An Evaluation of Current Malware Trends and Defense Techniques: A Scoping Review with Empirical Case Studies

Azaabi Cletus [1,2,*], Alex A. Opoku [3], and Benjamin Asubam Weyori [4]

[1] Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana
[2] Department of Mathematics and ICT, St. John Boscos College of Education, Navrongo, Ghana
[3] Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana
[4] Department of Computer and Electronic Engineering, University of Energy and Natural Resources, Sunyani, Ghana
Email: cleinhim@yahoo.com (A.C.); alex.opoku@uenr.edu.gh (A.A.O.); Benjamin.weyori@uenr.edu.gh (B.A.W.)
*Corresponding author

***Abstract*—The growing armed-race between malware attackers and defenders makes the malware ecosystems highly volatile, dynamic, stochastic, and unpredictable. The volatility of the ecosystem means that, both attackers and defenders are innovating to outwit each other, which requires regular evaluation to establish gaps for remediation. In this paper, the aim was to establish current malware trends, comparative weaknesses and strengths of existing malware defenses, the identification of research gaps and a proposal of future directions to malware defense. We adopted a scoping review with empirical case studies using data from extant literature and industrial sources for the study. The results revealed that, current malware are targeted, unknown, persistent and stealth and are increasing in volumes, variety and complexity. Attackers adopt innovative modes of transmission to spread malware from one network to another and use both anti—static and advanced forms of obfuscation to evade detection. The poor adaptability, learnability, memorability and generalizability of signature-based detection methods such as static, dynamic, hybrid makes ML algorithms the state-of-art, but they also show instability in classification, poor and redundant features, class imbalance and the associated "accuracy paradox", and poor resilience to detecting previously unknown malware. Additionally, user and organizational vulnerabilities also exacerbates the defense challenge. The paper concluded that with the increasing sophistication in malware, ensuring holistic malware defense requires novel techniques that addresses these gaps. This implies that, current research should refocus on providing hybrid defense approaches that are not only technical in nature but also non-technical leading to the provision of improved holistic malware defense.**

***Keywords*—vulnerability, malware, obfuscation, dynamic detection, static detection, hybrid detection, coping review, ransomware, scoping review**

I. Introduction

Globally, malware and its defense remain a major obstacle facing both researchers and practitioners [1, 2]. This is due to the growth and expansion of the internet, smart mobile telephony, the expansion in Internet of Things (IoTs), and the increased digitization and digitalization. Though these technology penetrations are all positive signals to the global cyber-inclusion, it has come with corresponding increase in cyber risk mainly using malware [3]. Malware (malicious software) and Potentially Unwanted Software (PUS) is a broad term used to describe software with malicious intent that causes harm to computing resources and information systems. There are varieties of malware types, including but not limited to Trojans, viruses, worms, rootkits, ransomware, bots, etc. Malware mainly steal confidential information, which can be used to disrupt systems, destroys systems, takes control and command of systems, spreads spam, and destroys critical infrastructure when they get into s system [4, 5]. A successful malware attack compromises the confidentiality, integrity and availability principles of information security and results in financial loss, reputational loss, and regulatory and compliance issues [6]. Thus, the concomitant effect of malware exposure is so dire that, stakeholders are spending hugely on providing defenses against malware attacks. Notwithstanding the huge investment, cyber exposure through malware attacks is still in ascendancy [7]. According to [8], on average, about 450,000 malware samples and PUS are recorded daily; this translates into 13,500,000 monthly and 162,000,000 annually. The report suggests that malware is not only increasing in volume but also in variety and complexity as it adopts novel techniques to mutate and change behavior depending on the medium and location. This adaptability, results in new malware with new signatures and characteristics that poses challenges to efficient and effective defense. In addition, with the growth of open-sourced tools and other techniques coupled with user and organizational vulnerabilities, malware assailants now adopt novel obfuscation techniques to evade detection [9, 10].

Interestingly, the rate of malware creation and deployment and the introduction of new and novel

Manuscript received June 20, 2023; revised September 18, 2023; accepted November 2, 2023; published May 28, 2024.

techniques to outwit defenses requires that malware defense teams equally develop corresponding defense techniques or countermeasures to contain these new malware attacks. This competition between the attackers and defenders has resulted in making the malware ecosystem highly volatile, revolving, evolving and dynamic ecosystem as both malware defenders and attackers constantly innovate to outwit each other. These rampant changes in approach by both the adversaries and the defenders implies a fluid, dynamic, and highly stochastic ecosystem. Consequently, this highly dynamic nature of the ecosystem makes it challenging researchers and industry to establish new developments and gaps that requires remediation in order to bring improvement. Understanding the current malware trends such as the volumes, variety and complexity that poses challenges to efficient and effective defense would assist researchers and industry to proposed improved techniques. Hence, the need for regular evaluation or reviews to establish the current state-of-art about the trends, attack and defense strategies and the research gaps prevalent in current research approaches, which is required regularly in malware research [11]. Therefore, this paper evaluated the current trends (volumes, variety, and complexity) and defense techniques to identify the weaknesses and research gaps for improved remediation. This was achieved using scoping review methodology supported with empirical case studies. The outcome led to better understanding of the variety of current malware and other socially engineered attacks, which results in building a light malware ontology. We also established the various modes by which malware is transmitted from one network to the other, and established the comparative characteristics of current and traditional malware that negatively affects defense method. In addition, the study led to better understanding of the various obfuscation techniques usually adopted by malware authors to conceal the identity of malware in order to evade detection and to exploit systems. Moreover, we established the limitations of the various signature-based malware defense methods such as static, dynamic, and hybrid approaches, which has made Artificial Intelligence (AI) and Machine Learning (ML) techniques as the state-of-the-art in automated malware defense. Finally, we identified and validated the research gaps in the use of ML techniques in malware defense and proposed projections to assist cybersecurity stakeholders in making the right decision when choosing tools and techniques for malware analysis.

II. Works

In recent times, a plethora of reviews on malware attacks and defenses has been explored. Raghaendra *et al*. [12] conducted a survey on Machine learning in malware detection. Their approach dwelt on only analysis techniques and sharing light on the use of ML in windows environment using portable executables. They concluded that, as malware authors innovate to evade detection, both industry and the research community should correspondingly react with superior techniques to stay above the competition. Faruki *et al*. [13] conducted a survey and evaluation of malware detection in android-based environments. Their concentration was on the detection techniques and detection frameworks. Bilot *et al.* [14] conducted a survey on malware detection based on graph representation learning. They mainly summarized the graph-based deep learning for malware detection and adversarial attacks. An overview of advances in malware detection in terms of methods, accuracy, and other evaluation metrics by Heena and Mehtre [11]. However, their approach did not indicate the future direction based on the study and the evasive techniques used by malware authors. In a similar vein presented a survey on malware classification using machine learning and deep learning techniques was presented in Ref. [2]. Their approach focused on malware features used and the stages in the machine learning and deep learning methods. A study on metamorphic malware and obfuscation techniques was conducted by and dwelt mainly on malware variants and generation kits. Merabet and Hajroui [15] provided a brief survey on malware detection techniques based on machine learning methods and mainly focused on the machine-learning pipeline, describing what to do at each stage. Similarly, Hamza *et al.* [16] did a survey on malware and dwelt mainly on the feature transformation techniques for data streams. They contend that many features lead to high computational resource usage and further provide a feature-dimensionality reduction method for malware analysis. A review of malware issues, challenges, and future directions was presented in Ref. [17] and concluded that generic detection and feature extraction methods are critical parts of malware detection systems and should be handled properly in building malware detectors. To understand the limitations and strengths of signature-based defense techniques, a brief survey on malware analysis techniques: static, dynamic, hybrid, and memory-based analysis techniques was conducted in Ref. [18]. They provided malware types, detection methods, analysis techniques and their limitations, malware obfuscation, and anti-analysis methods, including memory analysis. In addition, they presented the merits and demerits of each of them. Furthermore, Sihwail *et al*. [19] conducted a survey of machine-learning techniques for malware detection and concluded by demonstrating the extent of the use of ML in malware detection so far. Ucci *et al.* [20] did a survey on ML techniques for Android malware detection. They concluded that the use of static, dynamic, and hybrid approaches has its limitations at each point of use. Kouliarridis and Kambourakis [21] conducted a systematic literature review of Android malware detection using static analysis. They concluded that static methods still face many challenges and recommended novel techniques for enhancing the approach. Likewise, the conduct of a review on vulnerability mainly to assess the individual/user, organizational, software/operating system, and technical defense systems has also been employed [6, 22]. Similarly, Paula *et al.* [23] contend that, even though technical defense in malware attacks and social engineering is a requirement, the overreliance on only technical techniques to the neglect of non-technical controls leads to

compromise of security. A position supported by Alhashmi *et al*. [7] and Musah *et al*. [24]. They then advocated for the conduct of regular, routine, and well-organized user awareness, training, and education to build user resistance and immunity to malware attacks.

Notwithstanding the plethora of surveys or reviews on malware attacks and defense, an analysis of the forgoing literature, suggest some deficiencies and gaps that requires a new approach in handling the current topic as explained in the proceeding paragraphs.

Firstly, there appears to be no single review study making use of literature review outcomes supported by case study results. Hence, our approach where the literature results or outcome is augmented by a case study provides a holistic view of the problem. By adopting this approach, the review outcome is not only founded on the bases of only extant literature, but also supported by empirical results from industry or stakeholders. Thus, further improving the validity and reliability of the outcome of the study.

Similarly, there is less focus by various reviews and surveys on current malware trends in terms of volumes, variety, and complexity using hybrid data sources from extant literature and industry. By exploring current malware trends would reveal the volumes, variety, characteristics and their mode of transmission for efficient defense approaches and decision-making purposes.

Besides, there has been few or no recent studies focusing on various obfuscation techniques, their description and the integration or classification of obfuscation methods based on their characteristics. Exploring the various obfuscation techniques used by malware attackers and the categorization into anti-static methods and advanced methods will assist in the development of novel techniques to counter such attacks and assist in building immunity against such attacks.

Moreover, the available works rarely provides a comparative tracer study of the relative strengths and limitations of the various malware defense techniques in the face of the prevailing malware ecosystem. A comparative analysis of the strengths and limitations of the various defense techniques (static, dynamic, hybrid, and ML) will enable cyber-defense stakeholders in effective decision-making when choosing a tool for defense and assist the research community to propose patches to the identified vulnerabilities to improve malware defense.

Finally, there is no integration of the identified research gaps in the use of machine learning methods, which is the current research focus in automated malware detection. The identification of these research gaps in current research will assist the research community and industry to find novel defense tools for improved malware defense.

Consequently, this paper, presents an exploration of current malware trends, malware obfuscation methods, the comparative evaluation of the strengths and limitations of malware attack and defense strategies, and identifies gaps in ML techniques used for malware detection. In addition, the outcome of the literature review was augmented by the results of the empirical case studies on malware trends, and user and organizational vulnerabilities. This provides relevant, rich, and practical insights into malware defense, and assist stakeholders in defense choices. On the bases of the achieved objectives, this study made a modest contribution to knowledge in malware defense in particular and cybersecurity in general by highlighting key insights. Specifically, the following are the key highlights of the study:

- We highlighted the current malware trends from both industrial and extant literature contexts, established the various malware transmission modes, and conducted a comprehensive comparative analysis of the characteristics of current malware over traditional malware that make it difficult for signature-based detection systems to detect, leading to high false positives.
- The study presented an update on the various anti-static and advanced obfuscation techniques used by malware attackers to exploit detection systems. Knowledge of these obfuscation techniques assists malware analysts in identifying obfuscated malware and planning improved defenses against such methods.
- We highlighted the various limitations and vulnerabilities of signature-based detection systems at the technical defense level and user and organizational vulnerabilities relative to current malware attacks. This rich and practical insight offers malware defense teams the opportunity to improve their detection and preventive defenses against malware attacks and to develop alternative detection and preventive methods for improved malware defense.
- Finally, the study highlighted the key research gaps in current malware research. These include poor malware features for efficient malware detection, class imbalance problems leading to the accuracy paradox, the use of obfuscation techniques to evade detection, the problem of conventional ML techniques still in use by some legacy systems, and the problem of user and organizational vulnerabilities as challenges to efficient and effective malware defense. This knowledge affords researchers the opportunity to apply novel techniques to overcome the identified gaps, leading to improved malware detection.

The paper is organized as follows: In Section II, we present Related Works, the Background of the Study is presented in Section III; Section IV comprises the Materials and Methods, whiles section V contains the Results and Discussion of the study. In Section VI, we present Current Malware Research Approaches, whiles Section VII is the Identified Research Gaps in current research, Section VIII is made up of suggested Future Directions whiles Section IX details the Conclusion and Future Works of the study.

III. Background

In this section, we present the background materials of the study. These include a brief discussion of Malware and Malware Trends, malware obfuscation techniques, the

relative strengths and limitations of malware defense techniques, user/organizational vulnerabilities to malware attacks, the current research focus using ML and AI techniques.

A. Malware and Malware Trends

Globally, malware attacks and exposures continue to attract all relevant stakeholders in the cybersecurity value-chain as more and more exposures occur [7]. Malicious Software (Malware) are software with malicious intent. They are codes that propagate malicious actions when found in a computing system [5] and can be in the form of an executable file, code, a script and or any software. They mainly steal information such as credit card details, passwords, disrupt systems such as Denial of Service (DoS) and Distributed Denial of Service (DDoS), and can take control of an entire system such as Ransom attacks when compromised. Averagely, 450,000 malware and Potentially Unwanted Software (PUS) are recorded daily which translates into 13,500,000 monthly and 162,000,000 annually [8].

