International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Statistical Analysis of Network Traffic Techniques Using ML &Deep Learning Algorithms

ShrawanKumar Pandey[1], Bhupchand Kumhar[2], Nishchay Kumar[3], Trilok Rawat[4], Devendra Singh Mohan[5]

Submitted: 27/01/2023 **Accepted:** 07/04/2023

Abstract: The article is under the relationship of network traffic and using ML with deep learning for crucial for the classification of network traffic. This can help with lawful interceptions, maintain service quality, avoid identify characters. Furthermore, it exits the organization characterization strategies, for example, port-put together recognizable proof and those based with respect to profound bundle examination, factual elements related to AI, and profound learning calculations. In addition, we describe the applications, benefits, and drawbacks of these methods. Datasets used in the literature are also included in our analysis. Furthermore, we discussed about upcoming research directions as well as current and upcoming difficulties by using once way anova classification for the validity of the model.

Key words: *Network traffic, Machine learning, deep learning, anova one way*

1. Introduction

Worldwide, people's and businesses' reliance on the Internet has skyrocketed, making it an essential component of daily life [1]. This rise can be attributed to a number of factors, including the revolution in Internet technology, the convenience of accessing the Internet from a variety of devices, competitive prices, and a variety of services offered by the Internet [2]. Right now, it was used more than 5 billon people in 2023, according to the most recent ITU report. This represents an increase of 5.3% compared to 2020 and over 53% compared to 2022. This is accomplished by employing QoS, which prioritizes the traffic of real-time applications for processing by various network nodes [3]. The content is filtered in accordance with the regulations of various nations, which has to use of real life activities in borders. Skype, for instance, was banned in China in 2013 because its protocol did not comply with the country's laws [4]. This means that lawful interception is a requirement. It objectives are leading cybercriminals to focus on carrying out for financial gain. Anti-malware companies increase the confidence in their detection by integrating while the traffic flows with to find the nature of the transactions. For instance, it is possible to identify infected devices and cybercriminals by analyzing botnet network traffic [5].

In order to achieve the objective of classifying traffic is in the network model. The first option, which relied on classification based on packet ports, proved ineffective in applications employing port randomization techniques [6]. Later, the port-based solution was replaced by Deep Packet Inspection (DPI) of the content of network traffic. However, when tested against ciphered network traffic, it demonstrated its flaws. In addition, there has been an increase in the use of machine learning algorithms to classify traffic in a network without having to access the contents or port numbers of the packets [7]. For the purpose of establishing the solution, it has been demonstrated by machine learning frameworks that are supervised, unsupervised, or semi-supervised, each with its own strengths and weaknesses [8]. Convolution Neural Network (CNN) and other deep learning algorithms that use to output values have demonstrated their effectiveness by not needing to extract any statistical features [9].

It is such as the algorithms methods, motivated by the significance of network traffic classification. Adjacent to the conversation tech-niques, we give a top to bottom investigation of both the AI with measurable highlights and profound learning arrangements since they are viewed as the dog lease pattern of organization traffic characterization strategies [10]. There are a lot of studies that have been published about the methods that this

[1]*Buddha Institute of Technology, Gorakhpur shrawan458@bit.ac.in*
[2]*IES College of Technology & Management IES University Bhopal, bhupandra_sati09@yahoo.com*
[3]*Asia Pacific Institute of Information Technology SD India Panipat nishchay@apiit.edu.in*
[4]*Echelon Institute of Technology, Faridabad trilokrawat@eitfaridabad.co.in*
[5]*IIMT College of Engineering Greater Noida dev.mamo@gmail.om*

study looks at, but not all of them can be included in the literature. Therefore, for each technique in our work, we have chosen the published research based on the following criteria: novelty, as below the conditions presenting widely used methods for classifying for this model [11]:

- Describing benefits and drawbacks of the methods discussed.
- Examining the pertinent research regarding each technique.
- Bringing to light that are mentioned.
- It is utilized by literature on data sets.
- Discussing difficulties posed by the techniques used.
- Delivering a methodical form the algorithms
- Indicating potential future directions for network traffic classification research.

