iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55239

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Voting Classification Model for Network Traffic Classification

Jasmeet Kour[1], Prof. Lalit Sen Sharma[2]
Department of CS&IT, University of Jammu

Abstract: ***Network traffic classification 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. The network traffic classification has various phases which include pre-processing, feature extraction and classification. In the previous year's various machine techniques is designed for network traffic classification. The techniques which are already designed give low accuracy. In this research work, voting classification method is designed for network traffic classification which give high accuracy. The proposed model is implemented in python and results is analyzed in terms of accuracy, precision and recall.***

Keywords: INetwork Traffic, Classification, Machine learning, Voting.

.

I. INTRODUCTION

Network Traffic Classification is a significant point these days in a technological sector. It is fundamental for ISPs to deal with the general efficiency of internet. Traffic classification is the initial step to recognize and classify obscure network classes. Through this method, internet administrators can make a few moves, for example, to obstruct a few streams and oversee assets. Deficient traffic solution is the element that sways the efficiency of plans, for example, QoS and DAC, etc, delivering foundation scaling issues. In this manner, NTC is a critical instrument to tackle the traffic flow by giving the information to deciding the degrees of execution that are requested by applications [1]. To classify the traffic across internet is a fundamental component for framework management and to guarantee the Quality of Sensor of the various mechanisms. In all actuality, an exact traffic classification system permits the productive administration of existing network assets, consequently allowing more precise and hearty asset allotment plans. NTC joins the traffic flow with a created mechanism and considers as a fundamental initial stage for investigation. Significant data can be accumulated from traffic examination, particularly for security, for example, sifting traffic and distinguishing and recognizing malevolent action. Understanding the kind of utilization is streaming across network, its administrators can respond rapidly to potential episodes' dependent on their occurrence reaction plans. In the course of the most recent decade, traffic classification has been given a ton of consideration from both industry and the scholarly community. Network traffic classification is a significant issue of network resource management that emerges from analysing network patterns and network arranging and planning [2]. Ways to deal with network traffic classification differ as indicated by the properties of the packets utilized.

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:

1) 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.

©IJRASET: All Rights are Reserved

2) 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 business and openly available tools, such as for implementing Linux piece firewall [6].
3) 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.
4) 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.

II. LITERATURE REVIEW

Zhiyong Bu, et.al (2020) suggested an NN (neural network) with deep and parallel NIN (network-in-network) structures in order to classify the network traffic [21]. A global average pooling was employed in NIN prior to accomplish a final classification for mitigating the number of model parameters in efficient manner. The fixed-length packet vectors were mapped towards application or traffic labels by developing a deep NIN system with MLP (Multilayer perceptron) convolutional layers. The experimental outcomes attained on traffic dataset depicted that the suggested approach had potential for balancing the classification accuracy and complexity as compared to traditional models. In addition, the F1 score was computed0.983 for classifying the traffic and0.985 for recognizing the application.

Madhusoodhana Chari S., et.al (2019) intended a technique based on packet length signature extraction with the objective of classifying various classes of traffic [22]. A new feature set was introduced for training a J48 DT (decision tree) algorithm so that the classes of network traffic were recognized and the interpretability of the system was also described. The tree which was created through introduced set had provided more balance and assisted in producing the least number of rules for every class. The presented approach offered interpretability and easy deployment in a real time scenario using lesser resource requirements.

Jing Ran, et.al (2018) formulated a system on the basis of three-dimensional CNN to classify the network traffic [23]. The spatial and temporal attributes are extracted and the appropriate attributes were investigated when the iterations of validation was accomplished. A publicly available dataset named USTC-TFC2016 was employed to conduct a number of experiments. The proposed system provided more accuracy in contrast to other algorithms and efficiently classified the traffic. The next step would emphasize on integrating the feature extractor and traditional algorithms.

