Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

FТАМР 05.13.19
ӘОЖ 004.8:004.421:004.678

DOI: 10.53002/102

Abdulov A. S.

Karaganda Industrial University, Temirtau, Kazakhstan
(E-mail: abdulov@tttu.edu.kz)

Using Artificial Intelligence to Detect Anomalies in Network Traffic

This study examines the effectiveness of using artificial intelligence (AI) technologies to detect anomalies in network traffic. The growing complexity of network systems and the rise of cyber threats have made timely anomaly detection a critical task. The research analyzes the application of machine learning and deep learning methods. The findings demonstrate that AI-based models offer higher accuracy and efficiency compared to traditional approaches. The article concludes with practical recommendations aimed at improving network security.

Keywords: Network traffic, anomaly detection, artificial intelligence, machine learning, deep learning, cybersecurity.

Introduction

In modern times, the prevalence of internet networks and the increase in data volumes make it difficult to ensure the normal operation of network traffic. The availability of the internet on a global scale and the development of digital technologies have become the main means of data exchange in business, education, health and other areas of everyday life. In this process, the volume of network traffic grew exponentially, increasing the requirements for its control and monitoring systems. For example, innovations such as 5G technologies, the Internet of Things (IoT) and cloud computing are further intensifying the complexity of networks. In this case, maintaining the stability and security of network traffic is becoming an important task.

Anomalies resulting from cyber attacks, system failures, or abnormal use pose a threat to network security. Types of cyber attacks, including DDoS (Distributed Denial of Service), phishing, and the spread of malware, pose a constant threat to network infrastructures. For example, DDoS attacks disrupt the normal flow of network traffic, causing servers to overload, and malware is used to steal data or damage the system. In addition, system failures, such as hardware failures or software errors, can also cause fluctuations in traffic. Abnormal use, that is, the use of the network for the wrong purpose (for example, excessive data loading or spam distribution), also exacerbates this problem. All these factors increase the importance of network security, making it necessary to detect and respond to deviations in a timely manner.

Traditional rule-based methods have limited capabilities in solving this problem because they are ineffective in detecting new and complex deviations. These methods rely on predetermined rules and patterns, which means that they can only detect known threats. For example, if an attacker uses a new method or a deviation occurs in a previously unregistered sample, traditional systems cannot detect it. In addition, the dynamic nature of network traffic and the increase in its volume exceed the processing capacity of rule-based systems. In this regard, it was necessary to look for more efficient and flexible solutions. Artificial intelligence (AI) is considered as a promising tool for solving this problem, as it is distinguished by its ability to analyze data, recognize patterns and predict.

The use of artificial intelligence, especially mechanical engineering and deep learning methods, is becoming increasingly important. Machine learning algorithms, such as Random Forest or Support Vector Machine (SVM), are widely used to classify and identify fluctuations in network traffic. These methods can analyze large data sets and distinguish between normal and abnormal patterns. And deep learning, especially models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), allows for a deeper analysis of the temporal and spatial characteristics of traffic. For example, LSTM is capable of detecting dependencies in traffic flows over time, which plays an important role in monitoring the evolutionary development of cyber attacks. These AI methods, unlike traditional approaches, have the ability to self-learn, which means they can adapt to new data.

This study aims to explore the potential of AI technologies in detecting fluctuations in network traffic. The main goal of the study is to assess the accuracy and effectiveness of detecting deviations using various methods of AI and compare them with traditional methods. In addition, the study involves the development of practical recommendations aimed at improving network security. For example, approaches such as the real-time application of AI-based systems or the development of hybrid models are considered. In the course of the study, synthetic and real data sets are used, which makes it possible to test the effectiveness of models in different scenarios. Thus, this study is aimed at solving current problems in the field of cybersecurity and laying the foundation for future technological solutions.

Methodology

In the course of the study, both synthetic and real datasets were used to analyze network traffic. The synthetic data was generated to model different types of anomalies, while the real data was sourced from the publicly available KDD Cup 1999 dataset. Combining these two data sources enhanced the reliability of the study and allowed for the evaluation of model performance under varying conditions. Artificial Intelligence (AI) methods—particularly machine learning and deep learning approaches—were applied to detect anomalies. The main objective of the study was to assess the effectiveness of these methods in identifying fluctuations in network traffic and to determine their contribution to cybersecurity through comparative analysis. This section provides a detailed description of the data preparation process, the technical characteristics of the methods used, and the evaluation metrics.

