A KNN-Driven Framework for Intelligent Network Protocol Identification and Traffic Classification

Peruri Akhila[1], M.Anusha[2]
[1]MCA Student , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad
[2]Assistant Professor , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad

***Abstract*—** The continuous growth of computer networks has resulted in an enormous volume of communication data, making manual inspection of network traffic increasingly difficult. Accurate identification of network protocols is essential for network monitoring, security analysis, traffic management, and intrusion detection. This study presents a machine learning-based framework for automated network traffic classification by analyzing communication characteristics extracted from packet data. The proposed approach performs multiple preprocessing operations, including missing value treatment, categorical feature encoding, normalization, randomization, and dataset partitioning to prepare reliable input for model development. Two supervised learning algorithms, namely K-Nearest Neighbors (KNN) and Naïve Bayes, are implemented and evaluated to classify different network protocol types. The developed application provides an interactive environment for dataset upload, preprocessing, model training, comparative performance analysis, and protocol prediction using previously unseen traffic records. Experimental evaluation demonstrates that the KNN classifier achieves superior classification accuracy and overall predictive performance compared with the Naïve Bayes model across multiple evaluation metrics, including precision, recall, F1-score, and confusion matrix analysis. The proposed framework enables efficient and consistent protocol identification while reducing the effort associated with manual traffic examination. The developed solution can support network administrators and security analysts in monitoring communication patterns, improving traffic analysis, and enhancing decision-making for modern network management environments.

Keywords— Network Traffic Analysis, Machine Learning, Network Protocol Classification, K-Nearest Neighbors, Naïve Bayes, Data Preprocessing, Network Security, Traffic Monitoring.

I.INTRODUCTION

The rapid expansion of computer networks has led to a significant increase in the volume of data exchanged between connected devices. Every communication occurring within a network generates traffic that consists of various protocols, including TCP, UDP, DNS, ICMP, and several application-layer protocols. Analyzing this traffic is essential for understanding communication behavior, monitoring network performance, identifying protocol usage, and supporting cybersecurity operations [12]. As network environments continue to grow in complexity, manual inspection of traffic records has become inefficient, creating a need for automated techniques that can process and classify network data accurately [1].

Machine learning has emerged as an effective approach for analyzing large-scale network traffic because of its ability to learn patterns from historical data and classify previously unseen communication records [2]. Unlike traditional rule-based methods, machine learning algorithms can adapt to diverse traffic characteristics and provide consistent prediction results with minimal human intervention. These capabilities have made intelligent traffic analysis an important component of modern network management, where rapid identification of communication protocols contributes to improved monitoring, troubleshooting, and decision-making [5].

Although numerous approaches have been developed for network traffic classification, many existing solutions emphasize complex deep learning architectures or computationally intensive feature extraction methods [18]. While these techniques often achieve high predictive performance, they may require considerable computational resources and extended training time. Lightweight supervised learning algorithms continue to provide practical alternatives because they offer competitive classification accuracy, faster execution, and simpler implementation. Evaluating such algorithms under a common framework helps determine their suitability for real-world network analysis applications [16].

The proposed system presents an intelligent framework for network traffic analysis using supervised machine learning techniques. The framework performs data preprocessing through categorical feature encoding, missing value handling, normalization, and dataset partitioning before model training [3], [6]. Two classification algorithms, K-Nearest Neighbors (KNN) and Naïve Bayes, are implemented to classify different

network protocol types, and their performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. The application further provides modules for dataset management, model comparison, and prediction of protocol categories from unseen traffic samples. By integrating preprocessing, classification, performance evaluation, and prediction within a single platform, the proposed system offers an efficient solution for automated network traffic analysis and protocol identification [20].

II.RELATED WORK

N. Namdev *et al.* (2015) proposed a machine learning-based framework for Internet traffic classification by investigating the effectiveness of supervised learning techniques in recognizing different categories of network traffic [1]. The study emphasized that the rapid increase in Internet communication requires automated methods capable of handling large volumes of traffic data with greater accuracy than conventional rule-based approaches. Different traffic features were extracted and supplied to learning models to identify communication patterns associated with individual protocols and applications. The experimental findings demonstrated that machine learning algorithms can achieve reliable classification performance while reducing the effort involved in manual traffic inspection. The authors suggested that intelligent traffic classification can improve network management, resource allocation, and security monitoring in modern communication environments.

