Effects of Adversarial attack on Darknet Traffic Classsification

N. Kathiravan
Department of Computer Science & Engineering
National Institute of Technology Puducherry
Karaikal – 609609, India
kathiravannarayanan2@gmail.com

Jerome B
Department of Computer Science & Engineering
National Institute of Technology Puducherry
Karaikal – 609609, India
dbjerome7@gmail.com

Karthik N
Department of Computer Science & Engineering
National Institute of Technology Puducherry
Karaikal – 609609, India
nkarthikapce@gmail.com

Vani V
Department of Computer Science & Engineering
National Institute of Technology Puducherry
Karaikal – 609609, India
v.vani465@gmail.com

***Abstract*— Though internet anonymity in the darknet aims to protect privacy, it is often exploited for criminal purposes. Because of this, traffic classification of darknet using machine learning models is very crucial. This study demonstrates that Random Forest and XG-Boost achieves high f1-scores on CIC-Darknet2020 dataset. Also, to test robustness of the models, some portions of the dataset are obfuscated to simulate Adversarial attacks on machine learning models were demonstrated.**

Keywords—Darknet, Adversarial attacks, Random Forest, XG-Boost, Obfuscation.

I. INTRODUCTION

We are well aware of internet and WWW (World Wide web) that are used generally by all people for sharing various information. Here, the sender and the receiver are not anonymous. But there is a place called Dark Web which offers anonymity to the sender and the receiver to some level. This Dark Web can be accessed by using browsers created specifically for accessing Dark web [1]. Some of the examples of such browsers are Tor, I2P, Freenet etc. Another platform that offers such anonymity is VPN (Virtual Private Network). The TOR (Onion routing) browser was originally created by the US Navy to access anonymity. But now this feature of anonymity is exploited by various criminals for illegal purposes like human trafficking, hacking, child pornography, terrorism, media piracy [2]. Therefore, classification of darknet [3] traffic becomes really important to eradicate such illegal activities.

In this research, the CIC-Darknet2020 dataset has been used to train various machine learning classifiers which is able to predict if a traffic is from the Tor or VPN or Non-Tor or Non-VPN. And these were also trained such that they classify the application of the traffic into 8 classes namely Chat, Email, P2P, VOIP, Video-steaming, Audio-streaming, Browsing and File Transfer. CIC-Darknet dataset being an unbalanced dataset, SMOTE was used for oversampling. In this research, 0%, 50% and 100% synthetic oversampling using SMOTE was done. Out of all the models that were trained, Random Forest and XG-Boost performed really well giving 99.47% and 99.50% F1-scores on average respectively for traffic classification and 88.6% and 85.8% F1-scores on average respectively for application classification. On these model, adversarial attacks [4] were simulated by obfuscating 20% features such that the model gets confused.

The remaining pair is structured as given in the following. Section II is about the related work that was done on the topic and Section III talks about the model that was proposed by us. Whereas, Section IV talks about the setup that was needed to do the experiment.. Section V deals with the results and discussion, followed by Section VI is the conclusion of the paper and possible future work that can be done in this field.

II. RELATED WORK

Many works have been done in this field and multiple research papers were published in conference and in journals as this is a crucially important field. In the study " Darknet Traffic Classification using Machine Learning Techniques"[5], the authors showed that their Random Forest model achieved 98% accuracy in classifying Darknet traffic using the CIC-Darknet2020 dataset which was the highest among all the other classifiers they trained. The authors of the study "Detection of Tor traffic using deep learning" [6] observed that their Deep neural networks were able to achieve 99.89% accuracy in classifying Tor traffic and non-Tor traffic and 95.6% accuracy in classifying specific traffic types of Tor when trained on the UNB-CIC dataset.

In another study "DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory" [2], the authors have proposed a model consisting XG-Boost was used for feature selection and for recognition, CNN-LSTM was used and achieved 96% accuracy in darknet traffic detection and an accuracy of 89% in categorization using CIC-Darknet2020 dataset. And in the study "DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning"[7], the authors showed that 2D CNN can be used for classification and characterization which achieved 86% outperforming 1D CNN models using CXVPN2016 and ISCXTor2017 datasets.

