Journal Homepage: www.ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 02, Issue 08, pp 303-317, August 2025

The Evolution of Malware Sandboxing and Artificial Intelligence for Smart Threat Detection

Adeola, Olajide Olatunde[1]*, Alese, Boniface Kayode[2], Akinwonmi, Akintoba Emmanuel[1], Owolafe, Otasowie[2] and Omoniyi, Victoria Ibiyemi[3]

1Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria

2Department of Cyber Security, Federal University of Technology, Akure, Ondo State, Nigeria

3Department of Software Engineering, Federal University of Technology, Akure, Ondo State, Nigeria

**ooadeola@gmail.com*

ABSTRACT

One of the dangers of the increasing complex malware is the fact that new forms of cybersecurity protection should be created continuously. This report takes a look into how dynamic analysis, especially sandboxing has become a mandatory necessity in the detection of threats in the modern world, tracing its history dating back to an entry level analysis to complex threat hunting. It throws some light on how Artificial Intelligence (AI) and Machine Learning (ML) transform the methods of malware detection and enters the new level of the signature-based malware detection. The most relevant advanced developments of research, which relied on AI-assisted intrusion detection using sandboxing techniques, and literatures reviewed include emulated system artifacts and symbolic execution techniques. The other epic tragedies that this paper addressed include adversarial AI evasion, the insatiably broadening performance disparity between a regulated and real-world environment, and the fact that data quality and explainability are far greater hurdles than anticipated. In addition, it deals with the domain-specific application of these technologies within the sphere of the cybersecurity of the Industrial Control Systems (ICS). This paper has critically looked at the synergistic union between the AI/ML method and the malware sandboxing tool with the current problems, and what future research directions should be taken in a bid to overcome the challenge protecting cyber space against an opponent that is constantly evolving and globally becoming touchier than ever before.

Keywords: Artificial Intelligence, Machine Learning, Malware, Sandboxing, Threat Detection

1. INTRODUCTION

The era of digitalization has been characterized by an unprecedented rise of malware and malware is being found out every day and they are posing a threat to people, businesses and the infrastructure of nations on a daily changing and regular basis. The limitation of traditional signature-based detection is that they are not as effective when dealing with polymorphic, metamorphic and zero-day new malware they are specifically developed to avoid detection. The increasing level of sophistication of malware directly explains the need to change towards more dynamic and intelligent modes of detection as opposed to the older paradigms of static and signature-based detection. The financial cost of data breaches and the fact that the volume of new malware samples pales in comparison to the scale of it nowadays, with the AV Test Institute recording 450,000 new malware and potentially unwanted applications on a daily basis in 2022, and overall malware more than doubling between 2018 (450 million) and 2022 (970 million), reminds how necessary versatile and evolving are the detection methods (Liu, 2022). The ensuing cybersecurity arms race has thereby created a persistent game to react to the types of tricks criminals are using and offer appropriate protection.

Through sandbox environments in dynamic analysis, this has become key in achieving the real-life behavior of suspicious files by running them in a contained and controlled environment (Kaya, 2025). The method is important in solving the issue of obfuscation and interpreting the interactions at runtime that would be out of the scope of the static analysis. Sandboxes allow one to monitor system calls, network traffic, filesystems, and registry changes and supply critical information about how malware works (Guven, 2024). Switching and sandboxing of dynamic analysis is also one of the trends of behavior detection. This transition has been motivated by the poor performance of a static approach, which proves deficient against more malicious, slippery malware and thus the cybersecurity community is driven to be concerned with what malware does and not what it appears to be.

Artificial Intelligence (AI) and Machine Learning (ML) are becoming increasingly popular as useful assets that help to elevate the levels of malware detection, classification, and prevention (Kaya, 2025). They provide instruments of automated pattern recognition, anomaly detection and learning capability across large amounts of data therefore providing a way to detect malicious behaviors that cannot tall within the traditional rule-based systems. The use of AI is also limited to enhancing the quality of the analysis and increasing its accuracy in the fast-changing environment of threats. The number of new malware being detected is way too large and human analysis is thus not a viable option due to inherent flaws. Such automation and learning are the capabilities that AI and ML can provide, to process data of such volumes and respond to new threats as progressive and unpredictable, and thus become key elements (not auxiliary ones), of contemporary cybersecurity (Gupta, 2020). The incorporation of AI/ML is not just an addition to the improvement of scalable and adaptive cybersecurity services in case of excessive amounts of data and quickly altering threats.

This paper gives an interdisciplinary, in-depth examination and in-depth criticism of the synergistic combination of malware sandboxing with AI/ML. It will also discuss the current concepts/visions of sandboxing, how AI/ML algorithms may be used to detect pathogens in numerous ways in addition to the current limitations that continue to hinder their widespread usefulness. In detail, the following points will be addressed in this paper: the development of the sandbox technology and related analytical methods; the range of AI/ML models used to detect and classify malware and comments on Explainable AI; the serious constraints and challenges, especially the so-called adversarial AI and the performance gap of approaches in practice; the importance of high-quality malware repository and quality feature engineering; and application and peculiarities of the industrial control system protection. This report aims to propose the future lines of study and disperse important developments and ongoing challenges to strengthen cyber-shelter due to the ever-changing nature of the threats of the malware world, through a synthesis of recent literary works.

