



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: <https://doi.org/10.22214/ijraset.2024.61288>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Detection of Virtual Private Networks Encrypted Traffic using Machine and Ensemble Learning Algorithms

Sujatha S¹, Shivani P R², Sakthi Priya C³, Ahila R⁴

^{1, 2, 3}UG Students, ⁴Assistant Professor, Department of Computer Science and Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India.

Abstract: *The widespread adoption of Virtual Private Networks (VPNs) for secure communication over public networks has highlighted the need for accurate detection of encrypted network traffic. This study proposes a comprehensive approach utilizing machine and ensemble learning algorithms to address this detection challenge. Leveraging the ISCX VPN dataset, the project aims to develop a robust system capable of distinguishing between VPN and non-VPN traffic with high accuracy. Key stages include data acquisition, preprocessing, feature extraction, and model training using Decision Tree, Naive Bayes, Random Forest, and Adaboost algorithms. Through rigorous experimentation, the most effective algorithmic approach is identified. The project also offers insights into the diverse nature of encrypted network communication and its implications for network security. The anticipated outcomes include improved understanding and management of VPN traffic, enhancing overall network security and performance. This research contributes to advancing network security practices by offering practical solutions for encrypted traffic detection in real-world settings.*

Index Terms: *Virtual Private Network (VPN), Machine Learning, Ensemble Learning, Traffic Analysis, ISCX VPN Dataset.*

I. INTRODUCTION

The advent of Virtual Private Networks (VPNs) has revolutionized the realm of secure communication across public networks, presenting a pivotal paradigm shift in modern connectivity. VPNs facilitate encrypted connections, safeguarding sensitive data during transmission over untrusted networks like the internet. As such, they have become indispensable tools for ensuring privacy, security, and data integrity in an era characterized by escalating cyber threats and privacy concerns. This study delves into the intricate domain of VPN encrypted traffic detection, employing cutting-edge machine and ensemble learning algorithms to tackle the challenges inherent in accurately discerning VPN from non-VPN traffic. By harnessing the power of these advanced techniques and leveraging rich datasets, the project endeavors to forge innovative solutions that enhance network security and bolster data protection measures.

A. Existing Challenges

The detection of encrypted network traffic poses several formidable challenges, chief among them being the ability to accurately distinguish between VPN and non-VPN traffic amidst the complex and dynamic nature of modern network environments.

B. Disadvantages of Existing Systems

The disadvantages of existing Detection of Virtual Private Network systems include:

- 1) **Limited Accuracy:** Existing systems often struggle to achieve high levels of accuracy in detecting encrypted traffic, leading to misdetections and false positives.
- 2) **Scalability Issues:** Some systems may lack scalability, rendering them inefficient in handling large volumes of traffic or adapting to evolving network architectures.
- 3) **Dependency on Handcrafted Rules:** Many existing systems rely heavily on handcrafted rules or heuristics, which may be ineffective in capturing the nuanced characteristics of encrypted traffic.
- 4) **Vulnerability to Evasion Techniques:** Certain evasion techniques, such as traffic obfuscation or encryption protocol manipulation, can bypass detection mechanisms employed by existing systems, compromising their efficacy.

These shortcomings underscore the need for advanced technologies and methodologies to enhance the effectiveness of Detection of Virtual Private Networks encrypted traffic using Machine and Ensemble learning algorithms

C. Project Objectives

The main objective of this project is to develop a robust and accurate system for the detection of encrypted network traffic into Virtual Private Network (VPN) and non-VPN categories using machine and ensemble learning algorithms. With the pervasive adoption of VPNs for ensuring secure communication over public networks, there is a pressing need to accurately distinguish between VPN and non-VPN traffic to enhance network security measures. The project aims to address several key challenges in encrypted traffic detection. Firstly, as the use of VPNs continues to rise, understanding and managing the distinct characteristics of VPN traffic become paramount for network administrators, security professionals, and researchers. The existing methods for detecting encrypted traffic often lack accuracy and efficiency, leading to potential security vulnerabilities and operational inefficiencies. The project seeks to overcome these challenges by leveraging machine and ensemble learning algorithms to analyze encrypted network traffic comprehensively. By drawing upon the ISCX VPN dataset, encompassing a diverse range of network flows, the project endeavors to develop a detection system capable of accurately discerning between VPN and non-VPN traffic. The ultimate goal is to provide network administrators and security analysts with a reliable tool for proactive monitoring, detection, and mitigation of potential security threats posed by encrypted traffic. By improving the accuracy and efficiency of encrypted traffic detection, the project aims to enhance overall network security measures and ensure the integrity and confidentiality of data transmitted over public networks.

