A NEW APPROACH TO INTELLIGENT NETWORK TRAFFIC FLOW ANALYSIS

Kabulov Anvar
Professor of the National university of Uzbekistan named after Mirzo Ulugbek
E-mail: anvarkabulov@mail.ru
Juraev Maruf
Senior teacher of the Tashkent Institute of Technology, Management and Communication
E-mail: juraevmaruff@gmail.com

Abstract: Today, modern cyber threats are highly developed and new types are emerging, and the demand for systems aimed at detecting them in the protection of critical network infrastructure is also increasing. Rules-based systems have trouble detecting new threats as they emerge. Organizations are trying to use artificial intelligence-based threat detection methods as a solution. This research paper presents a method that detects network changes with artificial intelligence. Machine learning (ML) and deep learning (DL) methods, such as Support Vector Machines (SVM), Random Forest (RF), Autoencoders and Convolutional Neural Networks (CNNs), were used. The selected attribute based on the SHapley Additive Explanations (SHAP) analysis is used to identify the most important features in network traffic and is used to verify the understandability and effectiveness of the model. The proposed model achieved an accuracy of 80.4%, F1-score of 89.1%. Comparative analysis with previous studies has proven to provide high accuracy in detecting unexpected situations.

Keywords: packet features, network flow, accuracy, precision, recall, F1-score, Machine learning, Support Vector Machines, Random Forest

Annotatsiya: Hozirgi kunda zamonaviy kiber tahdidlar juda rivojlangan va yangi turlari paydo bo‘lmoqda, muhim tarmoq infratuzilmasini himoya qilishda ularni aniqlashga qaratilgan tizimlarga talab ham ortib bormoqda. Qoidalar asosida ishlaydigan tizimlar yangi tahdidlar paydo bo‘lganda ularni aniqlashda muammolar paydo bo‘ladi. Tashkilotlar buning yechimi sifatida sun’iy intellektga asoslangan tahdidlarni aniqlash usullarini qo‘llashga harakat qilishmoqda. Ushbu tadqiqot ishida sun’iy intellekt bilan tarmoqdagi o‘zgarishlarni aniqlaydigan usulni taqdim etadi. Mashinaviy o‘rganish (ML) va chuqur o‘rganish (DL) usullaridan Support Vector Machines (SVM), Random Forest (RF), Autoencoders va Convolutional Neural Networks (CNNs) algoritmlaridan foydalanildi. SHapley Additive Explanations (SHAP) tahlili asosida tanlangan atribut tarmoq trafigidagi eng muhim xususiyatlarni aniqlash uchun qo‘llanilib, modelning tushunarli bo‘lishi va samaradorligini tekshirishda qo‘llaniladi. Taklif etilgan model accuracy 80.4%, F1-

ISSN 3030-3907

score 89.1% erishdi. Oldingi tadqiqotlar bilan qiyosiy tahlil kutilmagan holatlarni aniqlashdagi yuqori aniqlik berishini isbotladi.

Kalit so‘zlar: paket xususiyatlari, tarmoq oqimi, accuracy, precision, recall, F1-score, Machine learning, Support Vector Machines, Random Forest

Аннотация: Сегодня современные киберугрозы высокоразвиты и появляются новые типы, и спрос на системы, направленные на их обнаружение при защите критической сетевой инфраструктуры, также растет. Системы, основанные на правилах, испытывают трудности с обнаружением новых угроз по мере их появления. Организации пытаются использовать методы обнаружения угроз на основе искусственного интеллекта в качестве решения. В этой исследовательской работе представлен метод, который обнаруживает изменения в сети с помощью искусственного интеллекта. Были использованы методы машинного обучения (Machine Learning – ML) и глубокого обучения (Deep Learning – DL), такие как опорные векторные машины (Support Vector Machines – SVM), случайные леса (Random Forest – RF), автокодировщики и сверточные нейронные сети (Convolutional Neural Networks – CNN). Выбранный атрибут на основе анализа SHapley Additive Explanations (SHAP) используется для определения наиболее важных особенностей в сетевом трафике и используется для проверки понятности и эффективности модели. Предложенная модель достигла точности 80,4%, F1-оценки 89,1%. Сравнительный анализ с предыдущими исследованиями доказал, что обеспечивает высокую точность обнаружения непредвиденных ситуаций.