2. LITERATURE REVIEW

2.1 Artificial Intelligence-Enhanced Malware Sandboxing

The use of malware sandboxes has developed and progressed along the way, having become a more complex infrastructure of high-class malware hunting. The principle that they are based on is the ability to create a controlled and isolated environment to safely run suspicious code and to study its behavior in a safe manner without touching the host system. It is required to isolate so as to bring into the fore dangerous functionality that is either dormant or hidden in the analysis of the path when the path is analyzed at the static analysis. A comprehensive account is provided by Debas *et al.* (2024), though such concerns like tracing the development of sandboxes and emphasising the beneficial aspect of sandboxes in constructing a safe and engaging environment of a profound analysis of malicious code is mentioned. Rahul *et al.* (2024) also explain why they are important, how they are used, and how to derive meaningful information using the sandboxes, since they keep a record of the system calls, network interactions, and file system and registry modifications. The idea of sandboxes moving towards complex threat hunting can be specifically explained by the growing complexity of the malware and malware evasion techniques. When malware became even better at concealing what malware is, sandboxes have become increasingly real and dynamic to get the malware to show its real intentions.

Dynamically analyzing in sandboxes is very dependent on finding a suitable feature within the observed behaviors. Guven (2024) emphasizes the process of deriving the features of network traffic logs (pcap files) acquired in the sandbox that are, in their turn, interpreted as malware by the machine learning models. That indicates the value of the network level indicators in behavior detection. In addition to network logs, Liu *et al.* (2022) mention the role of improving sandboxes by creating realistic system artifacts with the help of an emulation-based system called UBER. It has the aim of making the sandbox environment look and feel more like the system of a real user, and in that way luring evasive malware into running its complete malicious code. This would enhance fidelity of honeypots and sandboxes making it look more like real time user behavior. The emphasis on the point of view dubbed as the realistic system artifacts and the emulated user behavior demonstrates the more sophisticated knowledge of the anti-analysis power of the malware. It means that, nowadays, malware willingly fingerprints its execution context, and hence, successful sandboxing involves taking an aggressor approach on the model to avoid being caught by fingerprints techniques. It is also more general regarding design philosophy of future sandboxes stating that they would need to actively mislead malware instead of passively monitoring it.

Although sandboxing can be effective it can only do so much and in particular against the malware which prevents sandboxing by detecting and avoiding the virtualized environments. Researchers are considering add-on strategies in order to overcome these weaknesses. Vouvoutsis (2024) offers to implement symbolic execution frameworks along with the sandbox execution to facilitate identifying new kinds of malware. The symbolic execution provides an insightful analysis of the malware code by examining every execution pathway found, and in this manner, can be employed to provide more informative signatures that will be resistant to polymorphic malware. That is not the case with sandboxes since there always is only one execution path at a time. The other technique which is essential is the Dynamic Binary Instrumentation (DBI) which executes and analyses evasive malware and gets the actual behavior of it, giving you complete control over files instrumented. The key strength of DBI discussed by Gaber *et al.* (2024) is that it can extract real, top-quality characteristics of evasive malware, as well as its overall resistance to anti-analysis techniques, proving resistant most of the time but can be overcome when using anti-instrumentation techniques which are effective against it but possess a countermeasure. DBI is opined to be better at extracting real features than either static analysis or sandboxes (though it has been demonstrated that sandboxes can be easy beaten by anti-sandboxing methods, and static analysis is also limited by use of obfuscation). Symbolic execution and DBI underline that traditional sandboxing is conceptually rather limited because it cannot ensure the code coverage and evasion technique bypass. This implies that the use of sandboxing alone is not perfect and should be blended with more thorough ways of analysis that should lead to powerful detection, another optimization approach to complete analysis.

2.2 The Artificial Intelligence (AI) with Machine Learning (ML) in Malware Detection

Machine Learning (ML) and Deep Learning (DL) AI was successfully implemented in most areas of malware detection. According to the systematic review, Gaber *et al.* (2024) state the main progression regarding its features; they mention such common ML algorithms as Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). Out of these, Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), Deep Belief Network (DBN), Deep Neural Network (DNN), and Convolutional Neural Network (CNN) DL architectures are some of the examples. Representatively, SVM, LSTM, and CNN-LSTM are effective in the case of Android applications (Alkahtani 2022, 100% on CICAndMal2017, 99.4% on Drebin). Seneviratne *et al.* (2022) present SHERLOCK, a deep learning model of self-supervision based on the architecture of Vision Transformer (ViT) that is used to detect malware by binary representation of image. Ensemble learning and hybrid models are also on the rise to implement superior detection and interpretability. The variance in the applied AI/ ML algorithms displays an ongoing quest to obtain the optimum models capable of dealing with the complexity and the volume of the malware data. This evolution of ML models through deep learning and into top-specific architectures (ViT) can be explained by the fact that more abstract and subtle patterns should be found in strongly obfuscated or new malware, and it leads to the fact that the complexity of malware increases the complexity of the AI models.

