DOI: 10.24996/ijs.2022.63.9.33

ISSN: 0067-2904

Network Traffic Prediction Based on Boosting Learning

Mohammed A. Kashmoola *, Mohammed Kassim Ahmed, Naors Y. Anad Alsaleem

Department of Computer Sciences, College of Education, University of Al-Hamdaniya, Ninawa, Iraq

Received: 18/11/2021 Accepted: 1/2/2022 Published: 30/9/2022

Abstract

Classification of network traffic is an important topic for network management, traffic routing, safe traffic discrimination, and better service delivery. Traffic examination is the entire process of examining traffic data, from intercepting traffic data to discovering patterns, relationships, misconfigurations, and anomalies in a network. Between them, traffic classification is a sub-domain of this field, the purpose of which is to classify network traffic into predefined classes such as usual or abnormal traffic and application type. Most Internet applications encrypt data during traffic, and classifying encrypted data during traffic is not possible with traditional methods. Statistical and intelligence methods can find and model traffic patterns that can be categorized based on statistical characteristics. These methods help determine the type of traffic and protect user privacy at the same time. To classify encrypted traffic from end to end, this paper proposes using (XGboost) algorithms, finding the highest parameters using Bayesian optimization, and comparing the proposed model with machine learning algorithms (Nearest Neighbor, Logistic Regression, Decision Trees, Naive Bayes, Multilayer Neural Networks) to classify traffic from end to end. Network traffic has two classifications: whether the traffic is encrypted or not, and the target application. The research results showed the possibility of classifying dual and multiple traffic with high accuracy. The proposed model has a higher classification accuracy than the other models, and finding the optimal parameters increases the model accuracy.

Keywords: network traffic classification, machine learning algorithms, boosting, XGboost.

التنبؤ بحركة مرور الشبكة على أساس التعلم المعزز

محمد علاء الدين كشمولة*، محمد قاسم احمد، نورس يونس عناد

قسم علوم الحاسوب, كلية التربية, جامعة الحمدانية, نينوى, العراق

الخلاصة

يعد تصنيف حركة مرور الشبكة من المواضيع المهمة لإدارة الشبكة وتوجيه المرور وتمييز المرور الأمن وتوفير خدمة افضل. ان تحليل المرور هو عملية كاملة من اعتراض بيانات حركة المرور إلى اكتشاف العلاقات والأنماط والتشوهات والتكوينات الخاطئة في الشبكة. من بينها ، تصنيف حركة المرور وهو مجال

*Email: nawrasyounis@yahoo.com

فرعي في هذا المجال ، والغرض منه هو تصنيف حركة مرور الإنترنت إلى فئات محددة مسبقًا ، مثل حركة المرور العادية أو غير الطبيعية ، ونوع التطبيق. تقوم معظم تطبيقات الانترنت بتشفير البيانات اثناء حركة المرور و تصنيف البيانات المشفرة اثناء حركة المرور غير ممكن بالطرق التقليدية. ويمكن للطرق الاحصائية والذكائية العثور على انماط حركة المرور التي يمكن تصنيفها بناء على الخصائص الاحصائية و نمذجة تلك الخصائص. تساعد هذه الاساليب في تحديد نوع حركة المرور و حماية خصوصية المستخدم في نفس الوقت. ولتصنيف حركة مرور الشبكة من طرف الى طرف تقترح هذه الورقة , استخدام خوارزميات (ْ XGboost) وايجاد المعلمات العليا باستخدام امثلية Bayesian و مقارنة النموذج المقترح مع خوارزميات التعلم الالي (Nearest Neighbor, Logistic Regression, Decision Trees, Naive Bayes, Multilayer Neural Networks)لتصنيف حركة مرور الشبكة لنوعين من التصنيفات , تصنيف هل المرور مشفر ام لا وتصنيف وحهة المرور والتطبيق المستهدف. اظهرت نتائج البحث امكانية تصنيف المرور الثنائي والمتعدد بدقة عالية وان النموذج المقترح يتمتع بدقة تصنيف اعلى مقارنة بباقي النماذج , كما ان ايجاد المعلمات المثلى يزيد من دقة النموذج.

1. Introduction

Traffic classification is a key activity in network administration and cyber security since it categorizes network traffic into particular groups based on needs[1] . For example, in the realm of network management, traffic might be categorized based on different priorities to guarantee the network quality of service. Traffic can be identified as harmful or unauthorized traffic. Traffic can be separated into non-malicious or malicious traffic. Traffic encryption has recently become normal practice thanks to the extensive usage of encryption technologies in network applications[2, 3].

