IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 April 2024, accepted 10 May 2024, date of publication 15 May 2024, date of current version 23 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401578

SURVEY

Maleficent Neural Networks, the Embedding of Malware in Neural Networks: A Survey

STEPHANIE ZUBICUETA PORTALES[1] **AND MICHAEL ALEXANDER RIEGLER**[2]
[1]Independent Researcher, 5056 Bergen, Norway
[2]Department of Holistic Systems, Simulamet, 0167 Oslo, Norway

Corresponding author: Stephanie Zubicueta Portales (stephanieportales@yahoo.com)

ABSTRACT In this study, we address the evolving threat of Maleficent Neural Networks, also known as ''Evil'' Neural Networks, malicious neural networks embedded with malware. Due to the absence of effective detection mechanisms, these malicious models remain undetected, posing significant challenges to the security of users and systems in the rapidly expanding field of Artificial Intelligence and Machine Learning. This research provides a comprehensive examination of Maleficent Neural Networks, and their detection, mitigation, and security issues, based on recent foundational studies. A discussion of ethical and legal concerns surrounding the deliberate infusion of malware into neural networks is also included, emphasising the need for collaborative efforts among experts in the fields of AI, machine learning, and cyber security. The study shows that this new threat possesses several risks, and the number of works on the topic we identified confirms that more research is needed in this direction. Moreover, we propose promising future directions, including the creation of advanced adversarial defence mechanisms and the development of new methods to detect malware within neural networks.

INDEX TERMS Adversarial machine learning, cyber security, malware detection, neural network security.

I. INTRODUCTION

In this study, we investigate the threat of malicious models we call Maleficent Neural Networks (MNNs), or often ''Evil'' Neural Networks, which are neural networks embedded with malware. In the absence of detection mechanisms, this new threat remains undetected, making it difficult to protect users and systems.

With the development of new machine learning and deep learning techniques, malware detection has advanced over the years. Malware has become a significant cyber security threat over the last few decades, infiltrating computers, damaging them, and stealing sensitive data. Parallel to the growth of malware, anti-detection techniques have also evolved. Furthermore, as machine learning and artificial intelligence (AI) have grown rapidly, malicious models are spreading across multiple domains [1], [2].

However, security is overlooked in this field despite its critical importance in protecting systems and products from malicious intent and unauthorised access.

The associate editor coordinating the review of this manuscript and approving it for publication was Pietro Savazzi.

A recent study by Wang et al. [3] reveals that it is possible to embed malware into layers of neural networks, without affecting the performance of the model. This discovery not only renders malware undetectable, but also elevates the level of threat, emphasising the potential dangers posed by these evil models.

Therefore, the primary source of information on the creation of malware embedded within neural network layers is the studies conducted by Wang et al. in 2021 and 2022 [3], [4].

Although these studies serve as a foundational piece in this area, it is essential to contextualise them within the broader landscape of research related to malicious activities in machine learning and AI.

Several studies have contributed significantly to our understanding of security challenges in these domains. Despite these valuable contributions, a comprehensive exploration of the detection, mitigation, and security aspects of malware embedded within neural network layers remains an ongoing challenge. Existing research, including Wang's work, lays the foundation for understanding this threat, but further studies are required to fill gaps in knowledge and provide more

robust solutions. Therefore, this study represents the first comprehensive investigation of the detection, mitigation and security aspects surrounding this particular threat. In the following sections of this paper, we build on existing literature, incorporating insights from various sources, and propose possible approaches to address the challenges posed by MNNs. Our research extends beyond the current state of understanding, with the aim of contributing novel methodologies and strategies for the detection and mitigation of MNNs.

The main contributions of this work are:

- Comprehensive literature review and analysis of the state-of-the-art
- Analysis of existing approaches to embed malicious code into neural networks
- Analysis of possible counter measurements
- Propose future research directions and guidelines

The paper is structured as following. First, in Subsection II, we explore related work. Subsection II-A provides a comprehensive background on MNNs, exploring fundamental concepts such as the embedding of malware into neural network layers. Detection techniques are detailed in Subsection II-B, and Subsection II-C covers mitigation strategies, both extracted from an extensive literature review. Section III discusses challenges in identifying MNNs, ethical and legal considerations, and proposes future research directions. Finally, the paper concludes the work and the main findings in Section IV.