The systematic reviews conducted by Gaber *et al.* (2024) and Berrios *et al.* (2024) present relevant pieces of information that are crucial to the area. Gaber *et al.* mention that even though AI models are very promising, they can hardly be optimized to protect against novel, evasive and advanced malware due to their sensitivity to the quality and quantity of training features. They further talk of the impact of malware mechanism (malware is evasive, new, AI-aided) complexity on both the static and the dynamic analysis and inefficiency of the available malware data sets. Berrios *et al.* corroborate the fact that ML and DL far exceed traditional methods when it comes to enhanced malware and ransomware detection with the more advanced techniques demonstrating increased accuracies of finding zero-day malware types and ransomware. Nevertheless, they also identify the fact that some models are based on a particular collection of data which is not very representative of real-life diversity of threat activity, so generalization issues may arise. Despite having achieved high accuracies, deep learning and hybrid approaches have challenges in terms of cost and feasibility of calculation, interpretability and resistance to adversarial attacks. There is a stark difference between these high accuracy rates reported in research (they are frequently 90-99%) and performance problems in real life data and this leads to a high performance gap. Such indicates that the results of the research conducted in controlled conditions do not necessarily apply when it comes to the act of actually implementing them, which means a reviewing of the approach to evaluation is necessary, or some prioritization of robustness instead of the highest accuracy possible.

The dependence on the so-called black-box aspect of most of the AI models is a major challenge to cyber security because it complicates trust and the large-scale acceptance of the models. Galli (2024) introduces an Explainable AI (XAI) framework in behavioral malware detection and explains the approaches such as SHAP, LIME, LRP and Grad-CAM in making the decisions of AI models easily understood. XAI is transforming AI-related malware detection because it helps to demystify AI through the causal interpretation and comprehensibility of an AI process by reducing distrust, and enabling wide adoption. Such transparency is significant since it can allow the security analysts to know why a decision was chosen, which permits the incident response as well as model adjustments. The fact that XAI is emphasized means that there is a maturing industry in which technical performance (accuracy) is no longer enough. The right to question, to someone to be responsible and human beings to make interpretations on AI actions and take appropriate actions are also rising in importance. This is of a wider meaning to regulatory compliance and also human-AI collaboration model in cybersecurity, or systems have to not only be well performing but easily understandable to the human operators.

2.3 Threats and Constraints to the AI-Based Malware Detection and the Sandboxing

The development of adversarial malware binaries is a serious and increasingly difficult problem as such binaries are intended to disguise themselves against AI/ML detection. In lieu of this, Kolosnjaji *et al.* (2018) work on understanding how vulnerable deep networks trained on raw bytes are toward evasion attack, suggesting a gradient-based evasion attack that skews at-most 1 % of the bytes in a malware to attain a high probability of evasion whilst retaining functionality. The vulnerability illustrates a crucial weakness in which minor and in most cases barely noticeable transformations are exploitable by the advanced AI models. Malware developers are also increasingly putting AI frameworks into use, and developing what are dubbed as AI-powered malware, where the malware itself is reputed to be more evasive, and possessing a level of targeted attack where payload and trigger conditions are hidden using neural networks. It could be that such malware is immune to the analysis since the payload is encrypted and only decrypted on detection of the target and with the encryption key enclosed inside the black-boxes neural network. This new threat is also outlined in the Check Point Blog, and it points out the necessity of avant garde AI derived security systems to resist AI generated malware. The maturity of malicious AI and malware that improved with the help of AI is a predictor of the increase in cybersecurity in terms of AI arms race. There is a vicious cycle of the attackers using AI to counter AI defenses, and in turn, the good AI has to become more powerful and adaptive. This is a direct causal relationship in which developments of defensive AI would be the cause of developments of offensive AI, and the other way round.

One of the detailed issues that a recent study on the Kaya (2025) reported is that the proposed ML-based behavioral malware detection technique is by no means a silver bullet in practice. It is the first work that quantifies the performance on endpoints, meaning that it shows a large gap between performance during the training phase using sandbox traces (where most of the models claimed to achieve over 90% accuracy) and results when deployed on endpoint traces (with 20-50 True Positive Rates). The primary factors affecting this performance degradation are: distribution shift, such that there are massive differences in the execution traces between controlled sandboxed environments and diverse and uncontrolled real-world endpoint environments meaning that malware may be dissimilar due to environmental factors such as the hardware or network status; label noise, such as label errors or imprecision in the labeling malicious and benign samples in training datasets; and spurious features, unlikely to generalize to the real world because they appear distinctive in controlled settings, even though in real-world settings of diverse environments, they are not discriminating of the malicious A further widening of this gap is introduced by the presence of malware that has specifically been created in such a way as to not be detected in the situation where a system is not in operation under controlled conditions within a sandbox environment. The performance discrepancy is one of the possible flaws of the adopted AI/ML assessment strategies in the field of cybersecurity. This means that a lot of the claimed academic performance is not directly applicable (or may want to be directly applicable) to real-world security, and a paradigm shift needs to be adopted to train and test models directly on endpoints data. It is an urgent wake-up call to the field of research to be more in line with reality in the field of operation.

The restriction is also practical as certain AI models will demand a lot of computing resources and malware will constantly change. Great accuracy models can be very demanding in computation requirements as most deep learning models demand high computational consumption which is hindering deployment to space-limited environments. Model consistency is also threatened by the dynamism of malware which includes the possibility of behaving differently depending on the context or time. In addition to this, dynamic analysis is effective, but computationally more intensive, and slower than static methods. The computational complexities and resource requirements of an advanced AI model and the evolutions of the malware are the practical limitations to the effective deployment of the AI model in a computer system. This indicates an opportunity to study lightweight, computationally efficient, and flexible AI structures, and the techniques that are capable of dealing with the idea drifting and behavioral fluctuations in real-time to provide their continued efficiency in any operational environment.

