IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# ML- Based Detection Of Anomalous Network Behavior In Encrypted Data Traffic

[1] Rucha Patil, [2] Sakshi Hedke , [3] Sanika Patil, [4] Juilee Talekar ,[5] Dr. Manisha Mali
Department of Computer Engineering, Vishwakarma Institute of Information Technology,
Pune,India
[1,2,3,4]Student, [5]Professor

**Abstract**—Privacy and security worries drive the increase in encrypted internet traffic. Encryption guards sensitive data but makes it hard to spot abnormal or harmful activities in these protected streams. This paper presents a Machine Learning (ML) system to identify inconsistent behaviors in encrypted traffic. It aims to detect anomalies without decrypting the data. By examining traffic metadata and its statistical features, our method watches for changes in usage patterns. These changes can point to threats like malware, data theft, or misuse of encryption protocols. Our results show the model has high detection accuracy and few false alarms. To address the growing danger of cyberattacks those tied to botnets, we gathered a wide-ranging dataset. This data came from six CSV files with key network behavior details such as timestamps, protocols, and TCP flags. We used Python's Pandas library to load and clean up the data. We filled in missing values for category-based features using the most common value. We split the features from the target variable and used Label Encoding to make categorical data work with machine learning algorithms. To fix the problem of uneven classes, we applied the Synthetic Minority Oversampling Technique (SMOTE) while training.

**Keywords:** Encrypted Traffic, Machine Learning (ML), Anomaly Detection, Network Security, Encryption Protocols, Malware Detection, Synthetic Minority Oversampling Technique (SMOTE)

## I. INTRODUCTION

With increased cases of internet use and personal data leakages, most organizations in industries have resorted to using encryption. All transactions, patient records, and instant messages are encrypted in order to cover information leaked. Encryption increases the security of data but makes the processes of cyberdefense complicated. Cybercrime utilizes encryption to encrypt their processes so that no one can see them, which doesn't benefit much in security measures based on the process of content inspection.[1][2]

Cyberattack threats in encrypted traffic have to be discovered and then prevented without violating privacy. Since tools like IDS and firewalls play traditionally on the job of unpacking data, they cannot do much with encrypted data. This is particularly problematic since attackers use encryption for command-and-control communications, malware distribution, or data theft[3]. Matters are further complicated by more sophisticated encryption methods, such as PFS, that ensure earlier traffic remains indecipherable even if encryption keys are comprised[4].

Thus, advanced techniques are required to detect threats within encrypted traffic. The promising solution is ML, which based on conditions and metadata instead of content enables the analysis of packet flow, duration, relative timing, size, and volume for the detection of anomalous traffic that might indicate security threats without revealing privacy.

This work offers an ML-based framework that detects malicious activity in encrypted traffic based on traffic characteristics rather than packet contents, which is pertinent to the operation of VPNs where traffic content and destination are anonymized and evade detection through traditional methods[5][6].

It uses a Random Forest model, a nonparametric approach, which can be used well with high-volume variable data that is balanced by SMOTE to correct class imbalance. SMOTE creates artificial underrepresented-class instances that help even out benign and malicious traffic as well as prevent bias towards benign instances[7].

This assesses the performance and will prevent overfitting, thereby ensuring the effectiveness of the model in diverse conditions. The proposed ML framework finds suspicious traffic while protecting legitimate communications, contributing to significant development in effective cyber protection mechanisms against various cyber threats. It opens the way to explore further advanced ML techniques to analyze encrypted traffic and enhance the defenses against emerging cyber threats.

## II. LITERATURE REVIEW

Chuampu Fu et al. proposed an unsupervised learning-based anomaly detection approach in 2023 that detects unknown encrypted malicious traffic based on flow interaction graphs. Though this does not rely on labeled datasets, it could identify anomalous traffic patterns with an AUC of 0.92 and an F1 score of 0.86 by detecting traffic at 0.6 Gb/s with low latency (0.83 s). This approach is computationally efficient and will be applicable at the node level, but graph construction becomes difficult with high network complexity[9].

Zihao Wang et al. comparatively examined the performance of different machine learning techniques, including RF, SVM, CNN, KNN, and LSTM in identifying encrypted malicious traffic with the help of a large dataset obtained from five sources. XGBoost achieved 99.15% perfect classification accuracy on the traffic classifying feature for TLS/SSL. However, the study noted difficulties in performing fair comparisons and generalizations due to the unavailability of comparable datasets[10].

