Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN (O): 2320-5407
ISSN (P): 3107-4928

Article DOI:10.21474/IJAR01/23881
DOI URL: http://dx.doi.org/10.21474/IJAR01/23881

RESEARCH ARTICLE

MACHINE LEARNING ALGORITHMS IN NETWORK TRAFFIC CLASSIFICATION – A SURVEY

Maninder Singh Zandu[1] and Sandeep Kad[2]

1. Research Scholar, Department of CSE, Amritsar Group of Colleges, Amritsar.
2. Centre for Research and Innovation, Amritsar Group of Colleges, Amritsar.

Manuscript Info

Manuscript History
Received: 16 May 2026
Final Accepted: 18 June 2026
Published: July 2026

Key words:-
Machine Learning, Network Traffic, Feature extraction, Traffic analysis

Abstract

Effective NTC plays a vital role in enhancing bandwidth efficiency, ensuring network security, and maintaining Quality of Service (QoS) in today's communication systems. Conventional methods like port-based as well as payload-based classification are no more reliable because of the rise of encoded traffic and dynamic applications. As an alternative, Machine Learning (ML) approaches can automatically identify patterns in the statistical characteristics of network flows, offering more flexibility and accuracy. This paper presents a survey of Machine Learning that was used in NTC. The main focus on latest advances such as Protocol Detection, Traffic Behavior Analysis and Application Identification.Resent Trends, Evaluation Metrices, Benchmark Datasets and Machine Learning Algorithms are the core things that was discusse d in this review. Along with Class Imbalance, Concept drift and encrypt ed traffic, these kinds of problem were also discussed in it. Beside these things, to understand the ML algorithms and make baseline benchmark as KDD dataset to represent comparative analysis.

Introduction:-

HTTP/3, QUIC-based communication, cloud-native applications, edge computing and so many Internet of Things (IOT) devices rapidly adopted technologies in today's era changed the technology totally. Due to these methods, the use of Network Traffic is vast and using in various things, as a result of these Management, Cybersecurity, Quality of Services (QoS) and Intelligent Resource Allocation makes an important role in Network Traffic Classification. One of the major developments that has transformed modern communication to such an extent that the internet has a deep penetration level in all aspects of human life is the rapid advancement of secure communication technologies [1]. The progression along this path has led not only to the rise in volume and heterogeneity of traffic over the network but also in existence of these kinds of traffic, which in turn has made its management grow in complexity [2]. Concurrently, encryption technologies are being adopted to a large extent so as to protect privacy and thwart any unauthorized access. Encryption certainly adds an extra layer of defence; however, it makes the process of traffic monitoring and analysis more laborious and thus creates a peculiar problem for network operators and security specialists. In this changed scenario, network traffic classification (NTC) is referred to as a pivotal research area [3]. It denotes the process of associating network flows with a particular class,

