ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 7, July 2021

A Machine Learning Approach for Network Traffic Classification

[1] Isha, [2] Prof. Jasbir Singh Saini, [3] Prof. Kamaldeep Kaur
[1] Research Scholar, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India
[2] Associate Professor, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India
[3] Assistant Professor, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India

Abstract: The network traffic classification task is focused on recognizing diverse varieties of applications or traffic specifics for which the sustained data packets get analysed that is indispensable in transmission networks in these days. The traffic can be classed in several phases, in which the step of preprocessing get achieved, extracted and classification of the attributes is performed. Meanwhile, the processing of dataset is carried out as it is taken as input in the process of classification. The dataset is split into two positions, which are named as training and testing. The training set includes 70% of the entire dataset and testing set has 30%. The voting classification method is implemented in which the KNN (K-Nearest Neighbour) is integrated with the RF (Random forest) and SVM (Support Vector Machine). The suggested method is deployed in the programming language python and several parameters instance as recall, precision and accuracy are taken in the account for quantifying the outcome. This indicates that suggested method yields higher accuracy, precision and recall in comparison with the traditional classification models.

Keywords: Machine Learning, Cyber Security, Intrusion Detection System.

INTRODUCTION

The growing importance of the Internet since its birth has brought privacy and security assurances into the limelight. An effort has been made to meet these demands by designing a broad spectrum of privacy-preserving techniques, for example proxy servers, VPNs (virtual private networks) and AMs (anonymity mechanisms). The proxy sites play the role of a facilitator for web surfers and allow them not only to obscure the character of the content exchanged for information sharing besides any spying object [1]. Traffic classification (TC) is considered to be an elementary unit of extreme significance for QoS (quality-of-service) implementation, traffic production, network protection, in particular, that is, referring to the nature of traffic generated by a network object. Traffic classification also delivers a major contribution to the detection of miscellaneous attacks [2]. Nowadays, the advent of different forms of services and applications have accentuated the significance of network functioning and control. Figure 1 explains the operation of the network traffic classification model. This model consists of many steps such as collection of the data, extraction of the features meanwhile the reduction and selection of features and, finally model development[3]. This step-by-step process flow shows how network traffic classification methods identify/classify unknown forms of network traffic using machine learning algorithms.

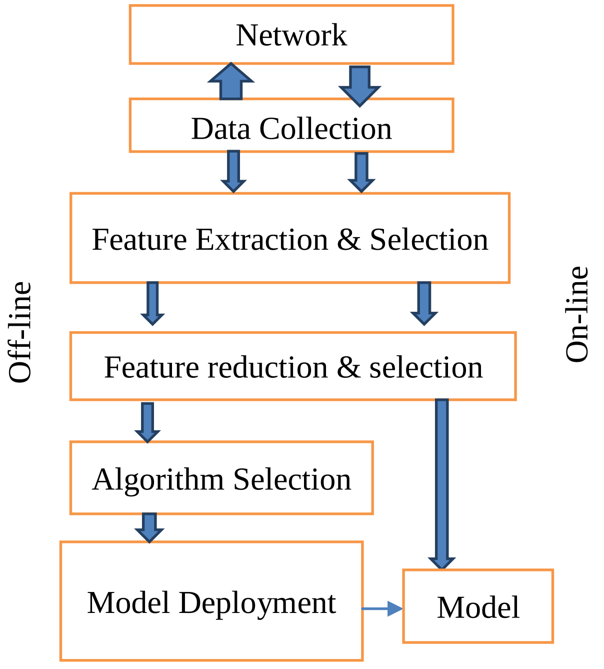


Figure 1 Network Traffic Classification Model

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

a. **Data Collection:** Classically, historical data has been a very important knowledge base for constricting machine learning models [4]. A plentiful and comprehensive set of conceptions about an issuehas 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[5]. 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.

b. **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 [6]. 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.

c. **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[7]. 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 [8].

d. **Classification:** A novel dataset is generated from the original dataset on the basis of selected attributes [9]. 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[10]. For a variety of ML methods, it is essential to discover the best model for classifying the traffic.