Port-based identification the first methods for classifying network traffic used the port numbers of the packets to assign the traffic to the appropriate protocol [12]. The Internet Assigned Numbers Authority (IANA) has assigned distinct port numbers to differentiate between various protocols or vices in network traffic. There are three categories for port numbers: system ports are the standard deployment port range for the protocol. The port numbers are typically inspected during the classification process. For instance, the HTTP protocol is represented by the inspected port number of 80% under process [13].

It has been evaluated. Their empirical evaluation yielded results that were < 80% accurate that were obtained. In the labs of their university, they compared the values % to the online connectionthey had collected. It isimplemented failed to found between 40 and 80 percent of the collected data [14]. Sen and co. It is for classifying P2P model was discussed. Their private data, which included, revealed that the solution offered poor performance and normally 40% of the data in real life.

The advantages of a port-based classification solution can be seen in its ease of implementation, low requirements for computing resources, and rapid [15]. There are two primary drawbacks to it. First, a lot of applications these days use masquerading, which is when they use to send other HTTP, like traffic model. Now, a lot of applications are put in place at random [16].

As a result, it avoids randomization and masquerading. examines the network traffic's payload to identify the underlying applications. The contents of the packets are used to extract a signature, which includes patterns. The signature record in a signature library is connected to a specific application. The classifier can use this method to check the contents of individual packets or groups of packets against a signature library; An alert is sent out in the event of a match, and the traffic is linked to an application [17]. It outperforms port-based identification in terms of accuracy.

They proposed a new framework for RegEx matching called Length-Based Matching, which includes a novel accelerating method. Stride-DFA is the solution. There are two ways , before forwarding it to StriD2FA, LBM begins. It is not constructed entirely from conventional RegEx. When compared to the conventional DFA, the evaluation experiments revealed a decrease in memory consumption. They suggested the less wight DPI framework. The objective of the solution was to maintain acceptable accuracy results while simultaneously reducing the detection process's overhead. It examines the extracting and converting the subsequent communication. Their system has hamming distance built in to cut down on collisions and make it easier to target more applications. A subsequent solution developed by the same authors, makes use of the idea. The n-bit signature for each flow is generated by the solution, which represents raw network traffic in the form of network flows [18]. They were used for the purpose of evaluation, resulting in low misclassification rates.

At present, encrypted network traffic is not categorized by DPI. This slows down the solution. Thirdly, in a number of nations, accessing content from network traffic constitutes a violation of privacy laws or policies. The models are valid from the data set and goal are summarized. It applications or protocols is shown in the table, with a focus on protocols [19][20][21].

2. Algorithms

(i) ML statistical analysis

When classifying network traffic, a statistical classification solution uses machine learning algorithms and statistical feature extraction. It is the second step. Using single or aggregate packets, packet level feature extraction yields features like during flow level feature extraction. The framework in Figure is generally followed by ML. The machine learning framework's general procedure can be summarized as follows [19] [22][23].

- Collection of data: Any method for classifying network traffic that requires a sufficient amount of traffic from targeted applications or protocols must start with this step. Most of the time, two approaches are used: use of public datasets or private data collection. The most recent publicly available datasets in the field of network classification are discussed in Section 7. PCAP files

are typically used to collect the raw traces for further processing.

