Differential Preserving in XGBoost Model for Encrypted Traffic Classification

Zhe Wang[1], BaiHe Ma[2], Yong Zeng[1], XiaoJie Lin[2], KaiChao Shi[1], ZiWen Wang[1]
1.School of Cyber Engineering, Xidian University, Xi'an, China
2.Global Big Data Technologies Centre, University of Technology Sydney, Sydney, Australia
Email: yzeng@mail.xidian.edu.cn

***Abstract*—The classification of encrypted traffic is becoming ever more relevant in the field of network security management and cybersecurity. Most users are currently using encrypted traffic, which can easily lead to privacy threats, and attackers can identify user behavior through the information obtained. VPN encrypted tunnel is the most popular encrypted tunnel method at present. This paper proposes to use the XGBoost model to classify VPNs and Non-VPNs, normalizing the features extracted from encrypted traffic. Experiments are performed on the public dataset ISCX VPN-nonVPN, and the results show that the XGBoost model has an accuracy of 92.4%. To illustrate the advantages of this model, it is compared with the other 5 classification algorithms. At the same time, this paper applies differential privacy technology to the classification model of encrypted traffic and reduces privacy threats by obfuscating data features.**

***Index Terms*—machine learning, encrypted traffic, classification, differential preserving, VPN.**

I. Introduction

In recent years, the public has paid more and more attention to network security and data protection. According to the report [1], more than half of online traffic has been encrypted. While encryption technology helps protect users' privacy, the research team has faced a large amount of traffic that is difficult to decrypt and detect. So far, many studies have analyzed encrypted traffic of a large number of applications, which improve network efficiency by identifying different types of traffic and allocating corresponding resources [2]. The traditional deep packet inspection (DPI) method has a relatively stable recognition rate in actual traffic classification and is widely used in the industry, but the DPI method is difficult to identify encrypted traffic. In the actual network management, it needs to classify different applications under different encryption protocols and application layer protocols using tunnel transmission. Machine learning has been widely used in the encrypted traffic classification and achieved high accuracy classification precision [3-4].

Researchers can identify traffic types by analyzing encrypted payloads or based on features extracted from encrypted traffic. Even if the data is encrypted, user behavior may be inferred from the identification of traffic types. Some users use encrypted tunnels for some online activities, and the ease of use of VPN makes it becoming a common method for attackers [5]. Attackers use VPN to hide their identities and steal users' behavioral information and even business secrets. Therefore, in addition to accurately identifying encrypted traffic, it also needs to prevent leakage of user privacy.

To provide privacy protection for users, studies process the data to ensure that the data is available without causing serious privacy leakage. A large number of literature [6-7] conduct comprehensive research on differential privacy (DP), and prove the availability of DP from a theoretical point of view. Early privacy-preserving models could not provide a rigorous method to prove the availability of their privacy-preserving, and could not quantitatively analyze the model when the model parameters were changed. DP can solve the shortcomings of traditional privacy protection models. In recent years, DP is wildly deployed in recommender systems, network trace analysis, search log protection, and so on.

This paper proposes a differential preserving machine learning classification model by combining differential privacy with machine learning (ML) to analyze time-based features extracted from encrypted traffic, and the accurate classification precision compared to the existing experimental model [8]. In this paper, we classify the regular encrypted data traffic and the encrypted VPN data traffic classification with the public dataset ISCX VPN-nonVPN [9].

The contributions of this paper are summarised as follows:

1) We classify the encrypted data into VPN traffic and Non-VPN traffic with the XGBoost model. Then the categorized encrypted traffic is further classified into 14 types of traffic as shown in Table I.
2) In our model, the row traffic is protected in differential preserving. The proposed model achieves high accurate classification precision. With the differential preserving protected dataset thus our model does not need to decrypt traffic, and our model does not diminish traffic privacy.

The experimented results show that our model achieves higher accurate classification precision than the existing model in [8].

To the best of our knowledge, none of the VPN-nonVPN encrypts traffic classification by considering differential privacy and XGBoost model to protect user privacy and achieve traffic classification.