Fig. 1 shows a 10-year trend or growth of malware and Pus. From the figure, both malware and PUS increases steadily over the past decade.

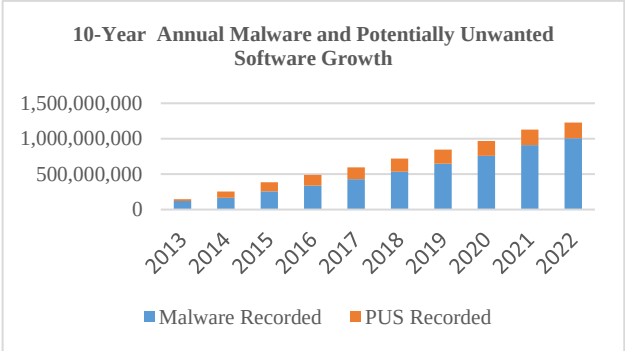

Fig. 1. A 10-year malware and pus growth.

In addition to the volumes, these malwares are also increasing in variety and complexity as malware authors adopt novel approaches to create new malware samples and launch them rapidly against their targets leading to exposures. Fig. 2 show the total number of vulnerabilities from 2013–2022. As depicted, there is an increasing trend of vulnerability exposures over the 10-year period. These increases and the need for improved defense makes the malware environment highly unstable and volatile making it cumbersome for effective and efficient defense.

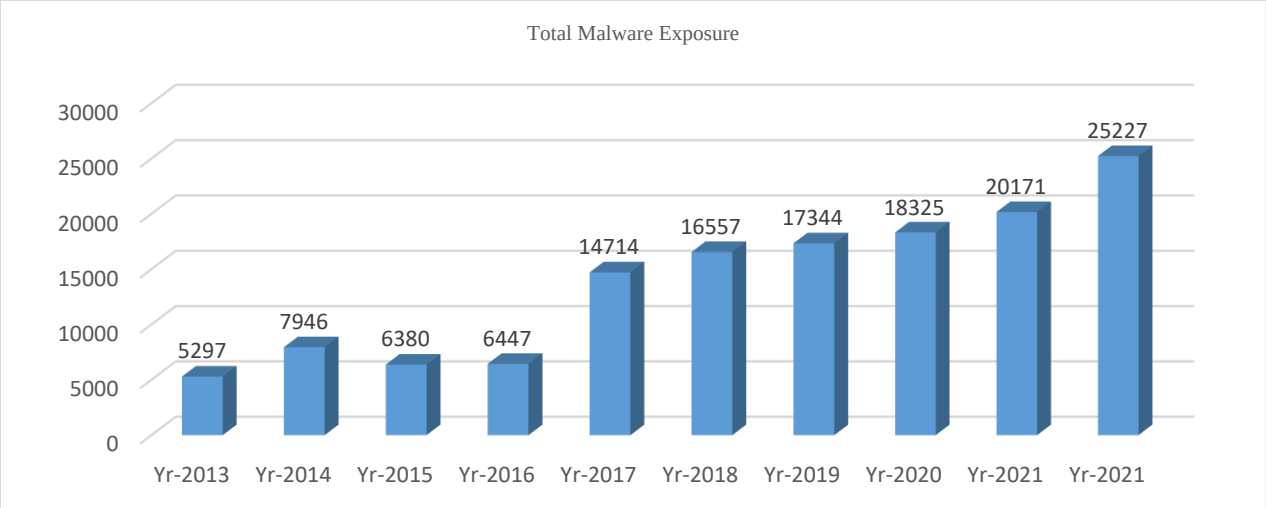

Fig. 2. 10-year annual malware vulnerability exposures.

Thus, the volatility in the ecosystem as a result of the exponential increase in volumes, variety and complexity requires regular evaluation of the ecosystem to establish the trends, variety, volumes and how these malwares are spread to their victims. Therefore, obtaining the current state of malware trends such as volumes, variety (Characteristics of new malware over traditional) is vital for planning innovative defenses and decision-making. Establishing current trend in malware volumes, variety, complexity and other adversarial activities has the potential to provide the needed and relevant rich and practical insights for industry and the research community for improved decision-making, incident management and other activities of Security Operation Centers (SOCs) and for other Incident Response Teams (IRTs).

B. Malware Obfuscation and Other Evasive Techniques

Generally, malware attackers use many evasive techniques to exploit vulnerabilities in existing defense systems using mainly obfuscation or evasive techniques. Obfuscation is the process by which the identity of malware is concealed with the aim of evading detection [18]. By hiding the details of the program code, encrypting, packing, or the use of other armor techniques, the disguised malware is able to evade detection. Malware obfuscation techniques are many and varied and include fragmentation, where the malware is broken into fragments and only executes after reassembly. Anti-sandbox techniques that detect simulated environments and refuses to execute, Stalling delays where the malware hibernates until an action, event, or time frame, the use of rootkits to attack or exploit operating system

vulnerabilities, and the use of action requirements such as clicking the mouse or tapping any key [11, 18]. Besides, malware authors use innovative obfuscation techniques such as dead code insertion, code transposition, and advanced approaches leading to polymorphic, metamorphic, and oligomorphic malware variants that are difficult to detect with current existing signature-based detection techniques [25]. Consequently, to be able to adopt innovative and novel ML defense techniques against malware obfuscation requires the exposure of the various obfuscation methods to the algorithm during the training or model building phase. Knowing the various obfuscation techniques and these are employed in malware attacks is of essence for all malware defense stakeholders in taking defense decisions and other remediation effort.

C. Historical Review of the Strengths and Weaknesses of Existing Defense Methods

In this section, we discuss the various signature-based methods such as static, dynamic, hybrid defense techniques, ML, and AI techniques, given their limitations and strengths.

1) Signature-based techniques

Over the years, the use of innovative signature-based malware defense techniques and approaches has been explored including static, dynamic, and hybrid approaches. Each of these techniques are shown to be efficient and effective depending on the task under consideration. However, with the growth, expansion and complexity of current malware, there appears to be limitations among these techniques as explained in the proceeding paragraphs.

Static techniques analyses malware binary without executing the malware code [26]. It was among the earliest techniques for malware analysis and usually the first point of call for analyst. It is a faster, cheaper, and easier analysis approach to establish a first-hand impression of malware. However, this technique fails to reveal enough information for effective decision-making and is susceptible to packing, crypting, resource obfuscation, and anti-disassembly techniques [19]. To overcome these weaknesses, the use of dynamic approaches has been explored.

In dynamic analysis and detection, malware binaries are executed in a safe and isolated environment [27]. This ensures that the runtime behavior of the malware is observed as it executes. Dynamic analysis methods are effective and precise in malware detection; their analysis method is relatively simpler (due to following single path analysis), gives runtime behavior of the malware, and can handle simple obfuscation compared with static techniques. However, dynamic analysis provides a limited view since it follows a single path, making the approach time-consuming. It is also limited due to time-dependent malware such as logic bombs and bots that only execute after a timed event; some malwares are also environment-aware, making the malware refuse to execute in a simulated environment; and it is resource-intensive. In addition, sophisticated malware such as polymorphic, metamorphic, and oligomorphic malware and their variants render dynamic analysis suboptimal, resulting in suboptimal exposures [28].

Moreover, the use of heterogeneous hybrid static and dynamic approaches has been explored [15, 29]. The usual heterogeneous hybrid is resource-hungry, such as in terms of processor time, high memory consumption, and time consumption. It is therefore apparent that heterogeneous hybridization involving static and dynamic techniques improves performances. However, the approach has limitations, including high execution time, high resource requirements such as processor and memory, poor analysis methods, being susceptible to high obfuscation techniques such as polymorphic and metamorphic malware, poor and redundant features, and a lack of adaptability with obfuscated malware, which remains a challenge for heterogeneous hybrids [15, 30, 31].

Consequently, due to the limitations of the signature-based techniques, the use of heuristic-based approaches has been explored. In heuristic methods, detection of malware is based on rules and patterns determined by experts. This with time and the growth and complexity of current malware renders the method error-prone and time consuming [11]. Thus, prompting the use of cloud-based techniques in which a server is protected using blacklisting to block malware, white listing for good ware whiles gray listing is used for indeterminate malware samples. the dexterity of current malware makes this approach inefficient and ineffective due to increasing numbers in the gray-list. Therefore, with these limitations coupled with the growing malware volumes, variety and complexity, there is the need for the use of automated techniques that demonstrates adaptability, learnability and generalizability and machine learning and Artificial Intelligence techniques are of essence [18].

2) The use of machine learning techniques

The use of machine learning approaches in malware detection has been widely explored with relative success [32–34]. The growth and exponential increase in malware coupled with the creation of new, novel malware makes defense highly challenging without automation. Thus, in the face of the limitations of the signature-based methods, the use of ML techniques has been explored as a means to automate the initial phases of malware attacks by detecting and classifying malware from benign ware thereby assisting in the triaging of threats. To achieve this requires the use of various ML algorithms including supervised, semi-supervised, unsupervised and reinforcement learning techniques. These techniques are used in clustering, classification and predictive analysis. These algorithms are mainly used for network protection, end-point protection, application security, suspect user behavior detection and many others [33]. Examples of ML algorithms include perceptron, Support Vector Machine (SVM), Neural Networks (NN), Random Forest (RF), Decision Trees (DT), Naïve Bayes, Logistic Regression (LR). Others include ensemble techniques such as Gradient Boost (GB), eXtreme Gradient Boost (XGB) and deep learning approaches such as Convolutional Neural Network (CNN). With the growth of AI and ML techniques, and owing to the limitations of the various classifiers, newer methods such as Generative Adversarial Networks (GANs), Long Short-Term in Memory (LSTM)

Networks, and others have been explored in malware detection.

Notwithstanding the relative success of the approach, the innovation and improvement in malware attack and defense requires innovative techniques that ensures reliability and efficiency. Knowing the relative weaknesses and the strengths of the various defense techniques provides an opportunity for methodological and design changes for improvement. It also led to effective, efficient and informed decision-making when choosing the right tool for remediation and defense. In addition, the need for non-technical security to augment these technical controls is of essence and leads to the provision of holistic security. Thus, apart from exploring the technical controls, the exploration of user and organizational variables that leads to malware exposures in addition to the technical, ones assist defense teams to apply holistic techniques to ensure optimum security against malware attacks.

D. User and Organizational Vulnerabilities to Malware Attacks

Generally, both industry and the research community have over-emphasize technical controls in malware defense at the neglect of the non-technical such as the user and organizational vulnerability. Contrary to this position, users have remained the weakest link in the cybersecurity value-chain [35, 36]. Thus, to ensure effective and efficient defense against malware, requires not only technical controls but with the augmentation of security conscious users and effective organizational policies and structures, that ensures optimum security. Moses and Sarah [22] suggested that, user vulnerability constitute a major risk to malware and other socially engineered attacks. Thus, defense stakeholders should perform user and organizational vulnerabilities regularly to identify the limitations in users and to provide innovative security training and awareness programs. Similarly, the organizational variables that malware authors exploit such as unpatched vulnerabilities, insider threats among others should be identified and remedial processes used to overcome them. According to Kerperski report 2022, there were 15.45% of users recorded malware-based attacks. The company blocked 687,861,449 attacks across the globe. During the period, 429,354 users were attacked for financial gain or to steal money. Fig. 3 shows the distribution of malware attacks according to the target medium or application. The data suggest that, malware exploits several media and applications to compromise systems leading to the loss of confidentiality, integrity and availability protocols.

Similarly, the Data Breach Investigation Report (DBIR), for the 2022 year also indicated that, there were four key drivers in vulnerabilities in the year and these included ransomware, users, errors/misconfigurations. As shown in Fig. 4, user vulnerability leads the pack with 82%. This makes user vulnerability a critical component in the fight against malware attacks as it exploits not only users, but also other organizational factors leading to compromises.

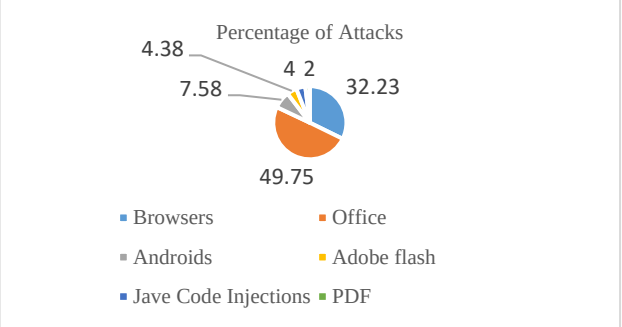

Fig. 3. Various malware media.

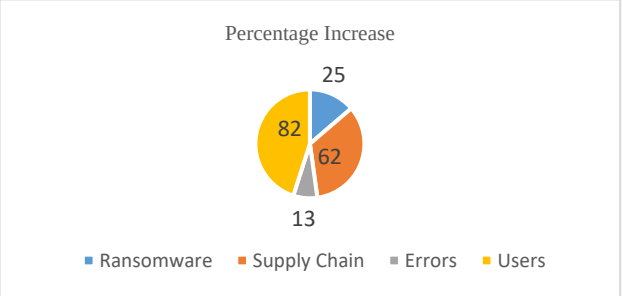

Fig. 4. Key vulnerabilities.

The increase vulnerability of users and organizations to malware attacks poses a challenge for effective and efficient malware defense. These staggering figures requires stakeholders to employ novel techniques to counter the onslaught. This effort can be achieved by providing novel and a holistic solution based on both technical and non-technical perspectives. Such integrated approach to the study of malware requires further research as it can lead to the improvement in the overall security architecture and improved holistic defense as required by current research.

