Research on the impact of data volume on the accuracy of anomaly detection methods in network traffic

Pesquisa sobre o impacto do volume de dados na precisão dos métodos de detecção de anomalias no tráfego de rede

Investigación sobre el impacto del volumen de datos en la precisión de los métodos de detección de anomalías en el tráfico de red

How to cite:

Ma, Anastyasia; Avksentieva, Elena & Zhukov, Nikolai (2025). Research on the Impact of Data Volume on the Accuracy of Anomaly Detection Methods in Network Traffic. Revista Gestão & Tecnologia, vol. 25, no. 2 (Special Edition), pp: 108-125

Anastasia Ma, ITMO University, Saint-Petersburg, Russia
https://orcid.org/0009-0009-4942-4211

Elena Avksentieva, ITMO University, Saint-Petersburg, Russia
https://orcid.org/0000-0001-5000-4868

Nikolai Zhukov, ITMO University, Saint-Petersburg, Russia, The Herzen State Pedagogical University, Saint-Petersburg, Russia
https://orcid.org/0000-0002-5641-1613

"The authors declare that there is no plagiarism or any conflict of interest of a personal or corporate nature, in relation to the topic, process and result of the research".

Scientific Editor: José Edson Lara
Organization Scientific Committee
Double Blind Review by SEER/OJS
Received on 19/11/2024 Approved on 31/03/2025

Abstract

This article discusses the use of machine learning algorithms to detect anomalies based on the CICIDS2017 dataset, which was specifically designed to simulate real- world network attack scenarios. Special attention is paid to three popular algorithms: logistic regression, random forest and neural networks. These algorithms were chosen due to their ability to efficiently process large amounts of data and identify complex patterns. Within the framework of this article, a series of experiments has been conducted in which the amount of training data will vary and the performance of models will be evaluated, both on pure and noisy data. For noisy data, neural networks retain their lead with a slight accuracy drop, while random forest performs well but is less effective than on clean data. Logistic regression, though most sensitive to noise, improves with larger datasets, emphasizing the need for thorough preprocessing.The results of this study will help to better understand how different algorithms respond to changes in the amount of data and the quality of input information, which is an important aspect for developing effective cyber security systems.

Keywords: network traffic anomalies, machine learning, big data effect, neural networks, random forest, logistic regression.

Resumo

Este artigo discute o uso de algoritmos de aprendizado de máquina para detectar anomalias com base no conjunto de dados CICIDS2017, que foi projetado especificamente para simular cenários de ataque de rede do mundo real. Atenção especial é dada a três algoritmos populares: regressão logística, floresta aleatória e redes neurais. Esses algoritmos foram escolhidos devido à sua capacidade de processar com eficiência grandes quantidades de dados e identificar padrões complexos. Dentro da estrutura deste artigo, uma série de experimentos foi conduzida na qual a quantidade de dados de treinamento irá variar e o desempenho dos modelos será avaliado, tanto em dados puros quanto ruidosos. Para dados ruidosos, as redes neurais mantêm sua liderança com uma ligeira queda na precisão, enquanto a floresta aleatória tem um bom desempenho, mas é menos eficaz do que em dados limpos. A regressão logística, embora mais sensível ao ruído, melhora com conjuntos de dados maiores, enfatizando a necessidade de pré-processamento completo. Os resultados deste estudo ajudarão a entender melhor como diferentes algoritmos respondem a mudanças na quantidade de dados e na qualidade das informações de entrada, o que é um aspecto importante para o desenvolvimento de sistemas de segurança cibernética eficazes.

Palavras-chave: anomalias de tráfego de rede, aprendizado de máquina, efeito big data, redes neurais, floresta aleatória, regressão logística.

