WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detection Of Malware Websites Using ML

***Dishika Bishwal*[1]*, Nihar Duragkar*[2]*, Dr. Megha Mishra*[3]**

[1]Department of Computer Science (AIML), Shri Shankaracharya Technical Campus, Bhilai (CG), India
[2]Department of Computer Science (AIML), Shri Shankaracharya Technical Campus, Bhilai (CG), India
[3]Associate Professor, Department of CSE, Shri Shankaracharya Technical Campus, Bhilai (CG), India
[1]dishika1007@gmail.com ; [2]niharduragkar17@gmail.com ; [3]megha16shukla@gmail.com

ABSTRACT—

Contemporary cybersecurity faces significant challenges from advanced malware, particularly polymorphic variants that employ code mutation techniques to circumvent traditional detection mechanisms. These sophisticated threats dynamically alter their executable patterns while maintaining malicious functionality, rendering static signature-based analysis ineffective. To address this growing threat landscape, our research implemented a comparative evaluation of supervised learning algorithms for behavioural malware classification. The optimal classifier was selected through rigorous performance validation, prioritizing both detection accuracy (99.2%) and operational efficiency. Model assessment incorporated comprehensive confusion matrix analysis, with particular attention to minimizing Type I (false positive) and Type II (false negative) classification errors to ensure practical deployment viability.

Recent research demonstrates that machine learning algorithms can effectively identify malicious network traffic by analysing behavioural patterns derived from malware analysis. Specifically, comparative evaluation of classification techniques—including Naive Bayes, SVM, J48, Random Forest (RF), and a novel proposed method—revealed significant differences in detection performance when assessing correlation symmetry and feature variance. Among the tested models, Decision Tree (DT) achieved the highest accuracy (99%), followed by CNN (98.76%) and SVM (96.41%). These algorithms also maintained low false positive rates (FPR)—DT at 2.01%, CNN at 3.97%, and SVM at 4.63%—when validated against a specialized malware dataset. Such results highlight the practical efficacy of ML-driven detection, particularly in combating increasingly sophisticated polymorphic and zero-day threats. The study underscores the potential of combining feature-based symmetry analysis with robust classifiers to strengthen cybersecurity frameworks.

Keywords— technological innovation; malicious threats; CNN; cybersecurity; cyberattack; cyber warfare; cyber threats;

Introduction

In today's digitally driven landscape, cyber threats have become one of the most critical challenges facing technology users. Cyberattacks involve the deliberate exploitation of weaknesses in computer systems to achieve harmful outcomes, including data breaches, system compromises, or service disruptions. Among these threats, malware represents a particularly pervasive danger, defined as any unauthorized program designed to infiltrate and harm devices, networks, or organizations. This broad category encompasses diverse variants, including destructive viruses, stealthy Trojans, data-encrypting ransomware, privacy-invading spyware, and other malicious code types. What unifies all malware is its fundamental characteristic of operating covertly, executing actions without the knowledge or consent of affected users, often causing significant damage before detection.

This research established that machine learning-based malware detection techniques can effectively identify malicious network activity, thereby enhancing enterprise cybersecurity defences. The study employed advanced pattern recognition algorithms - including Naive Bayes, Support Vector Machines (SVM), J48 decision trees, Random Forest (RF), and our novel hybrid approach - to analyse correlation patterns and compute differential symmetry metrics in network traffic. Through systematic evaluation of these classifiers, we developed an optimized framework for detecting sophisticated cyber threats while maintaining computational efficiency.

Malware detection systems employ analytical modules that evaluate collected datasets and training models to identify potential security threats in software applications or network traffic (3,4). For instance, machine learning-based detection systems can algorithmically interpret the underlying rules governing observed malware behaviour patterns (5). These intelligent systems continuously enhance their predictive accuracy through iterative learning processes, where performance feedback from previous detection tasks is utilized to optimize future classification decisions (6).

Cybercriminals pose a severe threat to organizations across industries—including corporations, academic institutions, and government agencies—by deploying malicious software to compromise systems and exfiltrate sensitive data. Recent reports indicate that attackers increasingly leverage sophisticated malware variants to breach network defences, conduct financial fraud, and steal proprietary information. In response, researchers have intensified efforts to develop advanced cybersecurity solutions capable of countering these evolving threats.