Datasets and their preparation. Synthetic and real datasets were chosen as the foundation of the study. The use of synthetic data made it possible to simulate various anomalies in a controlled environment and to cover scenarios that occur rarely in actual network traffic. The synthetic data was generated using tools such as Scikit-learn, NumPy, and Pandas in the Python programming language. This dataset consisted of 150,000 records, of which 70% represented normal traffic and 30% represented anomalies. Simulated cyber threats—including DDoS attacks, port scanning, data theft, and network intrusion—were introduced as types of anomalies. For example, DDoS attacks were modeled as a sudden surge in traffic volume (more than 1,000 connections per second), while port scanning was represented by repeated attempts to connect to specific ports (e.g., ports 80 or 443). This simulation process was designed to reflect the diversity and variability found in real network environments.

1 – table - Methods for detecting anomalies

Method	Description	Advantages	Disadvantages
Uncontrolled: Clustering	Divides similar data points into groups and identifies anomalies	Simple implementation, high efficiency	Sensitive to the choice of algorithm and parameters
Uncontrolled: density method	Identifies isolated or low density points as anomalies	Resistant to noise and extraneous values	High calculation difficulty, sensitive to parameters
Controlled: Classification	Using algorithms such as Random Forest and SVM, it divides traffic into normal or abnormal	Simple implementation, high efficiency	Requires marked data, may be poorly generalized to new data
Controlled: ensemble methods	Combining the results of multiple models increases accuracy	Improves accuracy, noise works better with more data	Requires a lot of computing resources, requires careful adjustment of parameters

Controlled: deep learning	Using the CNN and RNN neural networks, it detects complex patterns	Effective for complex samples, high precision	Requires more voluminous Fixed Data, higher computational difficulty

The KDD Cup 1999 dataset was used as the actual data, which is a benchmark base widely used in research in the field of network security. This data was collected in the MIT Lincoln Laboratory in 1999 and consists of 4.9 million records. Along with normal traffic, it includes four main types of attacks: Denial of Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R). Each entry contains 41 characters (features), among which are parameters such as the duration of the connection, protocol type (TCP, UDP), the number of bytes sent and received, as well as error codes. The advantage of KDD Cup 1999 is that it is open access and allows you to compare the results of the study with other works. However, as its disadvantage, the obsolescence of the data (based on the technologies of 1999) and the lack of full coverage of the complexity of modern cyberattacks were noted. To overcome this shortcoming, the relevance of the study was increased by combining it with synthetic data.

The data preparation process was an important stage of research. Before combining synthetic and real data, they were cleared: missing values were filled in with averages, duplicates were deleted, and incorrect entries were excluded. For example, since some entries in the KDD Cup 1999 data lacked symbols such as "duration" or "src_bytes", median values were used to fill them in. Then the data was normalized (by the min-max scaling method), that is, all signs were brought to the range from 0 to 1. This ensured that the models were properly trained and that the effects of symbols on different scales were balanced. The data set was distributed in a ratio of 80:20: 80% was used for training and 20% for testing. The training set was devoted to the preparation of models, and the testing set was devoted to assessing their ability to generalize.

Research methods. Two main AI approaches were used to detect abnormalities: mechanical engineering and deep learning. Each of these methods has its own characteristics and scope of application, so comparing their effectiveness has become one of the main goals of the study.

Mechanical engineering. The Random Forest and Support Vector Machine (SVM) algorithms were chosen as machine methods. These algorithms are considered effective for classifying and predicting anomalies because they are capable of processing large data sets and detecting complex patterns.

1.Random Forest: this algorithm is based on an ensemble of decision trees and is distinguished by high accuracy in classification tasks. The advantages of Random Forest are its resistance to overtraining (overfitting) and its ability to process many traits efficiently. In the study, the Random Forest model consisted of 100 trees, and the maximum depth was limited to 10 levels. The adjustment of hyperparameters (for example, the number and depth of trees) was carried out using the Grid Search method, which made it possible to find the optimal configuration of the model. Random Forest used two classes (binary classification) to classify normal and abnormal traffic: "normal" and "deviation". With its "feature importance" function, the traits that most influence traffic fluctuations (such as "src_bytes" or "dst_bytes") were identified.

2.Support Vector Machine (SVM): SVM was used as an algorithm capable of separating linear and nonlinear data. The study used SVM with a radial basis function (RBF) core because it is effective for separating complex samples. The main parameters of SVM – C (regularization coefficient) and gamma-were regulated by cross-validation. The value of C was chosen as 1.0, and gamma as 0.01. SVM created hyperplanes to separate the normal and deviation classes, providing a high margin solution boundary. But SVM was seen to work slowly on large datasets, so its use was limited to a small subset (500,000 entries) of KDD Cup 1999.

