WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

Feature Selection and Dynamic Network Traffic Congestion Classification based on Machine Learning for Internet of Things

Ahmed A. Elngar[1,*] and Adriana Burlea-Schiopoiu[2]

[1]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, Egypt
[2]Professor of Management at the University of Craiova, Romania

*Corresponding Author: Ahmed A. Elngar

DOI: https://doi.org/10.31185/wjcm.150
Received: April 2023; Accepted: June 2023; Available online: June 2023

ABSTRACT: The network traffic congestion classifier is essential for network monitoring systems. An approach to classifying traffic based on several attributes is known as network traffic characterization. An approach to characterization network traffic is presented in this paper that uses payload-based classification. It has a wide range of applications, including network security assessment, intrusion detection, QoS provider, etc.; moreover, it is useful in investigating suspicious network activity. Classifying traffic requires both supervised and unsupervised methods, including Support Vector Machines and K-Means clustering. In current network conditions, minimal supervised data and unfamiliar applications influence the usual classification procedure's performance. The Internet of Things (IoT) network traffic is classified using clustering, feature extraction, and variety in this paper. Further, K-Means is used for network traffic clustering datasets, and feature extraction is performed on grouped information. KNN, Naïve Bayes, and Decision Tree classification methods classify network traffic because of extracted features, These classification algorithms are compared based on their performance. The results discuss the best machine learning algorithm for network congestion classification. According to the outcome, clustering (k-means) with network classification (Decision Tree) generates a higher accuracy, 86.45 %, than other clustering and network classification.

Keywords: Machine learning, Privacy-preserving classification, data set composition, Network Traffic classification.

1. INTRODUCTION

Current network security and administration frameworks use network traffic classification to distinguish network applications and characterize the corresponding traffic. Or on the other hand, Traffic grouping is a programmed technique that classifies network traffic as per different requirements into various traffic. Application-related traffic classification is essential innovation for ongoing network security. The traffic order can be utilized to discover worm engendering, intrusion discovery, examples of DOS assaults, spam spread, suspicious activity around the network, QoS suppliers, et cetera. Payload-based deep classification is incorporated into network traffic classification strategies. This technique is also called Deep Packet Inspection (DPI) and allows for accurate classification results in network traffic classification [1]. Now focusing on the network traffic, such as mobile networks, Internet of Things (IoT) networks, and smart cities [2] data on traffic, this classification can be applied to various data types. Data mining, datasets, and traffic classification attributes are critical classification methods for Machine Learning (ML). Extraction, selection, and machine learning approaches must be considered to classify network traffic.

The Internet is getting to be focal in our life and work. From talking on Facebook to finding the remedy for disease, the Internet identifies almost every part of our life. Long gone are the days when the first ARPANET was conceived. From that point forward, the system has been in steady development, changing the underlying ARPANET in what we today know as

the Internet, a colossal aggregate of interconnected P.C. systems. Despite its indispensable role in our lives, most people lack a complete understanding of its role. Over the past several decades, new systems, conventions, and applications have consistently appeared, creating a constantly evolving material that is hard to comprehend and grasp. This has compelled the examination network to more readily break down the system traffic and realize some light the perplexing activity of the Internet. Especially another field of concentrate, more often than not alluded to as traffic classification, has become significant for understanding the Internet. The classification of system traffic satisfies our interest and has numerous critical applications for organized administrators and I.T. managers. Bi Organization administrators and I.T. managers need to be able to categorize system traffic, which satisfies our interest.Torrent, Skype (i.e., P2P), Youtube, Netflix (i.e., gushing) or Megaupload (i.e., coordinate download) are a few cases of system applications that sooner or later changed the ideal models of the Internet. The classification of system traffic helps in numerous different behaviour. For example, examining how new applications affect the system can all the more likely arrange new frameworks, models or conventions. An exact classification can likewise enable Internet To specialist co-ops (ISPs) to apply dependable systems to apply Quality of Service (QoS) approaches in light of the necessities of the applications (e.g., VoIP calls). At last, this opens another scope of charging conceivable outcomes for ISPs to take profit of their frameworks because of their genuine utilizations. System administrators' longing is to be able to precisely group all the traffic of their systems on the web [3].

The application consolidates with a port number. The regular system movement grouping strategy depends on the port number and utilizes a packet header. In this technique, the arrangement depends on the surely understood port number to a given activity composed. For instance, web movement is related to TCP port 80. The port number of an application doesn't default, which prompts disappointment with the port-based strategy.

Internet-connected sensory devices provide observations and measurements of the physical world thanks to software, hardware, and communication technology advances. A huge amount of traffic data is generated by modern communication networks and systems, such as the Internet of Things (IoT) and cellular networks. Devices connected to the Internet and the digital world are part of the Internet of Things (IoT). Managing big data in real-time presents challenges and issues in traditional network management techniques. Network traffic is complex in cellular networks due to some factors, such as device mobility and network heterogeneity. Research in networking uses deep learning models for applications such as traffic classification and prediction in Network Traffic Monitoring and Analysis (NTMA).

Examining, for example, how new applications affect the system can make new frameworks, models, or conventions even more likely. Exact classification may also enable Internet specialist co-ops (ISPs) to apply reliable service quality (QoS) approaches in the light of the requirements of the applications (e.g., VoIP calls). At last, this opens another scope for conceivable outcomes for ISPs to take profit of their frameworks because of their simple operations. System administrators can precisely group all traffic on their systems over the web. However, the continuous advancement of internet applications with their procedures refrains from recognition. It is challenging to identify them. A wide range of unsolicited functional issues arises when these methods are linked in real situations to manage large amounts of traffic and constrained assets [4–8].

Classifying Internet traffic requires the use of feature selection techniques. Traffic classification has widely used machine learning algorithms (ML). To classify different types of Internet traffic, three major approaches are presented in this paper: An approach based on the port, an approach based on the payload, and an approach based on statistics. This paper summarises how feature selection can be used to avoid problems related to class imbalance, classification, and low classification rates in several machine learning algorithms.

Here, focus on the network's classification issues of network traffic by machine learning and ML training and testing validation problems. Firstly, the influence on selecting datasets from the same/different networks. Secondly, in the case of online classification, the actual majority and minority classes are considered in the training datasets. Thirdly, collecting data from the same or specific geographic network affects training and research. Consequently, ML dataset features and classification performance were examined by analyzing real Internet traffic through experiments. Accurate traffic classification is required for various network activities, including traffic engineering, fault detection, and Quality of Service (QoS) parameters. The importance of Network Traffic Classification is increasing as the number of internet users grows rapidly.

This paper mentions several contributions, including:

1. This paper discusses the clustering and network traffic classification methods.
2. This paper discusses the comparative analysis of clustering and various machine learning algorithms.
3. Weka tools are used for feature extraction. Machine learning-based classifiers with k-means clustering are used to Identification of Internet traffic.
4. The classification is based on Naïve Bayes, KNN, and decision Trees by incorporating the features exaction implemented on network data.

5. This paper's outcome analysis is based on the method's accuracy. The classification accuracy among Naïve Bayes, KNN, and Decision Tree is demonstrated. With 86.45% classification accuracy, the Decision Tree is more accurate than the KNN, which is only 71.33% accurate.