II. RELATED WORK

In recent years, researchers have increasingly focused their attention toward the critical issue of malware detection in diverse domains. This surge of interest underscores the growing impact of these diverse threats with MNN being one of the latest and less explored ones. In this section, we explore different approaches to malware detection explored in several studies that are relevant to MNNs.

1) SEARCH STRATEGY

Our search strategy was designed to capture the depth of malware detection research to comprehensively survey the existing literature. To perform our search, we targeted relevant keywords and phrases across reputable databases and modified them to suit the syntax of each database. This can be seen in Table 1. For the literature search, the keywords 'Malware' and 'Neural Network' were used, while 'Classification' was excluded. Although the search strategy was modified to comply with the requirements of the individual databases, the keywords were selected to focus on the intersection of malware and neural network research without focusing on classification.

Adhering to this strategy and going through a rigorous, step-by-step elimination process as shown in Fig. 1, we systematically filtered through 5492 articles extracted from the various databases. This meticulous approach led to the identification and in-depth analysis of the 12 most pertinent articles.

TABLE 1. Overview of search strategy and results for relevant research papers.

Keywords	Database	Results	Unique
"Malware" AND "Neural Network" AND NOT "Classification"	ScienceDirect	288	288
	IEEE Xplore	423	395
	ACM Digital Library	196	193
	Web of Science	188	154
Malware "Neural Network" - Classification	Google Scholar	4400	2859
"Neural Network" AND (Malware) AND NOT (Classification)	SpringerLink	528	353
Total		5033	**4232**

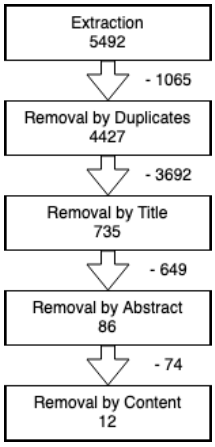

FIGURE 1. Overview of elimination strategy process.

This detailed analysis allowed us to uncover the inherent challenges associated with embedding malware in neural networks. Additionally, we gained insight into the techniques and methods employed for such embedding. Understanding these methodologies helps us to explore countermeasures effectively.

The findings highlighted neural networks' susceptibility to adversarial attacks, where malicious actors can manipulate inputs to deceive the network into producing incorrect outputs or directly embed malware into certain layers of neural networks with malicious intents. These findings demonstrate the urgent need to develop frameworks, guidelines, and detection mechanisms to mitigate, detect, and defend against such attacks. This makes it clear that further research and development in neural network security is crucial and urgent.

This comprehensive literature review revealed various approaches and methodologies used for embedding malware into neural networks. An overview of the literature review can be found in Table 2. The analysis of the final 12 articles not only provided valuable information on the topic, but also identified key research directions for future studies, offering a comprehensive contribution to the current research setting.

A. MALEFICENT NEURAL NETWORKS

In this research, the primary literature that improves our understanding of malware embedded within neural network layers consists of two important papers written by

TABLE 2. Overview of research papers.

Category	Year	Reference	Key Focus	Main Contribution/Technique
Malware Embedding	2021	[3], [4]	Fast Substitution, Batch Normalization, and Retraining	Embedding malware in neural networks while preserving functionality.
	2022		Countermeasures against EvilModels	Adjustments to parameters, model modification, and safeguarding the model supply chain
Detection Techniques	2018	[5], [6]	Detection methods (signature, ML, DL)	Signature-based, ML-based, and DL-based detection methods
	2020		Op-code analysis for static analysis	Improved accuracy through file-size-based segmentation and Multi-Layer Perceptron-Deep Neural Network
	2022	[7]	Ensemble methods, adversarial attacks	Use of FGSM, GANs, and ensemble methods to enhance robustness
	2020	[2]	Comprehensive exploration of detection techniques	Static (signature, heuristic) and dynamic (behaviour-based) detection
	2023	[8]	Dynamic detection with Instruc2Vec framework	Introduction of Instruc2Vec for dynamic detection of malicious code
	2019	[1]	DQEAF framework for exposing weaknesses	Proposal of DQEAF framework to expose weaknesses in supervised learning
	2018	[9]	Deep CNN for malware detection	Utilisation of deep CNN to analyse sequences of grouped instructions
	2022	[10]	Evolution of malware detection using deep learning	Achieving 100% accuracy with ANN and fuzzy mathematical model
	2017	[11]	Deep learning-based malware detection using static analysis	Addressing challenges posed by traditional methods in adapting to massive data
	2023	[12]	AMGmal for addressing vulnerabilities to adversarial examples	Use of saliency detection and mask guidance to prioritise critical bytes
Mitigation Strategies	2018	[5], [6]	Deep Learning architectures for feature extraction and classification	Efficacy of Auto-Encoders and Deep Neural Networks for malware defence
	2020		Feature engineering and LSTM configurations	Importance of precautions during model porting and transfer for resilience
	2022	[7]	Resilience in Adversarial Machine Learning	Advocating for resilience in malware detection models against adversarial attacks.
	2020	[2]	Challenges in anti-malware technologies	Emphasis on adaptive strategies due to developments in encryption and obfuscation
	2023	[8]	Resource conservation and accuracy maintenance	Focus on feature selection and hybrid deep learning frameworks
	2019	[1]	DQEAF framework for evading traditional detection engines	Use of reinforcement learning for evading detection engines
	2018	[9]	Lightweight CNN for efficient feature extraction	Utilisation of disassembled instructions for identifying polymorphic and zero-day malware
	2022	[10]	Multilayered ANN for comprehensive understanding	Incorporation of data from various sources for a comprehensive understanding
	2017	[11]	Deep learning with feature extraction	Use of grayscale images and OpCode 3-grams for automated and adaptable malware characterisation
	2023	[12]	AMGmal for adversarial attack mitigation	Strategy involving fooling visualisation-based detectors, reserving functionality, and minimising perturbation