Detecting encrypted traffic entails distinguishing encrypted from unencrypted traffic, and several types of research have been conducted on this topic[4]. The detailed categorization of encrypted traffic indicates that the communication is related to certain applications [5, 6]. This work is relatively challenging due to the huge diversity of programs and versions. Connecting encrypted communication to an application (such as a chat or broadcast) is known as encrypted profiling traffic, and this task has lately received a lot of attention.

This study is based on using statistics-based and behavior-based methods (machine learning) to classify traffic. The general workflow is as follows: first, manually design the traffic features; then extract, select, and configure these features to train the models. Finally, the traffic is categorized with these characteristics by the selected classifiers (Nearest Neighbor, Logistic Regression, Decision Trees, Naive Bayes, Multilayer Neural Networks).

This study proposes a comprehensive method for classifying encrypted traffic using XGboost, a learning method based on the (boosting) principle; finding hyperparameters using Bayesian optimization; and comparing the proposed model with the machine learning algorithms selected in this study.

2. Related works

Wang et al. [7]: Data transmission collects a vast quantity of information about Web users' access, which is crucial for network security. The initial step in categorizing network traffic in defect detection is frequently extracting characteristics from these records.

Gol et al. [8]: Approaches for categorizing traffic include DPI-based, port-based, behavioral, and statistical methods. These are typical machine learning approaches for classifying traffic, which need a large number of features. It is the first to employ data mining to identify malware, and it relies on three immutable features: strings, a portable executable (PE), and byte sequence.

Salman et al. [9]: For Deep Learning-based identification, they looked at two data representation methods: raw packet-based representation and quasi-raw stream-based representation. They discovered that traffic anonymization and the fact that several packet fields are data-dependent hurt the raw data representation. The flow-based representation, on the other hand, is sensitive to the number of packets utilized for categorization and traffic congestion.

Seddigh et al.[10] have conducted research into the classification of secured packets on high-speed networks. The dataset was gathered from the campus network, and despite the fact that there are over two hundred features, because of the high traffic bandwidth, feature selection was made. There are six classes in the dataset. Six different machine learning algorithms were utilized, and MLTAT (machine learning traffic analytics tool) was established as a framework. MLTAT was used to select hyperparameters, and binary and multiclassification were performed. The accuracy rate in the tests was above 88 percent for all courses, but the Web surfing class had a much lower average rate of success.

Uğurlu et al.[11]: The researchers recommended using extreme gradient boosting (XGBoost), decision trees, and random forest classification algorithms to categorize network traffic by examining incoming and outgoing data via encrypted traffic. They discovered that, without decryption, it is possible to classify packets traveling through encrypted communication based on information such as size and length, and take security precautions.

3. Machine learning model

A. Logistic Regression

It is a linear regression statistical model that allows modeling of a binomial variable in terms of a collection of predicted random variables, either numerical or categorical [12]. Forecasting in logistic regression is done by calculating the probability of an event with knowledge of the values associated with the event. Logistic regression uses several predicted variables, which can be numeric or categorical. Logistic regression is also known as the Logit model or the general classifier of entropy [13]. This modeling is widely used in many scientific and commercial applications. It is one of the most widely applied modeling methods in machine learning, as it is classified among the methods of controlled machine learning [14].

B. Nearest Neighbor

The Nearest Neighbor algorithm is considered one of the most important and simplest directed machine learning algorithms that works with a supervisor [15, 16]. The Nearest Neighbor algorithm is considered a descriptive and predictive classification algorithm. It can deal with anomalous data. The principle of work of this algorithm depends on calculating the Euclidean distance between points, where the less the distance between two points, the greater the possibility of the points belonging to each other, hence the name of the algorithm. The algorithm will measure the distance between the target point and the nearest points to it [17].

C. Decision Tree

The decision tree is one of the most important techniques through which the knowledge inherent in huge amounts of data is deduced and knowledge cases that support decision-making are reached. Decision trees are a supervised learning method used for classification and regression [18]. Its objective is to create a model for predicting a given value by learning simple rules deduced from the characteristics of the data. The classification process is applied through a set of rules or conditions that determine the path followed from the root node and ends with one of the final nodes representing the final decision [19]. For all non-final nodes, a decision should be made about the next node. The decision is to choose a solution from several solutions to a particular problem. Therefore, decision-making is the choice of one of the available alternatives, so The decision-making process is a series of stages and procedures that lead to selecting the best alternative options in the end [19, 20].