The study "Darknet traffic classification and adversarial attacks using machine learning[1]" demonstrated that Random Forest outperforms rest of the modes in darknet traffic classification and classification of traffic's

application using CIC-Darknet2020 dataset. The authors also used SMOTE technique to balance the dataset and they trained their models both on balanced and unbalanced. And the authors have also simulated adversarial attacks and demonstrated the performance degradation of Random Forest.

Through the Literature survey, we were able to find the popular dataset used, i.e, CIC-Darknet2020 dataset and was able to understand to simulate an adversarial attack by obfuscation of part of data.

III. Proposed Model

The model that was proposed in this study is the given in Fig.1 where the CIC-Darknet2020 dataset is preprocessed and the dataset being an unbalanced, is balanced using the SMOTE oversampling technique. Following which, the data is splitting into two as training and testing data, and are fed into the Machine Learning model and their performance is analyzed and compared with when adversarial attack is done.

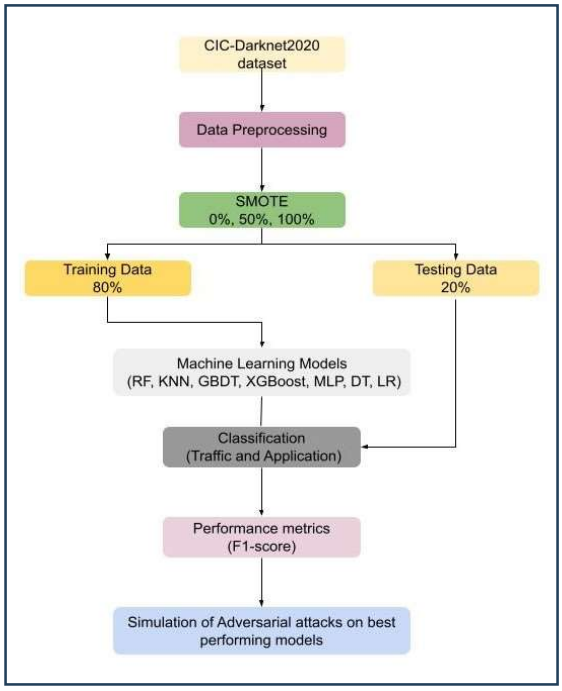


Fig. 1. Workflow diagram for traffic detection

A. Data Preprocessing

The dataset contain many NaN (Not a Number) values. Those tuples with NaN values were removed. And those with Infinity values were removed. Tuples with duplicate values were also removed. The features 'Flow Id' and 'Timestamp' were dropped. But instead of dropping 'Source IP' and 'Destination IP', their octets were converted as individual features. Following these processes, normalization and Principle component analysis(PCA) was done.

SMOTE [8] was implemented using the imblearn library in Python. Models were trained and their performance was tested with 0% SMOTE oversampling, 50% SMOTE oversampling and 100% SMOTE oversampling.

SMOTE(Synthetic Minority Over-sampling Technique) creates new samples by the process of interpolation between a minority class and its nearest neighbors. This helps the models to learn the pattern more effectively and improve the performance of the models.

B. Machine Leaning Models

The following machine learning models were trained and tested on the CIC-Darknet2020 dataset and their performance were compared:

1) Random Forest is an ensemble learning method and is widely used for tasks involving both classification and regression due to its robustness and high accuracy. It constructs multiple decision trees during training and merges their results to improve predictive performance and reduce overfitting. [9]

2) Decision Tree Classifier is a supervised ML algorithm which is utilized for regression and classification. Decision trees make decision by using the training dataset to create a tree like model. This classifier is prone to overfitting. [10]

3) K-Nearest Neighbor is an ML algorithm which works by identifying 'K' neighboring points of data in the feature space and predicts the value depending on majority class for classification and averaging value for regression tasks. [11]