The rest of this paper is organized as follows. The related works are reviewed in Section II. In Section III, the proposed classification model is presented, followed by the experimental content in Section IV. The simulation is summarised in Section V. Then conclusions and future research are provided in Section VI.

TABLE I
LIST OF ENCRYPTED APPLICATIONS

Traffic	Content
Browsing	Chrome and Firefox
Email	SMTP/S,POP3/SSL and IMAP/SSL
Chat	Facebook,Hangouts,Skype,AIM and ICQ
Streaming	Youtube and Vimeo
File Transfer	FTPS,SFTP and Skype using other service
VoIP	Facebook,Hangouts and Skype
P2P	BitTorrent, etc

II. RELATED WORK

With the advancement of technology, traffic classification methods are gradually proposed. Traditional traffic classification methods are based on packet size and flow-based features [10-11]. Deep Packet Inspection (DPI) is widely used due to its stable recognition rate [12]. DPI classifies encrypted traffic by using the information of the packet header and the payload of the traffic. However, the precision of traffic refined classification based on DPI technology gradually decreases, due to the rapid development of encryption technology, such as HTTPS and VPN tunnel technology. Therefore, we cannot identify encrypted traffic by pattern matching using fingerprints or signatures, etc. to search for keywords in the payload.

To effectively distinguish the types of encrypted traffic, the researchers analyzed the characteristics of the traffic with machine learning algorithms. Agrawal et al.[13] identify P2P traffic based on the port number by using five supervised ML algorithms, such as Naive Bayes Tree and NaiveBayes. Authors in [14] propose a statistical classification algorithm combining K-nearest neighbor (K-NN) and K-means, which can classify encrypted traffic in real time. Gil et al.[9] studied the impact of time characteristics on encrypted traffic and VPN traffic detection, proposed time-related features, and used two different algorithms KNN and C4.5 to test the accuracy of features. The experimental results show that based on time features can achieve more than 80% accuracy. Z Fan et al.[15] proposed a traffic classifier based on the SVM model, and analyzed the impact of various flow and packet-based features on the traffic classification system. The experiment optimized the features, finally selected 13 features, and achieves 98% accuracy in traffic classification, but the model did not consider encrypted traffic, such as the currently popular encrypted tunnel VPN technology. Wang et al.[16] obtained 89% accuracy by using the neural network 2D-CNN method. In the VPN-nonVPN classification problem, Deep Packet [17] achieved 93% accuracy using the CNN model.

In the classification of VPN-nonVPN encrypted traffic, most researchers focus on how to improve the accuracy and precision of the classifier and pay little attention to the problem of users' privacy leakage. In this paper, we combine the differential privacy technique and the XGBoost algorithm to propose a new encrypted traffic classification model to further classify VPN and Non-VPN traffic.

III. PROPOSED CLASSIFICATION MODEL

XGBoost is one of the boosting algorithms. The main idea of the boosting algorithm is to assemble the weak learners into strong learner, and through the iterative combination of the weak learners, adjust the weights of key samples, and iterate the model to obtain the final model. The XGBoost model is based on the residual to train the model to fit the real data scene, perform the efficient calculation of the gradient histogram, and realize the Boosting Tree of extreme parallel computing. The essence of XGBoost is a gradient boosting decision tree (GBDT) [18]. It continuously performs feature splitting to grow a tree. Learning a tree in each round is actually to fit the difference between the predicted value and the actual value of the model in the previous round. When n trees are obtained after training, the score of a sample is predicted according to the characteristics of the sample, and it will fall to a corresponding leaf node in each tree, and each leaf node corresponds to a score. The corresponding scores add up to the predicted value for that sample.

We process the dataset into a format in which XGBoost can be trained, and then set a series of parameters for the model. We use xgb to train, predict and save the model when we have the dataset and parameters. The main process of the XGBoost algorithm is shown in Algorithm 1. Through t iterations, calculate the first and second derivatives of the loss function L of the i sample in the current round and the first and second derivatives of all samples. Arrange the sample features k from large to small, calculate the sum of the first-order and second-order derivatives of the current split node, update the maximum score, and split the subtree based on the division features and eigenvalues corresponding to the maximum score. Until the maximum value of the score is 0, the weak learner $h_t(x)$ is obtained, the strong learner $f_t(x)$ is updated, and the next round of weak learner iteration is entered.

