. Новизна предлагаемых марковских моделей кластера состоит в учете двухэтапного восстановления его работоспособности с определением влияния на надежность кластера числа контейнеров, подлежащих миграции в процессе реконфигурации — до и после физического восстановления отказавших серверов. Рассмотрены два варианта миграции контейнеров при восстановлении кластера. В первом варианте, на этапе физического восстановления отказавшего сервера, миграция контейнеров на исправный сервер не происходит, а во втором — происходит. На втором этапе реконфигурации после физического восстановления отказавшего сервера осуществляется миграция контейнеров, при которой возможно как увеличение, так и уменьшение числа развернутых в них контейнеров. **Основные результаты.** На основе предлагаемых марковских моделей надежности кластера с контейнерной виртуализацией дана оценка его коэффициента готовности и определено влияние числа контейнеров, загружаемых при миграции на двух этапах реконфигурации, на надежность системы. **Обсуждение.** Предложенные марковские модели надежности кластера с контейнерной виртуализацией направлены на обоснование выбора проектных решений по организации и восстановлению работоспособности кластера после отказов серверов с учетом влияния вариантов реализации миграции виртуальных контейнеров на готовность системы. В дальнейших исследованиях предполагается анализ влияния вариантов миграции контейнеров как на готовность кластера, так и на задержки обслуживания запросов на двух рассматриваемых этапах реконфигурации.

Ключевые слова

отказоустойчивость, коэффициент готовности, контейнерная виртуализация, кластер, миграция контейнеров, марковская модель, надежность

Ссылка для цитирования: Богатырев В.А., Фунг В.К. Оценка надежности восстанавливаемого кластера контейнерной виртуализацией // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 988–995. doi: 10.17586/2226-1494-2025-25-5-988-995

Assessment of the reliability of a recoverable container virtualization cluster

Vladimir A. Bogatyrev[1]✉, Van Quy Phung[2]

[1] Saint Petersburg State University of Aerospace Instrumentation (SUAI), Saint Petersburg, 190000, Russian Federation

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] vladimir.bogatyrev@gmail.com✉, https://orcid.org/0000-0003-0213-0223

[2] phungvanquy97@gmail.com, https://orcid.org/0009-0006-3278-1106

Abstract

Container virtualization technology is increasingly being used in the development of fault-tolerant clusters with high availability and low request processing latency. In designing highly reliable clusters, a key task is the structural-parametric model-oriented synthesis which takes into account the impact of the number of deployed containers on performance, request processing latency, and system reliability. Justifying the choice of solutions to ensure high cluster reliability currently requires the development of reliability models for recoverable container virtualization clusters during reconfiguration, considering the migration of virtual containers. The basis for decisions to ensure high cluster availability is the development of models for a recoverable cluster during reconfiguration, taking into account the migration of virtual containers. The novelty of the proposed Markov model of a cluster lies in considering a two-stage recovery of its operability, determining the impact of the number of containers to be migrated during reconfiguration — both before and after the physical recovery of failed servers — on cluster reliability. Two options for container migration during cluster recovery are considered. In the first scenario, during the physical recovery phase of a failed server, container migration to a functional server does not occur, while in the second scenario it does. In the second stage of reconfiguration, following the physical recovery of a failed server, container migration takes place, allowing for either an increase or decrease in the number of containers deployed on them. Based on the proposed Markov models of cluster reliability with container virtualization, an evaluation of its readiness coefficient is provided, and the influence of the number of containers loaded during migration at the two reconfiguration stages on system reliability is determined. The proposed Markov models of cluster reliability with container virtualization are aimed at justifying design decisions for organizing and restoring cluster operability after server failures, considering the impact of container migration implementation options on system availability. Future research will analyze the impact of container migration options on both cluster availability and request processing latency at the two considered reconfiguration stages.

Keywords

fault tolerance, availability factor, container virtualization, cluster, container migration, Markov model, reliability

For citation: Bogatyrev V.A., Phung V.Q. Assessment of the reliability of a recoverable container virtualization cluster. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 988–995 (in Russian). doi: 10.17586/2226-1494-2025-25-5-988-995

Введение

Требования высокой надежности, доступности, отказоустойчивости и бесперебойности распределенных вычислений являются ключевыми для современных инфокоммуникационных систем [1–3].

Для распределенных компьютерных систем [4, 5] высокая доступность, надежность и отказоустойчивость при непрерывности вычислительных процессов и малых задержках обслуживания потока запросов достигаются при консолидации вычислительных ресурсов в кластерах, что предполагает их реконфигурацию после отказов с использованием репликации и динамической миграции виртуальных машин [6, 7] и контейнеров [8–10].

Построение отказоустойчивых кластеров требует обоснования и оптимизации решений, обеспечивающих надежность, производительность и высокоскоростной надежный доступ [11]. Эти решения должны быть взаимоувязаны на всех уровнях распределенной системы, включая резервированные средства передачи [12], обработки и хранения данных [13].

Выбор решений по организации высокоскоростного доступа и надежности сети может опираться на методы многопутевой маршрутизации и резервированных передач [14–16].

Обоснование и структурно-параметрическая оптимизация построения и эксплуатации кластера с целью обеспечения наибольшей надежности при малых задержках запросов в очередях должны опираться на расчеты надежности и задержек обслуживания запросов с учетом реконфигурации после отказов или изменений трафика. Известные марковские модели восстанавливаемых резервированных систем параллельной структуры не в полной мере отражают особенности функционирования кластеров, в том числе комплексное влияние на надежность и производительность кластера, организацию контроля, а также устройств обработки, хранения и внутрикластерного обмена [17–21].

Влияние организации тестового и оперативного контролей двухузлового кластера учитывается моделью, предложенной в работе [22]. Модель кластера, компонуемого из дублированных компьютерных узлов [23], отражает влияние на надежность и задержки обслуживания, связанное с организацией обмена внутри кластера, но не учитывает влияние различных дисциплин восстановления после отказов. Указанные модели не учитывают двухэтапность восстановления кластера после отказов, включающей сначала физическое, а затем информационное восстановления. Модель, отражающая двухэтапность восстановления узлов в дублированном кластере, объединяющем два компьютерных узла и два узла двухвходовой памяти при их полной связанности, предложена в [24] для различных вариантов разделения ресурсов процессорных узлов на решение функциональных задач и информационное восстановление памяти.

Представленные в работах [6, 7, 25] модели кластерных систем не отражают особенности технологии виртуальных машин, в том числе вопросы репликации и динамической миграции виртуальных машин при восстановлении физических серверов после их отказов [9, 26].

Модель двухузловой кластерной системы при репликации и динамической миграции виртуальных машин [27], а также при реконфигурации кластера после отказов его физических серверов, описанная в [28], не раскрывает особенностей реконфигурации кластера при контейнерной виртуализации.

В настоящее время существует потребность в разработке моделей кластеров контейнерной виртуализации, показывающих влияние реконфигурации системы при отказе серверов с их поэтапным физическим и информационным восстановлениями, предполагающим перераспределение трафика и миграцию виртуальных контейнеров. Эта потребность обусловлена необходимостью обоснования и оптимизации решений по построению кластера и его реконфигурации после отказов.

Цель данной работы — разработка моделей оценки надежности восстанавливаемого кластера при реконфигурации с учетом миграции виртуальных контейнеров, ориентированных на обоснование выбора решений по обеспечению высокой надежности кластера.

Новизна предлагаемой модели состоит в учете двухэтапного восстановления работоспособности кластера с оценкой влияния на надежность системы числа контейнеров, подлежащих миграции на этапах во время и после физического восстановления отказавших серверов.

Структура кластера с контейнерной виртуализацией

Рассмотрим простейшую реализацию кластера, включающую два сервера с контейнерной виртуализацией и узел — балансировщик нагрузки, осуществляющий распределение запросов между серверами и участвующий в развертывании и миграции виртуальных контейнеров, в том числе при реконфигурации после отказов. Балансировщик, распределяющий запросы, представляется одноканальной системой массового обслуживания с неограниченной очередью (M/M/1) [29].

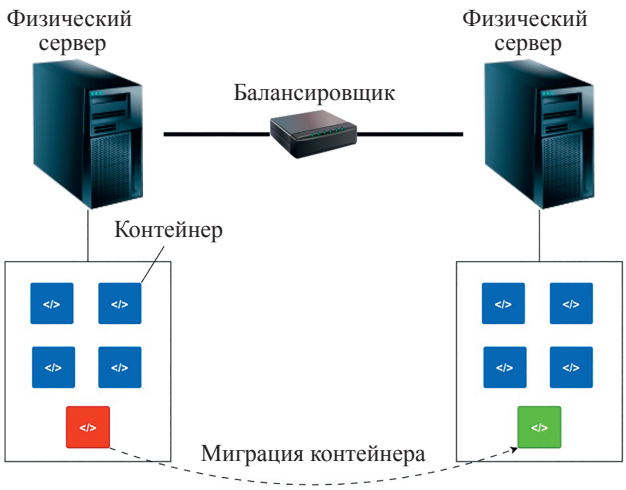

Рис. 1. Структура кластера
Fig. 1. Cluster Structure

Физический сервер, реализующий обслуживание потока запросов, моделируется многоканальной системой массового обслуживания с неограниченной очередью, при этом каждый канал соответствует контейнеру. Интенсивность обслуживания запросов в каждом канале (контейнере) зависит от числа загруженных и активных контейнеров, т. е. задействованных в выполнении поступающих запросов. Такой подход позволяет учесть разделение ограниченных вычислительных ресурсов сервера между контейнерами, участвующими и не участвующими в обслуживании запросов. Особенностью рассматриваемых систем с контейнерной виртуализацией является существование оптимального числа контейнеров, развернутых на узле, при котором достигается минимум задержки обслуживания запросов [30]. Причем оптимальное число развернутых контейнеров зависит от интенсивности потока запросов. Это позволяет предположить, что при отказе одного из серверов и перераспределении потока запросов на исправный сервер число развернутых на нем контейнеров до восстановления отказавшего сервера должно изменяться таким образом, чтобы минимизировать задержки обслуживания потока запросов при одновременном увеличении коэффициента готовности кластера.

