Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260503010
Volume 5, Issue3 (September-December 2026), PP: 79-83.
www.indjcst.com
ISSN No: 2583-5300

A Comprehensive Survey of Malware Detection and Classification

Santosh Patil[1], Vivekanand Koparde[2]
[1,2]Department of CSE, Angadi Institute of Technology and Management, Belagavi, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India.

To Cite this Article: *Santosh Patil[1], Vivekanand Koparde[2], "A Comprehensive Survey of Malware Detection and Classification", Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 79-83.*

Abstract: *As People become more dependent on digital technologies, the risk of cyber-attacks, particularly malware infections, escalates significantly. To prevent unauthorized access, data breaches and service disruption it is essential to have effective malware detection mechanisms is essential to prevent unlawful access, data breaches, and service disruptions. An effective malware detection approach helps users safeguard their digital assets, maintain user confidence, and comply with industry standards. This survey paper provides an overview of contemporary methodologies used in the detection and analysis of malicious code, like signature identification, behavioral analysis, heuristic analysis, sandbox analysis and machine learning methods.*

Key Word: *malware detection, signature identification; behavioral analysis; heuristic analysis; and sandbox analysis*.

I. INTRODUCTION

People's lives are getting more convenient as computers, smartphones, tablets, and other electronic devices become more and more common. Electronic gadgets and internet are now an integral part of people's everyday lives. As electronic devices become increasingly prevalent, they have introduced new security challenges like loss of user files, telecommunication fraud, malicious extortion and other issues. In early 2020,

AV-TEST [1] anticipated that the total number of malware samples would surpass 160 million by year-end, marking an unprecedented increase in scale. SonicWall [2] reported a significant surge in cyber threats between 2020 and 2021, with global cybersecurity expenditures doubling over the year. In the first half of 2022, SonicWall Capture Labs observed a notable 11% increase in global malware attacks, totaling 2.8 billion incidents. This uptick marked the first escalation in global malware volume in over three years, in 2024, StationX [3] reported that the global landscape of cybersecurity threats encompassed approximately 1.2 billion malicious programs and potentially unwanted applications (PUAs).

In 2024, the cybersecurity landscape was marked by a significant increase in malicious code attacks, with several high-profile incidents highlighting the evolving tactics and targets of cybercriminals. Here are some of the most notable attacks from that year: Around mid-2024, Snowflake Inc., a cloud data warehousing service, was at the center of a significant cyberattack that exposed confidential information from more than 160 clients, including AT&T, Ticketmaster, and Santander Bank. The cyberattack was linked to the group UNC5537, also referred to as Scattered Spider, which gained access to customer systems using compromised login credentials. The compromised information contained personal details and medical prescriber IDs, which were reportedly leveraged in extortion attempts. From June to August 2024, the Japanese platforms Kadokawa and Niconico were targeted in a ransomware attack carried out by the Russia-affiliated hacking group known as BlackSuit. The attack led to the leakage of personal data of over 250,000 users. The incident highlighted vulnerabilities in Japan's cybersecurity infrastructure and prompted calls for enhanced cyber defense measures. In June 2024, Indian telecom company BSNL suffered a major data breach that exposed more than 278 GB of confidential telecommunications information. The breach exposed IMSI numbers, SIM card details, and security keys. In October 2024, Star Health Insurance disclosed a large- scale data breach that revealed the personal data of nearly 31 million policyholders. The leaked data, which included insurance claims, government ID cards, and medical reports.

The proliferation of computer hardware and digital technologies has significantly contributed to the spike in the number and complexity of malicious software over recent years. This rising trend is expected to continue without slowing down anytime soon.

As the volume of malware continues to escalate, it poses substantial security risks to both individuals and organizations. The impact of these threats includes monetary losses, interruptions to operations, and harm to reputation.. For instance, the average cost of a malware attack on a company reached approximately $2.4 million. Moreover, the increasing complexity of malware, including the use of artificial intelligence to enhance attack capabilities, further complicates detection and mitigation efforts.

II.MATERIAL AND METHODS

Malicious code refers to different types of software or scripts designed deliberately to harm, disrupt, or illegally access computer systems. These malicious programs are designed to exploit vulnerabilities, steal sensitive information, corrupt data, or

cause other forms of harm to devices and networks. These malicious codes can infect various platforms, including operating systems like Windows, Android, and iOS, as well as applications such as booking systems, hotel management software, and student management platforms. Up until now, typical examples of malicious code are viruses, worms, Trojan horses, ransomware, and spyware. The traditional methods for identifying this malware include signature identification; behavioral analysis; heuristic analysis; and sandbox analysis.