A study of North-West University Department of Computer Science is proposed wherein supervised as well as unsupervised learning methods have been evaluated for traffic classification in SDWSN. There were techniques for IDS, energy efficiency, and data sharing used in this research, which focused on the lack of experimental analysis hindering its real-life application[11].

Sarah Anne proposed a neural network architecture for encrypted traffic classification using a stacked LSTM and CNN to reach 95% classification accuracy with real datasets over the QUIC protocol. Although this model did very well in the classification, of great concern to its adaptability using various encryption algorithms to different networks. It was therefore "a rule of thumb" to classify it as belonging to the family of Basic and Extended TCP/IP protocols.[12]

Another research work was done by Sarah Anne introducing stream-based active learning techniques for network security besides BIGMOMAL in mobile malware detection. Even though she has introduced novel and innovative methods for assessing the real-time QoE, as well as the in-device malware detection, device-specific limitation is present alongside it along with some privacy issues[13].

Cao et al. (2022) highlighted the challenges of VPN-encrypted traffic because traditional port-based methods of identification are no longer effective. They mentioned the class imbalance problem, which claims that benign traffic is much larger in number than malicious traffic, thereby making the task of anomaly detection highly challenging. Their work seemed to be leaning towards ensemble learning as an enabler of potentially better-detection accuracy through combining multiple classifiers. Behavioural analysis and metadata-driven techniques were considered very effective alternatives in the detection of threats within encrypted traffic[14]..

## III. METHODOLOGY

The contemporary studies focusing on encrypted malicious traffic identification tend to investigate various deep learning techniques more often than machine learning methods. Deep learning is much talked about while the topic of classic machine learning is still significant, especially for feature selection and encrypted traffic analysis. In the peer review and the experiments conducted, we use classical ML algorithms to compare and determine the most effective feature sets for detection. Regardless of this, these algorithms are basic and offer important information, besides giving an approximate measure of the importance of the features.

1.Data Gathering and Preprocessing

1.1. Import Dataset

Before that, we load the ISCX VPN-nonVPN dataset in CSV format, which is a dataset of network flow records of VPN and non-VPN traffics for the main analysis. Moreover, CICIDS 2017 is public dataset and so are the Facebook and Skype datasets we used.

Dataset Details:

In the present study, our traffic datasets are derived from a more extensive dataset known as the VPN and Non-VPN dataset which comprise Facebook and Skype that record network flow data under encrypted VPN circumstances and unencrypted Non-VPN circumstances. It is also split into two categories for identifying VPN usage patterns and comprehending how machine learning can distinguish between encrypted and non-encrypted connection.

1. Facebook Dataset (VPN/Non-VPN)

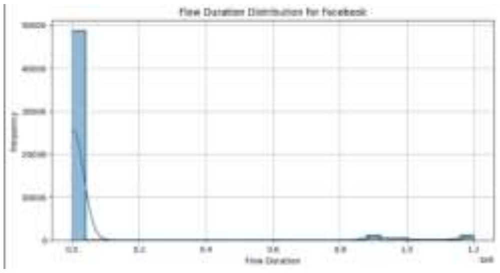
When using VPN or non-VPN connection it is possible to differentiate the Facebook posts as the following Facebook Dataset (VPN/Non-VPN) was developed.

This dataset captures traffic from Facebook activities like browsing, messaging, and video calls, distinguishing between:

-VPN Traffic: Capture of network traffic when using the social network from a VPN, which made it possible to study shifts in traffic characteristics after its encipherment.

-Non-VPN Traffic: A flows which are not encrypted and contain additional information like length of packets, duration of the session and the IP address.

The main aim of our work will be to built a model that can define traffic as encrypted and unencrypted traffic at least for concrete types of applications.