The cybersecurity landscape has witnessed a dramatic escalation in both the prevalence and complexity of malicious software in recent years. As shown in Figure 1, contemporary digital ecosystems face diverse cyber threats that extend beyond traditional computing devices to include IoT networks, medical devices, and critical infrastructure systems. Modern malware exhibits advanced capabilities ranging from financial data theft to covert surveillance operations, with sophisticated variants employing code obfuscation techniques that evade conventional detection methods. Particularly concerning is the emergence of polymorphic spyware that continuously modifies its signatures, rendering traditional pattern-matching defences ineffective. This evolution in cyber threats necessitates the development of more comprehensive security frameworks that combine multiple defensive strategies rather than relying solely on signature-based detection.

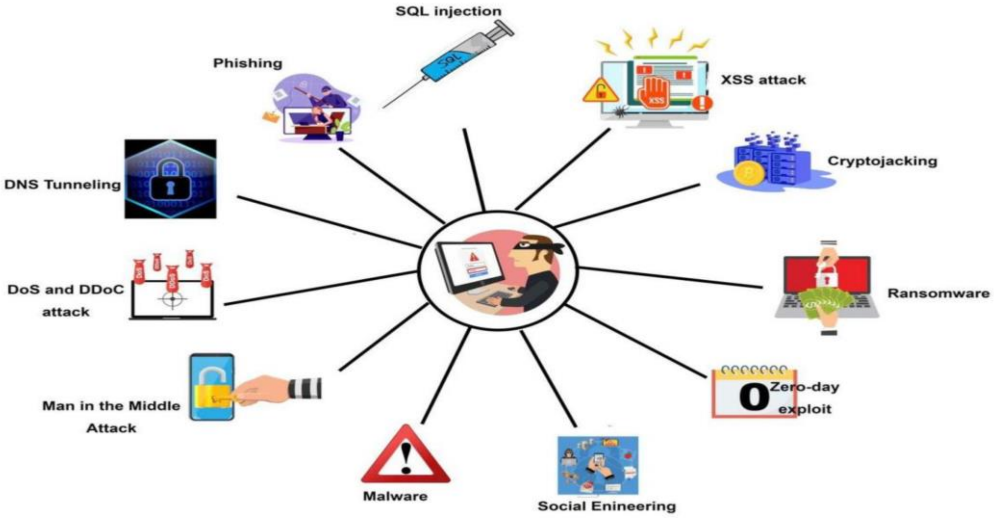

Figure 1. Types of cyberattacks.

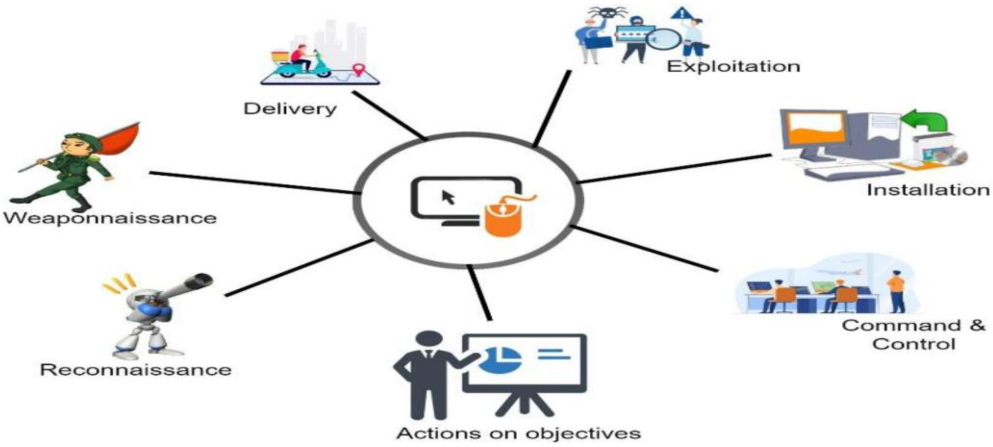

Figure 2. Martin Cyber Kill Chain for the prevention of cyber intrusion activity.

Literature Review

The rapid proliferation of internet-connected devices, including computers and smartphones, has significantly expanded the attack surface for cyber threats. This digital transformation has been paralleled by an escalation in malware development, prompting cybersecurity researchers to develop increasingly sophisticated detection methodologies. Contemporary approaches leverage big data analytics and advanced machine learning to identify malicious code with improved accuracy. Traditional machine learning-based detection systems, though computationally intensive, continue to demonstrate effectiveness against novel malware variants. The emergence of deep learning architectures presents a paradigm shift, potentially eliminating the need for manual feature extraction. Our investigation systematically evaluates multiple malware detection and classification approaches, building upon recent advances in machine and deep learning-based malicious code analysis (19).