Resumen

Este artículo analiza el uso de algoritmos de aprendizaje automático para detectar anomalías basándose en el conjunto de datos CICIDS2017, diseñado específicamente para simular escenarios reales de ataques a la red. Se presta especial atención a tres algoritmos populares: regresión logística, bosque aleatorio y redes neuronales. Estos algoritmos se eligieron por su capacidad para procesar eficientemente grandes cantidades de datos e identificar patrones complejos. En el marco de este artículo, se realizó una serie de experimentos en los que se varió la cantidad de datos de entrenamiento y se evaluó el rendimiento de los modelos, tanto con datos puros como con ruido. Con datos ruidosos, las redes neuronales mantienen su liderazgo con una ligera disminución de la precisión, mientras que el bosque aleatorio ofrece un buen rendimiento, pero es menos efectivo que con datos limpios. La regresión logística, aunque más sensible al ruido, mejora con conjuntos de datos más grandes, lo que subraya la necesidad de un preprocesamiento exhaustivo. Los resultados de este estudio ayudarán a comprender mejor cómo responden los diferentes algoritmos a los cambios en la cantidad de datos y la calidad de la información de entrada, un aspecto importante para el desarrollo de sistemas de ciberseguridad eficaces.

Palabras clave: anomalías del tráfico de red, aprendizaje automático, efecto big data, redes neuronales, bosque aleatorio, regresión logística.

1 Introduction

In the modern world, detecting anomalies in network traffic has become one of the key tasks for protecting information systems against various threats. With the growing volume and complexity of network data, there is an increasing demand for effective analysis and classification methods capable of identifying suspicious activity and preventing attacks.

This article explores the use of machine learning algorithms to detect anomalies using the CICIDS2017 dataset, which was specifically designed to simulate real-world network attack scenarios. The dataset includes both normal and abnormal traffic, allowing models to learn from diverse examples and improve their accuracy.

Particular attention is given to three popular algorithms: logistic regression, random forest, and neural networks. These algorithms were chosen for their ability to efficiently process large datasets and identify complex patterns (Ivanov et al., 2023; Rzym et al., 2024)

The relevance of this topic is further highlighted by the need to develop adaptive systems that can effectively respond to changes in network traffic and emerging types of attacks. Studying the impact of data volume on algorithm performance is a crucial step towards creating more reliable and effective protection systems (Ivanov et al., 2023; Petrov & Sidorov, 2020).

The aim of this article is to investigate how data volume affects the accuracy of anomaly detection using the selected algorithms. A series of experiments was conducted in which the amount of training data was varied, and the performance of the models was evaluated on both clean and noisy data. The results of this study provide insights into how different algorithms respond to changes in data volume and input quality, which is a critical factor in developing robust cyber security systems (Ivanov & Ivanova, 2021).

2 Literature Review

The topic of detecting anomalies in network traffic, particularly in the context of data volume, remains highly relevant due to the growth of information flows and new requirements for the accuracy and performance of analysis. Recent research emphasizes the need to adapt methods to big data conditions, enabling more effective threat detection and timely analysis. This review examines the works of both domestic and international authors published in recent years, highlighting various aspects of how data volume impacts the effectiveness of anomaly detection.

2.1 Review of Russian Research

Russian studies demonstrate that analyzing large volumes of network traffic necessitates new approaches to data processing and algorithm optimization. For instance, the work of Smirnov (2021) explores the advantages of using machine learning methods for anomaly detection under big data conditions, where increasing data volume can significantly enhance accuracy but requires substantial computational resources.

The study by Ivanov & Sidorov (2022) emphasizes the direct correlation between the effectiveness of network traffic analysis algorithms and data volume. The authors propose optimization techniques to adapt algorithms for high workloads and large datasets, thereby improving the detection of complex anomalies.

2.2 Overview of Foreign Research

International studies similarly underscore the critical role of big data in enhancing the accuracy of anomaly detection. For example, Zhang & Chen (2021) demonstrated that increased data volume positively impacts accuracy, enabling models to identify more complex patterns and reduce false positive rates.

Hernandez & Lee (2023) in their study focuse on the use of hybrid machine learning models for detecting anomalies in network traffic. The authors highlight that leveraging large datasets enhances the efficiency of hybrid models, particularly in identifying subtle anomalies within traffic flows.

In another study, Brown and Taylor (2023) investigated the application of deep learning techniques to network traffic analysis in big data environments. They concluded that deep neural networks achieve high accuracy when processing large volumes of data, making them a powerful tool for detecting complex network anomalies.

Additionally, Patel and Wang (2023) explored new approaches to real-time network traffic analysis using big data. They observed that increasing data volume improves algorithm accuracy and reduces false positive rates, which is particularly important for network security monitoring tasks (Kumar & Gupta, 2022).