Jiwon Yang, et.al (2019) introduced a new payload-based classifier using which unencrypted handshake packets were utilized to exchange among the end hosts to establish transport layer security [24]. The BNN (Bayesian neural network) was presented as the classification algorithm in which cipher suite, compression method and transport layer security were considered. The TLS extension information of the handshake packets were as the inputs. The experimental results demonstrated that the introduced classifier was performed more effectively than other models. The future work would concentrate on extending the introduced system to classify other secure protocols.

Pratibha Khandait, et.al (2020) introduced a system so that the internet traffic was classified with single scan of flow payloads for which a heuristic technique was implemented for obtaining a sub-linear search complexity [25]. This system focused on scanning some primary bytes of payload and determining the potential of application signature to match the succeeding signature. A dataset, in which 171873 network flows were included, utilized to carry out the experiments. The recommended system provided 98% accuracy while classifying the traffic.

Hyun-Kyo Lim, et.al (2019) discussed that a technology was required in the network management for classifying the network traffic in which the network operator was not interfered [26]. The network traffic was pre-trained for developing a dataset based on packet. The CNN (convolutional neural network) and ResNet (residual network) were deployed for training 5 DL (deep learning) systems so that the network traffic was classified.

In the end, the classification performance of packet-based datasets was quantified on the basis of f1 score of both the algorithms. The outcomes revealed that the presented approach was efficient.

Yu Zhang, et.al (2019) developed a new system named STNN (Stereo Transform Neural Network) for classifying the encrypted network traffic [27]. The LSTM (Long Short-Term Memory) was integrated with the CNN (convolutional neural network) on the basis of statistical attributes in the developed system. The convergence rate was enhanced, and the accuracy was increased using this system. The results obtained in experimentation indicated that the developed system was efficient for all the target applications. Furthermore, the average precision was computed 95%, recall was95% and its accuracy was 99.5%.

Xiao Wang, et.al (2019) projected a technique for optimizing the CNN (convolutional neural network) model in parallel on Spark platform [28]. The Spark Streaming model was utilized for deploying the requirements so that the network traffic was categorized in real-time. Moreover, the experiments were performed to compute the projected technique. The outcomes validated that the projected technique performed effectively in real-time and obtained higher accuracy for classifying the traffic. The task of classifying the network traffic was done in real time.

Ibraheem Saleh, et.al (2020) established a 2-D model of a stream of packet header lengths in which DCNN (deep convolutional neural network) was implemented in order to classify the network traffic [29]. An effective technique was put forward for so as 1-D packet-subflow was converted into a 2D image by planning 5 diverse network traffic image orientation mappings. The experiments were conducted on two diverse mapping strategies for mapping the packets related to a packet flow to the pixels in the image. The experimental results proved that the established system provided higher accuracy with least manual effort for which the traffic images were utilized in DL.

Rui Li, et.al (2018) examined a RNN to classify the internet traffic and an innovative algorithm named BSNN (Byte Segment Neural Network) was designed [30]. At first, the division of datagram was done into various byte segments. Thereafter, the RNN based encoders had employed these segments. The encoders were employed to extract the information which was later put together with a representation vector of the entire datagram. At last, the SoftMax function was executed for applying this vector so that the application protocol of this datagram was predicted. The data collected in real time with various protocols was utilized to conduct the experimentations. The results demonstrated that the introduced algorithm outperformed the conventional ML (machine learning) technique and attained 95.82% F1-measure.

III. RESEARCH METHODOLOGY

Classifying flowing network traffic is the main objective of this work. The network traffic classification has various steps which include data pre-processing, feature extraction and classification. The research methodology is explained below: -

A. Data set input and Pre-processing

The dataset input is the first step in which data is taken as input from the authentic source which is KDD. The NSL-KDD dataset which includes 42 attributes is used in this study. Improvements are made in the KDD'99 dataset by removing the duplicate instances so that the biased classification results can be eliminated from the dataset. Only 20% of the training data is used even though there exist different versions of the data set. The representation of this data is done in the form of KDDTrain+_20Percent. This dataset also includes around 22544 instances. Different configurations of this dataset are included with the variation in number of instances. However, it also includes 42 numbers of attributes. The attribute which labels 42 in the dataset such that it is possible can differentiate if the given instance includes a normal connection instance or an attack. In the phase of pre-processing missing and redundant values are removed from the dataset for the further processing. To remove missing and redundant values from the dataset mean of the whole dataset is taken and missing values will be removed with the mean value. This process leads to clean of the dataset which give high performance for the classification.