D. Artificial neural networks (ANN)

ANN is used for data mining to achieve high accuracy in many complex classification and prediction problems [21]. The ANN framework generates many patterns to solve various prediction and classification problems [22]. Electrochemical activity between a network of brain cells known as neurons occurs as a continuous activity in our brain, and as a result, simulations of artificial neural networks are common in artificial intelligence research [22, 23]. Artificial neural networks are made up of units connected by links. These connections are used to transport activity between the units, and each link has a weight that grows as the strength of the connection between the two units linked by this link increases. The stronger the connection between two units, the greater their weight. The information that we want to process is placed at the first layer of units, and the output of each neuron may be an input to another neuron, and each unit has a fictitious input whose value is equal to one, transmitted through a link that is also loaded with an initial weight [24, 25].

E. Naïve Bayes

The Bayes' algorithm is one of the most famous Machine Learning and Analytics algorithms used in classification problems, particularly as it is characterized by its speed in processing and efficiency in prediction operations [26]. This method is based on the statistical concept of Bayes' theorem, which calculates the probability of a certain result occurring by verifying what is available and known, and it is called (naive) because it adopts the principle of independent hypotheses, as it depends on the independence of the relationship between the characteristics (Attributes Features) and each other. Because the model ignores any possible relationships between the qualities, they all contribute to the probability computation. An outcome is a number that has no relevance in describing the dependence of one attribute on another or the order value [27].

F. XGBoost

XGBoost is a gradient-boosting tree model framework proposed by Chen. The basic idea of XGBoost is the same as GBDT [28]. For a dataset containing samples an feature $\mathcal{D} = \{(thm)\}$ among them $|\mathcal{D}| = n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$, an integrated treemodel can b used K Te predictive output of an addition function:

$$= \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\text{them}), f_k \in \mathcal{F}\text{s} \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

among them, $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is the regression tree (CART), q Represents the structure of each tree, which maps a sample to the final leaf node, T is the number of leaf nodes, each FW, the structure of a single corresponding tree, is q and the weight is w the tree. Different from decision trees, each leaf node of each regression tree contains a continuous score. The use wi Represents the first I Points on each leaf node [29].

4. practical framework:

Dataset

The ISCX VPN-NonVPN traffic dataset has been used [30-36]. The dataset includes two scenarios (A and B). Both scenarios contain a dataset of different times (15, 30, 60, and 129 seconds) and 23 attributes that include network traffic attributes. The output of data set A has two categories (VPN and Non-VPN) as shown in Table (1), and data set B has seven categories as shown in Table (2).

Table 1-VPN data category and content of scenario A

Label	Sample
VPN	5632
Non-VPN	5151

Table 2-VPN data category and content of scenario B

Label	Type	Sample
BROWSING	Firefox, Chrome	5000
CHAT	ICQ, AIM, Skype, Facebook, Hangouts	591
FT	FTP, SFTP	1340
MAIL	Email, SMTP, POP3, IMAP	907
P2P	Torrent	1813
STREAMING	Vimeo, YouTube, Netflix, Spotify	353

VOIP	Facebook, Skype, Hangouts	778

Proposed framework

At this point, the data has been initialized for classification. The initialization process included converting categorical data to digital and then generalizing it because some models that use the concept of distance are affected by non-generalized data. After this stage, the data set was divided into training data and test data in a ratio of 20:80, training the selected models and then measuring the performance of each model as shown in Figure 1.