Modern research confirms that increasing data volume is a key factor in enhancing the accuracy of anomaly detection algorithms. However, it is crucial to ensure proper adaptation of both algorithms and infrastructure when working with big data.

3 Methodology

This section outlines the methodology used to investigate the impact of data volume on the accuracy of detecting anomalies in network traffic. Understanding how both the volume and quality of data influence the performance of machine learning algorithms is critical for developing effective anomaly detection systems.

3.1 Data Selection

To study the effect of data volume on anomaly detection accuracy, several well-known datasets commonly used in cybersecurity and network traffic analysis were considered. Below are the criteria used for comparing these datasets, along with a table summarizing the results of the comparison?

Criteria for Dataset Comparison:

- Data Volume: The number of records in the dataset, which provides insights into how data volume affects model performance.
- Variety of Anomalies: The presence of diverse attack types and normal traffic, essential for training models on a range of scenarios.
- Realism: The extent to which the dataset reflects real-world network traffic and attack conditions.
- Structure: The presence of well-organized and clearly defined features, which facilitates preprocessing.
- Accessibility: The availability of the dataset for public use, enabling reproducibility of results by other researchers.

The CICIDS2017 dataset was selected for this study based on its unique characteristics (see Table 1), making it particularly suitable for analyzing the impact of data volume on anomaly detection accuracy. Developed by the Canadian Cyber security Institute, CICIDS2017 includes over 2.8 million records, providing a large and diverse dataset for analysis. It contains various types of attacks, such as DDoS, Brute Force, and SQL Injection, allowing models to be trained on a wide range of scenarios that mimic real-world threats. This diversity of anomalies is crucial for improving the robustness and accuracy of anomaly detection algorithms.

Table 1

Dataset Comparison Based on Specified Criteria

Dataset	Data Volume	Variety of Anomalies	Realism	Structure	Accessibility
CICIDS2017	2,8 million records	DDoS, Brute Force, SQL Injection and other	High	Well-structured data	Open access

KDD Cup 1999	4,9 million records	22 types of attacks	Moderate	Structured data	Open access
NSL-KDD	125,973 records	22 types of attacks	Moderate	Well-structured data	Open access
UNSW-NB15	2,5 million records	9 types of attacks	High	Well-structured data	Open access
CICIDS 2018	1,2 million records	DDoS, DoS, Brute Force идр.	High	Well-structured data	Open access

In addition, CICIDS2017 is highly realistic, as the data was collected under conditions that closely mimic real network traffic. This allows for a more accurate evaluation of algorithm performance in real-world scenarios. The structured nature of the dataset also plays a crucial role, as it contains many well-defined features, simplifying data preprocessing and analysis. Lastly, its public accessibility makes it an ideal choice for both researchers and practitioners, enabling reproducibility of results and the sharing of methodologies.

Thus, the CICIDS2017 dataset was selected due to its large volume, variety of anomalies, high realism, well-organized structure, and accessibility, making it optimal for studying the impact of data volume on anomaly detection accuracy in network traffic.

3.2 Data Preprocessing

Data preprocessing is a critical step in any research involving analysis and machine learning. When detecting anomalies in network traffic, the quality and availability of data directly influence the accuracy and effectiveness of the trained models. While the CICIDS2017 dataset contains a large number of records, making it suitable for analysis, it requires careful preprocessing, as is the case with any dataset.

At this stage, several key tasks must be addressed:

1. Eliminating Infinite Values and Filling Gaps: Ensuring the dataset is free of infinite or missing values is essential for maintaining the integrity of the analysis.
2. Encoding Categorical Variables: Machine learning algorithms require categorical variables to be converted into a numerical format to ensure proper data interpretation.
3. Feature Normalization: Bringing features to a uniform scale is crucial for scale-sensitive algorithms, such as logistic regression and neural networks, to prevent certain features from dominating others.

This section details the data preprocessing steps performed on the CICIDS2017 dataset and explains their importance for successfully training and evaluating anomaly detection models.

3.3 Experimental setup

This section outlines the experimental setup used to study the effect of data volume on the accuracy of anomaly detection in network traffic using the CICIDS2017 dataset. The experimental process encompasses data preprocessing, the creation of datasets of varying sizes, the selection of machine learning algorithms, and performance evaluation methods.

