УДК 04.032.26

Г.Д. Дюдюн, М.А. Лапина, М.Г. Бабенко

Исследование новых сценариев состязательных атак на нейронные сети распознавания образов в контексте поиска новых методов защиты

Нейронные сети (НС) являются эффективным инструментом решения трудно формализуемых задач, что сделало их незаменимым инструментом для их решения. Однако методики информационной защиты в данной области всё ещё не имеют достаточного уровня защиты, что делает их уязвимыми для киберпреступников. В данной статье исследуются состязательные атаки на НС, их особенности, а также предлагается новая методика обнаружения состязательных атак.

Ключевые слова: нейронные сети, машинное обучение, информационная безопасность, состязательные атаки.
DOI: 10.21293/1818-0442-2025-28-1-114-118

В настоящее время для распознавания изображений используются глубокие нейронные сети (DNN). Глубокие нейронные сети восприимчивы к шуму во входных данных. Шум, незаметный для человеческого глаза, может привести к сбоям в работе глубоких нейронных сетей. Атаки, основанные на зашумлении обучающей выборки, называют состязательными атаками, а искаженные или специально сгенерированные для них данные – состязательными примерами. В современном мире состязательные атаки (Adversarial attacks) становятся все более распространенным и опасным явлением, в данной работе мы рассматриваем влияние состязательной атаки на сверточные нейронные сети и предлагаем новый метод их обнаружения.

Постановка задачи

На данный момент ученые выделяют три основных типа состязательных атак (рис. 1).

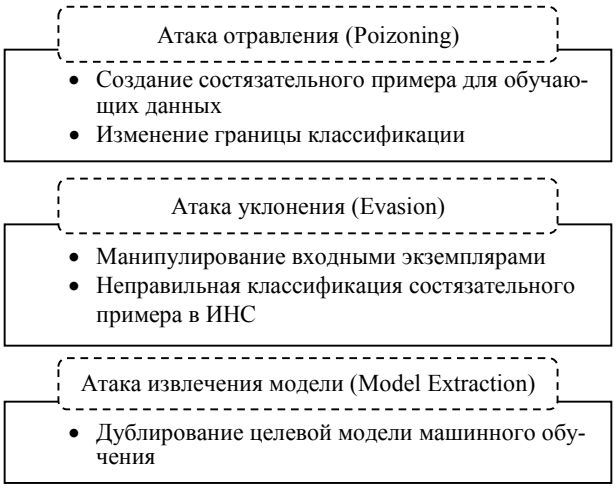


Рис. 1. Типы состязательных атак

В основе всех трех типов состязательных атак используется подход, основанный на внесении незначительных изменений в исходные данные, за счёт чего можно добиться существенного изменения результата классификации. Состязательные атаки стали активно изучаться в контексте различных задач машинного обучения и искусственного интеллекта [11].

В настоящее время существуют различные алгоритмы и модификации состязательных атак, включая FGSM (Fast Gradient Sign Method), DeepFool, C&W (Carlini & Wagner) и др., представленные в табл. 1.

Таблица 1

Методики проведения состязательных атак

Тип атаки	Название атаки
Атака уклонения	FGSM
	DeepFool
	JSMA
	PGD
	BIM
	Carlini & Wagner
Атака отравления	Feature collision
	SWM Poisoning
	Backdoor Attack
Атака извлечения модели	Knock off Nets
	MiFace
	Copycat CNN

В 2013 г. C.C. Szegedy et al. [1] доказали влияние незаметных глазу искажений на результат распознавания данных нейронной сетью.

I.J. Goodfellow, J. Shlens, C. Szegedy в 2014 г. исследуют влияние состязательных атак на модели машинного обучения и предлагают методы противодействия, являющиеся неактуальными к настоящему моменту [2].

В 2016 г. N. Papernot et al. [3] формализуют пространство злоумышленников против НС и представляют алгоритм создания состязательных примеров на основе сравнения входных и выходных данных НС.

В 2017 г. A. Madry et al. [4] представляют методику создания глубоких моделей, устойчивых к состязательным атакам. Они предлагают новый подход к обучению моделей, который предусматривает множество сценариев проведения атаки и способствует повышению устойчивости модели к ним. Минус данного подхода в его статичности: с появлением новых сценариев состязательных атак данный метод защиты будет нуждаться в улучшении и доработке.