E. Current Malware Research Approaches

The growing dexterity of new malware renders the existing signature-based methods inefficient and ineffective. Parisi [37] suggested that the application of AI and ML techniques in cyber defense is a new experimental research area with relative success but not without challenges. Among current approaches used include deep learning techniques, ensembles, Long Short-Term in Memory (LSTM) Networks used in sequence prediction. Bidirectional Long Short Terms in Memory (Bi-LSTM) used for sequence prediction in which one is used for forward processing and one for backward processing. Other approaches include the use of Generative Adversarial Networks (GANs) [37]. The results so far achieved hold promise, and future endeavors will adopt the use of these techniques for automated detection. Thus, the use of various ML techniques such as supervised learning, unsupervised learning, reinforcement learning, and others is employed in malware detection and cybersecurity in general [11, 37, 38]. However, notwithstanding the relative success of these techniques, they have some inherent challenges that impede the achievement of their full potential. Thus, there is a need for the use of innovative techniques that employ approaches that improve detection

performances. In addition to the challenges of ML methods, user and organizational vulnerabilities also significantly contribute to malware attack success [23]. This requires that not only do malware defenders require novel and efficient detection techniques but also cybersecurity-conscious users and organizational practices that inhibit the success of malware attacks. Therefore, in order to ensure effective and holistic malware defense, the new approaches should not only be on the vulnerabilities or limitations of the technical controls, but also on the other user and organizational factors, that promotes malware attack success. To understand these factors, requires review of both extant and industry to establish the current prevailing state and the identification of the gaps using empirical methods.

IV. Materials and Methods

In this section, we present the research methodology of the study. We describe how the needed and relevant papers were identified and selected for the study, the inclusion and exclusion criteria, the synthesis of various information based on the research questions and how the case studies were conducted to augment the outcome of the literature review.

A. Methodology

A scoping review methodology is the process of mapping literature, which gives a researcher the opportunity to survey and/or examine an area based on a research question of interest [39]. The interest of the researcher could be to explore the extent of the literature on the topic, identify the boundaries and parameters, or identify research gaps for filling in. Thus, we sought to identify gaps and explore the extent of the literature that could enable us to propose novel solutions. We used this technique because our research question is broad and does not lend itself to the usual systematic review as suggested by Tricco *et al*. [40] and Peters *et al.* [41]. The scoping review framework comprises the following as shown in Fig. 5.

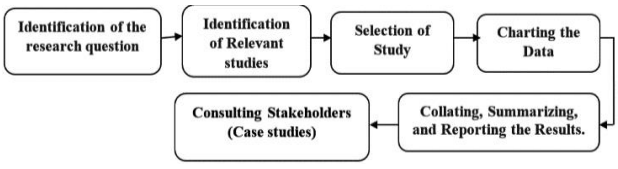

Fig. 5. Methodological framework of the study.

1) Identification of research objectives

The framework begins with the research question or objective. This is the objectives guiding the study. To this end, the following research objectives were set to guide this study:

(1) Present up-to-date or current malware types/variety, modes of transmission or spread, and comparative characteristics of current malware relative to traditional malware.
(2) Establish the various obfuscation techniques used by malware authors to evade detection by signature-based defenses.
(3) Evaluate the relative strengths and weaknesses of signature-based malware detection techniques.
(4) Evaluate the current research techniques that adopts ML techniques and to establish research gaps.
(5) Explore non-technical factors that leads to malware exploitations at user and organizational levels
(6) Propose future projections of malware trends based on the trends and industry requirements

2) Identification of relevant studies

We conducted a search in December 2022 in electronic databases. IEEE, MDPI, IJACSA, SCOPUS, and other internet sources. We downloaded peer-reviewed journal papers written in English that discuss issues of malware detection, obfuscation, and limitations, following our research question. The key search terms were malware detection, limitations of malware detection techniques, detection methods. Malware defense OR malware detection, malware defense AND detection, malware detection limitations OR vulnerabilities, static analysis, dynamic analysis, and hybrid analysis Limitations of static analysis OR dynamic analysis; limitations of static analysis AND dynamic analysis; limitations of hybrid analysis OR dynamic analysis; malware obfuscation techniques OR evasive techniques. "Current gaps in malware detection", "challenges facing automated ML techniques in cybersecurity" or "challenges of current cybersecurity defense techniques". Following these search keys, we downloaded a number of articles relating to the topic of the study.

3) Study selection criteria

The papers included in the study discussed issues relating to malware trends, types, and varieties. Malware obfuscation techniques, malware detection methods, limitations, and ML techniques in malware defense. The papers were mainly experimental in nature and described malware detection techniques, citing their limitations and proposing an innovation. All articles included in this study needed to discuss these issues relative to malware protection using rigorous scientific and experimental approaches. Studies that failed to mention the detection method, the limitations, and the somewhat evasive techniques that it sought to solve were not considered. Reviews of detection methods, limitations of malware detection methods, and obfuscation techniques were included as they directly contain relevant information pertaining to the topic under consideration. Thus, if a paper discussed malware detection, limitations, or obfuscation, that paper was considered since it contained information that is relevant to our research question.

4) Charting the data

We reviewed articles by titles and abstracts to identify duplicates by labeling them "included" or "excluded" and "potential for inclusion". The third reviewer, who cited reasons for inclusion and exclusion where the entire team must agree, resolved the variations in the ratings, or, by default, a vote is done. We created a template for the extraction of the relevant data for the included papers. We reviewed the selected papers, and the relevant information

was extracted based on the research objective of the study. Based on the analysis of the papers, we extracted information regarding the variables of interest to our research objectives. These included malware types or varieties, their transmission mode, comparative characteristics of the current malware verses traditional malware that impedes efficient and effective malware defense. Other relevant information extracted included the various obfuscation techniques and their categorization and how, malware authors evade detection by using these techniques. Furthermore, the comparative weaknesses and strengths of the various malware detection or analysis techniques such as static, dynamic, hybrid, and ML techniques were also extracted and presented. Finally, the various ML algorithms, used in malware detection, the types of features used in the training of algorithms, the performance evaluation matrices used in the evaluation of malware detectors, and the research gaps in each of the ML-based papers were extracted and presented. The aggregated information is presented in the results and discussion section in Section V.

5) Collating, summarizing, and reporting the results

In this section, we collate the results, make summaries, and give a narrative report of the study. This forms a coherent whole and provides meaning and context to this study. Analysis results were aggregated into themes and sub-themes according to the research questions. We present a brief overview of malware and trends, the variety and comparative characteristics of traditional malware verses the new and evolving malware. The various obfuscation techniques employed in the industry to conceal information leading to exploitation. Present the relative strengths and limitations of static, dynamic, hybrid, and ML techniques. Finally, a presentation of the identified research gaps in current research works. The second part of the results comprises the case studies as described in the next subsection.

B. Description of the Case Studies

Literature supports the view that scoping review outcomes can be improved and the result made more useful by involving practitioners and other stakeholders in the studies [41, 42]. Consequently, in light of this, we added this stage to the study by conducting empirical case studies to augment the outcome of the literature review. As a study involving humans, we obtained ethical clearance for the study, which is a requirement in such cases [43]. 450 participants who were regular users of internet resources were used for the study. We collected data on malware trends, variety, and volumes and analyzed the collected data using summary and inferential statistics. This section describes the case studies used to augment the outcome of the literature review. Transaction log analysis to establish current malware trends, variety, and complexity. User vulnerability studies to ascertain uses vulnerability to malware attacks and organizational factors exploited to perpetuate malware attacks.

1) Transaction log analysis for malware trends

The main purpose of this part of the study was to empirically identify current malware trends, types, variety, and prevalence from an industrial perspective using real environments and real systems [6]. To achieve this, it required the use and analysis of malware attacks from real systems and real environments as a means of establishing it using the real threats from a threat log system. The evaluation of the content of a log activity for identifying patterns is known as call log or transaction log analysis [6]. We extracted log content from a financial institution's threat log event systems for a period of 3 years (2020–2022). The logs were analyzed to identify malware and recorded. The type and number of malware were counted and recorded in a week, which was aggregated into months and into years. The type of malware and the trend for the three-year period are shown in the results section of the paper.

2) Approach to organizational vulnerability to malware attacks

The purpose of this part of the study was to identify the key organizational vulnerabilities inherent in organizational settings that malware could exploit to compromise systems, leading to the exposure of valuable assets. According to Ref. [44], malware exploits varied industries and sectors such as banking, manufacturing and others due to the value of information in such system. Every organization has High-Value Information Assets (HVIA) that they are required to protect against unauthorized access, disclosure, or modification to ensure confidentiality, integrity, and availability for authorized users [10]. However, many organizations have some fundamental flaws that are always exploited by cyber actors to perpetuate crimes, and after gaining access to the system, they now employ novel Tactics, Techniques, and Procedures (TTPs) to have lateral movement in the organization's system [10]. He opined that organizations should assess key and critical assets to identify their vulnerabilities and devise strategies to mitigate them by conducting vulnerability monitoring and management. To evaluate malware vulnerabilities in the organization, we adopted a retrospective case study using content analysis of the organization's records from 2020–2022. We reviewed the internal and external audit reports, cyber and information security reports, vulnerability assessment reports, and key records of the financial institution. The key findings in these documents were extracted and aggregated into themes. The themes were analyzed for cross-referencing and duplication of findings. We then compared the themes and synthesized them into five key headings. Based on our domain experience, we suggested defense strategies for each of the identified themes.

3) Approach to user vulnerability to malware attack study

User vulnerability is the inherent weakness in humans that makes them susceptible to deception and exploitation [6]. To evaluate the level of vulnerability among users, we proposed a Naturalistic Habitat-Based Field Experiment (NHBFE) in which a researcher is in a geographically distant location while interacting with study participants virtually. We purposively sampled 450 participants from the financial institution in Ghana for the study. Two malware attacks using spear phishing and waterhole approaches were launched on the users after

ethical clearance was obtained from the Chief Information Security Officer (CISO). The techniques were deployed at different times on all the users. The first one was a simulated phishing email about the appointments of heads of department. The emails were designed to have the look and feel of the company's communication structure after studying this with the CISO. Users who tried to open the mail were asked to upgrade their application before they could access the file. The vulnerable users clicked on the upgrade link, while others refused to upgrade even though they saw the message. Based on the action of the user, we could tell the level of vulnerability to malware attacks. Users who clicked the link to upgrade were considered vulnerable, while those who abstained were considered security-conscious users. Similarly, a water-hold type of attack was simulated. A water hole is a type of SE attack in which legitimate websites or systems of an institution are compromised by an attacker. The attacker, who now has access to the website or the system, laid ambush to gather confidential information by redirecting user traffic to an illegitimate site. We constructed a message titled "New Staff Schedules" as an attachment following the internal communication system of the financial institution and disseminated it to all participants in the study. When a user receives the mail and clicks to open it, they are asked to install a new version of Microsoft Excel since theirs is outdated. Users who proceeded to install were considered vulnerable users, while their counterparts who abstained at this level were considered security-conscious users. These results were extracted, analyzed, and summarized using summary statistics. We present the results of the scoping review and the case study results in the next section.

V. Results and Discussion

The aim of this paper was to provide an evaluation or a survey on current malware trends, evaluate the existing attack and defense strategies to establish their relative strengths and limitations in the face of current malware, and to identify research gaps in the current research. This section presents the results and discussion of the study using charts, figures, tables according the research objectives of the study.

A. Brief Overview of Malware, Trends and Variety from Literature and Case Study

This section presents definition and variety of malware, the characteristic of new malware compared with traditional malware that makes it challenging for existing defense techniques and the mode of transmission of malware across various networks and platforms.

1) Malware trends and variety

Malware refers to programs or codes with malicious intent that usually come in the form of an executable file, code, script, or any other unwanted software [37]. Cyber adversaries employ these techniques to steal sensitive information, take control of systems, disrupt systems, gain unauthorized access, spy on victims, lock up files, or take control of systems, including critical infrastructure [5]. There are diversity of malware and other and socially engineered malware techniques that are employed by malware authors to perpetuated an attack. These encompasses different types of malicious programs including but not limited to viruses, worms, rootkits, Trojans, adware, spyware, ransomware, downloaders, droppers, deep fakes and many others. Generally, these waves of attacks are known broadly as social engineering attacks, which comprises semantic (non-technical) and syntactic (Technical) methods. In semantic attacks also known as HUMINT (Human Intelligence), the attack targets the people or users in order to exploit their vulnerabilities to gain confidential information to compromise systems. On the other hand, the syntactic methods exploits vulnerabilities in technology systems using tech-based tools and techniques such as malware [9, 35]. Table I below shows the diversity of malware attack techniques or vectors. This diversity presents a challenge to existing malware defense architectures since the approaches varies. This position is in tandem with that of [8], which is one of the largest anti-virus companies in the world. According to Ref. [8], about 450,000 pieces of malware and Potentially Unwanted Software (PUS) are recorded daily. These reports indicated that the malware is not only increasing in volume but also in variety and complexity. An analysis of the literature suggests a global increasing trend of malware and other socially engineered attacks [6, 7, 22, 24]. This growth and expansion of malware attacks remain a major problem for all cyber defense stakeholders [7, 8, 45]. The high variety and complexity of these malware imply that defending against such attacks using traditional signature-based techniques becomes problematic [6, 10, 22].

TABLE I. Diversity of Malware and Other Socially Engineered Attack Techniques

S/N	Syntactic/Technical	Semantic/Non-technical
1	Phishing	Tailgating
2	Drive-by-download	Dumpster diving
3	Spyware	Eaves dropping
4	Adware	Pretexting
5	Ransomware	Information theft
6	Pharming	Reverse social engineering
7	Vishing	Shoulder surfing
8	Pop-ups	Online social Engineering
9	Trojans	Quid-Pro-Quo
10	Embedded links	Help-desk attacks
11	Worms	Piggy-backing
12	Viruses	Robocalls
13	Code injectors	
14	Rootkits	
15	Scareware	
16	Key loggers	
17	Browser hijackers	
18	Water-holing	
19	Logic bombs	
20	Bugs	
21	Botnets	
22	crime ware	
23	Backdoors etc.	