Такой подход отличается от традиционного тем, что позволяет учесть разделение ограниченных вычислительных ресурсов физического сервера между активными контейнерами.

Модель надежности кластера при реконфигурации с миграцией контейнеров

Построим марковскую модель надежности восстанавливаемого кластера с учетом этапов физического и информационного восстановлений серверов, связанных с загрузкой (миграцией) виртуальных контейнеров на сервер после его физического восстановления. Информационное восстановление происходит под управлением узла-распределителя (балансировщика нагрузки) и может включать изменение числа развернутых в узлах контейнеров (миграцию контейнеров). Миграция виртуальных контейнеров осуществляется на двух этапах реконфигурации.

Этап 1. После отказа сервера до его физического восстановления.

Этап 2. После физического восстановления отказавшего сервера до его информационного восстановления, включающего загрузку заданного числа контейнеров.

Диаграммы состояний и переходов для двух рассматриваемых вариантов реконфигурации кластера представлены на рис. 2 и 3. Состояние кластера охарактеризуем двумя строками, верхняя из которых отображает состояние серверов, а нижняя — распределителя запросов. Распределитель рассматривается в двух состояниях: «1» — исправен, «X» — отказал. Для каждого сервера выделены состояния: исправность с указанием числа развернутых в нем контейнеров (n_0

или n_1); завершение физического восстановления до начала загрузки контейнеров — «0». При этом n_0 — число контейнеров в исходном состоянии (при работоспособности двух серверов), а n_1 — число контейнеров, загружаемых на сервер, сохранявший работоспособность после отказа второго сервера. Временем, связанным с уменьшением числа загруженных контейнеров, пренебрежем.

При этом считаем, что время загрузки контейнеров пропорционально их числу.

Рассмотрим два варианта дисциплины реконфигурации системы.

При варианте 1 реконфигурации, после отказа одного из двух серверов, вся нагрузка распределяется на исправный сервер, при этом не происходит изменение числа развернутых в нем контейнеров, т. е. на этапе 1 реконфигурации миграция контейнеров не реализуется (рис. 2). После физического восстановления отказавшего сервера осуществляется развертывание в нем первоначально заданного числа контейнеров, таким образом

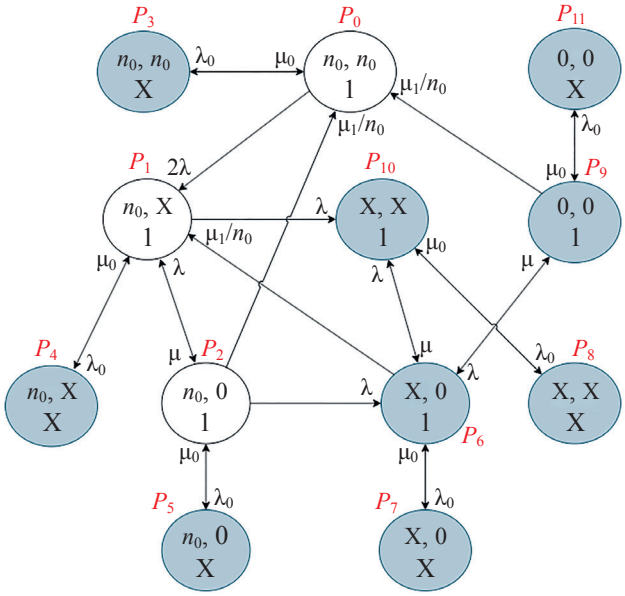

Рис. 2. Диаграмма состояний и переходов кластера без миграции контейнеров до этапа физического восстановления отказавшего сервера.

λ и λ_0 — интенсивности отказов физического сервера и балансировщика; μ_0 — интенсивность восстановления балансировщика; μ и μ_1 — интенсивности физического и информационного восстановлений серверов; $P_0, P_1, P_2, \ldots, P_{11}$ — вероятности состояний системы. Узлы, отмеченные серым фоном — состояния системы, при которых процесс выполнения требуемых функций невозможен из-за неработоспособности необходимых для этого ресурсов

Fig. 2. State and transition diagram of a cluster without migration prior to the physical recovery stage of the failed server.

λ and λ_0 — failure rates of the physical server and balancer; μ_0 — balancer recovery rate; μ and μ_1 — physical and information recovery rates of servers; $P_0, P_1, P_2, \ldots, P_{11}$ — system state probabilities.
Nodes marked with a gray background — system states in which the process of performing the required functions is impossible due to the inoperability of the resources required for this

на этапе 2 реконфигурации миграция контейнеров происходит. В результате система восстанавливается до исходного состояния.

При варианте 2 реконфигурации, после отказа одного из двух серверов, вся нагрузка распределяется на исправный сервер, но во время физического восстановления отказавшего сервера на работоспособном сервере под управлением балансировщика-распределителя нагрузки реализуется изменение числа (миграция) развернутых в нем контейнеров (рис. 3). На этапе 2 реконфигурации происходит изменение числа контейнеров на обоих серверах кластера, причем возможно как увеличение, так и уменьшение числа развернутых в них контейнеров.

Отметим, что при увеличении числа контейнеров на этапе 1 система может снижать [31] производительность сохранившего работоспособность сервера. Это связано с тем, что даже не принимающие участие в обслуживании запросов контейнеры при малой загрузке все равно потребляют определенную долю общих ресурсов, которые разделяются между активными и пассивными (простаивающими) контейнерами. Таким образом, чтобы минимизировать время пребывания запросов в системе, необходимо перенести лишь определенное число контейнеров, зависящее от интенсивности трафика. Анализ влияния числа развертываемых контейнеров на среднее время пребывания запросов в системе в зависимости от интенсивности потока трафика, определяющего число активных контейнеров, выполнен в работе [30].

При построении диаграмм состояний и переходов марковской модели предположим, что процесс физического восстановления узлов выполняется операто-

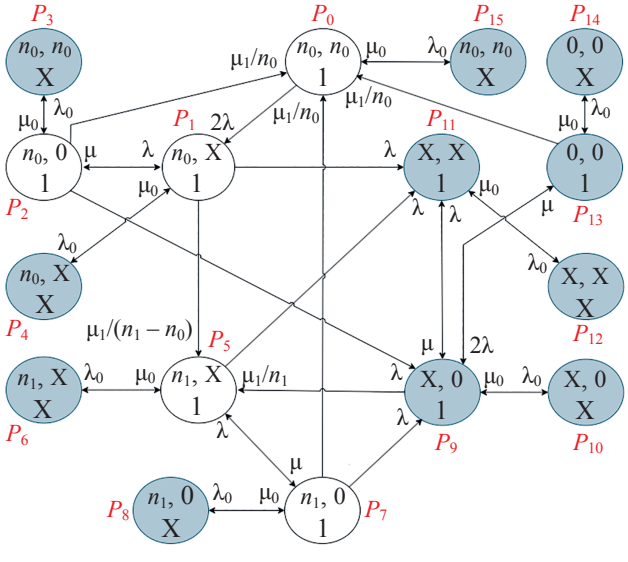

Рис. 3. Диаграмма состояний и переходов кластера с миграцией контейнеров на исправный сервер во время физического восстановления отказавшего сервера.

$P_0, P_1, P_2, \ldots, P_{15}$ — вероятности состояний системы.

Fig. 3. State and transition diagram of a cluster without migration prior to the physical recovery stage of the failed server.

$P_0, P_1, P_2, \ldots, P_{15}$ — system state probabilities

ром, при этом его работа не влияет на вычислительные процессы в работоспособном сервере и в том числе на миграцию контейнеров в нем, т. е. эти процессы могут быть совмещены по времени.

Для диспетчеризации и выполнения функциональных запросов требуется работоспособность балансировщика и хотя бы одного сервера. Миграция контейнеров требует физической работоспособности сервера и работоспособности балансировщика (диспетчера). Процесс загрузки контейнеров одновременно в два сервера может быть совмещен. Загрузка нескольких контейнеров в сервер производится последовательно.

При построении модели учтено, что накопления отказов серверов не происходит при отказе двух серверов или отказе одного из них и неработоспособности диспетчера, так как в этом случае выполнение запросов в системе невозможно и поэтому предполагаются отключения от серверов электропитания, что исключает отказы еще не отказавших серверов. При неработоспособности двух серверов отключение электропитания у диспетчера-балансировщика (несмотря на невозможность выполнения запросов в серверах) не происходит, ввиду того что в это время он должен выполнять оповещение для поступающих функциональных запросов о невозможности их выполнения.

Условием работоспособности кластера является работоспособность балансировщика нагрузки и хотя бы одного сервера с размещенными на нем контейнерами.

Для варианта 2 (рис. 3) переход в исходное состояние возможен при загрузке n_0 контейнеров параллельно в два либо в один сервер. Возможен также случай, когда в один сервер загружается n_0 контейнеров, а в другой меньшее их число.