LITERATURE SURVEY

Hassan Alizadeh,et.al(2020) suggested the innovative method in way to classed the network traffic with an implemented GMM (Gaussian Mixture Model) [11]. The Component wise expectation maximization abbreviated as CEM was exploited for making a separate GMM in way so that the trafficdistribution was go with similarity. The suggested had classified and verified the traffic on time efficiently using only preliminary packets of truncated flows. A publicly available dataset taken from a real network was utilized for conducting the experiments in order to compute this technique. The experimental outcomes demonstrated that the suggested technique had attained the accuracy around 97.7% for classifying the network flow in comparison with other methods.

Won-Ju Eom, et.al (2021) introduced a model recognized as LightGBM model with the help of SDN (software-defined network) architecture for classifying the network traffic [12]. This model was established in the network controller with the purpose of leveraging the better computational capacity of the SDN controller to classify the network traffic in real-time, adaptively and accurately. Four ensemble algorithms were deployed and their efficacy to classify the model was analyzed. Moreover, the suggested model performed more effectively in classifying the network traffic.

Madhusoodhana Chari S.,et.al(2019) intended the packet size signature extracting based method for the classification distinct classes including Audio and Video streaming ,the Browsing, Chat, P2P etc. [13]. For this purpose, the classes of network traffic were recognized by the training a J48 DT (decision tree) classification algorithm with a new feature set. The interpretability of the model was described. This set had provided a tree which was found more balanced and capable of producing the lower count of rules for individual class. The set provided interpretability to the intended method and easy deployment.

Jing Ran, et.al (2018) developed three-dimensional CNN (3D convolutional neural network) system in order to classify the network traffic [14]. These attributes were more representative as compared to others which were selected manually. When the feature extractor was integrated with classification technique, the global optimum was obtained as the effective classification algorithm was found the best extractor but had not provided satisfying outcomes in case the cooperation was not good. The USTC-TFC2016 dataset was applied to carry out the series of experiments. The experimental results confirmed the efficacy of developed system over the traditional algorithm with regard to accuracy.

Jiwon Yang, et.al (2019) projected a traffic classification technique to classify the encrypted traffic flows [15]. A new payload-based classification was put forward using which the unencrypted handshake packets were utilized that had exchanged amid the hosts to stable the transport layer security. The Bayesian neural network was employed as the classed technique where the cipher suite,compression technique related to the suite-packets was considered as the insertion. The investigation were carried out and outcomes depicted that the projected technique performed more efficiently in comparison with other conventional payload-based classification algorithms.The future work would focus on extending by classifying other secure protocols.

Yu Wu, et.al (2018) designed an approach for enhancing the classic,The time-division multiplexing Ethernet passive optical network framework [16]. Meanwhile designed approach made the deployment of two methods. Initially, the ML (machine-learning) models were deployed in order to classify the upstream traffic as useful and useless classes. Subsequently, sifting useless traffic was applied for avoiding the transmission of redundant EPON (Ethernet passive optical network) frames. The optimal outcomes were obtained by integrating baseness of 2 classifiers in the integrated method using 2 feature-selection techniques. In the second method, the hybrid bandwidth allocation system had utilized it as an input. The simulation outcomes revealed that the designed algorithm offered promising improvements with regard to per-RRH traffic load and SNR,the expaned form is signal-to-noise ratio and kept the E2E, the expanded form is end-to-end delay under 100 μs.

Pratibha Khandait, et.al (2020) formulated an efficient DPI based traffic classifier for classifying the network runs in a separate scan of payload [17]. The presented approach which is heuristic based had provided a sub-liner explore complication. The dataset consisted of traffic from ten diverse applications to perform the experiments. JnetPcap library. The experimental results indicated that the formulated approach provided higher accuracy. The future work would aim at implementing the formulated algorithm in C and the library LibPcap and comparing it with other works.