B. Feature Extraction

The feature extraction is the second step in which relationship is established between each attribute with the target set. In case when a sample is actually normal but classified as intrusion, the situation is called false positive. In case when a sample is classified as normal when it is actually an intrusion is called false negative. The earlier case does not detect the intrusion which means that this false negative scenario is bad. The layered approach uses most of the IDSs. This states that another layer might detect the intrusion if one is not detecting it. Also, a completely different working of layered approach can be seen. As many anomalies possible can be detected perhaps by the initial layer and then the data for which anomalies have been identified can be passed on to the other layers.

C. Classification

The voting classification method is applied for the network traffic classification. The voting classification is the combination of random forest, KNN and logistic regression. Combining the outputs of multiple predictors is in many cases of interest a successful strategy to improve the capabilities of artificial intelligence systems, ranging from agent architectures, to committee learning. A common approach is to build a collection of individual subsystems and then integrate their outputs into a final decision by means of a voting process. Specifically, in the machine learning literature, there is extensive empirical evidence on the improvements in generalization capacity that can be obtained using ensembles of learners. Voting ensemble is generally used for classification problems as it allows the combination of two or more learning models trained on the whole dataset. Each model predicts an outcome for a sample data point which is considered a "vote" in favor of the class that the model has predicted. Once each model predicts the outcome, the final prediction is based on the majority vote for a specific class.

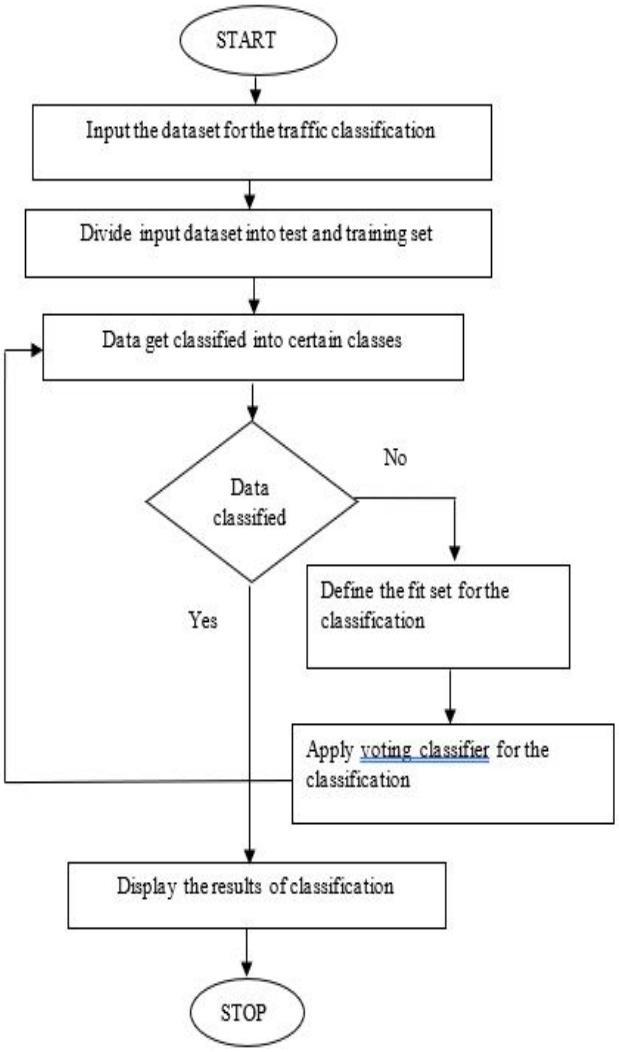


Figure 1: Proposed Methodology

IV. RESULT AND DISCUSSION

Python is an object-oriented programming language. This language is utilized to carry out simulation tests in various domains. The high-level data structures are incorporated with dynamic typing and binding for performing Rapid Application Development and linking existing components. The syntax of python is simple and can be learned easily due to which it gives more emphasis on readability. Thus, this decreases the program maintenance cost. Python helps several modules and packages and focuses on motivating the program modularity and code reuse. The python interpreter and wide-ranging standard library are offered s source or binary with no change in any important platform. Generally, python is implemented to increase productivity.