There are six sections in this paper. Section I introduction describes the introductory part about the method. Section II demonstrates a brief overview of the existing work with their author details. Sections III & IV presented the categorization of Network Traffic Classification Techniques and models. Section V & VI works on feature extraction and results from the analysis. At the last conclusion of the paper.

2. LITERATURE REVIEW

Using network traffic classification, traffic analysis is ordered to evaluate performance and security. Some ML methods for traffic analysis and A.I. procedures for detecting and analyzing malware behaviour are discussed [3]. Characterizing and classifying the traffic network requires high-speed transmission rates. DPI and port-based classification are the robust performance of operating systems such as the survival network, traffic engineering, QoS, dynamic access control, and standard traffic classification methods. The ML-based video streaming classification module provides systems requiring sufficient real-time network traffic processing. Using a novel approach, the procedure relaxes the presumption of independence between the algorithmic characteristics of Naive Bayes [4]. Obtaining observations and guidance about the study and practice of traffic classification by analyzing the advantages and disadvantages of each approach and evaluation.

Obtaining observations and guidance about the study and practice of traffic classification by analyzing the advantages and disadvantages of each approach and evaluation. Several networking problems have been solved using machine learning, including routing, traffic engineering, resource allocation, and security. It is becoming increasingly popular to use machine learning for improving IoT applications and providing IoT services, such as traffic engineering, network management, and traffic classification. As a preprocessing stage for machine learning, the selection function develops a subset of unique features that maximize learning precision with lower computational complexity. The performance of the algorithm has been improved with reduced classification accuracy. At an early stage, feature selection algorithms were introduced, such as deleting irrelevant [6] and redundant [7] features from Internet traffic classification. Computerized correlation-based filters (CFS) outperform the second filter process in accuracy and productivity [8].

A rapidly growing world population will increase urban mobility and result in traffic delays, fuel losses, noise and air pollution, and physiological and psychological problems. With traffic inference as a critical input, traffic control has maintained road serviceability. It has been examined by the scientific community how to gather sufficient data to infer traffic flows. Traffic control can be autonomously implemented using a modelling approach by harvesting and preprocessing vehicular data through interactive design [9, 10]. The CFS investigates the significance of every characteristic, i.e., strongly correlated with a particular class and limited correlation and the Greedy search [8]. The WEKA ML software [11] is used to estimate the most commonly used supervised ML methods [12–16], including Naive Bayes [15, 17], Naive Bayes Kernel Estimation [15], Bayesian Networks [8], C4.5 Decision Trees [8], k-Nearest Neighbours (KNN) [18], Neural Networks [8, 19, 20], Support Vector Machines [8, 21–23], and Sequential Minimal Optimization (SMO).

In previous works, clustering strategies are also commonly used [24]. The expectation-maximization algorithm [12, 25] is one of the earliest research projects in this field to split packet trails down into traffic clusters. Somewhere every cluster has specific traffic features. The unregulated Bayesian classification [26] is Auto Class to acquire the expected lessons characteristic in the training traffic dataset. The outcome presented according to precision and capability to discover before unfamiliar applications, the uncontrolled AutoClass [13] approach outperforms the controlled Naive Bayes Classifier. The clustering approach comprising K-means, AutoClass, and DBSCAN was compared [14]. [27], proposed a method for classifying TCP flows based on the size of the initial packets' payloads. The author used three clustering methods: the K-means [25], the Gaussian Mixture Model (GMM), and the spectral cluster.

The most commonly used techniques, such as port and payload-based, communication and numerical classification simulations, and traffic datasets with the graphic notation to improve research, have been addressed. Therefore, the systematic analysis of these classification models clarifies traffic classification, the QoS in numerous applications, and the problems in every classification system. Firewall access control, routing, policy-specific, and traffic QoS are just a few network services where traffic needs to be identified and classified [28, 29].

A machine learning algorithm's ability to learn is its most important feature. Experience and refinement are the keys to learning. The length of packets, the inter-arrival time of packets, etc., are statistical properties of each type of network traffic. This unique characteristic of the network pattern is learned through the training data set. ML algorithms are used to analyze the network data captured. Videos, file transfers, email, and browsing are among the diverse applications that are challenging to identify. A network management tool that can handle exponential growth in network traffic is necessary to maximize the use of network resources. Although reduced monitoring information and unfamiliar applications affect

classification efficiency, modern traffic classification methods aim to obtain statistical functions and ML methods. In addition, previous approaches do not detect anomalies in the flow level. Flow-level anomaly detection has been introduced by the author [30] under Unknown Flow Detection Approaches. A flow-level anomaly can be observed using the Synthetic Flow-level Traffic Trace Generation (SG-FLT) method. The two main challenges to this approach are normal and abnormal network performance and discovering a realistic system. A normal flow pattern can also be defined as an abnormal flow pattern. A real-world application-based network traffic dataset was used to demonstrate the unknown flow detection approach. As shown in Table 1, the proposed method is more efficient than current methods within the dynamic network context [30].

Table 1. ML-based traffic analysis summary

S. No.	Author & Year	Technique used	Descriptions
1	Mukkamala, S., et al., (2002) [31]	ML using (SVM)	To detect features that describe class behaviours to construct anomaly classification
2	Mahoney, M.V., (2003) [32]	ML Algorithm for Intrusion Detection	Identify the anomalies in network traffic.
3	McGregor, A. et al., (2004) [12]	ML with Flow Clustering	Clustering used bulk transfer, single and multiple transactions, and ML classifications.
4	Laskov, P., et al., (2005) [33]	ML-based on SVM and KNN classifier with k-means	A comparative study based on supervised and unsupervised ML for identifying malicious actions
5	Erman, J., et al., (2006) [13]	ML Using supervised ML approach	Identification of Internet traffic using ML
6	Liu, Y., et al., (2007) [25]	ML with K-means Clustering	Different levels in network traffic-analysis
7	Zamani, M., et al., (2009) [34]	Artificial immune approach	Intrusion detection (IDS) for distributed systems
8	Sommer, R. et al., (2010) [35]	ML technique	ML methods used for spam detection after that intrusion detection for detection of anomalies
9	Bujlow, T., et al., (2012) [36]	ML using the C5.0 method	Network traffic classification
10	Jamuna, A., et al., (2013) [37]	ML using Decision Tree and Naïve Bayes accuracy	Network traffic classification
11	Suthaharan, S. (2014) [38]	Supervised ML approach	IDS traffic classification by learning characteristics
12	Blowers, M. et al., (2014) [39]	ML with DBSCAN clustering	Anomaly discovers by clustering approach.
13	Zheng, N. et al., (2014) [40]	ANN with Voting Experts (VE) technique	Extract the main features of network traffic data from the VE method
14	Bartos, K. et al. (2016) [41]	Supervised ML approach	Detection based on unknown risks to security
15	Wang, P., et al., (2016) [42]	ML using SVM	Energy classification is used for data flows.
16	Furno, A., et al., (2017) [43]	ML with Exploratory factor analysis (EFA)	Analyzing spatial structures as well as connecting to mobile traffic using EFA technology
17	Lopez-Martin, Manuel, et al., (2017) [44]	Traffic Classifier with Convolutional (CNN) and Recurrent Neural Networks (RNN)	Classify the network traffic and feature selection using CNN and RNN for the Internet of Things (IOT)
18	Mirsky Y., et al., (2018) [45]	ANN with Kitsune	Malicious traffic detection at in and out network
19	Mohammed, Bushra, et al. (2020) [46]	Robust Feature Selection for Network Traffic Classification in the Internet of Things (IOT)	The ensemble Weight Approach (EWA) is used for feature selection on spatial and temporal Datasets.

