НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2025 Том 25 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2025 Vol. 25 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-1-128-139

Detection of L_0-optimized attacks via anomaly scores distribution analysis

Dmitry A. Esipov[1]✉, Mark I. Basov[2], Alyona D. Kletenkova[3]

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

[1] some1else.d.ma@gmail.com✉, https://orcid.org/0000-0003-4467-5117
[2] basovmark@gmail.com, https://orcid.org/0009-0000-0844-6881
[3] alyonka8855@gmail.com, https://orcid.org/0009-0001-8148-6764

Abstract

The spread of artificial intelligence and machine learning is accompanied by an increase in the number of vulnerabilities and threats in systems implementing such technologies. Attacks based on malicious perturbations pose a significant threat to such systems. Various solutions have been developed to protect against them, including an approach to detecting L_0-optimized attacks on neural image processing networks using statistical analysis methods and an algorithm for detecting such attacks by threshold clipping. The disadvantage of the threshold clipping algorithm is the need to determine the value of the parameter (cutoff threshold) to detect various attacks and take into account the specifics of the data sets, which makes it difficult to apply in practice. This article describes a method for detecting L_0-optimized attacks on neural image processing networks through statistical analysis of the distribution of anomaly scores. To identify the distortion inherent in L_0-optimized attacks, deviations from the nearest neighbors and Mahalanobis distances are determined. Based on their values, a matrix of pixel anomaly scores is calculated. It is assumed that the statistical distribution of pixel anomaly scores is different for attacked and non-attacked images and for perturbations embedded in various attacks. In this case, attacks can be detected by analyzing the statistical characteristics of the distribution of anomaly scores. The obtained characteristics are used as predictors for training anomaly detection and image classification models. The method was tested on the CIFAR-10, MNIST and ImageNet datasets. The developed method demonstrated the high quality of attack detection and classification. On the CIFAR-10 dataset, the accuracy of detecting attacks (anomalies) was 98.43 %, while the binary and multiclass classifications were 99.51 % and 99.07 %, respectively. Despite the fact that the accuracy of anomaly detection is lower than that of a multiclass classification, the method allows it to be used to distinguish fundamentally similar attacks that are not contained in the training sample. Only input data is used to detect and classify attacks, as a result of which the proposed method can potentially be used regardless of the architecture of the model or the presence of the target neural network. The method can be applied for detecting images distorted by L_0-optimized attacks in a training sample.

Keywords

artificial neural network, image processing, adversarial attack, attack detection, pseudonorm L_0, malicious perturbation, statistical analysis, anomaly score

For citation: Esipov D.A., Basov M.I., Kletenkova A.D. Detection of L_0-optimized attacks via anomaly scores distribution analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 1, pp. 128–139. doi: 10.17586/2226-1494-2025-25-1-128-139

УДК 004.056

Обнаружение неконвенциональных пиксельных атак посредством статистического анализа распределения оценок аномальности

Дмитрий Андреевич Есипов[1]✉, Марк Игоревич Басов[2], Алёна Дмитриевна Клетенкова[3]

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

[1] some1else.d.ma@gmail.com✉, https://orcid.org/0000-0003-4467-5117
[2] basovmark@gmail.com, https://orcid.org/0009-0000-0844-6881
[3] alyonka8855@gmail.com, https://orcid.org/0009-0001-8148-6764

Аннотация

Введение. Распространение искусственного интеллекта и методов машинного обучения сопровождается увеличением количества уязвимостей и угроз в системах, реализующих подобные технологии. Значительную опасность для таких систем представляют атаки на основе вредоносных возмущений. Для защиты от них разработаны различные решения, к числу которых относятся подход к обнаружению неконвенциональной пиксельной атаки на нейронные сети обработки изображений методами статистического анализа и алгоритм обнаружения таких атак посредством отсечения по порогу. Недостатком алгоритма отсечения по порогу является необходимость определения значения параметра (порога отсечения) для обнаружения различных атак и учета специфики наборов данных, что затрудняет его применение на практике. В работе изложен метод обнаружения неконвенциональных пиксельных атак на нейронные сети обработки изображений посредством статистического анализа распределения оценок аномальности. **Метод.** Для выявления искажения, свойственного неконвенциональным пиксельным атакам, определяются отклонения от ближайших соседей и расстояния Махаланобиса. По их значениям вычисляется матрица оценок аномальности пикселов изображения. Предполагается, что статистическое распределение оценок аномальности пикселов различно для атакованных и неатакованных изображений и для возмущений, встраиваемых при различных атаках. В этом случае атаки могут быть обнаружены посредством анализа статистических характеристик распределения оценок аномальности. Полученные характеристики используются в качестве предикторов для обучения моделей обнаружения аномалий и классификации изображений. **Основные результаты.** Апробация метода выполнена на наборах данных CIFAR-10, MNIST и ImageNet. Разработанный метод продемонстрировал высокое качество обнаружения и классификации атак. На наборе данных CIFAR-10 точность (accuracy) обнаружения атак (аномалий) составила 98,43 %, а бинарной и многоклассовой классификаций — 99,51 % и 99,07 % соответственно. **Обсуждение.** Несмотря на то, что точность обнаружения аномалий ниже аналогичного показателя многоклассовой классификации, предложенный метод позволяет успешно применять его для распознавания принципиально схожих атак, не содержащихся в обучающей выборке. Для обнаружения и классификации атак используются только входные данные, в результате чего предложенный метод потенциально может быть использован независимо от архитектуры модели или наличия целевой нейронной сети. Метод может быть рекомендован для обнаружения изображений, искаженных неконвенциональными пиксельными атаками в обучающей выборке.

Ключевые слова

искусственная нейронная сеть, обработка изображений, состязательная атака, обнаружение атак, вредоносное возмущение, псевдонорма возмущения L_0, статистический анализ, оценка аномальности

Ссылка для цитирования: Есипов Д.А., Басов М.И., Клетенкова А.Д. Обнаружение неконвенциональных пиксельных атак посредством статистического анализа распределения оценок аномальности // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 1. С. 128–139 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-1-128-139

Introduction

According to previous work [1, 2], high efficiency in solving various applied tasks has led to an increase in the spread of Machine Learning (ML) and Artificial Intelligence (AI). However, along with the growing popularity of AI and ML, the number of vulnerabilities and threats in systems implementing these technologies has also increased.

Due to the urgency of model evasion and backdoor, ML model threats through these attacks based on malicious perturbations, various approaches and methods have been developed [2, 3–7]. In previous work [2], an approach was proposed to detect L_0-optimized attacks on neural networks of image processing using statistical analysis methods, and a detection method was developed based on the approach. This method leads to the following disadvantages: different values of the cutoff threshold for different attacks, the need to classify attacks in order to detect perturbations. The elimination of indications is possible by analyzing the statistical distribution of the obtained pixel estimates of anomalies and their characteristics.