2.4 Malware Collections, Quality of Datasets and Feature Extraction

One of the biggest bottlenecks in the process of developing malware detection based on AI is the challenge in developing open public collections of clean and malicious files. Desire to be legal, copyright, security liability can also encourage researchers to use small, closed, curated and usually unbalanced data, which is a hindrance to growth and reproducibility of research. What this data can cause is poor AI models that only succeed when applied to certain data with overfitting, and has no chance of being generalized. Large, versioned, continuously updated open repositories of linked Dynamic Binary Instrumentation (DBI) frameworks are highlighted to be the future research potential in terms of extracting the authentic features thereby (Liu, 2022). Most of the limitations observed in AI-based malware detection, particularly the performance disparity within the real world, have their root cause in lack of high-quality, varied and openly accessible datasets. It is a systemic problem of the research community as a whole because the effectiveness of accuracy and other model performances undoubtedly relies on the upper quality and authenticity of training characteristics.

Nonetheless, there are a number of publicly available data. Having the five publicly accessible datasets that have been considered by Gaber *et al.* (2024), they include EMBER, which features the extraction of the feature vector out of 1.1 million PE files without intact ones and only static analysis, BODMAS that consists of very recent and categorized malware, timestamps, intact PE files to study changes over time, SOREL-20M that has 20 million samples including the extraction of the feature vector and disarming malware files, VirusShare that consists of over 55 million live malware files S. Considering the malware detection datasets the references to which are presented by Berrios *et al.* (2024), one will single out R2-D2 (Android apps as RGB images), CIC-InvesAndMal2019, CICMalDroid 2020, Microsoft Malware (2015 competition), Malimg (visualization-based), Drebin, Malgenome (Android), Edge-IIoTset (IoT network traffic), NSL-KDD, CICIDS-2017, and Bot-IoT. Although there are a lot of datasets, they have various limitations (e.g., the lack of dynamic features, intact files, imbalance, or out of date), which points out to a big disconnect when it comes to the actual ideal high-level AI research. This provides a possibility of creating new ways of dataset creation, potentially incorporating a usage of generative AI in solving data scarcity and abundance issues.

The feature extraction, engineering, processing and feature selection are important and very problematic processes because of anti-analysis techniques and imbalanced datasets. According to Gaber *et al.* (2024), some of the major discriminative features that can be tracked down to detect malware are Windows registry interactions (tracking the malware persistence), CPU registers (behaviors at the byte level), file interactions (creation, deletion, modification), API and system calls (characteristic of PE/DLL files), frequency and order of the opcodes, conversion of the malware files and network traffic to images, and network traffic details (IP, ports, and protocols). Berrios *et al.* (2024) also classify the techniques of feature extraction by N-gram, Graph-based, Vision-based (converting binaries to images), and Hashing techniques. Generalizable AI models would be most relevant to malware development by understanding optimal syntactic and semantic characteristics of the malware language and characteristics that can separate malware and ordinary software. It is observed that models with accuracy above 99 % frequently employed dynamically derived characteristics such as API calls, network traffic, opcode streams and memory dumps. The performance of the AI models will depend on that extracted or rather the characteristics and relevance. The issue is that, it is hard to extract features which are not based on malware evasiveness but resistant to obfuscation and anti-analysis, which is evidence of the interactions being between the evasiveness of malware and feature engineering process being complex.

2.5 Protection of Industrial Control System (ICS)

Critical infrastructure such as power grid and water treatment systems are run using what is known as Industrial Control Systems (ICS), and malware and cyberattacks are being used to compromise ICS. Industrial communication protocols are vulnerable by virtue of the absence of implementations of any security schemes. It is in such delicate surroundings that AI and ML are being applied to assist in detecting intrusion and malware analysis. Umer *et al.* (2022) provide a review of applying ML in ICS intrusion detection applications on a network level and on the level of physical processes. Gupta *et al.* (2020) focus on the training of cybersecurity professionals on the AI/ML application to malware analysis towards safeguarding critical infrastructure. The importance of using AI-stimulated cybersecurity in ICS has been a high-stakes field of application because the result of a cyberattack could lead to disastrous effects physically. This realm brings up very different issues of manually operating in real-time, legacy systems, and safety-critical constraints that are completely different than the more common IT settings, and as such specialized solutions are essential.

Varghese *et al.* (2022) proposes a digital twin framework of ICS security, and the illustration used is of real-time intrusion detection using stacked ensemble of classifier-based classifications on the basis of AI. Realtime monitoring of digital twins, the analysis of threats, and the simulation of attacks in an attack scenario within a safe controlled environment is an answer to the challenge of testing attacks on live ICS. They achieved higher F1-Score and accuracy of their stacked ensemble, and the detection and classification speed of intrusions is 0.1 second. In Kravchik and Shabtai (2018), the efficiency of Convolutional Neural Networks (CNNs) is discussed to detect anomalies in ICS that proves to be an efficient way of detecting cyberattacks with minimal false positive results. They used statistical deviation of predicted versus observed values as their method, which achieved 31 attacks on 36 on a Secure Water Treatment testbed, higher than recurrent networks, demonstrating that CNNs are simpler, smaller, and faster when time series have to be predicted in ICS. DigitalTwin and CNNs can present novel methods to address the very challenge of ICS security, namely, safe simulation environment and effective anomaly detection. The problem solved in the first place in this solution is the inability to test on live systems and the necessity to deploy high-performance and low-latency detection to sensitive infrastructure.

