Синтетические данные в задаче обнаружения аномалий в сфере информационной безопасности

А.И. Гурьянов[1]

[1] *Национальный исследовательский центр «Курчатовский институт»*

Аннотация. В настоящее время в машинном обучении высокую актуальность имеют синтетические данные. Современные алгоритмы генерации синтетических данных дают возможность генерации данных, очень близких по статистическим свойствам к исходным данным. Синтетические данные используются на практике в широком спектре задач, в том числе связанных с аугментацией данных. Предложен метод аугментации данных, совмещающий подходы увеличения объема выборки с помощью синтетических данных и генерации синтетических аномалий. Метод использован для решения задачи в сфере информационной безопасности, заключающейся в поиске аномалий в журналах сервера с целью обнаружения атак. Модель, обученная в рамках решения названной задачи, показала высокие результаты. Это демонстрирует эффективность использования синтетических данных для увеличения объема выборки и генерации аномалий, а также возможность с высокой результативностью использовать эти подходы совместно.

Ключевые слова: синтетические данные, обнаружение аномалий, информационная безопасность, генерация аномалий, аугментация данных, машинное обучение

Synthetic data in the problem of anomaly detection in the field of information security

A.I. Gurianov[1]

[1] *National Research Centre “Kurchatov Institute”*

Abstract. Currently, synthetic data is highly relevant in machine learning. Modern syn-thetic data generation algorithms make it possible to generate data that is very similar in statistical properties to the original data. Synthetic data is used in practice in a wide range of tasks, including those related to data augmentation. The author of the article proposes a data augmentation method that combines the approaches of increasing the sample size using synthetic data and synthetic anomaly generation. This method has been used to solve an

information security problem of anomaly detection in server logs in order to detect attacks. The model trained for the task shows high results. This demonstrates the effectiveness of using synthetic data to increase sample size and generate anomalies, as well as the ability to use these approaches together with high efficiency.

Keywords: synthetic data, anomaly detection, information security, anomaly generation, data augmentation, machine learning

1. Введение

В последние годы в сфере машинного обучения высокую востребованность приобрели синтетические данные. Благодаря значительному совершенствованию алгоритмов генерации данных существует возможность генерации синтетических данных, очень близких по статистическим свойствам к исходным данным.

Объем мирового рынка синтетических данных в 2022 году составил 163,8 млн долларов [1]. Объем этого рынка в настоящее время имеет стабильно высокие темпы роста (Рис. 1), что подтверждает высокую актуальность синтетических данных.

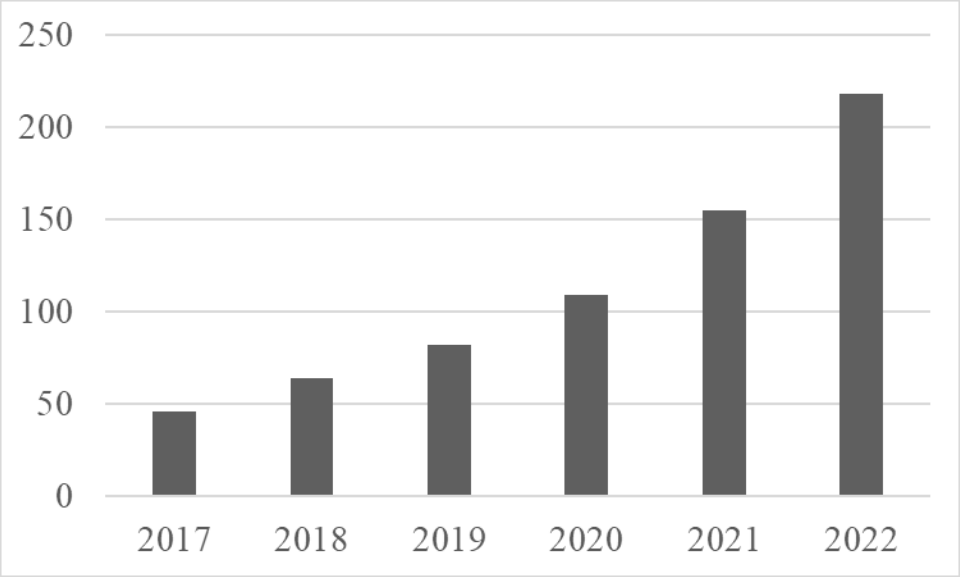

Рис. 1. Объем мирового рынка синтетических данных

Синтетические данные – это данные, являющиеся результатом работы генеративного процесса, обученного на свойствах реальных данных (см., например, [2]). Они могут быть применены на практике для решения широкого спектра задач в различных предметных областях [3, 4]. Далее рассмотрим ряд подходов применения синтетических данных в сфере машинного обучения.