Proposed method

The main idea. The proposed method is based on the assumption that the pixel anomaly scores matrices obtained from the preprocessing stage [2] have different frequency distributions for attacked [8–10] and clean (non-attacked) images. Then the task of detecting and classifying images can be shifted to the task of classifying distributions.

Examples of obtained pixel anomaly scores matrices [2] frequency distribution graphs for different classes of images are shown in Fig. 1.

According to Fig. 1, frequency distributions of clean and attacked images are similar, but have the following differences:

— One-pixel attack is characterized by the presence of a single value corresponding to a perturbed pixel significantly exceeding the rest;
— multi-pixel attacks (Jacobian Saliency Map Attack, JSMA; Localized and Visible Adversarial Noise, LaVAN) are characterized by a set of values that are out of the distribution characteristic of a clean image.

Then it is possible to detect and classify an attack based on the difference in distributions.

Considered statistical characteristics. The difference in distributions can be identified as a difference in their statistical characteristics. The list of considered statistical characteristics is given in Table 1.

Along with the known statistical characteristics, outliers beyond 5σ, outliers beyond 7σ, outliers beyond 3 interquantile range, outliers beyond 6 interquantile range were introduced and considered.

Anomaly detection. The detection of attacks can be performed as the detection of deviations of the listed

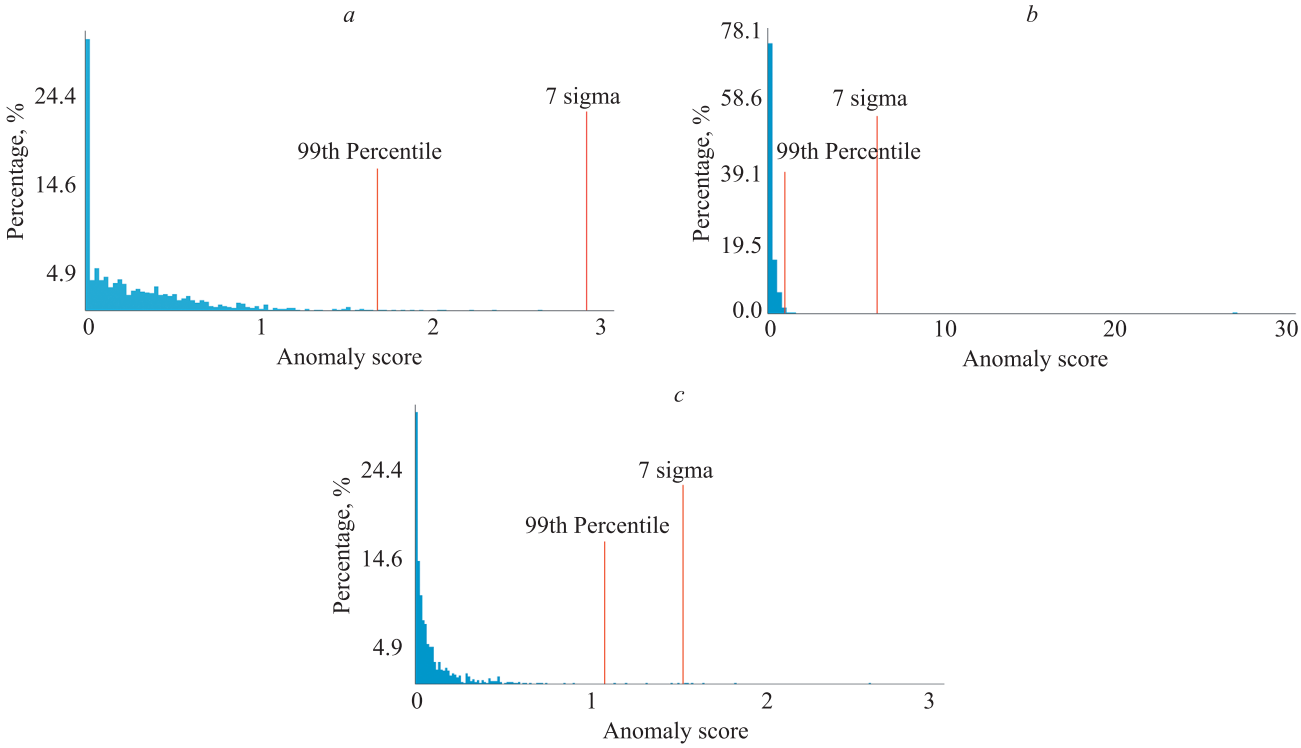

Fig. 1. Frequency distribution graphs: clean image (*a*); One-pixel attack (*b*); JSMA (*c*)

statistical characteristics from the values inherent in clean images (anomaly detection). Various types of unsupervised ML algorithms have been considered, including Support Vector Machine (SVM), density-based and tensor-based methods.

SVM includes two unsupervised ML algorithms: One-Class Support Vector Machine (OCSVM) [11, 12] and Support Vector Data Description (SVDD) [11, 13]. Since SVDD calculates a hypersphere that includes norm objects (points), it can be attributed to tensor-based methods.

Table 1. Considered statistical characteristics

Name	Abbreviation in the text	Description
Mean, μ	mean	The arithmetic mean of obtained values
Median	median	A value that divides an ordered list of values into two equal parts
Max	max	The maximum value in the image
Range	range	The difference between the maximum and minimum values
Variance	var	A measure of the spread of values relative to the mean
Standard deviation, σ	std	The square root of the variance
25th quantile	25q	The value below which 25 % of the data is located
75th quantile	75q	The value below which 75 % of the data is located
95th quantile	95q	The value below which 95 % of the data is located
97th quantile	97q	The value below which 97 % of the data is located
99th quantile	99q	The value below which 99 % of the data is located
Interquantile range	iqr	The difference between the 75th and 25th quantiles
Coefficient of variation, c_v	cv	The ratio of the standard deviation to the mean
Outliers beyond 3σ	3sigma_cnt	The number of values exceeding 3 standard deviations from the mean
Outliers beyond 5σ	5sigma_cnt	The number of values exceeding 5 standard deviations from the mean
Outliers beyond 7σ	7sigma_cnt	The number of values exceeding 7 standard deviations from the mean
Outliers beyond 3 interquantile ranges	3iqr_cnt	The number of values beyond 3 interquantile ranges from the 75q
Outliers beyond 6 interquantile ranges	6iqr_cnt	The number of values beyond 6 interquantile ranges from the 75q
Skewness	skew	A measure of the asymmetry of the data distribution relative to its mean
Kurtosis	kurt	A measure of the severity of the peak of the data distribution

Following density-based ML algorithms were considered: *k*-Nearest Neighbors, Local Outlier Factor and Isolation Forest (IF). According to a comparative analysis of algorithms [14], the IF demonstrates higher efficiency when working with multidimensional data.