D. Key Features and Innovations

Key Features and Advantages of the Detection of Virtual Private Network systems:

- 1) Integration of advanced machine and ensemble learning algorithms for enhanced detection accuracy.
- 2) Utilization of the ISCX VPN dataset to train and evaluate detection models.
- 3) Comprehensive analysis and evaluation of multiple algorithms, including Decision Tree, Naive Bayes, Random Forest, and Adaboost.
- 4) Emphasis on scalability, adaptability, and efficiency to address real-world challenges in encrypted traffic detection.

II. LITERATURE SURVEY

This application was developed based on the following papers:

- 1) "A Systematic Approach of Feature Selection for Encrypted Network Traffic classification" by D. McGaughey et al.:

In this paper, McGaughey et al. present a systematic approach to feature selection for classifying encrypted network traffic. They utilize the fast orthogonal search (FOS) algorithm to identify a subset of features with discriminative power from a large feature set derived from the data. Subsequently, a k-nearest neighbor (kNN) classifier is employed for traffic classification using the selected features. The FOS algorithm efficiently identifies a subset of features, leading to notable improvements in classification accuracy compared to using an arbitrary feature set. The approach not only enhances accuracy but also reduces computation time, making it scalable and applicable in real-world scenarios where timely classification is crucial. [1]

- 2) "Comparison of Machine-Learning Algorithms for Classification of VPN Network Traffic Flow Using Time-Related Features" by S. Bagui et al.:

Bagui et al. propose a framework for classifying VPN or non-VPN network traffic using time-related features and machine-learning techniques. The study compares six classification models, including logistic regression, support vector machine, Naïve Bayes, k-nearest neighbor, Random Forest, and Gradient Boosting Tree classifiers. Recommendations are provided based on optimized Random Forest and Gradient Boosting Tree models, which demonstrate high accuracy and low overfitting. The research contributes to advancing the understanding and capability of classifying encrypted network traffic, thereby enhancing network security measures. [2]

- 3) "A VPN-Encrypted Traffic Identification Method Based on Ensemble Learning" by Ying Cui et al.:

Cui et al. propose a method for identifying VPN-encrypted traffic based on ensemble learning to address challenges such as feature redundancy, data class imbalance, and low identification rates. The method involves feature selection using minimum Redundancy Maximum Relevance (mRMR), enhancing the XGBoost identification model with focal loss function for data class imbalance, and optimizing ensemble learning model parameters using optimal Bayesian. Experimental results demonstrate superior outcomes compared to existing methods, highlighting the method's effectiveness in identifying VPN-encrypted traffic and contributing to the advancement of network security practices. [3]

4) “Research on SDN Intrusion Detection Based on Online Ensemble Learning Algorithm” by Z. Lin et al.:

Lin et al. propose an adaptive SDN intrusion detection model based on online ensemble learning algorithm to address challenges posed by unbalanced data streams in Software-Defined Networking (SDN) environments. The model enhances the bagging algorithm to mitigate the impact of data stream imbalance on integrated classifier performance by incorporating features such as unbalanced detection, dynamic penalty factor adjustment, and selection integration. Experimental evaluation using the NSL-KDD dataset demonstrates enhanced detection accuracy, particularly in recognizing unknown intrusion behavior, contributing to the effectiveness of SDN intrusion detection systems and advancing network security practices in SDN environments. [4]

III. METHODOLOGY

The methodology employed in this study aims to develop a robust system for classifying encrypted network traffic into Virtual Private Network (VPN) and non-VPN categories using machine and ensemble learning algorithms. The primary objective is to enhance the accuracy and efficiency of VPN traffic classification, thereby contributing to improved network security measures.

The main objective of this project is to create an efficient and precise system that improves the accuracy of network traffic detection, ultimately enhancing network security measures and ensuring the efficient management of encrypted traffic in both VPN and non-VPN environments using machine and ensemble learning algorithms.

The project addresses the challenge of accurately classifying encrypted network traffic into VPN and non-VPN categories. As the adoption of VPNs continues to rise, understanding and managing the distinct characteristics of VPN traffic become paramount for network administrators, security professionals, and researchers alike.

1) *Data Acquisition*

Obtain the ISCX VPN dataset, a representative collection of encrypted network traffic.