Manufacturing processes are actively monitored and controlled with the help of CBI4.0, a novel cross-layer data collection approach. A multichannel and multi-radio sensor network is designed to provide quality of service (QoS) requirements by dynamically switching between various frequency bands in Multichannel Wireless Sensor Networks (MWSNs) to achieve higher data rates, low packet loss, throughput, delay, and corrupted packets [47]. A detailed comparison of the proposed work, including channel detection, channel assignment, and packet forwarding methods, in the intelligent grid Industry 4.0 described by the author [48] and WSN-based innovative grid applications [49].

3. CATEGORIZATION OF NETWORK TRAFFIC CLASSIFICATION TECHNIQUES

Categorizing network classes is the process of classifying network traffic. Over the last two decades, network traffic classification has become increasingly important. Many researchers have proposed the classification of network traffic. Traffic classification is divided into three techniques: post-based, Payload-based, and Statistics-based approaches, as shown in Figure 1. The complete details of these three traffic classification techniques are mentioned in the subsection.

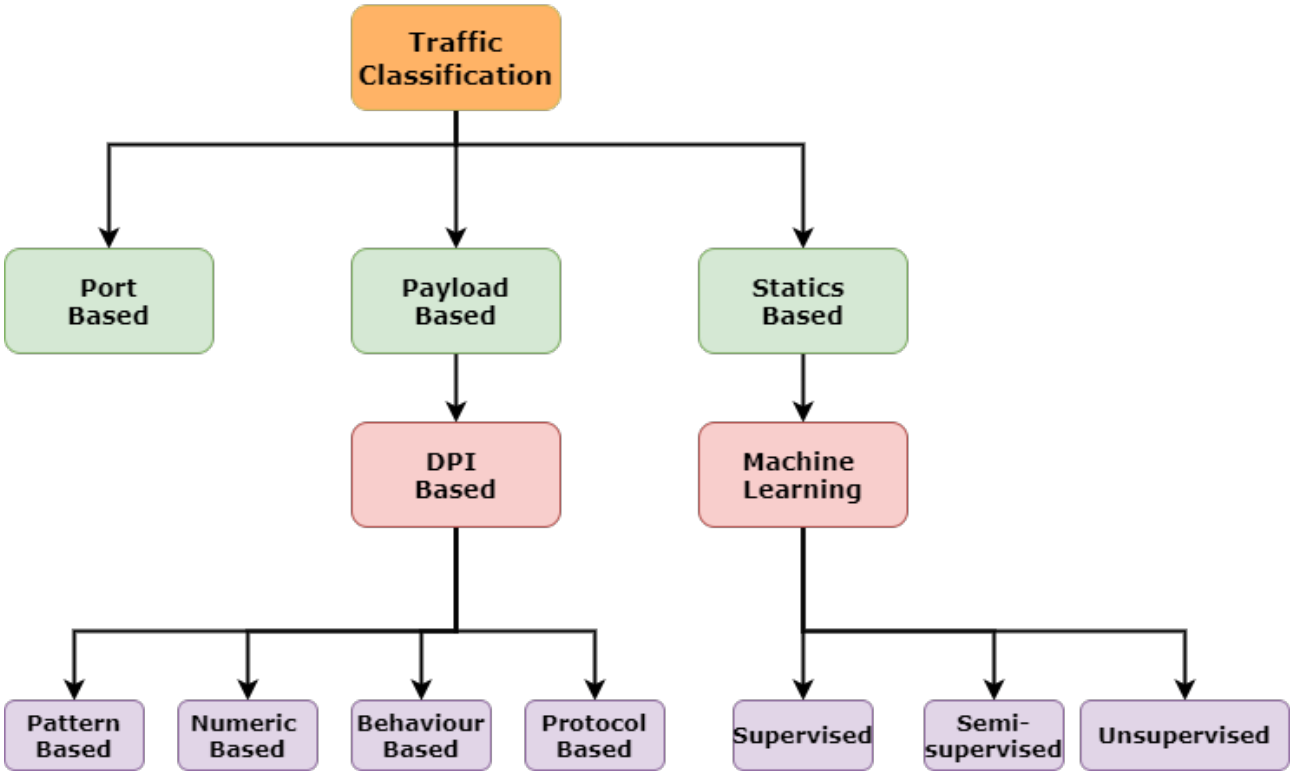

FIGURE 1. **The methods for classifying traffic**

3.1 PORT-BASED APPROACH

These techniques are arranged in light of the component that particular application organizations use of IANA (Internet Assigned Numbers Authority) consigned port numbers. This system encounters the going with insufficiencies. In any case, P2P applications use discretionary or dynamic port numbers. Secondly, general organization ports may be used by various organizations, for example, malware. Third, port numbers are different from the assigned ones. Fourth, coarsely cut. The transport layer or I.P. delivery can cover port numbers. The process coordinates an application with a port number, where the port number is associated with the application. The following table shows the ports associated with various types of applications.

This approach, therefore, does not have substantial results on classification accuracy [47, 48]. Due to the dynamic port number, this solution is not feasible.

Table 2. IANA allocated port-number organization for a few useful applications

Assigned Port	Application
20	FTP Data
21	FTP
22	SSH
23	TELNET
25	SMTP
53	DNS
80	HTTP
110	POP3
123	NTP
161	SNMP
3724	WoW

3.2 PAYLOAD-BASED APPROACH

Payload-based suggests the significant package survey DPI framework, which uses static application checks as a piece of payload to recognize traditions. Deep packet analysis further allows more use of the quantifiable properties of the packet payload. DPI is unfathomably hurt by encryption since the plaintext marks turn indistinct. In any case, it very well may be used as a piece of coarse gathering for specific mixed movement, for instance, SSL. SPI has a fine-grained course of action limit on a fundamental level since its components are considered specific for application-layer traditions. True components have drawn an impressive proportion of thought, and a wide range of measures have been employed to perceive different categories of action streams, including machine learning and neural networks.

In the I.T. industry, DPI sees and uses the consideration and usefulness of the DPI package level review to determine the authentic source, the complexity of the system, and the application's functionality quickly and precisely. DPI is mainly used to study product content and mastermind software applications. DPI was presented in four categories: (1) pattern-based, (2) numeric-based, (3) behavior-based, and (4) protocol-based. Table 2 presents the subsection of strings with TCP and UDP protocols. The most utilized ports for P2P protocols are also shown in Table 3—the more details with a list of strings presented in [47].

