S.Y.B.Tech Students' Object Oriented Programming Project Paper, SEM 3 A.Y. 2024-25
Vishwakarma Institute of Technology, Pune, INDIA.

VPN Detection Using Machine Learning

Ajay Talele, Shriraj Aher, Ananya Bharti, Aditya Chougule, Agrima Panwar, Arpit Dalal

Department of Multidisciplinary Engineering
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

***Abstract* — VPN detection involves identifying users who utilize virtual private networks to hide their real IP addresses, usually to bypass geographic restrictions or enhance privacy. Some detection techniques include analyzing traffic behavior, identifying IP addresses linked to known VPN servers, and inspecting packet metadata for signs of VPN usage. Accurate VPN detection has many advantages such as it improves network security, prevents fraud, enforces compliance with location-based restrictions, and ensures the integrity of services reliant on user location data.**

***Keywords* —** *Machine Learning, VPN, Virtual Private Network, IP, Network security,*

I. Introduction

With growing requirements in the field of cybersecurity and in regulatory compliance, virtual private networks (VPNs) have become a popular tool used to protect user privacy and bypass geographic restrictions. However, VPNs also introduce challenges, as they can be exploited by malicious actors to hide their identities and activities. Traditional VPN detection methods, such as IP blacklisting and traffic pattern analysis, often struggle to keep pace with evolving VPN technologies and obfuscation techniques. These methods may fail to accurately detect VPN traffic, leading to security loopholes and enforcement issues.

Machine learning (ML) offers a dynamic and adaptive solution to the limitations of traditional VPN detection. By analyzing network traffic data and learning from patterns, ML models can more effectively distinguish between VPN and non-VPN traffic, even when advanced encryption or stealth techniques are used. Our research focuses on developing an ML-based VPN detection system that enhances accuracy of detection, adapts to new VPN methods, and ensures robust network security. The proposed approach addresses the need for scalable and efficient VPN detection in modern network environments.

II. LITERATURE REVIEW

The growing use of Virtual Private Networks (VPNs) for privacy and bypassing restrictions has led to increasing interest in effective detection techniques, particularly with the aid of machine learning (ML) models. Various studies have investigated the potential of ML in enhancing VPN detection capabilities, addressing the limitations of traditional techniques like packet inspection and IP blacklisting.

Miller et al. [1] introduced an MLP neural network that classifies VPN and non-VPN traffic using flow statistics from TCP headers. Their model achieved a 93.71% accuracy rate for OpenVPN traffic and 97.82% for Stunnel OpenVPN traffic using 10-fold cross-validation. Similarly, Srivastava et al. [2] tested neural networks on the same protocols, attaining similar accuracy but pointing out the tendency of false negatives, indicating the need for further refinement in ML-based VPN detection models.

Deep learning has also gained traction in VPN detection research. Mahdi et al. [3] proposed a deep learning model that aggregates continuous packets into "Packet Block" images and utilizes CNNs for classification, achieving high accuracy rates on the OpenVPN and ISCX-Tor datasets. Sun et al. [4] further demonstrated the effectiveness of CNN-based packet block methods, achieving 97.20% accuracy on OpenVPN and 93.31% on ISCX-Tor data. These studies emphasize the potential of deep learning to handle encrypted VPN traffic more effectively than traditional methods.

Other studies have explored broader applications of ML for networking and traffic classification. Akinsanya et al. [5] reviewed VPN protocols, including IPSec, SSL/TLS, and WireGuard, discussing their roles in network security while highlighting ongoing challenges like scalability and compliance. Boutaba et al. [6] examined the evolution of ML in networking, covering supervised learning, clustering, and reinforcement learning techniques applied to traffic prediction and security, noting that ML has enhanced the accuracy and adaptability of network systems.

Fesl and Naas [7] highlighted the complexity of VPN traffic detection across diverse protocols, like PPTP, L2TP, and WireGuard. They found that Random Forests and Decision Trees performed well in controlled environments, while advanced models like CNNs and LSTMs were more capable of handling encrypted traffic. Goel et al. [8] reinforced the effectiveness of Random Forest models in detecting VPN traffic, advocating for further exploration of real-time detection in network security.

Jorgensen et al. [9] introduced a novel ML framework for encrypted network traffic classification, utilizing uncertainty quantification techniques to handle dynamic traffic environments and detect out-of-distribution data. Furdek Prekratic et al. [10] applied MLP neural networks to TCP flow statistics, underscoring their potential in optical network security, while Dulany et al. [11] focused on the performance challenges of VPN systems on Linux, highlighting the processing overhead of software-based VPNs.