Таким образом, решение проблемы противодействия состязательным атакам имеет важное значение

при обеспечении информационной безопасности систем машинного обучения. Разработка методов защиты является актуальной задачей из-за недостаточной эффективности существующих методик. Исходя из всего вышесказанного, особую актуальность приобретает решение задачи разработки методов обнаружения состязательных атак и защиты от них в условиях динамически изменяемой среды.

Предлагаемое решение

Большинство современных подходов к защите от состязательных атак основаны на повышении устойчивости нейронных сетей (НС) к состязательным атакам. Однако не менее важным является возможность своевременного устранения последствий атаки. Наша методика основана на гипотезе, что НС, которая была обучена на данных, подвергшихся искажению при состязательной атаке, будет совершать большее количество ошибок при распознавании неискаженных данных. При этом большинство ошибок будет совершено на правильных элементах, которые подвергались искажению при атаке. Таким образом, суть нашей методики заключается в проведении тестирования нейронной сети, на которую предварительно была совершена состязательная атака, подсчете «забракованных ответов» и определении характера искажения и самих искаженных данных для скорейшей минимизации последствий.

Моделирование

Для проведения моделирования нами была выбрана наиболее простая по своей структуре свёрточная нейронная сеть распознавания образов на основе базы данных изображений рукописных цифр MNIST.

Структурная схема моделирования предложенного метода представлена на рис. 2.

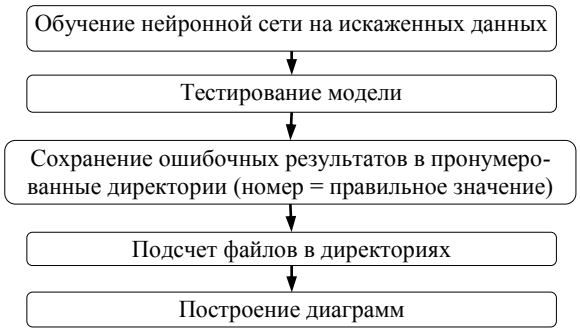


Рис. 2. Алгоритм проведения эксперимента на доказательство методики

По результатам подсчета количества элементов в каждой из конечных директорий, отмеченных цифрами соответствующих элементов, были построены диаграммы (рис. 3, 4).

Для диаграмм на рис. 3 по горизонтальной оси отложен номер директории, а по вертикальной – количество файлов в директории.

Из диаграммы на рис. 3 видно, что наибольшее количество ошибок совершено именно при распознавании изображений цифры «4», из чего неосведомленный об атаке наблюдатель сможет сделать вывод, что в результате атаки пострадала только 1 категория данных, и вовремя исправит ошибку. Таким образом, наша гипотеза верна, а значит, предлагаемая методика является применимой. В качестве проверки и окончательного доказательства теории необходимо также проверить гипотезу о том, что состязательная атака на любой элемент датасета окажет одинаковое влияние на общую работу нейросети. Для этого последовательно произведем атаку на каждый тип элементов и сверим показатели общей точности работы модели на тесте.

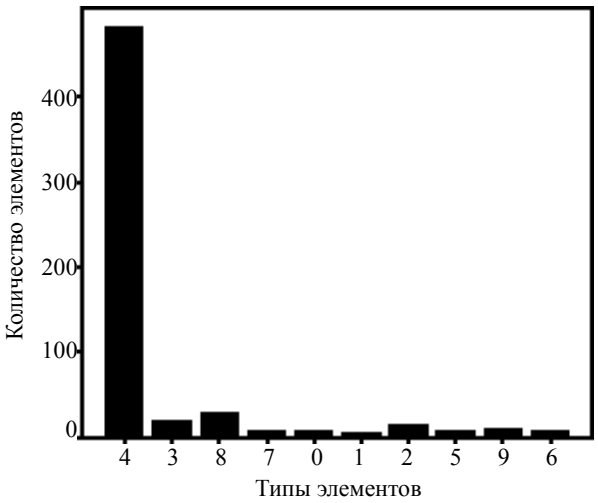


Рис. 3. Диаграмма результатов эксперимента (с элементом «4»)