GuangluWei, et.al(2020) recommended a DL (deep learning) model on the basis of convolutional neural network for the current complex network environment for the clasifcation of the the traffic [18]. A conversion load of network traffic get done into a 2-d (two-dimensional) gray image and the generated image was employed as the input of the model. The recommended model was capable of classifying the network traffic and learning the relevant attributes from traffic data in automatic manner. This algorithm assisted the researchers in classifying the network traffic and yielded higher accuracy in comparison with conventional techniques.

Fakhroddin Noorbehbahani,et.al(2018) investigated the novel semisupervised method and Xmeans clustering,label propagation algorithms for classification of the traffic [19]. This technique had provided accuracy of the label propagation 95% above. The dataset having 20% labeled data was employed to implement the Naïve Bayes and J48 DT.Meanwhile the evaluation results indicated that the investigated technique provided better accuracies using these algorithms whose training was done on datasets.

Xinxin Tong,et.al(2020) suggested an innovative classification called Bidirectional Flow Sequence Network on the basis of the long short-term memory [20]. Different from the conventional classifier, the BFSN was an E2E (end-to-end) classifier assisted in learning the representative attributes from the traffic and classification. Futhermore, the bidirectionaltraffic succession developed by the usage of the length and direction knowledge of encrypted traffic,processed this algorithm on the basis of LSTM. The ISCX VPN dataset was utilized for conducting experiments.And the experimental output depicted the suggested classifier yielded accuracy up to 91%.

RESEARCH METHODOLOGY

The research work performed on the foundation of classifying the traffic. It is classified in distinct phases including pre-processing, feature extraction, meanwhile theclassification and the analysis of the performance. These phases are:-

Step 1: Dataset Input and Pre-processing:-The real source of KDD (Knowledge Discovery in Databases) dataset is engaged for collecting the dataset. For the scrutinization of the KDD dataset, the train set contains 78% records in train set and 75% in test set. The studying algorithm is partial on the way to the records as enormous count of inefficacious records in the train set. The techniques have increased detection rates on the continous records assist in attaining biased outcomes considering the records in the test set. Moreover, this work does the execution 21 study machines for which the complexity level of records are analyzed in KDD for assigning labels to the records of entire training and testing sets of KDD (Knowledge Discovery in Databases). Hence, 21 declaration labels are supplied for every record. The KDD dataset has unstable form for preprocessing the dataa so that the data is cleaned. Meanwhile, under sampling is utilized for cleaning the input dataset. This technique is helpful to remove the inefficacious factors, to clean the set.

Step 2: Feature Extraction:- The current stage is performed for establishing the relation of individual attribute with targetset. The importance of thefeature is elaborated using the attribute facts depending on the target which the dataset has defined. This association is assisted in defining the target which has maximum reaction on target set. The features of the dataset is reduced using PCA algorithm. This algorithm is utilized to build a non-high dimension representation of the factors which defines variance in the specifics. Mathematically, following algorithm focuses on investigating a linear mapping$M$ to increase$M^T cov(X)M$ in which$cov(X)$denotes the covariance matrix X. As its displayed that the$d$principal eigen vectors regarding the covariance matrix of the mean data generates this linear mapping. Thus, the issue of eigen is resolved using Principal Components Analysis as:

$$cov(X)M = \lambda \leftthreetimes M$$

The eigen problem is tackled for the$d$principal eigenvalues $\leftthreetimes$.The low-dimensional data representations$y_i$of the datapoints $x_i$are calculated for which these values are mapped the linear basis M,i.e., $.Y = (X - \bar{X})M..$ Principal Components Analysis (PCA) is implemented thehuge count of domain to recognize the facts, classify the coin and analyze the seismic series. The major limitation of this algorithm is the proportionality of the size regarding matrix to data point.

