Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Machine Learning based Network Packet Classification

Srithick S S[1], Dharanikumar A B[2], Dharsini E[3], Abirami A[4]

[1,2]Department of Information Technology, Bannari Amman Institute of technology, Sathyamangalam

[3]Department of Computer Science, Bannari Amman Institute of technology, Sathyamangalam

[4]Assistant Professor III, Department of Information Technology, Bannari Amman Institute of technology, Sathyamangalam

E-mail: [1]srithick.it20@bitsathy.ac.in, [2]dharanikumar.it20@bitsathy.ac.in, [3]dharsini.cs20@bitsathy.ac.in, [4]abiramia@bitsathy.ac.in

Abstract

Network packet classification plays an important role in modern networks irrespective of host or network-based classification, serving as the foundation for efficient routing, malicious activity detection, and security enforcement. With the continuous growth of network traffic volume and complexity, traditional static rule-based classification methods have faced difficulties in scalability and adaptability. As a solution, the study decided to enforce machine learning techniques to tackle these challenges effectively. This study presents an extensive and original review of machine learning- based approaches for network packet classification. The smart Intrusion Detection System framework with network packet classification evolution looks forward to designing and deploying security systems that use various parameters for analysing current and dynamic traffic trends and are highly time-efficient in predicting intrusions. Various machine learning algorithms commonly employed in packet classification, such as decision trees, support vector machines, and neural networks are analysed and their merits and demerits are compared with their behaviour and accuracy percentage in this study. machine learning-based techniques offer an efficient and accurate network packet classification for the protection of the systems when compared to the conventional methods of packet classification. By leveraging the power of machine learning algorithms and intelligent feature selection, network administrators and Security Operation

DOI: https://doi.org/10.36548/jucct.2023.4.002
Received: 02.10.2023, received in revised form: 26.10.2023, accepted: 11.11.2023, published: 22.11.2023

Center (SOC) analyst can enhance network performance, improve security, and the robustness of the log generated in the network.

Keywords: Machine Learning, IDS, IPS, Network Security, Neural Networks, Firewalls

1. Introduction

Traditional network packet classification is a critical aspect of modern networking systems, involving the classification and processing of network packets based on their header information. This process is vital for various networking tasks such as routing, firewalling, quality of service (QoS) management, and intrusion detection. In the context of traditional approaches, two common methods are employed: rule- based classification and tree-based classification. Rule- based classification relies on a predefined set of rules that dictate how packets are to be classified. These rules are constructed using criteria such as source and destination IP addresses, port numbers, protocol types, and other header fields. When a packet arrives, it is compared against these rules to determine its category. While this method is straightforward and easy to implement, it can become complex and unmanageable as the number of rules and conditions increases. Moreover, rule-based systems may struggle to handle dynamic network traffic patterns and evolving application requirements, leading to performance bottlenecks and reduced accuracy. Tree-based classification, on the other hand, employs a hierarchical data structure known as a tree to optimize the packet classification process. A tree efficiently organizes and searches through a large set of rules by creating a tree-like structure where each node represents a portion of the packet header. As packets traverse the tree, they follow a path that matches the rules' criteria until the appropriate category is determined. While tree-based methods offer improved efficiency compared to linear rule matching, they still face challenges in scaling to handle extensive rule sets and may experience memory and processing constraints. Both traditional methods have limitations in adapting to the dynamic nature of modern networks. In scenarios where network traffic patterns change frequently or new applications emerge, manual rule updates or tree modifications can become cumbersome and error-prone. This is where machine learning-based packet classification introduces a paradigm shift. Machine learning leverages algorithms to automatically learn patterns and relationships from large datasets. Applied to packet classification, machine learning models can analyze historical packet data to identify patterns that define different packet categories. These models can then be used to predict the