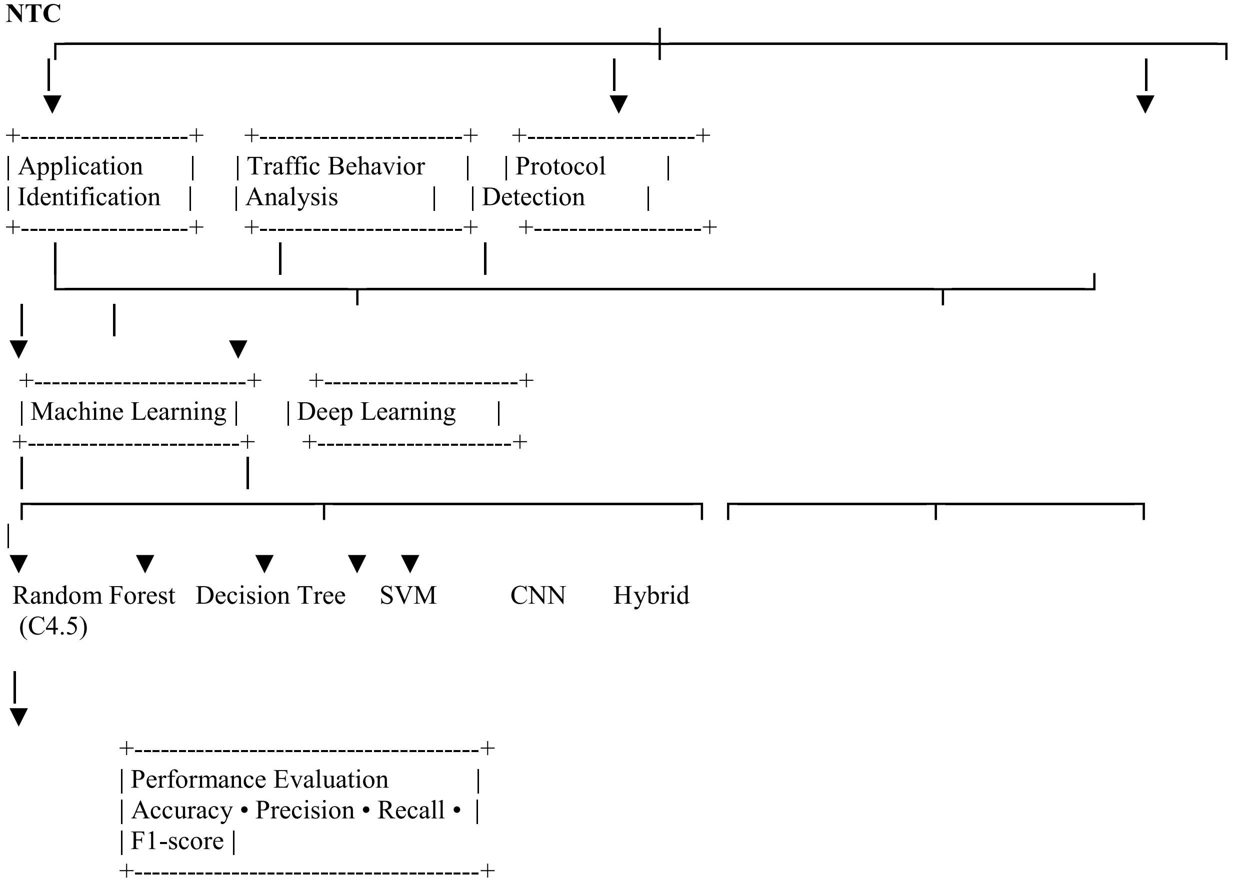
Corresponding Author:- Maninder Singh Zandu
Address:-Research Scholar, Department of CSE, Amritsar Group of Colleges, Amritsar.

e.g., that of applications, services, or threats, depending on the target. Indispensable can be the role of proper classification in elevating quality of service, facilitating bandwidth allocation, and corroborating the presence of cyber threats or malware attacks at an early stage. The lack of classification accuracy may lead to organizations suffering from degraded service performance, inefficient utilization of resources, and response to attacks being late [4].

Traditional methods for traffic classification, like port-based analysis or payload inspection, have a hard time keeping up with the complexity of modern traffic. Port-based analysis was once seen as a simple and efficient method. It relied on the idea that specific applications use fixed port numbers. This assumption is outdated because many modern applications now use dynamic or random ports [5]. Some applications even hide their traffic by using well-known ports, like port 80 for HTTP, to evade detection. Similarly, payload-based inspection looks at packet content to find application signatures. However, this method has become less effective with the advent of encryption protocols such as Hyper Text Transfer Protocol Secure, TLS& QUIC, which hide packet contents from direct inspection [6]. Additionally, statistical feature-based methods depend on flow-level features like packet length, inter-arrival time, or flow duration. This survey didn't cover only Intrusion detection, it also addresses the large part of Network Traffic Classification.

This review primarily focuses on recent papers especially from 2023 to 2025, so that we will highlight the latest techniques that was recently introduced.