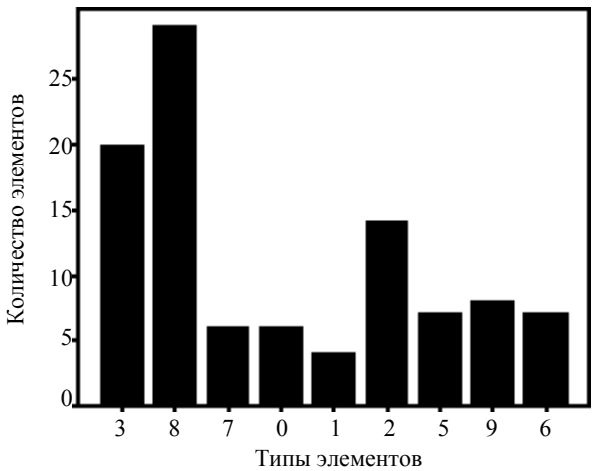


Рис. 4. Диаграмма результатов эксперимента (без элемента «4»)

Результаты замеров точности представлены в табл. 2. Названия строк – это атакуемые элементы, а названия столбцов – новые значения идентифицирующих меток, используемые для замены. Как видно из таблицы, значения точности различаются не более чем на 0,02%, следовательно, можно считать, что атаки оказали одинаковое воздействие на работу нейросети и что выбор элемента для атаки не влиял на результат основного эксперимента.

Таблица 2

Значения общей точности при атаках

	0	1	2	3	4	5	6	7	8	9
0	**0,980**	0,871	0,874	0,872	0,874	0,875	0,877	0,876	0,871	0,876
1	0,868	**0,980**	0,863	0,870	0,867	0,865	0,862	0,865	0,867	0,865
2	0,879	0,877	**0,980**	0,884	0,882	0,879	0,878	0,881	0,879	0,878
3	0,880	0,878	0,880	**0,980**	0,879	0,881	0,879	0,879	0,879	0,878
4	0,886	0,887	0,885	0,883	**0,980**	0,887	0,882	0,886	0,885	0,885
5	0,894	0,892	0,891	0,893	0,895	**0,980**	0,893	0,893	0,894	0,888
6	0,880	0,880	0,882	0,880	0,880	0,877	**0,980**	0,877	0,876	0,878
7	0,879	0,875	0,878	0,880	0,880	0,879	0,875	**0,980**	0,879	0,876
8	0,884	0,884	0,882	0,887	0,884	0,886	0,883	0,884	**0,980**	0,886
9	0,879	0,882	0,880	0,884	0,886	0,883	0,879	0,882	0,881	**0,980**

Методика образцов

Исходя из данных прошлого эксперимента, становится понятно, что атакованный элемент можно определить по резко возросшему количеству ошибок, совершаемых нейронной сетью при распознавании данного элемента на тесте. Однако данный подход обладает двумя существенными недостатками:

– Не все нейросети поддерживают возможность сравнения ответов нейронной сети с идентифицирующими метками элементов.

– Для предотвращения доступа злоумышленника к тестовым данным необходимо выделять часть данных для изолированного хранения.

С нашей точки зрения, оптимальным вариантом является использование обучаемой на неискаженных данных искусственной нейронной сети-образца (ИНС-образца), с высокой вероятностью дающей объективно правильные ответы. При этом обучение образца на полном объеме данных будет невыгодным и времязатратным, поэтому перед нами встает задача определения эффективного объёма данных, необходимых для обучения такого ИНС-образца.

Для решения данной задачи используется ранее обученная на искаженных данных нейросеть распознавания образов. На основе данных о её архитектуре была сделана программная копия данной нейронной сети, после чего данная копия была обучена на постепенно увеличиваемом объёме данных. Для простоты максимальное количество используемых данных ограничено 10% с шагом в 1% для каждого нового теста (табл. 3).

Таблица 3

Количество ошибок распознавания в зависимости от процента использованных обучающих данных

		Количество использованных обучающих данных (%)								
		1	2	3	4	5	6	7	8	9
Элементы (цифры)	0	4	4	0	3	3	2	1	1	0
	1	6	3	3	2	3	3	2	2	1
	2	19	12	15	13	18	16	8	8	8
	3	22	21	27	17	12	17	14	9	12
	4	**92**	**104**	**104**	**105**	**102**	**104**	**103**	**107**	**105**
	5	24	19	21	11	22	12	13	10	8
	6	8	9	7	6	8	5	5	5	5
	7	13	15	15	25	18	12	11	16	11
	8	19	22	16	15	13	15	13	14	12
	9	17	12	10	12	6	11	7	7	8

После завершения эксперимента была построена диаграмма количества «отбракованных» элементов в зависимости от процента использованных в обучении НС-образца данных (рис. 5, 6).

