Voting Classification Method for Network Traffic Classification Using SVM,KNN and Random Forest

Gunneet Kaur[1], Maninder Pal Singh[2]
[1]*Research Scholar, LCET Katani Kalan*
[2]*Head of Department CSE, LCET Katani Kalan*

ABSTRACT - The network traffic classification task is focused on recognizing diverse kinds of applications or traffic data for which the received data packets are analyzed that is essential in communication networks in these days. The network traffic can be classified in several stages, in which pre-processing is done, attributes are extracted and classification is performed. The processing of dataset is carried out as it is taken as input in the process of classification. The dataset is split into two sets: training and testing. The training set includes 60% of the entire dataset and testing set has 40%. The voting classification method is implemented in which the KNN (K-Nearest Neighbour) is integrated with the RF (random forest) and SVM. The suggested method is deployed in python and several parameters such as accuracy, precision and recall are taken in account for quantifying the results. This indicates that the suggested method yields higher accuracy, precision and recall in comparison with the traditional classification models.

KEYWORDS - Network Traffic , Random Forest, Support Vector Machine, K-Nearest neighbour, Voting Classifier

I. 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 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[3]. This step-by-step process flow shows how network traffic classification methods identify/classify unknown forms of network traffic using machine learning algorithms.

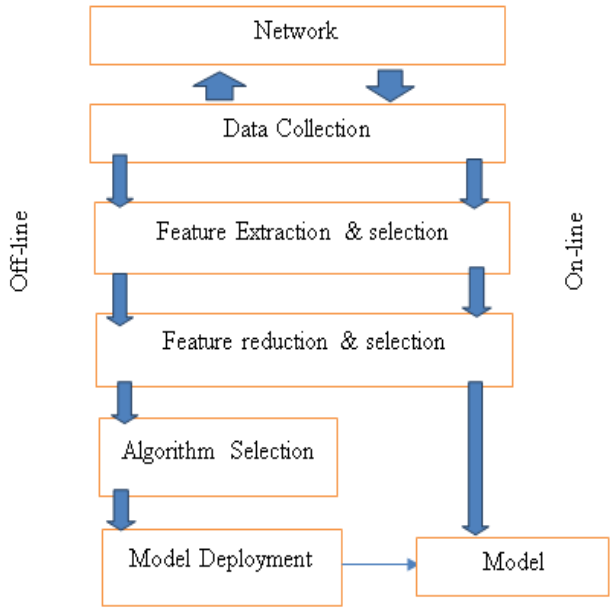

Figure 2:Network Traffic Classification Model

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.

II. LITERATURE SURVEY

Hassan Alizadeh, et.al (2020) suggested an innovative technique in order to classify the network traffic with the implemented GMM (Gaussian Mixture Model) [11]. The CEM (component-wise expectation-maximization) was exploited for creating a separate GMM so that the network traffic distribution was matched. 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. The experimental outcomes achieved on the real-world network traffic dataset validated that the superiority of introduced model over other classification algorithms. Moreover, the suggested model performed more effectively in classifying the network traffic.

Madhusoodhana Chari S., et.al (2019) intended a packet length signature extraction based method for classifying distinct classes of traffic including Audio streaming, Video streaming, 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. The run length of the packet length sequence was observed for every class of traffic to generate the intended feature set. This set had provided a tree which was found more balanced and capable of producing the lower number of rules for each class. This set provided interpretability to the intended method and easily deployment in a real time scenario with least resource requirements.

Jing Ran, et.al (2018) developed three-dimensional CNN (3D convolutional neural network) system in order to classify the network traffic [14]. This system was assisted in extracting the spatial and temporal attributes in automatic manner and determining the most appropriate attributes subsequent to the iterations of validation and classified the traffic. 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 end hosts to establish the transport layer security. The BNN (Bayesian neural network) was employed as the classification technique in which the cipher suite, compression technique and extension information related to the handshake packets was considered as the inputs. The experiments 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 TDM-EPON (time-division multiplexing Ethernet passive optical network) framework [16]. The 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 (signal-to-noise ratio) and kept the E2E (end-to-end) delay under 100 μs.

Pratibha Khandait, et.al (2020) formulated an efficient DPI based traffic classifier for classifying the network flows in a single scan of the payload [17]. The presented heuristic based approach had provided a sub-liner search complexity. A dataset consisted of traffic from ten diverse applications to perform the experiments. JnetPcap library, a wrapper function was utilized over LibPcap library to deploy this algorithm. The experimental results indicated that the formulated approach provided higher accuracy. The future work would aim at implementing the formulated algorithm in C using LibPcap library and comparing it with other works.