- Engineering features: Because it influences the classifier's overall performance, this procedure is an essential step. It reflects the characteristics of the traffic that was collected by calculating various metrics that are taken from each flow. Typically, help to distinguish the flows of various applications is known as feature extraction. After meticulously analyzing the collected flows, this procedure could be carried out either manually or with automatically. The process of extracting features produces a column represents a statistical feature and each row represents an application flow. Optional but preferable, feature reduction/selection reducing the classifier's. Wrapper and filter feature selection algorithms are the two most prevalent classifications. By iteratively evaluating the classifier with various feature subsets, Wrapper determines the best feature subset. By analyzing the training dataset, the without evaluating the classifier. Correlation–Based Feature Selection (CFS) selects a set by utilizing correlation and intercorrelation processes. By measuring how relevant features are determines which features are relevant. Thus the process between features and data as set, the cross-correlation process identifies redundant features. According to the algorithms, how relevant they are to the dataset's defined classes. Since the Information Gain (IG) algorithm relies on features with a large number of values, GR's performance is enhance. The difference in entropy between classes before and after observing features in the dataset is used by IG to rank the features in the dataset. The feature selection algorithm ranks features according to how relevant they are to the training dataset by comparing a feature's chi-square statistics to the class and the k-means are given by in Fig 1. [24][25][26].
- For the classifier to be built and evaluated, two datasets must be deployed in separate training and testing datasets. For optimal training, the training phase necessitates a substantial training set. In a similar vein, a suitable test dataset is necessary for effectively evaluating the constructed classifier. Since the results could be misleading, it is considered poor practice to use the same dataset for testing and training. The N-method, also known as semi-supervised solutions, has recently gained popularity. They are discussed in greater detail in subsequent subsects. It was used to illustrate the classifier's performance. The harmonic mean of recall and precision is the F-measure. The following are the equations for the four metrics:[27][28].

$$\text{accuracy} = \frac{(+)+(-)}{2(+)+2(-)}$$

$$\text{recall} = \frac{(+)}{(+)+(-)}$$

$$\text{precision} = \frac{(+)}{2(+)}$$

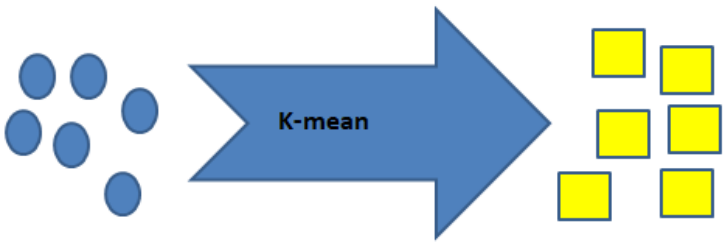

Fig 1. The structure of k-mean

(ii) Deep learning

MLP, Recurrent Neural Network (RNN), Auto En-coders (AE), and CNN are the deep learning methods that are used the most frequently, particularly for the classification of network traffic.MLP is regarded as a NN architecture feedforward and primarily, with several

neurons connected to the layers below it in each layer [19]. The sum of each neuron's inputs is used to determine each neuron's output. Due to the model's numerous parameters, MLP is thought to be difficult. Researchers are encouraged as a stand-alone solution as a result of this type of model is to determine an input's temporal correlation. Time series anomaly detection and speech recognition have both demonstrated the effectiveness of RNN. In contrast, the RNN, the inputs of a previous layer are fed back into the outputs of a subsequent.It provides a long-term learning relationship to solve RNN's vanishing gradient problem [29][30].

One type of deep neural network known as CNN has demonstrated its effectiveness in the field of image recognition by attempting to determine the input's spatial correlation. CNN avoids while simplifying the image for processing. This is accomplished by specifying a filter vector and denoting values. By moving in a specific way of the image, the filter vector extracts, after that, it descends and begins on the left with the same predetermined stride. Until the entire image is covered, this procedure is repeated. Parameter tuning is used to determine. The pooling layer uses the resulting feature map to shrink the convolved features' spatial size. [31][32].

By requiring, AE aims to reconstruct and accomplishes this by employing reduces dimensionality. As part of their architecture, AE is able to incorporate CNN, RNN, and MLP networks. In deep learning solutions, it is primarily utilized for weights initialization. The NN architecture known as stacked auto-encoders (SAEs) stacks multiple AEs, with each AE's output being used as input for the next AE. A greedy layer-by-layer strategy is used by SAE as a whole during the training phase. AE is typically utilized for feature reduction and extraction, and it is regarded as an unsupervised learning algorithm [33][34][35].