Step 3: Classification:-The major aim of the technique is the classification of the entire set in two class known as train set and test set. The train set has utilized 70% of the entire dataset and rest of the data is used in the test set. The voting classification is put forward for performing the final prediction. The voting classification algorithm take input of the Random Forest,K-Nearest Neighbor. The K-Nearest Neighbor is the instance-based learning in where classification get performed on the base of diverse the similarity of the measures. The Euclideanand Mahantt both are the suitable variables. While,Minkowski function performed more efficiently in comparison with the category variables. The Euclidean distance $\left(D_{ij}\right)$ amid two input vectors $\left(X_i, X_j\right)$is expressed as:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2 k = 1,2,\dots,n}$$

The Euclidean distance is measured from an input points to current point for individual datapoint in the dataset. Categorization of these intervals is done in increasing way and the k-items has lowest metrics are chosen for the input points. The class is found from such sources and KNN is classified for returning the majority class as the classification for the input point. The predictor altogether is generated with the implementation of RF in order to develop the Decision Trees in spaces of the data. The selection of these subspaces is done at random. The RF algorithm has adopted in easy nd fast way. Predictions have been achieved with high accuracy, it handles a number of input variables. First of all, as mall collection of the coordinates are chosen at some random at all nodes for building the tree in the collection. The train set makes a utilization of feature that is utilized to evaluate the best slip in order to develop a tree. The CART method is processed for expanding the size of the tree. Individual new tree get developed for the re-sampling the train dataset. The randomization is joined by the means the baggingg. The algorithm RF is produced for which the randomized-base RTs are executed$\{r_n(x, \Theta_{\mathrm{m}}, \mathrm{D}_{\mathrm{n})}, m \geq 1\}$ together. For the purpose of a randomized variableΘ, Θ_1, Θ_2,…. is worked for displaying thei.i.d outputs. These RTs are integrated in order to estimate the aggregated regression.

$$\overline{r_n}(X, D_n) = E_{\Theta}[r_n(X, \Theta, D_n)]$$

Following to this, the E_{Θ} denotes the expectation of the concerning some of the random parameters on X and data-set D_n.Meanwhile, estimation of the sample has not containedthe dependence and$\overline{r_n}(X)$ is be swapped with$\overline{r_n}(X, D_n)$. The SVM is the most common learning technique utilizes to identify the statistic pattern with applicance of the number of issues related to engineering. The fundamental intend is to separate two classes using a hyperplane which is denoted with the help of the simple vector (w) , the bias term (b). And length between the hyperplane and near points of classes is increased by optimal separating hyperplane. Kernel functions are often carried out with Support Vector Machine classifier that facilitated the non-linear decision boundaries. This demonstrated that a kernel function k causes the nonlinearity in the classification. The quantification of model is described and more accurate decision functions are facilitated using it. The formulation is defined as:

$$w.\Phi(x) + b = 0,$$

Using which the corresponding decision function is obtained that is expressed as:

$$f(x) = y^* = sgn(\langle w.\Phi(x)\rangle + b)$$

In which $y^* = +1$ if x belongs to the corresponding class otherwise$y^* = -1$.

Ahead of the application of kernel schemes, another generalization is suggested in which hard margins are replaced through soft margins with the help of the so-called slack-variablesζ_i, so that the inseparability is facilitated, the constraints are relaxed and the noisy data is handled. In addition, even though the original SVM paradigm was suggested for issues of binary classification, it is reformulated for addressing the multiclass problems for which the data is divided.

Step 4: Analysis of performance:- Quantification of performance of the current system performed on base of recall, precision and the accuracy. Various terms are used for scrutinizing the efficacious algorithms.