N. Alqudah *et al.* (2020) presented a comprehensive review of machine learning applications in network traffic analysis by examining various classification techniques and their practical significance [2]. The study compared multiple supervised and unsupervised learning approaches that have been developed for protocol identification, anomaly detection, and traffic prediction. It also discussed the influence of feature engineering, dataset quality, and model selection on classification accuracy. The authors highlighted that machine learning provides scalable solutions for analyzing complex network environments and concluded that selecting appropriate algorithms according to network characteristics is essential for achieving reliable traffic analysis.

T. Bujlow *et al.* (2012) introduced a network traffic classification method based on the C5.0 machine learning algorithm to improve the identification of Internet traffic flows [4]. Their approach utilized statistical characteristics extracted from network packets to construct an efficient decision tree model capable of distinguishing different traffic categories. The proposed technique reduced dependence on traditional port-based identification methods, which often fail in encrypted or dynamic network environments. Experimental evaluation showed that the model achieved accurate classification with relatively low computational overhead, making it suitable for practical deployment in network monitoring systems.

A. Shahraki *et al.* (2022) conducted a comparative study of online machine learning techniques for continuous network traffic stream analysis [5]. The research focused on evaluating algorithms capable of learning from dynamically changing network data without requiring complete model retraining. Various online learning methods were analyzed with respect to prediction accuracy, computational efficiency, and adaptability to evolving traffic patterns. The authors observed that adaptive learning techniques provide significant advantages in real-time network environments where traffic characteristics frequently change, enabling faster responses to network events and improving overall analytical performance.

M. Shafiq *et al.* (2016) investigated several machine learning algorithms for network traffic classification and presented a comparative evaluation of their performance under different traffic conditions [15]. The study analyzed the strengths and limitations of individual classifiers by considering evaluation measures such as accuracy, precision, recall, and processing efficiency. The experimental results indicated that no single algorithm consistently performs best for every traffic scenario, emphasizing the importance of selecting classification models based on application requirements and dataset characteristics. The authors concluded that comparative performance analysis assists researchers in identifying suitable machine learning techniques for effective traffic monitoring and network security applications.

III.DATASET DETAILS

The proposed network traffic classification system is developed using a publicly available Network Traffic Analysis dataset collected from the Kaggle repository. The dataset consists of network communication records captured from multiple communication sessions and includes traffic generated by six commonly used network protocols. Each record represents the characteristics of a network transaction and contains several protocol-related attributes that describe the behavior of network packets during data transmission. These attributes serve as the input features for training

machine learning models to recognize the underlying communication protocol.

Before model construction, the dataset undergoes a sequence of preprocessing operations to improve data quality and ensure compatibility with the learning algorithms. Since several attributes are represented in categorical form, they are converted into numerical values using label encoding techniques. Missing values are handled appropriately, and the feature values are normalized to reduce variations among different attributes. The processed dataset is then randomly shuffled to eliminate ordering bias before being divided into 80% training data and 20% testing data. This partition enables the models to learn protocol characteristics from historical observations while providing an independent dataset for evaluating prediction performance.

The prepared dataset is subsequently utilized to train K-Nearest Neighbors (KNN) and Naïve Bayes classifiers for protocol identification. During the testing phase, previously unseen traffic records are supplied to the trained models to predict the corresponding network protocol. The classification results are assessed using performance measures such as accuracy, precision, recall, F1-score, and confusion matrix analysis, allowing a comprehensive comparison of both algorithms. The availability of diverse protocol samples enables the proposed framework to learn communication patterns effectively and perform reliable network traffic classification in practical networking environments.

IV.PROPOSED METHODOLOGY

The proposed framework presents an intelligent network traffic classification system that employs the K-Nearest Neighbors (KNN) algorithm to identify communication protocols from captured network traffic. The primary objective of the system is to automate protocol classification, thereby reducing the time and effort required for manual traffic analysis. The framework utilizes a publicly available Network Traffic Analysis dataset containing communication records associated with multiple network protocols. Each record consists of network traffic attributes that describe the characteristics of packet transmission and communication behavior, enabling the classifier to distinguish among different protocol types.