2) *Data Preprocessing*

Refine the dataset by addressing missing values, outliers, and noise.

Normalize features to standardize their range and facilitate effective model training.

3) *Feature Extraction*

Identify and select relevant features from the dataset that contribute to VPN traffic detection.

4) *Data Splitting*

Partition the dataset into separate training and testing sets to ensure accurate evaluation of model performance.

5) *Model Training with Machine and Ensemble Learning Algorithms:*

Train the following algorithms on the pre-processed and feature-extracted dataset:

- Decision Tree
- Naive Bayes
- Random Forest
- Adaboost

6) *Model Validation:*

Validate the trained models to ensure they generalize well to unseen data and identify any overfitting or underfitting issues.

7) *Model Testing*

Evaluate the performance of trained models on a dedicated testing dataset not encountered during training or validation.

8) *Generating and Calculating Evaluation Metrics*

Quantitatively assess model performance using metrics such as accuracy, precision, recall, and F1 score.

C. Feature Extraction

- 1) *Identification of Relevant Features:* Analyze the dataset to identify features that contribute significantly to distinguishing between VPN and non-VPN traffic.
- 2) *Dimensionality Reduction:* Utilize techniques such as principal component analysis (PCA) or feature selection algorithms to reduce the number of features while preserving relevant information.
- 3) *Feature Engineering:* Create new features or transform existing ones to enhance the discriminatory power of the dataset, potentially using domain knowledge or heuristics.
- 4) *Validation of Extracted Features:* Validate the extracted features through exploratory data analysis and statistical tests to ensure their suitability for the detection task.

```
1 Final_dataframe.head()
```

	Fwd Pkt Len Min	Bwd Pkt Len Max	Bwd Pkt Len Mean	Bwd Pkt Len Std	Pkt Len Mean	Pkt Len Std	Down/Up Ratio	Pkt Size Avg	Bwd Seg Size Avg	Init Bwd Win Bytes	Label
0	0.0	0.008010	0.038076	0.033726	0.066290	0.035647	0.0	0.083497	0.038076	0.000000	1.0
1	0.0	0.004503	0.012576	0.015777	0.022997	0.021730	0.2	0.025016	0.012576	0.000458	0.0
2	0.0	0.006105	0.030566	0.023741	0.046871	0.025023	0.0	0.069525	0.030566	0.000000	1.0
3	0.0	0.014982	0.077375	0.055249	0.121102	0.058231	0.0	0.179634	0.077375	0.000000	1.0
4	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.005356	1.0

Fig 4: Feature Extraction

D. Data Splitting

- 1) *Training and Testing Set Division:* Split the dataset into training and testing sets using a predefined ratio (e.g., 70% training, 30% testing) to ensure unbiased evaluation of model performance.
- 2) *Randomization:* Randomly shuffle the dataset before splitting to prevent any inherent ordering effects from influencing the results.
- 3) *Stratification:* Ensure that the distribution of classes is maintained in both the training and testing sets to prevent bias, particularly in imbalanced datasets.

E. Model Training with Machine and Ensemble Learning Algorithms:

This project employs a range of Machine and Ensemble Learning algorithms, including Decision Tree, Naive Bayes, Random Forest, and Adaboost, for the detection of encrypted network traffic. The models undergo training on pre-processed and feature-extracted datasets to discern underlying patterns distinguishing VPN from non-VPN traffic. During training, the models adjust their internal parameters to optimize performance based on identified features.

Training time for each model is recorded as follows:

- 1) Decision Tree: 0.9541864139020537
- 2) Naive Bayes: 0.8560031595576619
- 3) Random Forest: 0.9541864139020537
- 4) Adaboost: 0.9119273301737757

Following training, model validation is conducted to ensure generalizability to unseen data. Validation assesses performance on a separate dataset, identifying overfitting or underfitting issues for subsequent fine-tuning.

F. Model Testing, Evaluation Metrics, Result Analysis:

Model testing involves evaluating trained models on a dedicated testing dataset not encountered during training or validation. This simulates real-world scenarios, providing insights into overall performance and the ability to classify encrypted traffic accurately.

The performance of each model during testing is as follows:

- 1) Decision Tree: 97%
- 2) Naive Bayes: 66%
- 3) Random Forest: 98%
- 4) Adaboost: 93%

Results analysis entails generating various evaluation metrics, such as accuracy, precision, recall, and F1 score, to quantitatively assess performance.