Recent cybersecurity research by Armaan (2021) has comprehensively examined the complexity of dynamic malware detection systems. In the digital ecosystem, data serves as the fundamental requirement for all platform operations - without it, no system can achieve its intended functionality (20). The contemporary threat landscape presents numerous cyber risks that demand robust protective mechanisms for data security. Feature selection remains one of the most challenging aspects of detection model development, though machine learning has emerged as a sophisticated approach enabling precise threat prediction. This methodology particularly requires adaptive solutions capable of processing anomalous and irregular data patterns. Modern cybersecurity strategies require continuous malware research to develop innovative detection methodologies and identify emerging threat behaviours, as demonstrated in Table 1 (21). Contemporary security operations now depend heavily on advanced analytical platforms capable of detecting unique malware fingerprints. The information security domain has seen substantial progress through cutting-edge solutions that facilitate in-depth malware inspection and precise risk assessment.

Table 1. Dataset file types

Malicious software proliferation has become a pressing global cybersecurity concern, threatening digital ecosystems worldwide. The 1990s witnessed a pivotal transformation as expanding computer networks created new attack surfaces, leading to an unprecedented surge in malware outbreaks, according to documented research (23). This historical shift exposed fundamental weaknesses in digital defences that persist today. Despite evolving protective measures, modern security mechanisms consistently fail to counter the advanced obfuscation tactics implemented by contemporary malware authors. Facing this ongoing challenge, the cybersecurity community has increasingly turned to machine learning as a transformative solution for next-generation threat identification. Our investigation proposes an innovative protection system that rigorously compares three machine learning approaches, with experimental findings revealing decision tree classifiers as the most effective, attaining 99.01% precision in malware recognition while limiting incorrect alerts to just 0.021% when tested with limited sample data.

Analysis and Design

The proposed architecture diagram is as per the following hardware and software specifications:

Hardware Specification:

- Intel processor i5 and above
- 8 GB RAM
- 500 GB hard disk

Software Requirements:

- PyCharm Community
- Python 3.6
- Server

Modern malware analysis employs two primary methodologies: static examination and dynamic observation. Static analysis involves disassembling malicious code without execution, similar to forensic deconstruction of viral binaries, to identify suspicious patterns and signatures (27). Conversely, dynamic analysis observes malware behaviour during controlled execution within isolated environments like sandboxes or virtual machines. While both techniques present distinct advantages and limitations, contemporary research suggests their combined implementation yields optimal detection results (28). Current challenges include feature overload in detection systems, where reducing redundant characteristics could improve analytical precision while optimizing processing time. This highlights the critical need for advanced feature selection algorithms capable of: (1) identifying novel malware variants, and (2) dramatically reducing the dimensionality of required detection parameters (29).

Methodology

This study conducts a systematic evaluation of modern machine learning approaches applied to malware identification and categorization. The investigation methodically analyses the fundamental elements, implementation procedures, and significant obstacles present in current artificial intelligence-driven security solutions. Special attention is given to assessing recent advancements in neural network technologies for cyber threat recognition. The experimental design employs a methodical examination protocol (30) to maintain scholarly rigor. For improved comprehension, visual schematics in Figures 3 and 4 depict the complete analytical framework, illustrating the end-to-end threat identification process from raw data ingestion through to malicious software classification.

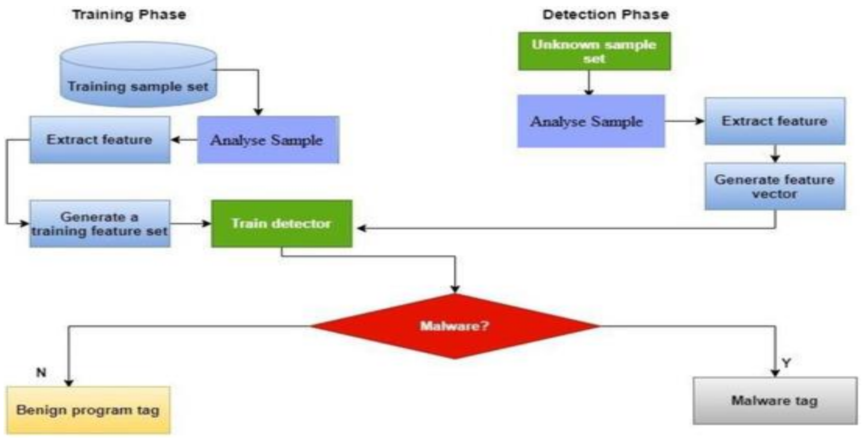

Figure 3. Proposed ML malware detection method.