Industrial threats Commercial products that apply AI to a greater extent in the field of the threat Industrial AI-based threat detection related products include MetaDefender Sandbox AI Threat Detection, FortiGuard AIBased Inline Malware Prevention Service, and Zscaler Cloud Sandbox. They exploit the real-time protection against unknown threats built on the basis of AI, thorough threat filtering, and zero-day attacks prevention. However, But with threats ever-changing, the latest frontier of malware technique is the AI Evasion. On a Check Point Blog one may read about the new threat of AI-spawned malware that has the ability to escape traditional sandbox detection mechanisms and focuses on the necessity of modern AI-based industrial-specific security solutions. In addition, Telefonicatech (2024) takes into consideration AI sandboxes to test AI modelling and securing them as they are also durable against hacking attacks and secure to operate them in industry. The recent appearance of commercial AI-based ICS sandboxes points to the fact that companies finally understood that sophisticated protection is highly necessary in this field. At the same time, AI-based malware that can exploit these systems is a considerable threat that will occur in the future, taking the AI arms race to deeply sensitive levels.

Table 1: Comparative Overview of Key Research on AI-Enhanced Malware Detection

S/N	Authors & Year	Paper Title	Problem Addressed	Method Used	Limitation	Findings
1	Guven *et al.* (2024)	Dynamic Malware Analysis Using a Sandbox Environment, Network Traffic Logs, and Artificial Intelligence	Malware classification based on network traffic	Feature extraction from pcap, ML/AI models	Not explicitly detailed in summary	Comprehensive approach to dynamic malware analysis; ML/AI models developed for classification
2	Liu *et al.* (2022)	Enhancing Malware Analysis Sandboxes with Emulated System Artifacts	Malware evasion of sandboxes	Emulation-based system (UBER) to generate realistic system artifacts	Not explicitly detailed in summary, but implies malware evasion	Enhances sandboxes by generating realistic system artifacts to improve fidelity
3	Debas *et al.* (2024)	Unveiling the Dynamic Landscape of Malware Sandboxing: A Comprehensive Review	Evolving malware analysis and threat detection	Review of sandbox progression, AI/ML integration, counter-evasion tactics	Ongoing evasive malware detection, need for comprehensive datasets, adaptability to zero-day	Traces maturation of sandbox technology from basic analysis to advanced threat hunting
4	Rahul *et al.* (2024)	Malware Analysis Using Sandbox	Combating sophisticated malware	Sandboxing, static and dynamic analysis, recording system calls, network activity, file/registry changes	Not explicitly detailed in summary	In-depth exploration of sandboxes, their significance, working principles, and best practices

5	Vouvoutsis (2024)	Beyond the Sandbox: Leveraging Symbolic Execution for Efficient Malware Detection	Detecting new malware strains efficiently	Complementing sandbox with symbolic execution frameworks	Not explicitly detailed in summary	Complements sandbox execution with symbolic execution for efficient new malware detection
6	Gaber *et al.* (2024)	Malware Detection with Artificial Intelligence: A Systematic Review	Key developments & core challenges in AI malware detection	Systematic review of AI in malware detection across 5 aspects	Malware sophistication (evasive, novel, AI-powered), dataset quality, analysis tool limitations	Comprehensive review of AI developments and challenges; DBI least impacted by anti-analysis
7	Galli (2024)	Explainability in AI-based Behavioral Malware Detection	Lack of transparent explanations in AI models	XAI framework (SHAP, LIME, LRP, Grad-CAM)	Global interpretability challenges, lack of comprehensive context, limited applicability	Proposes XAI framework; XAI revolutionizing AI malware detection by increasing transparency
8	Song *et al.* (2024)	A Study of the Relationship of Malware Detection and Artificial Intelligence	Combating various types of malware	AI implementation for malware detection	Lack of transparent explanations, dynamic analysis delays, evasion techniques	Explores AI for malware detection; some models achieve up to 99% accuracy with 90% consistent explanations

9	Alkahtani (2022)	Artificial Intelligence Algorithms for Malware Detection in Android Applications	Mobile malware detection efficiency	SVM, KNN, LDA, LSTM, CNN-LSTM, Autoencoder	ML cannot cope with huge data traffic; DL lacks optimization; some models not appropriate for complex data	SVM (100% on CICAndMal2017), LSTM (99.40% on Drebin) achieved high accuracy
10	Berrios *et al.* (2024)	Systematic Review: Malware Detection and Classification Using AI Techniques	Current trends and new methods for malware detection	Systematic review focusing on ML, DL, hybrid models	Generalization challenges due to specific datasets; computational cost, explainability, adversarial attacks	ML/DL outperform traditional methods; advanced techniques show superior accuracy for zero-day
11	Pfeffer *et al.* (2017)	Artificial Intelligence Based Malware Analysis	Volume, velocity, complexity, obfuscation of malware	MAAGI system (biologically/linguistically inspired), static/dynamic RE, hierarchical clustering, lineage analysis, component ID, trend prediction, functional analysis	Component uniqueness assumption, lack of data flow/temporal info, unpredictable attackers	MAAGI shows promising results in clustering, component identification, lineage, and trend prediction