Генерация синтетических данных может быть использована в машинном обучении для увеличения объема обучающей выборки [5, 6]. В

этом случае для обучения модели могут быть использованы как реальные данные, дополненные синтетическими, так и полностью синтетические данные. Кроме того, генерация синтетических данных применяется в задачах классификации в тех случаях, когда классы сильно не сбалансированы, с целью генерации дополнительных экземпляров редких классов [7, 8].

Также существует возможность применения синтетической генерации данных в задачах обнаружения аномалий для генерации аномальных значений. Этот подход имеет высокую актуальность, так как во многих случаях на практике аномалии очень редки, и собрать достаточное количество реальных аномальных данных крайне сложно или даже невозможно [9]. Данный подход неоднократно успешно применялся на практике [10–12], однако в настоящее время он недостаточно освещен в публикациях на русском языке.

Генерация данных для их использования в обучающей выборке называется аугментацией данных [13].

В работе предложен метод аугментации данных, совмещающий подходы увеличения объема выборки с помощью синтетических данных и генерации синтетических аномалий. Этот метод применен для задачи поиска аномалий в сфере информационной безопасности.

Статья имеет следующую структуру. В разделе 2 представлена постановка решенной в статье задачи. В разделе 3 описан механизм генерации синтетических данных, примененный при решении поставленной задачи. Раздел 4 посвящен обучению модели поиска аномалий. Раздел 5 является заключительным и содержит основные выводы.

2. Постановка задачи

Входные данные задачи представляют собой журналы сервера, включающие в себя информацию о событиях авторизации. Эти данные были собраны в процессе штатного функционирования рассматриваемой информационной системы. Необходимо обучить модель поиска аномалий, которая будет применяться в рамках системы обнаружения вторжений как составная часть эвристического алгоритма обнаружения атак. Действия, идентифицированные как аномалии, будут считаться подозрительными и дополнительно проверяться эвристическими методами информационной безопасности на вредоносность.

На практике данный эвристический алгоритм будет применяться совместно с сигнатурным методом. Роль сигнатурного метода будет заключаться в обнаружении атак известных типов, а эвристического метода – атак неизвестных типов. Применение эвристического метода необходимо для защиты информационной системы от угроз, информация о

которых не содержится в базе данных сигнатур из-за их новизны или редкости.

Для обучения модели поиска аномалий и проверки ее качества необходимы примеры аномальных данных. При этом, данные содержат недостаточное количество реальных примеров вредоносной активности. В то же время, составление экспертами репрезентативной выборки примеров вредоносной активности, адаптированной под особенности конкретной информационной системы, потребуют больших затрат, чрезмерных для рассматриваемой системы. В связи с этим, в данном случае имеет актуальность метод синтетической генерации аномальных значений.

На практике большая часть событий авторизации является нормальной, а атаки, как правило, являются редкими событиями. Поэтому выборка должна содержать малое количество аномалий, составляющее доли процента.

Однако в этом случае набор аномалий в выборке становится нерепрезентативным из-за их малого количества, и обучение модели на такой выборке приводит к переобучению. Поэтому возникла необходимость увеличения объема выборки за счет генерации синтетических данных на основе нормальных данных.

Все признаки данных, кроме временной метки события, являются категориальными, что существенно ограничивает круг применимых алгоритмов генерации аномалий.

3. Генерация синтетических данных

Для генерации синтетических данных была использована библиотека DataSynthesizer [14]. Механизм функционирования данной библиотеки описан ее авторами в статье [14], ее исходный код доступен на GitHub [15].

Также генерация аномальных данных реализована в рамках инструмента измерения качества алгоритмов поиска аномалий ADBench [16]. В этом инструменте реализована возможность генерации реалистичных аномалий с различными распределениями. Однако ADBench содержит возможность генерации аномалий только на основе числовых данных, категориальные же данные не поддерживаются, что является существенным ограничением и делают этот инструмент непригодным для решения поставленной выше задачи. Кроме того, в настоящее время функция генерации аномалий предназначена исключительно для работы в рамках инструмента и недостаточно адаптирована для применения к другим практическим задачам машинного обучения [17].