Deep learning has recently begun to be used by researchers to divide network traffic into its real life applications. In most cases, the deep learning algorithm takes as its input to determine its temporality or spatiality. In contrast to previous solutions, deep learning eliminates the feature engineering phase, making it is requiring the back details. It is type of deep neural network known as CNN has demonstrated its effectiveness in the field of image recognition by attempting. CNN reduces the image into a form that is easier to process it could affect accurate predictions. This is accomplished by specifying a filter vector and denoting. By moving in a specific way of the image, the filter vector extractswidth is covered. After that, it descends and begins on the left with the same predetermined stride. This procedure is repeated. Parameter tuning is used to determine. The pooling layer uses the resulting feature map to shrink the convolved features' spatial size. Eventually, for the classification process is in a fully connected neural network. If the process is outside lawer, then it has been extensively utilized in the literature. As it is research for manual feature engineering procedure has led to the widespread adoption of deep learning classifiers for the classification of network traffic. The traffic representation has been correlated spatially and temporally using CNN and RNN, respectively, by researchers. One of the first to use deep learning algorithms was Wang [24]. Stacked Auto Encoder (SAE) and Artificial Neutral Network (ANN) were used in the framework, according to its causal applications. More than 58 protocols, both encrypted and unencrypted, were contained in their privately collected network traffic. The evaluation's findings demonstrated by pooling model in Fig 2. [36].

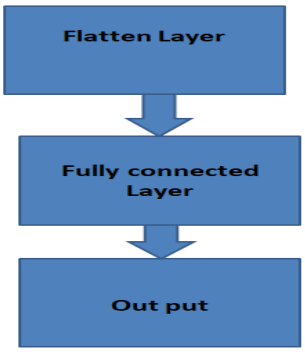

Fig 2 The structure of maximum pooling model

3. Results and discussion

In this study, the most common methods for classifying network traffic were discussed. In order to broaden the scope of this investigation, we intend to discuss additional approaches and frameworks in addition to the models that were specifically examined. The following

are some questions that the survey asks about possible future directions that the research could address:

The use of algorithms is discussed as the two models, particularly with regard to VoIP and Botnet traffic. The need to retrain the classifier in order to identify untrained versions would be less of a problem if this approach were tested in a variety of contexts. The evaluation of the built models was done offline. An important consideration for the built mode's scalability is the performance address in a real-world environment with a lot of those two models.

The combination of briefly discussed as the model, with encouraging outcomes and the one way classification given by Table 1.

Table 1: Analysis of Variance of MLand deep learning

Descriptive	Sum of Squares	df	Mean Square	F	Sig.
1	.2	1	.015	.005	.943
2	57	198	2.893		
3	60	199			
4	.09	1	.080	.243	.623
5	64	198	.330		
6	61	199			
7	.01	1	.012	.032	.859
8	76	198	.381		
9	75	199			
10	.1	1	.108	.372	.543
11	57	198	.291		

Testing of Inference

There is significant difference in Buying decision affect by social media factor (F=7.8502, P<.05) among the respondents of different gender. There is no significant difference in Most preferred social media factor (F=.005, P<.05), Preference of more time spent (F=.243, P>.05), Drive behind joining a brand (F=.032, P>.05), Use of social media sites factor (F=.372, P>.05), Driving force to social media than mass media factor (F=.134, P>.05), Source of data before purchase factor (F=.059, P>.05), Search of alternative in social media factor (F=.007, P>.05), Consumer review that affect purchase decision factor (.764, P>.05), Length of decision making process factor (F=2.228, P>.05) and Opinion on social networking sites factor (F=.456, P>.05), among the respondents of different gender.