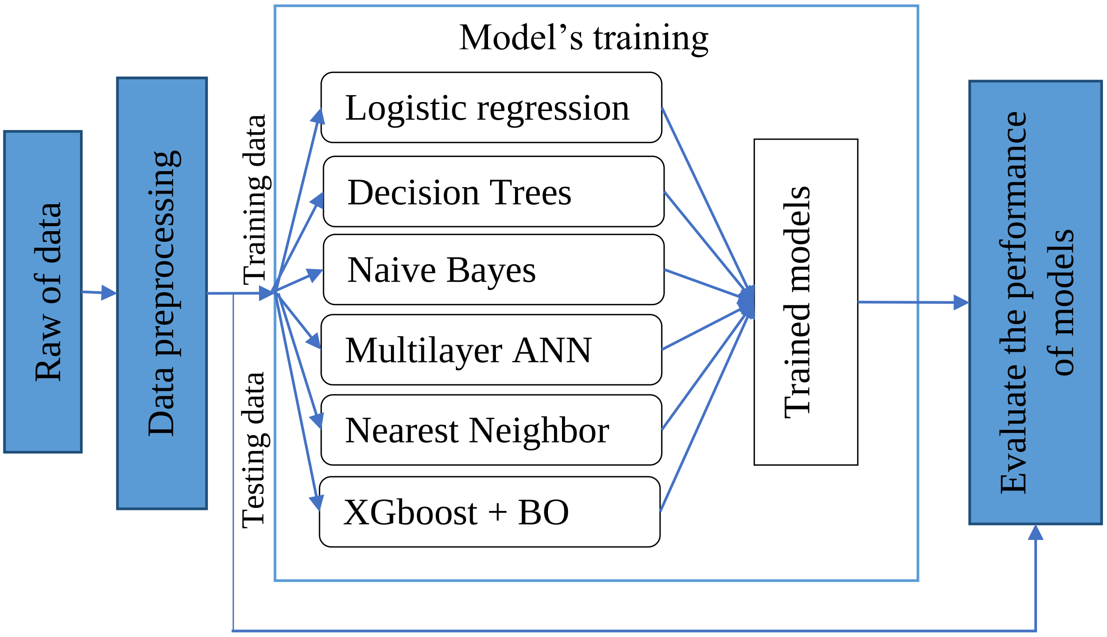

Figure 1-Proposed framework.

5. Evaluation criteria and experimental results

The results have been analyzed and the proposed models trained in the open-source Python tool. To evaluate the models, comparisons, and proposed results, the data has been divided into two groups: the first for training and the second for testing, using (5 folds cross) because this option is widely used, especially if we have a limited amount of data set. The data set is randomly divided into five subgroups. The first group was used for testing and the remaining four groups for training. Finally, the average product of the five groups is calculated. This process is done for the six proposed algorithms to compare them in terms of efficiency.

This paper uses the confusion matrix to represent the classification results to compare the classification model's performance, as shown in the table below:

Class I	*predicted positive*	*predicted negative*
The actual positive	True Positive	False Negative
The actual negative	False Positive	True Negative

True positive (TP) means that traffic belonging to category I is classified into a category I; false negative (FN) refers to traffic that belongs to category I and is classified as non-category I; false positive (FP) refers to traffic that is not a category I is classified as Category I. Based on the above concepts, give the accuracy, precision, and recall calculation method.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \ldots\ldots\ldots\ldots\ldots (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

The different results were examined based on the algorithms of the classification process that were applied to the network traffic data set. Figures 2, 3, 4, and 5 show the performance of machine learning algorithms in the case of binary classification of the selected metrics, and Figures 5, 6, and 7 represent the results of the algorithms for classifying the application used (multiclass classification).

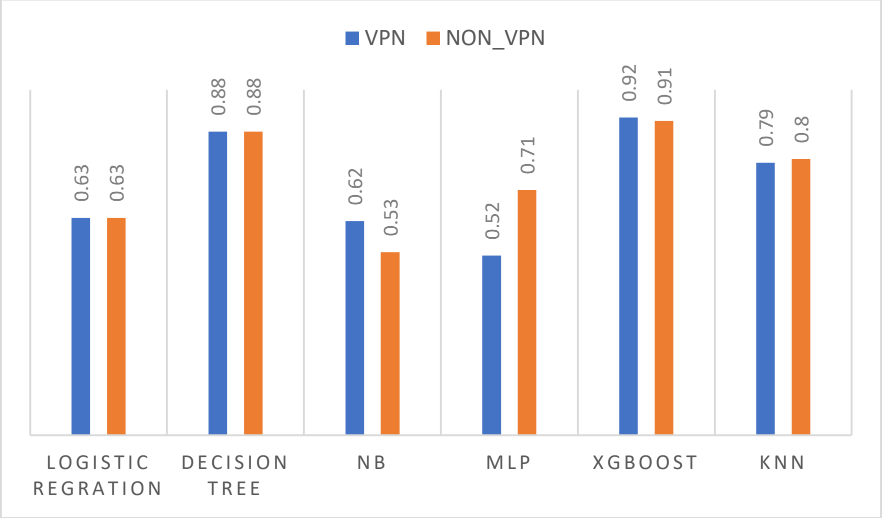

Figure 2- precision of binary classification scenario A 120s.