Signature Identification: This approach involves extracting distinctive features from known malware samples and storing them in a centralized database. When a new file is analyzed, its features are compared against those in the database. If a high degree of similarity is detected, the file is classified as malicious. Subsequently, the database is updated with the new sample to enhance future detection capabilities as shown in fig 1. For example, Yang et al.[4] used signature identification to analyze malicious apps in android. weight Zhang et al. [5] proposed a novel DAMBA approach to compare with signature-based recognition.

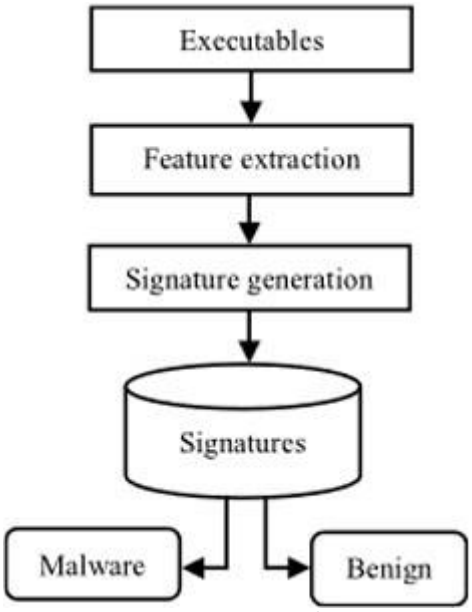

Fig 1: Signature Identification

U. Baldangombo et al. [6] proposes a static malware detection approach using data mining, focusing on feature extraction from Windows PE files. It applies Information Gain, PCA for dimensionality reduction, and classifies using SVM, J48, and Naïve Bayes, achieving 99.6% accuracy. A. Malhotra et al. [7] hybrid malware detection method that uses text mining and DBScan clustering on instruction sets extracted from executable files. By analyzing pattern similarities, it effectively classifies files as malicious or benign with improved accuracy and precision.

Behavior analysis: Behavioral detection is an advanced cybersecurity technique that monitors the actions of software during its execution to identify potential threats. Unlike traditional methods that rely on known signatures or predefined patterns, this approach focuses on observing real-time activities such as file modifications, network communications, and system changes as shown in fig 2. By analyzing these behaviors, security activities even if the malware has not been previously encountered or lacks a known signature.

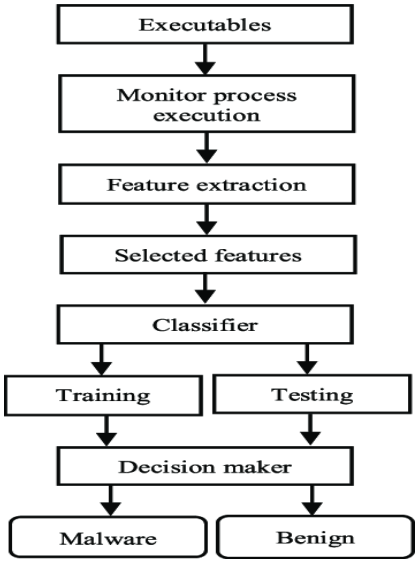

Fig 2: Behavior analysis

This approach is especially useful in defending against zero-day exploits and rapidly changing threats. For example, Rosli et al [8] Apply the K-means algorithm, an unsupervised learning technique, to perform clustering analysis on malware behavior. Ding et al. [9] Utilizes algorithms like KNN and SVM to classify behavior through a control flow-based analysis. S. Das et al. [10] proposes a real-time malware detection technique based on analyzing the semantic behavior of software during execution. It effectively identifies malicious activities by modeling execution patterns, enabling timely and accurate protection. M. Norouzi et al [11] suses dynamic analysis to capture malware behavior and converts the data for use with data mining tools. It applies cassifiers like Naive Bayes and J48 to accurately distinguish malicious activities from benign ones.

Heuristic analysis: This approach involves analyzing both static and dynamic characteristics of software to assess its behavior and determine if it is malicious. By examining the software's structural elements and runtime activities, this method can identify previously unknown malware variants as shown in fig 3. However, it is important to note that this technique is Exhaustive and tedious, requiring significant computational power and extended periods for thorough analysis.

