CGC International Journal of Contemporary Technology and Research
ISSN: 2582-0486 (online) Vol.-4, Issue-1 DOI: 10.46860/cgcijctr.2021.12.31.261

Analysis of Various Network Traffic Classification Techniques

Shivam Puri, Sukhpreet Kaur
Department of Computer Science,
Chandigarh Engineering College, Landran
Email Id: shivampuri6695@gmail.com

Abstract: **There are several interconnected entities present within the networked data for which the generation of inferences is important. For instance, hyperlinks are used to interconnect the web pages, calls are used to link the phone accounts, and references are used to connect the research papers and so on. Almost every existing application includes networks within it. The daily lives of individuals include social networking, making financial transactions, generating networks that show physical systems and so on. The manner in which the nodes present within the system influence each other can be known through this research. On the basis of observed attributed of an object within the system, another attributed is predicted using new model. The various network traffic classification techniques are reviewed in terms of certain parameters.**

Indexed Terms- Network Traffic Classification, Machine Learning, Data Analysis. (Keywords)

I. INTRODUCTION

Network Traffic Classification is an important topic nowadays in the field of Computer Science. It is very essential for Internet Service Providers (ISPs) to manage the overall performance of a network. Traffic classification is the first step to identity and classify unknown network classes. Through this technique, network operators can take some actions such as to block some flows and manage resources. Inadequate traffic treatment is one of the factors that impact the performance of schemes such as Network Survivability, Traffic Engineering, Quality of Service (QoS), Dynamic Access Control and so on, producing infrastructure scaling problems. Thus, Traffic Classification is a key mechanism for traffic treatment by providing a knowledge base for determining the levels of performance that are demanded by applications[1].

Network traffic classification is an essential feature for infrastructure management and to ensure the QoS of the different applications. In reality, a precise traffic classification procedure allows the efficient management of existing network resources, thereby permitting more accurate and robust resource allocation schemes. Network traffic classification is an important problem of network resource management that arises from analyzing network trends and network planning and designing. Approaches to network traffic classification vary according to the properties of the packets used. Some popular approaches for network traffic classification are identified as port-based approach, payload-based approach, host behavior-based approach and flow features-based approach. Port-based approach is the fastest and the simplest method to classify network traffic packets and as such has been extensively used[2].

The term 'Network traffic' defines the scale of operations in a network of computers. The computers linked to each other over a network send data back and forth data is sent back-and-forth as data packets. There are certain factors, such as volume of data packets, the time of broadcast and estimation of any delay, that determine the operation and traffic in the network. Most of the network sites and switches own traffic monitoring tool. This tool or software operates and maintains a record of every data packet, such as the original source, destination, distribution time and path followed. Network sniffer is a term used to represent a computer program accessing a network and reading data packets that are delivered on the network [3]. To detect such programs and alert users, it is possible to design a software that can monitor the traffic over network. Many approaches switch from their meaning when internet traffic is classified using machine learning. Figure 1 illustrates a simple network traffic framework.

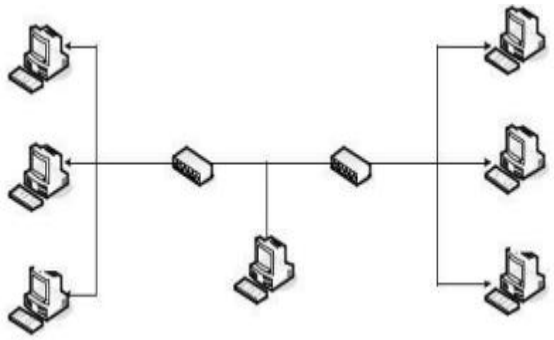

Figure 1: Network Traffic Model

Three terms related to the flow of network traffic have been defined below:
Stream or unidirectional stream: It denotes a sequence of packets that share the same pair of five components: IP address of source and destination, port numbers of source and destination, and protocol number.

➢ Bidirectional flow: It represents a tuple of unidirectional flows that runs in reverse directions among the IP addresses and ports of similar source and destination.

➢ Full flow: It is a bidirectional flow captured throughout its presence from making to end of the connection. A class usually specifies the internet traffic produced by a single application or an application set. Properties are typically numerical characteristics counted according to packets in the network packets [4].

Not all features equally affect the classification process, therefore, in practice, classifiers choose the smallest set of features that will lead to effective separation. All features

don't influence the classification process in same proportion, hence, in real time, classifier models select the minutest feature set for effectively classifying network traffic.

A. Network Traffic Classification Techniques

NTC has produced incredible concentration in the academic world alongside the industrial domain. A few procedures have been recommended and created in the course of the most recent twenty years. This segment makes a discussion on various classification strategies and partitions them into four classes dependent on their ordered development. There are basically four types of network traffic classification methods: port-based, payload-based, statistics-based, and behavioral-based. All these traffic classification techniques have been discussed below:

➢ Port-based classification: This approach often extracts the required value from the parcel header and afterward finds it in the table that has the port-application affiliations [5]. Tragically, Port-based classification has become generally inconsistent on the grounds that not all current applications utilize standard ports. A few applications even jumble themselves by utilizing the very much characterized ports of different applications. The payload-based technique looks for the application's signature in the payload of IP packets that can help keep away from the issue of dynamic ports. Henceforth, it is generally common in current industry items. Nonetheless, usually, the payload-based strategy comes up short with encrypted traffic.

