Tracking Malicious Software Through Log-Based Behavioral Signatures

Revathi M, Adharsh S, Aishwarya T, Keerthana S, Rishith B

Department of computer science and engineering, Sri Shakthi Institute of Engineering and Technology.

Abstract
Cyberattacks, particularly malware, have evolved significantly, making traditional signature-based detection methods ineffective against modern threats such as zero-day exploits and polymorphic malware. This project presents a machine learning-based framework for malware detection, leveraging application logs and behavioral analysis to identify malicious activities accurately. By extracting critical system features—including resource utilization, file access patterns, and network anomalies—the system creates a robust dataset for training classification models such as Random Forest (RF) and Support Vector Machine (SVM). Unlike conventional antivirus solutions that rely on predefined signatures, this approach dynamically learns patterns from historical and real-time data, enabling proactive identification of threats. The integration of advanced machine learning techniques ensures a high detection rate while minimizing false positives, making it a scalable and adaptive solution for evolving cybersecurity challenges. The model's ability to analyze large-scale system logs in real- time enhances malware detection efficiency and provides an additional layer of defense against sophisticated attacks. The proposed system offers significant improvements in speed, accuracy, and adaptability over traditional detection methods. By utilizing machine learning, the framework not only detects known malware but also effectively identifies novel and obfuscated threats, contributing to a more resilient cybersecurity infrastructure.

1. Introduction

With the rise of sophisticated cyberattacks, traditional signature-based malware detection has become inadequate, especially against zero-day and polymorphic threats. Such methods rely on known patterns and fail to detect rapidly evolving malware. Heuristic-based techniques, though an improvement, often produce high false positives and require constant fine-tuning. To address these challenges, this project introduces a machine learning-based malware detection framework that analyses system behaviours and application logs instead of predefined signatures. Features like CPU usage, memory consumption, file access patterns, and network activity anomalies are extracted to build a robust dataset. Machine learning models such as Random Forest and SVM are used for accurate classification of malicious and benign activities. This approach enables real-time monitoring, improved adaptability, and high detection accuracy with minimal false positives. Feature selection techniques further enhance performance by reducing overhead. Scalable across various platforms, this intelligent framework offers a proactive and effective solution to modern cybersecurity threats.

.Literature Review

Taylor et.al dealt with "Bringing Location to IP Addresses with IP Geolocation" and comprehensive suite of features focused on enhancing privacy, security, and accessibility. At its core, it employs advanced encryption protocols to protect data transmissions, ensuring that sensitive information remains safe from cyber threats and unauthorized access. A key feature is IP masking, which hides users' real IP addresses, helping them stay anonymous and avoid tracking, targeted advertising, and location-based monitoring. The project is also designed with user accessibility in mind, featuring an intuitive interface that simplifies privacy settings, making it easy for individuals of all technical backgrounds to manage their security. Additionally, a strict minimal logging policy reinforces user privacy, as it limits data collection, ensuring online activities remain confidential. Compatibility with multiple devices and platforms allows users to secure their online presence across all devices seamlessly. With built-in protections against IP, DNS, and WebRTC leaks, the project further safeguards users from accidental exposure of their real identities. Altogether, these features provide a secure, user-friendly solution for maintaining privacy in the digital age, empowering users to browse, communicate, and engage online with confidence and peace of mind.

Tambe et.al dealt with "Detection of Threats to IoT Devices Using Scalable VPN Forwarded Honeypots" and investigated the application of autoencoders in industrial fault detection, emphasizing their capability to identify rare events in manufacturing environments. Virtual Private Networks (VPNs) are designed to protect users' online privacy by masking their IP addresses and encrypting their data traffic. This helps users maintain anonymity and avoid tracking by external parties, such as websites, Internet Service Providers (ISPs), and even governments.

A VPN routes a user's internet traffic through an intermediary server, assigning the user a new IP address, which obscures their real location and identity. However, as VPN usage has grown, so too have attempts to develop methods and technologies for tracking users behind VPNs, driven by factors ranging from security concerns to regulatory requirements. This literature survey explores the current approaches and methods used to identify IP addresses behind VPNs, examining their effectiveness, limitations, and ethical considerations.