Table 3. Shows the strings for Protocol Payload

P2P Portal	String	Trans. Protocol
Edonkey 2000	Oxe 319010000	TCP/UDP
	Oxe 53f010000	
Fasttrack	"Get/.hash."	TCP
	0x27000000002980	UPD
BitTorrent	"0x13Bit"	TCP
Gnutella	"GNUT"," GIV"	TCP
Address	"GET hash"	UDP
	"Get Shal"	

3.3 STATISTIC-BASED CLASSIFICATION

Factual portrayal essentially suggests the techniques in perspective of quantifiable action properties, Where Machine Learning is the most preferred destination. ML technology [2, 3] is based on data collection. This approach trains The ML classifier as input and unknown classes are defined using the sample prediction learned.

An ML approach can be divided into supervised and unsupervised learning. In machine learning, supervised, semisupervised, and unsupervised learning are the main fields of study. Significant research on ML methods for traffic classification has been carried out over the last decade. ML is categorized as supervised Vs. Unsupervised methods.

4. NETWORK TRAFFIC CLASSIFICATION MODEL

We clarify the system movement arrangement structure, which incorporates a well-ordered process, as shown in Figure 3. This well-ordered process strategy will demonstrate best practices to utilize organized activity grouping procedures to character/characterize obscure system movement classes utilizing machine learning methods.

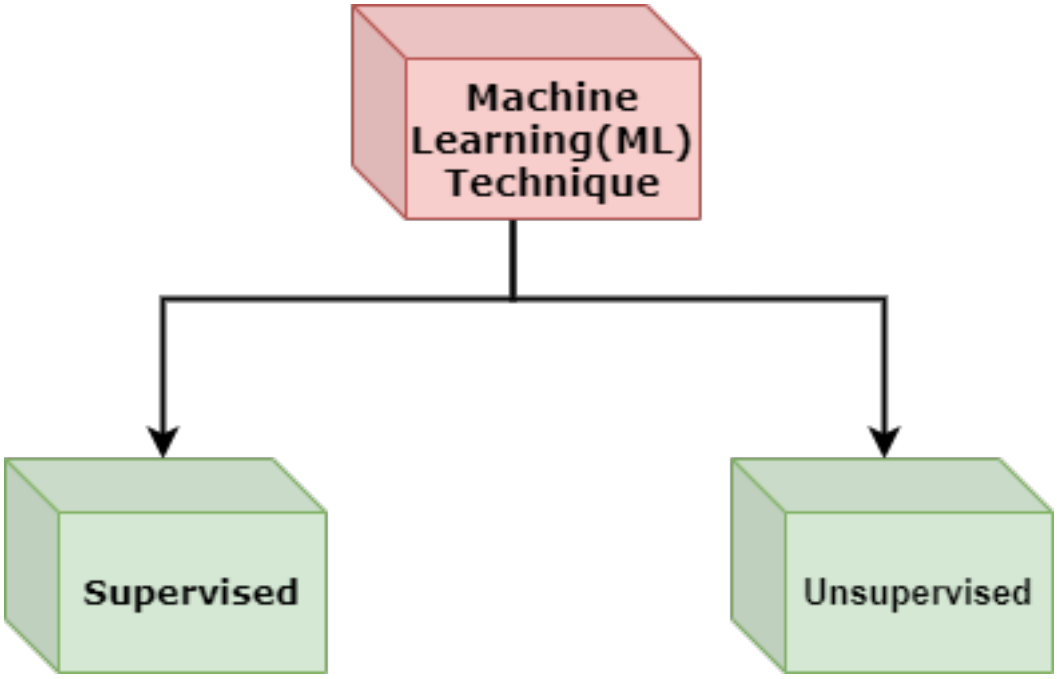

FIGURE 2. **Types of Machine Learning Classification**

• **Network Traffic Capturing**

It is only the most critical development, including the accumulation of knowledge. The ongoing machine movement is captured for this progression. Otherwise, it is called the step of gathering information. We use Wireshark [I.S.] to gather and analyze packets to organize travel. We capture the movement within a single moment of WWW, DNS, FTP, and P2P and use Telnet.

• **Feature Selection**

Organizing movement details is followed by component selection and extraction. The highlights of the data collected are isolated, for example, the package name, parcel length, and entomb package time convention. At that point, the extracted highlights are used to prepare the ML Classifier. The Perl material may delete the item from the captured information set in the extraction.

• **Training Process Sampling**

Information collection is being examined for a controlled learning system at this level. Information is first named in administered learning for characterizing obscure system applications. Training and testing data are utilized to build an ML model and measure the performance of an ML model. Training and testing data should be dissimilar, mutually independent, and created by random sampling.

• **Implementation of ML Algorithms**

It is the phase of use integrating either ML calculation or the description of occurrences into the specification. For example, direct, unsupervised, and semi-managed learning measurement, detailed data, time and review planning, etc. Predictive modelling generates a system with good presentation making predictions on new unseen data.

• **Performance and compassion**

After executing machine learning calculations, the reproduction device gives nitty-gritty outcomes about the connected calculations, such as definite precision data, preparing time and review, etc. This paper discusses, in the next section, three supervised learning algorithms.

5. PROPOSED METHOD

We introduced a new strategy to investigate network traffic activities using Java Framework. The proposed methodology investigates suspicious activities and malicious information flowing through the networks using various techniques such as clustering and classification. This system loads and processes traffic data by the proposed framework and examines the information and the network traffic differentiated as I.P. Wise, Port Wise, and Protocol Wise. The k-Means algorithm is utilized for clustering, followed by feature extraction. And for traffic classification, we utilized three classification