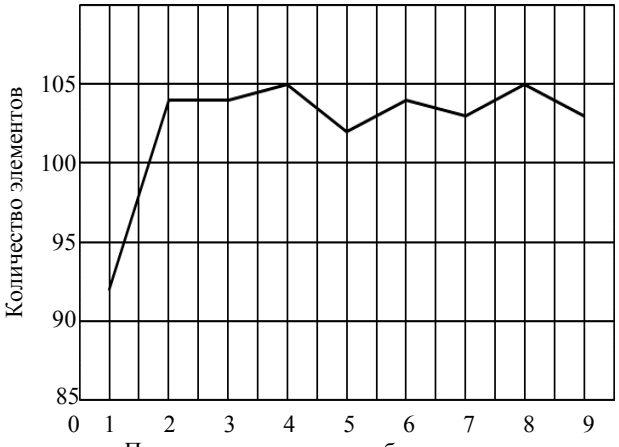
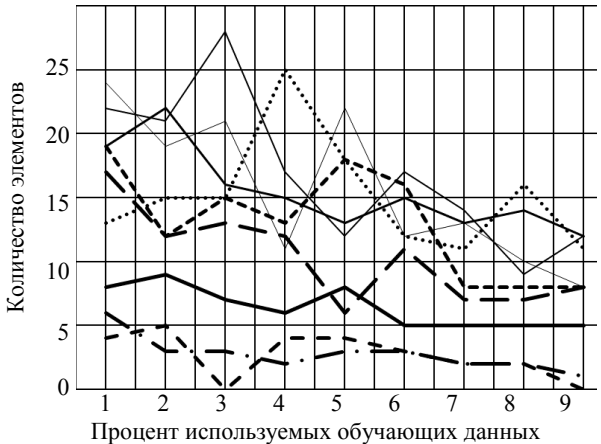


Рис. 5. График изменения количества отбракованных элементов вида «4» от процента обучающих данных



Элемент	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Тип линии									

Рис. 6. График изменения количества неатакованных отбракованных элементов от процента обучающих данных

Из таблицы видно, что наибольшее количество искаженных элементов было «отбраковано» при использовании 8 и 9% ИНС-образцов. Поскольку мы пытаемся детектировать атакованный элемент на

основе анализа ошибок нейронной сети и этим же способом определять влияние совершенной атаки на не атакованные элементы, то 9% ИНС-образец является наиболее подходящим благодаря четкой контрастности результатов, полученных при его использовании. Также использование столь малого количества данных позволит добиться высокой скорости подготовки образца к проверке и низкой загруженности вычислительных ресурсов.

Таким образом, метод детектирования состязательной атаки с использованием ИНС-образца способен показывать положительные результаты при детектировании состязательной атаки и определении её влияния на работу нейронной сети.

Выводы

По результатам эксперимента выяснено, что определение характера состязательной атаки на основе анализа ответов атакованной нейронной сети возможно и данная методика имеет право на существование. Однако требуется продолжение исследований в данной области, в частности, апробация методики на более сложных моделях нейронных сетей и использование для атаки более специфических и комплексных искажений обучающих данных.

Литература

1. Intriguing properties of neural networks / C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus // arXiv preprint arXiv:1312.6199.2013.

2. Explaining and harnessing adversarial examples / I.J. Goodfellow, J. Shlens, C. Szegedy // arXiv preprint arXiv:1412.6572.2014.

3. The limitations of deep learning in adversarial settings / N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, S. Ananthram // 2016 IEEE European symposium on security and privacy (EuroS&P). – IEEE. – 2016. – P. 372–387.

4. Towards deep learning models resistant to adversarial attacks / A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu // arXiv preprint arXiv:1706.06083.2017.

5. Adversarial attacks and defenses in images, graphs and text: A review / H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang, A.K. Jain // International Journal of Automation and Computing. – 2020. – Vol. 17. – P. 151–178.

6. Multi-column deep neural network for traffic sign classification / D. Ciresan, U. Meier, J. Masci, J. Schmidhuber // Neural Networks. – 2012. – Vol. 32. – P. 333–338.