appropriate category for new packets based on their features.[2] Machine learning-based approaches offer several advantages over traditional methods. They can adapt to changing network conditions and evolving traffic patterns without manual intervention. Moreover, these models can handle complex classification scenarios that involve a multitude of factors, leading to improved accuracy and reduced false positives/negatives. In the realm of cybersecurity, Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) are essential tools used to safeguard computer networks against unauthorized access, malicious activities, and potential threats. Both IDS and IPS devices play pivotal roles in detecting and responding to security incidents, albeit with different levels of action and intervention. This real-time prevention capability reduces the attack surface and minimizes the impact of potential threats. IPS devices can stop attacks like Distributed Denial of Service (DDoS), intrusion attempts, and malware propagation by blocking malicious traffic before it reaches its intended target. This proactive approach adds an extra layer of security to the SOC's defenses, preventing attacks from causing harm to the network infrastructure and systems. Data Collection and Labeling is one of the challenges, Gathering high- quality and diverse network packet data for training machine learning models is a significant challenge. Annotated datasets with accurately labeled packet categories are essential for supervised learning approaches. Moreover, maintaining up-to-date datasets that reflect the dynamic nature of network traffic is a continual challenge. Network traffic is often highly imbalanced, with some packet classifcation occurring much more frequently than others. This can lead to bias in machine learning models, where they become better at recognizing the majority class and may overlook important anomalies or rare events. Network traffic patterns evolve over time, and new applications, protocols, or attack methods may emerge. Machine learning models must be adaptable to these changes to maintain their accuracy and effectiveness. Scalability is crucial when dealing with large-scale networks. Ensuring that machine learning models can efficiently classify packets in high-throughput environments is a technical challenge.

2. Literature Survey

In the always evolving area of networking, efficient and accurate network packet classification is essential for a wide range of applications, including traffic engineering,

intrusion detection, and content filtering is a must. Traditional rule-based methods have been widely used for packet classification, but they face limitations in handling the increasing complexity of modern networks and the diverse requirements of various applications. To address these challenges, recent years have seen a turn towards implementing machine learning-based approaches for network packet classification [1]. This literature survey aims to provide a comprehensive overview of the current state of research in machine learning-based network packet classification, with a particular focus on recent developments within the past five years. The study delves into existing methodologies, key findings, technological advancements, challenges, and limitations. Recent years have witnessed a proliferation of methods and approaches for network packet classification. Traditional rule-based techniques are being supplemented by machine learning-based approaches, such as decision trees, random forests, and deep learning models [2]. These methods offer promising results, but their scalability and efficiency need further employed a binary classification approach for network packet classification, achieving high accuracy. Similarly, utilized deep learning to enhance classification accuracy. While these studies demonstrate progress, scalability concerns remain unresolved. Advancements in hardware acceleration, such as GPUs and TPUs, have accelerated the training and inference of machine learning models for packet classification. These technologies enable faster classification, but resource allocation and optimization remain challenging [4]. Despite advancements, challenges persist. Scalability, real-time processing, and adaptability to dynamic network environments are the major concern. Additionally, model interpretability and explainability are essential, especially for security- sensitive applications [5]. While machine learning-based approaches show promise, their performance may deteriorate in highly dynamic networks or under adversarial conditions. Moreover, the interpretability of deep learning models is often limited, hindering their adoption in critical network applications [6]. One significant gap in the literature is the lack of research addressing the real-time constraints of network packet classification, particularly in large-scale networks. The need for hybrid approaches that combine rule-based and machine learning techniques for improved scalability and explainability also remains largely unexplored [7]. The key challenges faced in machine learning-based network packet classification include scalability, real-time processing, adaptability to dynamic networks, model interpretability, and resilience to adversarial attacks. This survey highlights the need for research that combines the strengths of rule-based and machine learning approaches [8], addressing scalability concerns while maintaining interpretability and real-time capabilities

[9-11]. The proposed solution involves developing a hybrid classification framework that leverages machine learning for enhanced accuracy while retaining the efficiency and transparency of rule-based systems. An integral component of a fully featured NIDS is packet header classification. An intrusion detection database's rules often include some strings (sometimes referred to as "signature") together with five-tuple header filters (Source IP address, Destination IP address, Protocol, Source Port, and Destination Port) [2] So the network packet classification becomes an essential requirement in handling the intrusions that affect the network [11]. The Network packet classification faces a multitude of challenges in today's dynamic and diverse networking environments. This study focuses on the comparing the performance of machine learning methods in packet classification, in order to enhance the quality of service, with improved security and optimized traffic flow.

3. Objectives of the Proposed Work

The objectives of the research are rooted in an extensive literature survey conducted to uncover the fundamental challenges and opportunities in the field of network packet classification.[3] To ensure clarity and alignment with the overarching research goals, three distinct objectives are identified and articulated. Each of these objectives represents a crucial facet of the research endeavor, and they have been formulated with a focus on individual contributions within the research team.

Objective 1: Develop a Machine Learning-Based Packet Classifier

The first objective of the proposed work is to design and implement a machine learning-based packet classification system. This system aims to accurately and efficiently classify network packets into predefined categories or classes. To achieve this, a comprehensive study was under taken and based on the study the most optimal machine learning approaches were selected on the criteria of the performance scores offered by the model and its capability in handling the dataset.