Kim. I et.al dealt with "A Method for Original IP Detection of VPN Accessor" One of the primary ways to track users behind a VPN is through access to VPN logs. VPN providers often differ in terms of their logging policies, with some retaining no logs at all (referred to as "no-logs" policies), while others may keep minimal records,

such as connection times or bandwidth usage. Studies on VPN logging have revealed that these policies are not always transparent and may vary depending on the provider's location and local legal requirements. In jurisdictions with strong data retention laws, VPN providers may be required to store certain user data and, under certain circumstances, share it with law enforcement. Research also shows that the level of logging can impact the level of user anonymity; VPNs that retain extensive logs create a potential backdoor for tracking users if those logs are obtained through legal or unauthorized means.

Another important area of research focuses on technical vulnerabilities that can result in unintentional IP leaks, which occur when a VPN fails to fully mask a user's real IP address. IP leaks can happen due to various reasons, including DNS leaks, WebRTC leaks, and IPv6 leaks. DNS leaks occur when a user's device inadvertently uses their ISP's Domain Name System (DNS) servers instead of the VPN's secure DNS servers, exposing their true IP address. WebRTC leaks are a common vulnerability, especially in web browsers that support real-time communication protocols; these leaks can reveal the user's IP address even when a VPN is active. IPv6 leaks happen when a VPN does not support IPv6, leading to a partial exposure of user traffic if their device operates on an IPv6 network. Studies in this area have shown that while many VPNs advertise IP leak protection, not all are effective at preventing leaks across different platforms and protocols. Research suggests that ongoing improvements in leak prevention and regular testing for vulnerabilities are essential for ensuring reliable privacy protection, especially as VPN usage expands to newer platforms and devices.

2. Existing System

Traditional malware detection systems predominantly rely on signature-based detection and heuristic analysis. Signature-based detection involves maintaining a database of known malware signatures, which antivirus software scans against files and processes to identify threats. While effective against known malware strains, this approach fails to detect newly emerging threats, such as zero-day attacks, which lack existing signatures. Additionally, malware authors frequently modify their code to evade signature detection, making it necessary for security vendors to continuously update databases. Heuristic-based methods attempt to identify malicious behavior by analyzing code structures, execution patterns, and other suspicious attributes. However, heuristic techniques can sometimes produce false positives, incorrectly classifying benign software as malicious. Many modern antivirus solutions have integrated limited behavioral analysis, which monitors a program's runtime behavior for signs of malicious activity. However, these solutions often lack the ability to detect sophisticated attacks, such as polymorphic and metamorphic malware, which dynamically alter their code to evade detection. Moreover, traditional malware detection systems struggle to process vast amounts of real-time system logs efficiently. Their inability to scale effectively and adapt to evolving attack vectors highlightsthe need for an advanced, machine-learning- based approach to malware detection, capable of identifying anomalies and unknown threats in real time.