➢ Payload-based classification: In order to conquer the lack and dependence on initial approaches, numerous industry items and exploration mechanisms have been carried out, in light of assessment past the headers of the packets to contents, a procedure recognized as payload-based classification and at times known as DPI is used. This technique depends with respect to examining packet elements and to match them with a deterministic arrangement of signatures that are kept. The after effects of this strategy for classifying the traffic are very precise. Payload assessment is generally utilized in a few businesses and openly available tools, such as for implementing Linux piece firewall [6].

➢ Statistical classification: This method makes the use of statistical qualities of traffic stream to distinguish the request. This strategy uses various stream level estimations, for instance, the span of the packet, length of packets, and free time for traffic flow. These estimations are remarkable for explicit sort of utilizations; thus, this permits the classifier to separate various applications from one another.

➢ Behavioral classification: This classification strategy notices the entire internet traffic got through the host, looking to distinguish the kind of use investigating the created internet traffic designs from the intended host. For instance, the quantity of interacted host is tallied, considering the transport layer protocol and the quantity of ports [7]. Despite the fact that the behavioral classification procedure provides optimal outcomes with least computational cost, the greater works concentrate just the end hosts. The constraints of this exploration are adopted in the technique to be applied.

B. General Process of Network Traffic Classification

Nowadays, the advent of different forms of services and applications have accentuated the significance of network functioning and control. Figure 2 explains the operation of the network traffic classification model. This model consists of many steps such as data collection, feature extraction, feature reduction and selection and, finally model development. This step-by-step process flow shows how network traffic classification methods identify/classify unknown forms of network traffic using machine learning algorithms:

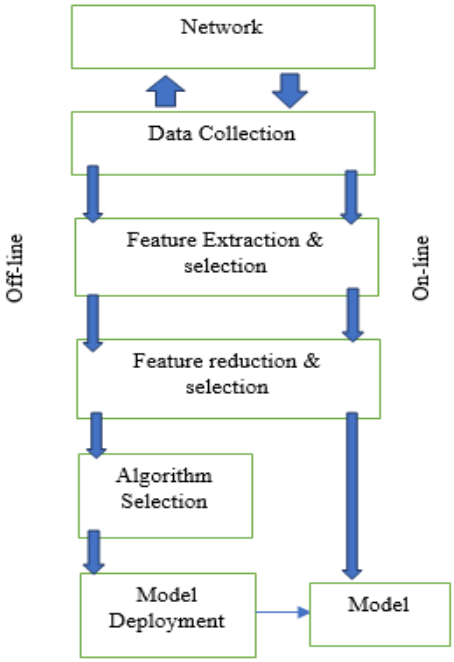

Figure 2: Network Traffic Classification Model

All tasks carried out in the above network classification models have been elaborated below:

➢ Data Collection: Classically, historical data has been a very important knowledge base for constricting machine learning models [8]. A plentiful and comprehensive set of conceptions about an issue has potential to upgrade the performance and generality of these paradigms. However, this factor is very important in the field of traffic classification due to several reasons. Some of these reasons include the complexity and scalability of web networks, the continual growth of traffic, and privacy rules not allowing the data collection. The phase of data collection allows the measurement of various conditions over the network. This phase mostly gathers IP runs within a timeframe. Moreover, this block consists of many tasks including packet management, flow reconstruction, and storage. It is essential to collect the historical dataset in offline flow. The online run, in contrast, constantly treats the packets' flow.

➢ Feature extraction: Appropriate features are extracted following the recording of the data that represents the

problem. It is a vital step as it permits to measure or compute features that might contain information concerning the process status [9]. Briefly, a feature extraction scheme calculates various metrics reflecting exclusive features in the collected data. Obtaining descriptors that better illustrate the issue is the major objective. The feature extraction process provides output as a structured table generated by feature columns. Every row is a pattern, with an extra random column representing each sample's current position (usually called a label or class). The patterns are not labelled when the status is not known. The online and offline stages involve computing attributes from past datasets and packets' flow [10], correspondingly. At this time, either feature selection or feature reduction schemes can treat the resultant attributes so that less space or a set of fresh attributes can be achieved. Following are the features used for network traffic classification.

i. Statistical based attributes: Features derived from the streams of packets are mostly statistically based attributes. The definition of these attributes is based on the belief that network layer' traffic has statistical properties (e.g., allocation of stream time, stream idle duration, packet inter-arrival time and length of packets). These features are exclusive to specific forms of services, and allow the separation of applications from various sources. Features e.g., IAT (Inter-Arrival Time) and packet length are regarded as the most considerable properties, along with their parameters including mean, maximum, mean, standard deviation, and so on.

ii. Graph based features: The internal structure of the Internet network enables modelling or representation of its communications in large interlinked graphs [11]. As a result, this scheme finds useful information from the network using graph premise. A network is considered to be a set of interlinked motes and assumes that nodes are hosts, and edges denote communication among hosts. These communications may be regarded as message sharing periods where packets are swapped. The usual process is making an observation end, for example a router. It is possible to aggregate packet sets propagating via an observation point into stream.

iii. Time-series based features: In general, time-series data can be regarded as a series of episodes sequenced in time sequence. In this context, the features are generally extracted on discrete-time data. The issue of network traffic has the appropriate features to be tackled as an episode-driven issue. The communication between duos (such as client-server) sturdily depends on time-sequenced episodes, e.g., opening or closing a interaction session, starting or ending the data transferring, and so on. This type of conditions prompts the usage of time-ordered attributes or data-driven schemes to find samples in networks [12].