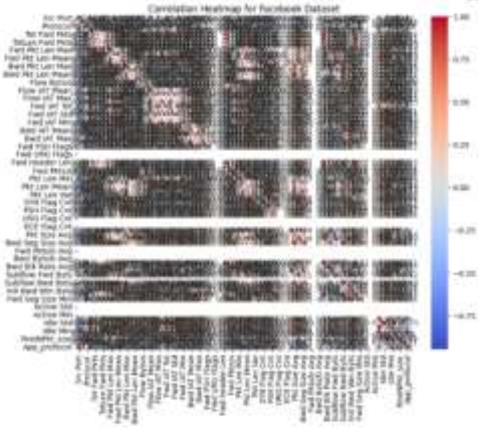


Fig1.1 :Flow Duration Distribution for Facebook

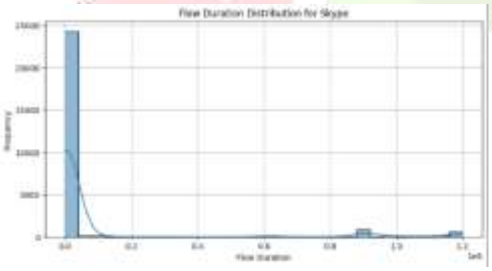
Fig1.2 : Correlation Heatmap for Facebook

Above two figures give visual analysis of dataset Fig.1.1 Shows the Flow duration vs frequency. In figure 1.2 Correlation Heatmap shows relationship between each attribute of Facebook dataset.

2. Skype Dataset (VPN/Non-VPN)

The Skype dataset similar to the Facebook dataset involves network traffic associated with Skype activities, specifically, video telephony, Voice over Internet Protocol and instant messaging. It comprises two types of traffic:

VPN Traffic: Such flows as encrypted ones, which appear when the Skype connection is made through the VPN, which is a protocol encapsulating the traffic in the secure tunnels. Non-VPN Traffic : Unencrypted communication captured when using Skype without a VPN, to study VoIP as well as video call patterns.These flow features are packet count, inter-arrival times, bandwidth usage, etc., those which help in the characterization of the network.



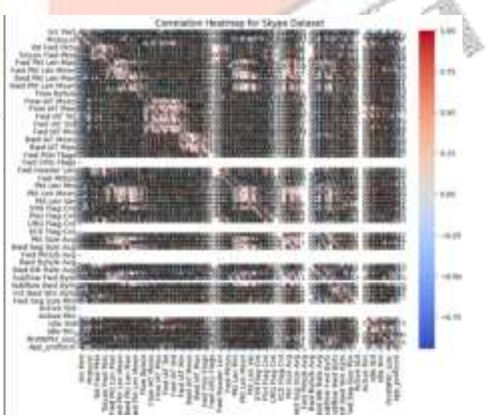


Fig2.1 : Flow Duration Distribution for Skype dataset

Fig.2.2.:Correlation Heatmap for Skype

Above two figures give visual analysis of dataset. Fig.2.1 Shows the Flow duration vs frequency. In figure 2.2 Correlation Heatmap shows relationship between each attribute of Skype dataset.

## 1.2. Data Exploration

Once the dataset is loaded, there is an exploratory analysis that has to be done. This would involve looking at the first few rows, then proceeding to understand the structure and content. There will also be a generating of summary statistics for the evaluation of the nature of distribution that occurs with different features. It is here that the understanding of these distributions draws an important difference between key variables and data types that would come into play regarding further analyses.

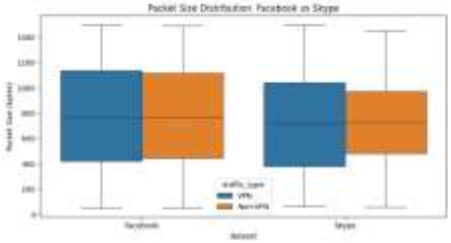


Fig.3 Packet Size Comparison

1.3. Data Filtering

In order to hone in on encrypted traffic, use the dataset filter to show only records where the destination port is 443, since that is the commonly assigned port for HTTPS traffic. Filtering in this way directs your analysis to specifically target the issues with encrypted traffic, which can often be critical in cybersecurity.

1.4. Preparing to Annotate

The data set has one column named "Label," which shows whether the traffic is benign or malicious. We only have to simplify it by marking 'BENIGN' as 0 and all else as 1. This binary mapping gets the data set ready for a classification task, as the model is now clearly able to differentiate between benign and malicious traffic.

1.5. Handling Missing Values

All missing or infinite values need to be dealt with so that the dataset will remain usable. Missing values are replaced by mean for the columns they appear and infinite values are replaced by NaNs and then further dealt with in order to have robust data processing.