Система уравнений, соответствующая диаграмме состояний и переходов на рис. 2, имеет следующий вид:

$$\begin{cases}
0 = -(2\lambda + \lambda_0)P_0 + \mu_0 P_3 + \frac{\mu_1}{n_0}P_2 + \frac{\mu_1}{n_0}P_9,\\
0 = -(\lambda_0 + \mu + \lambda)P_1 + 2\lambda P_0 + \frac{\mu_1}{n_0}P_6 + \lambda P_2 + \mu_0 P_4,\\
0 = -\left(\lambda + \lambda_0 + \lambda + \frac{\mu_1}{n_0}\right)P_2 + \mu P_1 + \mu_0 P_5,\\
0 = -\mu_0 P_3 + \lambda_0 P_0,\\
0 = -\mu_0 P_4 + \lambda_0 P_1,\\
0 = -\mu_0 P_5 + \lambda_0 P_2,\\
0 = -\left(\lambda_0 + \frac{\mu_1}{n_0} + \lambda + \mu\right)P_6 + \mu_0 P_7 + \lambda P_2 + \mu P_{10} + \lambda P_9,\\
0 = -\mu_0 P_7 + \lambda_0 P_6,\\
0 = -\mu_0 P_8 + \lambda_0 P_{10},\\
0 = -\left(\lambda_0 + \lambda + \frac{\mu_1}{n_0}\right)P_9 + \mu_0 P_{11} + \mu P_6,\\
0 = -(\mu + \lambda_0)P_{10} + \lambda P_1 + \lambda P_6 + \mu_0 P_8,\\
1 = \sum\limits_{i=0}^{11} P_i.
\end{cases}$$

Коэффициент готовности кластера в этом случае:

$$K = P_0 + P_1 + P_2.$$

Заметим, что при расчете коэффициента готовности учитываются все состояния, при которых возможно обслуживание потока запросов без их дифференциации по задержкам, независимо от работоспособности одного или двух серверов. При этом предполагается, что физическое и информационное восстановление отказавшего сервера не приводит к замедлению работы исправного сервера, а балансировщик не замедляет процессы в серверах.

Диаграмма состояний и переходов, представленных на рис. 3, приводит к системе уравнений:

$$\begin{cases}
0 = -(2\lambda + \lambda_0)P_0 + \frac{\mu_1}{n_0}P_2 + \mu_0 P_{15} + \frac{\mu_1}{n_0}P_{13} + \frac{\mu_1}{n_0}P_7,\\
0 = -\left(\lambda_0 + \mu + \frac{\mu_1}{n_1 - n_0} + \lambda\right)P_1 + \lambda P_2 + 2\lambda P_0 + \mu_0 P_4,\\
0 = -\left(\lambda + \lambda_0 + \lambda + \frac{\mu_1}{n_0}\right)P_2 + \mu P_1 + \mu_0 P_3,\\
0 = -\mu_0 P_3 + \lambda_0 P_2,\\
0 = -\mu_0 P_4 + \lambda_0 P_1,\\
0 = -\mu_0 P_5 + \lambda_0 P_2,\\
0 = -(\lambda_0 + \lambda + \mu)P_5 + \frac{\mu_1}{n_1 - n_0}P_1 + \mu_0 P_6 + \frac{\mu_1}{n_1}P_9 + \lambda P_7,\\
0 = -\mu_0 P_6 + \lambda_0 P_5,\\
0 = -\left(\lambda_0 + \mu + \frac{\mu_1}{n_0} + \mu\right)P_7 + \mu P_5 + \mu P_8,\\
0 = -\mu_0 P_8 + \lambda_0 P_7,\\
0 = -\left(\lambda_0 + \mu + \lambda + \frac{\mu_1}{n_1}\right)P_9 + \mu_0 P_{10} + \lambda P_{13} + \mu P_{11} +\\
+ \lambda P_2 + \lambda P_7,\\
0 = -\mu_0 P_{10} + \lambda_0 P_9,\\
0 = -(\lambda_0 + \mu)P_{11} + \mu_0 P_{12} + \lambda P_9 + \lambda P_5 + \lambda P_1,\\
0 = -\mu_0 P_{12} + \lambda_0 P_{11},\\
0 = -\left(\lambda_0 + \frac{\mu_1}{n_0} + \lambda\right)P_{13} + \mu_0 P_{14} + \mu P_9,\\
0 = -\mu_0 P_{14} + \lambda_0 P_{13},\\
1 = \sum\limits_{i=0}^{15} P_i.
\end{cases}$$

Отметим, что в состоянии, вероятность нахождения в котором равна P_1, выполнение запросов не происходит (так как в одном сервере вычислительный процесс и загрузка контейнеров не совмещены). Исходя из этого, стационарный коэффициент готовности кластера можно вычислить следующим образом:

$$K = P_0 + P_2 + P_5 + P_7.$$

Таблица. Зависимость коэффициента готовности кластера от числа контейнеров n_1, развернутых на этапе 1 реконфигурации в сохранившем работоспособность сервере

Table. Dependence of the cluster availability coefficient on the number of n_1 containers deployed at the first stage of reconfiguration in a functioning server

Число развернутых контейнеров n_1, шт.	Коэффициент готовности	Примечание
2	0,998995	На этапе 1 реконфигурации миграции контейнеров нет
3	0,998331	Число загруженных на этапе 1 реконфигурации контейнеров меньше числа контейнеров в отказавшем сервере
4	0,997999	Полная миграция всех контейнеров с отказавшего сервера на исправный
5	0,997799	Число контейнеров, дополнительно загруженных в исправный сервер, больше их числа в отказавшем сервере

Определим влияние числа контейнеров, загружаемых при миграции на двух этапах реконфигурации, на надежность кластера, определяемую по стационарному коэффициенту готовности. При расчетах примем, что в исходном состоянии в каждом из двух серверов загружено по $n_0 = 2$ контейнера, интенсивности отказов балансировщика и сервера $\lambda = \lambda_0 = 10^{-4}$ 1/ч, а интенсивности их восстановлений $\mu_0 = 0{,}1$ 1/ч и $\mu = 0{,}1$ 1/ч при интенсивности миграции контейнера $\mu_1 = 0{,}1$ 1/ч. Результаты расчета коэффициента готовности кластера в зависимости от числа контейнеров n_1, развернутых на этапе 1 реконфигурации в сохранившем работоспособность сервере, представлены в таблице.

Результаты расчета показывают отрицательное влияние на коэффициент готовности кластера увеличения числа загружаемых контейнеров при миграции на сервер, сохранивший работоспособность после отказа. Это связано с тем, что при миграции контейнеров на этапе 1 реконфигурации выполнение запросов на сервере, сохранившем работоспособность, не происходит. Миграция снижает коэффициент готовности, так как замедляет процесс восстановления: помимо времени на физическое восстановление добавляется время на информационное восстановление, что затягивает двухэтапный процесс реконфигурации и приводит к рискам перехода в неработоспособные состояния, требующие больше усилий (времени) для восстановления исправного состояния системы. Подчеркнем, что миграция контейнеров позволяет в состояниях с частичной потерей вычислительных ресурсов сохранить производительность системы, благодаря накоплению отказов, в том числе при обеспечении оптимального числа контейнеров, что приводит к максимальной производительности кластера.

Обсуждение результатов и направления развития

Предложенные модели надежности кластера направлены на обоснование выбора проектных решений по организации реконфигурации и восстановления работоспособности кластера после отказов, с учетом влияния различных вариантов реализации миграции виртуальных контейнеров на готовность системы.

Отличие предлагаемых марковских моделей заключается в учете двухэтапного восстановления работоспособности кластера при его реконфигурации, с определением влияния на надежность системы числа загружаемых при миграции контейнеров на двух выделенных этапах реконфигурации — до и после физического восстановления отказавших серверов.

В качестве дальнейшего развития работы предполагается исследование комплексного влияния различных вариантов миграции контейнеров как на готовность кластера, так и на задержки обслуживания запросов на двух рассматриваемых этапах реконфигурации.

Заключение

Предложены марковские модели надежности кластеров контейнерной виртуализации с двухэтапным восстановлением работоспособности отказавших серверов при их физическом и информационном восстановлении, с учетом влияния на готовность системы числа загружаемых при миграции контейнеров на каждом этапе восстановления.

Предложенные модели ориентированы на обоснование проектных решений по обеспечению надежности восстанавливаемых кластеров с контейнерной виртуализацией, включая дисциплины их восстановления и реконфигурации.