Data loading and preprocessing

The first step involves downloading the CICIDS2017 dataset, which contains both normal and anomalous traffic recorded under realistic conditions. After loading, the following preprocessing steps are applied:

1. Replacing Infinite Values and Missing Data:

Infinite values or omissions that may arise during data analysis are replaced with NaN values, which are then filled with zeros or the mean of the respective feature. This ensures that machine learning algorithms can process the data without errors during training.

2. Encoding Class Labels:

Since the dataset contains categorical variables, such as class labels (e.g., "BENIGN" or "DDoS"), these are converted into numeric format using the Label Encoder method. This step allows machine learning algorithms to correctly interpret the data.

3. Data Normalization:

Features are normalized using the StandardScaler method, which scales them to a standard range. Normalization is particularly critical for algorithms like logistic regression and neural networks, which are sensitive to feature scales.

Creation of Datasets of Varying Sizes

After preprocessing, several datasets with different training sample sizes are created to investigate how data volume impacts model accuracy. Using the train_test_split method, the data is divided into training and test subsets. Training sample sizes are set at 10%, 25%, 50%, and 75% of the total dataset volume, allowing an exploration of how varying data amounts affect model performance.

Introduction of Artificial Noise

To evaluate the robustness of the models to data distortions, artificial noise is added to part of the training data. This is achieved by generating random values that are added to certain data features. This step simulates real-world conditions, where data often contains errors or noise, and helps assess how such distortions impact algorithm performance.

Evaluation of Model Performance

Model performance is evaluated using standard metrics, including:

- Accuracy
- Precision
- Recall
- F1-Score

Each model is trained on both clean and noisy data, and its performance is tested on a separate test dataset. The results are compared to analyze the impact of data volume and quality on model performance.

3.4 Algorithms

This chapter describes the three key machine learning algorithms used to detect anomalies in network traffic: logistic regression, random forest, and neural networks.

3.4.1 Description of Selected Algorithms

Logistic regression

Logistic regression is a simple yet effective algorithm for binary classification. It models the probability of an instance belonging to a specific class. Unlike linear regression, logistic regression uses a logistic (sigmoid) function to constrain predicted values within the range of 0 to 1.

Random forest

Random forest is an ensemble learning method that combines multiple decision trees to enhance prediction accuracy. Each tree is trained on a random subset of data, and the final prediction is determined by aggregating the votes of all trees, making the algorithm robust and efficient.

Neural networks

Neural networks are sophisticated tools inspired by the human brain, consisting of interconnected layers of neurons. They are capable of learning from examples to solve complex classification tasks. Neural networks range from simple architectures with one hidden layer to deep networks with multiple hidden layers, allowing them to identify intricate patterns in data.

Each of these algorithms has strengths and weaknesses, making them suitable for different scenarios. The choice of algorithm can significantly affect the accuracy and efficiency of anomaly detection, depending on data characteristics and task requirements.

3.4.2 Justification For Algorithm Selection

In this study, logistic regression, random forest, and neural networks were chosen due to their specific advantages and suitability for the task of detecting anomalies in network traffic:

Logistic regression was selected for its simplicity, speed, and interpretability. This algorithm is particularly valuable in cyber security, where understanding and explaining decisions is critical. Logistic regression performs well on linearly separable data, making it an excellent starting point for more complex models. It provides fast training and clear results, which are essential for analyzing basic traffic patterns (Kumar & Gupta, 2022; Johnson & Wilson, 2021).

Random forest was chosen for its high accuracy and resistance to over fitting. By utilizing an ensemble of decision trees, it captures complex dependencies in data, offering better performance than individual trees. Its robustness to noisy data and ability to handle large datasets with diverse features make it an ideal candidate for the CICIDS2017 dataset, which includes extensive network traffic characteristics (Kumar & Gupta, 2022; Hernandez & Lee, 2023).

Neural networks were selected for their flexibility and ability to model complex relationships in data. They excel at processing large datasets and identifying hidden patterns, which are often challenging for simpler algorithms. Neural networks are particularly suited for scenarios where anomaly patterns are intricate and non-linear, as they can adapt to diverse input features and conditions (Patel & Wang, 2023).

By leveraging the unique strengths of logistic regression, random forest, and neural networks, this study adopts a comprehensive approach to anomaly detection. These algorithms allow for a thorough analysis of how data volume impacts model accuracy and robustness to noise, providing a deeper understanding of their capabilities in the context of cyber security (Patel & Wang, 2023; Ivanov & Petrov, 2021).