NTC



Literature Review:-

In the classification of Network Traffic, the use of machine learning is rising daily to prevent different types of intrusions using new innovative techniques.Existing innovations have expanded and spread across many applications such as identifying services or applications, analyzing headers and payloads of packets, detecting even the packets are encrypted andidentifying IoT traffic. In this part, the mentioned related areas give an overview of it, and themajor issues, techniques and contribution are discussed below.

ML Techniques for Application Identification:-

J. Kotak et al. [7] determined how well machine learning (ML) works to identify services or apps from network chatter. In the task of traffic identification based on statistical data and the headers and payloads of packets, machine learning algorithms achieved great accuracy even when network traffic was encrypted. H. Yzzogh et al. [8] investigated the possibilities of Convolutional Neural Networks (CNN), which were renowned for their exceptional outcomes in tasks to classifying images. In order to take use of Convolutional Neural Networks in this situation, they suggested employing the flow-to-bar chart conversion method to turn traffic flow data into pictures. H. Lu et al. [9] suggested a method for detecting unknown network traffic that used a cascade structure and confidence (difference) by examining the confidence distributions of the new and known classes. The suggested approach operated as follows. X. Xiao et al. [10] Suggested Robust Byte-Label Joint Attention Network (RBLJAN), a robust and effective deep learning-based architecture for classifying encrypted traffic over the network at the packet& flow stages. RBLJAN consists of an adversarial traffic generator and a classifier. By capturing implicit correlations between bytes and labels, the classifier used techniques like byte-label joint attention learning and header-payload parallel processing to create potent packet representations. Z. Wang et al. [11] proposed deep flow embedding (DFE), a feature compressor-based deep learning method based on metric learning. C.-Y. Shin et al. [12] offered an inductive refutation of the nebulous notion such deep learning models could do properly and beat decision machine learning models in every way over domains, particularly inside the area of classification of traffic over the network. In theTable 1 below, we will see the ML techniques for Application Identification in detail.

The review of recent papers indicates a clear that in the field of ML, rapidly changes techniques from Traditional handcrafted statistical features to Ensemble-based and Deep learning technology. These new techniques are more helpful in improving characteristics and analyzing properly. Classical Algorithms such as Random Forest, Decision Tree (C4.5) and Support Vector Machine are good and gives good result over Benchmark Datasets but their activity mostly depends on Handcrafted Features and in case of Encrypted as well as for various types of Heterogeneous Network performance may decline.

We evaluate recent studies below in the shape of table, selective machine learning techniques for Application Identification, emphasizing the traffic type and other parameters presented in the literature.

Table 1 Comparative Analysis of ML Techniques for Application

Ref.	ML Technique	Traffic Type	Dataset(s) Used	Evaluation Parameters	Key Outcomes
7	Time-series-based ML method for encrypted & VPN traffic identification	Encrypted &VPN traffic	ISCX VPN-nonVPN	Accuracy, Precision, Recall	Effective even for VPN-encrypted traffic.
8	CNN, SIFT (Scale-Invariant Feature Transform)	Network Traffic Flows converted into images for traffic classification	InSDN dataset	Classification Accuracy	NN maintained high accuracy while reducing computational overhead.
9	Cascade structure-based ML (adaptive threshold algorithm)	Encrypted traffic	Real-world network traffic	Overall accuracy and classification	Detected unknown traffic classes

The studies summarized in the table 1 indicates that classification accuracy varies from 90% to 99% and might be more than this. These all results depend upon ML techniques, characteristics of traffic and benchmark dataset.