Ключевые слова: характеристики пакета, сетевой поток, accuracy, precision, recall, F1-score, Machine learning, Support Vector Machines, Random Forest

Introduction

With the expansion of the digital infrastructure, the importance of the protection of information systems has also increased. The increasing interconnectedness of information systems, including the widespread adoption of cloud technologies and the rapid development of the Internet of Things (IoT), has led to an increase in the level of risk [6]. According to Cybersecurity Ventures, in 2024, the total cost of damages to information systems worldwide was 9.5 trillion dollars. Experts predict that by 2025, the cost of losses to the global economy from threats to information systems will reach $10.5 trillion.

Standard Intrusion Detection System (IDS) systems often produce 20% to 30% false positives (false alarms) and fail to detect complex and new types of threats. Artificial intelligence-based intrusion detection systems are considered an optimal solution to security needs, as they allow for effective detection of threats

ISSN 3030-3907

to information systems using ML and DL technologies. According to recent research, machine learning models such as SVM and RF are more effective in detecting threats than conventional methods [1]. SVM, RF, and Extreme Learning Machines have shown good results in classifying controlled data structures, while CNNs and Recurrent Neural Networks (RNNs) allow detecting complex attack signatures by processing unfiltered network traffic [2].

IDS systems based on artificial intelligence have the following disadvantages: false warning, complexity and the need for large computing resources. The main strength of IDS systems that detect threats based on signs is that they detect threats early, but detection problems arise when new types of threats emerge. Anomaly-based threat detection IDS systems are effective at detecting new threats, but their detection efficiency is reduced due to false alarms [8].

The research work aims to develop an artificial intelligence-based system that combines machine learning and deep learning approaches and to do this, record and analyze network flows. Although ready-made datasets are available, the dataset data in the research work was recorded in laboratory conditions. For analysis, the dataset data was converted to a csv file format based on the developed algorithm. To achieve the goals of this research, the following main aspects are focused on:

• Development and implementation of network threat detection system by combining machine learning and deep learning models.

• Evaluate the effectiveness of the proposed system by comparing it with existing intrusion detection methods.

LITERATURE REVIEW: Anomaly-based IDS is becoming increasingly popular because it can detect deviations from normal network behavior and identify unknown threats [10]. Recent advances in artificial intelligence and DL have spurred the development of IDS systems based on these [3].

Modern IDS systems use ML and DL methods, as researchers have studied their effectiveness in this area. SVM and RF algorithms, along with DL architectures, provide high accuracy [5, 9]. They have proven to be more effective than traditional methods in detecting anomalies in network traffic. Systematic analyses show the superiority of CNN, RNN, and hybrid models in modern IDS architectures.

Engineers have combined CNN and RNN architectures to create deep learning models, as each model has its own unique strengths [6]. CNN and LSTM have been shown to perform well in complex attack detection when used together in recent studies [4, 5]. However, the superior accuracy of CNNs and RNNs comes

ISSN 3030-3907

at the expense of increased computational complexity and memory consumption, which limits their application in real-time systems.

One of the major challenges in threat detection is the lack of access to modern, high-quality datasets. The open datasets KDD99 and NSL-KDD, which are widely used by researchers, do not reflect outdated attack methods, traffic characteristics, and modern threats [7]. Although the UNSW-NB15 dataset partially addresses these shortcomings, it does not fully address issues such as feature variability [9]. Therefore, researchers propose the use of artificial intelligence algorithms to improve the adaptability and resilience of threat detection systems.

In AI-based threat detection systems, SHAP approach is chosen to provide feature-level transparency. SHAP has both theoretical and practical advantages: it is based on cooperative game theory, provides global and local explanations, and efficiency in assessing the importance of features in network traffic.

Classical and effective models in the selection of ML and DL models are preferred in the research work SVM, RF, and Autoencoder. These models can be studied with new versions or device-specific adaptation options.

The research paper combines ML and DL methods to identify network conditions, reduce the number of false positives, and improve the efficiency of the operating system in the threat detection system. This research paper explores feature selection mechanisms, real-time traffic analysis, and adaptive learning methods.