Библиотека DataSynthesizer поддерживает следующие уровни генерации данных:

1. Режим коррелированных атрибутов – генерирует синтетические данные с сохранением как зависимостей между столбцами исходных данных, так и распределений данных в столбцах. Полученные таким

способом синтетические данные очень близки к исходным по статистическим свойствам. Для генерации данных применяются байесовские сети.

2. Режим независимых атрибутов – сохраняет распределение данных в столбцах, но не сохраняет зависимости между столбцами.

3. Режим случайных атрибутов – здесь не сохраняются ни зависимости между столбцами, ни распределение данных. Столбцы сгенерированных данных заполняются равномерно распределенными значениями, взятыми из исходных данных.

Для генерации синтетических нормальных данных был применен режим 1, а для генерации аномалий – режим 3.

В рамках библиотеки DataSynthesizer генерация синтетических данных состоит из двух этапов.

На первом этапе осуществляется анализ входных данных. Анализатор данных (класс DataDescriber) определяет различные статистические свойства исходных данных. По результатам анализа создается файл описания данных в формате json.

Анализ входных данных возможно запустить на любом из уровней генерации данных. Чем выше уровень, тем выше детализация полученного файла описания данных.

Файл описания данных можно использовать для генерации данных не только на уровне, непосредственно соответствующем файлу, но и на более низком уровне. Поэтому в рамках решения поставленной задачи производился анализ входных данных на уровне режима коррелированных атрибутов, а полученный файл применялся как для генерации синтетических нормальных данных, так и для генерации аномалий.

В библиотеке DataSynthesizer представлена возможность обработки как числовых, так и категориальных переменных. По умолчанию, анализатор данных автоматически определяет, к какому типу относится каждый отдельно взятый признак. Однако такое автоматическое определение не может быть безошибочным [14]. Поэтому существует необходимость явно указать, какие признаки данных являются категориальными.

Программный код первого этапа приведен на рисунке 2.

```
categorical_attributes_list = list(df.columns)
categorical_attributes = {}
for categorical_attribute in categorical_attributes_list:
    categorical_attributes[categorical_attribute] = True
categorical_attributes['logon_time'] = False

describer = DataDescriber()
describer.describe_dataset_in_correlated_attribute_mode(dataset_file=input_data,
                                                         k=2,
                                                         epsilon=0,
                                                         seed=random_state,
                                                         attribute_to_is_categorical=categorical_attributes)
```

Рис. 2. Анализ входных данных

На втором этапе производится непосредственно генерация синтетических данных. Генератор данных (класс DataGenerator) получает информацию о статистических свойствах данных из файла, созданного на первом этапе, и на основе этой информации генерирует данные. Программный код этого этапа приведен на рисунке 3.

```
from DataSynthesizer.DataGenerator import DataGenerator

generator = DataGenerator()
generator.generate_dataset_in_correlated_attribute_mode(normal_data_to_generate,
                                                         description_file,
                                                         seed=random_state)
generator.save_synthetic_data(synthetic_data_file)
```

Рис. 3. Генерация синтетических данных

После генерации осуществляется предобработка данных. Исходные данные, синтетические нормальные данные и синтетические аномалии были предобработаны одинаковым образом.

4. Обучение модели с применением синтетических данных

После предобработки формируются обучающая и тестовая выборки. Обучающая выборка формируется на основе синтетических нормальных данных, а тестовая выборка – на основе исходных данных, обе выборки в равных пропорциях содержат сгенерированные аномалии.

Для поиска аномалий применялся алгоритм Isolation Forest [18]. Для подбора оптимальных гиперпараметров модели был использован метод байесовского поиска (bayes search) [19]. Этот метод способен определять оптимальные значения гиперпараметров за меньшее количество итераций, чем поиск по сетке (grid search). Кроме того, байесовский поиск дает возможность поиска не только по дискретным, но и по непрерывным интервалам значений гиперпараметров, что дает позволяет обнаружить

более оптимальные значения [20]. Для повышения качества подбора гиперпараметров применялась перекрестная проверка (кросс-валидация). Размер валидационной выборки составлял 1/3 от общего размера обучающей выборки.

Преобразования, осуществляемые над данными, отражены на диаграмме потоков данных (data flow diagram, DFD), представленной на рисунке 4.