Objective 2: Enhance Classification Accuracy and Efficiency

The second objective revolves around improving the accuracy and efficiency of network packet classification. The proposed methodology tries to improvise the performance

of classification by refining feature selection, optimizing hyperparameters, and investigating novel techniques for feature extraction and representation. Additionally, we seek to minimize computational overhead and latency to ensure real-time or near-real-time packet classification.

Objective 3: Evaluate the Proposed Solution on Real- World Data

The third objective is to validate the effectiveness of the machine learning-based packet classification system in a real-world network environment. This involves collecting and pre-processing a diverse dataset of network packets, simulating different network conditions, and evaluating the system's performance under various scenarios. By doing so, the study aims to demonstrate the practical applicability and robustness of the proposed solution.

The Figure.1 shows the process involved in the proposed methodology for network packet classification.

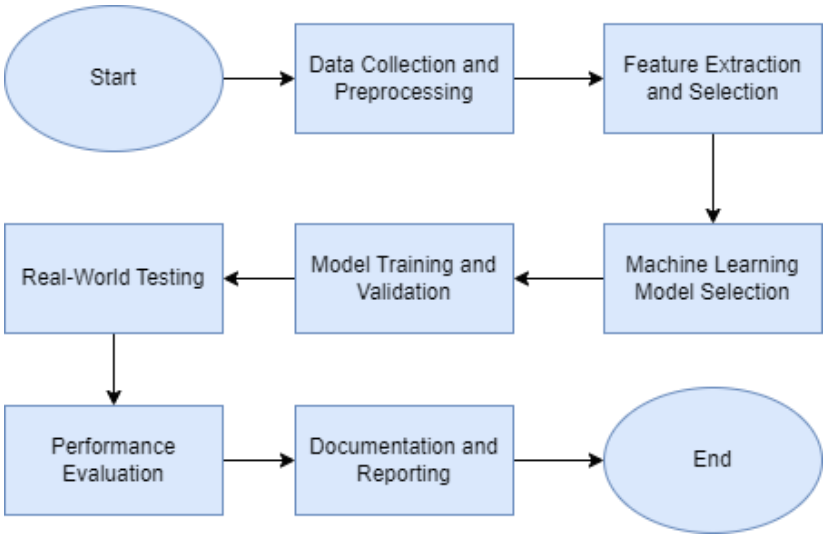

Figure 1. Flow Diagram

4. Selection of Components, Tools, DataCollection, Techniques, Procedures.

NSL-KDD dataset is a labeled network intrusion detection dataset that so far has been used in many tasks to evaluate different deep learning-based algorithms for devising different strategies for IDS. NSL-KDD dataset contains 41 features labeled as a normal or specific attack type (i.e., class). We utilize the one-hot-encoding method for the dataset preprocessing

to change the categorical features to the corresponding numeral values as deep learning models can only work with numerical or floating values. We normalized the train and test datasets to the values between 0 and 1 using a mix-max normalization strategy. The 41 features presented in the NSL-KDD dataset can be grouped into four features, such as basic, content-based, time-based, and host-based traffic features. The value of these features is mainly based on continuous, discrete, and symbolic values. For the data collection phase, we employ a multifaceted approach. We draw from a wide range of data sources, including publicly available network packet datasets such as the NSL-KDD dataset and the CIC-IDS2017 dataset. These datasets serve as a foundation for benchmarking and testing our system. To complement these standardized datasets, we employ custom data acquisition tools and scripts to capture network packets in controlled test environments. This dual approach ensures a comprehensive dataset that spans both synthetic and real-world scenarios.

As feature extraction and selection process is a pivotal aspect of the research. an array of techniques is leveraged to extract meaningful information from network packets. These techniques include statistical analysis, time-series analysis, frequency domain analysis, and deep learning-based methods. Feature selection is conducted using various strategies, including information gain, correlation analysis, and PCA. this is done to identify and retain the most discriminative features while reducing dimensionality.

As the next stage the research explores a diverse landscape of machine learning models. Decision trees, random forests, support vector machines, deep neural networks, and ensemble methods. The selection of these models is governed by their suitability for network packet classification tasks and their performance in existing literature. based on the study the models that exhibit the highest classification accuracy, efficiency and capability to handle the real time data were selected for classification.