Guanglu Wei, et.al (2020) recommended a DL (deep learning) model on the basis of CNN (convolutional neural network) for the current complex network environment for classifying the network traffic [18]. The conversion of load of network traffic was done into a 2-d (two-dimensional) gray image and the generated image was employed as the input of the model. The payload data of network traffic was a continuous 1-D (one-dimensional) byte stream organized in a hierarchical structure. 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 a novel semi-supervised technique and x-means clustering and label propagation algorithms for classifying 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 NB (Naïve Bayes) and J48 DT (decision tree). 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 classifier called BFSN (Bidirectional Flow Sequence Network) on the basis of LSTM (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 raw traffic and classifying them. In addition, the bidirectional traffic sequence was developed using the length and direction information of the encrypted traffic and processed this algorithm on the basis of LSTM. The ISCX VPN-NonVPN dataset was utilized to conduct the experiments. The experimental outcomes depicted that the suggested classifier yielded the accuracy up to 91%.

III. RESEARCH METHODOLOGY

This research work is carried out on the basis of classifying the network traffic. The network traffic is classified in distinct phases including pre-processing, feature extraction, classification and performance analysis. These phases are defined as:-

Step 1: **Dataset Input and Pre-processing:-**The authentic source of KDD (Knowledge Discovery in Databases) is employed for collecting the dataset. For the analysis of KDD, the training set contains 78% of forged records in training set and 75% of records in the testing set. The learning algorithm is partial towards the most common records as enormous numbers of inefficient records in the training set. Thus, the infrequent and risky records must be prevented. These records cause harm to the networks. The methods have enhanced detection rates on regular records assist in attaining biased outcomes considering the repeated records in the test set. Also, this work makes the execution of 21 learned 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 predicted labels are provided for every record. The KDD dataset has imbalanced form for pre-processing the data so that the data is cleaned. The under sampling is utilized for

cleaning the input dataset. This technique is helpful to remove the inefficient values and to clean the dataset.

Step 2: **Feature Extraction:-**This stage is executed for establishing the relation of each attribute with target set. The significance of feature is defined using the attribute values depending on the target which the dataset has defined. This association is assisted in defining the target which has maximum affect on target set. The features of the dataset is reduced using PCA algorithm. This algorithm is utilized to build a low-dimensional representation of the data which defines as much of the variance in the data as possible. Mathematically, this algorithm focuses on investigating a linear mapping M to increase $M^T cov(X)M$ in which$cov(X)$denotes the covariance matrix of the data X. It is demonstrated that the d principal eigenvectors of the covariance matrix of the zero-mean data generates this linear mapping. Thus, the issue of eigen is resolved using Principal Components Analysis as:

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

The eigen problem is tackled for the$d$principal eigenvalues $\curlywedge$ .The low-dimensional data representations y_i of the datapoints $x_i$are calculated for which these values are mapped onto the linear basis M , i.e., $Y = (X - \bar{X})M$. Principal Components Analysis (PCA) is implemented in a huge number of domains to recognize the face, classify the coin and analyze the seismic series. The major limitation of this algorithm is the proportionality of the size of the covariance matrix to the dimensionality of the data points.

Step 3: **Classification**:-The major goal of this technique is to classify the entire dataset in two classes known as training and testing. The training set has utilized 60% of the entire dataset and rest of the data is utilized in the testing set. The voting classification is put forward for performing the final prediction. The voting classification algorithm will take input of Random Forest and K-Nearest Neighbor for this purpose. The K-Nearest Neighbor is an instance-based learning in which classification is performed on the basis of diverse the similarity measures. The Euclidean and Mahantt are suitable continuous variables on the other hand, Minkowski distance function performed more efficiently in comparison with the categorical variables. The Euclidean distance (D_{ij}) amid two input vectors $(X_i, X_j)$is expressed as:

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

The Euclidean distance is quantified from an input data point to current point for each data point in the dataset. The sorting of these distances is done in increasing order and the k items that has lowest distances are chosen for the input data point. The majority class is found from these items and KNN is classified for returning this majority class as the classification for the input point. A predictor ensemble is developed with the implementation of RF in order to develop the Decision Trees in subspaces of data. The selection of these subspaces is done at random. The Random Forest algorithm is adopted in easy and fast way. The predictions are achieved with higher accuracy and it handles a number of input variables. First of all, as mall group of input coordinates are chosen at random at every node for building a tree in the collection. The training set makes the utilization of features that is utilized to evaluate the best slip in order to develop a tree. The CART technique is implemented for expanding the tree's size. Each new individual tree is developed for re-sampling the training data set every time. The subspace randomization is merged by the means the bagging. A RF is generated for which the randomized base RTs are executed $\{r_n(x, \Theta_m, D_{n)}, m \geq 1\}$ together. In this, for a randomized variable Θ, $\Theta_1, \Theta_2, \dots$. is employed for displaying thei.i.d outputs. These RTs are integrated in order to estimate the aggregated regression.