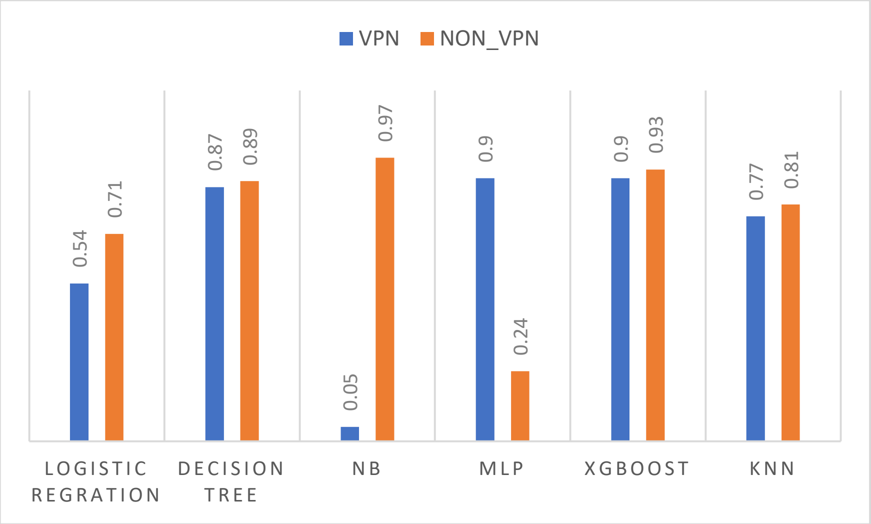

Figure 2- recall of binary classification scenario A 120s.

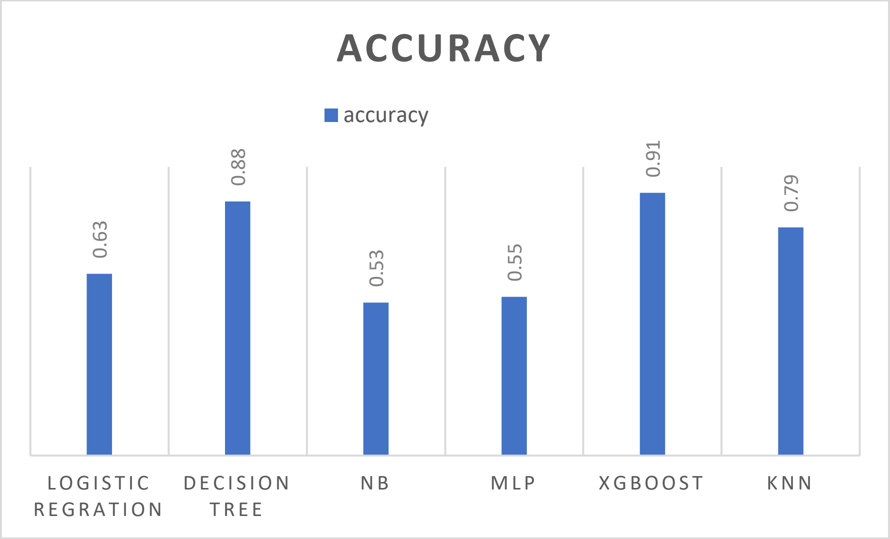

Figure 3-Accuracy of binary classification scenario A 120s

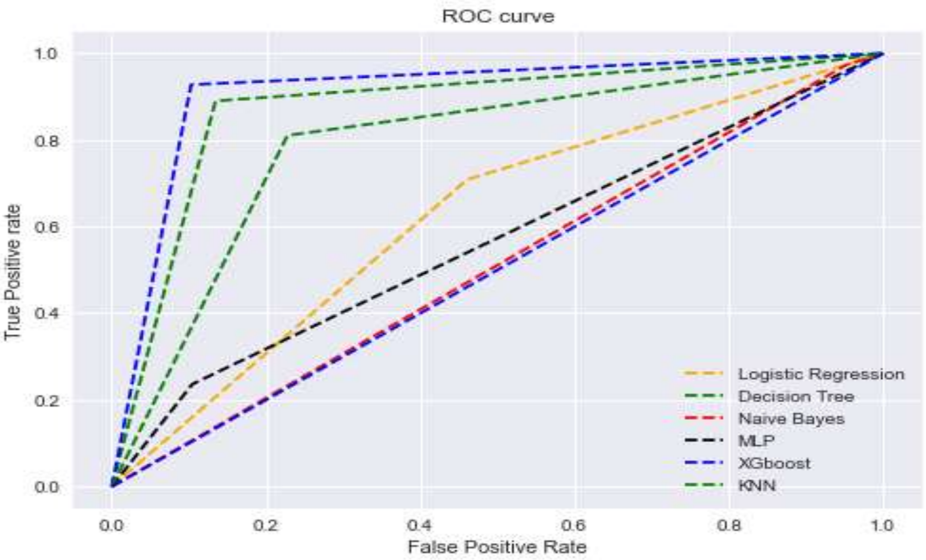

Figure 4- ROC of binary classification scenario A 120s.