Литература

1. Goyal P., Deora S.S. Reliability of Trust Management Systems in Cloud Computing // Indian Journal of Cryptography and Network Security. 2022. V. 2. N 1. P. 1–5. https://doi.org/10.54105/ijcns.C1417.051322
2. Chen G., Guan N., Huang K., Yi W. Fault-tolerant real-time tasks scheduling with dynamic fault handling // Journal of Systems Architecture. 2020. V. 102. P. 101688. https://doi.org/10.1016/j.sysarc.2019.101688
3. Shubinsky I.B., Rozenberg I.N., Papic L. Adaptive fault tolerance in real-time information systems // Reliability: Theory and Applications. 2017. V. 12. N 1 (44). P. 18–25.
4. Chinnaiah N.R., Niranjan N. Fault tolerant software systems using software configurations for cloud computing // Journal of Cloud Computing. 2018. V. 7. P. 3. https://doi.org/10.1186/s13677-018-0104-9
5. Srivastava A., Kumar N. Queueing model based dynamic scalability for containerized cloud // International Journal of Advanced Computer Science and Applications. 2023. V. 14. N 1. P. 465–472. https://doi.org/10.14569/IJACSA.2023.0140150
6. Shukur H.M., Zeebaree S.R.M., Zebari R.R., Zeebaree D.Q., Ahmed O.M., Salih A.A. Cloud computing virtualization of resources allocation for distributed systems // Journal of Applied Science and Technology Trends. 2020. V. 1. N 2. P. 98–105. https://doi.org/10.38094/jastt1331
7. Alam I., Sharif K., Li F., Latif Z., Karim M.M., Biswas S., Nour B., Wang Y. A survey of network virtualization techniques for Internet of things using SDN and NFV // ACM Computing Surveys. 2020. V. 53. N 2. P. 1–40. https://doi.org/10.1145/3379444
8. Chen H., Qin W., Wang L. Task partitioning and offloading in IoT cloud-edge collaborative computing framework: a survey // Journal of Cloud Computing. 2022. V. 11. P. 86. https://doi.org/10.1186/s13677-022-00365-8
9. Kushchazli A., Safargalieva A., Kochetkova I., Gorshenin A. Queuing model with customer class movement across server groups for analyzing virtual machine migration in cloud computing // Mathematics. 2024. V. 12. N 3. P. 468. https://doi.org/10.3390/math12030468
10. Kumari P., Kaur P. A survey of fault tolerance in cloud computing // Journal of King Saud University — Computer and Information Sciences. 2021. V. 33. N 10. P. 1159–1176. https://doi.org/10.1016/j.jksuci.2018.09.021
11. Tatarnikova T.M., Arkhiptsev E.D. Designing fault-tolerant systems with micro-service architecture // Proc. of the 27th International Conference on Soft Computing and Measurements (SCM). 2024. P. 348–351. https://doi.org/10.1109/SCM62608.2024.10554143
12. Bogatyrev V.A. Protocols for dynamic distribution of requests through a bus with variable logic ring for reception authority transfer // Automatic Control and Computer Sciences. 1999. V. 33. N 1. P. 57–63.
13. Sovetov B.Ya., Tatarnikova T.M., Poymanova E.D. Storage scaling management model // Information and Control Systems. 2020. N 5 (108). P. 43–49. https://doi.org/10.31799/1684-8853-2020-5-43-49
14. Bogatyrev A.V., Bogatyrev V.A., Bogatyrev S.V. The probability of timeliness of a fully connected exchange in a redundant real-time communication system // Proc. of the Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). 2020. P. 1–4. https://doi.org/10.1109/WECONF48837.2020.9131517
15. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Control of multipath transmissions in the nodes of switching segments of reserved paths // Proc. of the International Conference on Information, Control, and Communication Technologies (ICCT). 2022. P. 1–5. https://doi.org/10.1109/ICCT56057.2022.9976839
16. Terskov V., Sakash I. The reliability evaluation of local computer networks using markov model of multiple heterogeneous groups of switches // E3S Web of Conferences. 2024. V. 592. P. 3036. https://doi.org/10.1051/e3sconf/202459203036
17. Половко А.М., Гуров С.В. Основы теории надежности. СПб.: БХВ-Петербург, 2006. 702 с.
18. Koren I. Fault-Tolerant Systems. Morgan Kaufmann, 2007. 400 p.
19. Aysan H. Fault-tolerance strategies and probabilistic guarantees for real-time systems. Doctoral dissertation, Mälardalen University, 2012. 109 p.
20. Рахман П.А., Шарипов М.И. Модель надежности двухузлового кластера приложений высокой готовности в системах управления

References

1. Goyal P., Deora S.S. Reliability of Trust Management Systems in Cloud Computing. *Indian Journal of Cryptography and Network Security*, 2022, vol. 2, no. 1, pp. 1–5. https://doi.org/10.54105/ijcns.C1417.051322
2. Chen G., Guan N., Huang K., Yi W. Fault-tolerant real-time tasks scheduling with dynamic fault handling. *Journal of Systems Architecture*, 2020, vol. 102, pp. 101688. https://doi.org/10.1016/j.sysarc.2019.101688
3. Shubinsky I.B., Rozenberg I.N., Papic L. Adaptive fault tolerance in real-time information systems. *Reliability: Theory and Applications*, 2017, vol. 12, no. 1 (44), pp. 18–25.
4. Chinnaiah N.R., Niranjan N. Fault tolerant software systems using software configurations for cloud computing. *Journal of Cloud Computing*, 2018, vol. 7, pp. 3. https://doi.org/10.1186/s13677-018-0104-9
5. Srivastava A., Kumar N. Queueing model based dynamic scalability for containerized cloud. *International Journal of Advanced Computer Science and Applications*, 2023, vol. 14, no. 1, pp. 465–472. https://doi.org/10.14569/IJACSA.2023.0140150
6. Shukur H.M., Zeebaree S.R.M., Zebari R.R., Zeebaree D.Q., Ahmed O.M., Salih A.A. Cloud computing virtualization of resources allocation for distributed systems. *Journal of Applied Science and Technology Trends*, 2020, vol. 1, no. 2, pp. 98–105. https://doi.org/10.38094/jastt1331
7. Alam I., Sharif K., Li F., Latif Z., Karim M.M., Biswas S., Nour B., Wang Y. A survey of network virtualization techniques for Internet of things using SDN and NFV. *ACM Computing Surveys*, 2020, vol. 53, no. 2, pp. 1–40. https://doi.org/10.1145/3379444
8. Chen H., Qin W., Wang L. Task partitioning and offloading in IoT cloud-edge collaborative computing framework: a survey. *Journal of Cloud Computing*, 2022, vol. 11, pp. 86. https://doi.org/10.1186/s13677-022-00365-8
9. Kushchazli A., Safargalieva A., Kochetkova I., Gorshenin A. Queuing model with customer class movement across server groups for analyzing virtual machine migration in cloud computing. *Mathematics*, 2024, vol. 12, no. 3, pp. 468. https://doi.org/10.3390/math12030468
10. Kumari P., Kaur P. A survey of fault tolerance in cloud computing. *Journal of King Saud University — Computer and Information Sciences*, 2021, vol. 33, no. 10, pp. 1159–1176. https://doi.org/10.1016/j.jksuci.2018.09.021
11. Tatarnikova T.M., Arkhiptsev E.D. Designing fault-tolerant systems with micro-service architecture. *Proc. of the 27th International Conference on Soft Computing and Measurements (SCM)*, 2024, pp. 348–351. https://doi.org/10.1109/SCM62608.2024.10554143
12. Bogatyrev V.A. Protocols for dynamic distribution of requests through a bus with variable logic ring for reception authority transfer. *Automatic Control and Computer Sciences*, 1999, vol. 33, no. 1, pp. 57–63.
13. Sovetov B.Ya., Tatarnikova T.M., Poymanova E.D. Storage scaling management model. *Information and Control Systems*, 2020, no. 5 (108), pp. 43–49. https://doi.org/10.31799/1684-8853-2020-5-43-49
14. Bogatyrev A.V., Bogatyrev V.A., Bogatyrev S.V. The probability of timeliness of a fully connected exchange in a redundant real-time communication system. *Proc. of the Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*, 2020, pp. 1–4. https://doi.org/10.1109/WECONF48837.2020.9131517
15. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Control of multipath transmissions in the nodes of switching segments of reserved paths. *Proc. of the International Conference on Information, Control, and Communication Technologies (ICCT)*, 2022, pp. 1–5. https://doi.org/10.1109/ICCT56057.2022.9976839
16. Terskov V., Sakash I. The reliability evaluation of local computer networks using markov model of multiple heterogeneous groups of switches. *E3S Web of Conferences*, 2024, vol. 592, pp. 3036. https://doi.org/10.1051/e3sconf/202459203036
17. Polovko A.M., Gurov S.V. *Fundamentals of Reliability Theory*. St. Petersburg, BHV-Petersburg Publ., 2006, 702 p. (in Russian)
18. Koren I. *Fault-Tolerant Systems*. Morgan Kaufmann, 2007, 400 p.
19. Aysan H. *Fault-tolerance strategies and probabilistic guarantees for real-time systems*. Doctoral dissertation, Mälardalen University, 2012, 109 p.

предприятием // Экономика и менеджмент систем управления. 2015. № 3 (17). С. 85–102.
21. Хомоненко А.Д., Благовещенская Е.А., Проурзин О.В., Андрук А.А. Прогноз надежности кластерной вычислительной системы с помощью полумарковской модели альтернирующих процессов и мониторинга // Наукоемкие технологии в космических исследованиях Земли. 2018. Т. 10. № 4. С. 72–82. https://doi.org/10.24411/2409-5419-2018-10099
22. Bogatyrev V.A., Vinokurova M.S. Control and safety of operation of duplicated computer systems // Communications in Computer and Information Science. 2017. V. 700. P. 331–342. https://doi.org/10.1007/978-3-319-66836-9_28
23. Bogatyrev V.A. Exchange of duplicated computing complexes in fault-tolerant systems // Automatic Control and Computer Sciences. 2011. V. 45. N 5. P. 268–276. https://doi.org/10.3103/S014641161105004X
24. Богатырев В.А., Богатырев С.В., Богатырев А.В. Оценка готовности компьютерной системы к своевременному обслуживанию запросов при его совмещении с информационным восстановлением памяти после отказов // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 3. С. 608–617. https://doi.org/10.17586/2226-1494-2023-23-3-608-617
25. Compastié M., Badonnel R., Festor O., He R. From virtualization security issues to cloud protection opportunities: An in-depth analysis of system virtualization models // Computers & Security. 2020. V. 97. P. 101905. https://doi.org/10.1016/j.cose.2020.101905
26. Choudhary A., Govil M.C., Singh G., Awasthi L.K., Pilli E.S., Kapil D. A critical survey of live virtual machine migration techniques // Journal of Cloud Computing. 2017. V. 6. P. 23. https://doi.org/10.1186/s13677-017-0092-1
27. Алексанков С.М. Модели динамической миграции с итеративным подходом и сетевой миграции виртуальных машин // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 6. С. 1098–1104. https://doi.org/10.17586/2226-1494-2015-15-6-1098-1104
28. Bogatyrev V.A., Derkach A.N. Evaluation of a cyber-physical computing system with migration of virtual machines during continuous computing // Computers. 2020. V. 9. N 2. P. 42. https://doi.org/10.3390/computers9020042
29. Клейнрок Л. Теория массового обслуживания. М.: Машиностроение, 1979. 432 с.
30. Фунг В., Богатырев В.А., Кармановский Н.С., Лэ В.Х. Оценка вероятностно-временных характеристик компьютерной системы с контейнерной виртуализацией // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 2. С. 249–255. https://doi.org/10.17586/2226-1494-2024-24-2-249-255
31. Nguyen T.A., Kim D.S., Park J.S. A comprehensive availability modeling and analysis of a virtualized servers system using stochastic reward nets // The Scientific World Journal. 2014. V. 2014. P. 165316. https://doi.org/10.1155/2014/165316