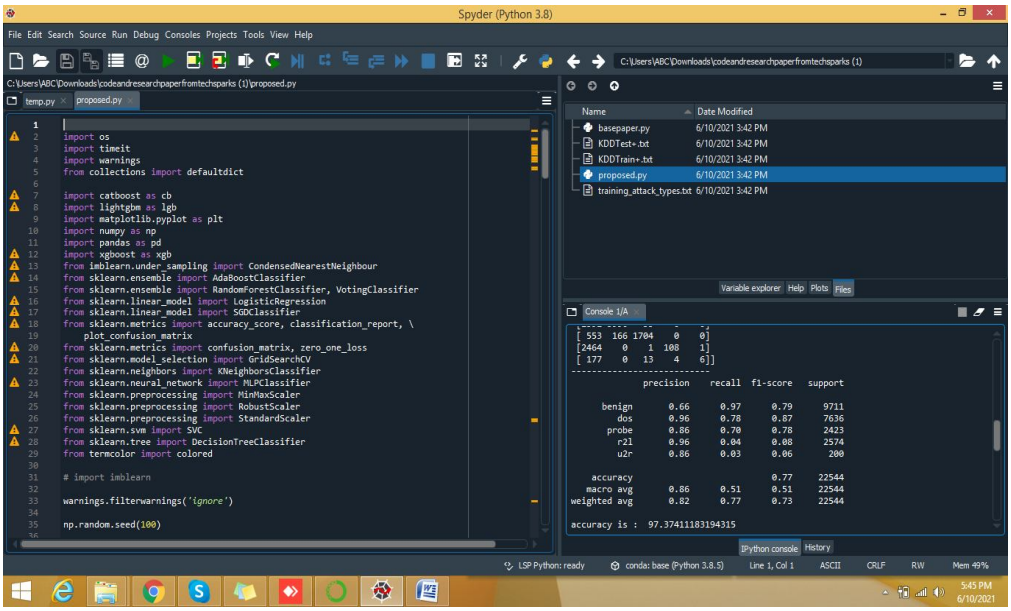


Figure 2: Proposed Results

As shown in figure 2, the proposed algorithm is implemented to classify the internet traffic. The accuracy of the model is 97 percent to classify the internet traffic using the voting classification. The voting classification is the combination of random forest, KNN and logistic regression.

Table I
Performance Analysis

Performance Parameters	SVM Classifier	Voting Classifier
Accuracy	75.74 percent	97.37 percent
Precision	81 percent	82 percent
Recall	76 percent	77 percent

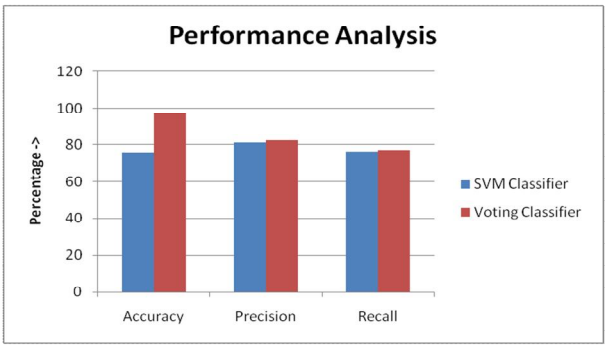


Figure 3: Result Analysis

The figure 3 depicts the comparison of the results of suggested work with the state-of-art work. The suggested approach utilizes the voting classification model in the network traffic classification. The existing work is svm for the network traffic classification.