- Feature selection and reduction: This step makes use of either feature selection or feature reduction schemes to treat resultant attributes to obtain less space or a set of new features. This is a voluntary process that allows to select or reduce the number of features extracted. Feature reduction is for creating new features using the original features, whereas feature selection is for finding a reduced set of attributes that better defines a procedure. These steps are intended to reduce issues, e.g., time expenditure and the obscenity of size and so on. These methods are usually classified into Filters, Wrappers and Embedded Schemes, which in turn can be devised by machine learning algorithms [13]. The supervised machine learning aims to discover the features that contribute most to defining the classification decision. The unsupervised learning, on the other hand, aims to determine the characteristics that enable the data clustering. Following are some popular approaches of feature selection:

i. Information Gain: This decision tree construction is one of the approaches used for ID3. After identifying the value of the features, it computes the number of information bits provided in the class prediction. The attribute with the maximal value of information gain is treated as the segmentation point, while the attribute with the lowest value indicates inaccuracy in data segmentation. It may be considered as the variance between the original information (which depends on the share of the class) and the fresh information (which is achieved after division).

ii. Gain Ratio: This scheme integrates the attribute's "split information" into the information gain measurement. Gain ratios uses a divided information measure to apply generalization of information gain in an attempt to remove information gain bias with lots of separate values. The ranking to the features is assigned according to the value of gain ratio. Therefore, if the value of the division point reaches zero, the proportion turns out to be uneven. Generally speaking, gain ratio as an information theoretical measure makes the feature selection with average-or-better gain. It has an upper hand over information gain because it does not treat features with multiple separate values [14].

iii. Principal Component Analysis (PCA):This approach denotes the original traffic data by exploring the K n-based vectors. Such data is presented in a very small space. This algorithm generates a small set of variables to combine the core of features. The input data signifies a linear mixture of the principal components, and explains the whole transformations with multiple components. It aims to give a real description by reducing dimensionality by means of linear equations. This algorithm transforms the data in a high-amplitude space into a space of low amplitudes.

The feature reduction or selection process is also concerned with algorithmic selection and model deployment steps, as some schemes choose the most applicable features on the basis of performance of a machine learning classifier.

- Algorithm selection: A novel dataset is generated from the original dataset on the basis of selected attributes [15]. The offline run makes the utilization of the new dataset for developing build models using which classification and regression tasks can be performed among other things. The Algorithm Selection block includes procedures and techniques for selecting the most adequate ML (machine learning) model. This approach is extensively executed for

discovering various solutions with the implementation of several ML models. For a variety of ML methods, it is essential to discover the best model for classifying the traffic. In general, the machine learning algorithm is capable of mapping the flows in accordance with the discriminative attributes. Thereafter, the learned rules are considered to classify the unknown traffic. The selected features are useful for classification performance and its complexity. Particularly, an enormous attribute set is necessitated to classify a huge number of applications so that the sufficient accuracy can be acquired. Nevertheless, the computational complexity of the classification process is enlarged due to the maximization of the size of the attribute set. Various works are reported on the application of ML (machine learning) models to classify the network traffic over the previous years. The classification of works is done in two classes: supervised techniques or unsupervised techniques.

i. Supervised Methods: The supervised traffic classifiers focuses on analyzing the supervised training data and generating an inferred function that assists in predicting the output class for any testing flow [16]. In this, the sufficient supervised training data is a general hypothesis to classify the traffic. Some popular supervised algorithms are NB (Naive Bayes) with discretization, NB with kernel density estimation, C4.5 DT (decision tree), Bayesian network, and NBT (Naive Bayes Tree).

ii. Unsupervised Methods: These techniques focus on discovering cluster structure in unlabeled traffic data and assigning any testing flow to the application-based class of its nearest cluster. The empirical research demonstrated that the traffic clustering is capable of generating high-purity clusters in case of greater number of clusters in contrast to the number of real applications. In general, the clustering methods are adaptable for discovering the traffic from earlier unknown applications.

Some well-known machine learning models have been discussed as follows:

i. Linear regression: This algorithm is considered as a part of the family of regression algorithms for investigating the relationships and dependencies among variables. A modelling association is presented amid a continuous scalar dependent variable y that is label or target in ML terminology and one or more explanatory variables such as independent variables, input variables, features, observed data, observations, attributes, dimensions, data point, etc. The X is used to represent the explanatory variables. The regression analysis emphasizes on predicting a continuous target variable [17]. On the other hand, another area known as classification is aimed to predict a label from a finite set. The model for a multiple regression in which linear combination of input variables is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + e \qquad \text{-------1}$$

LR also comes under the category of supervised learning algorithms. A set of labelled data is utilized for training this model. Afterward, the model is executed for predicting the labels on unlabeled data.

ii. Decision tree: DT is extensively utilized classification model adopted in various real-world applications. This symbolic learning method emphasizes on correlating the information that is achieved from a training dataset in which nodes and ramifications are contained in a stratified structure. The fundamental objective of this classification model is that the least squares error can be decreased for the next split of a node in the tree so that the average of the dependent variable of all training instances that covered for unseen instances in a leaf can be predicted. A Decision Tree model $T\left(x;\{R_j\}_{j=1}^{J}\right)$ is employed for partitioning the x- space into J disjoint regions denoted with $\{R_j\}$ and predicting a separate constant value in each one as:

$$x \in R_j \Rightarrow T\left(x;\{R_j\}_{j=1}^{J}\right) = \hat{y}_j \qquad \text{-----2}$$