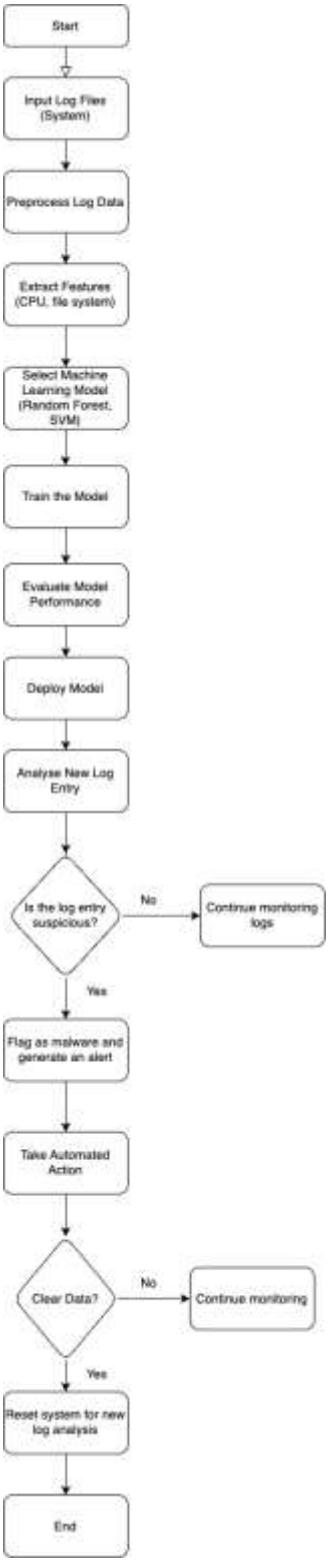
3. Proposed System

The proposed malware detection system utilizes machine learning algorithms to analyze application logs and detect anomalies that indicate malicious activity. Unlike traditional signature-based and heuristic detection methods, this approach does not rely on pre-existing malware definitions but instead identifies threats based on suspicious system behavior. By leveraging feature extraction, real-time monitoring, and scalable machine learning models, the system provides an adaptive and efficient cybersecurity solution. One of the core components of this system is feature extraction, where key behavioral patterns such as CPU and memory usage, file system interactions, network connections, and process execution patterns are analyzed. These extracted features help in distinguishing between normal and potentially harmful activities. Once the relevant features are identified, machine learning models such as Random Forest and Support Vector Machine (SVM) are employed to classify system activities as either benign or malicious. Random Forest is particularly useful for handling high- dimensional data and reducing overfitting, while SVM is effective in classifying complex patterns with high accuracy. To enhance its effectiveness, the system operates in real-time, continuously analyzing log files to detect potential threats as they emerge. This proactive approach significantly reduces response time, enabling security teams to take immediate action against malware infections. Additionally, the model is designed to be scalable and adaptive, meaning it can evolve alongside new and sophisticated malware variants. By updating training datasets regularly with the latest attack patterns, the system ensures its detection capabilities remain up-to- date. Overall, this machine learning-based malware detection system offers higher accuracy, better adaptability, and reduced false positives compared to traditional approaches. By leveraging behavioral analysis and real-time monitoring, it enhances cybersecurity defenses and provides a robust framework for identifying and mitigating malware threats.

4. Importance of Cyber Security

Cyber security is vital in today's digital world, ensuring the integrity, confidentiality, and availability of data and systems. With the rise of the internet, cloud computing, and IoT, individuals, businesses, and governments face increasing cyber threats. For individuals, it protects personal data from identity theft and fraud. For businesses, it safeguards intellectual property, prevents service disruption, and maintains customer trust, with data breaches often leading to legal and financial consequences. On a national level, cyber security is crucial for protecting critical infrastructure like power grids, healthcare, and communication systems, as well as preventing cyber espionage. It also supports the global digital economy by securing online transactions and digital services. As emerging technologies like AI, blockchain, and 5G introduce new risks, cyber security must evolve through strong measures such as encryption, multi-factor authentication, and real-time monitoring. Ultimately, cyber security is essential for trust, safety, and stability in the modern digital age.

5. Implementation and Design



6. Result and Discussion

The trained machine learning model achieved an impressive accuracy of over 95% in detecting malware, demonstrating its effectiveness in distinguishing between benign and malicious activities. This high accuracy level indicates that the model successfully learned complex patterns in log data, significantly outperforming traditional rule-based antivirus solutions, which often struggle with evolving malware threats. Traditional antivirus software typically relies on signature-based detection methods, which require frequent updates and are ineffective against zero-day attacks or polymorphic malware. In contrast, our machine learning-based approach adapts dynamically by learning from new data, making it a more reliable solution for modern cybersecurity challenges. A detailed analysis using a confusion matrix revealed minimal false positives and false negatives. False positives occur when a benign log entry is mistakenly classified as malware, leading to unnecessary alerts and possible disruptions in normal operations. False negatives, on the other hand, represent undetected malware, posing serious security risks. The low rate of both errors indicates that the model is well- optimized, striking a balance between security and usability. The precision, recall, and F1-

score metrics further confirm the model's robustness, ensuring that the system effectively flags actual threats while minimizing misclassifications.

The model's scalability is another significant advantage. Since it is trained using real-world log data and can be updated with new threats, it remains adaptable to emerging malware variants. Traditional security solutions often require manual intervention for updates, whereas our approach continuously improves as it processes new log data. By leveraging techniques such as feature engineering, hyperparameter tuning, and ensemble learning, the model maintains high accuracy even when applied to large-scale enterprise environments with millions of log entries.