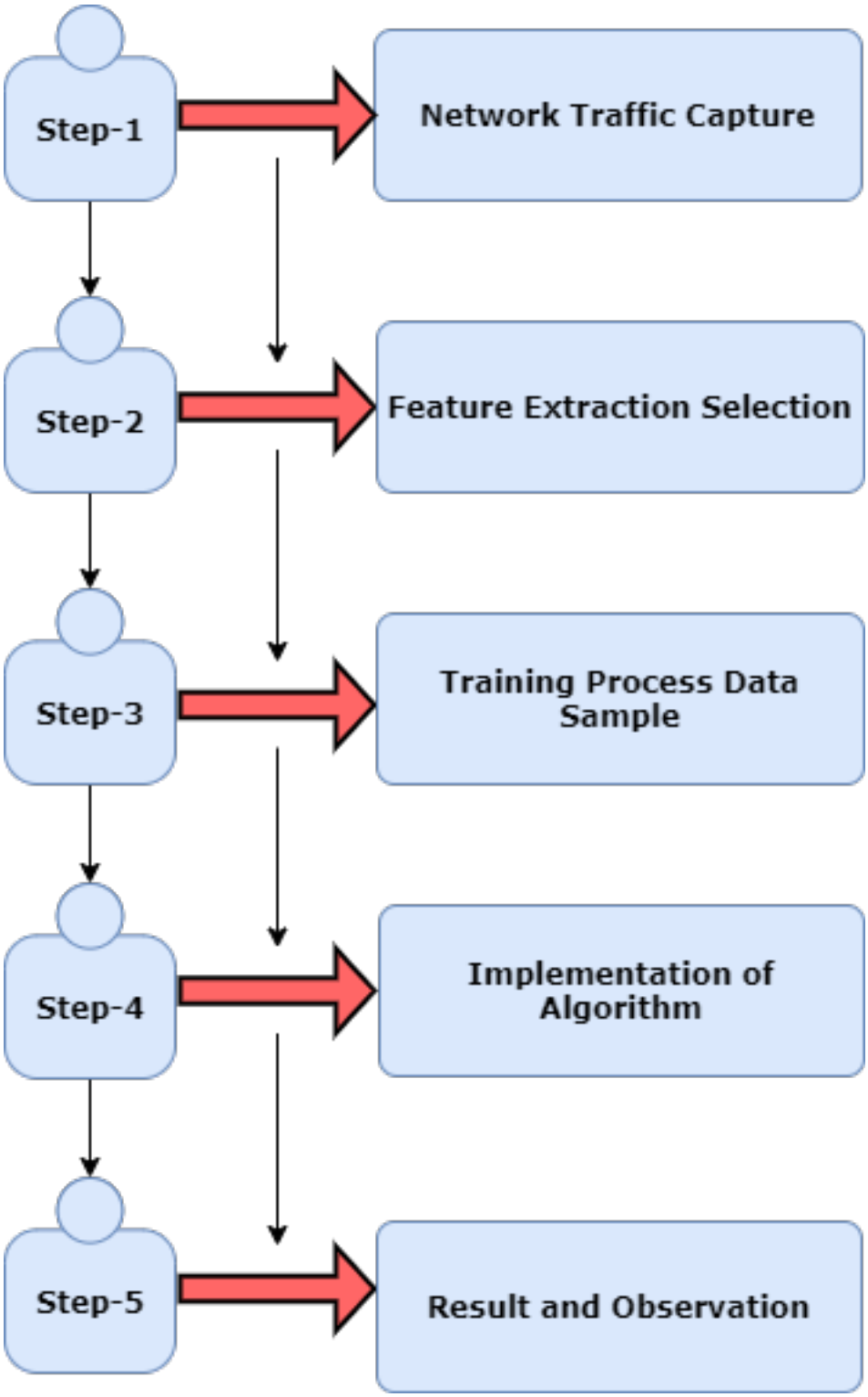

FIGURE 3. **The steps of Framework analyses for traffic classification**

algorithms KNN, Decision Tree, and Naïve Bayes. This paper also evaluates the classification performance between all three algorithms. The proposed framework is successfully implemented using the steps described below and presented in Figure 4.

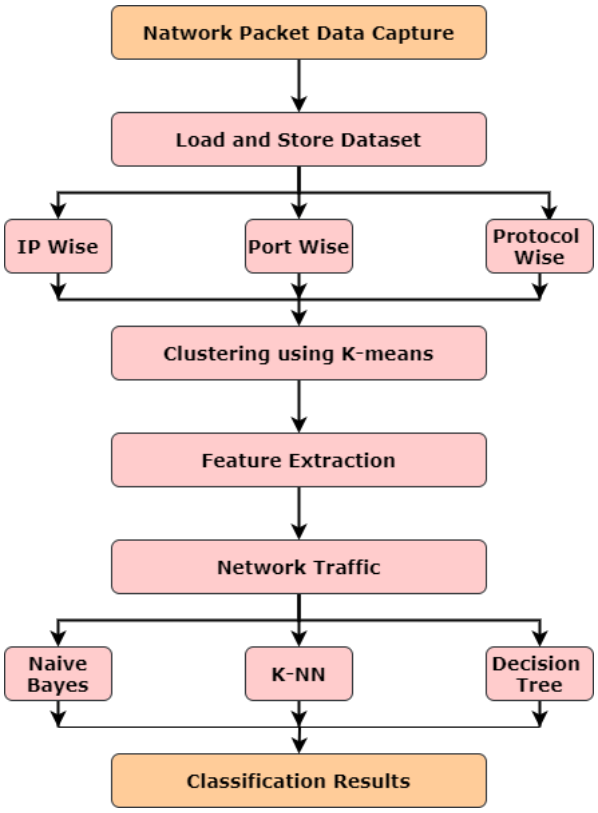

FIGURE 4. Flow of the Proposed Methodology

Step 1: Load Network Traffic Dataset

The proposed framework is loaded with the network traffic dataset in this phase. UCI Machine Learning provides the network traffic dataset. After loading the Dataset, arrange the Dataset into I.P.-wise, port-wise, and Protocol-wise. Similar work was designed with the help of SVM and Naïve Bayes [17, 23]. Data mining requires data preprocessing. Datasets with large amounts of network traffic were present. There are often errors in the network data, such as incomplete data, missing values, incorrect information, and many others. Many proven methods exist to resolve these errors, such as checking complete data and correcting all error fields.

Many data are incomplete, unreliable, and deficient in particular behaviours or patterns, and several errors are likely to occur. Preprocessing data is a proven way to solve these problems. The network traffic data set error field has been corrected several times during data preprocessing.

Step 2: Clustering

K-Means cluster data based on partitions in a fast, simple manner. Vectors are divided into disjoint subsets in a data set. A random selection of K cluster centres is made using the algorithm. Vectors in the data set are assigned to clusters based on their distance from the data set's centers. The new centres are reassigned iteratively as K-Means form the clusters. Once the cluster membership has stabilized, the partitioning results are produced.

This phase performs clustering on network traffic datasets utilizing the K-Means algorithm [25, 30]. The earlier work that reviewed the algorithms in network traffic classification indicated that the K-means are robust and fast for clustering and help the classification phase recognize unfamiliar applications, improving classification accuracy. K-means is an unsupervised ML method mainly used when the Dataset is unlabeled. It assigns every data point to a class iteratively based on its characteristics. K groups in data, and each data point are clustered based on feature measurement and assigned to a group K. The value of K can be found by calculating the average distance among data points and their cluster centroid by Eqn. (1).

$$J = \sum_{i=1}^{k} \sum_{j=1}^{k_i} \left(\left\| x_i - y_j \right\| \right)^2 \tag{1}$$

Where, '$\|x_i - y_j\|$' represent Euclidean distance among, x_i *and* y_j, 'k_i'represent the No. Of traffic data points in i^{th} group. 'k' indicates No. Of group centres. The K-means clustering approach's complete procedure is mentioned in Algorithm I.

Algorithm I: K-means clustering Algorithm

==========

Input: Network Traffic Data Instances

Output: Cluster or groups of Dataset

Begin

Initialization

No. of the network traffic data point $X = \{X_1, X_2, X_3, \ldots\ldots.., X_n\}$

Centers Points $Y = \{Y_1, Y_2, Y_3, \ldots\ldots.., Y_n\}$

Load Network Traffic Data

Randomly choose 'k' group centres.

For all the k value

Measure the distance among every network traffic data as well as group centres.

if the distance of traffic data < Among all the group centres

Allocate the traffic data to the group centre

else

Measure the new group centre by Eqn. (2) again:

$$y_i = \frac{1}{k_i}\sum_{j=1}^{k_i} x_i \tag{2}$$

Where k_i Indicates the No. Of traffic data in the i^{th} group.