METHODOLOGY:

A. Research model

The research includes the steps of network flow recording, data collection and feature processing in Figure 1, detection model generation and response evaluation. From feature extraction to classification analysis and anomaly assessment in network traffic. Enables better adaptability in real-time network flow analysis.

ISSN 3030-3907

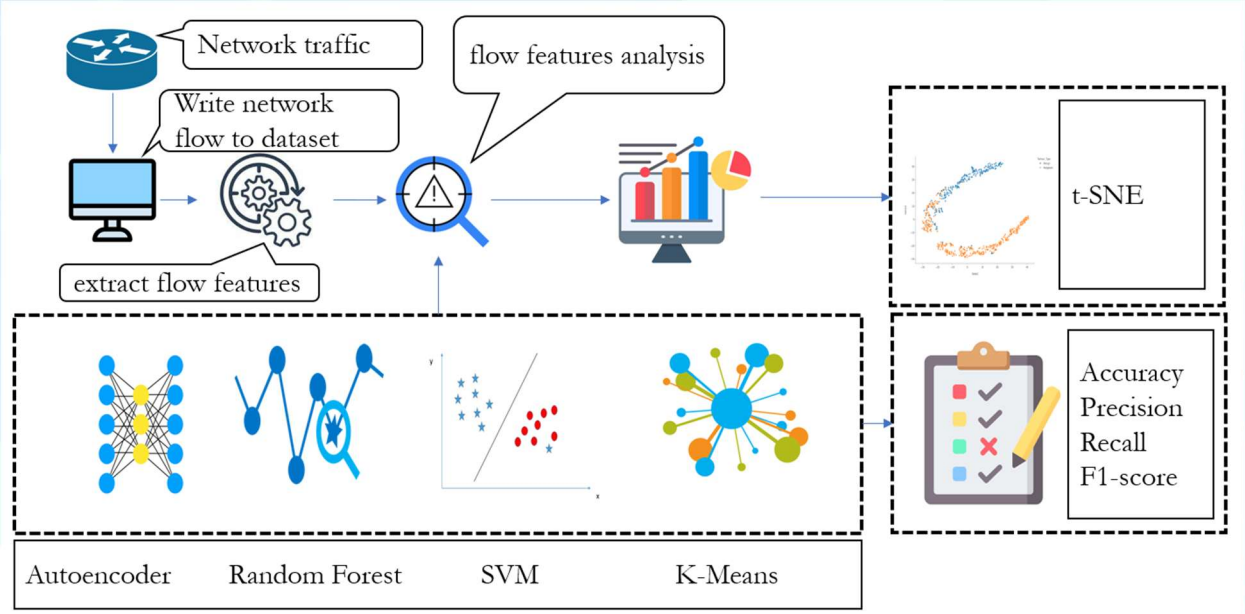


Figure 1. The process of analyzing network flow traffic.

B. Data collection

In the research work, network traffic was recorded to study modern attack types and network conditions. A network traffic monitoring program developed in the C# programming language in the laboratory was used for recording. The program allows you to save network traffic data in the pcap file format. The dataset has detailed attributes across multiple network layers, and packet features are extracted and analyzed using artificial intelligence algorithms. The research work did not demonstrate the extraction of network traffic packet features. However, feature extraction algorithms were developed. The main packet features used in the study are listed in Table 1.

Table 1.

Packet features.

Packet features	Packet size, stream duration, protocol type
Recipient and sender information	IP address, port numbers
Time features	Inter-packet arrival time, response time

C. The importance of features and their analysis

The effectiveness of a network fault detection system depends on the correct identification and prioritization of network flow characteristics. SHAP is used to determine the contribution of individual characteristics to the decision-making process. SHAP, based on its cooperative game theory, provides global and local interpretation capabilities, which are important in network flow analysis.

ISSN 3030-3907

$$\varphi_i = \sum_{S \subseteq N/\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (1)$$

Formula 1 is used here to calculate the SHAP value. φ_i denotes the contribution of the function, N is the set of all features, S is all subsets of N that contain element i, $|S|!$ is the factorial of the number of elements in S, and $v(S)$ is the predicted value.

SHAP beeswarm diagrams are available that can help identify important network features that indicate anomalous behavior. The difference between flow duration and packet length, as well as the time between packets arriving and the delay between packets, are important.