Wang et al. [3], [4]. These papers are the foundation of MNNs, elucidating the challenges and techniques related to the incorporation of malware while preserving the functionality of the neural network model. Therefore, these papers are valuable additions to the literature for this research, as understanding the techniques that can be employed is crucial for contributing to further studies aimed at finding ways to detect or mitigate these "evil" models.

1) UNDERSTANDING MNNS

In EvilModel [3], [4], malware can be covered up within neural networks. By utilising the complex structure of

neural networks, malicious actors can seamlessly replace a substantial number of neural network parameters with malware bytes, effectively concealing the the existence of the embedded payload while maintaining the model's functionality. This covert embedding is achieved through steganography, where segments of the model's parameters are replaced with malicious code, each segment meticulously deconstructed to 3 bytes to circumvent detection. The complexity of deep learning architectures, with their multi-layered neural structures comprising millions of interconnected parameters, additionally increases the challenge of identifying these changes. As neural networks are complex and are capable of generalising well, embedded malware can evade detection by anti-virus engines and is delivered evasively. The embedding takes advantage of the natural intricacy and flexibility of neural networks, making it difficult for traditional antivirus software to detect the hidden malware and enabling sophisticated delivery methods to avoid detection. Thus, the experiments demonstrate that malware can be embedded in a neural network model without raising suspicion from anti-virus engines, proving the feasibility of the concepts behind EvilModel.

This highlights the need for security researchers to prepare ahead and develop practical solutions to mitigate this threat. These papers present the discovery of a significant threat to network security. Thus, it is crucial to understand this new threat by conducting more research and experimentation to prepare in advance and develop practical solutions.

2) EMBEDDING MALWARE INTO NEURAL NETWORK LAYERS

In EvilModel [3], [4] steganography was utilised, a method where data pieces are replaced with hidden information. Initially, the malware was broken down into smaller segments, each only 3 bytes long, to avoid detection. These segments were then strategically swapped for parts of the neural network's parameters, exploiting the extensive interconnectedness of artificial neurons in deep learning models. Despite using mainstream deep learning frameworks like PyTorch and TensorFlow, which typically use 4-byte floating-point numbers for parameter values, the experimentation in the study managed to replace 3 bytes of a parameter with malware code, keeping the model's structure intact while embedding the malicious payload. Notably, replacing neurons with malware bytes did not alter the model's structure significantly, making it hard for typical antivirus software to detect the hidden malware. Furthermore, since there is redundancy in neurons within network layers, changes in some neurons had little effect on the model's performance. This capability enabled to hide significant amounts of malware within deep learning models, with minimal loss of accuracy, particularly focusing on Convolutional Neural Network (CNN) commonly used in various applications like image classification and processing. The study also highlights the possibility of spreading infected models through online repositories and supply chain attacks, stressing the immediate need for improved security measures in the machine learning development process.

There are several methods to hide malicious code in the layers of neural networks, each aiming to compromise the functionality of the model without compromising the security of the model, as explored in EvilModel [3], [4]. An example is the Least Significant Bit Substitution (LSB), which alters the last few bits of the neural network parameters with malware codes. By altering the least significant bits, a malicious payload can be incorporated without adversely affecting the model's performance.