ML Techniques for Traffic Behavior Analysis:-

Statistical feature-based methods depend on flow-level attributes like packet length, inter-arrival time, or flow duration. While these methods can classify traffic without examining payloads, they need extensive manual feature engineering and large labeled datasets for training. This process is both time-consuming and prone to errors. These methods also have trouble generalizing across different networks, devices, and application versions [10]. The statistical properties of traffic can change significantly with even small software updates. Another shortcoming of traditional techniques is their high computational demands. Payload inspection is particularly resource-intensive since it often requires deep packet inspection (DPI). This makes it unsuitable for high-speed networks where low latency matters [11]. Additionally, gathering traffic over long periods and depending on caching and storage can drive up infrastructure costs. It also raises issues about user privacy, as sensitive data might be exposed during inspection. These challenges emphasize the urgent need for smarter, automated, and encryption-resistant classification techniques. Researchers, in an attempt to bypass the constraints of classical methodologies, began exploring machine learning (ML) techniques to categorize network traffic [12]. Models like Decision Tree (C4.5), Random Forests, k-NN, Naïve Bayes&SVM have been commonly used to classify traffic flows derived from statistical features like packet size distribution, duration of flow, and times between arrivals. These models improved on port-based and payload-based inspection by allowing classification even when traffic was somewhat encrypted or disguised [13]. In the Table 2 below, we will see the ML techniques for Traffic Behavior Analysis in detail. The analysis of the reviewed papers indicates that not a single model of ML gives you accurate or good result in Traffic Behavior Analysis. In other words, Ensemble methods gives more robustness and classification accuracy, therefore DL techniques understands Behavior Patterns and Temporal to train the model in a better way. Recent researches shown some techniques of Machine Learning which are used in the Traffic Behavior Analysis that we will show in the table below.

Table 2 Comparative Analysis of ML Techniques for Traffic Behavior Analysis

Ref.	ML Technique	Traffic Type	Dataset(s) Used	Evaluation Parameters	Key Outcomes
14	Robust Byte-Label Joint Attention Network (RBLJAN) with adversarial training	Multiple real-world network traffic	X-APP, X-WEB, USTC-TFC, and ISCX-VPN	Precision, F1-score	High robustness, fast detection, strong resistance to adversarial attacks
15	Deep Flow Embedding (DFE) using metric learning	Network flow traffic	CIC-IDS2017	Precision, Recall	Robust to traffic variation, effective new-class identification
16	Tree-based ML vs lightweight DL with data augmentation	Network flow traffic	ISCX VPN-nonVPN 2016	Accuracy, Specification	Suitable for resource-constrained environments

ML Techniques for Protocol Detection:-

W. Lin et al. [17] presented a novel Information Bottleneck Neural Network (IBNN) to minimize extraneous information while effectively collecting critical features from raw traffic data. Y. Jang et al. [18] examined the based-on payload and port-numbernetwork traffic classification systems currently in use in order to determine their drawbacks within the current network context. Y.Wang et al.[19]proposed Ulfar, a multi-level feature attention method for classifying network traffic that used CNN as the foundation for the deep packet analysis model. To classify network traffic in Ulfar, they took a single packet per flow.Z. Bu et al. [20] offered a neural network model for the classification of encrypted network traffic that has deep and parallel network-in-network (NIN) structures. In order to improve local modeling, NIN adopted a micro network following each convolution layer in contrast to conventional CNN. In the Table 3 below, we will see the ML techniques for Protocol Detection in detail.

The Review of studies on Protocol Detection indicates that ML techniques improves Network Protocol specifically in the area of Encrypted Communications but used in the different Experimental Setup, Evaluation Matrices and Benchmark Datasets are not same that's why it's not easy to comparison the results. In recent papers, we observed more focus gives over Hybrid Classification Models and Robust Feature Learning. So that we can handle upcoming challenges in Encrypted Traffic and changes in the protocols.

Some protocol detection related techniques comparisons are used in the below table based on machine learning in detail.

Table 3 Comparative Analysis of ML Techniques for Protocol Detection

Ref.	ML Technique	Traffic Type	Dataset(s) Used	Evaluation Parameters	Key Outcomes
17	IBNN (Information Bottleneck Neural Network)	Encrypted Network Traffic	UNB CIS dataset	Accuracy, F1-score	Extracted critical features while minimizing redundant information
18	NFStream + ML models (GBTree, LightGBM, XGBoost)	Encrypted traffic dataset	ISCX VPN-nonVPN 2016	F1-score, Precision, Recall.	Very high classification performance, effective for encrypted traffic.
19	CNN for deep packet analysis	Single packet per flow	InSDN, CIC-DDoS2019	Classification Accuracy	Automatically recovering format-related bytes
20	NN with NIN	Encrypted dataset	ISCX VPN-nonVPN	Classification Accuracy, F1-score	Achieved a balance between accuracy and model complexity