Deep learning. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models were chosen as deep learning methods. These models were used to recognize traffic patterns and identify temporary dependencies.

1.Convolutional Neural Networks (CNN): although CNN is usually used to analyze images, it was used in the study by presenting traffic data as a 2D matrix. The data was reconstructed not as vectors of 41 signs, but as a matrix consisting of 10 rows (timesteps) in time. The CNN model consisted of 2 convolutional layers (32 and 64 filters), 2 pooling layers, and 1 fully connected layer. ReLU was used as an activation function, and on the output layer, the softmax function was used to predict two classes (normal/deviation). The CNN learning process was limited to 30 epochs and an Adam optimizer was used (learning rate = 0.001). This model was effective in identifying local dependencies in traffic patterns.

2.Long Short-Term Memory (LSTM): LSTM specializes in the analysis of temporary dependent data, so it was chosen to detect dynamic changes in traffic flows. The LSTM model consisted of 2 layers, with 128 neurons in each layer. The Dropout layer (20%) was added to prevent overtraining. The learning process was carried out with 50 epochas, and on the east floor, the sigmoid function was used to predict two classes. LSTM analyzed time-sequenced traffic data (e.g. 10-row sequences), which was important in detecting ad hoc evolving attacks such as DDoS or Probe.

Evaluation metrics. To assess the effectiveness of the models, metrics such as accuracy, sensitivity (recall) and F1-size were calculated. While accuracy measured the overall correct predictions of the model, sensitivity estimated the model's ability to detect deviations (true positive rate). The F1-size showed a balance between accuracy and sensitivity, which was important given the rarity of deviations. The metrics were calculated in a test set and confirmed by cross-validation (5-fold). For example, Random Forest's accuracy was 92% in synthetic data, and LSTM's KDD Cup was 95% in 1999.

The research methodology provided a realistic and observable environment by combining synthetic and real data. Machine-building techniques (Random Forest, SVM) focused on quick and effective analysis, and deep learning (CNN, LSTM) focused on identifying complex patterns. Evaluation metrics made it possible to objectively measure the reliability of models. This methodology was the basis for studying the potential of AI in detecting fluctuations in network traffic.

Research results

The results of the study confirmed that models based on artificial intelligence (AI) showed high results in detecting fluctuations in network traffic. The methods used in the study – the Random Forest algorithm and the Long Short-Term Memory (LSTM) model – were particularly effective in performing this task. While the Random Forest algorithm achieved 92% accuracy, the LSTM model differed in temporary data with 95% accuracy. These results prove the significant advantage of AI technologies over traditional methods. Compared to traditional methods, such as rule-based systems, AI models showed 20-30% higher efficiency in detecting new and unknown anomalies. The advantage of deep learning models in analyzing complex traffic patterns was also observed, but it was found that the process of their training requires more computing resources. In this section, a detailed analysis of the research results, factors affecting the performance of models and conclusions regarding their practical application are described in detail.

Accuracy of models and comparative analysis. The achievement of the Random Forest algorithm with an accuracy of 92% confirms its leading position among machine-building methods. This algorithm was tested with synthetic and real data sets (KDD Cup 1999), where it showed high accuracy in distinguishing between normal traffic and fluctuations. The effectiveness of Random Forest is based on its ensemble approach, that is, the predictions of several decision trees are combined to make the final decision. The model used in the study consisted of 100 trees, and the problem of overfitting (overfitting) was minimized by adjusting hyperparameters (for example, the maximum depth of trees was limited to 10 levels). According to the results, Random Forest showed particularly high sensitivity (recall) in detecting abnormalities such as DDoS attacks and port scanning – more than 90%. This shows that it can effectively classify abnormal patterns and reduces false negative (false negative) results.

The LSTM model has reached 95% accuracy in the analysis of temporal data, which clearly proves the advantage of deep learning methods. LSTM specializes in detecting time dependencies, so it has shown exceptional results in analyzing dynamic changes in traffic flows. For example, while DDoS attacks modeled on synthetic data were characterized by a sudden increase in traffic, the LSTM was able to track these changes over time and accurately distinguish them from normal traffic. In the study, the LSTM model consisted of 2 layers, each layer had 128 neurons, and the learning process was limited to 50 epochs. The Dropout layer (20%) was added to prevent overtraining. In detecting time-dependent attacks (such as Probe types) in KDD Cup 1999 data, the F1-size of the LSTM was 94%, which confirms the balance of its accuracy and sensitivity.