SVDD and Elliptic Envelope (EE) were considered from tensor-based ML algorithms [15–17]. Both algorithms are fundamentally similar, however, SVDD calculates a hypersphere, and EE calculates a multidimensional ellipsoid. However, the sphere is a special case of an ellipsoid, so EE was chosen.

Attack classification. For subsequent correction [2], the binary (single-pixel attack and JSMA) classification of detected attacks (anomalies) was also considered.

As an alternative, multiclass (clean, One-pixel attack and JSMA) image classification was also considered. In this case, anomaly detection is not required. The detection and classification of attacks are implemented by a single model. However, then it is possible to detect only those attacks that were present in the training sample of the model. Then it is impossible to detect attacks that are unknown at the time of training or missing from the sample.

The following ML algorithms are considered for model training: SVM [18] and Random Forest (RF) [19]. The selection of methods was performed similarly to the detection of anomalies. The Logistic Regression (LR) algorithm was also considered [20].

Design of the experiment

Attack algorithms. Three attack algorithms were chosen as L_0-optimized attacks: One-pixel attack [8], JSMA [9], and LaVAN [10].

Datasets used. Datasets from previous work were used to conduct the experiment [2]. In the current work, all the distorted images obtained were used.

The ImageNet[1] dataset was used to address the LaVAN attack. The publicly available part of the specified dataset contains 1,281,167 training, 50,000 validation and 100,000 test color images with an average size of 469 × 387 pixels corresponding to 1000 classes. The images used are 299 × 299 pixels in size. The area of the malicious patch is a rectangle of random size, covering no more than 10 % of the pixels of the image. 14,636 perturbed and the same count of clean images used in experiment.

All the data obtained were used to train and evaluate the model. The sets of perturbated images used in further experiments are available on GitHub[2].

Evaluation metrics. Because there is a class disbalance in the dataset being used, F1-score is selected as quality indicator of anomaly detection and binary classification. Accuracy was also calculated for comparison with analogues.

For the multiclass classification, similar metrices were selected as quality indicators: macro F1-score and accuracy.

Model development. The Scikit-learn [21] library of the Python programming language was chosen for model training. The dataset used was divided as follows:

— training sample — 80 %;
— test sample — 20 %.

Program code, trained models, and calculation results are available on GitHub[3].

Results and analysis

Selection of statistical characteristics. To assess the significance of statistical characteristics for model training, correlation matrices were constructed for the considered datasets. The calculation of the correlation matrix is possible through various methods [22–24]. According to the comparative analysis [25], Kendall's method was chosen. It is important to note that a low correlation between the parameters does not necessarily indicate independence between them. In this case dependence may be more complex.

ML models using various sets of predictors were trained and evaluated to assess the significance of statistical characteristics. This method was used for ambiguous situations:

— high correlation between parameters;
— low correlation with the target variable (flag).

Fig. 2 introduces a Kendall correlation matrix for the CIFAR-10 dataset. Correlation matrices for other considered datasets are available on GitHub[3].

According to the correlation matrix (Fig. 2), there is a strong relationship between some parameters. Each group of parameters with a correlation coefficient 0.75 or greater was tested using model training and evaluation. Statistical characteristics with correlation coefficient less than 0.25 with flag were tested in the similar manner.

The final sets of statistical characteristics for each dataset are given in Table 2.

Max, range, cv, 7sigma_cnt were selected for each dataset. At the same time, iqr, 3iqr_cnt, 6iqr_cnt were excluded for most datasets. All proposed characteristics demonstrated their significance and were chosen for at least one dataset.

The quality of anomaly detection and attack classification was tested using both the selected parameters and all the considered ones. Obtained quality indicators for both sets of statistical characteristics (selected and all) are available on GitHub[3].

Anomaly detection. Models were trained for each dataset based on the selected anomaly detection algorithms. Confusion matrices for detecting anomalies are shown in Fig. 3. The highest quality indicators for each model are shown in Table 3.

According to Table 3, the highest quality indicators on each of the considered datasets, except CIFAR-10, correspond to the IF algorithm. For CIFAR-10, the highest

[1] ImageNet. Available at: https://www.image-net.org/index.php, free access (accessed: 03.03.2024). http:// arxiv.org/abs/1711.04596 (accessed: 21.10.2024).

[2] GitHub. iNDm3802 / L0-optimized_attack_detection. Available at: https://github.com/iNDm3802/L0-optimized_attack_detection, free access (accessed: 02.10.2024).

[3] GitHub. iNDm3802 / L0-optimized_attack_detection_method: https://github.com/iNDm3802/L0-optimized_attack_detection_method, free access (accessed: 18.01.2024).