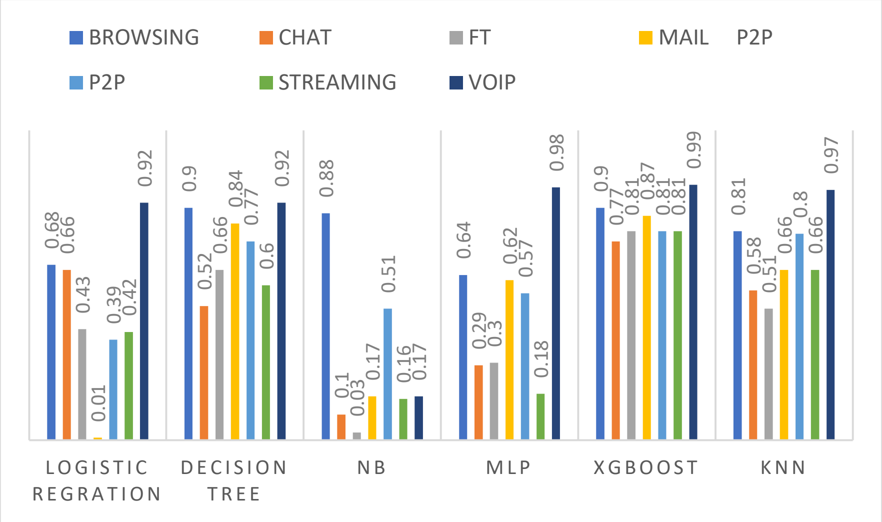

Figure 5- Precision of multiclass classification scenario B 120s.

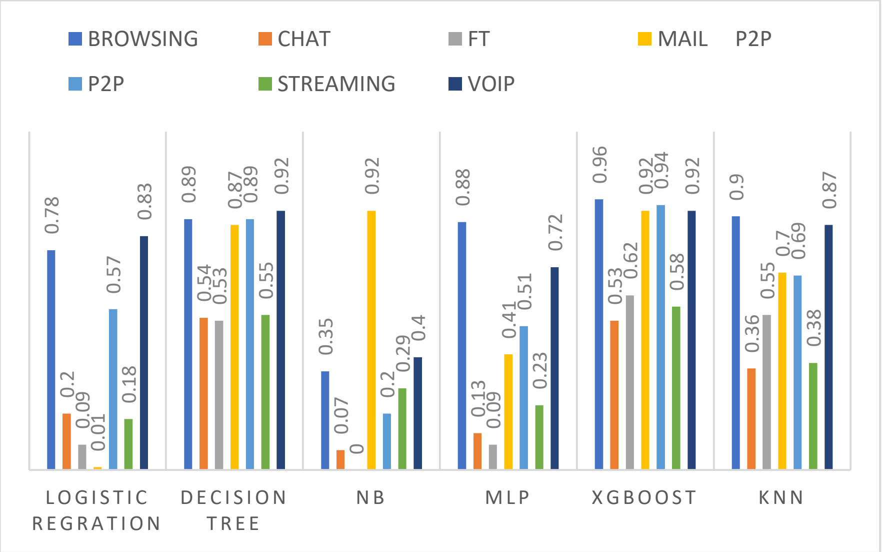

Figure 6- Recall of multiclass classification scenario B 120s.