2) Malware transmission modes

When new malware is created, the authors adopt a number of means or strategies to spread it and infect other targeted computers. The created malware can travel or reach other systems through vulnerabilities in network

services and architectures, by downloading from the internet sources, exploiting vulnerabilities in web browsers, or luring users to take malicious actions that can eventually harm their computers [46].

3) Comparison of current and traditional malware

To appreciate why it is becoming continually difficult to defend against current malware attacks, there is the need for comparison of the new malwares against the traditional malware. This is because existing malware defenses such as static, dynamic, and hybrid were designed with the traditional malware in mind. Hence, with the changing nature of the malware, the existing defense show some limitations. An analysis of the various malware samples from the study shows that, the current malware exhibits some features that makes it difficult for detection. With traditional malware attacks, it was easy to identify and detect since the malware maintained its form and shape throughout its lifespan. However, the complexity of the current malware shows the contrary, as new malware have features that makes it change overtime during its lifespan. A review of the papers revealed that current malware have adopted novel approaches to spread from one network to another, which presents a challenge to efficient detection and classification. In addition, the new malware also have new characteristics compared with the traditional malware that makes it challenging for efficient and effective detection and classification. These modes of spread or modes or transmission are as presented in Table II, which includes repackaging, backdoors, vulnerabilities, privilege escalation among others.

TABLE II. CURRENT MODES OF MALWARE TRANSMISSION

S/N	Spreading Technique	Description of Mode of Spread
1	Repackaging	Disassembling good ware, appending the malicious content and reassembling it using reverse engineering.
2	Vulnerabilities	Security defect in the defense architecture that enables illegal access
3	Backdoors	An accidental or intentional opening in software, hardware, network, or any part of the security architecture.
4	Privilege escalation	When an attacker gets escalated access to a computer or network and use same to launch an attack.
5	Blended Threats	This combines characteristics from multiple types of Malware making the detection very difficult because they exploit variety of vulnerabilities
6	Homogeneity	When a system used the same operating system, networks etc., it makes it easier for malware such as worms to spread across the network.
7	Dynamic Payload	Download an encrypted source file and after installation, the application decrypts the encrypted malicious payload and executes same
8	Drive By Download	Users visits a website containing malicious content and unknowingly download d malicious files or content into their computing devices.
9	Stealth Malware	Malware that exploits hardware vulnerability

Similarly, some of the variations or characteristics of the old malware compared with the new malware are as depicted in Table III, which shows that, current malware is targeted, persistent, are stealth in nature and is unknown or variants of known malware, including zero-day vulnerability. Whiles the traditional malware are broad, one-time, open, and whose signature is known already. This description is in harmony with [11], who proposed similar characteristics.

TABLE III. CURRENT MALWARE VS TRADITIONAL

S/N	Traditional Malware	Current Malware
1	Broad in Nature	Mainly Targeted
2	Already Known	Unknown/zero-day
3	One-time malware	Persistent in nature
4	Open	Stealth in nature

The current malware characteristics implies that, existing signature-based detection systems are unable to adequately detect malware, leading to poor accuracy and high false positives. Thus, not only that, malware assailants adopt other innovative obfuscation methods with which they evade detection including exploiting user/organizational vulnerabilities.

4) Malware trends and variety from case study perspective

To gain industrial insights about the current spate of malware attacks and the various forms of attacks, we used threat transaction log analysis to identify the variety of malware threats that targeted users in the organization. As depicted in Fig. 6, there are varied malware techniques deployed, with phishing, embedded links, and Trojans being the most prominent attacks in the studied organization. The variety and volumes mean that organizations must adopt new and novel techniques that can withstand such huge numbers. This is consistent with [6, 22]. The increased variety and complexity, requires that the existing defense techniques should have commensurate adaptability to withstand such attacks.

A detailed malware samples extracted from the threat log system is shown in Fig. 6. Similarly, the three-year trend analysis for the data gathered from the threat logs and aggregated into months is as shown in Fig. 7 with 2020 recoding the highest. The sharp increase of the malware attacks in the studied organization was possibly due to the COVID-19 that led to the lockdown of all nations forcing more people to go online. Thus, cybercriminals decided to cash in by launching several attacks. Notwithstanding the drop in numbers from the data, the global picture remains peaked with many malwares such as ransomware, deep fakes, phishing and others being employed to exploit system-making organizations to spend huge sums of their budgets in defense efforts [8, 45].

The results from this confirms the position of the literature that there is increasing malware volumes, variety and complexity. Different malware exploits vulnerabilities to spread across networks, and making use of obfuscation techniques to compromise systems [8, 27]. Thus, using these results supports and improves the position of the literature on the topic.

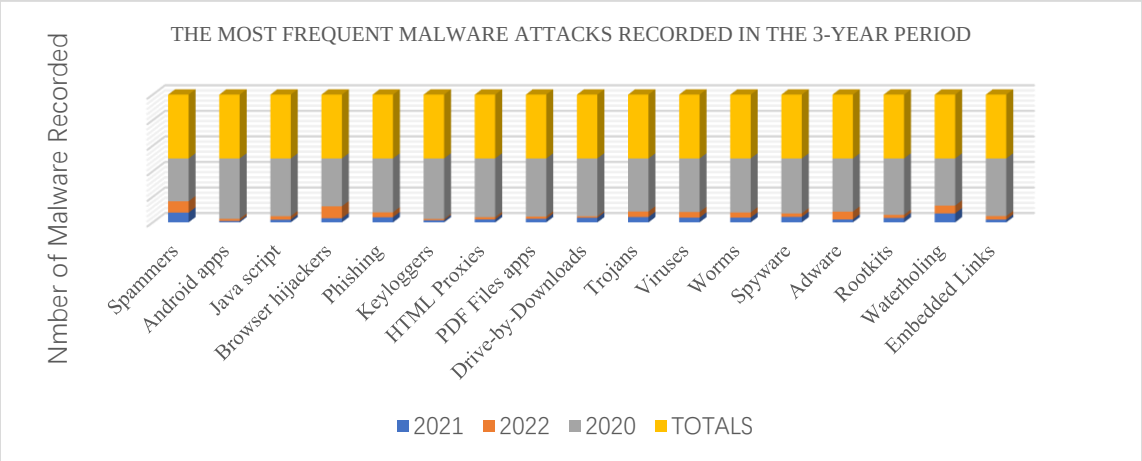

Fig. 6. Types of Malware recorded from 2020–2022.

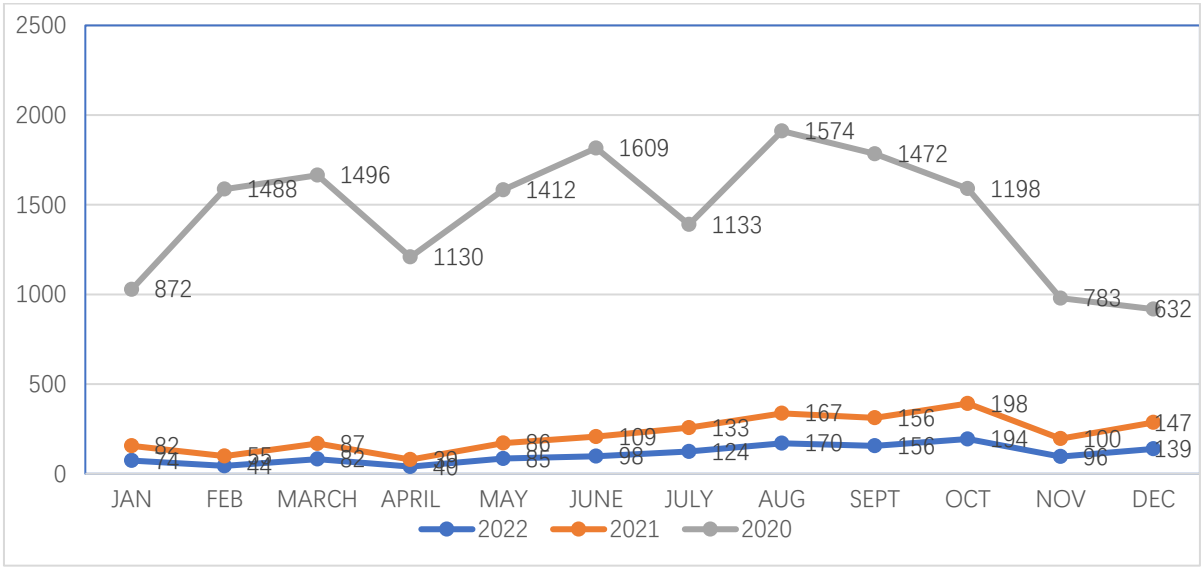

Fig. 7. Comparative monthly attacks for the three-years.

B. *Malware Obfuscation and Evasive Techniques*

In order to evade detection, malware authors adopt subtle ways of concealing the identity of the malware. To be able to defend against malware obfuscation requires a critical analysis of the various forms of obfuscation. Thus, we explored this by aggregating the various malware evasion techniques from the literature and categorized them into anti-static and advanced methods. As shown in the Fig. 8, the basic obfuscation methods used mostly to prevent anti-disassembly, while the highly sophisticated and/or advanced methods are employed to change the malware byte sequence, resulting in variants, new and novel malware that evade the normal, traditional, or conventional anti-virus and other scanners. This classification is consistent with the one suggested by Heena and Mehtre [11], Aboaoja *et al.* [18], Azaabi *et al.* [25].

To appreciate the various forms of the methods, we present the techniques and their description in Tables IV and V. The two Tables gives a brief description of each of the technique used during malware evasion as proposed by Azaabi *et al.* [25]. Thus, knowing the obfuscation methods and how the malware attackers execute their attacks places the analyst above the attacker.

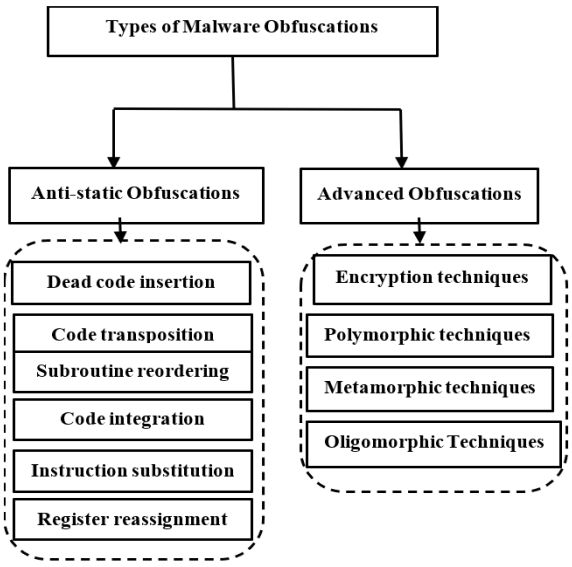

Fig. 8. Static and advanced malware obfuscation techniques.

TABLE IV. ANTI-STATIC TECHNIQUES

Obfuscation Technique	Description of Technique
Dead Code Insertion	Adds ineffective or meaningless codes into a program but does not change the true behaviour of the program, e.g., inserting NOP.
Register Reassignment	Switches registers from one generation to another while the program and behavior remains the same.
Sub-routine Reordering	Obfuscate by changing the order of the routines in random, which can generate N! Variants of malware.
Instruction Substitution	Replace the original code with equivalent ones making the original appear different.
Code Transposition	Replace the original code with equivalent ones making the original appear different.
Code integration	Code integration, in which the malware integrates itself in the target program and produces a new version of the target program.

TABLE V. ADVANCED OBFUSCATION TECHNIQUES

Advance Obfuscation Technique	Description
Encryption methods	Converts normal text to cipher text
Oligomorphism	Decryption key changes with each file infection
Polymorphism	Changes the behaviour with each copy with unlimited generated keys
Metamorphism	Content of the malware changes using mutating generation key

Table V presents some of the advanced obfuscation techniques used by malware authors to evade detection leading to compromises of confidentiality, integrity and availability principles of the information security.

In addition to the above classification, other armored techniques malware authors use to evade detection, according to Naseer [47], include compression of files, anti-patching techniques, anti-tracing techniques, anti-unpacking, anti-VMware, restrictive dates, and password-protected features. Thus, knowing the existing obfuscation or evasive techniques ensures that any attempt at malware defense will factor these into the design and implementation of innovative techniques that would be resilient against such obfuscation. By obfuscating or employing these evasive techniques, the malware's binary sequence is changed, making it difficult for signature-based detectors to detect the malware sample. This change renders signature-based techniques inefficient and ineffective [25]. Thus, the ability of the new variants of malware to mutate, create variants from the same malware, the use of high-level encryption techniques, anti-unpacking, anti-tracing and other evasion techniques makes the new malware challenging to defend with the existing signature-based methods. Therefore, the variety of the evasive techniques makes it challenging for existing defense methods to efficiently detect malware. Whiles anti-static techniques are largely easy to detect, the advanced form are highly difficult to detect leading to high false positives rates and the consequence of exposures such as financial loss, reputational loss and risks of non-compliance to regulatory and legal requirements. Thus, by exploring the various obfuscation methods and how these techniques are employed results in better understanding of the modus operandi and offer malware defenders the opportunity to proposed novel solutions against such evasive techniques.

C. Classification of Broad Malware Detection Techniques and Strengths and Limitations of Existing Detection Methods

In this subsection, we present the broad classification of the various malware detection methods, the weaknesses of some of the existing malware defense methods (static, dynamic, hybrid, and ML) techniques and the current research focus.