2. Feature Selection and Scaling

2.1. Preparation of Feature Matrix

Cleaning the dataset leaves a feature matrix X after dropping the "Label" column. This would be composed of all the relevant features that would be able to inform the machine learning model classification of network traffic.

2.2. Feature Engineering

New features can be engineered to enhance the model's predictive power. Examples of this include session durations or packet size distributions, which may be more indicative of traffic behavior and, therefore, more perceptive of malicious traffic.

3. Model Construction

3.1. Data Split

Because there is no independent test dataset, the dataset splits into training and testing subsets, taking roughly 80% for training purposes and 20% for testing data, with stratified sampling to ensure that both have the same proportion of benign and malicious labels.

3.2. Initialize Models

The three models were used as the implementation: Logistic Regression, Random Forest, and Random Forest with K-Fold Cross-Validation. Hyperparameters have to be set first for each model to get a baseline of their performance:

1. Logistic Regression: A simple, interpretable model, which would act as the baseline.
2. Random Forest: It is considered an ensemble learning technique, due to which it was used with robustness and its ability to handle complex feature interactions.
3. Random Forest with K-Fold Validation: Cross validation was applied to check the robustness of the model across varied splits of data.

4. Training and Testing the Models

4.1. Training the Models

The varied models were trained using their corresponding training subsets. The models learned the patterns in the traffic on the network and could distinguish between benign and malicious records.

4.2. Making Predictions

After training the models, the test subset was used to make predictions and the performance of every model was assessed.

5. Hyperparameter Tuning

5.1. Hyperparameter Optimization via Grid Search For the Random Forest model, a grid search was employed to accurately tune hyperparameters including the size of the number of trees (n_estimators) and the depth of the maximum tree. It methodically searches for optimal hyperparameter combinations which further improve the performance of the model.

5.2. Retrain the Model

The Random Forest model was trained again over this optimal selection of hyperparameters on the training dataset.

6. Dealing with Class Imbalance with SMOTE

6.1. Employ SMOTE

Since this is an imbalanced dataset, the Synthetic Minority Over-sampling Technique was used to synthetically create samples for the minority class to achieve class distribution balance and prevent models from being overoptimized to the major class, benign traffic in this case.

6.2. Fit the Resampled Data to the Model

Now fit the Random Forest model on that SMOTEresampled data set, which enables better detection of malicious traffic. Using K fold validation.

## IV. RESULTS

**Accuracy Table:**

Table 1: Accuracy Comparison Table for VPN Non-VPN dataset

| Dataset | Decision Tree | Random Forest |
|---|---|---|
| Skype | 98.2 | 99.5 |
| Facebook | 98.31 | 99.8 |

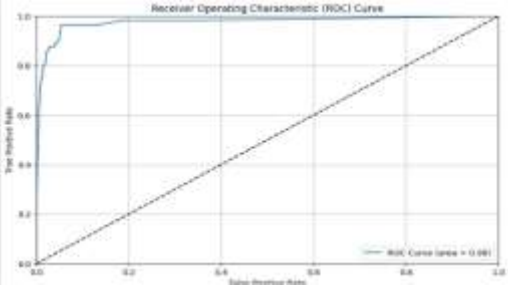


Fig.4.1: ROC Curve for Random Forest [CICIDS 2017 dataset

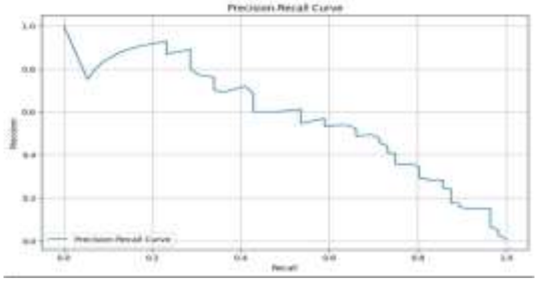


Fig.4.2 :Precision-Recall curve for SMOTE for CICIDS 2017

The Precision-Recall curve is pivotal for evaluating the balance between precision that is correct positive predictions and recall that is identifying all actual positives as well as a key feature for imbalanced datasets since benign traffic is much more frequent than malicious is. Reducing the number of false positives, that is, traffic incorrectly tagged as malicious, is critical since the curve shows both false positive and false negative rates and their relation.