The proposed model distinguishes VPN or Non-VPN traffic as shown in Figure 1.

1. The dataset is randomly selected, 70% of which is used as training datasets and 30% as test datasets. The training dataset is used to train the model, and the test dataset is used to test the accuracy of the model.
2. The normalized correlation coefficient Φ_k [19] is used to analyze the degree of association between the feature and the binary target (VPN or Non-VPN), and the monotonic relationship between the features was evaluated. Therefore, features with strong correlations are removed.
3. XGBoost model is used to select features that are important in the classification process. At the same time, the parameters of XGBoost are adjusted, and a set of parameters with the highest evaluation index is selected. Test the generated model and calculate the performance of the model.

4. Add Gaussian white noise to the original data, vague the impact of data features on user behavior, and analyze the relationship between classifier accuracy and signal-to-noise ratio value. The model can reduce privacy threats by introducing white Gaussian noise to features collected in encrypted traffic.

Algorithm 1 XGBoost Main Process Algorithm

Input:

Training set samples, I={(x_1, y_1),(x_2, y_2),...,(x_m, y_m)};
The maximum number of iterations, T;
Logloss, L;
Regularization factor, λ,γ.

Output:

The strong learner, $f(x)$.

for t = 1 to T **do**
  for i = $i-1$ to m **do**
    G_t=$\sum$g$_{ti}$
    H_t=$\sum$h$_{ti}$
  end for
  $score \leftarrow 0$
  for k = 1 to K **do**
    $G_L \leftarrow 0$
    $H_L \leftarrow 0$
    while $i \leq K$ **do**
      G_L = G_L + g_{ti}
      G_R = $G - G_L$
      H_L = H_L + h_{ti}
      H_R = $H - H_L$
      $score = max(score, \frac{1}{2}(\frac{G_L^2}{H_L+\lambda} + \frac{G_L^2}{H_R+\lambda} - \frac{(G_L+GR)^2}{H_L+H_R+\lambda}) - \gamma)$
      if $score = 0$ **then**
        return $h_t(x)$
      else
        $i++$
      end if
    end while
  end for
end for
return The strong learner, $f(x)$

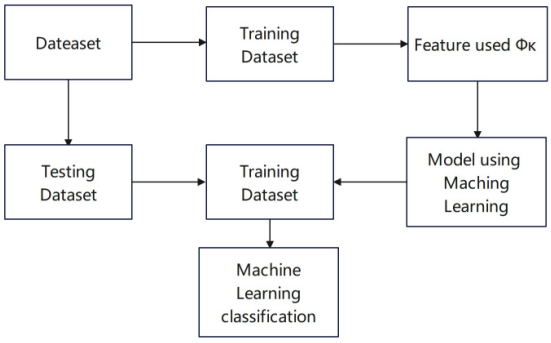

Fig. 1. System Model.

IV. EXPERIMENT

A. Environment

In this paper, we use Python 3.9, implemented under the Google Colab platform, the processor is AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz, with 16 CPU cores and 128GB memory.

B. Dataset

We use a publicly available UNB ISCX VPN-NonVPN network traffic dataset designed by the University of New Brunswick Research Center in Canada [9]. This dataset collects traffic sessions between different users over VPN usage. There are a total of 14 kinds of raw traffic data, 7 kinds of regular encrypted traffic captured by Wireshark and tcpdump, and 7 kinds of traffic encapsulated by VPN protocol, which was obtained by the author using an external VPN service provider. Table I describes the details of the ISCX dataset.

This paper mainly analyzes Time-Based features extracted from encrypted traffic. Like [9], time-based features are extracted, and the duration of the stream is divided into 15, 30, 60, and 120s. As shown in Table II, 8 main features and 23 indicators on the dataset are analyzed, e.g., the maximum value, minimum, mean, and standard deviation.

We first distinguish encrypted data traffic into VPN and Non-VPN networks. The VPN and Non-VPN traffic are then described separately. We divide data traffic into regular encrypted data traffic and encrypted VPN traffic. The proposed method can be extended to other scenarios.