1) Broad malware detection schemes

Due to the impact of malware exposure on all stakeholders, both the industry and academic community have adopted and employed a number of techniques [46]. These techniques are meant to prevent, mitigate, or detect malware. Malware detection is classified mainly into signature-based (using signatures or pattern matching of extracted features), behavior-based (observing what the program does during execution), and heuristic-based (the use of malware features to train algorithms), including deep-learning methods where neural networks and other Artificial Neural Networks (ANN) are used [11]. Other classifications include anomaly-based detection (using abnormal actions to identify malware) and statistics-based detection methods. These broad categorizations and their strengths and limitations are shown in Table VI.

TABLE VI. BROAD MALWARE DETECTION METHODS

Detection Technique	Strengths	Limitations
Signature-Based	Efficient in detecting known malware. They are very fast.	Susceptible to basic obfuscation techniques.
Heuristic-Based	Adaptability and resistance to malware obfuscation	Poor features, synthetic datasets used gives false performance
Behaviour-based	Show runtime behaviour of malware samples	Time and environment dependent
Statistical based	Good at probabilistic detection	High false positives

These techniques, the mostly used by the anti-virus community is the signature-based methods [48]. Notwithstanding the universal adoption of this technique, the current malware poses a challenge for such defense techniques. Hence, the need for the evaluation the defense techniques to identify their strengths and limitations and the adoption of novel defense techniques to be able to withstand these attacks.

2) Limitations and strengths of static, dynamic, hybrid and ML detection techniques

Among the defense techniques, the main technique used by the anti-virus community is the signature-based methods. In this method, a known malware is obtained. The features or signature of the malware is extracted and kept in a repository. When new malware arrives, the signature of the new malware is matched with the repository. If it matches, then it is malware; otherwise, it is not. This method has high efficiency and lower False Positive (FP) rates. However, this technique assumes that malware, once identified, remains the same throughout its

lifespan, which is not the case [18]. Thus, with new and revolving malware with mutating capabilities, these methods show some limitations. Examples of these techniques identified in the literature are static, dynamic, hybrid, and recently the use of ML algorithms. In static malware analysis, the malware binary is analyzed without the actual execution of the malware code. Static features such as IP addresses and others are observed. In dynamic analysis, the actual code of the malware is executed and the behavior of the malware activities is observed and used. Due to the limitations in dynamic and static analysis, the use of hybrid techniques has been proposed and used extensively [5]. This is where the static and dynamic techniques are used concurrently in the analysis of malware. Finally, with the growth and expansion of ML techniques and their ubiquitous use in all areas, malware detection has found solace in its use of ML techniques and has been widely explored including the use of hybrid features to improve classification performance [49]. To identify the limitations and strengths in the various signature-based malware defense techniques, we explored the literature and aggregated or synthesize these into themes and presented as shown in Table VII. From Table VII, it is apparent that, malware defense techniques have changed over the period; from static, dynamic and hybrid techniques to the use of ML methods. The change from one defense method to another was occasioned by new and innovative attack methods adopted by malware attackers to exploit known vulnerabilities. From the results, it is apparent that malware attacks exploit several vulnerabilities in the defense value chain and techniques. Thus, the prevailing limitations in these malware defense techniques require the use of novel and innovative techniques for effective and efficient remediation. Even though static and dynamic analysis are known to be inefficient and ineffective in the prevailing malware ecosystem, however, a careful disassembly of malware using these techniques can reveal relevant static and dynamic features for building ML models that can improve detection and classification of previously unknown malware [50]. Since, the success of ML techniques largely depends on the relevant features, extracting relevant features to train and test ML algorithms can lead to improvement [51]. Therefore, empirically identifying limitations in each of the methods presents an opportunity for improvement. In summary, from static, dynamic, hybrid and now the use of ML is geared towards getting efficient and robust methods.

There is no denying the fact that, it is the time of AI and ML techniques and malware requires the use of these techniques [5, 37].

However, technology alone cannot ensure adequate security and humans and other organizational factors are contingent on the success of malware attacks [6]. The next sections present the case study of user and organizational vulnerabilities to malware attacks. The purpose was to ascertain whether user vulnerabilities to malware attacks contributes to malware exposures, and finally, explore organizational factors that promotes the success of malware attacks to suggest solutions.

TABLE VII. COMPARATIVE WEAKNESSES AND STRENGTHS OF MALWARE DEFENSE TECHNIQUES

Analysis Technique	Strengths	Limitations	References
Static analysis	Computationally cheaper, Broader view of the binary during analysis, Stable and repeatable analysis, Security and independence of the data (once created, signature is stored for future analysis	Limited reverse engineering tools, susceptible to packing, rypting and other obfuscations, Cannot analyze runtime behaviour, Conservative approximation (little information revealed).	[9, 18, 26, 45]
Dynamic analysis	Precise and effective in detection Simplicity of analysis (Single path) Provides runtime behavior Efficient and some obfuscation	Provide only limited view of analysis (single path), Resources intensive, Time consuming task if the dataset is large, Some of the malware are environment aware	[5, 29, 50]
Hybrid analysis	Improved detection over individual static and dynamic Weaknesses from both are compensated	Resource intensive process, Presence of poor and redundant features, Relatively slower due to the combinations, Lack of adaptability	[5, 9, 31, 50]
Machine learning	Adaptability Generalizability Memorability Learnability	Poor detection rates sometimes, High false positives rates, Curse of dimensionality, Model weakness regarding some dataset	[5, 29, 35, 47, 50]

D. User and Organizational Vulnerabilities to Malware Attacks

This section presents the results and discussion of users and organizational vulnerabilities to malware attacks.

1) User vulnerability to malware attacks using spear phishing and water-holding technique

To explore the level of user vulnerability to malware attacks, we conducted and user vulnerability assessment using a simulated phishing and a water-hole attack approach. As depicted in Fig. 9, contract staff, other staff and management staff recoded the highest levels of vulnerabilities to the two attacks. Except the legal department that recorded zero vulnerability, all the others showed some form of vulnerability. Hence, with such levels of vulnerability, there is likely going to be exposures since some of the malware require user action or inaction to execute the attack. Even though the vulnerability level is lower compared to the industry benchmark of 35%, the average of 8.11% and 8.69% is significant to cause an exposure to the organization's HVAS. This consequently requires the organization and by extension all other security stakeholders to pay keen attention on users.

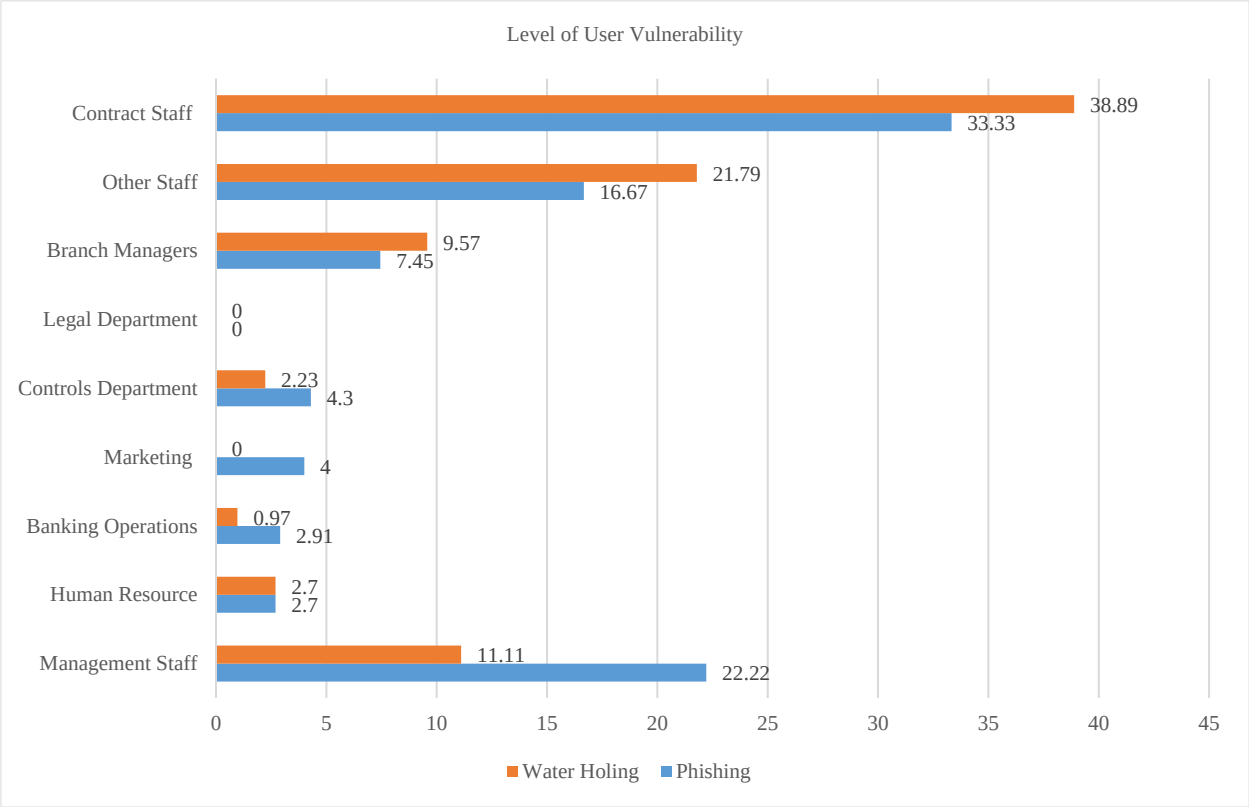

Fig. 9. User vulnerability to phishing and water holing techniques.

Similarly, a t-test statistic at 1%, 5%, and 10% showed that, there was statistical significance among management staff, banking operations, marketing, other staff and contract staff as shown in Table VII. Statistical Significance using t-test was presented as shown in Table VIII. The results show that users are largely vulnerable to some extend with some of the users showing very high vulnerabilities.

TABLE VIII. VULNERABILITY OF STAFF TO MALWARE ATTACKS

Staff Category	Sample size	Spear Phishing	Water-Holing	t-test
MGT staff	9	22.20%	11.11%	4.412**
H R	37	2.70%	2.7%	1.562
Banking operations	103	2.91%	0.97%	1.98**
Marketing	25	4%	0.00%	2.441**
Internal Controls	93	4.3%	3.23%	1.231
Legal Department	11	0.0%	0.00%	0.000
Branch Managers	94	7.45%	9.57%	1.544
Contract staff	18	16.67%	38.89%	1.789*
Other staff	60	33.33%	21.79%	2.789***

Note: ***, **, * represents 1%, 5%, and 10% significance level.

The implication of these results is that while users remain vulnerable to malware attacks, organizations should not only focus on building and purchasing technology products to defend themselves, but a considerable number of resources should be devoted to the non-technical part of malware attacks. This is supported by Aldawood and Skinner [6], Moses and Sarah [22]; who were of the view that user vulnerability, if not checked, leads to exposure and therefore proposed regular and routine user awareness programs. In addition, Paula *et al*. [23] suggested that organizations should adopt regular, innovative, and hybrid delivery methods with relevant content for ensuring Social Engineering Awareness, Training, and Education (SEATE) among users to build resilience and immunity to malware or social engineering attacks.

2) Organizational vulnerabilities

Whether organizational set-up contributes to malware attack success was explored using content analysis of documents from the studied organizations. This was conducted through the analysis of cyber and information security reports, internal audit reports, and IT systems audit reports for a period of 3 years (2020–2022). As depicted in Table IX, five key organizational vulnerability areas were identified as the key suspects leading to organizational vulnerability.

TABLE IX. IDENTIFIED ORGANIZATIONAL VULNERABILITIES AND RECOMMENDED DEFENSE STRATEGIES

Organizational Vulnerability	Recommended Strategies
Security Device Misconfiguration	The use of vulnerability management tools and processes. Avoiding default settings of devices and Vulnerability management monitoring.
Insider Threat	Separation of duties, Use of the least privilege rule. Regular monitoring. Auditing of systems.
Unpatched Vulnerability	The use of patches. Vulnerability management programs.
Social Engineering Methods	User education. User Awareness. User training. Threat Hunting/intelligence tools.
Bad Credential Management	Separation of duties use of the least privilege rule. Regular monitoring and auditing of systems.

The implication of this is that organizations of all sizes have some vulnerabilities that can be exploited by malware and other social engineering attackers. This requires that organizational leaders not only concentrate on the technical defenses but also regularly evaluate and assess the entire organization to identify loopholes and patch them before cybercriminals exploit them to compromise systems [10]. This result is consistent with the position in the literature that certain vulnerabilities in organizations are usually exploited by malware to gain entry into the system before they move laterally to compromise the systems. Thus, the case study results highly corroborate the outcomes of the literature review and improve the reliability and validity of the study since both industry and the research community seem to grapple with the same problem.

VI. Current Research Focus in Malware Defense

The section presents the synthesis of knowledge on current malware defense. The section focuses on the use of ML techniques in malware defense including the types of ML techniques, the features used, the various algorithms used and the deficiencies

A. The Use of ML in Malware Detection

In recent times, the use of AI and ML techniques for malware detection, generally known as data-mining techniques has been explored in the literature. The advantages of these techniques stems from the fact that, they are capable of detecting previously unknown malware samples, malware family classification. Malware detection can be in the form of classification or clustering [11]. In classification, the model is constructed and model usage. The classifiers include Artificial Neural Networks (ANN), SVM, RF and others, whiles in clustering technique; the task is to group like terms. The overall process of data mining technique is as shown in Fig. 10. Malware and benign ware files are obtained. The features from both are extracted and converted in feature vectors. The features are used as input to an identified ML algorithm such as ANN, SVM, DT, RF, and others. The training set is used for training the models and the test set is used for testing or evaluating the models. The performances of the models is measured using performance metrics. The approach has become the current research focus.