During training, the complexity of the model increases, it takes time to optimize the parameters, and this requires large data resources. In real-time systems, problems can arise if there is not enough processing power for the application. The recovery error is calculated by checking the network state.

$$E = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x_i})^2 \quad (2)$$

E is the mean squared reconstruction error, x_i is the input feature, $\hat{x}_i$ is the reconstructed output. To determine whether the reconstruction error indicates a change in the network flow, a fixed error threshold should be selected using a percentage approach. The threshold reconstruction error distribution was taken to be 95%.

RF is an effective tool for analyzing network traffic using packet features and time-related features. It learns the boundary using normal network traffic patterns through SVM and counts deviations as intrusions. It groups traffic patterns into specific clusters using K-Means, where high distances between clusters indicate unexpected situations in the network traffic. t-distributed stochastic neighbor embedding (t-SNE) visualizations are used to evaluate the clustering efficiency. Visualizations play an important role in improving the understanding of network traffic analysis. They provide network analysts with an intuitive way to show how malicious behavior occurs and how it evolves over time. Visualizing complex and high-dimensional network traffic t-SNE effectively shows cluster boundaries and behavior in low-dimensional space by creating nonlinear projections. t-SNE is used to generate 2D graphs that clearly distinguish normal and anomalous traffic.

D.Software and implementation

The proposed AI-based network protection system was implemented using a software package for machine learning, deep learning, data preprocessing, and

ISSN 3030-3907

visualization. The complete software environment settings are listed in Table 2. In addition to model development, Accuracy, Precision, Recall, and F1-score were used to evaluate the performance. The Python programming language was used to transparently rank features with the SHAP library.

Table 2.

Network flow analysis software architecture

Software	Task
Python	Programming language
TensorFlow/Keras	Implementing deep learning models
Scikit-learn	Machine learning algorithms
SHAP	Analysis of the importance of features
Matplotlib & Seaborn	Visualization
Scapy	Network flow packet analysis
Pandas & NumPy	Data preprocessing
OpenAI Gym	Designing and teaching learning environments

The developed software was developed using a 13th Gen Intel(R) Core(TM) i7 1355UCPU, 16384 MB RAM, Windows 11 pro computer.

THE RESULTS: The decision-making process of the model was explained using SHAP, which selected the most important features in determining the characteristics of the network flow. The combination of SHAP values and decision trees showed the importance of the attributes of the network flow characteristics, which are important in distinguishing normal and malicious network traffic. SHAP values were calculated in five different study-test units. The highest-rated features were selected as a constant. The 20 most important features used to determine the characteristics of the network flow were selected based on Table 1. Based on the research results, the maximum backward packet length and average backward segment size were considered the most important network flow characteristics. For the analysis, "Fwd Packet Length Mean" and "Average Packet Size" are selected as the main indicators.

In the study, network flow analysis methods were evaluated by using Autoencoder, Random Forest, SVM, and K-Means to detect new and unknown network states. The effectiveness of the models was evaluated using Accuracy, Precision, Recall, and F1-score. The methods evaluated in this study are listed in Table 3.

Table 3.

Performance indicators of detection models in network flow analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Autoencoder	74.0	80.2	89.8	84.8

ISSN 3030-3907

Random Forest	74.4	80.4	90.0	84.9
SVM	72.1	79.1	88.6	83.6
K-Means	80.4	80.4	100.0	84.8