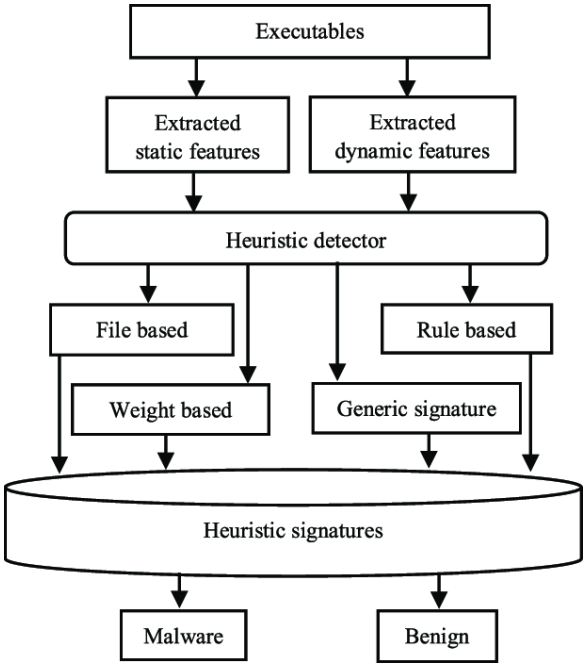

Fig 3: Heuristic analysis

For example, the SigMal [12] framework uses heuristic approach that utilizes signal processing techniques to analyze binary samples and identify similarities among them. Islam et al. [13] combines static and dynamic features into one classification model to improve malware detection accuracy and efficiency. It maintains strong performance with minimal accuracy loss on older malware samples. Naval et al. [14] presents a semantic analysis approach for malware detection that outperforms traditional signature-based methods. It effectively identifies malicious behavior, even in obfuscated code

Sandbox analysis: This technique involves executing software within a contained environment, such as virtual machines or containers, to observe its behavior and determine if it exhibits malicious activity. By monitoring actions like file modifications, network communications, and system changes, analysts can identify potential threats.

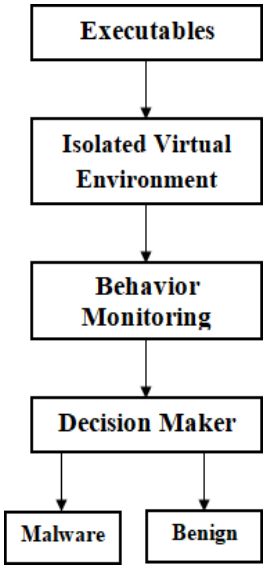

Fig 4: Sandbox analysis

However, this method is computationally intensive and time-consuming, requiring substantial resources and extended periods for comprehensive analysis. For example R.Rahul raj et al. [15] discusses the significance, working principles, and best practices for extracting valuable insights from malware samples within a controlled environment. As malware evolves, it increasingly employs advanced evasion techniques such as anti-virtualization and hook unhooking to bypass traditional detection methods. These sophisticated strategies enable malware to detect and circumvent analysis environments, rendering conventional sandboxing and debugging tools less effective. To address these challenges, researchers are integrating dynamic analysis with binary emulation frameworks, which offer faster and more scalable detection capabilities. While these emulation-based approaches may sacrifice some granularity, they provide a complementary layer of defense against evasive malware [16][17].

Conventional malware detection techniques, including signature-based and heuristic approaches, worked well when malware was simpler and easier to recognize. However, modern malware has evolved to operate at the kernel level, making it significantly more challenging to detect using conventional techniques. This advancement allows malware to execute with higher privileges and often remain hidden from standard detection methods. Consequently, these traditional approaches are increasingly inadequate for identifying sophisticated threats in today's cybersecurity landscape [18].

The emergence of machine learning and deep learning technologies presents a powerful opportunity to tackle the complexities of modern malware. These advanced techniques enable more efficient and accurate detection by analyzing complex patterns and behaviors within large datasets. For instance, ML algorithms can automatically extract relevant features from malware samples, while DL [19] models, like convolutional neural networks (CNNs), can learn hierarchical representations to identify intricate relationships that traditional methods might miss. This capability is particularly effective against sophisticated and previously unknown malware variants, enhancing the overall effectiveness of cybersecurity measures.

Hou et al. [20] introduced a machine learning-based technique to detect malicious web content, particularly targeting dynamic HTML (DHTML) attacks. Their method systematically analyzes web page characteristics and identifies key features, enabling effective detection even against obfuscated malicious code. D. Vasan et al [21] used CNN to classify malware by converting binary files into grayscale images and leveraging pre-trained CNNs for feature extraction. Kim et al. [22] introduced an Android malware detection model that utilizes strings, API calls, and permissions as key features. They trained a multi-modal deep learning model and evaluated it on a dataset of 41,260 samples, achieving a detection accuracy of 98.6%. This approach effectively identifies malicious applications by analyzing both static and dynamic characteristics. Zhenlong Yuan et al [23] Presented DroidDetector, sophisticated approach to identify malicious applications on Android devices based on deep learning that integrates both static and dynamic analysis features. Evaluated on thousands of apps, it achieved a detection accuracy of 96.76%, outperforming traditional machine learning methods.