4. Conclusion

The purpose of network traffic classification is to determine which protocols, applications, or services are utilized in a monitored network. It helps to apply QoS to the traffic of real-time applications, block the use of particular applications, comply with legal interception regulations, and detect malicious activities, making it an important field for ISPs, countries, and corporations. In order to achieve the classification objective, various solutions have been actively proposed by research communities. The earliest and simplest method is port-based identification, but it differs from randomization, masquerading, and general granularity identification. DPI is more accurate and has a finer grain than the port-based method. However, it necessitates expensive resource requirements and encryption.

References

[1] S. AlDaajeh, H. Saleous, S. Alrabaee, E. Barka, F. Breitinger,K.-K. R. Choo, The role of national cybersecurity strategieson the improvement of cybersecurity education, Computers&Security(2022)102754.

[2] A. Azab, Packing resistant solution to group malware bina-ries, International Journal of Security and Networks 15 (3)(2020)123–132.

[3] S. Alrabaee, A stratified approach to function fingerprintingin program binaries using diverse features, Expert SystemswithApplications193(2022)116384.

[4] S. Alrabaee, M. Debbabi, L. Wang, A survey of binary codefingerprinting approaches:Taxonomy, methodologies, andfeatures, ACM Computing Surveys (CSUR) 55 (1) (2022)1–41.

[5] M.Finsterbusch,C.Richter, E.Rocha, J.-A.Muller,

[6] K. Hanssgen, A survey of payload-based traffic classificationapproaches,IEEECommunicationsSurveysTutorials16(2)(2014)1135–1156.doi:10.1109/SURV.2013.100613.00161.

[7] O. Salman, I. H. Elhajj, A. Kayssi, A. Chehab, A review on machine learning–based approaches for internet traffic classification, Annals of Telecommunications 75 (11) (2020) 673–710.

[8] H. Tahaei, F. Afifi, A. Asemi, F. Zaki, N. B. Anuar, The rise of traffic classification in iot networks: A survey, Journal of Network and Computer Applications 154 (2020) 102538. doi:https://doi.org/10.1016/j.jnca.2020.102538.

[9] N.Hubballi,M.Swarnkar,$bitcoding$:Networktrafficclassificationthroughencodedbitlevelsignatures,IEEE/ACMTrans.Netw.26(5)(2018)2334–2346.doi:10.1109/TNET.2018.2868816.

[10] N. Hubballi, M. Swarnkar, M. Conti, Bitprob: Probabilistic bit signatures for accurate application identification, IEEE Transactions on Network and Service Management 17 (3) (2020) 1730–1741. doi:10.1109/TNSM.2020.2999856.

[11] G. Sun, T. Chen, Y. Su, C. Li, Internet traffic classificationbased on incremental support vector machines, Mobile NetworksandApplications23(4)(2018)789–796.

[12] S.Dong,MulticlassSVMalgorithmwithactivelearningfornetwork traffic classification, Expert Systems with Applica-tions 176 (2021) 114885.doi:https://doi.org/10.1016/j.eswa.2021.114885.

[13] A. Azab, The effectiveness of cost sensitive machine learningalgorithms in classifying zeus flows, Journal of InformationandComputerSecurity17(3-4)(2021)332–350,(InPress).doi:10.1504/IJICS.2020.10026851.

[14] A.Fahad,A.Almalawi,Z.Tari,K.Alharthi,F.S.AlQahtani,

[15] M. Cheriet, Semtra:A semi-supervised approach to trafficflow labeling with minimal human effort, Pattern Recogni-tion 91 (2019) 1–12.doi:https://doi.org/10.1016/j.patcog.2019.02.001.

[16] M. A. Lopez,D. M. Mattos,O. C. M. Duarte,G. Pujolle,Afastunsupervisedpreprocessingmethodfor networkmon-itoring, Annals of Telecommunications 74 (3) (2019) 139–155.

[17] G. Aceto, D. Ciuonzo, A. Montieri, A. Pescapè, Mimetic:Mobileencryptedtrafficclassificationusing multimodaldeeplearning,Computernetworks165(2019)106944.