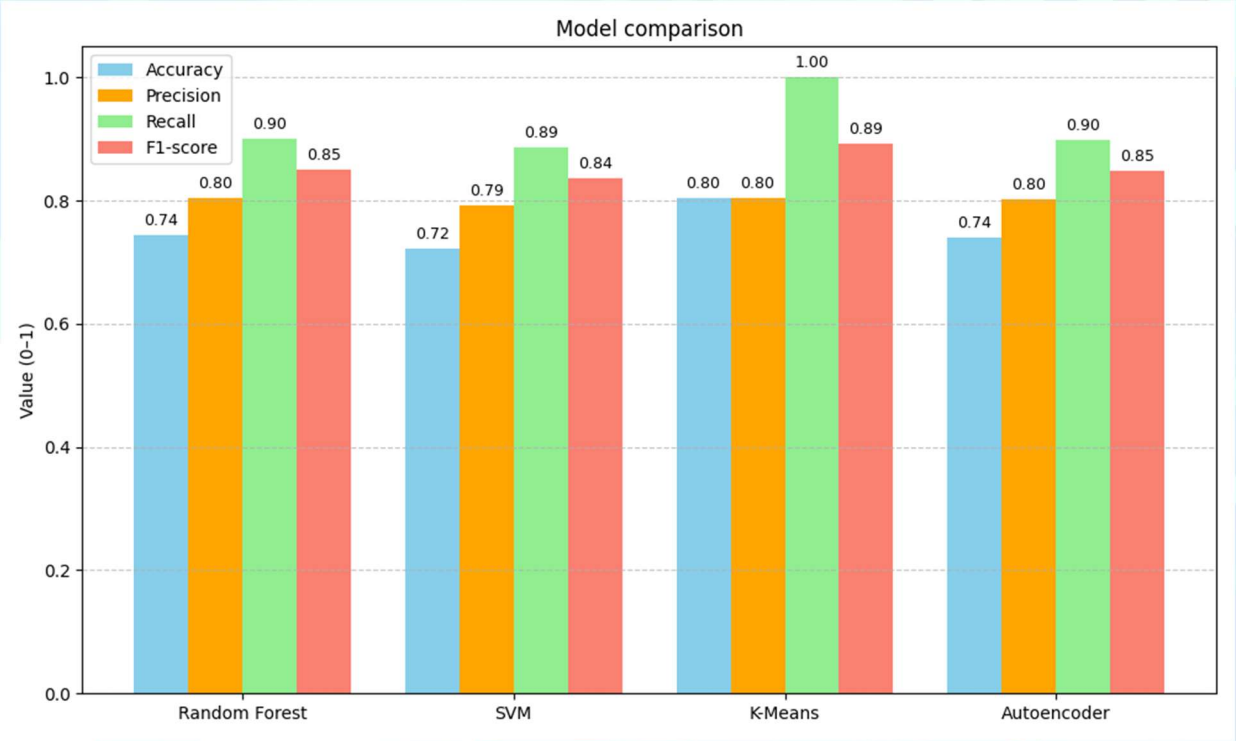


Figure 2. Comparison of anomaly detection models.

Autoencoder models show accuracy 74% precision 80.2% and recall 89.8% in network flow analysis, which confirms the ability to detect new attack patterns. K-Means showed the most effective result in the analysis of recorded network flow. Recall provided 100% accuracy. However, as the data increases, the accuracy indicator is affected

Figure 3 shows a scatterplot generated using t-SNE, where the normal state and the threat state are presented graphically. Network status monitoring systems allow analyzing the processes occurring in the network, therefore, a visual view was presented through t-SNE.