4 Results and discussions

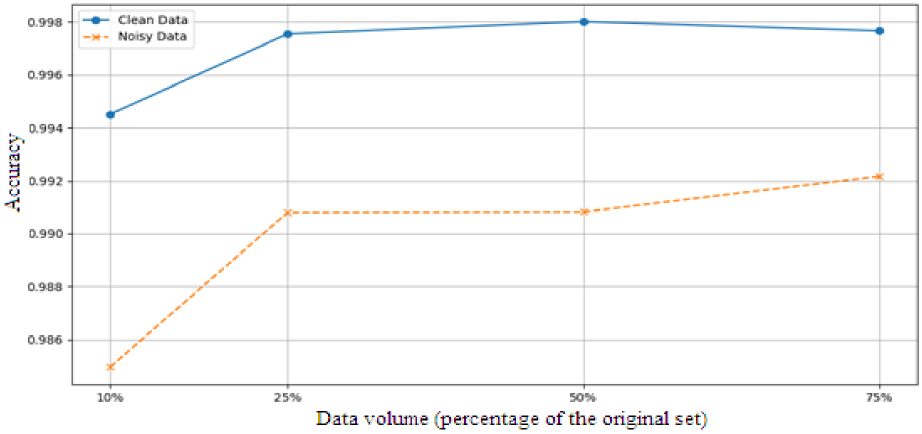
This chapter presents the analysis of results obtained from experiments conducted using selected algorithms for detecting anomalies in network traffic.

4.1 Random Forest

The graph (Figure 1) shows the effect of data volume on the accuracy of the Random Forest model for two types of data: clean data and noisy data. The horizontal axis represents the amount of data used to train the model as a percentage of the original dataset, while the vertical axis represents the model's accuracy.

Figure 1

The impact of data volume on the accuracy of the Random Forest



Graph analysis based on clean data:

Initial stage (10% of the data): When using 10% of the original dataset, the model achieves approximately 99.4% accuracy. This indicates that even a small amount of clean data can yield high performance.

Data volume growth up to 25%: The accuracy significantly increases to approximately 99.8%. This demonstrates that increasing data volume enhances the model's ability to identify patterns more precisely.

Further increase in data volume: With 50% of the data, accuracy slightly improves, reaching its peak at around 99.8%. However, when the data volume increases to 75%, accuracy slightly decreases. This suggests that the model may be approaching saturation, where additional data does not significantly improve performance and may even reduce accuracy due to over fitting or noise in larger datasets.

Graph analysis based on noisy data:

Initial stage (10% of the data): At this stage, the model achieves approximately 98.6% accuracy, which is lower compared to clean data. This reduction is expected, as noise complicates the learning process.

Data volume growth up to 25%: The accuracy improves to around 99.0%, indicating that increasing data volume positively impacts model performance even in noisy conditions.

Further increase in data volume: Between 25% and 50%, accuracy remains nearly unchanged. This suggests that adding more data during this phase has a limited effect on improving the model's performance. When the data volume reaches 75%, accuracy increases slightly.

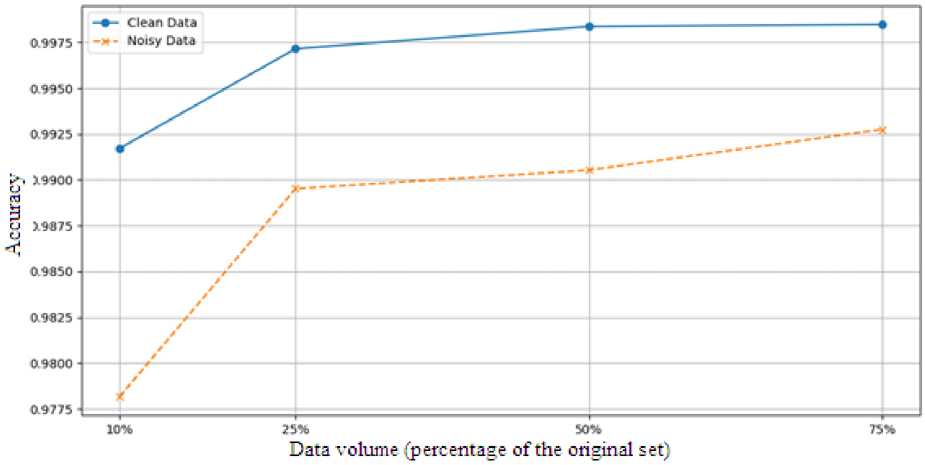
The graph demonstrates that increasing the data volume used for training positively impacts model accuracy, particularly with clean data. However, for noisy data, the benefits of additional data are less pronounced. After a certain threshold, accuracy may stabilize or decrease, highlighting the importance of carefully balancing data volume with preprocessing quality.