Or equivalently

$$T\left(x;\{R_j\}_{j=1}^{N}\right) = \sum_{j=1}^{J} \hat{y}_j \, I(x \in R_j) \qquad \text{------3}$$

In this, $\hat{y}_j = \frac{1}{a_j}\sum_{i=1}^{a_j} y_i$ denotes the mean of the response y in each region R_j, $y_i \in R_j$, a_j is used to represent the size of region R_j. Thus, a tree can predict a constant value y_j in each region R_j. The top-down iterative division is utilized on the basis of a least squares fitting criterion to develop the trees [18]. The identities of the predictor variables useful for splitting and their corresponding split points are considered in this model to deal with the regions $\{R_j\}_{j=1}^{J}$ of the partition.

iii. KNN algorithm: This algorithm is the earliest and simplest technique to classify the pattern. However, the K-Nearest Neighbour generates optimal results and in some specific domains, its integration with a prior knowledge leads to enhance the existing model. The KNN rule can classify each unlabeled example using the majority label among its k-nearest neighbors in the training set. The performance of this model is depending depends upon the distance metric which assists in recognizing the nearest neighbors [19]. When the prior knowledge is not available, a simple Euclidean metric is exploited in KNN for quantifying the dissimilarities among examples represented as vector inputs. Euclidean distance is described as:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r}\left(a_r(x_i) - a_r(x_j)\right)^2 \qquad \text{------4}$$

In this, an instance is defined as a vector $x = (a_1, a_2, a_3, \ldots, a_n)$, n denotes the dimensionality of the vector input called the number of an example's attributes. a_r reveals the rth attribute of an instance, w_r illustrates the weight of the rth attribute, $r$lies in 1 to n, the smaller $d(x_i, x_j)$ shows that two examples are more similar.

$$y(d_i) = \underset{k}{argmax} \sum_{x_j \in KNN} y(x_j, c_k) \quad \text{-----5}$$

In which d_i utilized as a test example, x_j denotes one of its k nearest neighbors in the training set, $y(x_j, c_k)$ is employed to determine whether x_j belongs to class c_k. The given equation means that the prediction is the class which contains most members in KNN. To illustrate, in order to generate 5-nearest neighbor algorithm the, three of these nearest neighbors comes under a class One and the other two belong to class Two. Subsequently, it is concluded that the test example comes under to class One [20].

- Deployment of Model: The main purpose of this block is to oversee the implementation and restricting efforts of machine learning strategies in real-time conditions of a network.
- Classifier Validation: This step gives an overview of the most commonly used validation strategies for classification tasks. Supervised learning needs prior knowledge of Patten labels. This information contributes a lot for validating machine learning classifiers. The general approach is to split the dataset into two subsets of training and testing. Training sets are used to construct machine learning classifiers, while the test sets evaluate the resultant models to assess their prediction potential. Machine learning models are validated by applying most common strategies that quantify their performance with respect to the classification abilities.

II. LITERATURE REVIEW

A. Network Traffic Classification using Deep Learning

Mingze Song, et.al (2019) suggested a NTC technique on the basis of text CNN (Convolutional Neural Network) to represent the traffic data as vectors so as the traffic was classified [21]. Afterward, the text CNN was deployed for extracting the key attributes to carry out the traffic classification. ISCX VPN-non-VPN dataset was applied to authenticate the suggested technique and performed more effectively as compared to earlier NTC technique. The issue of class imbalance was resolved using a new loss function and a suitable technique of allocating the class weight in multi-class classification task. A 3D-CNN (three dimensional-Convolutional Neural Network) was introduced byJing Ran, et.al (2018) to classify the traffic effectively [22]. This model was capable of extracting the spatial and temporal attributes in automatic manner and discovering the best attributes whenthe iterations were validated. Afterward, a reasonable classification was carried out. The USTC-TFC2016 was executed to conduct experiments for quantifying the introduced model. The results revealed the efficacy of the introduced model for classifying the network traffic. A DL (deep learning)-based method was designed by Ly Vu, et.al (2018)to utilize the correlation of time series data samples of the encrypted network applications [23]. The efficacy of the designed method was computed to address the issue related to classify the network classify. The VPN-nonVPN dataset was applied in the experimentation concerning precision, recall, and F1-score. The experimental outcomes exhibited that the designed method had potential for enhancing the performance to recognize the encrypted application traffic with regard to accuracy and computation efficacy.

A tree structural recurrent neural network (Tree-RNN) was projected by Xinming Ren, et.al (2020) in order to classify the network traffic. This algorithm, containing a number of classification models, had potential to robotically learn the nonlinear relationship established amid the input and output data without extracting attributes [24]. The experimental results obtained on ISCX traffic dataset indicated that the projected algorithm performed well in contrast to the benchmark methods and had utilized least training time. The accuracy of this algorithm was 4.88% above. A new system called SPCaps (session-packets-based capsule neural networks) was established by Susu Cui, et.al (2019) to classify the network traffic [25]. The interference traffic was lessened and the weight of effective traffic was boosted with the implementation of a twice-segmentation method. Thereafter, CapsNet was exploited to learn the spatial attributes of encrypted traffic.