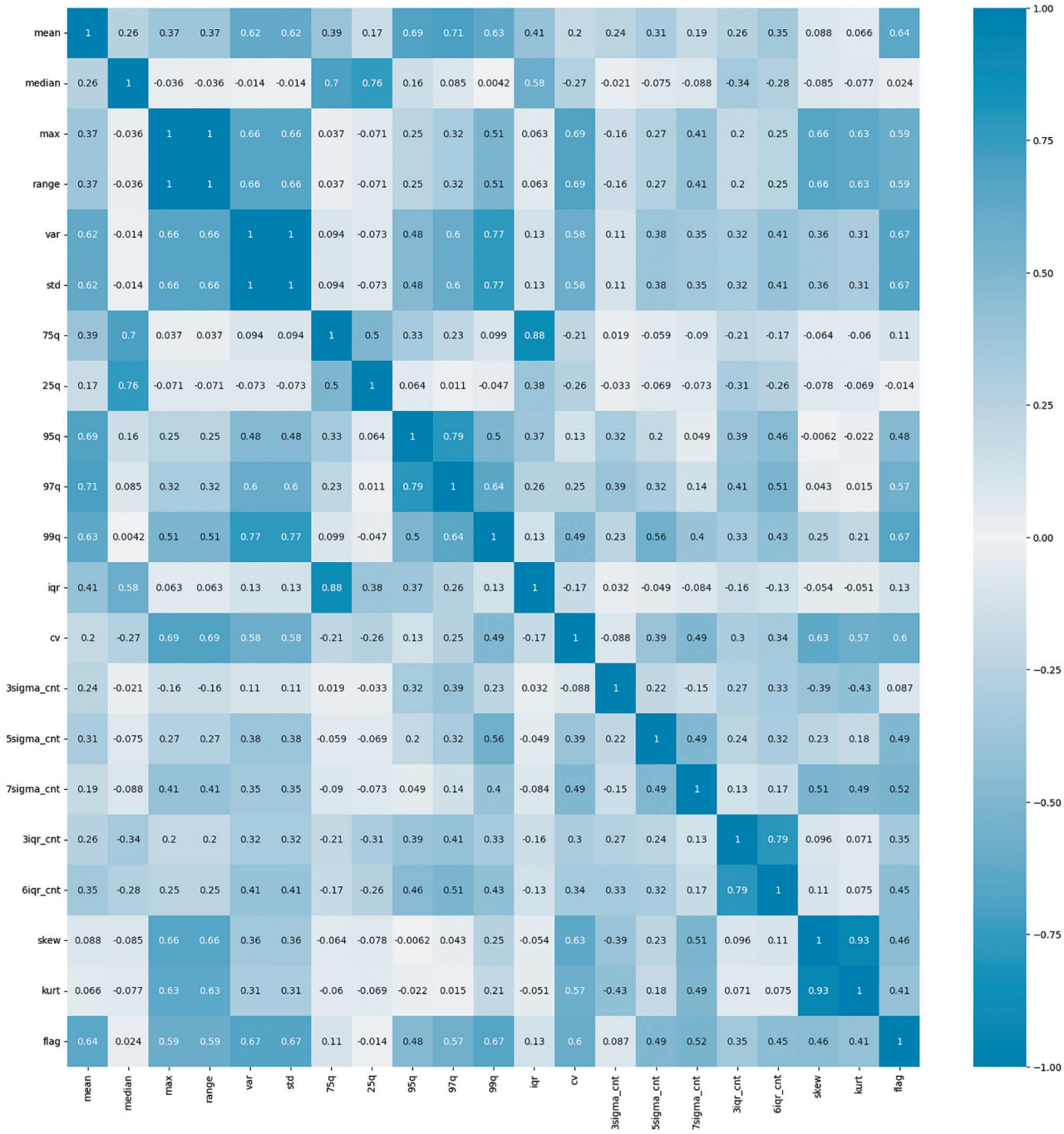

Fig. 2. Kendall correlation matrix – CIFAR-10

quality indicators correspond to the OCSVM algorithm. The proposed method demonstrates an anomaly detection accuracy of 98.43 % and F1-score of 97.71 % for CIFAR-10. There is a decrease in the detection quality for contrast images (MNIST) due to the limitations of the approach used.

Binary classification. Binary classification models (One-pixel attack and JSMA) were trained for each dataset

Table 2. Selected statistical characteristics

Dataset	CIFAR-10	CIFAR-10-G	MNIST	ImageNet
Selected characteristics	mean, max, rang, std, 75q, 95q, 99q, cv, 5sigma_cnt, 7sigma_cnt, skew	median, max, range, var, 25q, 95q, 99q, cv, 3sigma_cnt, 5sigma_cnt, 7sigma_cnt, kurt	mean, median, max, range, var, std, 25q, 97q, 99q, iqr, cv, 3sigma_cnt, 7sigma_cnt, skew, kurt	mean, median, max, range, var, std, 75q, 25q, 95q, 97q, 99q, iqr, cv, 3sigma, 5sigma, 7sigma, 3iqr, 6iqr, skew, kurt

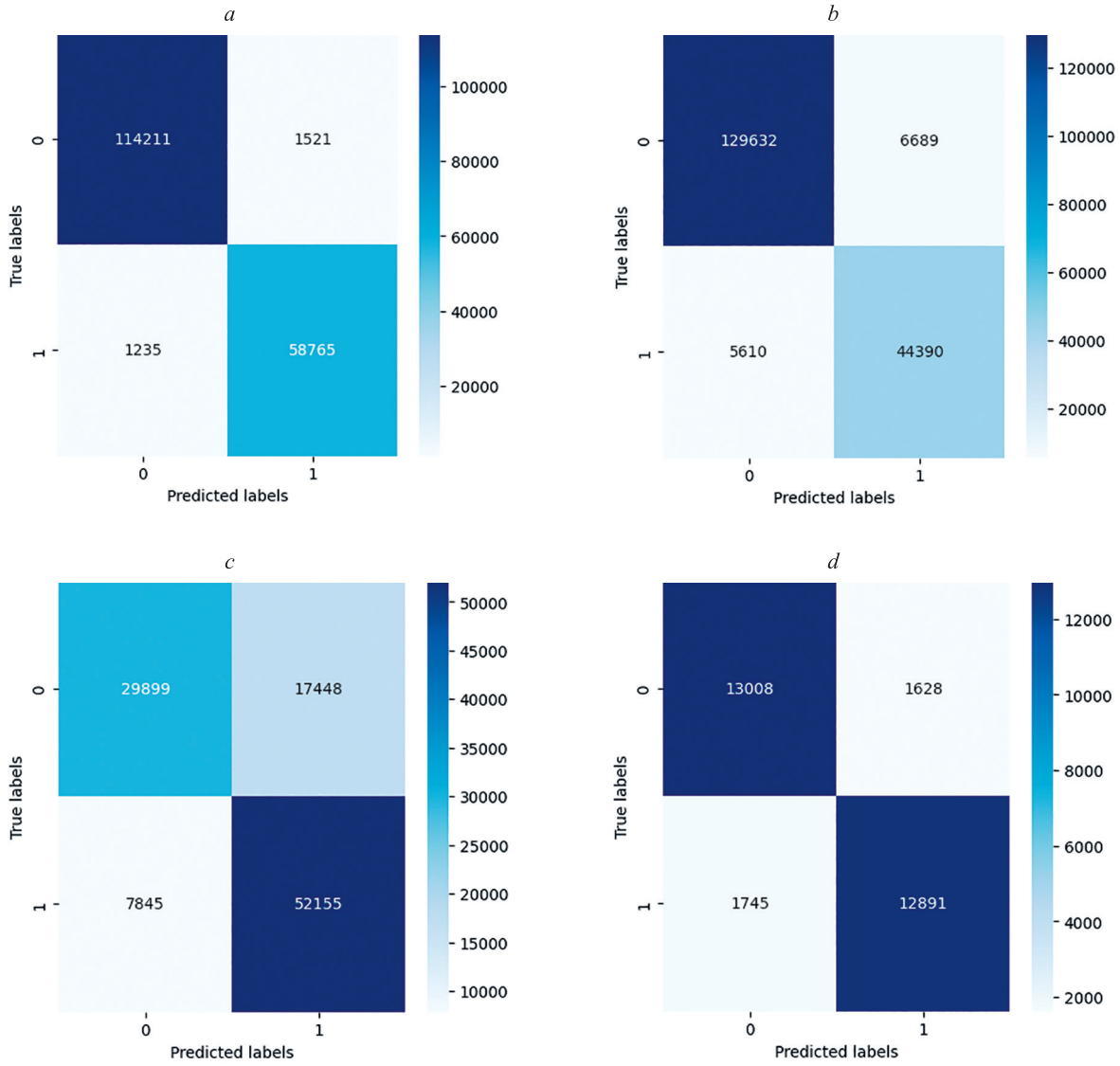

Fig. 3. Confusion matrices: CIFAR-10 (*a*); CIFAR-10-G (*b*); MNIST (*c*); ImageNet (*d*)