Allocate measure distance among every traffic data and new achieved group centres by Eqn. (3).

$$d\ (x,\ y) = \sqrt{\sum_{i=1}^{k}\ (x_i - y_i)^{\,2}} \tag{3}$$

end if

end for

if (! no traffic assigned to group centres) do

Return and redistribute the network traffic data

end if

End

==========

Clustering assigns a label to each traffic data to share similar characteristics that reflect their traffic application. After this phase, traffic data is divided into various clusters, where each traffic data in a cluster shares similar characteristics and thus strongly interrelates to a function of interest.

Step 3: Feature Extraction

This phase extracts the features from datasets. The features which are extracted are described in Table 4.

Step 4: Training and Classification

a) Naïve Bayes

Naïve Bayes is a practical and straightforward probabilistic classifier classifying network traffic data. The performance of any classification algorithm relies on how effectively features are extracted and trained. Naïve Bayes utilized the Bayes theorem for training the extracted features framework. The Naïve Bayes algorithm is highly scalable and requires several features in the training phase. Naïve bays classification can classify traffic data using even a few training data [28].

$$p(C = c \mid X = x) = \frac{p(C = c)p(X = x \mid C = c)}{p(X = x)} \tag{4}$$

This expression calculates the probability of each class having random features X, and c is a cluster of determined classes. The implementation and simulation procedures of the Naïve Bayes are mentioned step by step in algorithm II.

b) KNN

Among all the supervised machine learning, KNN is the most utilized. It gives a better outcome without the prior requirement of feature training data. It works by assuming all data points belong to n-dimensional space. This algorithm classifies the new data point by calculating the k-nearest neighbour, the most common class among them, by incorporating

Table 4. Feature Extraction

Features	Comment
Src_addr	source address
Src_port	source port
Dst_addr	destination address
Dst_port	destination port
Ip_prot	I.P. protocol
Tcp_ags	xor of TCP flags
N_ack	number of ACK flag
N_psh	number of PSH flag
Duration_sr	Duration
Pkt_max_lg	max of packet size
Pkt_mean_sr	mean of packet size
pktvar_sr	a variance in packet size
Packets_lg	number of packets
Octets_lg	number of bytes
Speed_sr	bytes per second
Pkt_rate_sr	packets per second

Euclidean distance by Eqn. (3). The KNN algorithm is an easy-to-use supervised ML approach that solves classification and regression problems. The complete procedure of classifying network traffic using the K-Nearest Neighbor approach is mentioned in Algorithm III.

c) Decision Tree

T Statistical classifiers are used in decision trees for classification. Based on all available features that differentiate target applications, it recursively selects the classes labelled based on each data point. The class decides based on the ratio between the features and the class.

$$RATIO\ (\ X\ )\ Y\ ) = \frac{H\ (\ X\ ) - H\ (\ X\ )\ Y\ )}{H\ (\ X\ )} \tag{5}$$

The marginal probability of $H\ (X|Y)$ is the probability of a probabilistic joint distribution

To calculate the value of the target variable, D.T.s use simple decision rules to build a learning model.

6. RESULT AND DISCUSSION

Implementing the proposed methodology generates the outcome of the applied algorithms, such as accuracy, measured by the classification method on the network traffic data. According to our proposed algorithm, the k-mean clustering algorithm extracts the first feature. After that, three classification algorithms were applied for accuracy. This work has three types of classifiers: Decision tree (D.T.), KNN, and Naïve Bayes.

This section discusses the results of execution clustering, feature extraction, and classification. **The accurate classification depends upon how effectively the clustering or features extraction and classification algorithm relies on reasonable accuracy**. Clustering and feature estimation are proposed to expand the classification accuracy. The implementation process is described phase by phase with GUI development.

Step 1: Design the home page for the proposed framework shown in Figure 5.

FIGURE 5. Home page

Step 2: Described the menu page where one by one classification steps as mentioned in Figure 6.

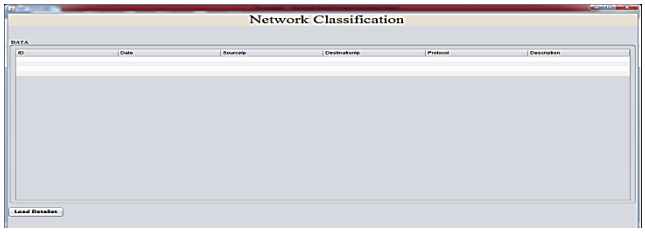

FIGURE 6. Available classification process options

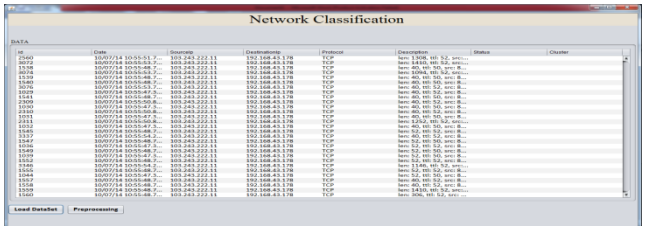

FIGURE 7. Dataset Load process

Step 3: Load the network traffic dataset captured using wire shark in the Weka tool, as mentioned in Figure 7.

Step 4: Filtering and cleaning data mining requires data preprocessing. Datasets with very high instances or extensive data were present in the network traffic dataset. The preprocessing step in Figure 8 involves removing unnecessary details from a loaded dataset using the Weka tool.

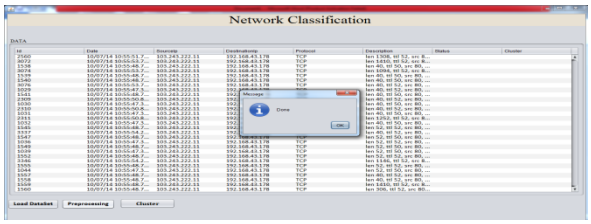

FIGURE 8. Dataset Preprocessing

Step 5: In Figure 9, clustering is the process of grouping objects or classes into groups so that objects in the same group (called a cluster) have a higher degree of similarity than those in other clusters. The method is used in various fields, such as data analysis and machine learning, and is essential to exploratory data mining. Traffic data is clustered during this phase, and centroid labels are assigned.

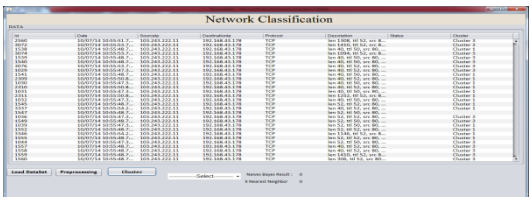

FIGURE 9. Presents clustering of captured data

Step 6: Now, network traffic data is prepared and tested based on training the test. The classification is done on the Weka tool by selecting the Naïve Byers. The attribute is selected as a class selector, and then press the Start button to build a model. This phase performs classification utilizing Naïve Bayes and presents the classification accuracy in Figure 10.