Research Gaps and Motivation:-

In this review, the literature identifies many difficulties of NTC. One of the most major problem that we understand is Manual Feature Engineering, because traditional machine learning techniques rely on handcrafted Statistical Features. That's why so much time has been to make a model and adaptability is limited not fit for all environments[13].As recent papers including Shin et al. and Lu et al. [16] indicates that statistical handcrafted features especially in the case of Encrypted Traffic decrease the efficiency of generalization of the model. It represents that the manual feature engineering is the main obstacle as plays a roleof bottleneck in NTC.

The typical NTC process has various steps which include data input, pre-processing, extraction of features, classification, and performance analysis. The data are collected as input from the authentic source. During the pre-processing phase missing and redundant values from the dataset will be removed which will clean the input and help to achieve maximum accuracy for the classification. In the phase of feature extraction relation among each attribute with the target, the set will be established and it helps to identify main features from the dataset. In the phase of classification, a model of classification will be applied to the training and test set which will provide output as the predicted set. In the last phase, various performance analysis metrics like accuracy, precision, recall will be used to analyze the performance of the network traffic classification model. Network traffic classification is a crucial part of network management and rapid advances in machine learning have driven the utility of learning methods to categorize traffic across a network. Inherent traits of internet networks initiate imbalanced class distributions in the case of consistent databases. This phenomenon is called class imbalance and it is attracting growing attention in a large number of research fields. The propagation of encrypted traffic has emerged machine learning (ML) based network traffic categorization (NTC) as a mainstream approach. However, the goal of machine learning algorithms is usually to get the maximum overall accuracy while ignoring class imbalance. This results in the serious downfall in the execution of existent machine learning-centric NTC approaches when encountering unbalanced cases. In the research task, existing network traffic classification models will be optimized so that high accuracy, precision, as well as recall values will be achieved.

Future Research Directions and Open Challenges:-

Networks in today's era changed in a very fast way due to the updated technology. New Application, updates in protocols, user behaviour changes and new ways of cyber-attacks as a result of these characteristics changes often.

This process is known as Concept drift. That's why, one-time trained model of Static ML is not sufficient to handle or facing lack in accuracy [10] [11]. Another important challenge in NTC is class imbalance where certain traffic categoriesunderrepresented and having more data from rest of the categories. That's why ML model biased with majority classes only and skip minority classes as a result some sort of applications as well as didn't able to recognize the cyber-attacks.

We need to build a model thatgeneralize across different networks, devices and environments, therefore, we need high quality and diverse data. Obtaining labeled data is often very difficult, therefore we aim to train models using a small amount of labeled data or even without labeled data. The Majority of deep learning models are black boxes, but we need to understand how theymake their decisions. Network traffic arrives very quickly, so the model must also perform quickly. We need to build a model that performsefficiently with network traffic using minimal time and resources. Intrudersmay try to attack deep learning models to cheat, so the model should be secure and robust. It is still challenging to detect patterns in raw network data (packet or flow data) which is why we need to design a high-quality model. We can use anintegration of deep learning and machine learning towardsbuild a better modelthat can accurately classifysystem traffic.

Conclusion:-

This survey presents a review of ML techniques used for NTC. Protocol detection, Traffic Behavior analysis and application identification are included in it. In this paper, systematically comparison is done of recent research trends, evaluation metrices, benchmark datasets and main representative algorithms.A detailed preprocessing pipeline was implemented for fair evaluation. The inclusion of statistical validation confirms the reliability of findings. This work contributes by offering a reproducible, statistically sound comparison of widely used ML techniques, guiding future research in intelligent by classifying traffic over the network. ML plays a vital role in understanding network traffic. These techniques are helpful in recognizingeven encrypted and anonymous data. As a result, it becomes easier to classify different applications and services effectively even when the data is highly digitalized. These ML techniques are very helpful in detecting patterns and relationshipwithin traffic data, even when the input is in the form of sequence or pictures. That’s why these ML models are a very good and reliable way to classify network traffic.