[18] G.Aceto,D.Ciuonzo,A.Montieri,A.Pescapé,Towardeffectivemobileencryptedtrafficclassificationthroughdeeplearning,Neurocomputing409(2020)306–315.doi:https://doi.org/10.1016/j.neucom.2020.05.036.

[19] G. Aceto, D. Ciuonzo, A. Montieri, A. Pescapè, Mimetic:Mobileencryptedtrafficclassificationusing multimodaldeeplearning,ComputerNetworks165(2019)106944.doi:https://doi.org/10.1016/j.comnet.2019.106944.

[20] Kumar, V. and Kumar, R., 2015. An adaptive approach for detection of blackhole attack in mobile ad hoc network. Procedia Computer Science, 48, pp.472-479.

[21] Kumar, V. and Kumar, R., 2015, April. Detection of phishing attack using visual cryptography in ad hoc network. In 2015 International Conference on Communications and Signal Processing (ICCSP) (pp. 1021-1025). IEEE.

[22] Kumar, V. and Kumar, R., 2015. An optimal authentication protocol using certificateless ID-based signature in MANET. In Security in Computing and Communications: Third International Symposium, SSCC 2015, Kochi, India, August 10-13, 2015. Proceedings 3 (pp. 110-121). Springer International Publishing.

[23] Kumar, Vimal, and Rakesh Kumar. "A cooperative black hole node detection and mitigation approach for MANETs." In Innovative Security Solutions for Information Technology and Communications: 8th International Conference, SECITC 2015, Bucharest, Romania, June 11-12, 2015. Revised Selected Papers 8, pp. 171-183. Springer International Publishing, 2015.

[24] Kumar, V., Shankar, M., Tripathi, A.M., Yadav, V., Rai, A.K., Khan, U. and Rahul, M., 2022. Prevention of Blackhole Attack in MANET using Certificateless Signature Scheme. Journal of Scientific & Industrial Research, 81(10), pp.1061-1072.

[25] Pentyala, S., Liu, M., & Dreyer, M. (2019). Multi-task networks with universe, group, and task feature learning. arXiv preprint arXiv:1907.01791.

[26] Srivastava, Swapnita, and P. K. Singh. "Proof of Optimality based on Greedy Algorithm for Offline Cache Replacement Algorithm." International Journal of Next-Generation Computing 13.3 (2022).

[27] Smiti, Puja, Swapnita Srivastava, and Nitin Rakesh. "Video and audio streaming issues in multimedia application." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018.

[28] Srivastava, Swapnita, and P. K. Singh. "HCIP: Hybrid Short Long History Table-based Cache Instruction Prefetcher." International Journal of Next-Generation Computing 13.3 (2022).

[29] Srivastava, Swapnita, and Shilpi Sharma. "Analysis of cyber related issues by implementing data mining Algorithm." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019.

[30] P Mall and P. Singh, "Credence-Net: a semi-supervised deep learning approach for medical images," Int. J. Nanotechnol., vol. 20, 2022.

[31] Narayan, Vipul, et al. "Deep Learning Approaches for Human Gait Recognition: A Review." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.

[32] Narayan, Vipul, et al. "FuzzyNet: Medical Image Classification based on GLCM Texture Feature." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.

[33] "Keyboard invariant biometric authentication." 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT). IEEE, 2018.

[34] Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." Journal of Pharmaceutical Negative Results (2023): 2607-2615.

[35] Pentyala, S., Liu, M., & Dreyer, M. (2019). Multi-task networks with universe, group, and task feature learning. arXiv preprint arXiv:1907.01791.

[36] Choudhary, Shubham, et al. "Fuzzy approach-based stable energy-efficient AODV routing protocol in mobile ad hoc networks." Software Defined Networking for Ad Hoc Networks. Cham: Springer International Publishing, 2022. 125-139.