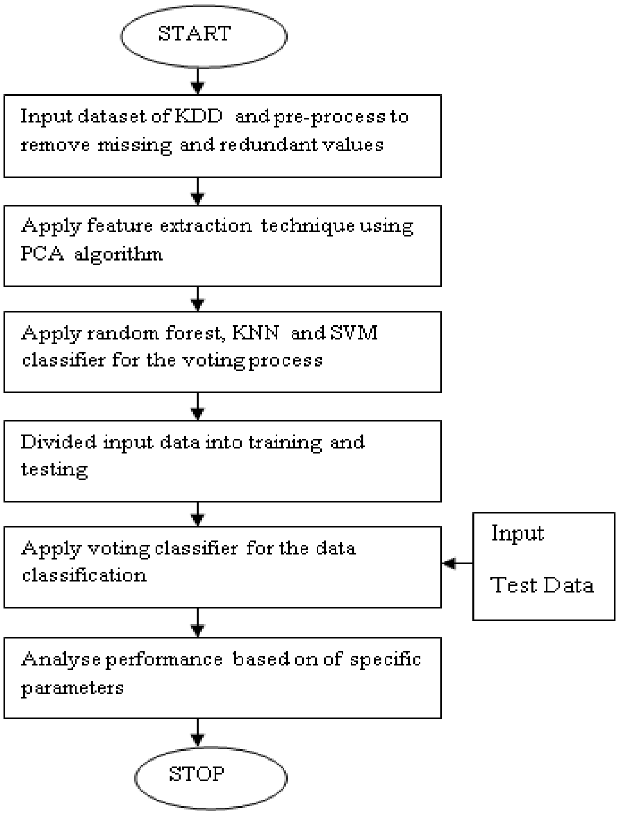

START
Input dataset of KDD and pre-process to remove missing and redundant values
Apply feature extraction technique using PCA algorithm
Apply random forest, KNN and SVM classifier for the voting process
Divided input data into training and testing
Apply voting classifier for the data classification
Input
Test Data
Analyse performance based on of specific parameters
STOP

Figure 2 The Proposed work Flowchart

RESULT AND DISCUSSION

The work classified the traffic by the implementation of the two models. The foremost is the SVM(Support Vector Machine) classification while the second is an ensembled classifier model. The ensembled model is generated with the integration of the KNN(K-Nearest Neighbor) and RF(Random Forest) classification models altogether. Different metric instance as precision, recall, accuracy considered the computation the performing of both of these models. All these metrics are described as below:

1. Precision: The quantity of data which get obtained from a count that is depending on the statistics. Closeness of dimensions to each other respectively is presented with the precision. It's don't relied on the accuracy-data. In case, the count of TPs divide by the no. of TPs(true positives) plus the number of FPs, the precision is obtained. False positives are cases which are labelled using the model non-correct as positive however its negative.

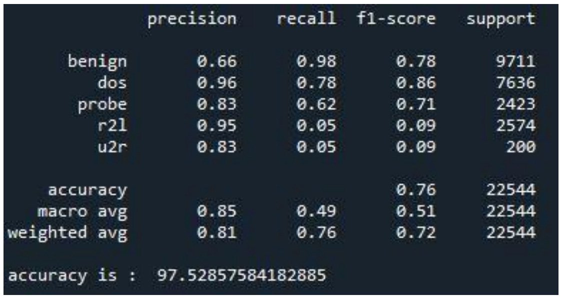
Precision= (True Positive) / (True Positive + False Positive)

2. Recall: Itis defines as the proportion of the whole quantity of draw out models. It is the part of count of TPs and the count of TPs and the number ofFPs (false positives). TPs(true positives) are data point. Where the classification is performed with the model as positive that true are positive and false negatives are points which are indetified as negative hoever, are positive.

Recall= (True Positive) / (True Positive + False Negative)

3. Accuracy: It can be regarded as the proportion of number of points which are correct after classification and the total count of points and multiplication with 100.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$



```
              precision    recall  f1-score   support

      benign       0.66      0.98      0.78      9711
         dos       0.96      0.78      0.86      7636
       probe       0.83      0.62      0.71      2423
         r2l       0.95      0.05      0.09      2574
         u2r       0.83      0.05      0.09       200

    accuracy                           0.76     22544
   macro avg       0.85      0.49      0.51     22544
weighted avg       0.81      0.76      0.72     22544

accuracy is :  97.52857584182885
```

Figure 3 Voting Classifier- performance

The figure 3 displays the processing of voting classification where the KNN is combined with RF and SVM. This algorithm possess the accuracy 97.52%.