C. Evaluation metrics

To further validate and analyze the proposed model, we use a set of conventional evaluation metrics, such as accuracy, precision, recall, and F-Measure. The efficiency of the classifier is calculated according to the confusion matrix (CM), as shown in Figure 2 below. The following evaluation metrics are used for each record in the test dataset:

- True-positive CM [1][1]: The number of VPN data correctly classified as VPN.
- True-negative CM [0][0]: The number of non-VPN data correctly classified as non-VPN.
- False-positive CM [0][1]: The number of non-VPN data that was misclassified as VPN.
- False-negative CM [1][0]: The number of VPN data that was misclassified as non-VPN.

According to the previous evaluation, we give the following four commonly used evaluation formulas:

Accuracy is defined as the number of samples correctly classified by the classifier to the total number of samples for a given test data set. The formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

Precision is defined as the proportion of all classified traffic types that should be classified correctly. The formula is:

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

TABLE II
TABLE TYPE STYLES

Feature	Table Column Head
duration	Duration of a flow
fiat	Forward Inter Arrival Time, time between two packets sent forward (mean, min, max, std)
biat	Backward Inter Arrival Time, time between two packets sent backward (mean, min, max, std)
flowiat	Flow Inter Arrival Time, time between two packets sent in either direction (mean, min, max, std)
active	Amount of time flow was active before going idle (mean, min, max, std)
idle	Amount of time flow was idle before going active (mean, min, max, std)
fb_psec	Flow bytes per second
fb_psec	Flow packets per second

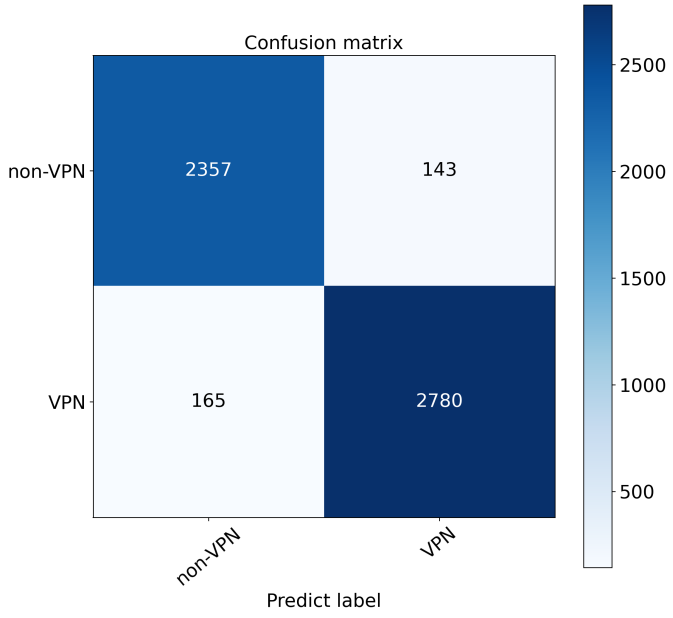

Fig. 2. Confusion Matrix.

The definition of recall is the proportion of all classified traffic types to all traffic types that should be classified. The formula is:

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

F-Measure is precision and recall weighted harmonic mean. The formula is:

$$F - Measure = \frac{2TP}{2TP + FP + FN}. \tag{4}$$

V. SIMULATION RESULTS

A. Comparison of XGBoost and other machine learning in binary classification

By using the normalized correlation coefficient Φ_k to study the degree of correlation between the feature and the binary target (VPN or Non-VPN), and to evaluate the relationship between the features, it was found that no single feature had a significant relationship with the target. We removed features with Φ_k =1 (Figure 3), and for Kendall Tau correlation coefficients, we removed features with correlations over 0.87 (Figure 4). We chose the 15s and 30s datasets to compare XGBoost with other machine learning methods. Among the evaluation metrics described in IV, XGboost is the best performing binary classification problem (Figure 5,6).