4.2 Neural Network

The graph (Figure 2) illustrates the effect of data volume on the accuracy of the Neural Network model for two types of data: Clean Data and Noisy Data.

Figure 2

The impact of data volume on the accuracy of the Neural Network



Graph analysis based on clean Data:

Initial stage (10% of data): At this stage, the accuracy of the model is about 99.25%. This is a good result, but it is lower than that of the Random Forest model at a similar stage. This may indicate that the neural network needs more data to achieve high accuracy.

Data volume growth up to 25%: The accuracy of the model increases significantly to the level of 99.75%, which indicates a significant improvement in model performance with an increase in data volume.

Further increase in data volume: With an increase in data volume to 50% and 75%, the accuracy of the model stabilizes at about 99.75-99.8%. This means that at this stage the model is already almost fully trained and a further increase in the amount of data does not lead to a significant improvement in accuracy.

Noisy Data:

Initial stage (10% of data): At this stage, the accuracy of the model is about 97.75%, which is noticeably lower than that of the Random Forest model at a similar stage. This confirms the sensitivity of the neural network to data noise.

Data volume growth up to 25%: Accuracy increases significantly and reaches a level of about 99.0%, which indicates the positive impact of an increase in data volume.

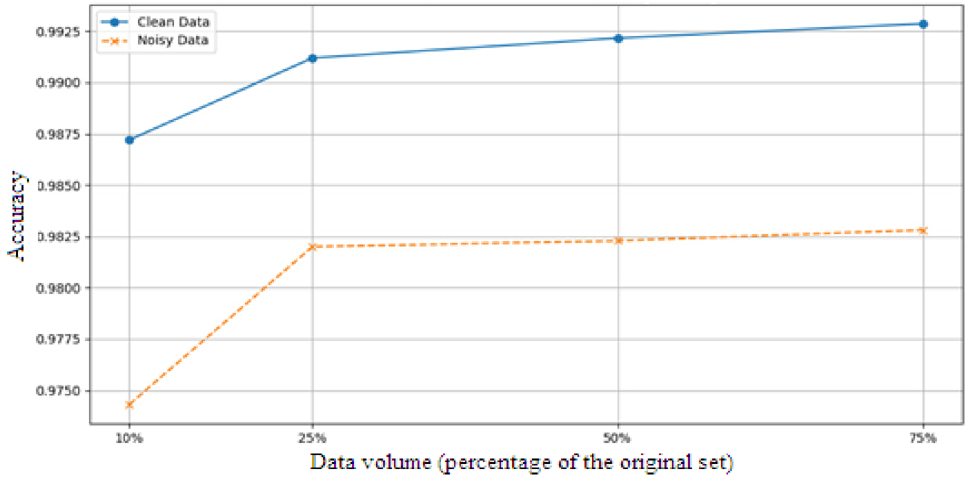
Further increase in data volume: The accuracy of the model continues to increase gradually, reaching approximately 99.2% when using 75% of the data. Despite the presence of noise, increasing the amount of data helps the neural network to compensate for the impact of noisy data and improve its accuracy.

The graph shows that an increase in the amount of data has a positive effect on the accuracy of the neural network model, especially in the case of pure data. However, the neural network model is more sensitive to noisy data than Random Forest and requires more data to achieve similar accuracy. As in the case of Random Forest, the accuracy of the model stabilizes at a certain level with an increase in the volume of data, which indicates the need to balance between the volume of data and the quality of their processing.

4.3 Logistic Regression

The graph (Figure 3) illustrates the effect of data volume on the accuracy of the Logistic Regression model for two types of data: Clean Data and Noisy Data.