Visualization of results, including charts, graphs, and confusion matrices, aids in comprehending algorithmic strengths and weaknesses, enabling stakeholders to make informed decisions.

This detailed system model outlines the step-by-step process involved in building a detection system for encrypted network traffic using machine and ensemble learning algorithms, ensuring thoroughness and clarity in implementation.

V. FUTURE WORK

Moving forward, there are several avenues for further exploration and enhancement of the proposed detection system for encrypted network traffic. Future research could focus on expanding the scope of the system to accommodate emerging encryption techniques and evolving network protocols. Moreover, exploring the application of deep learning techniques, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), may offer insights into more complex patterns in encrypted traffic data, potentially improving detection performance. Furthermore, research efforts could delve into optimizing the system's scalability and efficiency to handle large-scale network environments effectively.

VI. CONCLUSION

In conclusion, this project has showcased the efficacy of machine and ensemble learning algorithms in accurately classifying encrypted network traffic into VPN and non-VPN categories. Through rigorous experimentation and leveraging the ISCX VPN dataset, Random Forest emerged as the top performer with 98% accuracy, closely followed by Decision Tree at 97%. These findings underscore the robustness of ensemble learning techniques in handling complex detection tasks. The project contributes to advancing network security by offering a reliable detection system for encrypted traffic, enabling proactive threat mitigation and regulatory compliance. Future endeavors may focus on expanding the system's scope, integrating anomaly detection techniques, and optimizing scalability to address evolving challenges in encrypted traffic analysis and bolster network security measures.

REFERENCES

- [1] Cao, J.; Yuan, X.-L.; Cui, Y.; Fan, J.-C.; Chen, C.-L. A VPN-Encrypted Traffic Identification Method Based on Ensemble Learning. *Appl. Sci.* 2022, 12, 6434. <https://doi.org/10.3390/app12136434>.
- [2] S. Bagui, X. Fang, E. Kalaimannan, S.C. Bagui, J. Sheehan, "Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features", *Journal of Cyber Security Technology*, 1 (2) (2017), pp. 108-126
- [3] Z. Cao, G. Xiong, Y. Zhao, Z. Li, L. Guo "A survey on encrypted traffic classification", Springer (2014), pp. 73-81
- [4] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 17-36; *JMLR Workshop and Conference Proceedings*.
- [5] W. Wang, M. Zhu, X. Zeng, X. Ye, Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning", *IEEE* (2017), pp. 712-717.
- [6] W. Wang, M. Zhu, J. Wang, X. Zeng, Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks", *IEEE* (2017), pp. 43-48
- [7] D. McGaughey, T. Semeniuk, R. Smith, S. Knight, "A systematic approach of feature selection for encrypted network traffic classification", *IEEE* (2018), pp. 1-8
- [8] O. Salman, I.H. Elhajj, A. Chehab, A. Kayssi, "A multi-level internet traffic classifier using deep learning", *IEEE* (2018), pp. 68-75
- [9] M. Lotfollahi, M.J. Siavoshani, R.S.H. Zade, M. Saberian "Deep packet: A novel approach for encrypted traffic classification using deep learning", *Soft Comput.* 24 (3) (2020), pp. 1999-2012
- [10] S. Cui, B. Jiang, Z. Cai, Z. Lu, S. Liu, J. Liu, "A session-packets-based encrypted traffic classification using capsule neural networks", *IEEE* (2019), pp. 429-436.
- [11] Sengupta, S.; Chowdhary, A.; Sabur, A.; Alshamrani, A.; Huang, D.; Kambhampati, S. A survey of moving target defenses for network security. *IEEE Commun. Surv. Tutor.* 2020, 22, 1909–1941.
- [12] Tahaei, H.; Afifi, F.; Asemi, A.; Zaki, F.; Anuar, N.B. The rise of traffic classification in IoT networks: A survey. *J. Netw. Comput. Appl.* 2020, 154, 102538.
- [13] Pacheco, F.; Exposito, E.; Gineste, M.; Baudoin, C.; Aguilar, J. Towards the deployment of machine learning solutions in network traffic classification: A systematic survey. *IEEE Commun. Surv. Tutor.* 2018, 21, 1988–2014.
- [14] Masdari, M.; Khezri, H. A survey and taxonomy of the fuzzy signature-based Intrusion Detection Systems. *Appl. Soft Comput.* 2020, 92, 106301.
- [15] Khatouni, A.S.; Heywood, N.Z. How much training data is enough to move a ML-based classifier to a different network? *Procedia Comput. Sci.* 2019, 155, 378–385.
- [16] Juma, M.; Monem, A.A.; Shaalan, K. Hybrid end-to-end VPN security approach for smart IoT objects. *J. Netw. Comput. Appl.* 2020, 158, 102598.
- [17] Aceto, G.; Ciunzo, D.; Montieri, A.; Pescapé, A. Toward effective mobile encrypted traffic classification through deep learning. *Neurocomputing* 2020, 409, 306–315.
- [18] Bu, Z.; Zhou, B.; Cheng, P.; Zhang, K.; Ling, Z.-H. Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models. *IEEE Access* 2020, 8, 132950–132959.
- [19] Cao, Z.; Xiong, G.; Zhao, Y.; Li, Z.; Guo, L. A Survey on Encrypted Traffic Classification; *International Conference on Applications and Techniques in Information Security*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 73–81.