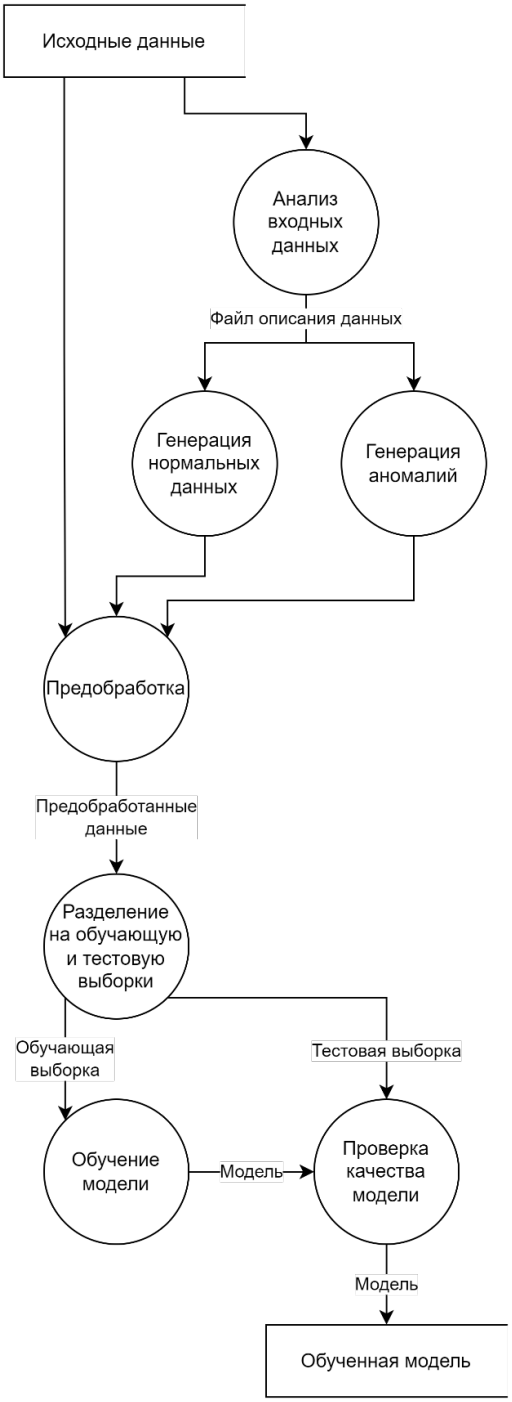

Рис. 4. Диаграмма потоков данных в рамках решения поставленной задачи

При подборе гиперпараметров максимизировалась полнота (recall). В качестве оптимизируемой метрики была выбрана именно полнота, так как при поиске потенциально вредоносных событий ложноположительные результаты значительно опаснее ложноотрицательных. В пределах разумного лучше ошибочно пометить событие как аномалию, чем не идентифицировать аномалию, значит, потенциальную атаку.

Обученная модель показала высокие результаты и успешно обнаруживает сгенерированные аномалии. Полученные значения метрик качества приведены в таблице 1.

Таблица 1

Значения метрик качества обученной модели

Метрика	Значение на обучающей выборке	Значение на тестовой выборке
Полнота (recall)	0,96	0,97
Точность (precision)	0,32	0,3

5. Заключение

Проведенное исследование показало, что генерация синтетических данных в настоящее время имеет высокую актуальность в значительном количестве предметных областей.

В статье генерация синтетических данных применена автором с целью аугментации данных для решения задачи поиска аномалий в сфере информационной безопасности. В рамках решения этой задачи синтетические данные одновременно были использованы двумя различными способами: для генерации дополнительных нормальных данных с целью увеличения объема выборки, а также генерации аномальных данных, так как в исходной выборке аномальные данные отсутствовали. Такая методика показала хорошие результаты, так как дала возможность обучить модель поиска аномалий высокого качества.

В сфере информационной безопасности данный подход следует применять в составе эвристических алгоритмов, в сочетании с сигнатурными методами.

При этом, разработанный подход потенциально возможно применить в любой сфере, где имеет актуальность задача поиска аномалий.

Таким образом, исследование продемонстрировало эффективность использования синтетических данных для увеличения объема выборки, а также генерации аномальных данных. Кроме того, было установлено, что описанные подходы могут с высокой результативностью использоваться совместно.

Работа выполнена в рамках государственного задания FNEF-2024-0014.