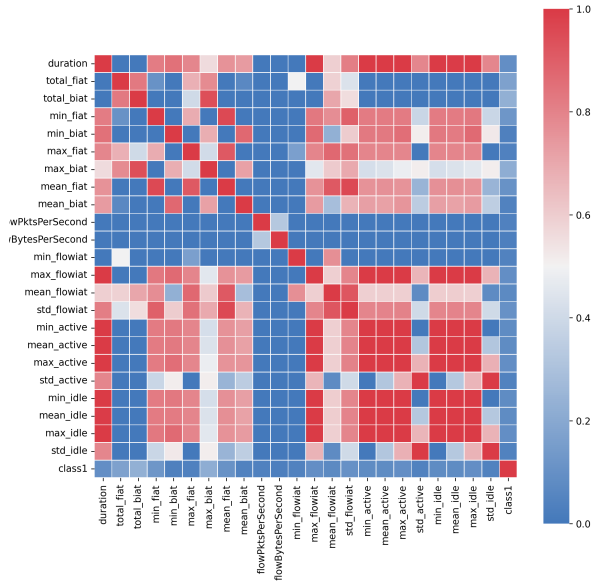

Fig. 3. Heat map using Φ_k correlation coefficient.

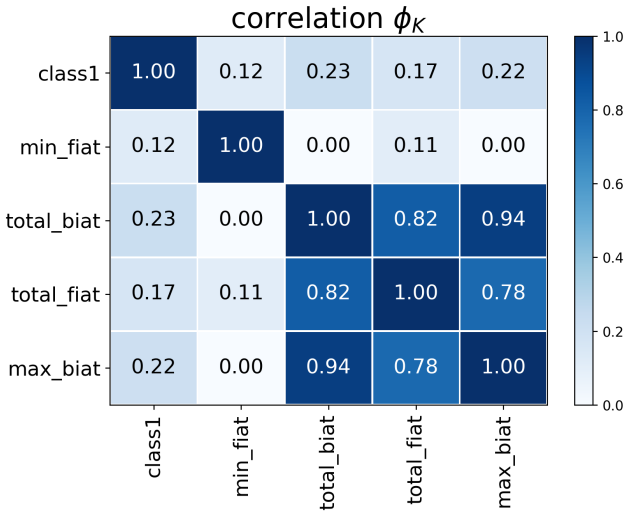

Fig. 4. Correlation Φ_k between features.

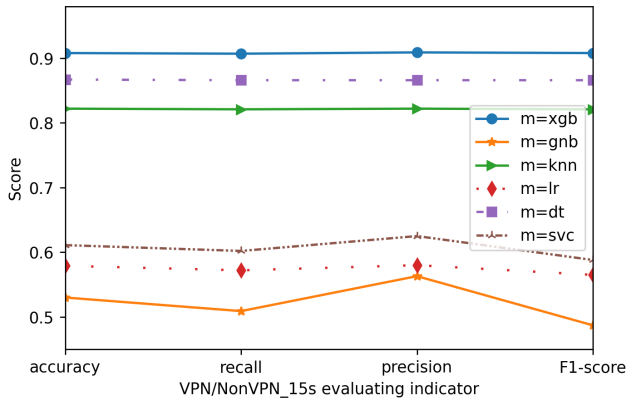

Fig. 5. Results of binary classification on the 15s dataset.

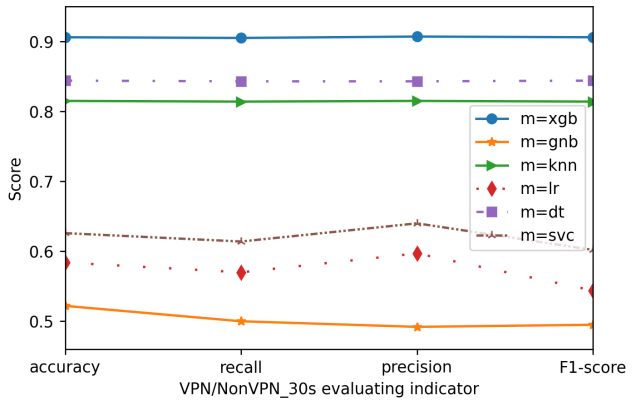

Fig. 6. Results of binary classification on the 30s dataset.