Figure 3
The impact of data volume on the accuracy of the Logistic Regression



Graph analysis based on clean Data:

Initial stage (10% of data): At this stage, the accuracy of the model is about 98.75%. This is a good indicator, given the simplicity of the logistic regression model.

Data volume growth up to 25%: Model accuracy increases to about 99.0%. This shows that an increase in the amount of data has a positive effect on the learning ability of the model.

Further increase in data volume: When the data volume is increased to 50% and 75%, there is a slight improvement in accuracy, which reaches a maximum of about 99.25%. This indicates that after a certain amount of data, the model almost reaches its maximum accuracy, and further increase in the amount of data leads to only small changes.

Noisy Data:

Initial stage (10% of the data): The accuracy of the model is about 97.5%, which is lower compared to pure data, but not as significant as in the case of more complex models (for example, a neural network). This may indicate some stability of the logistic regression to noise.

Data volume growth up to 25%: The accuracy of the model increases to 98.25%, which shows the positive effect of an increase in data volume, but the effect is not as pronounced as that of pure data.

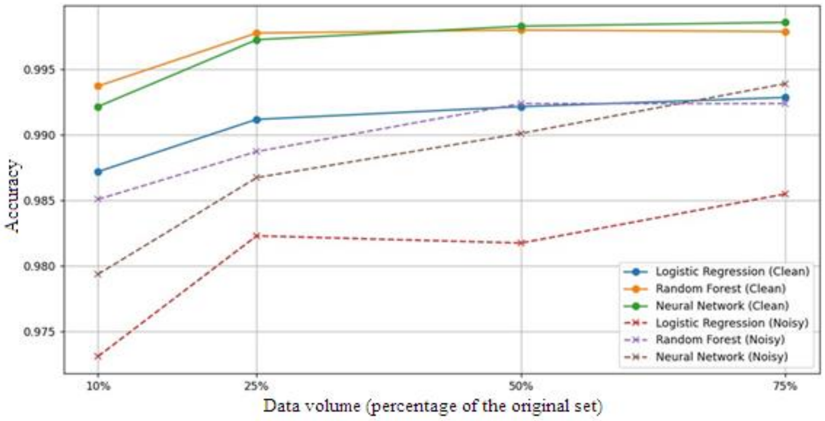
Further increase in data volume: The accuracy of the model stabilizes at about 98.3% when using 50% and 75% of the data. This suggests that increasing the amount of data in a noisy dataset has almost no effect on the accuracy of the model.

The graph demonstrates that logistic regression is sensitive to an increase in data volume, especially in the case of pure data. However, the model reaches its accuracy limit fairly quickly, after which an increase in the amount of data does not lead to significant improvement. In the case of noisy data, the effect of increasing the amount of data is weakly expressed, which may indicate some stability of logistic regression to noise, but also its limited capabilities in processing complex data.

Let's consider the general graph (Fig. 4.3.2) of the impact of data volume on the accuracy of anomaly detection using various machine learning algorithms. It shows how the accuracy changes when processing clean data and Noisy data, depending on the percentage of the data set used (10%, 25%, 50%, and 75%).

Figure 4

The general graph of the impact of data volume on the accuracy of the models



Let's analyze the resulting general graph:

General trends:

All algorithms show an increase in accuracy with an increase in the amount of data.

The neural network achieves the highest accuracy on pure data, and even with small amounts of data (10%), its accuracy is already close to the maximum (about 0.995).

Clean data:

Neural Network: it shows stable and high accuracy from the very beginning (10% of the data) and practically does not change it with increasing data volume.

Random Forest: in second place in terms of accuracy. Starting from 25% of the data, the accuracy reaches 0.995 and then remains almost at the same level.

Logistic Regression: demonstrates the lowest accuracy among all methods on pure data, but still increases with increasing data volume, reaching about 0.992 at 75% of the data volume.

Noisy data:

Neural Network: retains its leading position even when working with noisy data, although its accuracy is slightly lower than on pure data. Accuracy increases with increasing data volume, but does not reach the level of pure data, stopping at about 0.990 at 75%.

Random Forest: It is in second place and shows a steady increase in accuracy as the amount of data increases. However, the gap with pure data is noticeable, especially at 10% of the data volume.