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

In this, E_Θ denotes the expectation concerning random parameter on X and data set D_n. The estimation in the sample has not containedthe dependency and $\bar{r_n}(X)$ is be replaced with $\bar{r_n}(X, D_n)$. The SVM is the most common learning technique utilizes to identify the statistical pattern with applications in a number of issues related to engineering. The fundamental intend is to separate two classes using a hyperplane which is denoted with the help of its normal vector w and a bias term b. The distance between the hyperplane and the nearest points of both classes is increased by the optimal separating hyperplane. Kernel functions are often carried out in conjunction with the 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-variables ζ_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.

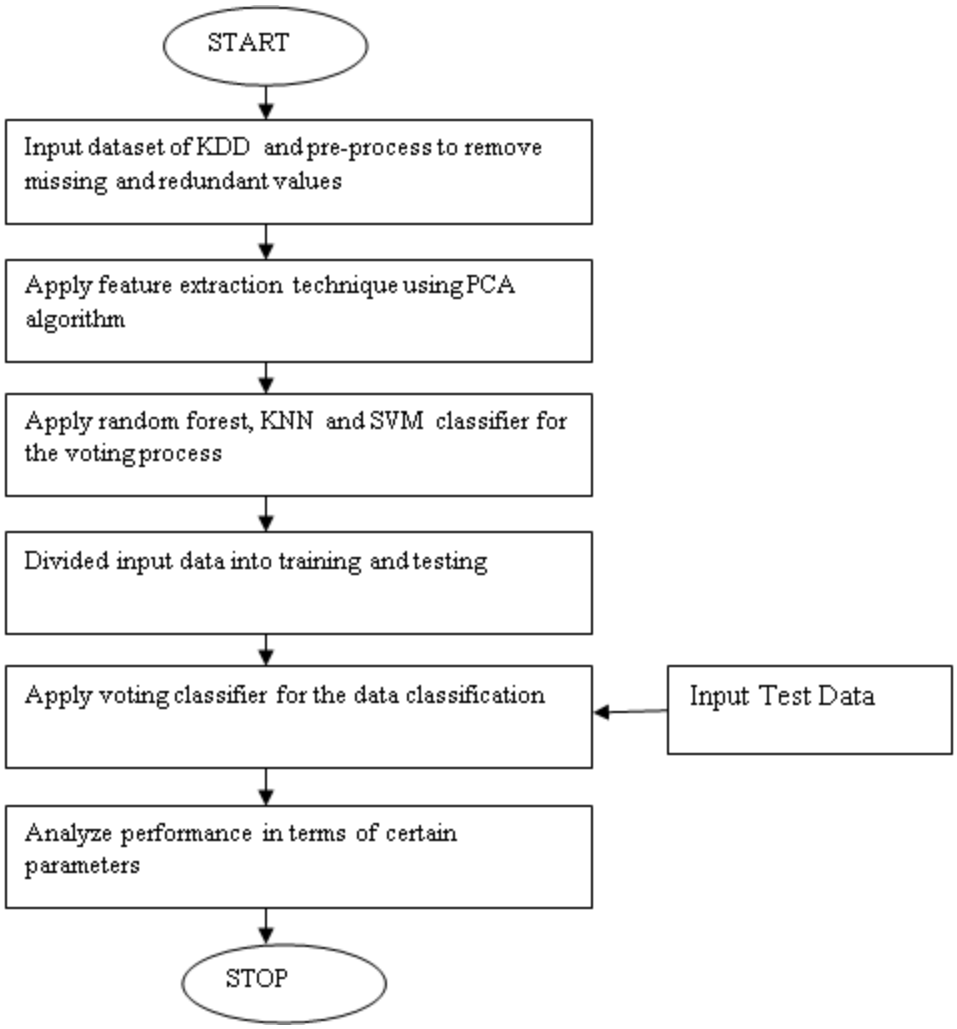

Figure 2: Proposed Flowchart

Step 4: **Performance Analysis:-**The quantification of the performance of the introduced system is performed concerning the accuracy, precision and recall. Various metrics are used for analyzing the efficacy of these algorithms.