B. Comparison of XGBoost and other machine learning in Multiclass classification

This subsection describes further traffic descriptions for VPN and Non-VPN on the 15s and 30s datasets. Some results are shown in Figure 7,8, XGBoost still has a good classification effect in multi-classification problems.

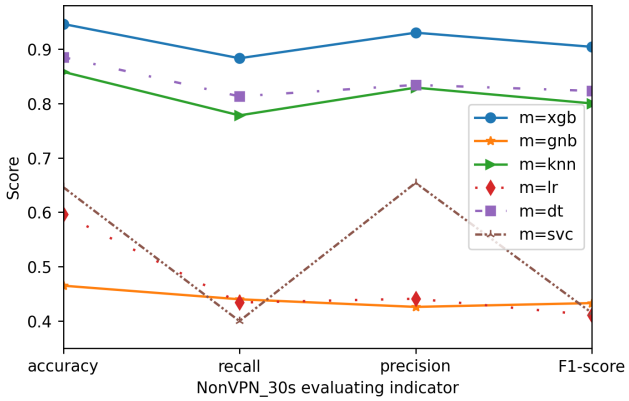

Fig. 7. Results of multiclass classification on the NonVPN-30s dataset.

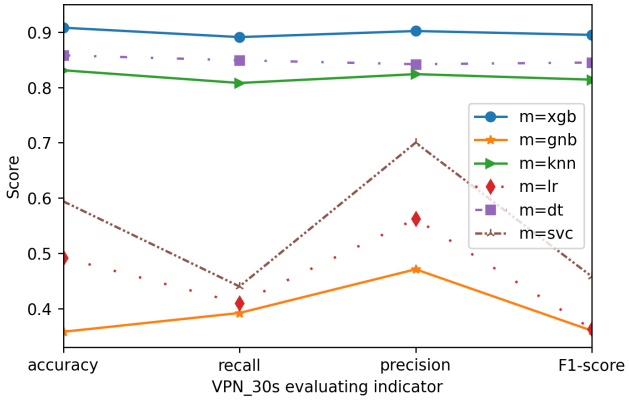

Fig. 8. Results of binary classification on the 30s dataset.

C. Combining XGBoost with Differential Privacy

In the process of analyzing encrypted traffic, privacy leakage is prone to occur, and user behavior can be deduced from the traffic type. The proposed model reduces privacy threats by adding white Gaussian noise (AWGN) to obfuscate user behavior. Adding additive white Gaussian noise is a way to confuse user behavior. It can be done by adding white Gaussian noise to the original dataset and then analyzing the accuracy of the classifier based on the signal-to-noise ratio (SNR) value. It can also be done by adding Laplacian noise. Figure 9 shows the classification accuracy of the XGBoost algorithm under different signal-to-noise ratios. The results show that Gaussian white noise can be introduced to the features in encrypted traffic to reduce privacy threats. Most of the time-frequency representations are degraded by the addition of Gaussian white noise.

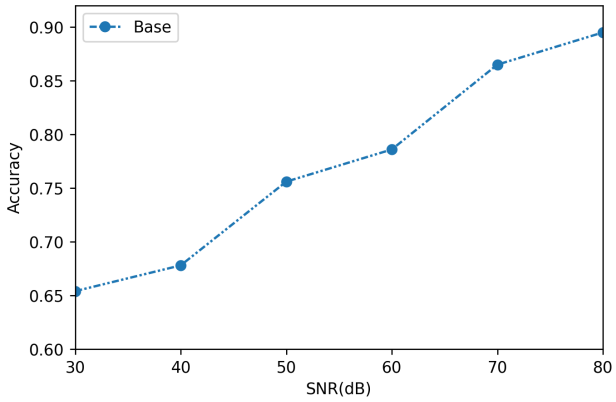

Fig. 9. The recognition accuracy of XGBoost under different SNR values.

VI. Conclusion

In this paper, we used the XGBoost model to classify VPNs and Non-VPNs, and XGBoost showed better classification accuracy compared to other machine learning methods. We also apply Gaussian white noise in encrypted traffic classification. The results show that the introduction of noise has little effect on accuracy and effectively reduces the privacy threat. In the future, we will further study the application of differential privacy in encrypted traffic classification, which has a great impact on maintaining user privacy security.