based on the selected algorithms. The confusion matrices of the binary classification are shown in Fig. 4. The highest quality indicators for each model are shown in Table 4.

According to Table 4, the highest quality indicators on each of the considered datasets correspond to the RF ML algorithm. The accuracy of the binary attack classification for the considered datasets varies from 95.31 % (MNIST) to 99.51 % (CIFAR-10), and the F1-score ranges from 97.27 % (MNIST) to 99.73 % (CIFAR-10). On each dataset the quality indicators, when using selected characteristics, are slightly less than when using all the parameters considered.

Multiclass classification. Multiclass classification models (clean, One-pixel attack and JSMA) were trained for each dataset based on the selected algorithms. The single-pixel attack is not relevant for ImageNet and is not represented in the sample for this dataset, so the specified dataset was not used for multiclass classification. The confusion matrices of the multiclass classification are shown in Fig. 5. The highest quality indicators for each model are given in Table 5.

According to Table 5, the highest quality indicators on each of the considered datasets correspond to the RF algorithm. There is a decrease in the classification quality for contrast images (MNIST) due to the limitations of the approach used.

Performance evaluation. The estimation of the computational complexity of calculation of statistical characteristics is O(n), where n corresponds to the number of pixels of the image taking into account the number of color channels, that is, its shape. Model forward propagation performance depends on the parameters of the model as well as on the ML algorithm. For performance evaluation, models with the highest quality indicators of anomaly detection or classification were used.

Calculations were performed on the following hardware:

— CPU: Intel(R) Core (TM) i7-10700 CPU @ 2.90GHz, 2904 MHz, cores: 8, logical processors: 16;
— RAM: 32.0 GB.

Performance evaluation of the proposed method is shown in Table 6.

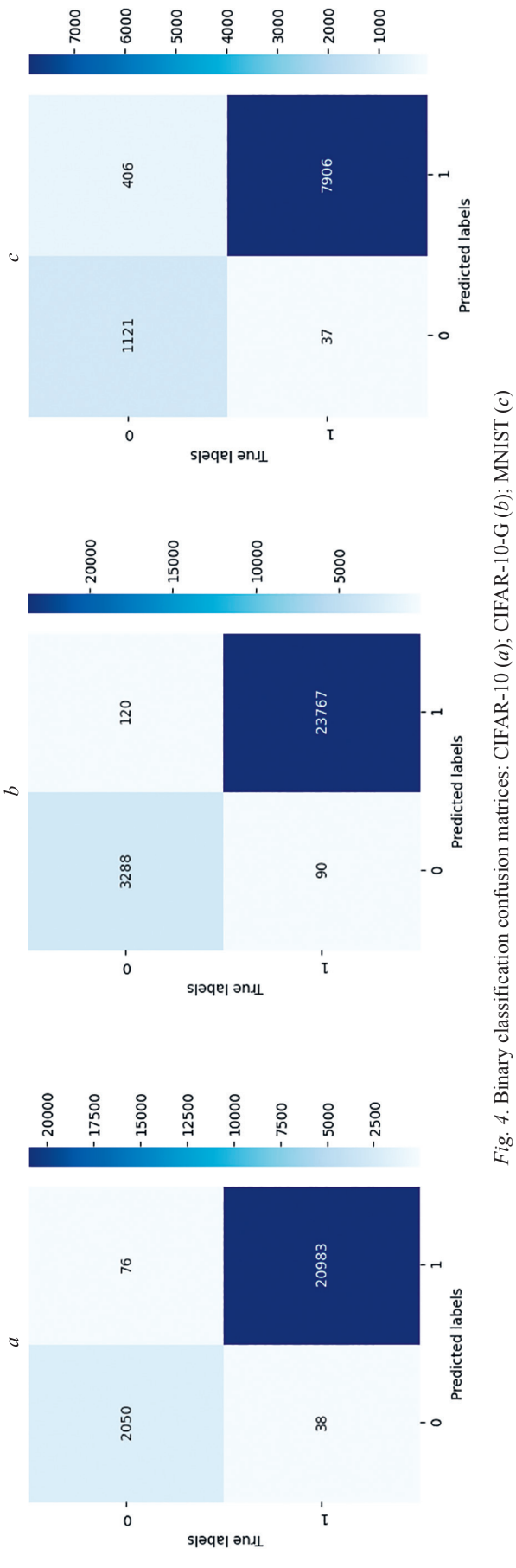

Fig. 4. Binary classification confusion matrices: CIFAR-10 (*a*); CIFAR-10-G (*b*); MNIST (*c*)