ISSN 3030-3907

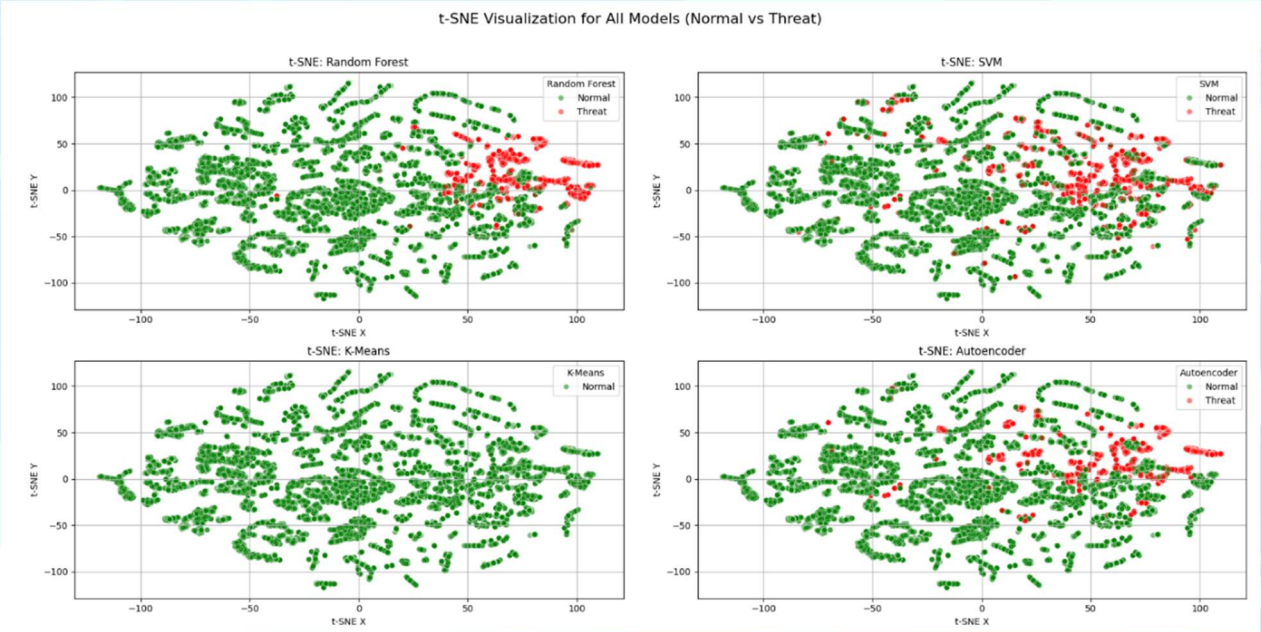


Figure 3. t-SNE visualization of network traffic data.

CONCLUSION: This research combines machine learning and deep learning methods into an artificial intelligence network detection system. The research work helps to distinguish between normal and malicious traffic in the network flow. The proposed hybrid artificial intelligence model provides a better security solution compared to standard network detection systems that use conventional feature-based detection as the main method.

Complex detection is difficult to implement through deep learning models, as such models require significant computational resources. Explaining DL-based decisions using SHAP is less efficient than traditional rule-based systems.

Further research will be aimed at developing real systems and improving protection, while improving operational reliability using modern understanding-based artificial intelligence methods. The completed research will serve as a basis for the development of future network detection systems.

The proposed research presents a network detection system by combining machine learning and deep learning methods with artificial intelligence. It demonstrated this by distinguishing between normal and anomalous traffic using t-SNE visualizations.

REFERENCES

1. Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE access, 6, 33789-33795.

2. Al-Qatf, M., Lasheng, Y., Al-Habib, M., & Al-Sabahi, K. (2018). Deep learning approach combining sparse autoencoder with SVM for network intrusion detection. Ieee Access, 6, 52843-52856.

ISSN 3030-3907

3. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., and Seale, M., "Explainable Intrusion Detection Systems (X-IDS): A survey of current methods, challenges, and opportunities," IEEE Access, vol. 10, pp. 112392–112415, 2022. doi: 10.1109/ACCESS.2022.3216617.
4. M. Ahsan and K. Nygard, "Convolutional neural networks with LSTM for intrusion detection," ResearchGate Preprint, 2020, doi:10.13140/RG.2.2.24796.82567.
5. M. Ahsan and K. Nygard, "Convolutional neural networks with LSTM for intrusion detection," ResearchGate Preprint, 2020, doi:10.13140/RG.2.2.24796.82567.
6. Khan, L. U., Yaqoob, I., Tran, N. H., Kazmi, S. A., Dang, T. N., & Hong, C. S. (2020). Edge-computing-enabled smart cities: A comprehensive survey. IEEE Internet of Things journal, 7(10), 10200-10232.
7. R. Lazzarini, H. Tianfield, and V. Charissis, "Federated learning for IoT intrusion detection," AI, vol. 4, no. 3, pp. 509–530, 2023, doi:10.3390/ai4030028.
8. A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," Knowledge-Based Systems, vol. 189, p. 105124, Jan. 2020, doi:10.1016/j.knosys.2019.105124.
9. Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems, 189, 105124.
10. Zhuo, S., Hong, Y. Y., & Palaoag, T. D. (2022, December 14). AN INTELLIGENT CYBER SECURITY DETECTION AND RESPONSE PLATFORM. International Journal for Research in Advanced Computer Science and Engineering, 8(12), 1–10. https://doi.org/10.53555/cse.v8i12.2167

ISSN 3030-3907