Huang et al. [24] proposed a Windows malware detection technique that combines static and dynamic analysis through hybrid malware visualization. Utilizing Cuckoo Sandbox for dynamic analysis and converting results into visual images, they trained a VGG16-based convolutional neural network, achieving effective detection of unknown malware. Cui et al. [25] introduced a malware detection method that converts executable files into grayscale images, which are then analyzed using CNN. To address class imbalance in malware datasets, enhancing detection accuracy and reducing loss. Xu et al. [26] introduced DCEL, a classifier fusion method for Android malware detection that blends deep neural networks (DNNs) and convolutional neural networks (CNNs) to enhance detection accuracy. Patil S et al. [27] introduced SE-MADnet, a multiclass classification of malware images able to achieve more than 90 acciracy across different dataset.

III.CONCLUSION

The escalating threat landscape of cyberattacks underscores the urgent need for advanced malware detection and mitigation strategies. conventional malware detection methods, such as signature-based and heuristic analyses, are increasingly inadequate against sophisticated threats that operate at the kernel level and employ evasion techniques like anti-virtualization and hook unhooking. The integration of machine learning and deep learning technologies offers promising solutions, enabling more efficient and accurate detection by analyzing complex patterns and behaviors within large datasets. As cyber threats continue to evolve in complexity and scale, it is imperative to adopt advanced detection methodologies and proactive defense mechanisms to safeguard against potential risks and ensure the integrity of digital infrastructures.

REFERENCES

1. Av-Test: Security Report 2019/2020. [Online]. Available:ttps://www.avtest.org/fileadmin/pdf/security_report/AV-TEST_Security_Report_2019-2020.pdf
2. SonicWall: 2022 SonicwallCyber Threat Report. Accessed: 2022. [Online]. Available: ttps://www.sonicwall.com/resources/whitepapers/2022-sonicwall-cyber-threat-report.
3. Stationx: malware staticts report. https://www.stationx.net/malware-statistics/
4. X. Yang, D. Lo, L. Li, X. Xia, T. F. Bissyandé, and J. Klein, ‘‘Characterizing malicious Android apps by mining topic-specific data flow signatures,’’ Inf. Softw. Technol., vol. 90, pp. 27–39, Oct. 2017, doi: 10.1016/j.infsof.2017.04.007.
5. W. Zhang, H. Wang, H. He, and P. Liu, ‘‘DAMBA: Detecting Android malware by ORGB analysis,’’ IEEE Trans. Rel., vol. 69, no. 1, pp. 55–69, Mar. 2020, doi: 10.1109/TR.2019.2924677.
6. U. Baldangombo, N. Jambaljav, and S.-J. Horng, ‘‘A static malware detection system using data mining methods,’’ 2013, arXiv:1308.2831. [Online]. Available: https://arxiv.org/abs/1308.2831.
7. A. Malhotra and K. Bajaj, ‘‘A hybrid pattern based text mining approach for malware detection using DBScan,’’ CSI Trans., vol. 4, nos. 2–4, pp. 141–149, Dec. 2016.
8. N. A. Rosli, W. Yassin, F. M. A, and S. Rahayu, ‘‘Clustering analysis for malware behavior detection using registry data,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 12, 2019, doi: 10.14569/ijacsa.2019.0101213.
9. Y. Ding, W. Dai, S. Yan, and Y. Zhang, ‘‘Control flow-based opcode behavior analysis for malware detection,’’ Comput. Secur., vol. 44, pp. 65–74, Jul. 2014, doi: 10.1016/j.cose.2014.04.003.