A softmax classifier was applied to generate the outcomes of encrypted NTC (network traffic classification). ISCX VPN-nonVPN was implemented for evaluating the designed system while classifying the encrypted network traffic concerning service and application. The experimental outcomes validated that the established system performed more effectively as compared to other methods. A semi-supervised learning technique was investigated by Auwal Sani Iliyasu, et.al (2020) in which DCGAN (Deep Convolutional Generative Adversarial Network) was deployed to classify the internet traffic [26]. This approach focused on applying the samples that this technique had produced for boosting the performance of a classifier whose training was done on a few labeled samples. Hence, the complexities related to gather and label the dataset were mitigated. The suggested approach obtained the accuracy around 89% on QUIC and 78% on ISCX VPN-NonVPN datasets.

TABLE -1: COMPARISON TABLE

Author	Year	Technique used	Dataset	Outcomes
Mingze Song, et al.	2019	text CNN (Convolutional Neural Network)	ISCX VPN-non VPN dataset	The suggested technique performed more effectively as compared to earlier NTC technique.
Jing Ran, et al.	2018	3D convolutional neural network	USTC-TFC2016	The results revealed the efficacy of the introduced model for classifying the network traffic.

Ly Vu, et al.	2018	DL (deep learning)-based method	VPN-nonVPN dataset	The designed method had potential for enhancing the performance to recognize the encrypted application traffic with regard to accuracy and computation efficacy.
Xinming Ren, et al.	2020	tree structural recurrent neural network (Tree-RNN)	ISCX traffic dataset	This algorithm enhanced the accuracy by 4.88%.
Susu Cui, et al.	2019	SPCaps (session-packets-based capsule neural networks)	ISCX VPN-nonVPN	The established system performed more effectively as compared to other methods.
Auwal Sani Iliyasu, et al.	2020	DCGAN (Deep Convolutional Generative Adversarial Network)	QUIC protocol and ISCX VPN-NonVPN	The suggested approach obtained the accuracy around 89% on QUIC and 78% on ISCX VPN-NonVPN datasets.

B. Network Traffic Classification using Machine Learning

Shi Dong, et.al (2021) presented CMSVM (Cost-Sensitive SVM) which was the improved variant of the original SVM (Support Vector Machine) as a solution of the imbalance issue in network traffic classification [27]. The new approach used a multi-class SVM algorithm with active learning to assign a weight for applications dynamically. This work used two different datasets (MOORE_SET and NOC_SET) to analyze the classification accuracy and performance of the new algorithm.

The obtained outcomes confirmed that the CMSVM algorithm was capable enough to decrease computing overhead, enhance classification accuracy and dealing with the imbalance issue as compared to other ML (Machine Learning) methods. An improved stacked auto-encoder was developed by Peng Li, et.al (2018) for learning the complex connections across the multi-source network flows. For this purpose, various basic Bayesian auto-encoders were stacked [28]. In addition, the BP (back-propagation) model was implemented to train the stacked auto-encoder so that the complex associations were observed on the network flows. In the end, the MAWI and DARPA 99 were applied for carrying out the experiments on the developed system. The outcomes validated the superiority of the developed model over the conventional models with regard to accuracy while classifying the internet traffic.

A new method named GMM (Gaussian Mixture Models) was put forward by Hassan Alizadeh, et.al (2020) to classify and verify the network traffic [29]. A separate GMM was created for every class of applications with the help of CEM (component-wise expectation-maximization) during matching of network traffic distribution generated through these applications. The traffic was classified more effectively and at time using only the first initial packets of truncated flows. A dataset which was publicly available was executed to perform the experiments. The results proved that the presented method was more efficient as compared to other techniques and provided the accuracy around 97.7%. The FLAGB (focal loss based adaptive gradient boosting) system was recommended by Yu Guo, et.al (2020) to classify the imbalanced traffic [30]. This system was adapted to classify the internet traffic with diverse imbalance levels and dealing with imbalance without any prior knowledge of data distribution. The experiments were carried out on two network traffic datasets that had binary class and multiple-class. The experimental outcomes exhibited that the recommended system was more applicable in contrast to the existing models.

The time consumed by this system in training phase was found lower and it performed well to classify the imbalanced traffic. James MsughterAdeke, et.al (2020) intended to implement ML (Machine Learning) technique based on parameter optimization [31]. Three ML algorithms were selected at random for validating the presented technique on WEKA (Waikato Environment for Knowledge Analysis) software. The USTC-TFC2016 dataset was applied for the experimentation on 3 scenarios. The outcomes demonstrated that the RF (Random Forest) provided a superior accuracy up to 99.52% in contrast to the existing approach. Furthermore, the accuracy obtained for classifying traffic from the Decision Tree (J48) was computed greater in comparison with other techniques. A supervised scheme called XGBoost (eXtreme Gradient Boosting) was formulated by Iyad Lahsen Cherif, et.al (2019) in order to classify the traffic [32]. The outcomes of evaluation demonstrated that the accuracy of formulated algorithm was calculated 99.5% on a dataset which had real flows. This algorithm yielded superior accurately as compared to other ML algorithms.