Литература

1. Synthetic Data Generation Market by End-user, Type, and Geography – Analysis and Forecast // Technavio. – 2023. – URL: https://www.technavio.com/report/synthetic-data-generation-market-analysis (дата обращения 04.02.2024)
2. Assefa S., Dervovic D., Mahfouz M., Balch T., Reddy P., Veloso M. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls // Proceedings of the First ACM International Conference on AI in Finance. – 2020. – https://doi.org/10.1145/3383455.3422554
3. James S., Harbron C., Branson J., Sundler M. Synthetic data use: exploring use cases to optimise data utility // Discover Artificial Intelligence. – 2021. – Vol. 1. – https://doi.org/10.1007/s44163-021-00016-y
4. Jordon J., Szpruch L. et al. Synthetic Data – what, why and how? // ArXiv. – 2022. – https://doi.org/10.48550/arXiv.2205.03257
5. Хафизов А.В., Григорьев М.В. Генерирование синтетических пористых изображений для аугментации данных с целью тренировки алгоритмов машинного обучения // Сенсорные системы. – 2021. – Т. 35, № 4. – С. 340–347. – https://doi.org/10.31857/S023500922104003X
6. Heine J., Fowler E.E.E., Berglund A., Schell M.J., Eschrich S. Techniques to produce and evaluate realistic multivariate synthetic data // Scientific Reports. – 2023. – Vol. 13. – https://doi.org/10.1038/s41598-023-38832-0
7. Vicente C., Muzo D., Jiménez I., Fabelo H., Gram I.T., Løchen M., Granja C., Ruiz C. Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors // Applied Sciences. – 2023. – Vol. 13(7). – https://doi.org/10.3390/app13074119
8. Wang Z., Wang H. Global Data Distribution Weighted Synthetic Oversampling Technique for Imbalanced Learning // IEEE Access. – 2021. – Vol. 9. – P. 44770–44783. – https://doi.org/10.1109/ACCESS.2021.3067060
9. Astrid M., Zaheer M., Lee S. Synthetic Temporal Anomaly Guided End-to-End Video Anomaly Detection // 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). – 2021. – P. 207–214. – https://doi.org/10.1109/ICCVW54120.2021.00028
10. Luo M., Wang K., Cai Z., Liu A., Li Y., Cheang C.F. Using Imbalanced Triangle Synthetic Data for Machine Learning Anomaly Detection // Computers, Materials & Continua. – 2019. – Vol. 58(1). – P. 15–26. – https://doi.org/10.32604/cmc.2019.03708
11. Salem M., Taheri S., Yuan J.S. Anomaly Generation Using Generative Adversarial Networks in Host-Based Intrusion Detection // 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference. – 2018. – P. 683–687. – https://doi.org/10.1109/UEMCON.2018.8796769

12. Smolyakov D., Sviridenko N., Ishimtsev V., Burikov E., Burnaev E. Learning Ensembles of Anomaly Detectors on Synthetic Data // International Symposium on Neural Networks. – 2019. – https://doi.org/10.1007/978-3-030-22808-8_30
13. Емельянов С.О., Иванова А.А., Швец Е.А., Николаев Д.П. Методы аугментации обучающих выборок в задачах классификации изображений // Сенсорные системы. – 2018. – Т. 32, № 3. – С. 236–245. – https://doi.org/10.1134/S0235009218030058
14. Ping H., Stoyanovich J., Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets // Proceedings of the 29th International Conference on Scientific and Statistical Database Management. – 2017. – P. 1–5. – https://doi.org/10.1145/3085504.3091117
15. DataResponsibly / DataSynthesizer // GitHub. – URL: https://github.com/DataResponsibly/DataSynthesizer (дата обращения 12.01.2024)
16. Han S., Hu X., Huang H., Jiang M., Zhao Y. ADBench: Anomaly Detection Benchmark // Neural Information Processing Systems (NeurIPS). – 2022.
17. Minqi824 / ADBench // GitHub. – URL: https://github.com/Minqi824/ADBench (дата обращения 23.01.2024)
18. Liu F.T., Ting K.M., Zhou Z. Isolation Forest // Eighth IEEE International Conference on Data Mining. – 2008. – P. 413–422. https://doi.org/10.1109/ICDM.2008.17
19. Snoek J., Larochelle H., Adams R.P. Practical Bayesian Optimization of Machine Learning Algorithms // Advances in Neural Information Processing Systems 25. – 2012.
20. Yang L., Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice // Neurocomputing. – 2020. – Vol. 415. – P. 295–316. – https://doi.org/10.1016/j.neucom.2020.07.061