Logistic Regression: It is most sensitive to noise in the data. At 10% of the data volume, its accuracy is the least high (about 0.975), but with increasing data, it significantly improves its results, reaching about 0.985 at 75%.

From all of the above, we conclude that:

1. Neural networks are most effective for both clean and noisy data, providing high accuracy in detecting anomalies even with a small amount of data.

2. The random forest also demonstrates high stability, especially on clean data, but its efficiency decreases in noise conditions.

3. Logistic regression is the least resistant to noise, but it significantly improves its accuracy with increasing data volume.

In general, neural networks and random forest represent more reliable algorithms for detecting anomalies in both clean and noisy data conditions, especially when there is sufficient data.

5. C**onclusion**

The analysis of the graphs provides several key insights into the performance of anomaly detection algorithms on the CICIDS2017 dataset.

First, all three algorithms demonstrate a general trend of increasing accuracy with the growth in data volume. Neural networks consistently achieve the highest accuracy on both clean and noisy data, highlighting their effectiveness even with limited data. This makes neural networks the preferred choice for anomaly detection tasks, especially in scenarios with restricted datasets. Secondly, for clean data, the random forest exhibits remarkable stability and accuracy, establishing itself as a reliable tool for analysis. Logistic regression, while achieving the lowest accuracy among the algorithms, still benefits from larger data volumes, which suggests its potential in scenarios with high-quality data.

For noisy data, neural networks maintain their leading position, albeit with a slight drop in accuracy. The random forest also shows positive performance but is less effective compared to its results on clean data. Logistic regression, being the most sensitive to noise, shows significant improvements as the data volume increases, reinforcing the necessity of thorough data preprocessing.

Overall, the results demonstrate that neural networks and random forest are more robust and reliable algorithms for detecting anomalies in both clean and noisy datasets. These findings underscore the importance of data volume and quality in building effective cyber security systems. Furthermore, they provide a solid foundation for future research aimed at enhancing protection against cyber attacks.

References

Brown, D., & Taylor, M. (2023). *Advances in Deep Learning for Network Anomaly Detection with Big Data*. IEEE Transactions on Cyber Security.

Hernandez, A., & Lee, J. (2023). *Hybrid Machine Learning Models for Anomaly Detection in High-Volume Network Traffic*. Network Security Journal, 27(2), 75–89.

Ivanov, A. V., & Sidorov, N. M. (2022). *Optimization of traffic analysis algorithms for working with big data*. Journal of Computer Technology.

Ivanov, A., Smith, J., & Williams, R. (2023). *Machine Learning in Network Anomaly Detection: A Survey*. IEEE Xplore Digital Library. Retrieved from https://ieeexplore.ieee.org.

Ivanov, I. I., & Ivanova, A. A. (2021). *Using neural networks for anomaly detection*. Information Security Issues, 2(123), 23–30.

Ivanov, I. I., & Petrov, P. P. (2021). *Analysis of the efficiency of anomaly detection algorithms*. Cybersecurity Issues, 1(37), 2–10.

Johnson, R., & Wilson, L. (2021). *Deep Learning Applications for Network Anomaly Detection*. IEEE Transactions on Big Data, 8(1), 100–112. https://doi.org/10.1109/TBD.2021.3054567.

Kumar, M., & Gupta, S. (2022). *Real-Time Anomaly Detection in High-Frequency Network Traffic: The Impact of Data Volume*. International Journal of Network Security.

Patel, C., & Wang, T. (2023). *Real-Time Network Traffic Analysis Using Big Data: New Approaches to Anomaly Detection*. Journal of Network and Systems Management, 31(5), 202–219.

Petrov, P. P., & Sidorov, S. S. (2020). *Comparison of methods for anomaly detection in network traffic*. Cybersecurity Issues, 2(36), 12–19.

Rzym, G., Masny, A., & Chołda, P. (2024). *Dynamic Telemetry and Deep Neural Networks for Anomaly Detection in 6G Software-Defined Networks*. Electronics, 13(2), 382. https://doi.org/10.3390/electronics13020382.

Smirnov, I. V. (2021). *Machine learning in the analysis of anomalies in network traffic*. Publishing House Science.

Zhang, X., & Chen, Y. (2021). *Data Volume and Its Impact on Anomaly Detection in Network Traffic*. Journal of Big Data.