TABLE -2: COMPARISON TABLE

Author	**Year**	**Technique used**	**Dataset**	**Outcomes**
Shi Dong, et al.	2021	CMSVM (Cost-Sensitive SVM)	MOORE_SET and NOC_SET	The CMSVM algorithm was capable enough to decrease computing overhead, enhance classification accuracy.
Peng Li, et al.	2018	Improved stacked auto-encoder	MAWI and DARPA 99	The outcomes validated the superiority of the developed

				model over the conventional models with regard to accuracy.
Hassan Alizadeh, et al.	2020	Gaussian Mixture Models (GMMs)	UNIBS-2009 dataset	The presented method was proved more efficient as compared to other techniques and provided the accuracy around 97.7%.
Yu Guo, et al.	2020	Focal loss based adaptive gradient boosting framework (FLAGB)	BOT dataset and KDD99' dataset.	The time consumed by this system in training phase was found lower and it performed well to classify the imbalanced traffic.
James Msughter Adeke, et al.	2020	Random Forest (RF), Decision Tree (J48)	USTC-TFC2016 dataset	The outcomes demonstrated that the RF (Random Forest) provided a superior accuracy up to 99.52% in contrast to the existing approach.
Iyad LahsenCherif, et al.	2019	eXtreme Gradient Boosting (XGBoost)	R-Studio software	The accuracy of formulated algorithm was calculated 99.5% on a dataset which had real flows.

C. Network Traffic Classification using Hybrid Techniques

Abid Saber, et.al (2018) established a technique in order to classify the traffic in one step [33]. The PCA (Principal Component Analysis) was adopted to integrate the over-sampling with under-sampling. This approach was useful to select the optimum feature subset prior to classify the network traffic with the help of SVM (Support Vectors Machine). The evaluations were carried out using 14 kinds of traffic. Only the time-based flow-based attributes were employed. Various experiments were conducted to compute the efficiency of established technique using the UNB ISCX dataset. The established technique performed more effectively to classify the traffic in a single process with shorter flow timeout values. An innovative technique of classifying the encryption traffic was projected by Boyu Sun, et.al (2020) for learning the feature representation from the traffic structure and the traffic flow data [34].

A KNN (K-Nearest Neighbor) traffic graph was developed for representing the structure of traffic data. The traffic attributes were extracted and the encrypted traffic was classified using two-layer GCN (Graph Convolutional Network) architecture. Moreover, the auto-encoder was implemented in order to learn the representation of the flow data. The results of experiment conducted on two publicly available datasets demonstrated that the projected technique had generated the promising outcomes in contrast to other algorithms. Yu Wu, et.al (2018) emphasized on improving the traditional TDM-EPON (time-division multiplexing Ethernet passive optical network) model [35]. This model employed ML (machine-learning) technique for classifying the traffic initially and shifted the inadequate traffic for avoiding the transmission of unnecessary EPON frames in second method. Two feature-selection techniques were utilized with classifier biasedness in the initial method for acquiring the best classification outcomes. The sifting-based hybrid bandwidth allocation technique deployed these outcomes in the second method.

The simulation outcomes revealed that the presented model performed efficiently with regard to significant per-RRH traffic load and SNR (signal-to-noise ratio) and E2E (end-to-end) delay of this model was found under 100 µs. A new system known as MGHMM was introduced by Zhongjiang Yao, et.al (2020) on the basis of GMM (Gaussian mixture model) and HMM (hidden Markov model) [36]. Initially, the efficiency of the introduced system was computed by classifying the protocols and recognized the obfuscated traffic in the experimentation. Subsequently, a comparative analysis was conducted on this system against another classifier. The results indicated that the introduced system had generated optimal outcomes and least computational cost. An approach was suggested by Maonan Wang, et.al (2020) in which CNN (convolutional neural network) and SAE (stacked autoencoder) models were implemented [37]. The high-level attributes were extracted from the raw network traffic using CNN. Twenty-six statistical attributes counted from raw traffic in direct manner were encoded through SAE. The integration of outcomes attained from both the models was done into novel high-level features. The ISCX VPNnonVPN dataset was adopted to determine that the suggested approach was feasible. The suggested approach assisted in enhancing the performance to classify the encrypted traffic and yielded the f1-score around 0.98.

TABLE -3: COMPARISON TABLE

Author	**Year**	**Technique used**	**Dataset**	**Outcomes**
Abid Saber, et al.	2018	PCA (Principal Component Analysis)	UNB ISCX dataset	The established technique performed more effectively to classify the traffic in a single process with shorter flow timeout values.
Boyu Sun, et al.	2020	Graph Convolutional Network (GCN)-autoencoder	ISCX VPN-nonVPN	The projected technique had generated the promising outcomes in contrast to other

				algorithms.
Yu Wu, et al.	2018	TDM-EPON	MATLAB 2017a LTE System Toolbox	The presented model performed efficiently with regard to significant per-RRH traffic load and SNR (signal-to-noise ratio) and its E2E (end-to-end) delay was found under 100 μs.
Zhongjiang Yao, et al.	2020	Gaussian mixture models and hidden Markov models (MGHMM)	UNB opentraffic dataset	The results indicated that the introduced system had generated optimal outcomes and leastcomputational cost.
Maonan Wang, et al.	2020	Convolutional neural network (CNN) and a stacked autoencoder (SAE)	ISCX VPNnonVPN dataset	The suggested approach assisted in enhancing the performance to classify the encrypted traffic and yielded the f1-score around 0.98.

III. REVIEW METHODOLOGY

This section presents the state-of-the-art review layout, a step-by-step method for the literature discussed in the previous sections. This research focuses on categorizing the current literature on network traffic detection assessing the current trends. This evaluation finds relevant research articles from reputable electronic databases and the top conferences in the field. After then, inclusion and exclusion criteria were used to reduce the number of papers that were considered. Following that, final research studies were chosen based on a variety of variables. The information given here is the product of a thorough investigation. For this review study, various electronic database sources were investigated; some of the popular electronic databases used in this search like google scholar, Elsevier, Science direct etc.