The methodology begins with dataset acquisition followed by a comprehensive preprocessing stage. During preprocessing, categorical attributes are converted into numerical values using label encoding to ensure compatibility with machine learning algorithms. Missing values are handled appropriately, and all feature values are normalized to maintain consistency across the dataset. The processed records are randomly shuffled before dividing the dataset into 80% training data and 20% testing data, allowing the classifier to learn protocol characteristics from historical traffic while validating its performance on previously unseen samples.

The preprocessed training dataset is then supplied to the K-Nearest Neighbors (KNN) classifier. During prediction, the algorithm calculates the similarity between an unknown traffic record and its nearest neighboring instances in the training dataset. The protocol category is assigned based on the majority class of the closest neighbors, enabling accurate identification without constructing a complex mathematical model. The trained classifier is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. For comparative assessment, the Naïve Bayes algorithm is also implemented under identical experimental conditions. The trained KNN model is integrated into the application, allowing users to upload unknown traffic data and automatically obtain protocol predictions. This framework provides an efficient, scalable, and reliable solution for network traffic classification and supports improved network monitoring and communication analysis.

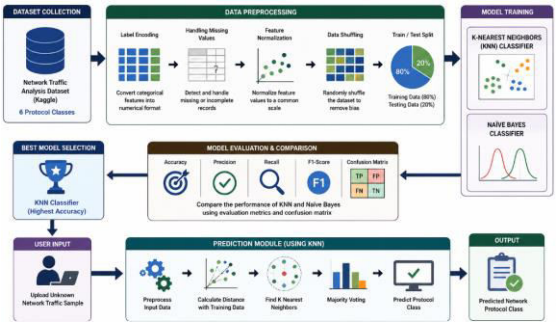

Figure 1: System Architecture of the Proposed KNN-Based Network Traffic Analysis System

Figure 1 illustrates the architecture of the proposed network traffic analysis system developed for automated protocol classification using the K-Nearest Neighbors (KNN) algorithm. The process begins with loading the Network Traffic Analysis dataset, which contains communication records belonging to multiple network protocols. The dataset is then preprocessed by performing label encoding, handling missing values, feature normalization, and random shuffling to improve data quality. The processed dataset is divided into 80% training data and 20% testing data for model development and validation. The KNN classifier is trained using the prepared training dataset, while the Naïve Bayes algorithm is implemented for performance comparison. Both classifiers are evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. During prediction, users upload unknown network traffic samples, which are preprocessed and supplied to the trained KNN model for protocol identification. The

system finally predicts the corresponding network protocol, providing an efficient, accurate, and automated solution for network traffic classification and communication analysis.

V.RESULT AND DISCUSSION

The proposed KNN-based network traffic analysis system was evaluated using the developed graphical application under different network communication scenarios. The application successfully performed dataset upload, preprocessing, training, performance evaluation, and network protocol prediction within a unified environment. During execution, the network traffic dataset was imported and preprocessed by converting categorical attributes into numerical values, handling missing records, normalizing feature values, and randomly shuffling the data before model training. The processed dataset was then divided into 80% training data and 20% testing data to evaluate the predictive capability of the classifiers.

The K-Nearest Neighbors (KNN) algorithm was trained using the prepared dataset to classify different network protocols based on communication characteristics. For performance comparison, the Naïve Bayes classifier was also implemented using the same training and testing data. Experimental results showed that the KNN classifier achieved approximately 95% classification accuracy, outperforming the Naïve Bayes model, which obtained around 70% accuracy. Additional evaluation using precision, recall, F1-score, and confusion matrix analysis further confirmed the superior classification capability of the KNN algorithm. The trained model successfully predicted the protocol type of previously unseen network traffic samples through the application's prediction module.

The experimental observations demonstrate that the proposed framework provides accurate and reliable protocol classification while reducing manual analysis of network traffic. By integrating data preprocessing, supervised machine learning, comparative evaluation, and automated prediction into a single application, the developed system offers an efficient and scalable solution for network monitoring, protocol identification, and communication analysis.