20. Rakhman P.A., Sharipov M.I. Reliability model of a two-node cluster of high-availability applications in enterprise management systems. *Ekonomika i menedzhment sistem upravleniya*, 2015, no. 3 (17), pp. 85–102. (in Russian)
21. Khomonenko A.D., Blagoveshchenskaya E.A., Prourzin O.V., Andruk A.A. Forecasting the reliability of a cluster computing system using a semi-Markov model of alternating processes and monitoring. *High Technologies in Earth Space Research. H&ES Research*, 2018, vol. 10, no. 4, pp. 72–82. (in Russian). https://doi.org/10.24411/2409-5419-2018-10099
22. Bogatyrev V.A., Vinokurova M.S. Control and safety of operation of duplicated computer systems. *Communications in Computer and Information Science*, 2017, vol. 700, pp. 331–342. https://doi.org/10.1007/978-3-319-66836-9_28
23. Bogatyrev V.A. Exchange of duplicated computing complexes in fault-tolerant systems. *Automatic Control and Computer Sciences*, 2011, vol. 45, no. 5, pp. 268–276. https://doi.org/10.3103/S014641161105004X
24. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Assessment of the readiness of a computer system for timely servicing of requests when combined with information recovery of memory after failures. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 608–617. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-3-608-617
25. Compastié M., Badonnel R., Festor O., He R. From virtualization security issues to cloud protection opportunities: An in-depth analysis of system virtualization models. *Computers & Security*, 2020, vol. 97, pp. 101905. https://doi.org/10.1016/j.cose.2020.101905
26. Choudhary A., Govil M.C., Singh G., Awasthi L.K., Pilli E.S., Kapil D. A critical survey of live virtual machine migration techniques. *Journal of Cloud Computing*, 2017, vol. 6, pp. 23. https://doi.org/10.1186/s13677-017-0092-1
27. Aleksankov S.M. Models of live migration with iterative approach and move of virtual machines. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol.15, no. 6, pp. 1098–1104. (in Russian). https://doi.org/10.17586/2226-1494-2015-15-6-1098-1104
28. Bogatyrev V.A., Derkach A.N. Evaluation of a cyber-physical computing system with migration of virtual machines during continuous computing. *Computers*, 2020, vol. 9, no. 2, pp. 42. https://doi.org/10.3390/computers9020042
29. Kleinrock L. *Queueing Systems*. Volume 1: Theory. Wiley-Interscience, 1975, 417 p.
30. Phung V.Q., Bogatyrev V.F., Karmanovskiy N.S., Le V.H. Evaluation of probabilistic- temporal characteristics of a computer system with container virtualization. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 2, pp. 249–255. (in Russian). https://doi.org/10.17586/2226-1494-2024-24-2-249-255
31. Nguyen T.A., Kim D.S., Park J.S. A comprehensive availability modeling and analysis of a virtualized servers system using stochastic reward nets. *The Scientific World Journal*. 2014. V. 2014. P. 165316. https://doi.org/10.1155/2014/165316

Авторы

Богатырев Владимир Анатольевич — доктор технических наук, профессор, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация; профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 7006571069, https://orcid.org/0000-0003-0213-0223, vladimir.bogatyrev@gmail.com

Фунг Ван Кю — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0006-3278-1106, phungvanquy97@gmail.com

Authors

Vladimir A. Bogatyrev — D.Sc., Professor, Saint Petersburg State University of Aerospace Instrumentation (SUAI), Saint Petersburg, 190000, Russian Federation; Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 7006571069, https://orcid.org/0000-0003-0213-0223, vladimir.bogatyrev@gmail.com

Van Quy Phung — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0006-3278-1106, phungvanquy97@gmail.com

Статья поступила в редакцию 08.05.2025
Одобрена после рецензирования 24.08.2025
Принята к печати 21.09.2025

Received 08.05.2025
Approved after reviewing 24.08.2025
Accepted 21.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2025-25-5-996-998
УДК 514.124.1

Вычисление объема симплекса в барицентрических координатах в многомерном евклидовом пространстве

Марина Александровна Степанова✉

Российский государственный педагогический университет им. А. И. Герцена, Санкт-Петербург, 191186, Российская Федерация

ratkebug@yandex.ru✉, https://orcid.org/0009-0000-0856-9507

Аннотация

Представлено описание трех способов вычисления k-мерного объема k-мерного симплекса в n-мерном евклидовом пространстве ($n \geq k$) в канонической барицентрической системе координат. Первый способ основан на вычислении для n-мерного симплекса с помощью определителя барицентрической матрицы, столбцами которой являются барицентрические координаты вершин симплекса. Второй способ представляет вычисление объема для k-мерного симплекса с помощью определителя Кэли–Менгера через длины ребер симплекса, которые можно найти по барицентрическим координатам вершин. Третьим способом является вычисление с помощью определителя Грама для построенной по вершинам k-мерного симплекса системы векторов в $(n + 1)$-мерном евклидовом пространстве.

Ключевые слова

барицентрическая система координат, барицентрическая матрица, базисный симплекс, объем симплекса

Ссылка для цитирования: Степанова М.А. Вычисление объема симплекса в барицентрических координатах в многомерном евклидовом пространстве // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 996–998. doi: 10.17586/2226-1494-2025-25-5-996-998

Calculation of the volume of simplex in barycentric coordinates in a multidimensional Euclidean space

Marina A. Stepanova✉

Herzen University, Saint Petersburg, 191186, Russian Federation

ratkebug@yandex.ru✉, https://orcid.org/0009-0000-0856-9507

Abstract

The paper describes three ways of calculating the k-dimensional volume of the k-dimensional simplex in the n-dimensional Euclidean space ($n \geq k$) in the canonical barycentric coordinate system. The first method is to calculate for the n-dimensional simplex using the determinant of the barycentric matrix, the columns of which are the barycentric coordinates of the simplex vertices. The second method is to calculate the volume for k-dimensional simplex using the Cayley–Menger determinant through the lengths of the simplex edges which can be found from the barycentric coordinates of the vertices. The third method is to compute using a Gram determinant for a system of vectors constructed from the vertices of a given simplex in a $(n + 1)$-dimensional Euclidean space.

Keywords

barycentric coordinate system, barycentric matrix, basic simplex, simplex volume

For citation: Stepanova M.A. Calculation of the volume of simplex in barycentric coordinates in a multidimensional Euclidean space. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 996–998 (in Russian). doi: 10.17586/2226-1494-2025-25-5-996-998

Во многих современных прикладных задачах используется барицентрическая система координат для определения положения одних объектов относительно других (базисных) объектов или при описании распределения долей объекта. Например, барицентрические координаты можно встретить в научных работах по проектированию расположения мобильных интеллектуальных устройств, в геологических исследованиях, в исследованиях свойств химических растворов, при моделировании цветного изображения, в исследованиях по обобщению классических задач механики, в классических задачах линейного программирования и других (например, [1–4]). С каждым годом область применения барицентрического метода расширяется.

Существует некоторый недостаток в вычислениях при применении барицентрических координат. Вычисления длин, площадей и объемов осуществляются с помощью перехода к декартовым или аффинным координатам, и очень редко производятся непосредственно в барицентрических координатах, и то в малой размерности (два или три). Основная цель настоящей работы — устранить упущение в вычислениях и получить явные формулы для расчета объемов симплексов.

Рассмотрим три способа вычислений k-мерного симплекса.

Пусть $\mathbb{E}^n$ — n-мерное евклидово пространство. Зафиксируем в пространстве $\mathbb{E}^n$ каноническую барицентрическую систему координат с правильным базисным симплексом Δ, все ребра которого равны 1 [5–8]. Пусть P – k-мерный симплекс с вершинами C_1, C_2, …, C_{k+1} ($n \ge k$), и известны координаты точек C_j в барицентрической системе координат: C_j (x_{ij}), где $i = 1, \ldots, n+1$; $j = 1, \ldots, k+1$; $\sum\limits_{i=1}^{n+1} x_{ij} = 1$. Необходимо вычислить $V^{(k)}(P)$ — k-мерный объем симплекса P. Для канонической барицентрической системы координат в работах [6, 7] для случаев $n = 2$ и $n = 3$ получены формулы для косого, смешанного и векторного произведений векторов, следовательно, задача в этих размерностях решена.