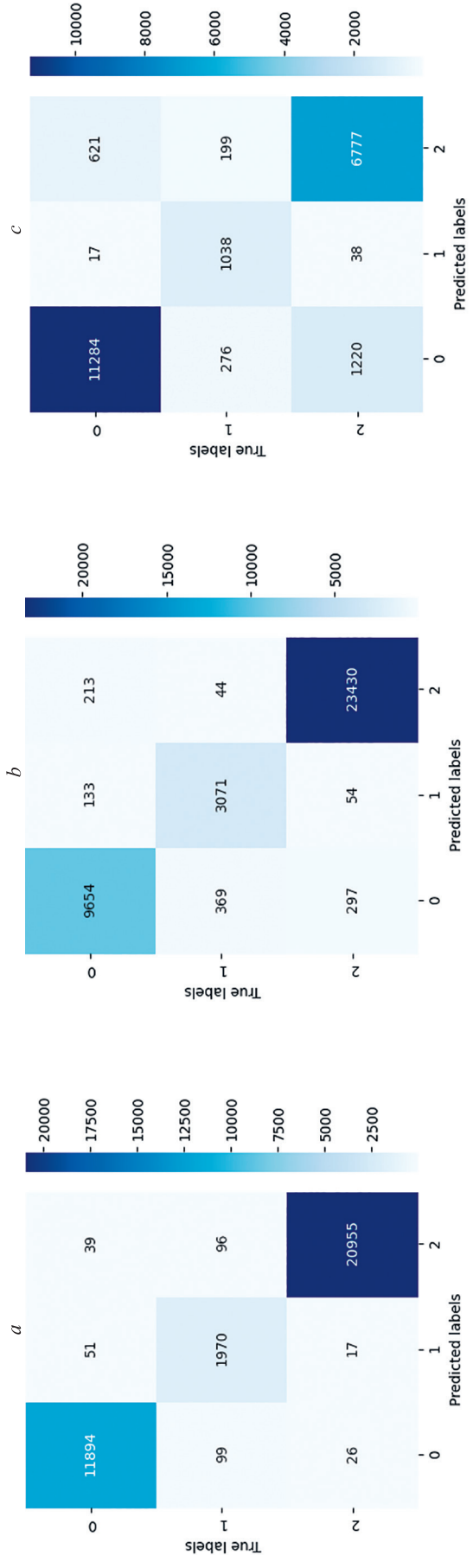

Fig. 5. Multiclass classification confusion matrices: CIFAR-10 (*a*); CIFAR-10-G (*b*); MNIST (*c*)

Table 3. Quality indicators for anomaly detection

Dataset	Algorithm (characteristics)	Parameters	Obtained quality indicators	
			Accuracy, %	F1-score, %
CIFAR-10	OCSVM (selected)	nu=0.02 kernel= ‘rbf’	**98.43**	**97.71**
	IF (selected)	n_estimators=10 random_state=3,802	95.44	92.87
	EE (selected)	random_state=3,802	96.38	94.43
CIFAR-10-G	OCSVM (all)	nu=0.14 kernel=‘rbf’	87.58	78.79
	IF (selected)	n_estimators=63 random_state=3,802	**93.40**	**87.83**
	EE (all)	random_state=3,802	91.86	85.57
MNIST	OCSVM (all)	nu=0.05 kernel= ‘rbf’	62.42	73.86
	IF (selected)	n_estimators=23 random_state=3,802	**76.44**	**80.48**
	EE (selected)	random_state=3,802	62.71	72.95
ImageNet	OCSVM (all)	nu=0.01 kernel=‘rbf’	50.64	66.72
	IF (all)	n_estimators=25 random_state=3,802	**88.48**	**88.43**
	EE (all)	random_state=3,802	85.62	86.23

According to Table 6, the method demonstrated attack detection speed from 0.19 to 65.51 images per second for ImageNet and MNIST, respectively, depending on their characteristics and parameters of a model. Similar patterns are observed for binary and multiclass classification.

Discussion

A comparative analysis of the developed L_0-optimized attack detection method with previous method [2] is shown in Table 7.

Table 4. Quality indicators for binary image classification

Dataset	Algorithm (characteristics)	Parameters	Obtained quality indicators	
			Accuracy, %	F1-score, %
CIFAR-10	SVM (all)	kernel=‘linear’	98.82	99.35
	RF (all)	all: n_estimators=12	**99.51**	**99.73**
	LR (all)	penalty=‘l2’, solver=‘ ’newton- cholesky’, random_state=3,802	98.79	99.33
CIFAR-10-G	SVM (all)	kernel=‘linear’	98.40	99.08
	RF (all)	n_estimators=69	**99.23**	**99.56**
	LR (all)	penalty=‘l2’, solver=‘liblinear’, random_state=3,802	98.20	98.97
MNIST	SVM (all)	kernel=‘linear’	93.74	96.36
	RF (all)	n_estimators=68	**95.31**	**97.27**
	LR (all)	penalty=‘l2’, solver=‘newton–cg’, random_state=3,802	93.25	96.09

Table 5. Quality indicators for multiclass image classification

Dataset	Algorithm (characteristics) [parameters]	Image class	Obtained quality indicators		
			F1-score, %	Accuracy, %	Macro F1-score, %
CIFAR-10	SVM (all) [kernel='linear']	Clean	98.78	98.45	96.22
		One-pixel	90.84		
		JSMA	99.05		
	RF (all) [n_estimators=56]	Clean	99.11	**99.07**	**97.47**
		One-pixel	93.74		
		JSMA	99.58		
	LR (all) [penalty='l2', solver='newtoncg', random_state=3,802]	Clean	98.69	98.38	96.07
		One-pixel	90.50		
		JSMA	99.01		
CIFAR-10-G	SVM (all) [kernel='linear']	Clean	91.19	94.44	92.17
		One-pixel	88.61		
		JSMA	96.72		
	RF (all) [n_estimators=96]	Clean	95.02	**97.02**	**94.95**
		One-pixel	91.10		
		JSMA	98.72		
	LR (all) [penalty='l2', solver='newton–cg', random_state=3,802]	Clean	89.97	93.64	90.83
		One-pixel	86.22		
		JSMA	96.31		
MNIST	SVM (all) [kernel='linear']	Clean	88.61	85.05	81.92
		One-pixel	76.51		
		JSMA	80.65		
	RF (all) [n_estimators=77]	Clean	91.36	**88.96**	**85.91**
		One-pixel	79.66		
		JSMA	86.71		
	LR (all) [penalty='l2', solver='newton–cg', random_state=3,802]	Clean	88.40	84.79	81.45
		One-pixel	75.39		
		JSMA	80.55		

Table 6. Performance evaluation of the L_0-optimized attack detection method

Task	Dataset	Count of images	Image shape		Time, s
			color channels	pixels	
Anomaly detection	CIFAR-10	10,000	3	32 × 32	547.53
	CIFAR-10-G		1	32 × 32	227.08
	MNIST		1	28 × 28	152.64
	ImageNet		3	299 × 299	51,911.74
Binary classification	CIFAR-10		3	32 × 32	576.26
	CIFAR-10-G		1	32 × 32	277.48
	MNIST		1	28 × 28	223.44
Multiclass classification	CIFAR-10		3	32 × 32	606.29
	CIFAR-10-G		1	32 × 32	299.09
	MNIST		1	28 × 28	209.33

Table 7. Comparative analysis of L_0-optimized attack detection methods

Method	Dataset	Attack	Accuracy, %	F1-score, %
Esipov D. [2]	CIFAR-10	Both One-pixel attack and JSMA	96.94	96.92
	MNIST		75.14	74.43
Developed (selected characteristics)	CIFAR-10	Both One-pixel attack and JSMA	98.43	97.71
	MNIST		76.44	80.48
	ImageNet	LaVAN	84.73	85.50
Developed (all characteristics)	CIFAR-10	Both One-pixel attack and JSMA	97.82	96.84
	MNIST		72.95	77.80
	ImageNet	LaVAN	88.48	88.43

The developed method demonstrates quality indicators comparable to analogues [2–7]. Since the developed method uses only input data to detect and classify L_0-optimized attacks, it can potentially be used regardless of the architecture of the model or the presence of a target neural network. In addition, the method allows detecting various L_0-optimized attacks (One-pixel attack and JSMA). Due to the use of anomaly detection, the method can also detect other fundamentally similar attacks that are not represented in the data sets used and are not considered in the current work.