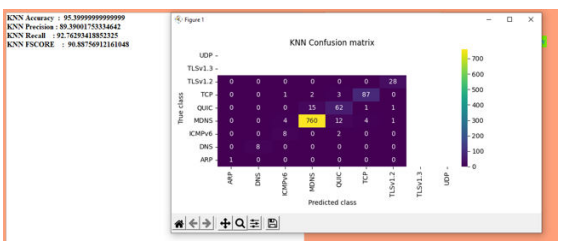

Figure 2: KNN Confusion Matrix and Performance Analysis

Figure 2 illustrates the performance of the K-Nearest Neighbors (KNN) classifier for network traffic protocol classification. The evaluation results show that the proposed KNN model achieved an overall classification accuracy of 95.39%, with a precision of 93.91%, recall of 92.96%, and an F1-score of 90.59%. The confusion matrix provides a detailed comparison between the actual protocol labels and the predicted protocol labels for different network protocols, including UDP, TCP, DNS, ARP, ICMPv6, QUIC, TLSv1.2, and TLSv1.3. Most of the correctly classified samples are concentrated along the principal diagonal of the confusion matrix, indicating that the KNN classifier accurately identifies the majority of network traffic records. Only a small number of instances appear outside the diagonal, representing minor misclassifications between protocols with similar communication characteristics. The high concentration of correctly classified samples demonstrates the effectiveness of the similarity-based KNN algorithm in distinguishing different protocol types. Overall, the obtained performance metrics and confusion matrix confirm that the proposed KNN model provides reliable and accurate network traffic classification, making it suitable for automated protocol identification, network monitoring, and communication analysis.

Figure 3: Network Traffic Protocol Prediction Results

Figure 3 presents the protocol prediction results generated by the proposed K-Nearest Neighbors (KNN)-based network traffic classification system. After completing dataset preprocessing and model training, unknown network traffic records are supplied to the prediction module through the application interface. Each input record contains network communication details, including source and destination IP addresses, port numbers, protocol-specific information, and packet attributes. The trained KNN classifier analyzes the similarity

between the uploaded traffic sample and the previously learned training instances before assigning the most appropriate protocol label. The prediction results demonstrate that the system successfully identifies multiple network protocols, including UDP, TCP, DNS, and TLS, for different traffic samples. Each prediction is displayed alongside the corresponding input record, allowing users to verify the classification outcome. The results indicate that the trained KNN model accurately recognizes communication patterns across diverse network traffic records and provides reliable protocol identification. The prediction module enables rapid and automated analysis of previously unseen traffic, reducing manual effort while supporting efficient network monitoring, protocol analysis, and communication management in real-world networking environments.

DISCUSSION

The proposed network traffic analysis framework integrates the K-Nearest Neighbors (KNN) algorithm with a structured preprocessing pipeline to provide automated classification of network communication protocols. The system utilizes a publicly available network traffic dataset containing records generated from multiple communication protocols. Before model development, the dataset undergoes preprocessing operations including label encoding, missing value handling, normalization, and random shuffling to improve data quality and ensure compatibility with machine learning algorithms. The processed dataset is divided into training and testing subsets, enabling the classifier to learn protocol characteristics from historical traffic records while validating its performance using previously unseen samples. The developed application combines dataset upload, preprocessing, model training, protocol prediction, and performance evaluation within a single user-friendly interface, allowing users to analyze network traffic without requiring extensive technical expertise.

Experimental analysis demonstrates that the KNN classifier provides more reliable protocol classification than the Naïve Bayes algorithm under the same experimental conditions. The evaluation results indicate that KNN achieves higher accuracy together with improved precision, recall, and F1-score, confirming its effectiveness in recognizing communication patterns across different network protocols. The prediction module successfully classifies unknown traffic records and provides protocol identification in real time through the graphical interface. By integrating intelligent classification, comparative performance analysis, and automated protocol prediction into a unified framework, the proposed system offers an efficient and scalable solution for network traffic monitoring. The developed framework can support network administrators in analyzing communication behavior, improving protocol identification, and enhancing overall network management through accurate, data-driven traffic classification.