12	Seneviratne *et al.* (2022)	Self-Supervised Vision Transformers for Malware Detection	Detecting previously unseen malware from unlabeled data	SHERLOCK (self-supervision, ViT architecture, image-based binary representation)	Not explicitly detailed in summary	Experimental results using 1.2 million Android applications
13	Kolosnjaji *et al.* (2018)	Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables	Vulnerability of deep networks to evasion attacks	Gradient-based attack (modifies <1% bytes)	Preserving functionality, avoiding syntax breakage, detecting/sanitizing added content	Achieves high evasion probability with <1% byte modification
14	Varghese *et al.* (2022)	Digital Twin-based Intrusion Detection for Industrial Control Systems	IDS deployment challenges in ICS, lack of testbeds	Digital twin framework, stacked ensemble classifier, process-aware attack scenarios	Unbalanced datasets (more normal than anomalous samples)	Stacked model outperforms previous methods (F1-Score, accuracy), near real-time detection (0.1s)
15	Kravchik, *et al.* (2018)	Detecting Cyberattacks in Industrial Control Systems Using Convolutional Neural Networks	Detecting cyber attacks in ICS	Anomaly detection via statistical deviation, 1D CNNs	Not explicitly detailed in summary	Detected 31/36 attacks with 3 false positives; 1D CNNs outperform RNNs for time series prediction in ICS

16	Gupta *et al.* (2020)	AI-Assisted Malware Analysis: A Course for Next Generation Cybersecurity Workforce	Shortfall of AI/ML trained cybersecurity professionals	Lab-intensive modules on CTI, malware analysis, classification, adversarial ML, APTs	Malware sophistication, classification complexity, API rate limits, resource constraints in labs	Development of comprehensive course curriculum to bridge talent gap

3. FINDINGS

The reviewed works clearly show that AI/ML and malware sandboxing have a strong synergy, leaving the static model of signature-based detection in favor of the dynamic model, which is based on observation (behavior). Such a movement is a major paradigm change in malware detection where malicious activities have become pro-active and intelligent in nature and based on behavior. This kind of transition is blamed on the sophistication of malware and inadequacy of the past approaches to it. An important progress involves the creation of advanced sandbox functionalities such as emulated system artifacts (UBER system) to mislead presumably evasive malware and the use of complementary methods like symbolic execution and Dynamic Binary Instrumentation (DBI) to identify more pronounced pieces of information. The use of a wide range of different AI/ML models, such as SVM and RF and the more sophisticated ones, like LSTM, CNN, and ViT, has substantially increased detection and classification rates, especially on new and polymorphic malware. Also, the development of Explainable AI (XAI) systems is another big step on the way to developing trust and comprehensibility in these complex systems, which is the problem of a black box.

Most of the AI/ML algorithms display good performance but it is not applicable in all settings. In the case of Android malware, SVM models, as well as LSTM, have displayed a remarkably high accuracy of 100 % and 99.40 % respectively on certain sets of data such as CICAndMal2017 and Drebin. DNN and CNNs are also deep learning models that have shown an improvement in malware detection in Windows using both static and dynamic features compared to the traditional ML. However, the article ML-Based Behavioral Malware Detection Is Far From a Silver Bullet shows that the difference between the results obtained by sandbox datasets and truly harmful malware is chilling since its True Positive Rate dropped more than 10 times (20-50%) higher when deployed on real-world end-point traces due to distribution shift, label noise and spurious features. This gives rise to what may be one of the greatest shortcomings of the existing models and questioning their applicability and validity outside of controlled conditions. The difference between this contextual performance and generalization challenge that has been observed means that despite AI/ML providing tools of such power, the ultimate goal of delivering strong and generalizable detection under the different operational circumstances is a remarkably challenging task. Achieved accuracies of AI/ML models are subject to context and do not necessarily generalize to varied and realistic inferences, which implies that performance is highly influenced by the environment where the algorithm was trained and tested.

An ongoing AI arms race can be found in the cybersecurity landscape as the development of AI-based defensive technology is matched by the emergence of adversarial AI and AI-powered malware aimed at devising methods that cannot be detected. Attacks based on adversarial examples have the capability to fool deep learning models without significant changes, and malware applications through the use of AI have the ability to conceal their payloads until certain conditions have been satisfied and thus hard to analyse. Besides such adversarial threats, deployment logistics in actuality are not very easy. One of the crucial concerns is the problem of the distribution change and environmental differences that existed between the values of sandbox and endpoint performance. The complexities in the implementation of deep learning models are also inflated by the fact that some deep learning models are computationally intensive and that malware is dynamic and always evolving. The real issue is the availability of rich, balanced and authentic data sets of malware, which is a big bottleneck on its own and this causes the model to be fragile and not portable. These difficulties are not independent of each other but are interrelated in the process of creating a feedback loop This level of malware sophistication is what causes the advancement of AI to rise to a higher level and this is what has led to the development of adversarial AI which continues to increase in depth and depth. This fluidity entails that the implementation of changes in the real-world is characterised by in-built complexity, and the process necessitates constant adjustments.

The weakness of static analysis when it comes to obfuscated malware has driven the significance of behavioral analysis which is the process of monitoring the actions of malware as it runs on the system to reveal its real purpose. It is more difficult to object to such an approach when it comes to such veiling on the code level. Also, the opacity (or black-box) of many AI models has necessitated the pressing need of Explainable AI (XAI), which offers transparency and increases the level of trustworthiness, elucidating the decisions made by the model. This is crucial to the human analyst who wants to take action based on the information produced by AI. Finally, the argument of direct connection between quality and authenticity of training features and the performance of the AI models supports the idea that such a practice as the development of large and diverse malware collection with an appropriate level of representation is important. Future malware detection systems will need to focus on having behavioral analysis, XAI and robust practices on the datasets in order to achieve higher success rates. They are the most essential infrastructures behind succeeding in the establishment of trustful, competent, and flexible AI-powered cybersecurity systems. The repetition of these issues shows that they are not just the small changes that can be made but essential ones to achieve a good high in the field.