4) Gradient Boosted Decision Trees is an advanced ensemble learning technique which combines the strengths of both decision tree classifier and boosting. In GBDT, decision trees are built sequentially with each new tree formed would aim to correct the errors made by the previous trees. [12]

5) Extreme Gradient Boosting is the optimization of GBDT. XGBoost has several enhancements including regularization to reduce overfitting, handling of missing data, and parallel processing for faster training of the model. [13]

6) Multi-Layer Perceptron is a type of feedforward ANN (artificial neural network) which is made up of multiple layers of nodes which includes single input layer, single or multiple hidden layers and single output layer. [14]

7) Logistic Regression is a method of statistics which is mainly used for tasks involving binary classification which involves the prediction of probability of a categorical outcome. Logistic regression applies a sigmoid function to the output so that it is transformed to an output range of 0 to 1. [15]

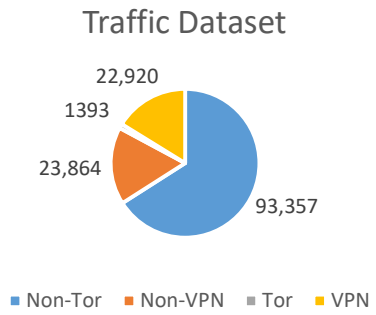
IV. EXPERIMENTAL SETUP

The models were trained in Google Colab using the T4 GPU for faster training times. The experiments were completed using the Python language. The Imblearn library (imblearn) was used to implement SMOTE oversampling technique to balance the dataset, while the package of Scikit-learn was utilized to implement the machine learning and evaluation. And the common libraries of python like Numpy and Pandas were also used data manipulation and analysis. And the xgboost library was installed and used to implement the XGBoost model.

A. Dataset

FIGURE I

DESCRIPTION OF CIC-DARKNET2020 DATASET(TRAFFIC)



The dataset of CIC-Darknet2020 [7] is used in this study. It has 158,659 hierarchically labeled samples. The top categories are Tor, VPN, Non-Tor and Non-VPN. While, the subcategories which are categorized within these top levels are audio-streaming, browsing, chat, email, file transfer, P2P, video-streaming, and VOIP. It is available publicly. It is a unbalanced dataset having 24,313 darknet samples and 1,17,221 clearnet samples. This is why oversampling is performed using SMOTE (Synthetic minority oversampling technique).

TABLE I
DESCRIPTION OF CIC-DARKNET2020 DATASET(APPLICATION)

Class	Application types	Samples
0	Audio-streaming	18,065
1	Browsing	32,809
2	Chat	11,479
3	Email	6146
4	File Transfer	11,183
5	P2P	48,521
6	Video-streaming	9768
7	VOIP	3567

B. Performance metrics

The proposed model involves classification of a network traffic into darknet and Clearnet traffic. The classification of a Clearnet traffic as darknet would lead to false alarms in security systems which might to a waste in the resources that would be involved in the investigation and the security personnel might overlook for some kind of darknet threats. On the other hand, classification of a darknet as a Clearnet would involve security risks and might to lead to data leakage as the darknet sites are designed to maintain the anonymity sensitive information may be shared or stored inappropriately, exposing users and organizations to privacy risks. So, in this case both precision and recall need to be measured equally. To balance both the measures equally we utilize the F1-score which acts as a single metric that considers both the measures equally.

C. Adversarial attacks

The data obfuscation technique involves modifying 20 percentage of a particular class by altering the values based on the mean difference between the corresponding feature of the two classes. This results in the change in the dataset and might potentially affect the results in the prediction of the model. The type of attacks involved in this process include evasion, poisoning and simultaneous attacks.