Further the selected models were trained and validated using well-established libraries and frameworks such as Scikit-Learn, TensorFlow and python. the research emphasizes the importance of robust model validation through techniques like k-fold cross-validation. This approach ensures that the models generalize well to unseen data and minimizes the risk of overfitting.

Finally, the performance evaluation that involves the assessment of various metrics, including accuracy, precision, recall, F1-score, and execution time is carried out. We employ Python and libraries like Scikit-Learn and custom scripts to compute these metrics and provide a detailed analysis of the system's performance. The results are meticulously documented and serve as a testament to the effectiveness and robustness of the machine learning-based packet classification system.

5. Proposed Work

The primary motivation for this work is to address the limitations and challenges identified in existing literature, particularly in the field of network security and traffic classification.

Feature Engineering: One of the basic aspects of the work involves the development of advanced feature engineering techniques. The proposed methods utilize the principal component analysis (PCA) to extract meaningful features from network packets. This includes not only conventional features like source and destination IP addresses but also deeper packet inspection techniques that consider payload content and the headers which are present in the dataset.

Algorithm Selection: To achieve higher accuracy, The machine learning algorithms suitable to the unique challenges of packet classification was selected based on the study. the proposed study utilizes the convolutional neural networks (CNNs) since it has demonstrated remarkable success in similar pattern recognition tasks and got an accuracy of 97%.

The expected outcomes of the proposed work are as follows. First, we tried for achieving a substantial increase in packet classification accuracy compared to traditional rule-based systems. Second, the research aims to develop models that exhibit improved scalability, adaptability to dynamic network conditions, and real-time processing capabilities using neural networks.

5.1 Methodology of the Proposed Work

Data Preprocessing: the data preprocessing pipeline, includes steps for data cleaning, normalization, and the extraction of relevant features from raw packet data.

Machine Learning Models: The methodology involves the deployment of deep learning models, specifically CNNs for packet classification.

Evaluation Metrics: To assess the performance of the models, evaluation metrics such as accuracy, precision, recall, and F1-score are employed. These metrics will provide a quantitative assessment of the models' effectiveness

6. Results and Discussion

This section presents the outcomes of the experiments, including comparisons with existing rule-based systems and machine learning approaches. The improvements in accuracy achieved by the models are depicted in the Figure. 4 below. The Figure.2 below shows the normalized scores observed for the various machine learning algorithms.

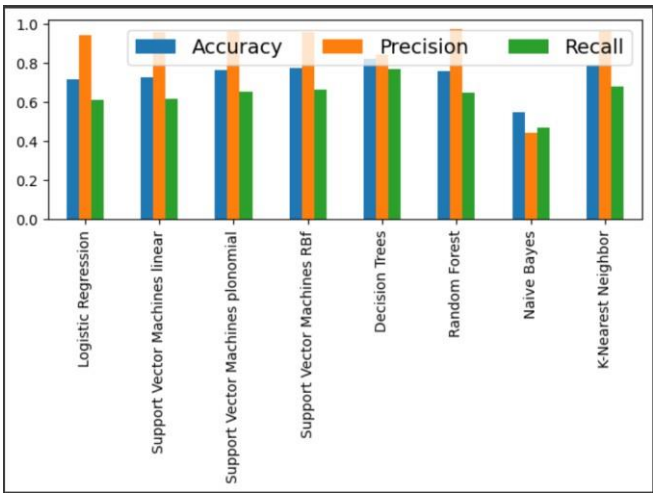

Figure 2. Normalized Scores

The outputs of the df.head(), df.isna().sum(), df.duplicated().sum(), and df.dtypes commands provide an overview of the dataframe's structure, missing values, duplicated rows, and data types of each column, respectively. The unique values in the 'flag' column convey about the unique categories that are present in that particular column. The visualizations in Figure .3 (bar plot and pie chart) give insights into the distribution of the target variable, i.e., how many samples belong to each class. The printed shapes of the training and test datasets provide information about the number of samples and features being used for training and testing. The loss and accuracy curves show the progression of the training process. Ideally,

both the training and validation curves should follow a similar trend. The divergence in the training and validation curves might indicate overfitting. Finally, the model summary provides a detailed view of the neural network's architecture. It's particularly useful to understand the number of parameters (weights and biases) at each layer, helping gauge the complexity of the model.