IV. RESULT AND DISCUSSION

This work classifies network traffic by implementing two models. The first is the SVM (Support Vector Machine) classifier while second is an ensemble classifier model. The ensemble model is generated with the integration of the KNN (K-Nearest Neighbor) and RF (Random Forest) classifier models together. Different metrics such as precision, recall and accuracy are considered to compute the performance of both of these models. All these metrics are described as below:

1. Precision: It is the amount of information which is obtained from a number that is depending on its digits. The closeness of two or more measurements to each other is represented with precision. It is not relied on the accuracy. In case, the number of TPs divided by the number of TPs (true positives) plus the

number of FPs, the precision is obtained. False positives are cases which are labelled using the model incorrectly as positive but they are negative in reality.

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

2. Recall: Itis defines as the ratio of the total amount of extracted pertinent models. It is the fraction of number of TPs and the number of TPs plus the number ofFPs (false positives). TPs(true positives) are data point whose classification is done with the model as positive that actually are positive and false negatives are data points which are recognized as negative but, in fact, theyare positive.

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

3. Accuracy: It can be defined as the ratio of number of points which are classified correctly and the total number of points and its multiplication is done with 100.

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

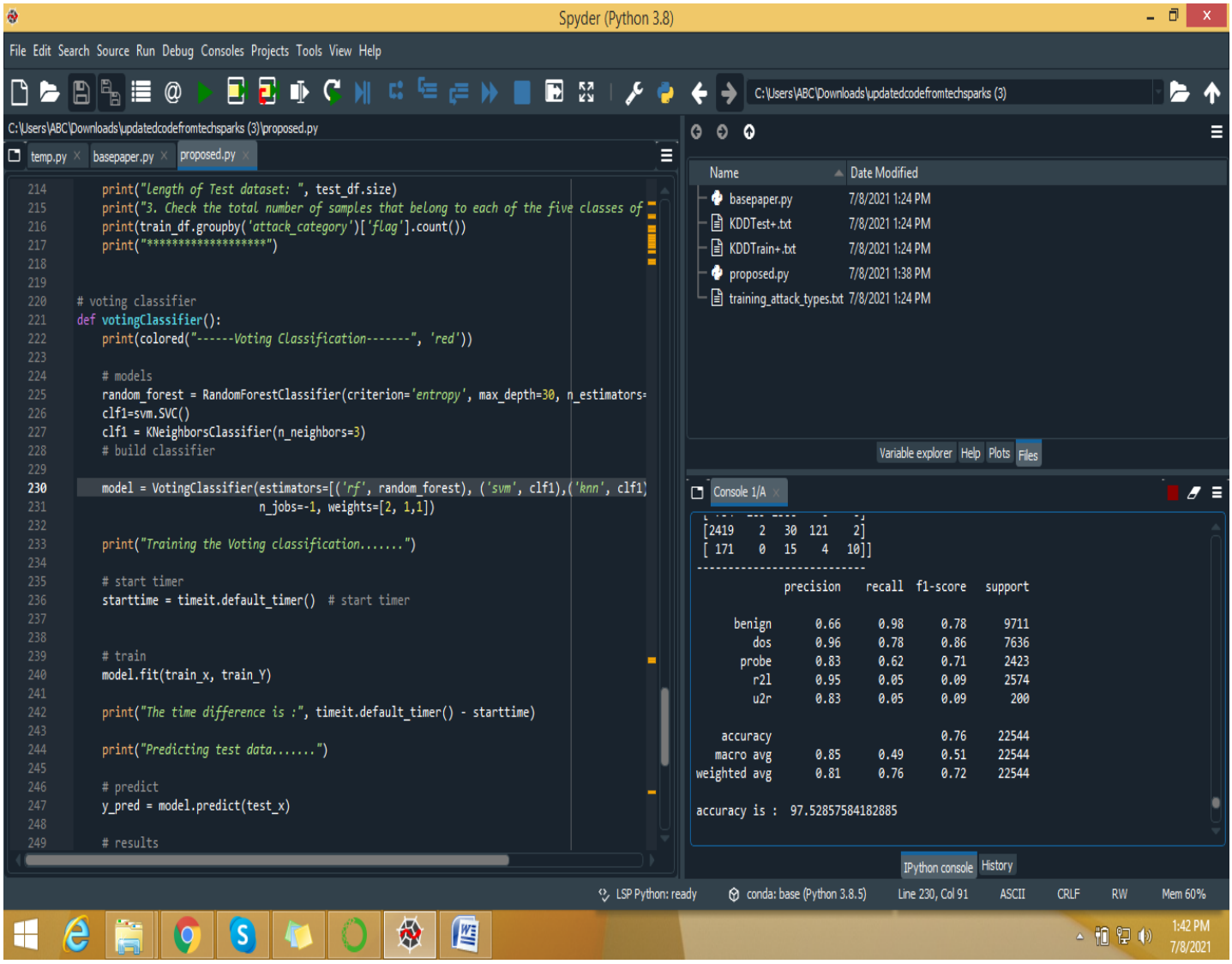

Figure 2 Performance of Voting Classifier.