In terms of classification techniques, Bagui et al. [12] compared machine learning algorithms, such as Random Forest and Gradient Boosting Trees, for VPN traffic classification using time-related features. They found that Random Forest models outperformed simpler models, although further optimization of feature selection was necessary. Liu et al. [13] emphasized the integration of blockchain and ML for enhanced security in communication networks, suggesting that ML can aid in resource management and secure data transmission.

Hussain et al. [14] detailed how ML is applied to resource management in IoT networks, discussing its potential in optimizing power allocation and data aggregation. Lastly, Peter Dulany et al. [15] provided insights into the performance overhead of gateway-to-gateway VPNs on Linux, noting the substantial processing burden under high traffic conditions and suggesting that machine learning optimization could mitigate this issue.

Overall, the literature demonstrates that machine learning, especially deep learning, offers promising solutions for VPN detection by improving accuracy and scalability. However, challenges such as traffic obfuscation, real-time detection, and handling encrypted traffic remain key areas for future research.

III. Methodology

The methodology adopted for this project ensures a systematic and reliable approach to achieving the classification objectives. The process begins with data cleaning, during which inconsistencies are addressed, and necessary preprocessing steps, such as data type conversion, are performed to ensure the dataset's integrity.

Following this, feature selection is conducted using the SelectKBest method. This technique evaluates feature vectors to identify the most relevant variables for predicting the target class.

With the optimal features determined, the next step involves training the classification model. Considering the nature of the problem, which is a classification task, a Support Vector Machine (SVM) is employed to predict whether the given feature vector corresponds to VPN usage.

Finally, the classifier's performance is evaluated using standard evaluation metrics, with model accuracy serving as the primary measure of effectiveness.

IV. Results and Discussions

The dataset was trained using a Logistic Regression model, achieving an accuracy of 0.9988. While this high accuracy initially appears promising, it is indicative of overfitting, as such performance is unlikely with the limited size of the dataset. Overfitting suggests the model has learned patterns specific to the training data rather than generalizable features, undermining its predictive robustness.

This overfitting is attributed to the dataset's constraints, including a limited number of samples and insufficient feature diversity. A lack of comprehensive feature vectors limits the model's ability to generalize across unseen data, highlighting the need for dataset augmentation. Adding more meaningful columns with relevant features can provide the model with a richer, more representative dataset for training, thereby addressing this limitation.

Future work will focus on expanding the dataset by incorporating additional attributes and increasing its size. This enhancement aims to mitigate overfitting, improve the model's generalizability, and ensure robust performance on diverse data. Such modifications will strengthen the applicability of the model in real-world scenarios and advance its reliability for VPN detection tasks.

The confusion matrix provides further insights into the model's performance:

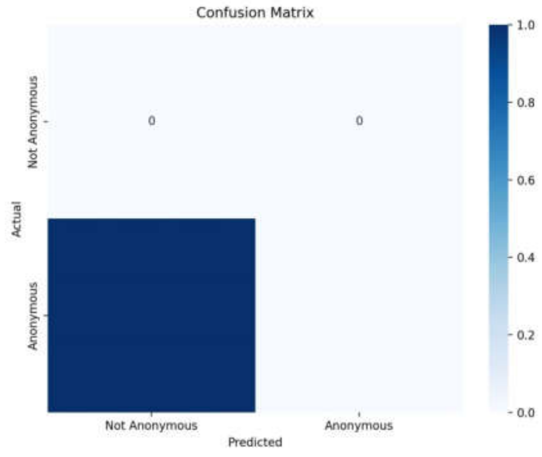

Fig 1: Confusion matrix

V. Future Scope

Currently, the model has been trained using IPv4 datasets. Future work could incorporate IPv6 datasets, enabling the model to adapt to modern network infrastructures and improve detection accuracy across a broader spectrum of internet traffic. This expansion could also enhance processing speeds, scalability, and applicability in dynamic real-world environments with evolving network protocols.

Furthermore, integrating the model with real-time traffic analysis systems could enable proactive detection of VPN usage in live environments. This would require optimizing the model for low-latency predictions and handling large volumes of continuous network traffic, ensuring it can operate effectively in high-demand scenarios. Such an extension could be particularly useful for enterprise-level security and network management systems.

Another promising direction involves exploring ensemble methods or hybrid approaches that combine multiple machine learning models to boost detection accuracy and robustness. Additionally, incorporating advanced techniques like deep learning could enable the system to uncover hidden patterns in encrypted traffic, further enhancing its reliability and adaptability to new VPN protocols.