V. CONCLUSION

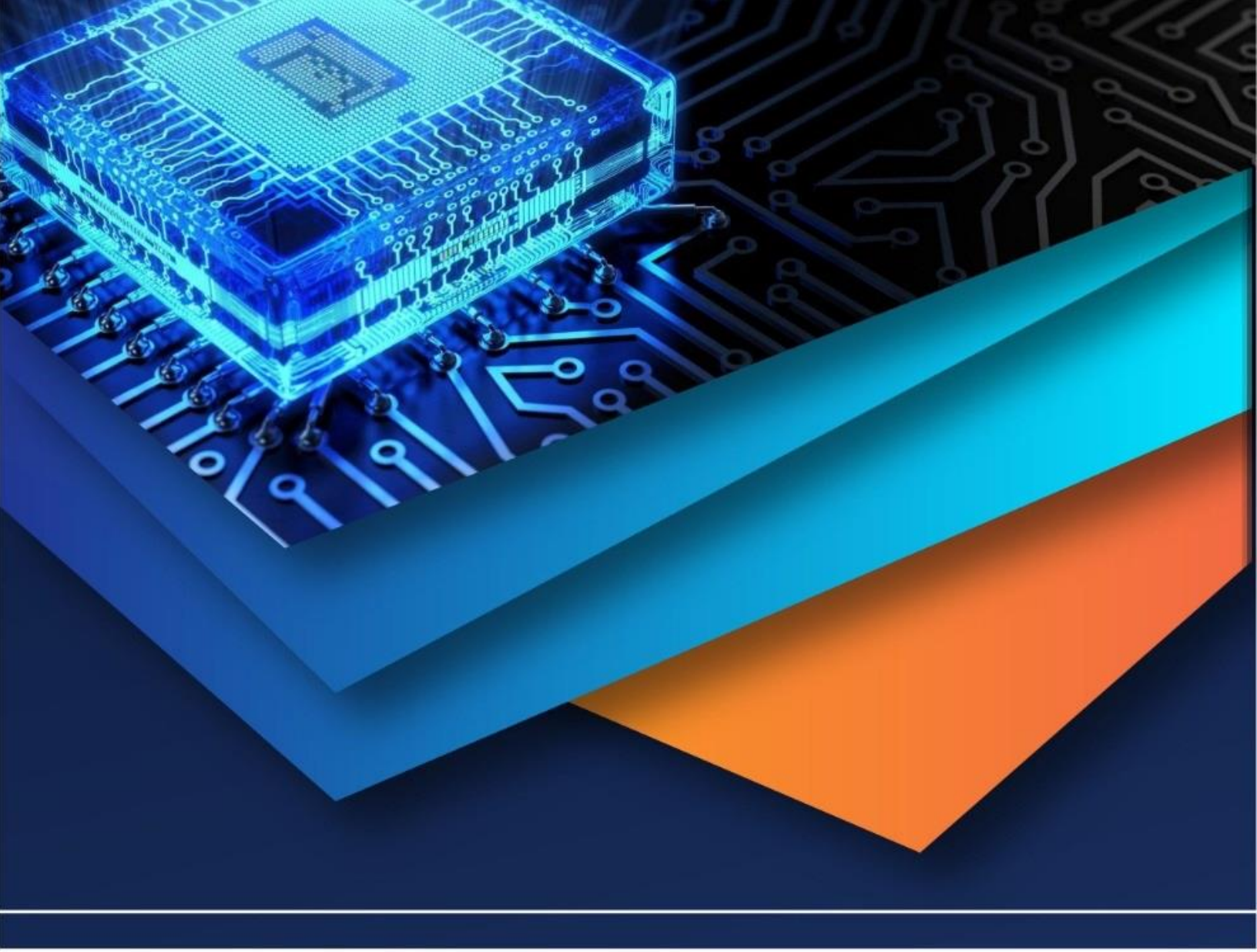
The systems which monitor the device on which they are installed are known as host-based IDSs. The states of main system through audit logs to the program execution are monitored by this approach to execute the monitoring program. The audit logs can be limited since HIDS relies on them at huge scale. the network analyst can improve the data through these activities and detecting the intrusions becomes difficult.