Using the inclusion criterion, which mainly depends on the techniques, the relevant work of network traffic classification algorithms is retrieved from the enormous collection of data given by search engines. The data shows that journals account for most of the work in this study (51%), with conferences accounting for 40% of the work and book chapters accounting for 9%. In addition, the data depicts a year-by-year study of work relevant to network traffic classification. The major data is available on the google scholar as compared to Elsevier and Science direct. The google scholar has 60 percent data, Elsevier has approx. 10 percent and Science direct has approx. 30 data on network traffic classification. The data division is shown in figure 3:

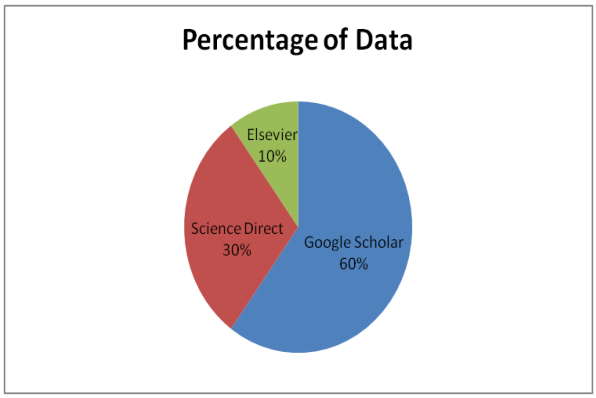

Figure 3: Percentage of Data Sharing

As shown in figure 3, the percentage of data sharing is shown in figure approx. 60 percent data is available on Google scholar, 30 percent is available on science direct and very less amount of data that is 10 percent is available on Elsevier.

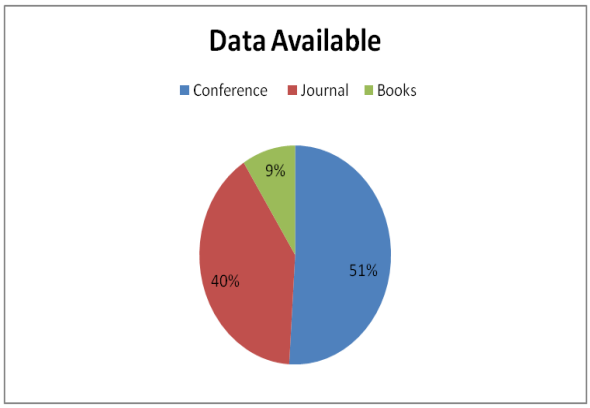

Figure 4: Data Available

As shown in figure 4, the data is available which is available through conference, journal and books. The conference has approx. 51 percent of total data, 40 percent data has available through journal and 9 percent is available through books.

IV. CONCLUSION

An important machine learning task that collects the information from labeled data sets and predicts the unlabeled samples is known as data classification. There are several data classification mechanisms such as Naïve Bayes, KNN and SVM proposed so that different applications can apply them. Several data classification systems like SVM, KNN, Naïve Bayes, etc. have been proposed over the years since there are large numbers of applications. The tuning of parameters is very important to consider a classifier. For example, in C4.5 decision tree, least number of cases needed to divide a set and the confidence factor are important to be considered. The various machine learning algorithms are reviewed in this paper for the network traffic classification. It is analyzed that ensemble classification model needs to derive in future for the network traffic classification

REFERENCES

[1] Jaehwa Park, JunSeong Kim, "A classification of network traffic status for various scale networks", 2013, The International Conference on Information Networking 2013 (ICOIN)

[2] Ji-hye Kim, Sung-Ho Yoon, Myung-Sup Kim, "Study on traffic classification taxonomy for multilateral and hierarchical traffic classification", 2012, 14th Asia-Pacific Network Operations and Management Symposium (APNOMS)

[3] Rui Yang, "The Comparison of Split-Flow Algorithms in Network Traffic Classification: Sequential Mode vs. Parallel Model", 2013, International Conference on Information Technology and Applications

[4] Zeba Atique Shaikh, Dinesh G. Harkut, "A Novel Framework for Network Traffic Classification Using Unknown Flow Detection", 2015, Fifth International Conference on Communication Systems and Network Technologies

[5] Shashikala Tapaswi, Arpit S. Gupta, "Flow-Based P2P Network Traffic Classification Using Machine Learning", 2013, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery

[6] Sung-Ho Lee, Jun-Sang Park, Sung-Ho Yoon, Myung-Sup Kim, "High performance payload signature-based Internet traffic classification system", 2015, 17th Asia-Pacific Network Operations and Management Symposium (APNOMS)

[7] Yaojun Ding, "Imbalanced network traffic classification based on ensemble feature selection", 2016, IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)

[8] Zhengwu Yuan, Chaozheng Wang, "An improved network traffic classification algorithm based on Hadoop decision tree", 2016, IEEE International Conference of Online Analysis and Computing Science (ICOACS)

[9] Yang Hong, Changcheng Huang, Biswajit Nandy, Nabil Seddigh, "Iterative-tuning support vector machine for network traffic classification", 2015, IFIP/IEEE International Symposium on Integrated Network Management (IM)

[10] Chao Wang, Tongge Xu, Xi Qin, "Network Traffic Classification with Improved Random Forest", 2015, 11th International Conference on Computational Intelligence and Security (CIS)

[11] Jie Yang, Zheng Ma, Chao Dong, Gang Cheng, "An empirical investigation into CDMA network traffic classification based on feature selection", 2012, The 15th International Symposium on Wireless Personal Multimedia Communications