- [20] Aceto, G.; Ciunzo, D.; Montieri, A.; Pescape, A. Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges. *IEEE Trans. Netw. Serv. Manag.* 2019, 16, 445–458.
- [21] Rezaei, S.; Liu, X. Deep learning for encrypted traffic classification: An overview. *IEEE Commun. Mag.* 2019, 57, 76–81.
- [22] Handa, A.; Sharma, A.; Shukla, S.K. Machine learning in cybersecurity: A review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2019, 9, e1306.
- [23] Ribeiro, V.H.A.; Reynoso-Meza, G. Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. *Expert Syst. Appl.* 2020, 147, 113232.
- [24] Meng, F.; Cheng, W.; Wang, J. Semi-supervised Software Defect Prediction Model Based on Tri-training. *KSII Trans. Internet Inf. Syst. (TIIS)* 2021, 15, 4028–4042.
- [25] Xibin, D.; Zhiwen, Y.; Wenming, C.; Yifan, S.; Qianli, M. A survey on ensemble learning. *Front. Comput. Sci.* 2020, 14, 241–258.
- [26] Paxson, V. Empirically derived analytic models of wide-area TCP connections. *IEEE/ACM Trans. Netw.* 1994, 2, 316–336.
- [27] Sen, S.; Spatscheck, O.; Wang, D. Accurate, scalable in-network identification of p2p traffic using application signatures. In *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, USA, 17 May 2004; pp. 512–521.
- [28] Lotfollahi, M.; Siavoshani, M.J.; Zade, R.S.H.; Saberian, M. Deep packet: A novel approach for encrypted traffic classification using deep learning. *Soft Comput.* 2020, 24, 1999–2012.
- [29] Dutt, I.; Borah, S.; Maitra, I.K. Multiple Immune-based Approaches for Network Traffic Analysis. *Procedia Comput. Sci.* 2020, 167, 2111–2123.
- [30] Yao, Z.; Ge, J.; Wu, Y.; Lin, X.; He, R.; Ma, Y. Encrypted traffic classification based on Gaussian mixture models and Hidden Markov Models. *J. Netw. Comput. Appl.* 2020, 166, 102711.
- [31] Chang, L.; Zigang, C.; Gang, X.; Gaopeng, G.; Siu-Ming, Y.; Longtao, H. MaMPF: Encrypted Traffic Classification Based on Multi-Attribute Markov Probability Fingerprints. In *Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, Banff, AB, Canada, 4–6 June 2018; pp. 1–10.
- [32] Gijon, C.; Toril, M.; Solera, M.; Luna-Ramirez, S.; Jimenez, L.R. Encrypted Traffic Classification Based on Unsupervised Learning in Cellular Radio Access Networks. *IEEE Access* 2020, 8, 167252–167263.
- [33] Draper-Gil, G.; Habibi Lashkari, A.; Mamun, M.S.; Ghorbani, A.A. Characterization of encrypted and VPN traffic using time-related. In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP)*, Rome, Italy, 19–21 February 2016; pp. 407–414. Available online: <https://www.unb.ca/cic/datasets/vpn.html> (accessed on 1 June 2022).
- [34] Raikar, M.M.; Meena, M.S.; Mulla, M.M.; Shetti, N.S.; Karanandi, M. Data Traffic Classification in Software Defined Networks (SDN) using supervised-learning. *Procedia Comput. Sci.* 2020, 171, 2750–2759.
- [35] Dias, K.; Pongelupe, M.A.; Caminhas, W.M.; de Errico, L. An innovative approach for real-time network traffic classification. *Comput. Netw.* 2019, 158, 143–157.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)