10. S. Das, Y. Liu, W. Zhang, and M. Chandramohan, ''Semantics– based online malware detection: Towards efficient real–time protection against malware,'' IEEE Trans. Inf. Forensics Security, svol. 11, no. 2, pp. 289– 302, Feb. 2016.
11. M. Norouzi, A. Souri, and S. M. Zamini, ''A data mining classification approach for behavioral malware detection,'' J. Comput. Netw. Commun., vol. 2016, p. 1, Mar. 2016.
12. D. Kirat, L. Nataraj, G. Vigna, and B. S. Manjunath, ''SigMal: A static signal processing based malware triage,'' in Proc. 29th Annu. Comput. Secur. Appl. Conf., Dec. 2013, pp. 89–98, doi: 10.1145/2523649.2523682.
13. R. Islam, R. Tian, L. M. Batten, and S. Versteeg, ''Classification of malware based on integrated static and dynamic features,'' J. Netw. Comput. Appl., vol. 36, no. 2, pp. 646–656, Mar. 2013.
14. S. Naval, V. Laxmi, M. Rajarajan, M. S. Gaur, and M. Conti, ''Employing program semantics for malware detection,'' IEEE Trans. Inf. Forensics Security, vol. 10, no. 12, pp. 2591–2604, Dec. 2015.
15. Rahul Raj R, Naveen S, Subhikshan R, and Tarun S," Malware Analysis Using Sandbox" http://dx.doi.org/10.2139/ssrn.4708146.
16. S. Liu, P. Feng, S. Wang, K. Sun, and J. Cao, ''Enhancing malware analysis sandboxes with emulated user behavior,'' Comput. Secur., vol. 115, Apr. 2022, Art. no. 102613, doi: 10.1016/j.cose.2022.102613.
17. V. Vouvoutsis, F. Casino, and C. Patsakis, ''On the effectiveness of binary emulation in malware classification,'' J. Inf. Secur. Appl., vol. 68, Aug. 2022, Art. no. 103258, doi: 10.1016/j.jisa.2022.103258.
18. E. Gandotra, D. Bansal, and S. Sofat, ''Malware analysis and classification: A survey,'' J. Inf. Secur., vol. 5, no. 2, pp. 56–64, 2014, doi:10.4236/jis.2014.52006.
19. G. Gopinath and S. C. Sethuraman, ''A comprehensive survey on deep learning based malware detection techniques,'' Comput. Sci. Rev., vol. 47,Feb. 2023, Art. no. 100529, doi: 10.1016/j.cosrev.2022.100529.
20. Y.-T. Hou, Y. Chang, T. Chen, C.-S. Laih, and C.-M. Chen, ''Malicious Web content detection by machine learning,'' Expert Syst. Appl., vol. 37, no. 1, pp. 55–60, Jan. 2010, doi: 10.1016/j.eswa.2009.05.023
21. D. Vasan, M. Alazab, S. Wassan, B. Safaei, and Q. Zheng, ''Imagebased malware classification using ensemble of CNN architectures (IMCEC),'' Comput. Secur., vol. 92, May 2020, Art. no. 101748, doi: 10.1016/j.cose.2020.101748.
22. T. Kim, B. Kang, M. Rho, S. Sezer, and E. G. Im, ''A multimodal deep learning method for Android malware detection using various features,'' IEEE Trans. Inf. Forensics Security, vol. 14, no. 3, pp. 773–788, Mar. 2019, doi: 10.1109/TIFS.2018.2866319.
23. Z. Yuan, Y. Lu, and Y. Xue, ''Droiddetector: Android malware characterization and detection using deep learning,'' Tsinghua Sci. Technol., vol. 21, no. 1, pp. 114–123, Feb. 2016, doi: 10.1109/TST.2016.7399288.
24. X. Huang, L. Ma, W. Yang, and Y. Zhong, ''A method for windows malware detection based on deep learning,'' J. Signal Process. Syst., vol. 93, nos. 2–3, pp. 265–273, Mar. 2021, doi: 10.1007/s11265-020- 01588-1.
25. Z. Cui, L. Du, P. Wang, X. Cai, and W. Zhang, ''Malicious code detection based on CNNs and multi-objective algorithm,'' J. Parallel Distrib. Comput., vol. 129, pp. 50–58, Jul. 2019, doi: 10.1016/j.jpdc.2019.03.010.
26. X. Xu, S. Jiang, J. Zhao, and X. Wang, ''DCEL: Classifier fusion model for Android malware detection,'' J. Syst. Eng. Electron., vol. 35, no. 1, pp. 163–177, Feb. 2024, doi: 10.23919/jsee.2024.000018.
27. Santosh Patil, Prakash K Sonwalkar and Sagargouda Patil ," SE-MADnet: A Balanced Deep Learning Framework for Robust Malware Classification with Imbalanced Code Samples", 2nd International Conference on Computational and Cognitive Systems Science (ICCCSS 2025)- Volume 1, pages 96-104 ISBN: 978-989-758-824-2.