Table1 Results Scrutinization

Parameters	SVM Classifier	Voting Classifier
Accuracy	75.51 %	97.52 %
Precision	81.45 %	81.67 %
Recall	75 %	76 %

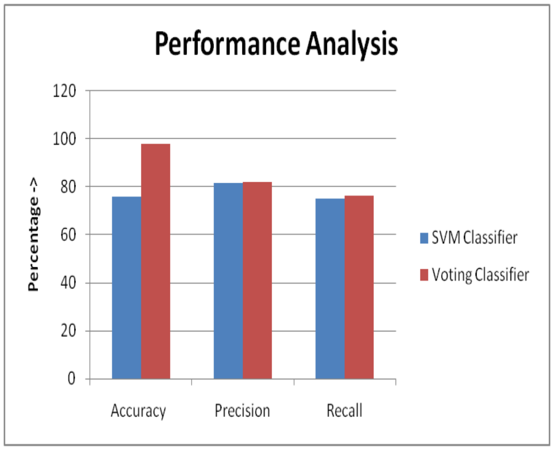


Figure4 Performance Scrutzation.

The Figure 4 reveals an analysis regarding performance Scrutinization of SVMClassifier and Voting Classifierand their comparison in order to classify the traffic.Voting classifier attains better accuracy, precisionand recall values comparision to SVM for classifying the network traffic.

CONCLUSION

The network traffic classification techniques posses the three phases namely port-based,payload-based andflow statistics-based. The process to classify the network traffic is deal with recognizing distinct types of applications or traffic data-part for which the obtained data packets scrutinized that is crucial in the transmission of networks of real global world. The traditional port-based course of action based on contributing the standard ports that the famed functions deploy. The traffic can be categorized in several stages phases instance as pre-processing, to draw out the attributes and to succeed in doing the classification. This research work makes the usage of voting classification algorithm in way to classify the traffic. The suggested algorithm supply the greater precision,accuracy and recall in contrast to the existing SVM (Support Vector Machine) algorithm.

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

[12] Won-Ju Eom, Yeong-Jun Song, Chang-Hoon Park,

Jeong-Keun Kim, Geon-Hwan Kim, You-Ze Cho, "Network Traffic Classification Using Ensemble Learning in Software-Defined Networks", 2021, International Conference on Artificial Intelligence in Information and Communication (ICAIIC)

[13] Madhusoodhana Chari S., Srinidhi H., Tamil Esai Somu, "Network Traffic Classification by Packet Length Signature Extraction", 2019, IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)

[14] 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)

[15] Jiwon Yang, Jargalsaikhan Narantuya, Hyuk Lim, "Bayesian Neural Network Based Encrypted Traffic Classification using Initial Handshake Packets", 2019, 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks – Supplemental Volume (DSN-S)

[16] 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

[17] Pratibha Khandait, Neminath Hubballi, Bodhisatwa Mazumdar, "Efficient Keyword Matching for Deep Packet Inspection based Network Traffic Classification", 2020, International Conference on COMmunication Systems & NETworkS (COMSNETS)

[18] Guanglu Wei, "Deep Learning Model under Complex Network and its Application in Traffic Detection and Analysis", 2020, IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)

[19] Fakhroddin Noorbehbahani, Sadeq Mansoori, "A New Semi-Supervised Method for Network Traffic Classification Based on X-Means Clustering and Label Propagation", 2018, 8th International Conference on Computer and Knowledge Engineering (ICCKE)

[20] Xinxin Tong, Xiaobin Tan, Lingan Chen, Jian Yang, Quan Zheng, "BFSN: A Novel Method of Encrypted Traffic Classification Based on Bidirectional Flow Sequence Network", 2020, 3rd International Conference on Hot Information-Centric Networking (HotICN)