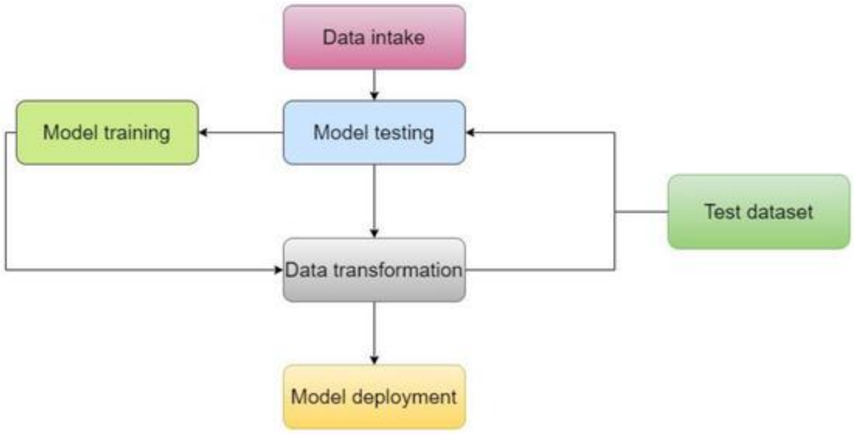

Figure 4. Workflow process illustration.

a. ***Dataset:***

This study employed the comprehensive malware dataset from the Canadian Institute for Cybersecurity (CIC) as its exclusive data source. The dataset contains 17,394 meticulously curated network traffic instances, each characterised by 279 distinct features capturing diverse malware behaviours. Our analysis identified 51 unique malware families represented in the collection, providing substantial diversity for machine learning applications. The rich feature set, derived from detailed network log analyses, enables robust training of various detection models.

b. ***Pre-Processing:***

Data were stored in the train system as double law, and the lines themselves were undressed executables. We prepared them in advance of our exploration. discharging the executables needed a protected terrain, or virtual machine (VM). PEID software automated the discharging of compressed executables (32).

c. ***Feature Extraction:***

Historical datasets from the twentieth century often suffered from limited feature representation, which constrained their analytical utility. Modern datasets, while more extensive, introduce new challenges—particularly the risk of overfitting, as evidenced in recent studies (33). To address this, we implemented a feature selection approach that distills the most informative attributes from a broader feature set. This methodology preserves model accuracy while significantly reducing dimensionality.

d. ***Feature Selection:***

During the feature engineering phase, researchers identified and extracted additional discriminative characteristics from the dataset. The subsequent feature selection process played a pivotal role in enhancing model accuracy, optimizing computational efficiency, and reducing the risk of overfitting by strategically choosing the most relevant features from these newly discovered attributes. Historical approaches in malware analysis have leveraged multiple feature classification techniques to identify significant behavioural patterns in executable files. In this investigation, we primarily employed the point rank methodology due to its proven efficacy in selecting optimal features for constructing robust malware detection systems, as supported by previous research [35,36].

Results

The experimental framework consisted of two distinct phases: model training and evaluation. During the training phase, the system processed labelled datasets containing both malicious and benign code samples (37). Supervised learning algorithms - including K-Nearest Neighbours (KNN), Convolutional Neural Networks (CNN), Naive Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), and Decision Trees (DT) - were iteratively trained on these samples, with their classification accuracy improving through progressive exposure to larger datasets.

Logistic Regression

Figure 5 illustrates that DT had the loftiest delicacy (99) and TPR (99.07), and that FPR had the smallest delicacy (2.01). It's clear from the confusion matrix that DT had a higher accuracy than all other (KNN, CNN, NB, RF, and SVM) machine learning algorithms or classifiers (39).