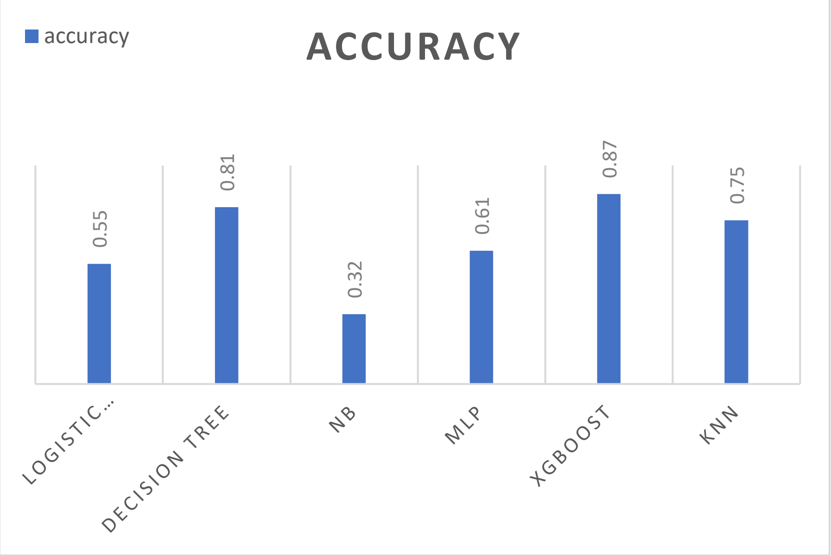

Figure 7-Accuracy of multiclass classification scenario B 120s.

From the results shown above, it is clear that the performance of XGboost and the decision tree were to some extent good, and the performance of the XGboost model has been 91% in the binary classification and 87% in the multiple classifications, while the decision trees achieved an accuracy of 88% in the binary classification and 81% in the multiple classifications The superiority of the algorithm is due to the use of the principle of gradient boosting learning, which gives more accurate results and addresses the problem of data imbalance at the same time.

6. Conclusion

Classification of encrypted traffic is a critical task in network management and cyber security. A network traffic classification model has been built based on six machine learning algorithms: Nearest Neighbor, Logistic Regression, Decision Trees, Naive Bayes, Multilayer Neural Networks, and XGboost. The study explored the possibility of modeling network traffic using supervised machine learning algorithms to determine encrypted and unencrypted traffic in addition to finding the type of traffic in terms of the requested website. The results showed that these features could be modeled and categorized. The best performance in classification was for the XGBoost algorithm, which reached an accuracy of 91% in binary classification and 87% in multiple classifications. Testing methods for features, selecting and finding different hyperparameters using swarms can give different results in future work.

References

[1] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, *End-to-end encrypted traffic classification with one-dimensional convolution neural networks*, 2017.

[2] L. Kong, G. Huang, K. Wu, Q. Tang, and S. Ye, *Comparison of Internet Traffic Identification on Machine Learning Methods*, 2018.

[3] L. Peng, B. Yang, Y. Chen, and Z. Chen, "Effectiveness of Statistical Features for Early Stage Internet Traffic Identification," *International Journal of Parallel Programming*, vol. 44, 01/18, 2015.

[4] T.-H. Jeng, W.-Y. Luo, C.-C. Huang, C.-C. Chen, K.-H. Chang, and Y.-M. Chen, "Cloud Computing for Malicious Encrypted Traffic Analysis and Collaboration," *International Journal of Grid and High Performance Computing*, vol. 13, pp. 12-29, 07/01, 2021.

[5] A. Sawabe, T. Iwai, and K. Satoda, *Identification of Smartphone Applications by Encrypted Traffic Analysis*, 2019.

[6] S. Deng, M. Wei, M. Xu, and W. Cai, “Prediction of the rate of penetration using logistic regression algorithm of machine learning model,” *Arabian Journal of Geosciences,* vol. 14, 11/01, 2021.

[7] F. Wang, X. Xiang, J. Cheng, and A. Yuille, *NormFace: L2 Hypersphere Embedding for Face Verification*, 2017.

[8] B. Atli, Y. Miche, A. Kalliola, I. Oliver, S. Holtmanns, and A. Lendasse, “Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space,” *Cognitive Computation,* vol. 10, 10/01, 2018.

[9] O. Salman, I. Elhajj, A. Kayssi, and A. Chehab, “Data Representation for CNN Based Internet Traffic Classification: A Comparative Study,” *Multimedia Tools and Applications,* vol. 80, 05/01, 2021.

[10] S. Dolgikh, N. Seddigh, B. Nandy, D. Bennett, C. Zeidler, Y. Ren, J. Knoetze, and N. Muthyala, *A Framework & System for Classification of Encrypted Network Traffic using Machine Learning*, 2019.

[11] M. Uğurlu, İ. Doğru, and R. S. Arslan, “A new classification method for encrypted internet traffic using machine learning,” *Turkish Journal of Electrical Engineering and Computer Sciences,* vol. 29, pp. 2450-2468, 04/19, 2021.

[12] J. Savoy, *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*, 2020.

[13] J. Žižka, F. Dařena, and A. Svoboda, *Text mining with machine learning: principles and techniques*: CRC Press, 2019.