7. Ivanyuk V.A. Neural networks and their analysis // Chronoeconomics. – 2021. – № 4 (32) [Электронный ресурс]. – URL: https://cyber leninka.ru/article/n/neyronnye-seti-i-ih-analiz, свободный (дата обращения: 28.09.2023).

8. Neuron networks – development prospects / K.S. Kachagina, A.D. Safarova // E-Scio. – 2021. – No. 2 (53) [Электронный ресурс]. – URL: https://cyberleninka.ru/article/n/ neyronnye-seti-perspektivy-razvitiya, свободный (дата обращения: 28.09.2023).

9. Investigation of adversarial attacks on pattern recognition neural networks / D.V. Kotlyarov, G.D. Dyudyun, N.V. Rzhevskaya, M.A. Lapina, M.G. Babenko // Proceedings of the Institute for System Programming of the RAS. – 2023. – Vol. 35, No. 2. – P. 35–48 [Электронный ресурс]. – URL: http://syrcose.ispras.ru/2023/submissions/SYRCoSE_2023_paper_1044.pdf, свободный (дата обращения: 15.01.2025).

10. Особенности организации атак на нейронные сети для распознавания образов / М.А. Лапина, Н.В. Ржевская, Д.В. Котляров, Г.Д. Дюдюн // Auditorium. – 2023. – № 2 (38). – С. 97–103 [Электронный ресурс]. – URL: https://www.elibrary.ru/item.asp?id=54117348, свободный (дата обращения: 28.09.2023).

11. Анализ методов обнаружения состязательных атак на глубокие нейронные сети / М.А. Лапина, Г.Д. Дюдюн, Д.В. Котляров // Матер. VIII Междунар. науч.-практ. конф. «Дистанционные образовательные технологии», «ДОТ–2023», 19–21 сентября 2023 г., Ялта. – С. 345–348 [Электронный ресурс]. – URL: http://elibrary.ru/item.asp?id=54606341, свободный (дата обращения: 28.09.2023).

12. Analysis of an existing method for detecting adversarial attacks on deep neural networks / M.A. Lapina, G.D. Dudun, D.V. Kotlyarov, N.V. Rjevskaya, S.J. Subramanian // Current Problems of Applied Mathematics and Computer Systems. CPAMCS 2023. Lecture Notes in Networks and Systems, – Vol. 1044. – Springer, – Cham [Электронный ресурс]. – URL: https://doi.org/10.1007/978-3-031-64010-0_29, свободный (дата обращения: 28.09.2023).

Дюдюн Глеб Дмитриевич
Студент каф. информационной безопасности автоматизированных систем
Северо-Кавказского федерального университета (СКФУ)
Пушкина ул., 1, г. Ставрополь, Россия, 355017
ORCID: 0009-0008-1256-0204
Тел.: +7-906-467-86-36
Эл. почта: gleb.dudun@gmail.com

Лапина Мария Анатольевна
Канд. физ.-мат. наук, доцент каф. вычислительной математики и кибернетики фак-та математики и компьютерных наук им. проф. Н.И. Червякова СКФУ
Пушкина ул., 1, г. Ставрополь, Россия, 355017
ORCID: 0000-0001-8117-9142
Тел.: +7-918-761-00-38
Эл. почта: mlapina@ncfu.ru

Бабенко Михаил Григорьевич
Науч. рук., д-р ф.-м.н. доцент, зав. каф. вычислительной математики и кибернетики фак-та математики и компьютерных наук им. проф. Н.И. Червякова СКФУ
Пушкина ул., 1, г. Ставрополь, Россия, 355017
ORCID: 0000-0001-7066-0061
Тел.: +7-906-440-02-19
Эл. почта: mgbabenko@ncfu.ru

Поступила в редакцию: 27.11.2024.
Принята к публикации: 17.04.2025.

Dyudyun G.D., Lapina M.A., Babenko M.G.

Exploring new scenarios of adversarial attacks on pattern recognition neural networks in the context of finding new defense methods

Neural networks (NNs) are an effective tool for solving hard-to-formalize problems, which has made them indispensable tools for solving them. However, information defense techniques in this area still lack sufficient protection, making them vulnerable to cybercriminals. This paper investigates adversarial attacks on neural networks, their characteristics, and proposes a new technique for detecting adversarial attacks.