[12] Hui Dong, Guang-Lu Sun, Dan-Dan Li, "A hybrid method for network traffic classification", 2013, Proceedings of 2013 2nd International Conference on Measurement, Information and Control, Volume: 01

[13] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, Nabin Kumar Karn, FoudilAbdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms", 2016, 2nd IEEE International Conference on Computer and Communications (ICCC)

[14] Hardeep Singh, "Performance Analysis of Unsupervised Machine Learning Techniques for Network Traffic Classification", 2015, Fifth International Conference on Advanced Computing & Communication Technologies

[15] Matija Stevanovic, Jens Myrup Pedersen, "An analysis of network traffic classification for botnet detection", 2015, International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)

[16] Hyun-Kyo Lim, Ju-Bong Kim, Kwihoon Kim, Yong-Geun Hong and Youn-Hee Han, "Payload-Based Traffic Classification Using Multi-Layer LSTM in Software Defined Networks", 2019, Applied Sciences

[17] Jie Cao, Da Wang, Zhaoyang Qu, Hongyu Sun, Bin Li and Chin-Ling Chen, "An Improved Network Traffic Classification Model Based on a Support Vector Machine", 2020, Symmetry.

[18] Aafa J S, Soja Salim, "A Survey on Network Traffic Classification Techniques", 2014, International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 3

[19] Argha Ghosh, Dr. A. Senthilrajan, "Classifying Network Traffic Using DPI And DFI", 2019, International Journal of Scientific & Technology Research, Vol. 8, No. 11

[20] Brian Schmidt, DionysiosKountanis, Ala Al-Fuqaha, "A Biologically-Inspired Approach to Network Traffic Classification for Resource-Constrained Systems", 2014, IEEE/ACM International Symposium on Big Data Computing

[21] Mingze Song, Jing Ran, Shulan Li, "Encrypted Traffic Classification Based on Text Convolution Neural Networks", 2019, IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)

[22] Jing Ran, Yexin Chen, Shulan Li, "Three-Dimensional Convolutional Neural Network based Traffic Classification for Wireless Communications", 2018, IEEE Global Conference on Signal and Information Processing (GlobalSIP)

[23] Ly Vu, Hoang V. Thuy, Quang Uy Nguyen, Tran N. Ngoc, Diep N. Nguyen, Dinh Thai Hoang, Eryk Dutkiewicz, "Time Series Analysis for Encrypted Traffic Classification: A Deep Learning Approach", 2018, 18th International Symposium on Communications and Information Technologies (ISCIT)

[24] Xinming Ren, Huaxi Gu, Wenting Wei, "Tree-RNN: Tree structural recurrent neural network for network traffic classification", 2020, Expert Systems with Applications

[25] Susu Cui, Bo Jiang, Zhenzhen Cai, Zhigang Lu, Song Liu, Jian Liu, "A Session-Packets-Based Encrypted Traffic Classification Using Capsule Neural Networks", 2019, IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)

[26] Auwal Sani Iliyasu, Huifang Deng, "Semi-Supervised Encrypted Traffic Classification With Deep Convolutional Generative Adversarial Networks", 2020, IEEE Access

[27] Shi Dong, "Multi class SVM algorithm with active learning for network traffic classification", 2021, Expert Systems with Applications

[28] Peng Li, Zhikui Chen, Laurence T. Yang, Jing Gao, Qingchen Zhang, M. Jamal Deen, "An Improved Stacked Auto-Encoder for Network Traffic Flow Classification", 2018, IEEE Network

[29] Hassan Alizadeh, Harald Vranken, André Zúquete, Ali Miri, "Timely Classification and Verification of Network Traffic Using Gaussian Mixture Models", 2020, IEEE Access

[30] Yu Guo, Zhenzhen Li, Zhen Li, Gang Xiong, Minghao Jiang, Gaopeng Gou, "FLAGB: Focal Loss based Adaptive Gradient Boosting for Imbalanced Traffic Classification", 2020, International Joint Conference on Neural Networks (IJCNN)

[31] James MsughterAdeke, Jinfu Chen, Lei Zhang, Richard N. K. Mensah, Kun 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)

[32] Iyad LahsenCherif, AbdesselemKortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification", 2019, Wireless Days (WD)

[33] Abid Saber, BelkacemFergani, Moncef Abbas, "Encrypted Traffic Classification: Combining Over-and Under-Sampling through a PCA-SVM", 2018, 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)

[34] Boyu Sun, Wenyuan Yang, Mengqi Yan, Dehao Wu, Yuesheng Zhu, Zhiqiang Bai, "An Encrypted Traffic Classification Method Combining Graph Convolutional Network and Autoencoder", 2020, IEEE 39th International Performance Computing and Communications Conference (IPCCC)

[35] Yu Wu, Massimo Tornatore, Yongli Zhao, Biswanath Mukherjee, "Traffic classification and sifting to improve TDM-EPON fronthaul upstream efficiency", 2018, IEEE/OSA Journal of Optical Communications and Networking

[36] Zhongjiang Yao, Jingguo Ge, Yuxiang Ma, "Encrypted traffic classification based on Gaussian mixture models and Hidden Markov Models", 2020, Journal of Network and Computer Applications

[37] Maonan Wang, Kangfeng Zheng, Dan Luo, Yanqing Yang, Xiujuan Wang, "An Encrypted Traffic Classification Framework Based on Convolutional Neural Networks and Stacked Autoencoders", 2020, IEEE 6th International Conference on Computer and Communications (ICCC)

.