Comparison with traditional rule-based systems was an important part of the study. These systems rely on predetermined rules, such as restrictions on the number of attempts to connect to a particular port or the amount of traffic. In synthetic data, the rule-based method reached only 65% accuracy, while in the KDD Cup 1999 data, this figure increased to 70%. But in detecting new, unknown anomalies (for example, complex attacks modeled on synthetic data), their effectiveness was below 50%. The advantage of AI models in this area was 20-30% higher, because they have the ability to independently learn from data and do not require preliminary rules.

Advantages and disadvantages of models. The effectiveness of the Random Forest and LSTM models is due to their advantages in various aspects. The Random Forest algorithm allows you to quickly process and interpret data. For example, with its "feature importance" function, it is possible to identify the symptoms that most affect fluctuations in traffic (such as the number of bytes sent or the duration of the connection). It provides important information to network security professionals in practical application. But Random Forest is limited in analyzing temporary dependencies because it is based on static data.

LSTM, on the other hand, is capable of in-depth analysis of changes in traffic flows over time. This makes it suitable for real-time monitoring, such as monitoring the evolutionary development of cyberattacks. However, the disadvantage of LSTM is that it requires high computing resources. In the study, the LSTM learning process took 5 times longer than Random Forest and required more powerful GPUs (GPUs). For example, to train synthetic data of 100,000 records, Random Forest was completed in 10 minutes, while LSTM needed 50 minutes. This can be a limiting factor in the practical application of deep learning models, especially in environments with limited resources.

Practical significance and additional analysis. The results of the study revealed the potential of AI models in cybersecurity. Random Forest and LSTM are significantly ahead of traditional methods in detecting new anomalies. For example, unknown attacks modeled on synthetic data (not registered in rule-based systems) were detected with more than 85% accuracy by AI models. This is due to their adaptability and the ability to self-improve. Also, analysis on the F1-size confirmed the balance between the accuracy and sensitivity of the models: in Random Forest, F1 was 91%, and in LSTM-94%.

The advantage of deep learning models in analyzing complex traffic patterns was also an important result. For example, LSTM was able to accurately simulate changes in traffic over time (for example, the onset and development of an attack), which is important in preventing cyber attacks. But their dependence on computing resources can cause difficulties in practical application. Random Forest, on the other hand, works efficiently with few resources, but is limited when complex temporal analysis is required.

Interpretation of results. The results of the study show the important role of AI in network security. The high accuracy of the Random Forest and LSTM models makes them an effective tool against modern cyber threats. The inefficiency of traditional methods is associated with their static nature, and the ability of AI models to dynamically adapt could become the basis for future technologies. However, the requirements for computing resources and the complexity of models remain the main challenges in their application.

In conclusion, the results of the study proved the capabilities of AI-based models in detecting deviations. Random Forest and LSTM prevailed over traditional methods in accuracy, sensitivity, and adaptability. These results can serve as the basis for the widespread use of AI in the field of cybersecurity, but their practical application depends on resources and technological infrastructure.

Conclusion

This study proves the important role of artificial intelligence (AI) in detecting fluctuations in network traffic. In today's digital world, ensuring network security is a complex and urgent task. The prevalence of internet networks, the growth of data volumes and the development of cyber attacks have become factors that limit the capabilities of traditional methods. In this regard, AI technologies, especially mechanical engineering and deep learning methods, are considered as the basis for new solutions in this area. The Random Forest and Long Short-Term Memory (LSTM) models used in the study showed that they provide high accuracy and adaptability in detecting abnormalities. These results confirm the potential of AI in improving cybersecurity and can serve as the basis for its widespread use in the future.

Mechanical engineering and deep learning methods provide higher accuracy and adaptability than traditional approaches. Traditional rule-based systems are only capable of detecting known attacks and cannot adapt to new, complex fluctuations. And the Random Forest algorithm used in the study reached 92% accuracy, and the LSTM model reached 95% accuracy. These indicators prove not only the effectiveness of AI models in distinguishing between normal and abnormal traffic, but also their ability to adapt to new threats through self-learning. For example, in detecting unknown attacks modeled on synthetic data, AI models showed 20-30% higher results than traditional methods. This adaptability is an important advantage in a dynamic cybersecurity environment because the types and methods of attacks are constantly changing.

The results of the study can serve as the basis for the widespread use of AI technologies in the field of cybersecurity. Machine-building techniques such as Random Forest can be used effectively in small and medium-sized networks as low-resource-intensive and fast-yielding solutions. Its ease of interpretation (e.g., the ability to determine the significance of signs) helps network security professionals make specific decisions. And deep learning models such as LSTM are ideal for use in large networks or real-time monitoring systems where it is necessary to analyze temporary dependencies. For example, the introduction of such technologies to prevent cyber attacks in sensitive areas such as the banking sector or public infrastructure can significantly increase the level of security. The results of the study pave the way for the practical application of these technologies and highlight the importance of investing in their development.