V. RESULTS

TABLE II
F1-SCORE FOR TRAFFIC CLASSIFICATION

SMOTE	0%	50%	100%
DT	98.7%	98.3%	98.4%
LR	96.4%	92.7%	91.6%
KNN	99.3%	99.1%	99.3%
RF	99.5%	99.4%	99.5%
GBDT	98.5%	97.0%	96.8%
XGBoost	99.6%	99.4%	99.5%
MLP	99.6%	99.0%	99.2%

As it can be seen clearly from TABLE III & IV, the Random Forest model and XGBoost model perform better in both the classification. In traffic classification, Random Forest has got 99.5% F1-score with 0% SMOTE oversampling, 99.4% F1-score with 50% SMOTE oversampling and 99.5% F1-score with 100% SMOTE oversampling. While, XGBoost has got 99.6% F1-score with 0% SMOTE oversampling, 99.4% F1-score with 50% SMOTE oversampling, 99.5% with 100% SMOTE oversampling. In application classification, Random Forest has got 87.8% F1-score with 0% SMOTE, 88.5% F1-score with 50% SMOTE, 89.5% F1-score with 100% SMOTE oversampling. While, XGBoost has got 86.2% F1-score with 0% SMOTE, 85.7% F1-score with 50% SMOTE oversampling, 85.5% F1-score with 100% SMOTE. Thus, adversarial attack has been simulated in Random Forest and XGBoost models.

TABLE III
F1- SCORE FOR APPLICATION CLASSIFICATION

SMOTE	0%	50%	100%
DT	83.3%	83.2%	84.2%
LR	64.5%	62.1%	60.2%
KNN	84.4%	84.7%	85.5%
RF	87.8%	88.5%	89.5%
GBDT	78.6%	77.1%	75.5%
XGBoost	86.2%	85.7%	85.5%
MLP	83.1%	83.2%	84.2%

Here, in the simulation of adversarial attack, P2P class has been obfuscated as Video-streaming and Browsing as Email. In normal circumstances, Random Forest gave F1-score as 97.2% and 92.4% for P2P class and Browsing class respectively. While, XGBoost gave F1-score as 97.4% and 92.0% for P2P class and Browsing class respectively. It can be observed from TABLE V that it has been degraded when training data is obfuscated. This shows that both the models are very vulnerable to adversarial attacks. Between the two, Random Forest performed better but not enough to be used in real life applications. And it can also be observed that both the models perform well when the testing data or both the testing and training data is obfuscated. This indicated that to mitigate the effects of adversarial attack, training data must be partially obfuscated.

TABLE IV
F1- SCORE ON P2P CLASS WHEN ADVERSARIAL ATTACK IS SIMULATED

Application types obfuscation	P2P to Video Streaming		
Models	**Training**	**Testing**	**Both**
Random Forest	63.7%	96.0%	99.0%
XG-Boost	55.6%	94.3%	99.0%

TABLE V
F1- SCORE ON BROWSING CLASS WHEN ADVERSARIAL ATTACK IS SIMULATED

Application types obfuscation	Browsing to Email		
Models	**Training**	**Testing**	**Both**
Random Forest	21.5	85.8%	78.4%
XG-Boost	03.9%	90.7%	61.7%

Efforts have been put to decrease the effects of adversarial attack by training the model partially (30%, 50%) by obfuscated data to check the difference in performance when tested with 100% obfuscation of testing data. And as it can be seen from Table VII and Table VIII, both performs amazingly well compared to when trained on normal data. But here also, Random Forest stands above XG-Boost.