Аналогичная задача для декартовой системы координат решена много лет назад: можно воспользоваться внешним произведением векторов или определителем Грама [9], либо определителем Кэли–Менгера [10–13].

В случае $k = n$ симплекс P можно считать базисным симплексом другой барицентрической системы координат, тогда $\mathbf{B} = (\mathbf{x}_{ij})$ — матрица перехода от симплекса Δ к симплексу P [14]. В работе [15] доказано, что $V^{(n)}(P) = V^{(n)}(\Delta)|\det \mathbf{B}|$, при этом $V^{(n)}(\Delta) = \dfrac{\sqrt{n+1}}{n!2^{n/2}}$ [12, 16]. В результате получаем формулу:

$$V^{(n)}(P) = \frac{\sqrt{n+1}}{n!2^{n/2}}|\det\mathbf{B}|, \qquad (1)$$

где в матрице $\mathbf{B}$ по столбцам заданы координаты вершин симплекса P.

Длина ребра C_lC_m симплекса P, как было показано в [5], вычисляется по формуле $|C_lC_m|^2 = \dfrac{1}{2}\sum\limits_{i=1}^{n+1}(x_{il} - x_{im})^2$.

Рассчитаем по длинам ребер объем P с помощью определителя Кэли–Менгера, и получим:

$$[V^{(k)}(P)]^2 = \frac{(-1)^{k+1}}{2^k(k!)^2}\begin{vmatrix} 0 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & \cdots & d_{1\,k+1}^2 \\ 1 & d_{21}^2 & 0 & d_{23}^2 & \cdots & d_{2\,k+1}^2 \\ 1 & d_{31}^2 & d_{32}^2 & 0 & \cdots & d_{3\,k+1}^2 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & d_{k+1\,1}^2 & d_{k+1\,2}^2 & d_{k+1\,3}^2 & \cdots & 0 \end{vmatrix}, \qquad (2)$$

где $d_{lm}^2 = \dfrac{1}{2}\sum\limits_{i=1}^{n+1}(x_{il} - x_{im})^2$.

Рассмотрим пространство $\mathbb{E}^n$ как гиперплоскость α в пространстве $\mathbb{E}^{n+1}$. Пусть $A_1, A_2, \ldots, A_{n+1}$ — вершины базисного симплекса Δ. Существует точка $O \notin \alpha$ такая, что $OA_i \perp OA_j$ $(i \ne j)$ и $|OA_i|^2 = 1/2$. Введем аффинную систему координат в $\mathbb{E}^{n+1}$ с началом в точке O и с базисными векторами $\mathbf{a}_i = \mathbf{OA}_i$. В этой системе координат плоскость α задается уравнением $x_1 + x_2 + \ldots + x_{n+1} = 1$. Заметим, что для любой точки $M \in \alpha$ ее барицентрические координаты в α совпадают с координатами радиус-вектора $\mathbf{OM}$ в базисе $\mathbf{a}_i$, и барицентрические координаты любого вектора, параллельного α, совпадают с его координатами в базисе $\mathbf{a}_i$. Матрица скалярного произведения (матрица Грама) в базисе $\mathbf{a}_i$ следующая: $\Gamma(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{n+1}) = \dfrac{1}{2}\mathbf{E}_{n+1}$, где $\mathbf{E}_{n+1}$ — единичная матрица $(n + 1) \times (n + 1)$. Исходя из этого, скалярное произведение векторов $\mathbf{v} = (x_1, x_2, \ldots, x_{n+1})$ и $\mathbf{u} = (y_1, y_2, \ldots, y_{n+1})$ в данном базисе вычислим по формуле:

$$\mathbf{v}\cdot\mathbf{u} = \frac{1}{2}\sum_i x_i y_i.$$

Симплекс P натянут на векторы $\mathbf{v}_j = \mathbf{C}_1\mathbf{C}_{j+1}$ $(j = 1, \ldots, k)$ и $\mathbf{v}_j = \sum\limits_{m=1}^{n+1}(x_{mj+1} - x_{m1})\mathbf{a}_m$. Пусть $(\mathbf{g}_{ij}) = (\mathbf{v}_i\mathbf{v}_j)$ — матрица Грама векторов $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$. Найдем объем симплекса P с помощью определителя матрицы Грама для векторов $\mathbf{v}_j$ [9, С. 315], тогда:

$$V^{(k)}(P) = \frac{1}{k!}\sqrt{\det(\mathbf{g}_{ij})}, \qquad (3)$$

где $\mathbf{g}_{ij} = \dfrac{1}{2}\sum\limits_{m=1}^{n+1}(x_{mi+1} - x_{m1})(x_{mj+1} - x_{m1})$.

Полученные формулы (1)–(3) для вычисления объемов симплексов могут расширить применение барицентрических координат и упростить вычисления в уже имеющихся математических моделях. Прежде всего, следует ожидать применение представленных формул в задачах оптимизации расположения объектов (например, при проектировании расположения устройств приема и преобразования сигналов), в расчетах, связанных с интерполяцией на многомерных объектах, при обработке статистических данных в многомерном случае, при обобщении классических задач линейного программирования.

Литература

1. Барышников В.Д., Качальский В.Г., Лудцев К.Б. Определение неизвестных точечных свойств массива горных пород методом интерполяции с использованием барицентрических координат // Фундаментальные и прикладные вопросы горных наук. 2014. Т. 1. № 1. С. 51–55.
2. Кудрявченко И.В., Карлусов В.Ю. Измерение параметров движения мобильных объектов с «коллективным» поведением // Автоматизация и измерения в машино-приборостроении. 2018. № 3 (3). С. 92–99.
3. Никонов В.И. Относительные равновесия в задаче о движении треугольника и точки под действием сил взаимного притяжения // Вестник Московского университета. Серия 1: Математика. Механика. 2014. № 2. С. 45–51.
4. Степанова М.А. Применение барицентрических комбинаций точек в теории выпуклых многогранников // Методика преподавания в современной школе: актуальные проблемы и инновационные решения: материалы II Российско-узбекской научно-практической конференции. СПб.: РГПУ им. А. И. Герцена, 2024. С. 263–268.
5. Степанова МА., Малинникова К.С. Барицентрические матрицы и канонические барицентрические координаты в евклидовом пространстве // Современные проблемы математики и математического образования: Международная научная конференция «78 Герценовские чтения». СПб: Издательско-полиграфическая ассоциация высших учебных заведений, 2025. С. 360–364.
6. Понарин Я.П. Основные метрические задачи планиметрии в барицентрических координатах // Математический вестник педвузов Волго-Вятского региона. 2002. № 4. С. 114–132.
7. Понарин Я.П. Метод барицентрических координат в метрических задачах стереометрии // Математический вестник педвузов Волго-Вятского региона. 2004. № 6. С. 189–200.
8. Ungar A.A. Barycentric Calculus in Euclidean and Hyperbolic Geometry a Comparative Introduction. North Dakota State University, 2010. 360 p. https://doi.org/10.1142/7740
9. Берже М. Геометрия. Т. 1. М.: Мир, 1984. 560 с.
10. Сабитов И.Х. Объем многогранника как функция длин его ребер // Фундаментальная и прикладная математика. 1996. Т. 2. № 4. С. 305–307.
11. D'Andrea C., Sombra M. The Cayley-Menger determinant is irreducible for n ≥ 3 // Siberian Mathematical Journal. 2005. V. 46. N 1. P. 71–76. https://doi.org/10.1007/s11202-005-0007-0
12. Fiedler M. Matrices and Graphs in Geometry. Cambridge University Press, 2011. 206 p.
13. Sabitov I.K. The volume as a metric invariant of polyhedra // Discrete and Computational Geometry. 1998. V. 20. N 4. P. 405–425. https://doi.org/10.1007/PL00009393
14. Степанова М.А. Барицентрическая система координат барицентрическая группа // Современные проблемы математики и математического образования: сборник научных трудов международной научной конференции. СПб: РГПУ им. А.И. Герцена, 2024. С. 356–360.
15. Степанова М.А. Геометрический смысл матрицы перехода от барицентрической системы координат к барицентрической системе координат // Актуальные аспекты развития науки и общества в эпоху цифровой трансформации: сборник материалов XX Международной научно-практической конференции. М.: АНО ДПО «Центр развития образования и науки», 2025. С. 428–433.
16. Buchholz R.H. Perfect Pyramids // Bulletin of the Australian Mathematical Society. 1991. V. 45. P. 353–368.