References

1. Synthetic Data Generation Market by End-user, Type, and Geography – Analysis and Forecast // Technavio. – 2023. – URL: https://www.technavio.com/report/synthetic-data-generation-market-analysis (дата обращения 04.02.2024)
2. Assefa S., Dervovic D., Mahfouz M., Balch T., Reddy P., Veloso M. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls // Proceedings of the First ACM International Conference on AI in Finance. – 2020. – https://doi.org/10.1145/3383455.3422554
3. James S., Harbron C., Branson J., Sundler M. Synthetic data use: exploring use cases to optimise data utility // Discover Artificial Intelligence. – 2021. – Vol. 1. – https://doi.org/10.1007/s44163-021-00016-y
4. Jordon J., Szpruch L. et al. Synthetic Data - what, why and how? // ArXiv. – 2022. – https://doi.org/10.48550/arXiv.2205.03257

5. Khafizov A.V., Grigor'ev M.V. Generirovanie sinteticheskikh poristykh izobrazhenii dlia augmentatsii dannykh s tsel'iu trenirovki algoritmov mashinnogo obucheniia // Sensornye sistemy. – 2021. – T. 35, № 4. – S. 340–347. – https://doi.org/10.31857/S023500922104003X
6. Heine J., Fowler E.E.E., Berglund A., Schell M.J., Eschrich S. Techniques to produce and evaluate realistic multivariate synthetic data // Scientific Reports. – 2023. – Vol. 13. – https://doi.org/10.1038/s41598-023-38832-0
7. Vicente C., Muzo D., Jiménez I., Fabelo H., Gram I.T., Løchen M., Granja C., Ruiz C. Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors // Applied Sciences. – 2023. – Vol. 13(7). – https://doi.org/10.3390/app13074119
8. Wang Z., Wang H. Global Data Distribution Weighted Synthetic Oversampling Technique for Imbalanced Learning // IEEE Access. – 2021. – Vol. 9. – P. 44770–44783. – https://doi.org/10.1109/ACCESS.2021.3067060
9. Astrid M., Zaheer M., Lee S. Synthetic Temporal Anomaly Guided End-to-End Video Anomaly Detection // 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). – 2021. – P. 207–214. – https://doi.org/10.1109/ICCVW54120.2021.00028
10. Luo M., Wang K., Cai Z., Liu A., Li Y., Cheang C.F. Using Imbalanced Triangle Synthetic Data for Machine Learning Anomaly Detection // Computers, Materials & Continua. – 2019. – Vol. 58(1). – P. 15–26. – https://doi.org/10.32604/cmc.2019.03708
11. Salem M., Taheri S., Yuan J.S. Anomaly Generation Using Generative Adversarial Networks in Host-Based Intrusion Detection // 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference. – 2018. – P. 683–687. – https://doi.org/10.1109/UEMCON.2018.8796769
12. Smolyakov D., Sviridenko N., Ishimtsev V., Burikov E., Burnaev E. Learning Ensembles of Anomaly Detectors on Synthetic Data // International Symposium on Neural Networks. – 2019. – https://doi.org/10.1007/978-3-030-22808-8_30
13. Emel'ianov S.O., Ivanova A.A., Shvets E.A., Nikolaev D.P. Metody augmentatsii obuchaiushchikh vyborok v zadachakh klassifikatsii izobrazhenii // Sensornye sistemy. – 2018. – T. 32, № 3. – S. 236–245. – https://doi.org/10.1134/S0235009218030058
14. Ping H., Stoyanovich J., Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets // Proceedings of the 29th International Conference on Scientific and Statistical Database Management. – 2017. – P. 1–5. – https://doi.org/10.1145/3085504.3091117

15. DataResponsibly / DataSynthesizer // GitHub. – URL: https://github.com/DataResponsibly/DataSynthesizer (дата обращения 12.01.2024)
16. Han S., Hu X., Huang H., Jiang M., Zhao Y. ADBench: Anomaly Detection Benchmark // Neural Information Processing Systems (NeurIPS). – 2022.
17. Minqi824 / ADBench // GitHub. – URL: https://github.com/Minqi824/ADBench (дата обращения 23.01.2024)
18. Liu F.T., Ting K.M., Zhou Z. Isolation Forest // Eighth IEEE International Conference on Data Mining. – 2008. – P. 413–422. – https://doi.org/10.1109/ICDM.2008.17
19. Snoek J., Larochelle H., Adams R.P. Practical Bayesian Optimization of Machine Learning Algorithms // Advances in Neural Information Processing Systems 25. – 2012.
20. Yang L., Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice // Neurocomputing. – 2020. – Vol. 415. – P. 295–316. – https://doi.org/10.1016/j.neucom.2020.07.061