The figure 2 denotes the implementation of voting classifier in which KNN is integrated with RF. This algorithm yields the accuracy around 97.52%.

Table 1: Results Analysis

Parameters	SVM Classifier	Voting Classifier
Accuracy	75.74 percent	97.52 percent
Precision	81.45 percent	81.67 percent
Recall	75 percent	76 percent

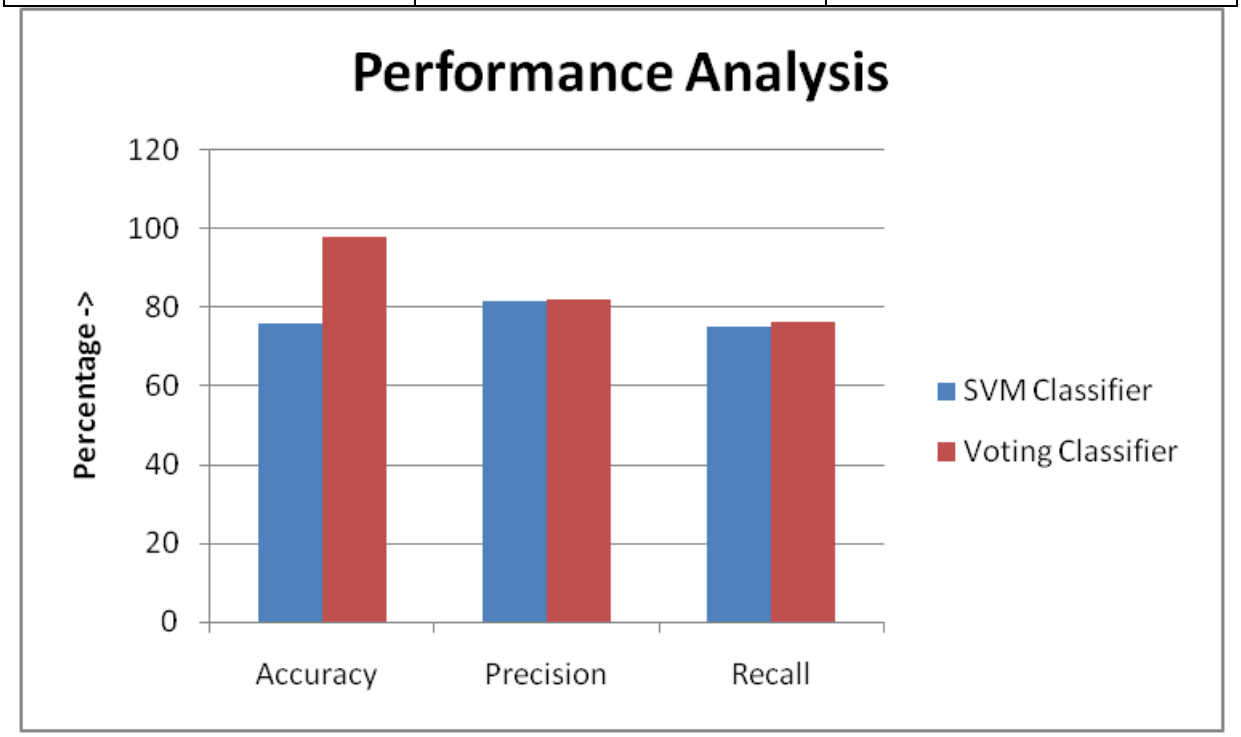

Figure 4.6. Performance Analysis

The figure 4.6 reveals the analysis of performance of Support Vector Machine and Voting Classifier and their comparison in order to classify the network traffic. The voting classifier attains better accuracy, precision and recall values as compared to SVM for classifying the network traffic.

V. CONCLUSION

The network traffic classification techniques have 3 categories namely port-based, payload-based and flow statistics-based. The process to classify the network traffic is deal with recognizing distinct kinds of applications or traffic data for which the received data packets are investigated that is essential in the communication networks of real world. The conventional port-based scheme based on investing the standard ports that the renowned applications deploy. The network traffic can be classified in different stages phases such as pre-processing, for extracting the attributes and accomplishing the classification. This research work makes the utilization of voting classification algorithm in order to classify the network traffic. The suggested algorithm provides greater accuracy, precision and recall in contrast to the existing SVM (Support Vector Machine) algorithm.

VI. 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)