VI. Conclusion

The proliferation of misinformation and disinformation Our research explores how machine learning can detect VPN traffic, using techniques like Logistic Regression and SelectKBest to identify key network characteristics. The results were impressive, with the model reaching 99.88% accuracy in identifying patterns.

While this shows promise, we found the model may be too closely fitted to our specific dataset, suggesting we need more diverse training data to make it work well in real-world situations.

We took a careful step-by-step approach to prepare our data, choose the most important features, and build our classification model. By focusing on the most relevant network attributes, we've laid groundwork that others can build upon. This matters because as VPNs become more common for privacy and security, being able to tell VPN traffic from regular traffic is increasingly important for network security teams.

To make the model better, we need to include IPv6 network data and gather more types of network information. We could also try more sophisticated machine learning approaches, like ensemble methods or neural networks, to improve how well we can spot VPN usage. This work is just the beginning of developing better tools for modern networks.

Overall, our project shows that machine learning can effectively tackle VPN detection, while also pointing the way forward for future improvements in this field.

VII . Acknowledgment

We express our deepest gratitude to Prof. Ajay Talele, whose guidance and expertise were invaluable throughout the research process. His insightful feedback and unwavering support significantly enriched the quality of this work.

We extend our gratitude to the developers and contributors of the tools and libraries that enabled the seamless implementation of this project. The data preprocessing, feature selection using methods like SelectKBest, and model training with classification algorithms such as Support Vector Machines were made possible by reliable open-source frameworks. These resources played a vital role in ensuring the efficiency and accuracy of our VPN detection system

We extend our gratitude to Vishwakarma Institute of Technology, Pune for providing the necessary infrastructure, resources, and a conducive environment for carrying out this research.

References

[1] S. Miller et al., "Detection of Virtual Private Network Traffic using Machine Learning," *Int. J. Wireless Netw. Broadband Technol.*, vol. 9, no. 1, pp. 13-23, 2020.

[2] S. Srivastava et al., "Detection of VPNs using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 7, pp. 12-20, 2022.

[3] R. A. K. Mahdi and M. Ilyas, "Using Deep Learning Technology to Optimize VPN Networks Based on Security Performance," *J. Elect. Syst.*, vol. 10, no. 5, pp. 55-63, 2024.

[4] W. Sun et al., "A Deep Learning-Based Encrypted VPN Traffic Classification Method Using Packet Block Image," *MDPI J. Electron.*, vol. 12, no. 1, pp. 123-135, 2023.

[5] M. O. Akinsanya et al., "Virtual Private Networks (VPN): A Conceptual Review of Security Protocols and Their Application in Modern Networks," *Eng. Sci. Technol. J.*, vol. 5, no. 3, pp. 78-89, 2024.

[6] R. Boutaba et al., "A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities," *J. Internet Serv. Appl.*, vol. 10, no. 2, pp. 1-23, 2018.

[7] J. Fesl and M. Naas, "A Complex ML-Based Approach for VPN Detection and Identification," *Int. J. Intell. Works*, vol. 8, no. 2, pp. 45-52, 2024.

[8] A. Goel et al., "Detection of VPN Network Traffic," *IEEE Delhi Sect. Conf.*, pp. 1-5, 2022.

[9] S. Jorgensen et al., "Extensible Machine Learning for Encrypted Network Traffic Application Labeling via Uncertainty Quantification," *IEEE Trans. Artif. Intell.*, vol. 5, no. 1, pp. 234-245, 2024.

[10] M. Furdek Prekratic et al., "Machine Learning for Optical Network Security Monitoring: A Practical Perspective," *J. Lightwave Technol.*, vol. 38, no. 7, pp. 1024-1034, 2020.

[11] P. Dulany, C. S. Kim, and J. T. Yu, "A Performance Analysis of Gateway-to-Gateway VPN on the Linux Platform," in *Proc. IEEE Conf.*, pp. 78-85, 2006.

[12] S. Bagui et al., "Comparison of Machine-Learning Algorithms for Classification of VPN Network Traffic Flow Using Time-Related Features," *J. Cyber Secur. Technol.*, vol. 3, no. 1, pp. 45-55, 2017.

[13] Y. Liu et al., "Blockchain and Machine Learning for Communications and Networking Systems," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 2, pp. 1112-1134, 2020.

[14] F. Hussain et al., "Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 1, pp. 117-134, 2020.

[15] P. Dulany, C. S. Kim, and J. T. Yu, "A Performance Analysis of Gateway-to-Gateway VPN on the Linux Platform," *Article*, 2006.