Another strategy is resilience training, which focuses on embedding malware by making the neural network model resilient to parametric changes. By subtly modifying the model parameters, this method conceals the malware while maintaining the model's efficiency.

Malware can also be embedded into neural network models using Value Mapping. This method involves associating the model parameters with malware bytes.

A similar approach is used in Sign Mapping to embed malware without affecting the performance of the model, as the model parameters are mapped with malware bytes.

Together, these methods are intricately designed to deliver malware in an undetectable manner without raising suspicion through the use of neural networks. A visualisation of these malware embedding methods can be seen in Fig. 2.

EvilModel [3] introduce methods such as fast substitution, batch normalisation and retraining. These methods embed malware by replacing neurons in fully connected layers while preserving the model's structure. This allows the incorporation of malware bytes without compromising the model's functionality. This sheds light on effective techniques for covertly embedding malware without impacting the model's performance.

Based on this foundation, EvilModel 2.0 [4] explores a potential threat scenario and evaluates the performance and evasiveness of the proposed embedding methods. The limitations of existing methods are discussed, and countermeasures are proposed, offering valuable guidance in the development of effective defences against embedded malware attacks.

By analysing the techniques provided in the first paper and the findings and experiments presented in the second paper, we can gain valuable insights into the vulnerabilities associated with malware embedding and use the results as a valuable resource in designing strategies to develop guidance, prevention and detection mechanisms to identify embedded malware.

Furthermore, by focusing on these insights, the exploration of detection techniques becomes a crucial step, which adds importance to advancing our understanding of malware countermeasures and contributing to the ongoing evolution of these threats.

B. DETECTION TECHNIQUES

Detecting malware embedded within neural networks poses distinctive obstacles compared to conventional techniques for

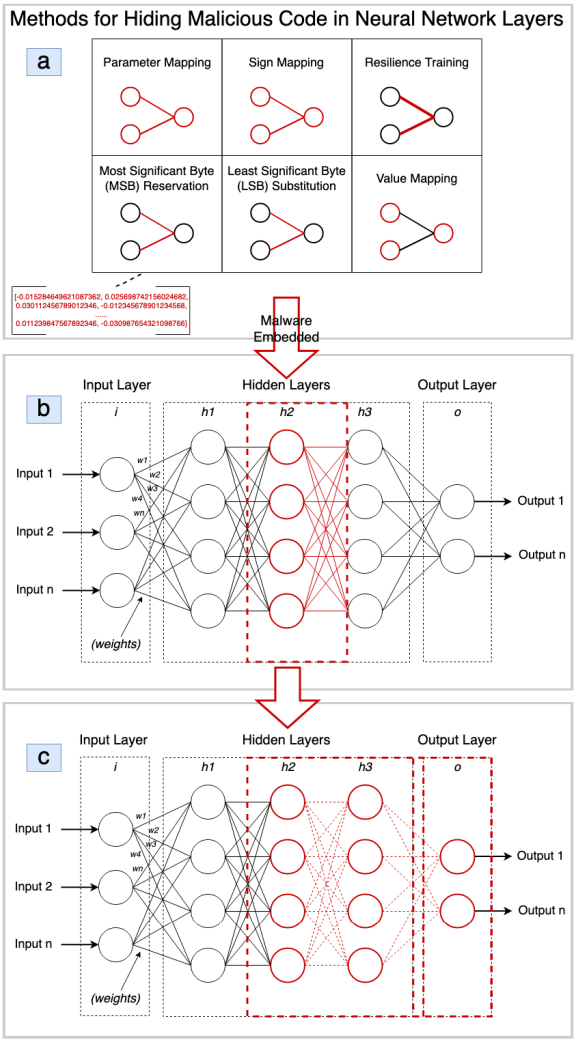

FIGURE 2. Visualisation of techniques for embedding malicious code within neural network layers. In (a), the embedding process is visually highlighted in red, highlighting where this method takes effect. (b) Illustration of the subsequent impact, with infected neurons and altered parameters in hidden layers visibly highlighted. For illustrative purposes, the entire layer is marked in red as infected, indicating that the layer has been embedded by one, multiple, or all methods combined. Finally, in (c), once a layer is infected, the malware progresses through the network, affecting connected neurons until it reaches the output layer. This progression demonstrates the flow of malicious code within the neural network. Ultimately, the evil model achieves its objective by reaching the output, impacting the system and the user, and effectively embedding the system with the malicious code through the methods indicated in a).