FIGURE II
F1-SCORE ON P2P CLASS AFTER ADVERSARIAL ATTACK MITIGATION EFFORTS

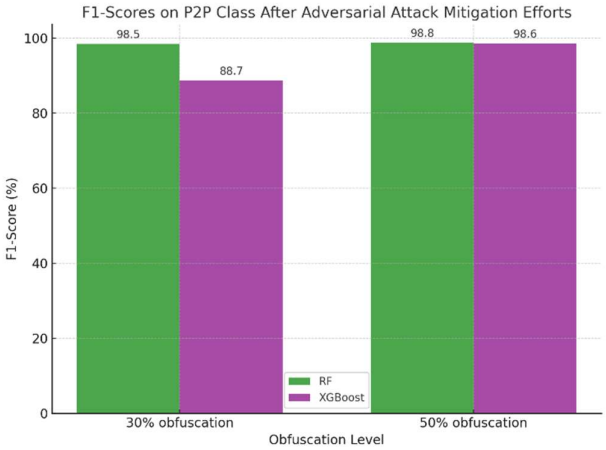
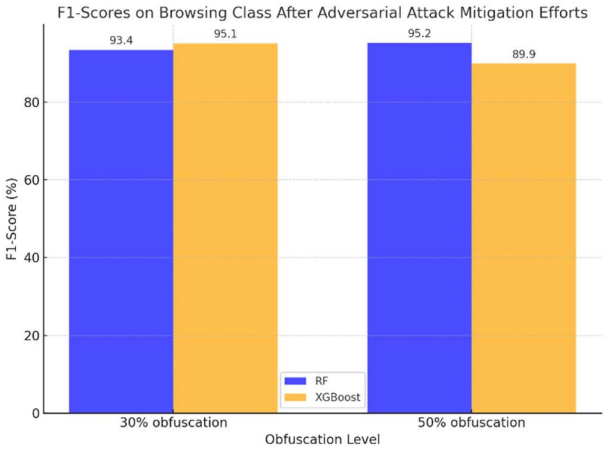


TABLE III
F1-SCORE ON BROWSING CLASS AFTER ADVERSARIAL ATTACK MITIGATION EFFORTS



VI. CONCLUSION AND FUTURE WORK

In this research, a study on classification of darknet traffic is done using CIC-Darknet2020 dataset. Since it is an imbalanced dataset, SMOTE oversampling technique was used. And it was observed that Random Forest and XGBoost model outperforms other models in both traffic and application classification. But when adversarial attack was simulated on both the models, it was easily observed that performance of both the models degraded. Out of the two, Random Forest performed better than XGBoost classifier. Also, when the efforts for mitigation of adversarial attacks were done, Random Forest showed its superiority over XGBoost. This demonstrates the importance of Random Forest for studies related to these areas of research.

This study can be further expanded by experimenting a real time classification of the darknet traffic. In this paper, only on the best performing models, adversarial attack was simulated. This can be simulated in other models and their performance can be compared. Also, Importance of each feature can be found. This helps in

further demonstrating a powerful adversarial attack by obfuscation of important features that determine the class easily. Further, more complex deep learning models can be trained and compared with these supervised machine learning models. It would be interesting to explore how those DL models perform under adversarial attacks. And finding other methods and techniques to mitigate the effects of adversarial attack would be really helpful in preventing criminal actions done with the help of Darknet.

REFERENCES

[1] N. Rust-Nguyen, S. Sharma, and M. Stamp, "Darknet traffic classification and adversarial attacks using machine learning," *Computers & Security*, vol. 127, p. 103098, 2023. https://doi.org/10.1016/j.cose.2023.103098.

[2] M. B. Sarwar, M. K. Hanif, R. Talib, M. Younas, and M. U. Sarwar, "DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory," *IEEE Access*, vol. 9, pp. 113705-113713, 2021, doi: 10.1109/ACCESS.2021.3105000.

[3] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1197-1227, 2015.

[4] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25-45, 2021..

[5] L. A. Iliadis and T. Kaifas, "Darknet traffic classification using machine learning techniques," in *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, Jul. 2021, pp. 1-4, IEEE.

[6] D. Sarkar, P. Vinod, and S. Y. Yerima, "Detection of Tor traffic using deep learning," in *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, 2020, pp. 1-8.

[7] A. Habibi Lashkari, G. Kaur, and A. Rahali, "Didarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning," in *Proceedings of the 2020 10th International Conference on Communication and Network Security*, 2020, pp. 1-13.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[10] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986

[11] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[15] D. R. Cox, "The regression analysis of binary sequences," *J. Royal Stat. Soc. Ser. B: Stat. Methodol.*, vol. 20, no. 2, pp. 215–232, 1958.