Keywords: Neural networks, machine learning, information security, adversarial attacks.
DOI: 10.21293/1818-0442-2025-28-1-114-118

References

1. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks arXiv preprint arXiv:1312.6199.2013.

2. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.2014.

3. Papernot N., McDaniel P, Jha S., Fredrikson M., Celik Z.B., Ananthram Swami. The limitations of deep learning in adversarial settings. *2016 IEEE European symposium on security and privacy* (*EuroS&P*), IEEE, 2016, pp. 372–387.

4. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.2017.

5. Xu H., Ma Y., Liu H., Deb D., Liu H., Tang J., Jain A.K. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*. 2020. vol. 17, pp. 151–178.

6. Ciresan D., Meier U., Masci J., Schmidhuber J. Multi-column deep neural network for traffic sign classification *Neural Networks*, vol. 32, 2012, pp. 333–338.

7. Ivanyuk V.A. Neural networks and their analysis. *Chronoeconomics.* 2021, no. 4 (32). URL: https://cyberleninka.ru/article/n/neyronnye-seti-i-ih-analiz, free (Accessed: September 28, 2023).

8. Kachagina K.S., Safarova A.D. Neuron networks – development prospects. *E-Scio.* 2021, no. 2 (53). URL: https://cyberleninka.ru/article/n/neyronnye-seti-perspektivy-razvitiya, free (Accessed: September 28, 2023).

9. Kotlyarov D.V., Dyudyun G.D., Rzhevskaya N.V., Lapina M.A., Babenko M.G. Investigation of adversarial attacks on pattern recognition neural networks. *Proceedings of the Institute for System Programming of the RAS*, – 2023, vol. 35, no. 2, pp. 35–48. URL: http://syrcose.ispras.ru/2023/submissions/SYRCoSE_2023_paper_1044.pdf., free (Accessed: September 28, 2023).

10. Lapina M.A., Rzhevskaya N.V., Kotlyarov D.V., Dyudyun G.D. Features of organization of attacks on neural networks for pattern recognition. *Auditorium.* 2023, no. 2 (38). pp. 97–103 (in Russ.). URL: https://www.elibrary.ru/item.asp?id=54117348, free (Accessed: September 28, 2023).

11. Lapina M.A., Dyudyun G.D., Kotlyarov D.V. Analysis of methods for detecting adversarial attacks on deep neural networks. *Proceedings of the VIII International Scientific and Practical Conference «Distance Education Technologies», «DET-2023»*, section «Information Security and Cyber Resistance», September 19–21, 2023, Yalta, pp. 345–348. (in Russ.) URL: http://elibrary.ru/item.asp?id=54606341, free (Accessed: September 28, 2023).

12. Lapina M.A., Dyudyun G.D., Kotlyarov D.V., Rjevskaya N., Subramanian S.J. Analysis of an existing method for detecting adversarial attacks on deep neural networks. *Current Problems of Applied Mathematics and Computer Systems*. CPAMCS 2023. Lecture Notes in Networks and Systems, vol. 1044. Springer, Cham. URL: https://doi.org/ 10.1007/978-3-031-64010-0_29, free (Accessed: September 28, 2023).

Gleb D. Dyudyun
Student, Department of Information Security
of Automated Systems, North Caucasus Federal University
1, Pushkin st., Stavropol, Russia, 355017
ORCID: 0009-0008-1256-0204
Phone: +7-906-467-86-36
Email: gleb.dudun@gmail.com

Maria A. Lapina
Candidate Sciences in Physics and Mathematics,
Associate Professor, Department of Computational
Mathematics and Cybernetics, Faculty of Mathematics and
Computer Sciences named after professor N.I. Cherviakov,
North Caucasus Federal University, Associate Professor
1, Pushkin st., Stavropol, Russia, 355017
ORCID: 0000-0001-8117-9142
Phone: +7-918-761-00-38
Email: mlapina@ncfu.ru

Mikhail G. Babenko
Supervisor, Doctor of Physical and Mathematical Sciences,
Head of the Department of Computational Mathematics and
Cybernetics, Faculty of Mathematics and Computer Sciences
named after professor N.I. Cherviakov,
North Caucasus Federal University, Associate Professor
1, Pushkin st., Stavropol, Russia, 355017
ORCID: 0000-0001-7066-0061
Phone: +7-906-440-02-19
Email: mgbabenko@ncfu.ru

Received: 27.11.2024.
Accepted: 17.04.2025.