[14] O. Borgohain, M. Dasgupta, P. Kumar, and G. Talukdar, "Performance Analysis of Nearest Neighbor, K-Nearest Neighbor and Weighted K-Nearest Neighbor for the Classification of Alzheimer Disease," pp. 295-304, 2021.

[15] V. Prasath, H. A. A. Alfeilat, A. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, and H. S. E. Salman, “Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier--A Review,” *arXiv preprint arXiv:1708.04321,* 2017.

[16] H. Zhou, "K-Nearest Neighbors," *Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods*, H. Zhou, ed., pp. 93-108, Berkeley, CA: Apress, 2020.

[17] Z.-H. Zhou, "Decision Trees," *Machine Learning*, Z.-H. Zhou, ed., pp. 79-102, Singapore: Springer Singapore, 2021.

[18] J. Suzuki, "Decision Trees," *Statistical Learning with Math and Python: 100 Exercises for Building Logic,* J. Suzuki, ed., pp. 171-198, Singapore: Springer Singapore, 2021.

[19] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis, *On the Explanatory Power of Decision Trees,* 2021.

[20] M. P. Del Rosso, S. Ullo, A. Sebastianelli, D. Spiller, E. Puglisi, F. Biondi, and D. Orlando, "Artificial neural network," pp. 63-90, 2021.

[21] D. Zhang, "Artificial Neural Network," pp. 229-270, 2021.

[22] F. Feng, W. Na, J. Jin, J. Zhang, W. Zhang, and Q.-J. Zhang, “Artificial Neural Networks for Microwave Computer-Aided Design: The State of the Art,” *IEEE Transactions on Microwave Theory and Techniques*, 2022.

[23] H. e. e. h. i. l. g. v. r. e. Cartwright, *Artificial Neural Networks,* 3rd 2021. ed., 2021.

[24] M. O. Okwu, and L. K. Tartibu, "Artificial Neural Network," *Metaheuristic Optimization: Nature-Inspired Algorithms Swarm and Computational Intelligence, Theory and Applications*, M. O. Okwu and L. K. Tartibu, eds., pp. 133-145, Cham: Springer International Publishing, 2021.

[25] D.-H. Vu, “Privacy-preserving Naive Bayes classification in semi-fully distributed data model,” *Computers & Security,* vol. 115, pp. 102630, 2022.

[26] B. Şeref, and E. Bostanci, “Performance Comparison of Naïve Bayes and Complement Naïve Bayes Algorithms,” *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE),* pp. 131-138, 2019.

[27] A. M. Sieminski, and C. R. Donovan, “Forecasting overhead distribution line failures using weather data and gradient-boosted location, scale, and shape models,” *arXiv preprint arXiv:2209.03495,* 2022.

[28] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009.

[29] N. Mohammadi, and A. Shirmarz, "A Novel Deep Encrypted Network Traffic Discriminator in Software Defined Network (SDN)," 2022.

[30] M. Kashmoola, M. Anad, N. Alsaleem, and M. Moskalets, "Model of dynamics of the grouping states of radio electronic means in the problems of ensuring electromagnetic compatibility," *Eastern-European Journal of Enterprise Technologies,* vol. 6, 12/27, 2019.

[31] M. Anad, and S. Hasoon, "COMPARISON OF DT& GBDT ALGORITHMS FOR PREDICTIVE MODELING OF CURRENCY EXCHANGE RATES," *EUREKA: Physics and Engineering,* vol. 1, pp. 56-61, 01/31, 2020.

[32] S. Aziz, and M. Ahmed, "Method for determining the responses from a non-linear system using the Volterra series," *Eastern-European Journal of Enterprise Technologies,* vol. 4, pp. 106, 09/12, 2020.

[33] N. Zakar, N. Alsaleem, and M. Kashmoola, *Multi-agent Models Solution to Achieve EMC In Wireless Telecommunication Systems*, 2018.

[34] E. Nasser, and F. Dawood, "Diagnosis and Classification of Type II Diabetes based on Multilayer Neural Network," *Iraqi Journal of Science*, pp. 3744-3758, 10/30, 2021.

[35] Z. Abood, H. Taher, and R. Ghani, "Detection of Road Traffic Congestion Using V2V Communication Based on IoT," *Iraqi Journal of Science,* pp. 335-345, 01/30, 2021.

[36] Z. Shah Weli, *Deep Belief Network for Predicting the Predisposition to Lung Cancer in TP53 Gene*, 2020.