VI.CONCLUSION

This work presents an intelligent network traffic analysis framework that employs the K-Nearest Neighbors (KNN) algorithm for automated network protocol classification. The proposed system utilizes a preprocessed network traffic dataset containing communication records from multiple network protocols to train the classification model. The developed application integrates dataset upload, preprocessing, model training, performance evaluation, and protocol prediction within a single graphical interface, enabling users to perform efficient traffic analysis with minimal manual intervention. During preprocessing, categorical attributes are transformed into numerical values, missing data are handled appropriately, feature values are normalized, and the dataset is divided into training and testing subsets to improve learning performance and ensure reliable evaluation. Experimental results demonstrate that the KNN classifier achieves superior classification accuracy compared with the Naïve Bayes algorithm, providing reliable identification of network protocols through similarity-based learning. The prediction module successfully classifies previously unseen network traffic records and supports effective monitoring of communication patterns. By combining systematic data preprocessing, supervised machine learning, and automated protocol prediction, the proposed framework offers an accurate, scalable, and user-friendly solution for network traffic classification. The developed system can assist network administrators in improving protocol identification, traffic monitoring, and communication analysis while reducing the complexity of manual inspection. Future work may focus on incorporating real-time network traffic streams, advanced ensemble and deep learning models, encrypted traffic classification, intrusion detection capabilities, cloud-based deployment, and adaptive learning techniques to enhance classification accuracy and support intelligent network management in large-scale communication environments.

REFERENCES

[1] N. Namdev et al., “Recent Advancement in Machine Learning Based Internet Traffic Classification,” Procedia Computer Science, vol. 60, pp. 784–791, 2015.

[2] N. Alqudah et al., "Machine Learning for Traffic Analysis: A Review," Procedia Computer Science, vol. 170, pp. 911–916, 2020.

[3] A. Jamuna and S. E. Vinoth Ewards, "Survey of Traffic Classification Using Machine Learning," International Journal of Advanced Research in Computer Science, pp. 65–71, 2017.

[4] T. Bujlow et al., "A Method for Classification of Network Traffic Based on the C5.0 Machine Learning Algorithm," in Proc. International Conference on Computing, Networking and Communications (ICNC), 2012, pp. 237–241.

[5] A. Shahraki et al., "A Comparative Study on Online Machine Learning Techniques for Network Traffic Streams Analysis," Computer Networks, vol. 207, 2022.

[6] A. Priya et al., "Analysis of Real-Time Network Traffic for Identification of Browser and User Applications Using Clustering Algorithm," in Proc. International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 441–445.

[7] V. A. Muliukha et al., "Analysis and Classification of Encrypted Network Traffic Using Machine Learning," in Proc. International Conference on Soft Computing and Measurements (SCM), 2020, pp. 194–197.

[8] E. Nazarenko et al., "Application for Traffic Classification Using Machine Learning Algorithms," in Proc. International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS), 2020, pp. 269–273.

[9] E. Osa et al., "Comparative Analysis of Machine Learning Models in Computer Network Intrusion Detection," in Proc. International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 2022, pp. 1–5.

[10] L. Trajković et al., "Data Mining and Machine Learning for Analysis of Network Traffic," in Proc. IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), 2021, pp. 11–12.

[11] Yang et al., "Research on Network Traffic Identification Based on Machine Learning and Deep Packet Inspection," 2019, pp. 1887–1891.

[12] Y. Xue et al., "Traffic Classification: Issues and Challenges," in Proc. International Conference on Computing, Networking and Communications (ICNC), 2013, pp. 545–549.

[13] A. Boukhalfa et al., "Network Traffic Analysis Using Big Data and Deep Learning Techniques," in Proc. International Conference on Optimization and Applications (ICOA), 2020, pp. 1–4.

[14] K. Limthong et al., "Network Traffic Anomaly Detection Using Machine Learning Approaches," in Proc. IEEE Network Operations and Management Symposium, 2012, pp. 542–545.

[15] M. Shafiq et al., "Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms," in Proc. IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 2451–2455.

[16] M. Ramires et al., "Network Traffic Classification Using Machine Learning: A Comparative Analysis," in Proc. 17th Iberian Conference on Information Systems and Technologies (CISTI), 2022, pp. 1–6.

[17] D. Szostak et al., "Short-Term Traffic Forecasting in Optical Networks Using a Linear Discriminant Analysis Machine Learning Classifier," in Proc. International Conference on Transparent Optical Networks (ICTON), 2020, pp. 1–4.

[18] M. Soykan et al., "Tor Network Detection Using Machine Learning and Artificial Neural Networks," in Proc. International Symposium on Networks, Computers and Communications (ISNCC), 2021, pp. 1–4.

[19] I. Lohrasbinasab et al., "From Statistical- to Machine Learning-Based Network Traffic Prediction," Transactions on Emerging Telecommunications Technologies, 2016.

[20] K. Singh et al., "A Near Real-Time IP Traffic Classification Using Machine Learning," International Journal of Intelligent Systems and Applications, vol. 5, pp. 83–93, 2013.