Unlike the previous method [2], the current one does not have such limitations, as need for different algorithm parameters (cut-off threshold) for detecting different attacks and its parameters (gamma). The developed method also performs classification for further perturbation detection. The disadvantage of the developed method is a decline in attack detection and image classification quality on contrasting images. This disadvantage is due to the limitations of the approach used.

Conclusion

The proposed method allows detecting the fact of an attack based on L_0-optimized perturbation as well as the classification of specified attacks. The method demonstrates high accuracy and F1-score when detecting various L_0-optimized attacks. The use of anomaly detection allows detecting other similar attacks. The classification of attacks allows selecting the parameters of the algorithm to detect the perturbation introduced by these attacks. Since the developed method uses only input data to detect and classify L_0-optimized attacks, it can potentially be used regardless of the architecture of the model or the presence of a target neural network.

The direction of further work is to enhance the algorithm for detecting perturbed pixels by the use the statistical distribution of the obtained pixel anomaly scores and its characteristics along with the values of the obtained values. Another direction is to develop approaches, algorithms or methods for detecting other types of attacks based on malicious perturbation on image processing neural networks.

References

1. Esipov D.A., Buchaev A.Y., Kerimbay A., Puzikova Y.V., Saidumarov S.K., Sulimenko N.S., Popov I.Yu., Karmanovskiy N.S. Attacks based on malicious perturbations on image processing systems and defense methods against them. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 720–733. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-4-720-733
2. Esipov D.A. An approach to detecting L_0-optimized attacks on image processing neural networks via means of mathematical statistics. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 3, pp. 490–499. https://doi.org/10.17586/2226-1494-2024-24-3-490-499
3. Nguyen-Son H.Q., Thao T.P., Hidano S., Bracamonte V., Kiyomoto S., Yamaguchi R.S. Opa2d: One-pixel attack, detection, and defense in deep neural networks. *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–10. https://doi.org/10.1109/IJCNN52387.2021.9534332
4. Alatalo J., Sipola T., Kokkonen T. Detecting One-Pixel Attacks Using Variational Autoencoders. *Lecture Notes in Networks and Systems*, 2022, vol. 468, pp. 611–623. https://doi.org/10.1007/978-3-031-04826-5_60
5. Wang P., Cai Z., Kim D., Li W. Detection mechanisms of one-pixel attack. *Wireless Communications and Mobile Computing*, 2021, vol. 2021, no. 1, pp. 8891204. https://doi.org/10.1155/2021/8891204
6. Grosse K., Manoharan P., Papernot N., Backes M., McDaniel P. On the (statistical) detection of adversarial examples. *arXiv*, 2017, arXiv:1702.06280. https://doi.org/10.48550/arXiv.1702.06280

Литература

1. Есипов Д.А., Бучаев А.Я., Керимбай А., Пузикова Я.В., Сайдумаров С.К., Сулименко Н.С., Попов И.Ю., Кармановский Н.С. Атаки на основе вредоносных возмущений на системы обработки изображений и методы защиты от них // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 4. С. 720–733. https://doi.org/10.17586/2226-1494-2023-23-4-720-733
2. Esipov D.A. An approach to detecting L_0-optimized attacks on image processing neural networks via means of mathematical statistics // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2024. V. 24. N 3. P. 490–499. https://doi.org/10.17586/2226-1494-2024-24-3-490-499f
3. Nguyen-Son H.Q., Thao T.P., Hidano S., Bracamonte V., Kiyomoto S., Yamaguchi R.S. Opa2d: One-pixel attack, detection, and defense in deep neural networks // Proc. of the International Joint Conference on Neural Networks (IJCNN). 2021. P. 1–10. https://doi.org/10.1109/IJCNN52387.2021.9534332
4. Alatalo J., Sipola T., Kokkonen T. Detecting One-Pixel Attacks Using Variational Autoencoders // Lecture Notes in Networks and Systems. 2022. V. 468 P. 611–623. https://doi.org/10.1007/978-3-031-04826-5_60
5. Wang P., Cai Z., Kim D., Li W. Detection mechanisms of one-pixel attack // Wireless Communications and Mobile Computing. 2021. V. 2021. N 1. P. 8891204. https://doi.org/10.1155/2021/8891204
6. Grosse K., Manoharan P., Papernot N., Backes M., McDaniel P. On the (statistical) detection of adversarial examples // arXiv. 2017. arXiv:1702.06280. https://doi.org/10.48550/arXiv.1702.06280