VII. Acknowledgment

We would like to thank the anonymous reviewers for their insightful comments. This work was sponsored by The National Natural Science Foundation of China (No.61941105).

References

[1] The encrption that protects your online data can also hide malware.Detecting these harmful threats has been a problem until now[EB/OL].[2018-03-03]. https:// newsroom. cisco. com/feature-content? type = web content & article Id = 1853370.

[2] P. Wang, F. Ye, X. Chen, and Y. Qian, “Datanet: Deep learning based encrypted network traffic classification in sdn home gateway,” IEEE Access, vol. 6, pp. 55 380–55 391, 2018.

[3] Jin Y,Duffield N,et al. A modular machine learning system for flow-level traffic classification in large networks[J]. ACM Transactions on Knowledge Discovery from Data,2012,6(1):4.

[4] L. Grimaudo, M. Mellia, E. Baralis and R. Keralapura, ”Self-learning classifier for Internet traffic,” 2013 Proceedings IEEE INFOCOM, 2013, pp. 3381-3386, doi: 10.1109/INFCOM.2013.6567168.

[5] S. Miller, K. Curran, T. Lunney, Multilayer perceptron neural network for detection of encrypted vpn network traffic, in: 2018 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), IEEE, 2018, pp. 1–8. doi:https://doi.org/10.1109/CyberSA.2018. 8551395.

[6] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun and A. A. Ghorbani, ”Characterization of encrypted and VPN traffic using time-related features”, In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP), pp. 407-414, 2016.

[7] Jaewoo Lee and Christopher W. Clifton. Top-k frequent itemsets via differentially private fptrees. In KDD’14, pages 931–940, 2014. DOI: 10.1145/2623330.2623723. 93-96, 102-112.

[8] Sami Smadi et al.,VPN Encrypted Traffic classification using XGBoost, International Journal of Emerging Trends in Engineering Research, 9(7), July 2021, 960 – 966.

[9] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun and A. A. Ghorbani, ”Characterization of encrypted and VPN traffic using time-related features”, In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP), pp. 407-414, 2016.

[10] V. Paxson, Empirically derived analytic models of wideareatcp connections, IEEE/ACM transactions on Networking 2 (4) (1994) 316336. doi:https://doi.org/10.1109/90.330413.

[11] V. Paxson, S. Floyd, Wide area traffic: the failure of poisson modeling, IEEE/ACM Transactions on networking 3 (3) (1995).

[12] T. AbuHmed, A. Mohaisen, and D. Nyang, “A survey on deep packet inspection for intrusion detection systems,” arXiv preprint arXiv:0803.0037,2008.

[13] S. Agrawal and B. S. Sohi, “Feature optimization and performance evaluation of machine learning algorithms for identification of p2p traffic,” Journal of Advances in Information Technology, vol. 3, no. 2, pp. 107–114, 2012.

[14] R. Bar-Yanai, M. Langberg, D. Peleg, and L. Roditty, “Realtime classification for encrypted traffic,” in International Symposium on Experimental Algorithms. Springer, 2010, pp. 373–385.

[15] Z. Fan, R. Liu, Investigation of machine learning based network traffic classification, in: 2017 International Symposium on Wireless Communication Systems (ISWCS), IEEE, 2017, pp. 1–6. doi:https://doi.org/10.1109/ISWCS.2017. 8108090.

[16] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, “Malware traffic classi-fication using convolutional neural network for representation learning,”inProc. IEEE Int. Conf. Inf. Netw., 2017, pp. 712–717.

[17] Lotfollahi M. et al. Deep packet: A novel approach for encrypted traffic classification using deep learning //Soft Computing. 2020.24. No.3. 1999-2012.

[18] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232. URL http://www.jstor.org/stable/2699986.

[19] Baak, M., Koopman, R., Snoek, H., Klous, S., 2020. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. Computational Statistics & Data Analysis.. doi:10.1016/j.csda.2020.107043.