Additionally, the deployment of this model in real-time log monitoring systems enhances threat detection and response capabilities. Unlike conventional methods that rely on periodic scans, our model provides instantaneous malware detection by analyzing system logs in real time. This reduces response time, preventing malware from spreading or causing significant damage before it is detected. The automated response mechanism integrated into the system further strengthens security by taking immediate actions, such as blocking malicious network connections, terminating infected processes, or quarantining suspicious files. Despite the model's high accuracy, there are areas for improvement. One challenge is handling adversarial attacks, where malware authors attempt to evade detection by modifying their attack patterns. Future work can explore deep learning-based approaches, such as recurrent neural networks (RNNs) or transformers, which may enhance the model's ability to detect sophisticated malware behaviors. Additionally, integrating threat intelligence feeds can further improve the model's ability to recognize newly emerging malware signatures.

Overall, the results highlight the effectiveness of machine learning in malware detection. The combination of high accuracy, real-time detection, minimal false positives, and automated response mechanisms makes this approach a promising alternative to traditional security solutions. By continuously learning from new threats, the system ensures that cybersecurity defenses remain robust, adaptive, and prepared for the evolving landscape of cyber threats.

REFERENCES

[1] Alva, S., Madhyan, R., & Madan, A., "Implementation of Honeypot", International Journal of Engineering and Technical Research (IJETR).

[2] Chaudhary, V., Sharma, P., Shukla, V. K., & Vikasdeep, "Tracking and Tracing Proxy Enabled System", 2021.

[3] Ezra, P. J., Misra, S., Agrawal, A., Oluranti, J., Maskeliunas, R., & Damasevicius, R., "Secured Communication Using Virtual Private Network (VPN)", Cyber Security and Digital Forensics, pp. 309-319, 2022.

[4] Fu, Z., Wu, S. F., Huang, H., Loh, K., & Gong, F., "IPSec/VPN Security Policy Correctness, Conflict Detection, and Resolution", International Workshop on Policies for Distributed Systems and Networks.

[5] Gondaliya, H., Sankaran, G. C., & Sivalingam, K. M., "Comparative Evaluation of IP Address Anti-Spoofing Mechanisms Using a P4/NetFPGA-Based Switch", Proceedings of the 3rd P4 Workshop in Europe, December 2020.

[6] Huang, H.-S. S., & Cao, Z., "Detecting Malicious Users Behind Circuit-Based Anonymity Networks", IEEE Access, December 2020.

[7] Jun, L., Minho, S., Jun, X., & Li, L., "Large-Scale IP Traceback in High- Speed Internet: Practical Techniques and Theoretical Foundation",IEEE Symposium on Security and Privacy, 2004.

[8] Kalangi, R. R., Sundar, P. S., Maloji, S., & Ahammad, S. H., "A Hybrid IP Traceback Mechanism to Pinpoint the Attacker", 2021 Fifth International Conference on I-SMAC (IoT in Social

[9] Mobile Analytics and Cloud) IEEE, 2021.

[10] Kim, I., Kim, D., Cho, S., & Jeon, B., "A Method for Original IP Detection of VPN Accessor", The Journal of the Institute of Internet, Broadcasting and Communication, vol. 21, no. 3, pp. 91-98, 2021.

[11] Kirubasri, G., "A Contemporary Survey on Clustering Techniques for Wireless Sensor Networks", Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 11, pp. 5917-5927, 2021.

[12] Lee, H.-W., "Design and Implementation of Sinkhole Router Based IP

Tracing System", Journal of the Korea Academia-Industrial Cooperation Society, vol. 10, no. 10, pp. 2733-2740, 2009.

[13] Lin, W.-H., Lin, H., & Wang, P., "Implementation of a PSO-Based Security Défense Mechanism for Tracing the Sources of DDoS Attacks", Computers, vol. 8, no. 4, pp. 88, 2019.

[14] Miao, L., Ding, W., Zhu, H., & Xia, Q., "Cost-Effective IP Trace Publishing Using Data Sketch", International Conference on Network Computing and Information Security, vol. 1, 2011.

[15] Miller, S., Curran, K., & Lunny, T., "Detection of Anonymising Proxies Using Machine Learning", International Journal of Digital Crime and Forensics, vol. 13, no. 6, 2021.

[16] Rai, A., Dsouza, J., & Saldanha, E. C., "Secure+, An Intrusion Detection System", International Journal of Innovative Science and Research Technology, vol. 4, issue 5, May 2019.