7. Guo F., Zhao Q., Li X., Kuang X., Zhang J., Han Y., Tan Y.A. Detecting adversarial examples via prediction difference for deep neural networks. *Information Sciences*, 2019, vol. 501, pp. 182–192. https://doi.org/10.1016/j.ins.2019.05.084
8. Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019, vol. 23, no. 5, pp. 828–841. https://doi.org/10.1109/TEVC.2019.2890858
9. Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A. The limitations of deep learning in adversarial settings. *Proc. of the IEEE European symposium on security and privacy (EuroS&P)*, 2016, pp. 372–387. https://doi.org/10.1109/EuroSP.2016.36
10. Karmon D., Zoran D., Goldberg Y. Lavan: Localized and visible adversarial noise. *arXiv*, 2018, arXiv:1801.02608. https://doi.org/10.48550/arXiv.1801.02608
11. Lampert C.H. Kernel methods in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 2009, vol. 4, no. 3, pp. 193–285. http://dx.doi.org/10.1561/0600000027
12. Bounsiar A., Madden M.G. One-class support vector machines revisited. *Proc. of the 5th International Conference on Information Science & Applications (ICISA)*, 2014, pp. 1–4. https://doi.org/10.1109/ICISA.2014.6847442
13. Tax D.M.J., Duin R.P.W. Support vector data description. *Machine Learning*, 2004, vol. 54, no. 1, pp. 45–66. https://doi.org/10.1023/B:MACH.0000008084.60811.49
14. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest. *Proc. of the 8th IEEE International Conference on Data Mining*, 2008, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17
15. Ji Y., Wang Q., Li X., Liu J. A survey on tensor techniques and applications in machine learning. *IEEE Access*. 2019, vol. 7, pp. 162950–162990. https://doi.org/10.1109/ACCESS.2019.2949814
16. Howard S. The Elliptical Envelope. *arXiv*, 2007, arXiv:math/0703048. https://doi.org/10.48550/arXiv.math/0703048
17. Ashrafuzzaman M., Das S., Jillepalli A.A., Chakhchoukh Y., Sheldon F.T. Elliptic envelope based detection of stealthy false data injection attacks in smart grid control systems. *Proc. of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 1131–1137. https://doi.org/10.1109/SSCI47803.2020.9308523
18. Hearst M.A., Dumais S.T., Osuna E., Platt J., Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their applications*, 1998, vol. 13, no. 4, pp. 18–28. https://doi.org/10.1109/5254.708428
19. Ho T.K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 1998, vol. 20, no. 8, pp. 832–844. https://doi.org/10.1109/34.709601
20. Wright R.E. Logistic regression. *Reading and understanding multivariate statistics. American Psychological Association*, 1995, pp. 217–244.
21. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2825–2830.
22. Sedgwick P. Pearson's correlation coefficient. *British Medical Journal*, 2012, vol. 345, pp. e4483. https://doi.org/10.1136/bmj.e4483
23. Abd Al-Hameeda K.A. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, 2022, vol. 13, no. 1, pp. 3249–3255. https://doi.org/10.22075/ijnaa.2022.6079
24. Abdi H. The Kendall rank correlation coefficient. *Encyclopedia of measurement and statistics. SAGE Publications*, 2007. vol. 2, pp. 508–510.
25. Xu W., Hou Y., Hung Y.S., Zou Y. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 2013, vol. 93, no. 1, pp. 261–276. https://doi.org/10.1016/j.sigpro.2012.08.005

7. Guo F., Zhao Q., Li X., Kuang X., Zhang J., Han Y., Tan Y.A. Detecting adversarial examples via prediction difference for deep neural networks // Information Sciences. 2019. V. 501. P. 182–192. https://doi.org/10.1016/j.ins.2019.05.084
8. Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. 2019. V. 23. N 5. P. 828–841. https://doi.org/10.1109/TEVC.2019.2890858
9. Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A. The limitations of deep learning in adversarial settings // Proc. of the IEEE European symposium on security and privacy (EuroS&P). 2016. P. 372–387. https://doi.org/10.1109/EuroSP.2016.36
10. Karmon D., Zoran D., Goldberg Y. Lavan: Localized and visible adversarial noise // arXiv. 2018. arXiv:1801.02608. https://doi.org/10.48550/arXiv.1801.02608
11. Lampert C.H. Kernel methods in computer vision // Foundations and Trends in Computer Graphics and Vision. 2009. V. 4. N 3. P. 193–285. http://dx.doi.org/10.1561/0600000027
12. Bounsiar A., Madden M.G. One-class support vector machines revisited // Proc. of the 5th International Conference on Information Science & Applications (ICISA). 2014. P. 1–4. https://doi.org/10.1109/ICISA.2014.6847442
13. Tax D.M.J., Duin R.P.W. Support vector data description. Machine Learning. 2004. V. 54. N 1. P. 45–66. https://doi.org/10.1023/B:MACH.0000008084.60811.49
14. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest // Proc. of the 8th IEEE International Conference on Data Mining. 2008. P. 413–422. https://doi.org/10.1109/ICDM.2008.17
15. Ji Y., Wang Q., Li X., Liu J. A survey on tensor techniques and applications in machine learning // IEEE Access. 2019. V. 7. P. 162950–162990. https://doi.org/10.1109/ACCESS.2019.2949814
16. Howard S. The Elliptical Envelope // arXiv. 2007. arXiv:math/0703048. https://doi.org/10.48550/arXiv.math/0703048
17. Ashrafuzzaman M., Das S., Jillepalli A.A., Chakhchoukh Y., Sheldon F.T. Elliptic envelope based detection of stealthy false data injection attacks in smart grid control systems // Proc. of the IEEE Symposium Series on Computational Intelligence (SSCI). 2020. P. 1131–1137. https://doi.org/10.1109/SSCI47803.2020.9308523
18. Hearst M.A., Dumais S.T., Osuna E., Platt J., Scholkopf B. Support vector machines // IEEE Intelligent Systems and their applications. 1998. V. 13. N 4. P. 18–28. https://doi.org/10.1109/5254.708428
19. Ho T.K. The random subspace method for constructing decision forests // IEEE transactions on pattern analysis and machine intelligence. 1998. V. 20. N 8. P. 832–844. https://doi.org/10.1109/34.709601
20. Wright R.E. Logistic regression // Reading and understanding multivariate statistics. American Psychological Association, 1995. P. 217–244.
21. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. 2011. V. 12. P. 2825–2830.
22. Sedgwick P. Pearson's correlation coefficient // British Medical Journal. 2012. V. 345. P. e4483. https://doi.org/10.1136/bmj.e4483
23. Abd Al-Hameeda K.A. Spearman's correlation coefficient in statistical analysis // International Journal of Nonlinear Analysis and Applications. 2022. V. 13. N 1. P. 3249–3255. https://doi.org/10.22075/ijnaa.2022.6079
24. Abdi H. The Kendall rank correlation coefficient // Encyclopedia of measurement and statistics. SAGE Publications, 2007. V. 2. P. 508–510.
25. Xu W., Hou Y., Hung Y.S., Zou Y. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models // Signal Processing. 2013. V. 93. N 1. P. 261–276. https://doi.org/10.1016/j.sigpro.2012.08.005

Authors

Dmitry A. Esipov — Assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57954958600, https://orcid.org/0000-0003-4467-5117, some1else.d.ma@gmail.com

Авторы

Есипов Дмитрий Андреевич — ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57954958600, https://orcid.org/0000-0003-4467-5117, some1else.d.ma@gmail.com

Mark I. Basov — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0000-0844-6881, basovmark@gmail.com
Alyona D. Kletenkova — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0001-8148-6764, alyonka8855@gmail.com

Басов Марк Игоревич — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0000-0844-6881, basovmark@gmail.com
Клетенкова Алёна Дмитриевна — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0001-8148-6764, alyonka8855@gmail.com

Received 02.09.2024
Approved after reviewing 04.01.2025
Accepted 24.01.2025

Статья поступила в редакцию 02.09.2024
Одобрена после рецензирования 04.01.2025
Принята к печати 24.01.2025