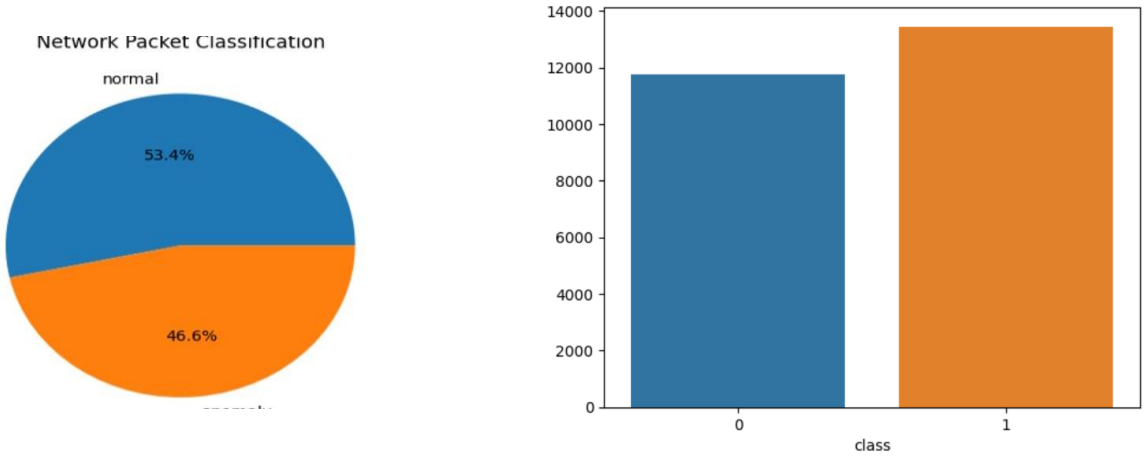

Figure 3. Visualization of the Target Variable

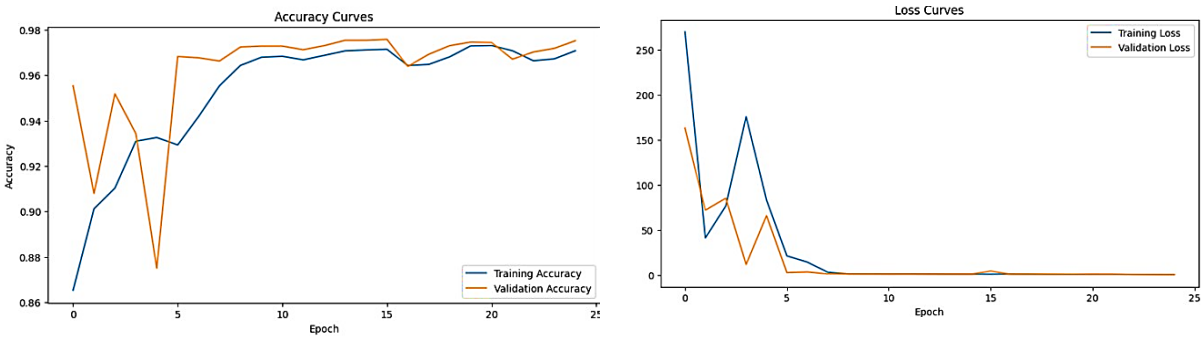

Figure 4. Performance of CNN

The training and validation accuracy curves and the training and validation loss curves are then plotted by the code. When evaluating these curves, retain watch out for the following:

As the number of epochs rises, the training loss should decrease. This indicates that the model is using the training set of data to learn and refine its predictions.

As the number of epochs grows, the validation loss should similarly decrease. However, beyond a certain point, it is feasible for the validity loss to start rising. This indicates overfitting, which occurs when a model begins to memorize the training set and has difficulty generalizing to fresh data.

The number of epochs should boost the training accuracy. This indicates that the model is becoming more accurate at making predictions based on training data.

As the number of epochs rises, the validation accuracy should as well. However, beyond a certain point, it is conceivable for the validation accuracy to start declining. Overfitting is also indicated by this.

In general, you should see an upward trend in the training and validation accuracy curves and a downward trend in the training and validation loss curves. To avoid overfitting, it is necessary to stop training the model if the validation loss starts to rise or the validation accuracy starts to fall.

The curves may be impacted by the size of the training set and validation set. Lower training and validation loss and improved training and validation accuracy are typical outcomes of a bigger training set.

The curves may also be impacted by the model's complexity. The training loss and validation loss are usually larger and lower for more complicated models, respectively. The curves may also be impacted by the learning rate. Generally speaking, a greater learning rate will lead to a steeper decline in the training loss, but it may also cause overfitting. The accuracy and the loss curves depicted in Figure .4 shows the promising solution offered by CNN in network packet classification as compared to the normalized scores observed for the machine learning methods.