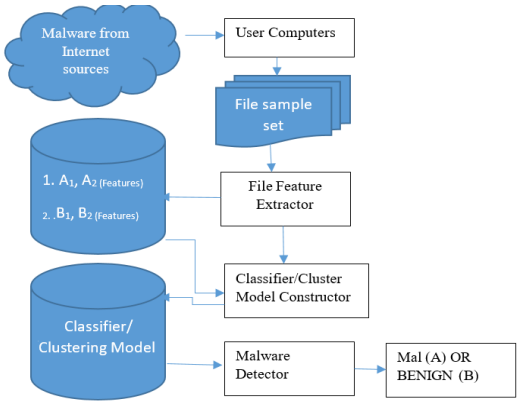

Fig. 10. Overall data mining (ML) approach to malware detection.

The application of ML techniques has become ubiquitous and is being adopted in malware fields. The use of AI and ML techniques in cybersecurity domains remains experimental, with some challenges [37]. He suggested that, notwithstanding the challenges, ML holds very high promise in cybersecurity and malware detection in general. In the following paragraph, we discuss the various issues related to malware detection using ML techniques and identify research gaps. The first thing that comes to mind when ML is mentioned is the use and need for relevant features. Features play a critical role in malware detection using ML techniques [52].

To obtain relevant features, the extracted features should be Feature Engineered (FE). FE is the use of domain knowledge to select and transform features or variables from raw data into vector form for predictive modeling. It embodies exploratory data analysis, where the data is visualized. Feature understanding, in which we get to know the shape of the features; Feature improvement, where the values of columns are changed; feature construction, in which the new feature columns are combined to form new informative features; feature selection, in which noise or irrelevant features are removed from the dataset; and finally, feature transformation, which deals with dimensionality reduction techniques [52].

From the review, the main features used in malware environments are either static or dynamic features, or, in some circumstances, a hybrid of these features [5]. The features used in malware classification include Application Programming Interface (APIs), Dynamic Link Libraries (DLL) and how to extract these relevant features for ML works remains challenging task for both researchers and industry [53]. As depicted in Table X, the features used in ML include static, dynamic, and hybrid features. These are features extracted from static and dynamic environments and can be used to train or build malware detectors [5, 52]. Due to the limitations of these individual features, the use of hybrid features by combining static and dynamic features has proven to be efficient for malware classification. Consequently, this review synthesizes knowledge on the various features available for the use of ML and other AI techniques. The review shows that several features exist that can be extracted to model algorithms; however, these features need to be reduced or selected so that they can lead to high prediction. Thus, by reviewing this literature, we ought to identify very relevant features for training and testing ML methods. This is consistent with [54] demonstration of an efficient feature dimensionality reduction technique for malware detection using malware datasets with ensemble techniques.

In real-life, it is not possible that all features extracted have relevance and contributes to the prediction. This requires that features extracted needs to be selected. This is done by using filter methods, wrappers and embedded techniques. Each of these methods have their own trade-offs [52]. Thus, the goal of features selection is to improve model performance by reducing overfitting, improving accuracy and reducing training time of models [53].

1) ML algorithms in malware detection and features used

From Table X, a number of ML techniques are used in malware detection. The use of ML techniques arises because of the weaknesses in static, dynamic, and hybrid methods [31, 37]. In Table X, a number of ML techniques are used. These can be classified as traditional or conventional MLs and ensemble techniques. The traditional MLs include logistic regression, J48, Nave Bayes, and decision trees. Ensemble ML is also called committee-based learning, in which two or more weaker classifiers are integrated to form a stronger classifier. This could involve bagging or boosting techniques. Examples of ensembles include Support Vector Machines (SVM), Random Forests (RF), gradient boosting, eXtremeGradientBoost [55–59]. They also indicated that Deep Learning (DL) techniques have been used in malware detection by organizations such as CNN [51]. Thus, when using MLs or ensembles, it is necessary to evaluate the performance of the models to ascertain how they perform on unseen data. In addition, we provided some recent works, the analysis technique, the features used, and the type of algorithm used. As depicted in Table X, there are static, dynamic, and hybrid approaches with base algorithms and hybrid approaches where base algorithms are combined with other techniques.

TABLE X. ML ALGORITHMS, STATIC, DYNAMIC, HYBRID AND EVALUATION METHODS

ML Algorithms	Static Features	Dynamic Features	Hybrid Features	Metric
Random Forest	permissions	API Calls	API/API Calls	Accuracy
SVM	API Calls	file system registry activities	FLF/API calls	Precision
J48 Decision Tree	byte code	network activities	import function/functions	Recall
Decision Tree	byte sequence	API sequence	API/API Calls/strings	Sensitivity
Logistic Regression	system calls	DLL function calls	pefile/String	Specificity
Bayesian Networks	n-grams	IP Address		TPR
Neural Networks	API arguments	network traffic		FPR
K-Nearest Neighbour	API Sequence	CPU usage		TNR
Multi-layer Perceptron	Opcode Sequence	Meomory usage		FPR
Gradient Boost	Naïve APIs	processor consumption		F1–Score
Extreme Gradient Boost	IP address	batery usage		AUC
CART	system calls	processor usage		ROC
Radial Bases Function	byte code fragments	syste calla		

As depicted in Table XI, whiles heterogeneous hybridization of malware features is widely explored, not much is done using homogeneous hybridization of two or more features extracted from static environments or dynamic environments and hybridized. In addition, there appears to little or no much work on the use of data augmentation methods to improve malware detection by overcoming the class imbalances, and finally little is done by exposing models to obfuscation methods as a means to improve generalizability. Thus, there is the need for the use of ensemble and data augmentation methods for improved malware detection.

TABLE XI. SAMPLE MALWARE ANALYSIS/DETECTION TECHNIQUES, FEATURES AND THE ML METHODS USED

Authors	Analysis Technique	Features	ML Techniques
Ucci *et al.* [20]	Static	API Calls	Base ML Models
Tahir *et al.* [60]	Dynamic	CPU, Memory, network usage	Base +Ensemble
Cai *et al.* [61]	Static	Communication network traffic	Base +Ensemble
Scalas *et al.* [62]	Hybrid	Permissions, APIs	Base +feature importance
Diwakar [63]	Static	Permission/API/Intents	Base + feature importance
Zhou and Wang [64]	Static	APIs, intents	Base Models

2) Ensemble classification methods

Ensembles are also called committee-based learning or learning multiple classification systems. They are a combination of weaker classifiers to make improved predictions over the individual classifiers [59, 65, 66]. Ensembles are believed to outperform individual classifiers [59]. They combine a number of tree-based algorithms to produce a better- prediction performance than the individual predictors. In machine learning, the performance of a model is influenced by certain factors; the major influence of ML prediction is noise, variability and the usual bias. Thus, the use of ensembles or committee-based learner reduces these factors to the barest. Fig. 11 below is the architecture of an ensemble-based system.

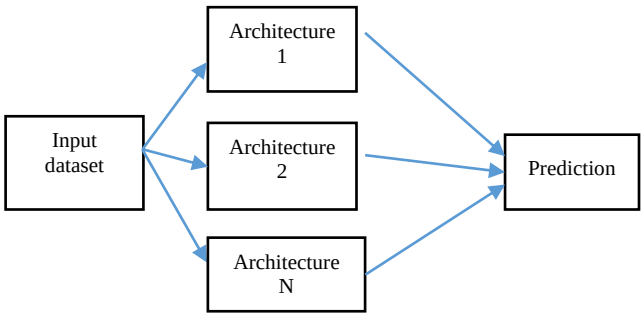

Fig. 11. Ensemble technique.

Due to the noise, variability and bias that is always found in data, under fitting and overfitting occurs which has an impact on the model. The error of training and that of generalization, leads to a generalization Gap that indicates the under-fit or over-fit model as shown in Fig. 12 below.

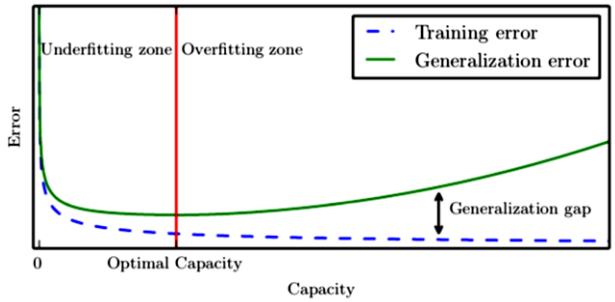

Fig. 12. Overfitting and under fitting in a machine learning model adopted from [67].

3) Constructing ensembles and types of ensembles

Constructing an ensemble learners can take many approaches all with the aim of improving the performance of the model. This is done by tuning the training dataset, by re-sampling from the original dataset. This is usually applied in unstable methods such as NNs, DTs and the rule-based learning algorithms [50]. Examples of these are Bagging and boosting. The other approach is manipulating the features of the dataset using random or a specified method. This process is applicable when the given dataset has many redundant features that do not add to the prediction function of the model features. Example of such model is the Random Forest (RF). The third way of constructing and improving model performance is to manipulate the class labels such as error correcting output coding method, random partitioning of the class labels. This led to the development of two disjointed subsets of the data. Finally, it is also possible to construct and improve model performance by altering the algorithm by reconstruction or changing the topology of the algorithm [59]. Ensembles are classified into two; homogeneous and Heterogeneous ensembles. Homogeneous methods classify models that are created from the same base classifiers. Bagging and Boosting are the examples in this category [66]. Bagging (boostrap Aggregation) is an ensemble for generating predictions and combining them in a simple way to make an improved prediction. The classifiers use only portions of the data and combine them using a simple averaging method. In bagging, a dataset is used to generate similar datasets by sampling with replacement [59]. This concept as demonstrated in Fig. 13 below.

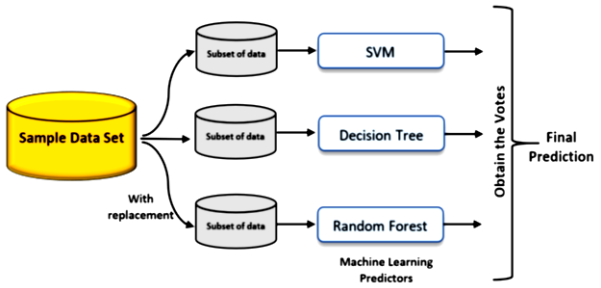

Fig. 13. Bagging approach.

In boosting (Hypothesis Boosting), weak learners' performance is improved by means of iterations a number of times; thus, boosting the strength of the learning algorithm. Examples of these approaches are AdaBoost ensemble algorithm as shown in Fig. 14.

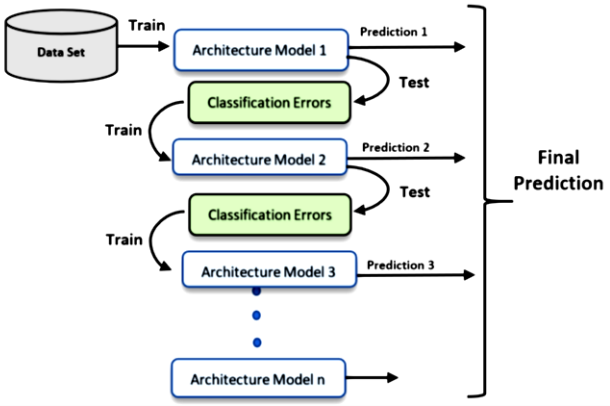

Fig. 14. Boosting approach.

They combine models that are created using different base classifiers. They are usually used when we are not aware which classifier will be useful for a given task. Thus, a number of these models are put together to see the one given the highest performance prediction. The main advantage of these models is that, they each view the data differently and have different assumptions of the data. A lot of studies have been done on the superiority of the heterogeneous methods [30, 67].

4) Ensemble methods in malware classification

Over the years, a number of ensemble method have been developed. Ironically, they are all variants or groups of the known and established algorithms that has been extensively evaluated and the capabilities determined [59]. They indicated that, they are categorized into "hard voting" and "soft voting" approaches. In hard voting the most predicted is the predicted value whiles in soft voting the class labels are predicted mainly by measuring the class.

Hard voting remains the simplest form of majority voting in ensemble languages. Given a class label as:

$$Y^1 = \text{Mode}\ \{C_1(X), C_1(X), \ldots, C_m(X)\} \quad (1)$$

Thus, if given that the three classifiers:

C_1 predict 1

C_2 predict 2

C_3 predict 1

It implies that:

$$Y^1 = \text{mode}\ (1, 2, 1) = 1 \quad (2)$$

Hence, following the majority voting approach then, the sample is classified as Class 1.

B. Weighted Majority Voting

In the case of weighted Majority voting approach, it is calculated by associating a weight W_j with a classifier C_j.