Step 7: Data mining techniques such as KNN can be used to predict the outcome value of a new data instance by utilizing past data instances with known values. KNN classification applied, and respective accuracy is shown in Figure 11.

Different machine learning algorithms are shown in Table 5 and Figure 12 show the classification accuracy obtained

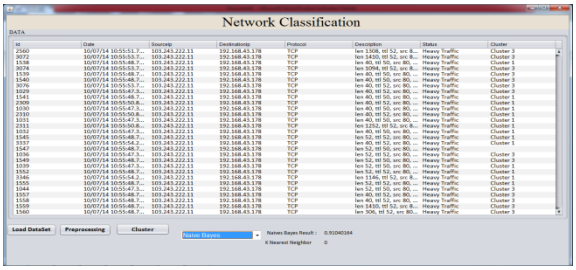

FIGURE 10. Output after implementing Naïve Bayes Classification

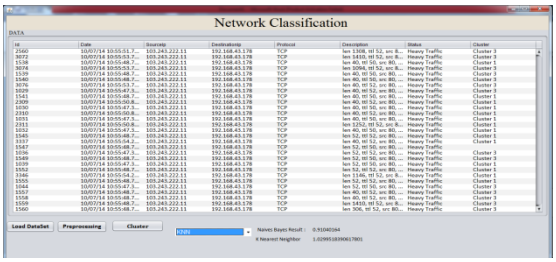

FIGURE 11. Sample output of KNN Classification

through different machine learning algorithms. The classification accuracy depicts that the Decision Tree has superior classification accuracy among the other implemented algorithms. The classification is done on a dataset obtained from WireShark. Figure 12, Shows the recall and precision performance comparison on three different ML approaches. The decision three prediction outcome was completed with an average cost of 0.3268 per test example and took 10.567 seconds to complete the process. The obtained confusion matrix is presented in Table 5.

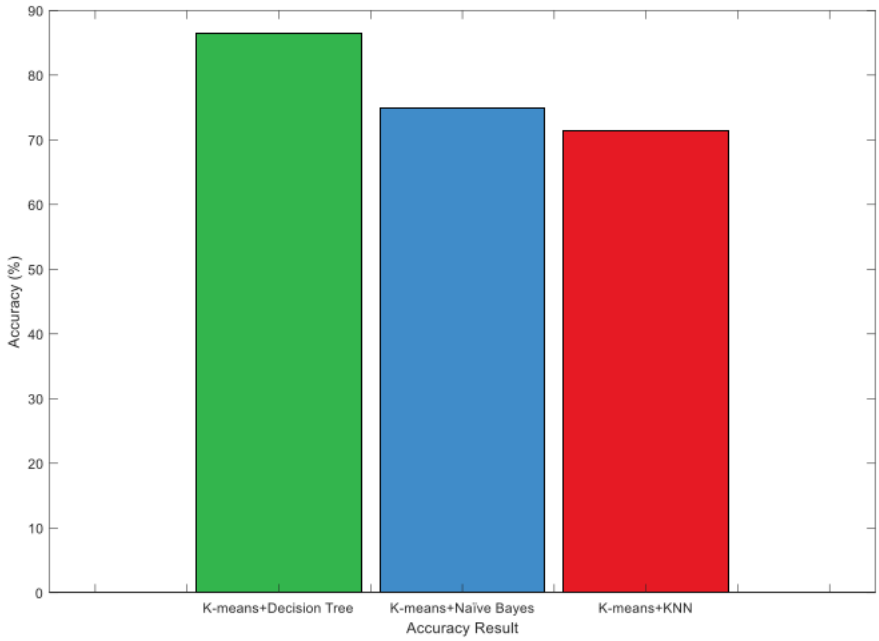

FIGURE 12. Performance Graph for classification accuracy

Table 5. Performance evaluation for different ML algorithms

Dataset	Clustering	Classifiers	Accuracy (%)
Retrieved from Wire-Shark	K-means	KNN	71.33
Retrieved from Wire-Shark	K-means	Naïve Bayes	74.86
Retrieved from Wire-Shark	K-means	Decision Tree	86.45

7. CONCLUSION

In this paper, Network traffic datasets are used for clustering and classification. K-mean algorithms are used for clustering, and Decision trees, KNN, and Naïve Bayes are used for the classifier. This paper performs clustering and feature extraction used in classification like Src_port, Dst_port, Src_addr, Dst_addr, and so on. And the effective feature extraction to help in enhancing the classification accuracy and network performance. The classification is based on Naïve Bayes, KNN, and decision Trees by incorporating the features exaction implemented on network data. The classification performance of all techniques is measured. The classification accuracy among Naïve Bayes, KNN, and decision Trees is demonstrated. 86.45% is the classification accuracy of the decision tree, while 71.33% is the classification accuracy of the KNN. The decision tree is the best algorithm for network traffic among Naïve Bayes and KNN.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] M. M. Raikar, S. M. Meena, M. M. Mulla, N. S. Shetti, and M. Karanandi, "Data Traffic Classification in Software Defined Networks (SDN) using supervised learning," *Procedia Computer Science*, vol. 171, pp. 2750–2759, 2020.

[2] M. Shafiq, Z. Tian, A. K. Bashir, A. Jolfaei, and X. Yu *Data Mining and Machine Learning Methods for Sustainable Smart Cities Traffic Classification: A Survey. Sustainable Cities and Society*, pp. 102177–102177, 2020.

[3] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary *A Deep Comprehensive Research Architecture, Characteristics, Challenges, Issues, and Benefits of Routing Protocol for Vehicular Ad-Hoc Networks. International Journal of Distributed Systems and Technologies (IJDST)*, vol. 13, pp. 1–23, 2022.

[4] K. L. Dias, M. A. Pongelupe, W. M. Caminhas, and L. D. Errico, "An innovative approach for real-time network traffic classification," *Computer Networks*, vol. 158, pp. 143–157, 2019.

[5] H. Kim, K. C. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified: myths, caveats, and the best practices in," *Proceedings of the 2008 ACM CoNEXT conference*, pp. 1–12, 2008.

[6] V. Pervouchine and G. Leedham, "Extraction and analysis of forensic document examiner features used for writer identification," *Pattern Recognition*, vol. 40, no. 3, pp. 1004–1013, 2007.

[7] A. Appice, M. Ceci, S. Rawles, and P. Flach, "Redundant feature elimination for multi-class problems," *Proceedings of the twenty-first international conference on Machine Learning*, pp. 5–5, 2004.

[8] N. Williams and S. Zander *Evaluating machine learning algorithms for automated network application identification*, 2006.

[9] N. Hussain and P. Rani, "Comparative Studies Based on Attack Resilient and Efficient Protocol with Intrusion Detection System Based on Deep Neural Network for Vehicular System Security," *Distributed Artificial Intelligence*, pp. 217–236, 2020.

[10] P. Rani, N. Hussain, R. A. H. Khan, Y. Sharma, and P. K. Shukla, "Vehicular Intelligence System: Time-Based Vehicle Next Location Prediction in Software-Defined Internet of Vehicles (SDN-IOV) for the Smart Cities," in *Intelligence of Things: AI-IoT Based Critical-Applications and Innovations*, pp. 35–54, Springer, 2021.