References:-

[1] F. Alam, R. Kashef and M. Jaseemuddin, "Enhancing the Performance of Network Traffic Classification Methods Using Efficient Feature Selection Models," 2021 IEEE International Systems Conference (SysCon), Vancouver, BC, Canada, 2021, pp. 1-6, doi: 10.1109/SysCon48628.2021.9447076.
[2] Mamta, M. Poriye and S. Upadhyaya, "Hybrid Classification Model for Network Traffic Density Estimation," 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2023, pp. 1051-1056, doi: 10.1109/ICCCIS60361.2023.10425147.
[3] J. M. Adeke, J. Chen, L. Zhang, R. N. K. Mensah and K. Tong, "An Efficient Approach Based on Parameter Optimization for Network Traffic Classification Using Machine Learning," 2020 7th International Conference on Dependable Systems and Their Applications (DSA), Xi'an, China, 2020, pp. 99-107, doi: 10.1109/DSA51864.2020.00021.
[4] Y. Zhu and Z. Han, "Distributionally Robust Federated Learning for Network Traffic Classification with Noisy Labels," in IEEE Transactions on Mobile Computing, vol. 23, no. 5, pp. 6212-6226, May 2024, doi: 10.1109/TMC.2023.3319657
[5] H. Liu and B. Lang, "Network Traffic Classification Method Supporting Unknown Protocol Detection," 2021 IEEE 46th Conference on Local Computer Networks (LCN), Edmonton, AB, Canada, 2021, pp. 311-314, doi: 10.1109/LCN52139.2021.9525009.
[6] A. Ariffin, F. Zaki and N. B. Anuar, "Leveraging Federated Learning and XAI for Privacy-Aware and Lightweight Edge Training in Network Traffic Classification," 2023 IEEE International Conference on Computing (ICOCO), Langkawi, Malaysia, 2023, pp. 47-52, doi: 10.1109/ICOCO59262.2023.10397836.
[7] J. Kotak, I. Yankelev, I. Bibi, Y. Elovici and A. Shabtai, "VPN-Encrypted Network Traffic Classification Using a Time-Series Approach," in IEEE Transactions on Network and Service Management, vol. 22, no. 2, pp. 2225-2242, April 2025, doi: 10.1109/TNSM.2025.3543903.
[8] H. Yzzogh and H. Benaboud, "Evaluating ORB and SIFT with Neural Network as Alternatives to CNN for Traffic Classification in SDN Environments," in IEEE Access, vol. 13, pp. 51484-51498, 2025, doi: 10.1109/ACCESS.2025.3553449.