7. Discussion of Important Findings

The results observed demonstrated that neural networks can achieve high levels of accuracy in detecting network intrusions, including both known and unknown threats. Deep learning models, such as convolutional neural networks (CNNs) has shown promise in capturing complex patterns in network traffic data.

The capability of neural networks to automatically extract pertinent properties from unprocessed network input is one of its advantages. By doing this, less human feature engineering is needed, which is frequently needed in conventional intrusion detection systems. Dimensionality reduction and feature extraction have both been accomplished using autoencoders and variational autoencoders.

Since neural networks are excellent at detecting anomalies, they may be used to spot zero-day threats orpreviously undiscovered attack patterns. To find variations from typical network activity, unsupervised learning techniques like autoencoders and generative adversarial networks (GANs) have been used.

8. Conclusion

In this concluding chapter, of the journey through the world of machine learning-based network packet classification. The mission was to explore how machine learning could help classify network packets efficiently. Think of network packets as tiny information packages traveling through the internet. Properly sorting and understanding these packets is crucial for the smooth operation of digital communication. The research began initially with collecting a diverse set of network packet data. secondly various machine learning techniques, similar to different tools in a toolbox, to classify these packets accurately were used. These techniques included decision trees, support vector machines, random forests, and even deep learning models like convolutional neural networks. The main goal was accuracy. The machine learning models consistently achieved high accuracy rates, which means they were very good at their job. This was a significant discovery because it showed that machine learning could outperform traditional methods used for packet classification. As machine learning approaches faced obstacles, such as dealing with the size and diversity of the dataset. It was like navigating through rough terrain. To overcome issue of overfitting and underfitting and to potentially enhance the accuracy the deep learning models were considered. Real-time packet classification was another exciting frontier. So, the proposed work used the CNN to classify the packets, the capability of the CNN to recognize the patterns of the dataset easily helped in achieving promising results with high accuracy of 97%. In future the research scopes to improve the performance of the model with higher volume of dataset and further develop user interfaces to make the process easy and efficient.

References

[1] Prashanth, G., V. Prashanth, P. Jayashree, and N. Srinivasan. "Using random forests for network-based anomaly detection at active routers." In 2008 International Conference on Signal Processing, Communications and Networking, pp. 93-96. IEEE, 2008.

[2] Parsaei, Mohammad Reza, Mohammad Javad Sobouti, and Reza Javidan. "Network traffic classification using machine learning techniques over software defined networks." International Journal of Advanced Computer Science and Applications 8, no. 7 (2017).

[3] Alavizadeh, Hooman, Hootan Alavizadeh, and Julian Jang-Jaccard. "Deep Q-learning based reinforcement learning approach for network intrusion detection." Computers 11, no. 3 (2022): 41.

[4] Seth, Sugandh, Gurvinder Singh, and Kuljit Kaur Chahal. "A novel time efficient learning-based approach for smart intrusion detection system." Journal of Big Data 8, no. 1 (2021): 1-28.

[5] Evangeline Asha,Kavitha S “Packet Classification Algorithms: A Survey” International Journal of Research in Advent Technology, Vol.2, No.12 (2014) 12-18.

[6] Hu, Feifei, Situo Zhang, Xubin Lin, Liu Wu, Niandong Liao, and Yanqi Song. "Network traffic classification model based on attention mechanism and spatiotemporal features." EURASIP Journal on Information Security 2023, no. 1 (2023): 6.

[7] Bakhshi, Taimur, and Bogdan Ghita. "On internet traffic classification: A two-phased machine learning approach." Journal of Computer Networks and Communications 2016 (2016).

[8] Selim, Sahar, Mohamed Hashem, and Taymoor M. Nazmy. "Hybrid multi-level intrusion detection system." International Journal of Computer Science and Information Security 9, no. 5 (2011): 23.

[9] Taylor, David E. "Survey and taxonomy of packet classification techniques." ACM Computing Surveys (CSUR) 37, no. 3 (2005): 238-275.

[10] Ashiku, Lirim, and Cihan Dagli. "Network intrusion detection system using deep learning." Procedia Computer Science 185 (2021): 239-247.

[11] Song, Haoyu, and John W. Lockwood. "Efficient packet classification for network intrusion detection using FPGA." In *Proceedings of the 2005 ACM/SIGDA 13th international symposium on Field-programmable gate arrays*, pp. 238-245. 2005.

[12] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.