3.1 Detection Accuracies and Performance

In many cases, the AI/ML models have also been demonstrated to be accurate on detecting in controlled environments. A case in point is that SVM on the CICAndMal2017 dataset achieved an accuracy of 100 % when detecting Android malware and LSTM had an accuracy of 99.4% when detecting Android malware on the Drebin dataset. Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) achieved an outstandingly high accuracy levels of 98.9% and 96.6% respectively by using static and dynamic feature representations in Windows malwares. Moreover, ensemble classifiers have achieved very high accuracy of 99 % with 90 % consistency on dynamic malware dataset explanations. These results indicate that AI/ML will be very promising in malicious software detection.

3.2 The Gap in Performance

Notwithstanding those great lab-based levels of accuracy, there is also an important methodological gap that we can discern when these models are put to use in real life. An empirical analysis of the ML behavioural detection of malicious malwares identified an enormous performance decline upon actual deployment to an endpoint detection (where performance was measured by False Positive Rate FPR and True Positive Rate TPR). Such sharp contrast highlights the fact that the profitability of the studies models might not be train and measured under conditions realistic to the operational environment with all its complexity and variety. This implies that the process by which assessment is carried out should also be critically re-visited in order to come up with findings that can really be implemented in real-life security scenario.

The Sandbox Performance: When comparing Cuckoo Sandbox and Process Monitor (Procmon), Cuckoo Sandbox was faster (average time on a comparative number of samples was about 530 seconds with Cuckoo compared to 989 seconds (highest score) with Procmon) and identified an average sample with a higher level of accuracy (99.35 % average accuracy versus 94.48 %) and ROC (0.97 versus 0.91). This implies that there are sandbox environments that are more effective and practical in delivering data to be analyzed by means of AI.

Adversarial Evasion: The performance of the AI models is antagonistic to attacks. Indeed gradient-based attacks have proven evasive to detection by deep learning using only around 1 byte of modifications with a high probability of evasion. The digital artefact proves the fact that the current AI systems are fragile against sophisticated informed attacks.

ICS Detection: Digital twins of stacked ensemble classifiers in ICS world have outperformed F1 -Score and accuracy of invading detection, in our implementation of the human in the loop (HITL) in real-time environment, we recorded average latency determination\ delay of 0.1 second. Further, we found that 1D Convolutional Neural Networks identified 31 out of 36 cyberattacks in ICS with low false positives and therefore, seems to be an effective neural network that outperforms recurrent networks in specific contexts. The possible implications of these findings are that tailored AI can prove to have a very big impact in small-scale, high-stakes contexts.

MAAGI System: The MAAGI system, presented by Pfeffer *et al.* (2017) and based on the methods inspired by biology, promises certain results in a number of areas. It overcame the challenge of hierarchical clustering as it also offered low negative accuracy when compared to batch clustering but at a speed that was much faster (less than an hour compared to more than five hours). The system was effective with the identification of shared components in component identification by applying the features of gen code and gen semantics. In addition, lineage analysis conducted by it showed high accuracy related to lineage reconstruction of malware. Such qualitative results lead to the future potential of novel and interdisciplinary forms of analysis of malware.

The comparison between the very high accuracy rates that are obtained in the laboratory and the real world measure of performance which has been lower by far represents an awful methodological gap in the field. It means that more realistic assessment criterion and settings should be placed on the priority list of further research to guarantee applicable relevance.

4. CONCLUSION AND RECOMMENDATIONS

Contemporary status of AI-aided malware sandboxing is a significant step in the development of security measures in the sphere of cybersecurity. There has been an enormous leap forward in the realm of integrating Artificial Intelligence and Machine Learning with dynamic analysis techniques and ceasing to rely solely on signature-based approaches to threat detection and expanding towards behavioral analysis as a means of enhancing robust threat detection. This ongoing research can be summed up by the creation of sandboxes capable of dealing with emulated artifacts of the target system, and by blending the related techniques like symbolic execution and Dynamic Binary Instrumentation. Although AI/ML models can achieve remarkable accuracies under controlled settings, a combination of challenges that can be described as both persistent and interconnected challenges marks the field, namely the growing AI arms race with malware that is adversarial in nature and the severe performance discrepancy between lab and in-the-wild endpoints deployments that has been noted before. The malware dataset quality and accessibility and a desire to have better explainability in malware tasks remains central impeding blocks to overall efficacy and trust.

In order to contribute to the development of the field and to reinforce cyber defense against the ever-changing land scapes of the malware threats, future research are coming forth by focusing on Greater Explainability AI(XAI), Highly Resistant to Adversarial AI, Public Quality Datasets and Industry Control System (ICS)-Specific Solutions.