References

1. Baryshnikov V.D., Kachalsky V.G., Ludtsev K.B. Point estimation of unknown properties of rock mass using interpolation and barycentric coordinates. *Fundamental'nye i Prikladnye Voprosy Gornyh Nauk*, 2014, vol. 1, no. 1, pp. 51–55. (in Russian)
2. Kudryavchenko I.V., Karlusov V.Yu. Measurement of mobile objects motion parameters with "collective" behavior. *Automation and Measurement in Mechanical Engineering and Instrument Engineering*, 2018, no. 3 (3), pp. 92–99. (in Russian)
3. Nikonov V.I. Relative equilibria in the motion of a triangle and a point under mutual attraction. *Moscow University Mechanics Bulletin*, 2014, vol. 69, no. 2, pp. 44–50. https://doi.org/10.3103/S0027133014020034
4. Stepanova M.A. Barycentric point combinations and convex polyhedral. *Proc. of the 2nd Russian-Uzbek Scientific and Practical Conference*, 2024, pp. 263–268. (in Russian)
5. Stepanova M.A., Malinnikova K.S. Barycentric matrices and canonical barycentric coordinates in euclidean space. *Proc. of the Contemporary Problems of Mathematics and Mathematical Education*, 2025, pp. 360–364. (in Russian)
6. Ponarin Y.P. Basic metric problems of planimetry in barycentric coordinates. *Mathematical Bulletin of Pedagogical Universities of the Volga-Vyatka Region*, 2002, vol. 4, pp. 114-132. (in Russian)
7. Ponarin Y.P. The method of barycentric coordinates in metric problems of stereometry. *Mathematical Bulletin of Pedagogical Universities of the Volga-Vyatka Region*, 2004, no. 6, pp. 189–200. (in Russian)
8. Ungar A.A. *Barycentric Calculus in Euclidean and Hyperbolic Geometry a Comparative Introduction*. North Dakota State University, 2010, 360 p. https://doi.org/10.1142/7740
9. Berger M. *Géométrie. Action de Groupes, Espaces Affines et Projectifs*. Cedic/Fernand Nathan, 1979, 200 p. (in French)
10. Sabitov Kh. The polyhedron's volume as a function of length of its edges. *Fundamentalnaya i prikladnaya matematika*, 1996, vol. 2, no. 4, pp. 305–307. (in Russian)
11. D'Andrea C., Sombra M. The Cayley-Menger determinant is irreducible for n ≥ 3. *Siberian Mathematical Journal*, 2005, vol. 46, no. 1, pp. 71–76. https://doi.org/10.1007/s11202-005-0007-0
12. Fiedler M. *Matrices and Graphs in Geometry*. Cambridge University Press, 2011, 206 p.
13. Sabitov I.K. The volume as a metric invariant of polyhedral. *Discrete and Computational Geometry*, 1998, vol. 20, no. 4, pp. 405–425. https://doi.org/10.1007/PL00009393
14. Stepanova M. Barycentric coordinate system barycentric group. *Proc. of the 77th Herzen Readings*, 2024, pp. 356–360. (in Russian)
15. Stepanova M. The geometric meaning of the transition matrix from barycentric coordinate system to barycentric coordinate system. *Proc. of the 20th International Scientific and Practical Conference Current aspects of the development of science and society in the Era of Digital Transformation*, 2025, 428–433. (in Russian)
16. Buchholz R.H. Perfect Pyramids. *Bulletin of the Australian Mathematical Society*, 1991, vol. 45, pp. 353–368.

Автор

Степанова Марина Александровна — кандидат физико-математических наук, доцент, доцент, Российский государственный педагогический университет им. А. И. Герцена, Санкт-Петербург, 191186, Российская Федерация, sc 8840198800, https://orcid.org/0009-0000-0856-9507, ratkebug@yandex.ru

Author

Marina A. Stepanova — PhD (Physics & Mathematics), Associate Professor, Associate Professor, Herzen University, Saint Petersburg, 191186, Russian Federation, sc 8840198800, https://orcid.org/0009-0000-0856-9507, ratkebug@yandex.ru

Статья поступила в редакцию 07.07.2025
Одобрена после рецензирования 23.07.2025
Принята к печати 23.09.2025

Received 07.07.2025
Approved after reviewing 23.07.2025
Accepted 23.09.2025

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-5-999-1001
УДК 004.961

Вероятностный метод матричной кластеризации с априорным распределением признаков для формирования несмещенной контрольной группы

Дмитрий Андреевич Усольцев✉

Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США
Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

dusoltsev.27@gmail.com✉, https://orcid.org/0000-0001-8072-310X

Аннотация

Предложен метод вероятностной кластеризации матричных данных с использованием априорного распределения признаков и понижения размерности (Singular Value Decomposition, SVD). Метод позволяет выделять в большой контрольной группе кластер, статистически сопоставимый с тестовой группой, что снижает систематические искажения при дальнейшем сравнительном анализе. Показано, что предлагаемый метод позволяет корректно подбирать контрольную группу в случаях, когда известный метод ближайшего соседа дает ложноположительные результаты. Представленный метод применялся для отбора контрольных групп в исследованиях на основе медико-генетической базы данных, проводимых в Национальном медицинском исследовательском центре имени В.А. Алмазова.

Ключевые слова

матричная кластеризация, SVD, априорное распределение, расстояние Махаланобиса, χ^2-критерий

Ссылка для цитирования: Усольцев Д.А. Вероятностный метод матричной кластеризации с априорным распределением признаков для формирования несмещенной контрольной группы // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 999–1001. doi: 10.17586/2226-1494-2025-25-5-999-1001

Probabilistic matrix clustering with feature priors for unbiased control selection

Dmitrii A. Usoltsev✉

Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA
ITMO University, Saint Petersburg, 197101, Russian Federation

dusoltsev.27@gmail.com✉, https://orcid.org/0000-0001-8072-310X

Abstract

We propose a probabilistic matrix-clustering method that leverages a prior distribution of features and dimensionality reduction (Singular Value Decomposition, SVD). The approach identifies, within a large control pool, a cluster statistically comparable to the test cohort, thereby reducing systematic bias in downstream comparative analyses. We show that the method correctly selects control groups in scenarios where standard nearest-neighbor matching produces false positives. The method has been used to construct control groups in studies based on the Russian Biobank at the Almazov National Medical Research Centre (Ministry of Health of the Russian Federation).

Keywords

matrix clustering, SVD, prior (feature) distribution, Mahalanobis distance, χ^2-criterion

For citation: Usoltsev D.A. Probabilistic matrix clustering with feature priors for unbiased control selection. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 999–1001 (in Russian). doi: 10.17586/2226-1494-2025-25-5-999-1001

В исследованиях по типу «случай-контроль» требуется корректно сопоставить тестовую группу пациентов (с заболеванием/фенотипом) с контрольной группой большой размерности [1]. Несовпадение распределений ключевых факторов между группами приводит к систематическим ошибкам при сравнительном анализе [2]. Распространенные методы подбора контрольной группы — ближайшие соседи по евклидову расстоянию и по расстоянию Махаланобиса [3] — в условиях высокой размерности часто дают сбои: возвращают ограниченное число сопоставимых наблюдений (образцов) и часто подбирают нерелевантные (ложноположительные) наблюдения. В настоящей работе предлагается формализовать подбор контрольной группы как вероятностную задачу на матрицах признаков $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m}$ и $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m}$, где $\mathbf{A}_1$ — тестовая матрица; $\mathbf{A}_2$ — контрольная матрица; n_1 и n_2 — число наблюдений в матрицах $\mathbf{A}_1$ и $\mathbf{A}_2$ соответственно; m — число характеристик.

Пусть требуется найти подмножество $\hat{S}_2 \subseteq \mathbf{A}_2$ такое, чтобы проекции тестовой и выбранной контрольной групп не отличались по ключевым статистикам и корреляциям в общем пространстве признаков, тем самым минимизируя смещения при последующем сравнительном анализе.

На первом этапе задача сводится к приведению двух наборов данных — тестового и контрольного — к единому признаковому пространству. Для этого проводится стандартизация признаков. В тестовом наборе $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m}$ вычисляются средние значения по каждому столбцу. Эти значения затем вычитаются из обеих матриц $\mathbf{A}_1$ и $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m}$, что обеспечивает выравнивание выборок по среднему значению признаков и устраняет глобальные линейные смещения.

Далее применяется такой метод снижения размерности, как сингулярное разложение матриц [4], позволяющий выделить наиболее информативные направления изменения данных. К матрице $\mathbf{A}_1$ применяется сингулярное разложение: $\mathbf{A}_1 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, где $\mathbf{U} \in \mathbb{R}^{n_1 \times n_1}$ — матрица левых сингулярных векторов; $\mathbf{\Sigma}$ — диагональная матрица сингулярных чисел; $\mathbf{V} \in \mathbb{R}^{m \times m}$ — матрица правых сингулярных векторов. В матрице $\mathbf{U}$ выбираются первые k столбцов, соответствующие наибольшим сингулярным числам. Полученная матрица $\hat{\mathbf{U}} \in \mathbb{R}^{n_1 \times k}$ определяет проекцию в подпространство наибольшей дисперсии. Проекции обоих наборов данных вычисляются следующим образом: $\hat{\mathbf{A}}_1 = \hat{\mathbf{U}}\mathbf{A}_1$ и $\hat{\mathbf{A}}_2 = \hat{\mathbf{U}}\mathbf{A}_2$. Каждые строки матриц $\hat{\mathbf{A}}_1$ и $\hat{\mathbf{A}}_2$ теперь представляют наблюдение в новом пространстве признаков размерности k общем для обеих выборок. Это обеспечивает сопоставимость полученных наборов данных.

После проекции на пространство главных компонент $\hat{\mathbf{A}}_1$, компоненты тестовой выборки рассматриваются как реализация случайной величины, подчиняющейся многомерному нормальному распределению. Оцениваются вектор математических ожиданий $\mathbf{m} \in \mathbb{R}^k$ и ковариационная матрица $\mathbf{\Sigma} \in \mathbb{R}^{k \times k}$, задающие априорное распределение признаков: $\hat{\mathbf{A}}_1 \sim \mathcal{N}(m, \mathbf{\Sigma})$.