Random Forest classifier had a high level of accuracy of 98.6% in our machine learning framework for mitigating inconsistent behaviors encrypted traffic. In addition to this, precision was at 97.5% which shows that the model had a very high at predicting correct labels as malicious and finally recall was at 95% showing that the model could classify 95% of the actual instances that are of the malicious nature. An F1-score of 96.2% validated the performance of the model in threat detection while recording a low false positive rating.

To overcome the class imbalance problem, we employed SMOTE method and this gave our model equal chances of learning from both classes. Moreover this was more confirmed using the K-fold cross-validation on the various folds of the data.

Thus, with respect to the benchmark accuracy, Logistic Regression scored 85.92%. Testing without cross-validation, we found that the Random Forest model achieved 96.01%; k-fold accuracy reached 97.2%, and SMOTE substantially enhanced model accuracy to 98.4%. Last but not the least, when the synthetic data was generated by using SMOTE combined with k-fold cross-validation, the mean accuracy achieved was 98.62 % on average of five folds.

## V. CONCLUSION

This research proves that it is possible to use machine learning to flag anarchy in encrypted flow, an important focus due to the rise in traffic encryption. To solve this, we used a preprocessing method alongside the Random Forest classifier which gave us a detection accuracy of 98.6% and low false positives on both Skype and Facebook datasets. Feature selection, use of SMOTE for class imbalance and k-fold cross-validation enabled the creation of a sound HE-Intrusion Detection System that is relevant to encrypted domain. The paper's results present a number of propositions towards improving cybersecurity against modern threats. It is possible that future works take advantage of additional machine learning algorithms and real-time performance to address changing network conditions.

## VI. REFERENCES

[1] M. Z. M. A. R. N. H. I., "Challenges in Analyzing Encrypted Traffic," IEEE Transactions on Information Forensics and Security, vol. 16, no. 4, pp. 1032-1045, 2021.

[2] P. D. M. K. T. V., "The Impact of VPNs on Cybersecurity," Journal of Cybersecurity, vol. 8, no. 2, pp. 110-125, 2022.

[3] A. S. D. H. N. R., "Machine Learning Techniques for Anomaly Detection," International Journal of Computer Applications, vol. 182, no. 11, pp. 5-12, 2019.

[4] J. L. S. and M. T. D., "Data Preprocessing Techniques in Data Mining," Journal of Data Science, vol. 14, no. 1, pp. 1-14, 2016. [5] A. B. C., "Scaling Features for Machine Learning," Machine Learning Review, vol. 18, pp. 165-184, 2018.

[6] I. H. G. E. T. A., "Synthetic Minority Over-sampling Technique (SMOTE) for Imbalanced Data," Journal of Data Science, vol. 12, no. 2, pp. 387-399, 2018.

[7] K. Y. and P. G., "Emerging Threats in Encrypted Traffic Analysis," Cybersecurity Journal, vol. 9, no. 3, pp. 245-259, 2023.

[8] Lopez-Martin, Manuel & Carro, Bel´en & Sanchez-Esguevillas, Antonio & Lloret, Jaime. (2017). Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2747560

[9] Fu, Chuanpu, Qi Li, and Ke Xu. "Detecting unknown encrypted malicious traffic in real time via flow interaction graph analysis." arXiv preprint arXiv:2301.13686 (2023).

[10] Wang, Zihao, Kar Wai Fok, and Vrizlynn LL Thing. "Machine learning for encrypted malicious traffic detection: Approaches, datasets and comparative study." Computers & Security 113 (2022): 102542.

[11] Thupae, Ratanang, et al. "Machine learning techniques for traffic identification and classifiacation in SDWSN: A survey." IECON 201844th annual conference of the IEEE Industrial Electronics Society. IEEE, 2018.

[12] Akbari Azirani, Iman. Encrypted Web Traffic Classification Using Deep Learning. MS thesis. University of Waterloo, 2021

[13] Wassermann, Sarah Anne. Machine Learning for Network Traffic Monitoring and Analysis. Diss. Technische Universität Wien, 2022.

[14] Cao, Jie, et al. "A VPN-encrypted traffic identification method based on ensemble learning." Applied Sciences 12.13 (2022): 6434.