References

Alkahtani, H. (2022). Artificial Intelligence Algorithms for Malware Detection in Android Applications. *Journal of Computer Virology and Hacking Techniques*, *18*(3), 349–361. https://doi.org/10.1007/s11416-022-00444-7

Berrios, S., Leiva, D., Olivares, B., Allende-Cid, H. & Hermosilla, P. (2024). Systematic Review: Malware Detection and Classification Using AI Techniques. *Applied Sciences*, *15*(14), 7747. https://doi.org/10.3390/app15147747

Check Point Blog. (2025). AI Evasion: The Next Frontier of Malware Techniques. *Check Point Research*. Retrieved from: https://blog.checkpoint.com/artificial-intelligence/ai-evasion-the-next-frontier-of-malware-techniques/

Debas, E., Alhumam, N., & Riad, K. (2024). Unveiling the Dynamic Landscape of Malware Sandboxing: A Comprehensive Review. *International Journal of Advanced Computer Science and Applications*, *15*(3). https://doi.org/10.14569/IJACSA.2024.01503137

Fortinet. (2024). FortiGuard AI-powered Security Services counter threats in real-time with AI-powered, coordinated protection across your entire attack surface. *FortiGuard AI-based Inline Malware Prevention Service*. Retrieved from: https://www.avfirewalls.com/Fortiguard.asp

Gaber, M. G., Ahmed, M., & Janicke, H. (2024). Malware Detection with Artificial Intelligence: A Systematic Review. *ACM Computing Surveys*, *56*(6), 1–37. https://doi.org/10.1145/3638552

Galli, A., Gatta, V. L., Moscato, V., Postiglione, M., & Sperlì, G. (2024). Explainability in AI-based behavioral malware detection systems. *Computers & Security*, *141*, 103842. https://doi.org/10.1016/j.cose.2024.103842

Gupta, M., Mittal, S., & Abdelsalam, M. (2020). *AI assisted Malware Analysis: A Course for Next Generation Cybersecurity Workforce*. arXiv. https://doi.org/10.48550/arXiv.2009.11101

Guven, M. (2024). Dynamic Malware Analysis Using a Sandbox Environment, Network Traffic Logs, and Artificial Intelligence. *International Journal of Computational and Experimental Science and Engineering*, *10*(3). https://doi.org/10.22399/ijcesen.460

Kaya, Y., Chen, Y., Botacin, M., Saha, S., Pierazzi, F., Cavallaro, L., Wagner, D. & Dumitras, T. (2025). ML-Based Behavioral Malware Detection Is Far From a Solved Problem. Cryptography and Security. Available at: https://doi.org/10.48550/arXiv.2405.06124

Kolosnjaji, B., Demetrio, D., & Schiele, G. (2018). *Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables*. arXiv. https://doi.org/10.48550/arXiv.1803.04173

Kravchik, M., & Shabtai, A. (2018). Detecting Cyberattacks in Industrial Control Systems Using Convolutional Neural Networks. In *CPS-SPC 2018 - Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy, co-located with CCS 2018*, 72–83. Association for Computing Machinery. https://doi.org/10.1145/3264888.3264896

Liu, S., Feng, P., Wang, S., Sun, K., & Cao, J. (2022). Enhancing Malware Analysis Sandboxes with Emulated System Artifacts. *Computers & Security*, *115*, 102613. https://doi.org/10.1016/j.cose.2022.102533

OPSWAT (2024). MetaDefender Sandbox AI Threat Detection. Retrieved from https://www.opswat.com/docs/filescan/datasheet/url-analysis

Pfeffer, A., Ruttenberg, B., Kellogg, L., Howard, M., Call, C., O'Connor, A., Takata, G., Reilly, S. N., Patten, T., Taylor, J., Hall, R., Lakhotia, A., Miles, C., Scofield, D., & Frank, J. (2017). *Artificial Intelligence Based Malware Analysis*. arXiv. https://doi.org/10.48550/arXiv.1704.08716

Rahul, R.R., Naveen, S., Subhikshan, R. & Tarun, S. (2024). Malware Analysis Using Sandbox. *SSRN*. https://doi.org/10.2139/ssrn.4708146

Seneviratne, S., Shariffdeen, R., Rasnayaka, S., & Kasthuriarachchi, N. (2022). Self-Supervised Vision Transformers for Malware Detection. *arXiv*. https://doi.org/10.48550/arXiv.2208.07049

Song, J., Choi, S., Kim, J., Park, K., Park, C., Kim, J., & Kim, I. (2024). A Study of the Relationship of Malware Detection and Artificial Intelligence. *ICT Express*, *10*(3), 632–649. https://doi.org/10.1016/j.icte.2024.03.005

Telefonicatech (2024). AI Sandboxes: Secure Environments for Testing and Protecting Artificial Intelligence Models. Retrieved from: https://telefonicatech.com/en/blog/ai-sandbox-secure-environments-for-evaluating-and-protecting-artificial-intelligence-models

Umer, M. A., Junejo, K. N., Jilani, M. T., & Mathur, A. P. (2022). Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations. *International Journal of Critical Infrastructure Protection*, *38*, 100516. https://doi.org/10.1016/j.ijcip.2022.100516

Varghese, S. A., Ghadim, A. D., Balador, A., Alimadadi, Z., & Papadimitratos, P. (2022). Digital Twin-based Intrusion Detection for Industrial Control Systems. *arXiv*. https://doi.org/10.48550/arXiv.2207.09999

Vouvoutsis, V., Casino, F., & Patsakis, C. (2024). Beyond the sandbox: Leveraging symbolic execution for evasive malware classification. *Computers & Security*, *149*, 104193. https://doi.org/10.1016/j.cose.2024.104193

WebAsha. (2024). AI-Powered Malware Analysis: How Artificial Intelligence Detects and Prevents Malware Attacks. Retrieved from: https://www.webasha.com/blog/ai-powered-malware-analysis-how-artificial-intelligence-detects-and-prevents-malware-attacks

Zscaler (2024). Zscaler Cloud Sandbox - AI-Powered Malware Defense. Retrieved from: https://www.zscaler.com/products-and-solutions/cloud-sandbox