[9] H. Lu, Y. Dong, Z. Wu, H. -L. Wei and G. Lu, "New Class Detection in Network Traffic Classification Using Confidence Information Embedded Cascade Structure," in IEEE Transactions on Network Science and Engineering, vol. 12, no. 3, pp. 1692-1706, May-June 2025, doi: 10.1109/TNSE.2025.3538564.
[10] I. Fosić, D. Žagar and K. Grgić, "Network traffic verification based on a public dataset for IDS systems and machine learning classification algorithms," 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2022, pp. 1037-1041, doi: 10.23919/MIPRO55190.2022.9803674.
[11] M. Usama, A. Qayyum, J. Qadir and A. Al-Fuqaha, "Black-box Adversarial Machine Learning Attack on Network Traffic Classification," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 2019, pp. 84-89, doi: 10.1109/IWCMC.2019.8766505.
[12] C. Wang, T. Xu and X. Qin, "Network Traffic Classification with Improved Random Forest," 2015 11th International Conference on Computational Intelligence and Security (CIS), Shenzhen, China, 2015, pp. 78-81, doi: 10.1109/CIS.2015.27.
[13] M. M. K. Abid Hussain, M. Jaseemuddin and R. Kashef, "Network Traffic Classification Using Distributed ML-Based Data Parallelization Approach," 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Regina, SK, Canada, 2023, pp. 539-546, doi: 10.1109/CCECE58730.2023.10288743.
[14] X. Xiao et al., "RBLJAN: Robust Byte-Label Joint Attention Network for Network Traffic Classification," in IEEE Transactions on Dependable and Secure Computing, vol. 22, no. 3, pp. 2161-2178, May-June 2025, doi: 10.1109/TDSC.2024.3478838.
[15] Z. Wang et al., "DFE: Deep Flow Embedding for Robust Network Traffic Classification," in IEEE Transactions on Network Science and Engineering, vol. 12, no. 3, pp. 1597-1612, May-June 2025, doi: 10.1109/TNSE.2025.3535577.
[16] C. -Y. Shin, Y. -S. Choi and M. -S. Kim, "Data Augmentation-Based Enhancement for Efficient Network Traffic Classification," in IEEE Access, vol. 13, pp. 6006-6028, 2025, doi: 10.1109/ACCESS.2024.3525000
[17] W. Lin and Y. Chen, "Robust Network Traffic Classification Based on Information Bottleneck Neural Network," in IEEE Access, vol. 12, pp. 150169-150179, 2024, doi: 10.1109/ACCESS.2024.3477466.
[18] Y. Jang et al., "An Investigation of Learning Model Technologies for Network Traffic Classification Design in Cyber Security Exercises," in IEEE Access, vol. 11, pp. 138712-138731, 2023, doi: 10.1109/ACCESS.2023.3336674.
[19] Y. Wang, X. Yun, Y. Zhang, C. Zhao and X. Liu, "A Multi-Scale Feature Attention Approach to Network Traffic Classification and Its Model Explanation," in IEEE Transactions on Network and Service Management, vol. 19, no. 2, pp. 875-889, June 2022, doi: 10.1109/TNSM.2022.3149933.
[20] Z. Bu, B. Zhou, P. Cheng, K. Zhang and Z. -H. Ling, "Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models," in IEEE Access, vol. 8, pp. 132950-132959, 2020, doi: 10.1109/ACCESS.2020.3010637.
[21] S. Wei, X. Chen and B. Zhou, "Network Traffic Classification Based on AR-ELM Algorithm," 2018 13th World Congress on Intelligent Control and Automation (WCICA), Changsha, China, 2018, pp. 1500-1505, doi: 10.1109/WCICA.2018.8630436.
[22] O. Bader, A. Lichy, Amit Dvir, R. Dubin, and Chen Hajaj, "OSF-EIMTC: An open-source framework for standardized encrypted internet traffic classification," Computer Communications, vol. 213, pp. 271–284, Jan. 2024, doi: 10.1016/j.comcom.2023.10.011.
[23] Y. Geng, S. Cai, S. Qin, H. Chen and S. Yin, "An Efficient Network Traffic Classification Method based on Combined Feature Dimensionality Reduction," 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), Hainan, China, 2021, pp. 407-414, doi: 10.1109/QRS-C55045.2021.00067.
[24] X. Wang, S. Chen and J. Su, "App-Net: A Hybrid Neural Network for Encrypted Mobile Traffic Classification," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 2020, pp. 424-429, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162891.
[25] J. Koumar, K. Hynek and T. Čejka, "Network Traffic Classification Based on Single Flow Time Series Analysis," 2023 19th International Conference on Network and Service Management (CNSM), Niagara Falls, ON, Canada, 2023, pp. 1-7, doi: 10.23919/CNSM59352.2023.10327876.
[26] Y. P. Kumar S, S. Mishra and V. K. Chaithanya Manam, "A Comparative Study of Unsupervised Learning Techniques and Natural Language Processing in Network Traffic Classification," 2023 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Jaipur, India, 2023, pp. 138-143, doi: 10.1109/ANTS59832.2023.10469018.