detecting malware in traditional software. Various techniques have been explored in the literature to identify and counteract malware embedded in neural networks. Incorporating these methods contributes to the development of a diverse set of detection methodologies, enhancing cyber security measures against Maleficent Neural Networks. The dynamic and complex nature of neural networks presents significant challenges for detecting embedded malware. Attackers can conceal malicious behaviour within the structure and parameters of the model, making it difficult to distinguish it from legitimate activity. Consequently, detecting embedded malware requires a deep understanding of the model architecture and behaviour, as well as the development of specialised detection techniques tailored to neural network environments.

Sewak et al. [5] discuss methods for detecting and classifying malware, including signature-based detection, machine learning-based detection, and deep learning-based detection. While signature-based detection relies on predefined signatures, machine learning-based detection demands extensive feature engineering, and deep learning-based detection often requires custom domain features. These methods collectively improve cyber security by identifying malicious behaviour and software. Specialised methods such as anomaly detection and behavioural analysis are relevant for identifying embedded malware within neural networks. Traditional signature-based detection methods may not adequately recognise malicious activity embedded within neural network parameters. Although traditional detection methods such as patching vulnerabilities, updating antivirus definitions, and implementing network firewalls are effective for many types of malware, they may not be suitable for detecting embedded malware within neural networks.

Simion et al. [7] explore ensemble methods and frequent retraining to improve robustness against adversarial attacks. Adversarial Machine Learning (AML) vulnerabilities are addressed through techniques such as the Fast Gradient Sign Method (FGSM) and Generative Adversarial Networks (GANs). Simion emphasises the effectiveness of evasion techniques when multiple detection models are combined, contributing to the advancement of AML.

Both embedding malware in neural network models and adversarial attacks exploit vulnerabilities in machine learning systems, particularly deep neural networks, to evade detection and cause malicious outcomes. Malware embedding and adversarial attacks aim to manipulate the model's decision-making process, leading to erroneous classifications or actions. They both aim to avoid detection by security mechanisms, achieving it by crafting inputs of malware or adversarial examples that are indistinguishable from benign data or by introducing minimal perturbations that alter the model's predictions without significantly altering the input's appearance.

Malware embedding involves inserting malicious code or payloads into neural network models with the intent of causing harm or achieving specific malicious objectives, such as compromising system integrity or stealing sensitive information. Adversarial attacks, on the other hand, are often designed to manipulate model predictions without necessarily introducing harmful payloads, with the goal of deceiving the model rather than directly causing harm. Malware embedding typically involves modifying the parameters or architecture of the neural network model itself to incorporate malicious functionality. Adversarial attacks, on the other hand, manipulate input data to generate adversarial examples that exploit vulnerabilities in the model's decision boundary.

The adversarial attack detection method contributes to addressing the similarities and differences between embedding malware in neural network models and adversarial

attacks by providing a means to identify and mitigate both types of threats. Adversarial attack detection methods can identify instances where the model's predictions have been manipulated, whether by adversarial examples or embedded malware. By recognising deviations from expected behaviour, these detection methods can flag potential security threats and trigger appropriate responses. Although adversarial attack detection methods can effectively identify manipulated model outputs, they may require different strategies to detect malware embedded within neural network models. Techniques such as anomaly detection, behavioural analysis, and model introspection may be necessary to uncover subtle indications of malicious code or functionality within the model itself. Potentially, the combination of neural network architectures such as RNNs, LSTMs, and ESNs to detect sophisticated malware embeddings with advanced adversarial defence mechanisms and reinforcement learning techniques could enhance the effectiveness of surveillance systems against MNNs.

Yang et al. [2] present a comprehensive exploration of malware detection techniques, including static detection (based on signature and heuristics) and dynamic detection (based on behaviour and pattern checking methods). The study underscores the growing importance of deep learning-based detection in future malware research. Where traditional signature-based detection methods fail to identify a new variant of malware, deep learning-based detection models can analyse the underlying structure and behaviour, identifying subtle patterns and anomalies. By continuously learning from new samples and evolving threats, these models can adapt and improve their detection capabilities, providing more robust protection against emerging malware threats.