The weighted majority vote can be computed by associating a weight wj in the following form, with a classifier C_j as follows:

$$Y' = \arg \text{Max} \sum_{j=1}^{m} WJXA(CJ(X) = i) \quad (3)$$

where XA is the characteristic function ($C_J(X) = i \in A$, and A represents the set of unique class labels. Hence, for classifiers C_1, C_2, and C_3, given that,

C_1 predict 0

C_2 predict 0

C_3 predict 1,

If the weights associated with them are (0.2, 0.2, 0.6), then

$$Y' = \text{argmax}\ (0.2 \times i_0 + 0.2 \times i_0 + 0.6 \times i_1) = 1 \quad (4)$$

C. Soft Voting Approach

In this approach, the labels of the class are predicted based on the predicted probabilities p for the classifier. We use this approach if the classifiers are well calibrated**.**

Y'= argmax$_i$ $\sum_{j=1}^{m} Wjpij$ where W_j is the weight assigned to the jth classifier. For example, for a binary classification and given C_1, C_2, and C_3 at (0.9, 0.1), (0.8, 0.2), and (0.4, 0.6), respectively, using uniform weights, the probability averages are as follows:

$$P(i_0/X) = (0.9+0.8.0.4)/3 = 0.7$$

$$P(i_1/X) = (0.1+0.2+0.6)/3 = 0.\ 3.$$

On the other hand, if the weights are given as (0.1, 0.1.0.8), the prediction, Y' = 1, i.e., $P(i_0/X) = 0.1 \times 0.9 + 0.1 \times 0.8 + 0.8 \times 0.4 = 0.49$.

$P(i_1/X) = 0.1 \times 0.1 + 0.2 \times 0.1 + 0.8 \times 0.6 = 0.51$. But, $P(i_0/X)$, $P(i_1/X) = 1$. Consequently, depending on the type of ensemble technique used, noise, variability and bias are eliminated leading to improved prediction and classification performance.

D. Performance Evaluation of ML Models (Metrics and Formulae)

The efficiency of an ML technique is usually measured using different metrics, as shown in Eqs. (5)–(14). A careful understanding of the metrics, the algorithms, and the data size, type and variety is key in selecting a metric for the evaluation of learning models. For example, when there are imbalances in the dataset, the use of the accuracy metric is not good as it leads to the ‘accuracy paradox," where the model is skewed towards the majority class at the expense of the minority [63]. In addition, several factors influence the performance of the techniques, such as the type of features, the feature selection method or dimensionality reduction techniques, the type of algorithm, and the classifier parameters or hyper-parameters. To evaluate the performance of ML models, a test dataset is used after the use of the training dataset. This set should contain the right labels or the observed labels for all the data points or instances. These observed labels are used to compare with the predicted labels to determine the performance evaluation of the models after calculation using the various variables from the confusion matrix as shown in Table XII for a binary classification problem.

TABLE XII. CONFUSION MATRIX

Predicted	Actual Malware	Non-Mlaware
Malware	True Positive(TP)	False Positive(FP)
Non-Malware	False Negative(FN)	True Negative (TN)

Note: FN = False Negative, TP = True Positive, FP = False Positive, TN = True Negative, PPV = Positive Predictive Value, NPV = Negative Predictive Value.

- Accuracy: It measures the cases a model predicted correctly, i.e., it measures the number of all the correctly classified samples over the total sum of the dataset used. The best of the accuracy metric is 1or 100%, whiles the worse accuracy of a model is 0 or 0%.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

- Error Rate: This is a measure of the number of all the incorrect classifications over the total number of the dataset. The best error rate is 0.0 whiles the worse is 1.0.

$$Error\ Rate = 1 - Accuracy \quad (6)$$

- Positive Predictive Value/Precision: This metric measures the fraction of truly positive samples from all the cases the model predicted as positive. In other words, Precision is measured as the number of correctly predicted or classified divided by the total number of positive predictions. Known also as the positive predictive value with the best being measure being 1 and the worse being 0. Thus, the correctly classified malware over the total number of the positively classified malware.

$$Precison\ or\ Precision = \frac{TP}{TP+FP} \quad (7)$$

- Sensitivity/Recall: Recall refers to the model’s ability to classify or predict all the relevant cases inherent in a dataset. It is the ration of the True Positives over the true positives plus the false negatives. Sensitivity/Recall/True positive is measured as the correct negative classification value over the total number of the positives. Best value of this metric is 1 and the worse is 0.

$$\text{Sensitivity/Recall} = \frac{TP}{TP+FN} \quad (8)$$

- False Positive Rate/Fall out Rate or probability of false alarm: refers to the fraction of sample predicted as positive, but the sample were actually negative.

- Specificity/True Negative: This metric is calculated based on the number of correctly negative classification over the total number of the negatives. Its best value is 1, whiles the worse is 0.

$$Specificity/(TNR) = \frac{TN}{TN+FP} = \frac{TN}{N} \quad (9)$$

- F1-Score: The harmonic mean of positive predictive value and the sensitivity of a given sample used in the machine learning activity.

$$F1 - Score = \frac{2\times TP}{2(TP+FP+FN)} \quad (10)$$

- Negative Predictive Value (NPV): The truly negative samples predicted as negative.

$$NPV/TNR = \frac{TN}{TN+FN} \quad (11)$$

- False Negative Rate/Miss Rate: This refers to the fraction of the sample that are predicted negative, but are actually positive.
- Root Mean Square Error: The RMSE of a model is the average of the squared differences between the predicted and the observed outcomes in the ML model usually used in regression model. A lower RMSE shows better the model 'accuracy. When the variation is wide, it implies greater discrepancies between the predicted and the outcome variable.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(t1 - y1)} \quad (12)$$

- Mean Absolute Error: MAE is a statistical metric that measures the average size of the mistakes found in a collection of predictions without taking into account the directions.

$$MAE = \frac{1}{n}\sum_{1}^{n}(t1 - y1) \quad (13)$$

- Area Under the Curve: AUC in ML domain measures the whole two-dimensional area under the ROC usually from points (0,0) to (1,1). It provides an aggregate measure of the performance of the model(s) across all the likely classification thresholds.

$$AUC = \int_0^1 \frac{TP}{TP+FN}\, d\frac{FP}{FP+TN} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \quad (14)$$

- Root Mean Square Logarithmic Error (RMSLE): RMSLE is determined or calculated by the application of log to the actual and the predicted values and calculating the differences. This metric is robust in cases where there are outliers where the small and large errors are treated evenly. This metric is given as

$$RMSLE\sqrt{MSE(\log(yn+1), \log(yn+1)} \quad (15)$$

In conclusion, the performance measures used in modelling the phenomenon would determine whether it meets the required precision required for deployment. This performance over the years has been shown to be influenced by certain variables in the ML modelling process that militates against the achievement of the full potential of the algorithms. The next section of presents the gaps identified from the scoping review supported by the results of the empirical case studies about ML performance issues.

VII. Identified Research Gaps

Overall, the current malware explosion in volumes, variety, and complexity, coupled with the new and innovative techniques to evade detection, remains a major challenge facing both industry and the academic community. As a result, all cybersecurity stakeholders are investing heavily to ensure optimum security and protection of their High-Value Assets (HVA), as exposure to these assets has disastrous consequences [10]. Nevertheless, the existing defense methods and tools exhibit some limitations that are exploited by malware attackers. They use innovative obfuscation techniques and other advanced methods, making it cumbersome for defenders to protect their information assets and resources against ever-growing, revolving, and evolving malware attacks. This requires innovative techniques, and the use of machine learning, ensemble learning, and deep learning has become the new approach that has come to the rescue of signature-based detection systems. However, the review of the papers shows that there remain some challenges facing the use and adoption of ML techniques in malware detection. The proceeding paragraphs discusses the various limitations and or research gaps that impede the utilization of the full potentials of ML techniques in cybersecurity in general and malware defense in particular.

The problem of poor malware Features: Features play critical roles in the success of every machine-learning algorithm. One of the problems identified from the studies is poor features that can lead to the efficient classification of malware [23, 31].

The problem of class imbalances and the 'accuracy paradox': most malware environments and the datasets used in malware experiments are largely imbalanced between the benign and malware classes. However, many authors still use the accuracy metric as their performance measure. This results in most of the algorithms being biased or skewed towards the majority sample or class and usually treating the minority class as noise, leading to misclassification and inaccurate classification [36, 37, 63].

Another gap identified in the literature is the fact that most of the current malware adopts high-level obfuscation techniques such as polymorphic, metamorphic, and others to evade detection by the existing malware defense architectures. Such techniques with self-mutating capability easily pose a challenge as authors fail to empirically experiment with their datasets with real-time obfuscation methods [25, 33].

The use of relatively smaller data sizes and synthetic malware datasets: The malware ecosystem is rapidly

evolving, leading to a volatile and unstable environment with huge and humongous malware attacks. Thus, the use of smaller datasets in the experiments leads to overfitting. Similarly, with the changes in malware, the use of synthetic datasets and not real-time threats poses a challenge [37].

Lastly, the use of traditional or conventional ML techniques without the use of data augmentation techniques results in inefficient performances [37, 67].

Finally, the over-reliance on only technical controls at the expense of non-technical techniques (users, organizational) constitutes limitations that are exploited by the malware adversarial groups to compromise systems [36].

VIII. Suggested Future Directions

In this paper, we conducted an extensive evaluation of current malware trends, the attack and defense strategies, current research approaches and the research gaps. Notwithstanding the achievement of the objectives of the study, there remain some open issues and trends that needs attention from the both industry and the research community. In this section, we discuss these issues and the suggested research directions as a means towards improving the effectiveness, efficiency and robustness of Malware defense.

(1) **The exponential increase in malware volumes, variety, and complexity.** The malware scare at both the industry and individual levels is assuming pandemic levels [8]. Malware is not only increasing in volume but also in variety and complexity as more daily records of malware reach 450,000, according to [8]. This translates to 13,500,000 monthly and 162,000,000 annually. These numbers, coupled with the increased variety and the high level of mutating ability, pose a major challenge for traditional and conventional machine learning and signature-based detection techniques [37, 54]. This is expected to increase as the world is digitalized because of the growth and expansion of the internet and related technologies.

(2) **The exponential growth and expansion of the Internet of Things (IOTs)** is on. Almost every device is connected, such as home appliances, smart energy systems, and other devices connected to scanners and sensors at personal, corporate, national, and international levels and dimensions. This trend is expected to increase with the corresponding malware risk in particular and cyber security in general. Exploring novel defenses in this area should be a subject of concern for both industry and the academic community [45].

(3) **The drift to cloud-based products and services and the security implications.** There is no denying the fact that cloud products are expanding at an alarming rate. As more data and other information resources become available, including critical infrastructure such as energy, health records, banking, and others, it will attract the attention of more cyberattacks. How do we fare in the face of a marauding malware onslaught? This has high security implications going into the future, and the trend is expected to increase as more data resources migrate to the cloud [45].

(4) **The Increased use and penetration of mobile telephony:** It is apparent that the world is currently inundated with mobile phones and has high mobile connectivity and usage. The trend is expected to increase over the years as governments and other stakeholders support and encourage the use of mobile phones for learning (M-learning), banking (M-banking), and other purposes. This ubiquitous adoption and use is expected to increase. Thus widening the attack surface. These attacks are highly targeted and sophisticated. The security implications of such an invasion are expected to be dire for both users and organizations that fail to employ the right defense techniques [6, 22]. This trend has already taken a toll on the Android OS as more malware targets this application system.

(5) With such huge numbers of malware and the sophistication and dexterity with which they deploy them, how will data mining and other AI techniques fare? Understanding the limitations of the current malware tools and how we can deploy alternative techniques, including Generative Adversarial Networks (GANS), ensemble techniques, and other deep learning methods, is likely to offer a lifeline for the future [37, 68].

(6) **Identity as a Security Perimeter**: Digitization and the digitalization of goods and services are bringing a new dimension to how we use technologies [69]. This concept is defying the usual physical and logical perimeters as users consume services anywhere, at any moment, and on the go, making use of increasingly interconnected devices, cloud systems, and things. Though this phenomenon is improving the global cyber inclusion drive, it comes with an inherent risk to people, systems, and things that border outside the usual traditional security perimeter, rendering the traditional perimeter obsolete. With highly mobile users making use of several connected mobile devices all connected to the internet, a security model focusing on user identity and access control is eminent. With this Identity-Based Security Model (ISM), the emphasis is on identity across all interconnected devices across platforms, which enables organizations to provide holistic authentication and authorization and manage users, things, and systems. We propose that this new mode can be improved by embedding machine learning and AI to provide insights on user identity threats for quick response and remediation efforts.

(7) **Insider Threat as a Service**: Globally, the cybersecurity crisis is assuming pandemic levels, and all stakeholders must keep one thing in mind: the best way to exploit and infiltrate the security perimeter of organizations is from the inside. This

is because it is estimated that two-thirds of all data breaches are known to be caused by insiders [70]. The report indicated that the cost of insider threat events increased by 34% from 2020 to 2022, peaking at $15.38 million. The time to resolve the incident also peaked at 85 days, up from 77. With the current volatile malware and cybersecurity landscape, insider threats are not only increasing but also providing a new and motivated attack option for attackers. Cyber adversaries that compromise organizations in exchange for incentives recruit malicious insiders. The report also cited improved security and resistance to traditional attack methods, thriving dark web markets, growth in remote employment avenues, and geopolitical reasons underscoring the growth of Insider Threat as a Service. This is expected to increase going into the future as organizations implement an Insider Threat Program (ITP) that can leverage intelligence to provide insights about threat attacks, improve collaboration with other cybersecurity stakeholders, and invest heavily in employee training, awareness, and education programs.