[11] *WEKA: Data Mining Software in Java*.

[12] A. Mcgregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," in *International workshop on passive and active network measurement*, pp. 205–214, Springer, 2004.

[13] J. Erman, A. Mahanti, and M. Arlitt, "Qrp05-4: Internet traffic identification using machine learning," *IEEE Globecom*, pp. 1–6, 2006.

[14] P. Rani and R. Sharma *Intelligent transportation system for internet of vehicles based vehicular networks for smart cities. Computers and Electrical Engineering*, vol. 105, pp. 108543–108543, 2023.

[15] A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and Modeling of computer systems*, pp. 50–60, 2005.

[16] T. T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimize the use of machine learning classifiers in real-world ip networks," *Proceedings. 2006 31st IEEE Conference on Local Computer Networks*, pp. 369–376, 2006.

[17] R. Gowsalya and S. M. J. Amali, "Naive Bayes-based network traffic classification using correlation information," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 3, 2014.
[18] G. Ansari, P. Rani, and V. Kumar, "A Novel Technique of Mixed Gas Identification Based on the Group Method of Data Handling (GMDH) on Time-Dependent MOX Gas Sensor Data," in *InProceedings of International Conference on Recent Trends in Computing: ICRTC*, pp. 641–654, Springer Nature, 2022.
[19] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," *IEEE Transactions on neural networks*, vol. 18, no. 1, pp. 223–239, 2007.
[20] A. Pradhan, "Network Traffic Classification using Support Vector Machine and Artificial Neural Network," *International Journal of Computer Applications*, vol. 8, pp. 8–12, 2011.
[21] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?," *Acm Sigkdd Explorations Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.
[22] Z. Li, R. Yuan, and X. Guan, "Accurate classification of the internet traffic based on the SVM method," *2007 IEEE International Conference on Communications*, pp. 1373–1378, 2007.
[23] R. A. Gowsalya and S. M. J. Amali, "SVM-Based Network Traffic Classification Using Correlation Information," *International Journal of Research in Electronics and Communication Technology*, pp. 2348–9065, 2014.
[24] Y. Wang, Y. Xiang, J. Zhang, and S. Yu, "A novel semi-supervised approach for network traffic clustering," *5th International Conference on Network and System Security*, pp. 169–175, 2011.
[25] Y. Liu, W. Li, and Y. Li, "Network traffic classification using k-means clustering," *Second international multi-symposiums on computer and computational sciences*, pp. 360–365, 2007.
[26] S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) l*, pp. 250–257, 2005.
[27] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 2, pp. 23–26, 2006.
[28] Z. A. Shaikh and D. Harkut, "An overview of network traffic classification methods," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 2, pp. 482–488, 2015.
[29] P. Singhal, R. Mathur, and H. Vyas, "State the Art Review of Network Traffic Classification based on Machine Learning Approach," *International Journal of Computer Applications*, vol. 975, pp. 8887–8887, 2013.
[30] G. Suganya, "An efficient network traffic classification based on unknown and anomaly flow detection mechanisms," *Int. J. Comput. Trends Technol. (IJCTT)*, vol. 10, no. 4, 2014.
[31] S. Mukkamala, G. Janoski, and A. Sung, "Intrusion detection: support vector machines and neural networks," *Proceedings of the IEEE international joint conference on neural networks (ANNIE)*, pp. 1702–1707, 2002.
[32] M. V. Mahoney *A machine learning approach to detecting attacks by identifying anomalies in network traffic*, 2003.
[33] P. Laskov, P. Düssel, C. Schäfer, and K. Rieck, "Learning intrusion detection: supervised or unsupervised," in *International Conference on Image Analysis and Processing*, pp. 50–57, Springer, 2005.
[34] M. Zamani, M. Movahedi, M. Ebadzadeh, and H. Pedram, "A DDoS-aware IDS model based on danger theory and mobile agents," *2009 International Conference on Computational Intelligence and Security*, vol. 1, pp. 516–520, 2009.
[35] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," *2010 IEEE symposium on security and privacy*, pp. 305–316, 2010.
[36] T. Bujlow, T. Riaz, and J. M. Pedersen, "A method for network traffic classification based on C5. 0 Machine Learning Algorithm," *2012 international conference on computing, networking and communications (ICNC)*, pp. 237–241, 2012.
[37] A. Jamuna and V. Ewards, "Survey of traffic classification using machine learning," *International journal of advanced research in computer science*, vol. 4, no. 4, 2013.
[38] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 70–73, 2014.
[39] M. Blowers and J. Williams, "Machine learning applied to cyber operations," in *Network science and cybersecurity*, pp. 155–175, Springer, 2014.
[40] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are how you touch: User verification on smartphones via tapping behaviours," *2014 IEEE 22nd International Conference on Network Protocols*, pp. 221–232, 2014.
[41] K. Bartos, M. Sofka, and V. Franc, "Optimized invariant representation of network traffic for detecting unseen malware variants," *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 807–822, 2016.
[42] P. Wang, S. C. Lin, and M. Luo, "A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs," *2016 IEEE international conference on services computing (SCC)*, pp. 760–765, 2016.
[43] A. Furno, M. Fiore, and R. Stanica, "Joint spatial and temporal classification of mobile traffic demands," *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, 2017.
[44] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai *Kitsune: an ensemble of autoencoders for online network intrusion detection*, 2018.
[45] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network traffic classifier with convolutional and recurrent neural networks for the Internet of Things," *IEEE Access*, vol. 5, pp. 18042–18050, 2017.
[46] B. Mohammed, M. Hamdan, J. S. Bassi, H. A. Jamil, S. Khan, A. Elhigazi, D. B. Rawat, I. B. Ismail, and M. N. Marsono, "Edge Computing Intelligence Using Robust Feature Selection for Network Traffic Classification in Internet-of-Things," *IEEE Access*, vol. 8, pp. 224059–224070, 2020.
[47] M. Faheem, R. A. Butt, R. Ali, B. Raza, M. A. Ngadi, and V. C. Gungor, "CBI4. 0: A Cross-layer Approach for Big Data Gathering for Active Monitoring and Maintenance in the Manufacturing Industry 4.0," *Journal of Industrial Information Integration*, pp. 100236–100236, 2021.
[48] M. Faheem, G. Fizza, M. W. Ashraf, R. A. Butt, M. A. Ngadi, and V. C. Gungor *Big Data acquired by the Internet of Things-enabled industrial multichannel wireless sensors networks for active monitoring and control in the smart grid*, vol. 35, pp. 106854–106854, 2021.
[49] M. Faheem, M. W. Ashraf, R. A. Butt, B. Raza, M. A. Ngadi, and V. C. Gungor, "Ambient energy harvesting for low-powered wireless sensor network-based smart grid applications," *7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)*, pp. 26–30, 2019.
[50] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos, *File-sharing in the Internet: A characterization of P2P traffic in the backbone*. 2003.
[51] A. W. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in *International Workshop on Passive and Active Network Measurement*, pp. 41–54, Springer, 2005.