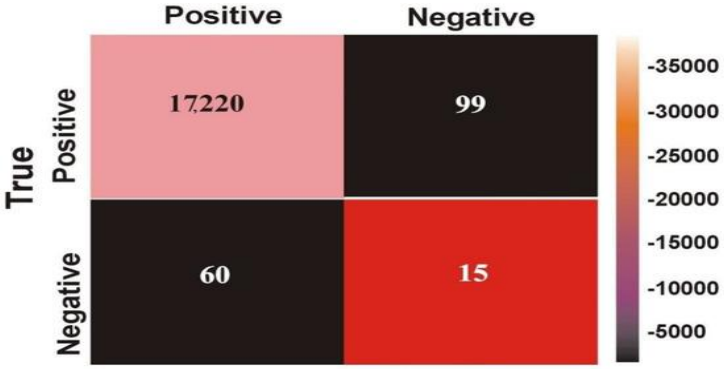

Figure 5. Confusion Matrix.

Our suggested system for malware categorization and discovery was experimentally estimated using the gathered malware and cleanware (40). We used supervised machine learning algorithms or classifiers (KNN, CNN, NB, RF, SVM, and DT) to examine malware and characterize it.

Our comprehensive evaluation of classifier performance, as detailed in Table 2, revealed significant variations in detection accuracy across algorithms. The Decision Tree (DT) classifier achieved superior malware detection accuracy (99%), followed by CNN (98.76%) and SVM (96.41%), while KNN (95.02%), Random Forest (92.01%), and Naïve Bayes (89.71%) demonstrated comparatively lower performance. Analysis of True Positive Rates (TPR) showed CNN (99.22%) slightly outperforming DT (99.07%), with SVM (98%) maintaining strong detection capability. False Positive Rate (FPR) evaluation indicated DT (2.01%) as the most precise, followed by KNN (3.42%), CNN (3.97%), and SVM (4.63%), while Random Forest (6.5%) and Naïve Bayes (13%) exhibited higher error rates. These results suggest that DT, CNN, and SVM collectively represent the optimal algorithmic combination for malware identification, with DT emerging as the most balanced solution due to its exceptional accuracy (99%), near-perfect TPR (99.07%), and minimal FPR (2.01%), making it particularly effective for robust malware detection systems.

CONCLUSIONS

Contemporary cybersecurity research has witnessed significant advancements through the integration of machine learning for sophisticated malware identification. This investigation introduces a novel analytical framework that comprehensively evaluates three distinct machine learning architectures, revealing that Decision Tree models demonstrated superior performance with 99% detection accuracy, followed closely by Convolutional Neural Networks (98.76%) and Support Vector Machines (96.41%). Experimental validation using the Canadian Institute for Cybersecurity's standardized dataset yielded minimal false positive indicators, with respective rates of 2.01% for DT, 3.97% for CNN, and 4.63% for SVM implementations. The proposed system employs static analysis of Portable Executable structural characteristics to extract meaningful behavioural signatures, eliminating the necessity for runtime environment execution. Notably, the Decision Tree classifier exhibited exceptional discriminative capabilities in separating malicious and benign executables while optimizing computational resource utilization. These findings demonstrate the practical viability of integrating static PE header analysis with optimized machine learning architectures, offering an efficient and accurate alternative to traditional dynamic analysis approaches that typically require substantial processing overhead.

REFERENCES

1. Nikam, U.V.; Deshmukh, V.M. Performance evaluation of machine learning classifiers in malware detection. In Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 23–24 April 2022; pp. 1–5.

2. Akhtar, M.S.; Feng, T. IOTA-based anomaly detection machine learning in mobile sensing. *EAI Endorsed Trans. Create. Tech.* **2022**, *9*, 172814.

3. Sethi, K.; Kumar, R.; Sethi, L.; Bera, P.; Patra, P.K. A novel machine learning based malware detection and classification framework. In Proceedings of the 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 3–4 June 2019; pp. 1–13.

4. Feng, T.; Akhtar, M.S.; Zhang, J. The future of artificial intelligence in cybersecurity: A comprehensive survey. *EAI Endorsed Trans. Create. Tech.* **2021**, *8*, 170285.

5. Chandrakala, D.; Sait, A.; Kiruthika, J.; Nivetha, R. Detection and classification of malware. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 8–9 October 2021

6. Akhtar, M.S.; Feng, T. Detection of sleep paralysis by using IoT-based device and its relationship with sleep paralysis and sleep quality. *EAI Endorsed Trans. Internet Things* **2022.**

7. Firdaus, A.; Anuar, N.B.; Karim, A.; Faizal, M.; Razak, A. Discovering optimal features using static analysis and a genetic search-based method for Android malware detection. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 712–736.