IX. Conclusion and Future Works

The armed race between malware assailants and defenders makes the ecosystem highly volatile, dynamic and stochastic in nature. To understand the malware phenomenon at any given time requires an evaluation or survey to establish the state-of-affairs for remediation and other decision-making efforts. To this end, this scoping review supported by empirical case studies was conducted. The results show that malware attacks are increasing in volumes, variety and complexity making existing defense system inefficient and ineffective. Current malware are targeted, persistent, unknown and stealth in nature compared with the traditional malware that were open, known, broad and one-time, which poses challenges to effective defense by existing signature based methods. The new malware adopt both anti-static and advanced obfuscation techniques to evade detection leading to exploitation and compromise of confidentiality, integrity and availability. In addition, the study established the comparative weaknesses and strengths of existing malware defense methods including static, dynamic, hybrid and ML techniques relative to effective and efficient defense. Finally, the study revealed that, the use of conventional ML techniques in malware defense, poor and redundant malware features, class imbalances and the resultant 'accuracy paradox', poor resilience and robustness in detecting unknown malware coupled with user and organizational vulnerabilities constitutes the research gaps and challenges facing effective and efficient malware defense. Therefore, to improve malware defense would require the adoption of novel techniques that addresses these gaps. Consequently, the paper concluded that, the use ML techniques alone is necessary but not sufficient condition of providing holistic malware defense since user and organizational vulnerabilities constitute a part of the challenge. This implies that both industry and the research community needs to refocus on the use of hybrid approaches involving both technical and non-technical controls to ensure holistic malware security. Notwithstanding the relative success of the study, some limitations need future consideration. Thus, future works would adopt other review methods such as systematic literature review approaches and narrow the topic to explore the phenomenon further. In addition, apply novel techniques to improve upon the gaps identified in current research, such as the use of identity-as-security-perimeter, the use of data augmentation techniques, improved malware features, the use of Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GANS) with effective user control and behavior change programs and models for efficient detection and preventive defense

Conflict of Interest

The authors declare no conflict of interest.

Author Contribution

Azaabi Cletus formulated the topic, developed the scoping review format, obtained the required papers, and conducted the analysis; Alex Akwasi Opoku did part of the data extraction, review of the manuscript and set the structure; Benjamin Asubam Weyori did part of the data extraction and analysis, review of the manuscript for grammatical errors and others; all authors had approved the final version.

References

[1] B. Kenneth and F. Ken, "Metamorphic malware and obfuscations: A survey of techniques, variants and generation kits," *Security and Communication Networks*, 2023

[2] M. Goyal and R. Kumar, "A survey on malware classification using machine learning and deep learning," *International Journal of Computer Networks and Applications*, vol. 8, no. 6. 2021.

[3] A. A. Hamza, I. T. A. Halim, M. A. Sobh, and A. M. B. Eldin, "A survey and taxonomy of program analysis for IoT platforms," *Ain Shams Engineering Journal*, vol. 12, no. 4, 2021.

[4] D. Airehrour, N. V. Nair, and S. Madanian, "Social engineering attacks and countermeasures, in the New Zealand banking system: Advancing a user-reflective mitigation," *Information, and Austria*, vol. 9, no. 5, 110, 2018.

[5] K. A. Monnappa, *Learning Malware Analysis: Explore the Concepts, Tools and the Techniques*, Packt Publishing Ltd, 2018.

[6] H. Aldawood and G. Skinner, "Reviewing cyber security social engineering training and awareness programs-Pitfalls and ongoing issues," *Future Internet*, vol. 11, no. 3, 2020.

[7] A. A. Alhashmi, A. Darem, and J. H. Abawajy, "Taxonomy of cybersecurity awareness delivery methods: A countermeasure for phishing threats," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 10, 2021.

[8] AV-Test Institute. Annual Malware statistics, Malware Statistics. [Online]. Available: http://www.av-test.org/en/statistics/malware

[9] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, 2019.

[10] T. Rains, *Cybersecurity Threats, Malware Trends, and Strategies, Mitigate Exploits, Malware, Phishing and Other Social Engineering Attacks*, Packt Publishing, 2020.

[11] A. Heena and M. Mehtre, "Advances in malware detection-An overview," *Institute for Development and Research in Banking Technology*, pp. 5371–5396, 2021.

[12] R. Raghaendra and M. V. Dutta, "Machine learning in malware detection: A survey of analysis techniques," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 12, no. 4, 2023.

[13] P. Faruki *et al*., "A survey and evaluation of android-based malware evasion techniques and detection frameworks," *Information*, vol. 14, 2023.

[14] T. Bilot, E. N. Madhoun, A. K. Agha, and Z. Anis, "A survey on malware detection with graph representation learning," arXiv preprint, arXiv:2303.16004, 2023.

[15] H. E. Merabet and A. Hajraoui, "A survey of malware detection techniques based on machine learning," *International Journal of Advance Computer Science and Applications*, vol. 10, no. 1, 2019.

[16] A. A. Hamza *et al*., "HSAS-MD analyzer: A hybrid security analysis system using model-checking technique and deep learning for malware detection in IoT apps," *Sensors*, vol. 22, no. 3, 1079, 2022.

[17] M. Bahri *et al*., "Efficient Batch-incremental classification using umap for evolving data streams," *Advances in Intelligent Data Analysis XVIII*, pp. 40–53, 2020.

[18] F. A. Aboaoja *et al*., "Dynamic extraction of initial behavior for evasive malware detection," *Mathematics*, vol. 11, no. 2, 2023.

[19] R. Sihwail, K. Omar, K. A. Z. Ari, and S. A. Afghani, "Malware detection approach based on artefacts in memory image and dynamic analysis," *Applied Sciences*, vol. 9, no. 18, 2019.

[20] D. Ucci, L. Aniello, and R. Baldoni, "Survey on the usage of machine learning techniques for malware analysis," arXiv preprint, arXiv 1710.08189, pp. 1–67, 2018.

[21] V. Kouliarridis and G. Kambourakis, "A comprehensive survey on machine learning techniques for android malware detection," *Information*, vol. 12, no. 5, 2021.

[22] A. Moses and M. Sarah, "Analaysi of android malware detection techniques: A systematic review," *International Journal of Cybersecurity and Forensics*, vol. 8, no. 3, pp. 177–187, 2019.

[23] M. Paula, C. Christopher, and G. Kathering, "A naturalistic methodology for assessing susceptibility to social engineering through phishing," *The African Journal of Information Systems*, vol. 11, no. 3, 2019.

[24] S. Musah, A. George, and R. S. Weir, "Predicting individuals' vulnerability to social engineering in social networks," *Cybersecurity*, vol. 3, no. 7, 2020.

[25] C. Azaabi, A. O. Alex, and B. A. Weyori, "Improving Social Engineering Awareness, Training and Education (SEATE)," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022.

[26] C. S. Bhusal, "Systematic review on social engineering: Hacking by manipulating humans," *J. Inf. Security*, vol. 12, 2021.

[27] J. Singh and J. Singh, "Challenges of malware analysis: Obfuscation techniques," *International Journal of Information Security Science*, vol, 7, no. 3. 2018.

[28] W. N. P. Pongkham and K. Sripanidkulchai, "Comprehensive detection of vulnerable personal information leaks in android applications," in *Proc. IEEE Conference on computer Communications Workshop*, 2020, pp. 121–126.

[29] D. Chaulagain *et al*., "Hybrid analysis of android apps for security vetting using deep learning," in *Proc. IEEE Conference on Communication and Network Security*, 2020.

[30] J. Saxe and H. Sanders, *Malware Data Science: Attack Detection and Attribution*, No Starch Press, Inc., 2018.

[31] E. Masabo, K. S. Kaawaase, J. S. Otim, J. Ngubiri, and D. Hanyurwimfura, "Improvement of malware classification using hybrid feature engineering," *SN Computer Science*, vol. 17, 2019.

[32] K. B. M. Yunus and S. B. Ngah, "Review of hybrid analysis technique for malware detection," in *Proc. IOP Conference Series: Materials Science and Engineering*, 2022.

[33] C. Lu, "Malware detection methods," in *Proc. International Conference on Computing and Data Science*, 2018, pp. 7–18.

[34] C. Azaabi, A. O. Alex, and B. A. Weyori, "Improving Social Engineering Awareness, Training and Education (SEATE)," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022.

[35] C. Hadnagy, *Social Engineering: The Science of Human Hacking*, John Wiley & Sons, Inc., 2018.

[36] C. Azaabi, A. O. Alex, and B. A. Weyori, "Exploring the performance of feature dimensionality reduction technique using Malware Dataset," *International Journal of Computer Science and Network Security*, vol. 22, no. 6, 2022.

[37] A. Parisi, *Hands-on Artificial Intelligence for Cybersecurity. Implement Smart AI System for Preventing Cyber-Attacks and Detecting Threats and Network Anomalies*, Packt Publishing, 2020.

[38] F. Shihab, *et al*., "Preliminary analysis of malware detection in opcode sequences within IOT environment," *Journal of Computer Science*, vol. 16, no. 9, 2020.

[39] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *International Journal Social Research Methodology*, pp. 19–32, 2005.

[40] Tricco *et al*., "PRISMA extension for scoring reviews (PRISMA-ScR): Checklist and explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[41] M. D. J. Peters *et al*., "Updated methodological guidance for the conduct of systematic reviews," *JBI Evid Synth*, vol. 18, no. 10, 2020.

[42] S. Oliver, "Marking research more useful: Integrating different perspectives and different methods," *Buckingham: Open University Press*, pp. 167–179, 2001.

[43] O. Awotipe, "Log analysis in cyber threat detection," *Creative Components*, vol. 468, 2020.

[44] G. Bassett, C. D. Hylender, P. Langloise, A. Pinto, and S. Widup "Verizon data breach investigations report 2022," *Comput. Fraud Secur*, vol. 4, 2020.

[45] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–40, 2018.

[46] H. Darabian *et al*., "Detecting cryptomining malware: Deep learning approach for static and dynamic analysis," *Journal of Grid Computing*, vol. 18, 2020.

[47] M. Naseer, "Malware detection: Issues and challenges," *J. Phys: Conf. Serv*., 1807, 2021.

[48] J. B. Higuera *et al.*, "Systematic Approach to Malware Analysis (SAMA)," *Applied Science*, vol. 10, no. 4, 2021.

[49] C. Azaabi, A. O. Alex, and B. A. Weyori, "Exploring the performance of feature dimensionality reduction technique using Malware Dataset," *International Journal of Computer Science and Network Security*, vol. 22, no. 6, 2022.

[50] D. Kim. (2019). Improving existing static and dynamic malware detection techniques with intrusion-level behaviour. Digital Repository at the University of Maryland. [Online]. Available: https://doi.org/10.13016/m21q-qhlu

[51] P. Duboue. (2020). The art of feature engineering. Essentials for machine learning. [Online]. Available: https://doi.org/10.1017/9781108671682

[52] S. Ozdemir and D. Susarla, *Feature Engineering Made Easy: Identifying Unique Features from Your Dataset in Order to Build a Powerful Machine Learning Systems*, Birmingham, Mumbai, 2018.

[53] C. Azaabi, A. O. Alex, and B. A. Weyori, "Exploring the performance of feature dimensionality reduction technique using Malware Dataset," *International Journal of Computer Science and Network Security*, vol. 22, no. 6, 2022.

[54] D. Su, J. Liu, X. Wang, and W. Wang, "detecting android locker-ransomware on Chinese social networks," *IEEE Access*, vol. 1, no. 7, 2018.

[55] S. Alsoghyer and I. Almohaeni, "On the effectiveness of application of permissions for android ransomware detection," in *Proc. 2020 6th Conference on Data Science and Machine learning Applications*, 2020, pp. 94–99.

[56] F. Shihab *et al.*, "Preliminary analysis of malware detection in opcode sequences within IOT environment," *Journal of Computer Science*, vol. 16, no. 9, 2020.

[57] S. Sumathi *et al., Advance Decision Sciences Based on Deep Learning Algorithms: A Practical Approach Using Python*, Nova Science Publishers, New York, 2021.

[58] N. Potha, V. Kouliaridis, and G. Kambourakis, "An extrinsic random-based ensemble approach for android malware detection," *Connect. Sci*., pp. 1–17, 2020.

[59] M. K. Alzaylaee, S. Y. Yerima, and S. Sezer, "DL-Droid: Deep leaning based android malware detection using real devices," *Comput. Secur*., 101663, 2020.

[60] R. Tahir *et al*., "Similarity-Based android malware detection using hamming distance of static binary features," *Future Genr. Comput Sys.*, vol. 105, 2020.

[61] L. Cai, Y. Li, and Z. Xiong, "JOWMDriod: Andriod malware detection based on feature weighting with joint optimization of weight-mappping and classifier parameters," *Comput. Secur*., vol. 100, 2021.
[62] M. Scalas *et al*., "Practical on-service detection of android ransomware," arXiv preprint, arXiv:1805.09563v1, 2018.
[63] R. Diwakar, "Handling imbalance data with imbalance-learn in python," *Data Science Blogathon*, vol. 101, 2023.
[64] Y. Zhou and P. C. Wang, "An ensemble learning approach for XSS attack detection with domain knowledge and threat intelligence," *Computers and Security*, vol. 82, 2019.
[65] H. Dhamija and A. K. Dhamija, "Malware detection using machine learning classification algorithms," *International Journal of Computational Intelligence Research*, vol. 30, no. 4, 2021.
[66] A. Maryam *et al*., "Chybridoid: A machine learning-based hybrid technique for securing the edge computing," *Security and Communication Networks*, 2020.
[67] H. Hallqvist and J. Luhr, "Fast classification of obfuscated malware with and artificial neural network," Thesis, RKT Royal Institute of Technology, 2022.
[68] A. Cletus, A. A. Opoku, and B. A. Weyori, "A homogeneous multi-static hybrid features with ensemble and data augmentation for efficient malware variant detection," *Journal of Theoretical and Applied Information Technology*, vol. 10, 2023.
[69] R. G. Shende. (2023). Identity as a new security perimeter. [Online]. Available: https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2023/volume-21/identity-as-a-new-security-perimeter
[70] M. Williams and B. Kohy. Detecting insider threat behaviors using social media platforms. [Online]. Available: https://www.isaca.org/resources/news-and-trends/industry-news/2022/detecting-insider-threat-behaviors-using-social-media-platforms