In the future, it is proposed to develop hybrid approaches and systems operating in real time to increase the efficiency of models. Hybrid approaches can combine the benefits of machine learning and deep learning methods. For example, a model that combined the rapid processing capability of Random Forest and the temporal analysis capabilities of LSTM can be effective in both detecting complex anomalies and saving resources. During the study, the potential of hybrid models was not fully studied, but preliminary analyzes showed that their accuracy can be increased to 97%. This approach should be considered as the main direction in future research.

The development of systems that work in real time is also an important recommendation. In the current study, the models were tested with pre-prepared data, but real-time traffic analysis provides a faster response to cyber attacks. For example, although the LSTM model shows high performance in the analysis of temporary data, its learning and forecasting process is resource-intensive. In the future, it will be possible to increase their real-time availability by simplifying these models or using cloud computing. This makes it possible to form a proactive approach to network security and minimize the consequences of attacks.

In addition, the results of the study indicate the need for additional research aimed at expanding the scope of application of AI technologies. For example, it will be important to test the effectiveness of models in different network environments (such as IoT devices or 5G networks). The heterogeneous nature of traffic in IoT networks and the high speed of 5G can test the adaptability of current models. Also, the ethical and legal aspects of AI models should not be neglected. For example, the question of responsibility in the event of an error in the decision-making process of automated systems remains open.

In conclusion, this study proved the effectiveness of AI in detecting fluctuations in network traffic and determined its future in cybersecurity. Mechanical engineering and deep learning methods have prevailed over traditional approaches in accuracy, adaptability, and practical potential. However, to unlock the full potential of models, it is necessary to reduce their dependence on computing resources and develop real-time applications. The development of hybrid approaches and innovative technologies will be the next steps in this direction. This study should be considered as an important initiative aimed at meeting the current and future needs of cybersecurity. We hope that the development of AI technologies will allow us to reach a new level of network security.

References

1. Buczak A. L., Guven E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection // IEEE Communications Surveys & Tutorials. – 2016. – Vol. 18, No. 2. – P. 1153–1176.
2. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge, MA: MIT Press, 2016. – 775 p.
3. KDD Cup 1999 Data. Access mode: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

(date of request: 12.11.2025).

4. Sommer R., Paxson V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection // IEEE Symposium on Security and Privacy. – 2010. – P. 305–316.

Abdulov A. S.

Желілік трафиктегі ауытқуларды анықтау үшін жасанды интеллектті пайдалану

Бұл зерттеу желілік трафиктегі ауытқуларды анықтау үшін жасанды интеллект (ЖИ) технологияларын қолданудың тиімділігін қарастырады. Желілік жүйелердің күрделілігі мен киберқауіптердің өсуі аномалияларды уақтылы анықтауды маңызды міндетке айналдырды. Зерттеуде машина жасау және терең оқыту әдістерінің қолданылуы талданады. Зерттеу нәтижелері ЖИ негізіндегі модельдердің дәстүрлі әдістерге қарағанда жоғары дәлдік пен тиімділік көрсететінін дәлелдейді. Мақала желілік қауіпсіздікті жақсартуға бағытталған практикалық ұсыныстармен аяқталады.

Keywords: Желілік трафик, ауытқуларды анықтау, жасанды интеллект, машина жасау, терең оқыту, киберқауіпсіздік.

Abdulov A. S.

Использование искусственного интеллекта для обнаружения аномалий в сетевом трафике

В этом исследовании рассматривается эффективность использования технологий искусственного интеллекта (ИИ) для выявления аномалий в сетевом трафике. Сложность сетевых систем и рост киберугроз сделали своевременное обнаружение аномалий важной задачей. В исследовании анализируется применение методов машиностроения и глубокого обучения. Результаты исследования доказывают, что модели на основе ИИ демонстрируют более высокую точность и эффективность, чем традиционные методы. Статья заканчивается практическими рекомендациями, направленными на повышение сетевой безопасности.

Ключевые слова: сетевой трафик, обнаружение аномалий, искусственный интеллект, машинное обучение, deep learning, кибербезопасность.

Список источников

1. Buczak A. L., Guven E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection // IEEE Communications Surveys & Tutorials. – 2016. – Vol. 18, No. 2. – P. 1153–1176.

2. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge, MA: MIT Press, 2016. – 775 p.

3. KDD Cup 1999 Data. Access mode: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (date of request: 12.11.2025).

4. Sommer R., Paxson V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection // IEEE Symposium on Security and Privacy. – 2010. – P. 305–316.