Далее формулируется вероятностный критерий принадлежности. Для каждого объекта $x \in \hat{\mathbf{A}}_2$ вычисляется расстояние Махаланобиса как мера правдоподобия его принадлежности к распределению тестовой выборки: $D_M^2(x) = (x - m)^T\mathbf{\Sigma}^{-1}(x - m)$. Таким образом, каждый объект из контрольной выборки получает количественную оценку степени принадлежности к априорному распределению. В отличие от детерминированного отнесения объектов к кластерам, используется вероятностный критерий: объект включается в кластер $\hat{S}_2$, если значение $D_M(x)$ не превышает порог, соответствующий 95%-квантилю распределения χ^2 с k степенями свободы. Это соответствует включению в доверительную область уровня 0,95 для многомерного нормального распределения. Таким образом, кластер $\hat{S}_2 \subseteq \mathbf{A}_2$ формируется как $\hat{S}_2 = \{x \in \mathbf{A}_2 | D_M(\hat{\mathbf{U}}^T x) \le \chi^2_{k,1-\alpha}\}$, где α — уровень значимости. Неравенство означает, что x принадлежит 95%-ной доверительной области (при $\alpha = 0{,}05$) распределения $\mathcal{N}(m, \mathbf{\Sigma})$, которая в k-мерном пространстве имеет форму эллипсоида (в двумерном случае — эллипса).

Для воспроизводимой оценки предложенного метода было проведено два эксперимента. В эксперименте 1 тестовая и контрольная выборки были симулированы из одного многомерного нормального распределения, в эксперименте 2 — из разных. Предполагается, что практически все образцы из контрольной выборки будут отобраны предлагаемым методом в эксперименте 1, и никакие из образцов не будут отобраны во эксперименте 2. Образцы, отобранные в эксперименте 2, считаются ложноположительными, так как относятся к другому распределению.

В эксперименте 1 была промоделирована тестовая матрица $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m}$, где $n_1 = 200$ и $m = 30$, как выборка из многомерного нормального распределения $\mathcal{N}(0, \mathbf{\Sigma})$. В качестве матрицы ковариаций $\mathbf{\Sigma}$ с параметром корреляции Пирсона r [5] была использована матрица Тёплица [6] размером $m \times m$, где каждый элемент равен $\mathbf{\Sigma}_{ij} = r^{|i-j|}$. Такая структура ковариаций была выбрана для того, чтобы промоделировать корреляции, характерные для реальных данных. Контрольная матрица $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m}$, $n_2 = 20\,000$ генерировалась из того же распределения, что и матрица $\mathbf{A}_1$. В эксперименте 2 тестовый набор был тот же, а контрольная матрица смещалась на константу равную 2,7 относительно тестового набора в пространстве главных компонент, которая была выбрана экспериментально. Затем в обоих экспериментах был применен предлагаемый метод с целью поиска в матрице $\mathbf{A}_2$ контрольной подгруппы. Декомпозиция Singular Value Decomposition (SVD) была применена к матрице $\mathbf{A}_1$ и обе матрицы проектировались в k-мерное подпространство ($k = 9$), объясняющее более 70 % дисперсии. В этом пространстве для тестовой группы оценивались вектор средних $\boldsymbol{\mu}$ и ковариационная матрица $\mathbf{\Sigma}$, и предлагаемый метод отбирал все наблюдения $x \in \mathbf{A}_2$, удовлетворяющие вероятностному критерию $D_M^2(x) = (x - m)^T\mathbf{\Sigma}^{-1}(x - m) \le \chi^2_{k,1-\alpha}$, где $\alpha = 0{,}05$.

В качестве метода для сравнения надежности отбора контрольной группы был использован метод ближайшего соседа по расстоянию Махаланобиса в исходном пространстве признаков. По объединенным данным $[\mathbf{A}_1; \mathbf{A}_2]$ строилась матрица попарных расстояний Махаланобиса, после чего выполнялось сопоставление жадным алгоритмом каждого тестового наблюдения с

ближайшим доступным контрольным наблюдением по формуле «один к одному».

Предлагаемый метод показал более корректные результаты по сравнению с методом ближайшего соседа. В эксперименте 1 предлагаемый метод сформировал широкий набор (N = 19 744) статистически сопоставимых контрольных наблюдений, тогда как метод ближайшего соседа вернул число контрольных наблюдений равное числу тестовых наблюдений (N = 200). В эксперименте 2 предлагаемый метод не включил ни одного наблюдения из контрольной группы, избегая ложноположительных совпадений. Напротив, метод ближайшего соседа ошибочно отобрал N = 200 контрольных наблюдений, формируя ложные пары с тестовой выборкой.

В результате следует, что в настоящей работе был представлен метод формирования несмещенной контрольной выборки для исследований по типу «случай-контроль» [1]. Комбинация SVD и вероятностного отбора по Махаланобису с χ^2-критерием обеспечивает статистическую сопоставимость групп и улучшает корректность выводов. Предлагаемый метод внедрен в исследования, проводимые в Национальном медицинском исследовательском центре имени В.А. Алмазова. В частности, он позволил отобрать контрольную группу из медико-генетической базы данных для сравнительного исследования межпоколенческих эффектов потомков жителей блокадного Ленинграда [7].

Литература

1. Artomov M., Loboda A.A., Artyomov M.N., Daly M.J. Public platform with 39,472 exome control samples enables association studies without genotype sharing // Nature Genetics. 2024. V. 56. N 2. P. 327–335. https://doi.org/10.1038/s41588-023-01637-y
2. Pearce N. Analysis of matched case-control studies // BMJ Online. 2016. V. 352. P. i969. https://doi.org/10.1136/bmj.i969
3. Ghosh A., Ghosh A.K., SahaRay R., Sarkar S. Classification using global and local Mahalanobis distances // Journal of Multivariate Analysis. 2025. V. 207. P. 105417. https://doi.org/10.1016/j.jmva.2025.105417
4. Brunton S.L., Kutz J.N. Singular Value Decomposition (SVD) // Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. 2019. P. 3–46. https://doi.org/10.1017/9781108380690.002
5. Rovetta A. Raiders of the lost correlation: a guide on using pearson and spearman coefficients to detect hidden correlations in medical sciences // Cureus. 2020. V. 12. N 11. P. e11794. https://doi.org/10.7759/cureus.11794
6. Wang Z., Li G., Hu F., Chi N. Toeplitz concatenated matrix aided ICA algorithm for super-Nyquist multiband CAP VLC systems // Optics Express. 2020. V. 28. N 20. P. 29876–29894. https://doi.org/10.1364/OE.404925
7. Tolkunova K., Usoltsev D., Moguchaia E., Boyarinova M., Kolesova E., Erina A., et al. Transgenerational and intergenerational effects of early childhood famine exposure in the cohort of offspring of Leningrad Siege survivors // Scientific Reports. 2023. V. 13. N 1. P. 11188. https://doi.org/10.1038/s41598-023-37119-8

References

1. Artomov M., Loboda A.A., Artyomov M.N., Daly M.J. Public platform with 39,472 exome control samples enables association studies without genotype sharing. *Nature Genetics*, 2024, vol. 56, no. 2, pp. 327–335. https://doi.org/10.1038/s41588-023-01637-y
2. Pearce N. Analysis of matched case-control studies. *BMJ Online*, 2016, vol. 352, pp. i969. https://doi.org/10.1136/bmj.i969
3. Ghosh A., Ghosh A.K., SahaRay R., Sarkar S. Classification using global and local Mahalanobis distances. *Journal of Multivariate Analysis*, 2025, vol. 207, pp. 105417. https://doi.org/10.1016/j.jmva.2025.105417
4. Brunton S.L., Kutz J.N. Singular Value Decomposition (SVD). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, 2019, pp. 3–46. https://doi.org/10.1017/9781108380690.002
5. Rovetta A. Raiders of the lost correlation: a guide on using pearson and spearman coefficients to detect hidden correlations in medical sciences. *Cureus*, 2020, vol. 12, no. 11, pp. e11794. https://doi.org/10.7759/cureus.11794
6. Wang Z., Li G., Hu F., Chi N. Toeplitz concatenated matrix aided ICA algorithm for super-Nyquist multiband CAP VLC systems. *Optics Express*, 2020, vol. 28, no. 20, pp. 29876–29894. https://doi.org/10.1364/OE.404925
7. Tolkunova K., Usoltsev D., Moguchaia E., Boyarinova M., Kolesova E., Erina A., et al. Transgenerational and intergenerational effects of early childhood famine exposure in the cohort of offspring of Leningrad Siege survivors. *Scientific Reports*, 2023, vol. 13, no. 1, pp. 11188. https://doi.org/10.1038/s41598-023-37119-8

Автор

Усольцев Дмитрий Андреевич — старший научный сотрудник, Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57279360300, https://orcid.org/0000-0001-8072-310X, dusoltsev.27@gmail.com

Author

Dmitrii A. Usoltsev — Senior Researcher, Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57279360300, https://orcid.org/0000-0001-8072-310X, dusoltsev.27@gmail.com

Статья поступила в редакцию 15.08.2025
Одобрена после рецензирования 05.09.2025
Принята к печати 21.09.2025

Received 15.08.2025
Approved after reviewing 05.09.2025
Accepted 21.09.2025

Уважаемые подписчики научно-технической литературы!

Журнал выходит 6 раз в год.

На журнал «Научно-технический вестник информационных технологий, механики и оптики» можно оформить подписку в почтовых отделениях по объединенному каталогу «Пресса России», подписные индексы 47197 (полугодовая подписка),

а также online
по объединенному каталогу «Пресса России» и по каталогу «Пресса по подписке»,
подписные индексы Э47197 (полугодовая подписка)
и по электронному каталогу Почты России, подписной индекс ПС543.

Корпоративная подписка и подписка физических лиц возможна
по каталогу компаний «Урал-Пресс»,
подписные индексы 47197 (полугодовая подписка) и 70522 (годовая подписка).

Сведения о подписке можно уточнить в редакции журнала по адресу:

Санкт-Петербург, ул. Ломоносова, д.9, литера А, комн. 2136.

Тел.: +7(812) 480 02 75