It is possible to perform learning from those systems and detect any new kinds of intrusions existing in them. The various intrusion detection techniques which are highly dynamic, adaptive and perform in the presence of huge network traffic have been introduced here. The voting classifier is applied in this research work for the network traffic classification. The voting classifier improve accuracy, precision and recall as compared to SVM classifier.

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] ZebaAtique 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] Zhiyong Bu, Bin Zhou, Pengyu Cheng, Kecheng Zhang, Zhen-Hua Ling, "Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models", 2020, IEEE Access
[9] 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)
[10] 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)
[11] Jiwon Yang, JargalsaikhanNarantuya, 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)
[12] Pratibha Khandait, NeminathHubballi, Bodhisatwa Mazumdar, "Efficient Keyword Matching for Deep Packet Inspection based Network Traffic Classification", 2020, International Conference on COMmunication Systems &NETworkS (COMSNETS)
[13] Hyun-Kyo Lim, Ju-Bong Kim, Joo-SeongHeo, Kwihoon Kim, Yong-Geun Hong, Youn-Hee Han, "Packet-based Network Traffic Classification Using Deep Learning", 2019, International Conference on Artificial Intelligence in Information and Communication (ICAIIC)
[14] Yu Zhang, Shiman Zhao, Jianzhong Zhang, Xiaowei Ma, Feilong Huang, "STNN: A Novel TLS/SSL Encrypted Traffic Classification System Based on Stereo Transform Neural Network", 2019, IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)
[15] Xiao Wang, Ying Liu, Wei Su, "Real-Time Classification Method of Network Traffic Based on Parallelized CNN", 2019, IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)
[16] Ibraheem Saleh, Hao Ji, "Network Traffic Images: A Deep Learning Approach to the Challenge of Internet Traffic Classification", 2020, 10th Annual Computing and Communication Workshop and Conference (CCWC)
[17] Rui Li, Xi Xiao, Shiguang Ni, Haitao Zheng, Shutao Xia, "Byte Segment Neural Network for Network Traffic Classification", 2018, IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)
[18] Akkaya, K., Senel, F., & Ulusar, U. D. (2019). A comprehensive survey on anomaly detection in IoT systems. Journal of Network and Computer Applications, 127, 48-67.
[19] Akkaya, K., Ulusar, U. D., & Şenel, F. (2020). A survey on machine learning techniques for anomaly detection in IoT systems. Journal of Network and Computer Applications, 150, 102508.
[20] Liu, F., Zhang, C., & Chen, C. (2018). Anomaly detection in IoT data using deep learning approaches. Future Generation Computer Systems, 87, 278-287.
[21] Pan, S., Chen, X., Zhu, X., & Long, G. (2020). Deep anomaly detection with outlier exposure. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), 6292-6299.
[22] Ahmad, I., Lloret, J., Cano, J. C., & Macià-Pérez, F. (2020). Machine learning-based intrusion detection system for the Internet of things in edge computing environments. Sensors, 20(16), 4497.
[23] Özdemir, S., Uluagac, A. S., & Beyah, R. (2017). A survey on anomaly detection for cyber-physical systems. ACM Computing Surveys (CSUR), 50(3), 40.
[24] Papadimitriou, S., Shilton, A., Thakker, D., Lepri, B., & Kostakos, V. (2018). Anomaly detection in IoT for urban spaces: A survey. ACM Computing Surveys (CSUR), 51(2), 1-34.
[25] Li, M., Zheng, Y., Li, S., & Li, H. (2019). Deep learning for anomaly detection: A review. Neurocomputing, 335, 98-112.
[26] M. Mallick, A. Misra, N. Ganguly, and Y. Lee, "DETECTIVES: Unified Detection & Correction of IoT Faults in Smart Homes," 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), 2020, pp. 78-87.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)