Sewak et al. [6] explore opcode analysis for static malware analysis, achieving improved accuracy through file size-based segmentation and the application of deep learning methods such as multi-layer perceptron deep neural networks. By learning the intricate patterns within opcode sequences, these networks can effectively identify malicious behaviour and distinguish it from legitimate software. Segmentation based on file sizes ensures that neural networks can handle files of varying complexity, improving their versatility and robustness in detecting malware in different file types and sizes.

Poornima and Subramanian [8] introduce the Instruc2Vec framework for dynamic detection of malicious code within open-source software, highlighting the limitations of static approaches in the identification of emerging threats. By capturing the contextual information of opcode sequences, this framework can effectively identify previously unseen malware variants and detect suspicious activities in real time, enhancing the overall security of software systems.

Fang et al. [1] emphasise the vulnerabilities of conventional malware detection methods and propose the Deep Q-network to Evade Anti-Malware Engines Framework (DQEAF) to expose weaknesses in supervised learning-based methods, especially against sophisticated attackers. By utilising deep reinforcement learning techniques, DQEAF can systematically generate adversarial examples that evade detection by traditional anti-malware engines. Neural networks trained with DQEAF may adapt and evolve to counteract evolving evasion tactics employed by attackers, providing a more resilient defence against advanced malware threats.

Kan et al. [9] propose a malware detection system that uses a deep neural network (CNN) to analyse sequences of grouped instructions. The system automatically learns high-level representations of low-level data, achieving an impressive overall accuracy of 95%.

Venkatramulu et al. [10] highlight the evolution of malware detection using deep learning, achieving 100% precision with an artificial neural network (ANN) and a fuzzy mathematical model. By combining the power of neural networks with fuzzy logic, this approach can effectively capture the uncertainty and ambiguity inherent in malware detection tasks. The ANN learns to extract meaningful features from raw data, while the fuzzy mathematical model provides a framework for reasoning and decision-making in uncertain environments. Together, these components enable the system to achieve high precision in detecting malware, minimising false positives and false negatives, and improving overall detection accuracy.

Liu and Wang [11] propose a deep learning-based approach to malware detection using static analysis, addressing the challenges posed by traditional methods in adapting to massive data. By leveraging deep neural networks, this approach can automatically extract relevant features from malware samples and classify them into benign or malicious categories. The deep learning model learns to identify subtle patterns and anomalies indicative of malware behaviour, thus improving detection accuracy and efficiency. Furthermore, the scalability of deep learning allows this approach to handle large volumes of malware samples with ease, making it suitable for real-world deployment in malware detection systems.

Zhan et al. [12] address the vulnerability of deep neural networks to adversarial examples in malware detection, presenting AMGmal, that is, adaptive mask-guided adversarial attack against malware detection with minimal perturbation, a novel strategy that uses saliency detection and mask guidance to prioritise critical bytes in slack areas. Experimental results demonstrate AMGmal's effectiveness in evading detection and reducing perturbations.

In conclusion, the examination of malware detection techniques not only highlights key insights for fortifying cyber security against Maleficent Neural Networks but also provides a nuanced understanding of the challenges and opportunities associated with enhancing cyber resilience against MNNs. Signature-based detection, as highlighted in some studies, is a fundamental tool that shows effectiveness but shows limitations in the face of newly emerging malware, suggesting the need for continuous evolution and improvement in detection capabilities.

Deep learning methods, extensively explored in various research studies, such as those conducted by Liu and Wang [11] and Sewak et al. [5], contribute significantly to cyber security. Their ability to learn from extensive datasets and identify complex patterns aligns well with the requirements to detect sophisticated MNNs. Although these methods demonstrate a balance between advanced capabilities and practical usability, they are also complex and require custom features.

Based on the exploration of ensemble methods and the use of GANs carried out by Simion et al. [7], malware detection systems can be made more robust. In the context of MNNs, where attackers constantly evolve their strategies to evade detection, this approach is especially applicable.

A shift towards adaptive and responsive malware detection strategies, essential for identifying and countering MNNs effectively, is also highlighted by the use of op-code analysis and dynamic detection methods.

It is obvious from the revelation of Fang et al. [1], that supervised learning-based malware detection methods are vulnerable. Therefore, continued vigilance and innovation are critical to malware detection. Innovations like EvilModels [3], [4] not only assess evasiveness through antivirus engine detection and stegoanalysis but also challenge the current state of cyber security by demonstrating how malware can be embedded within neural networks in a way that evades traditional detection methods. The evolution of deep learning, as demonstrated by EvilModels, shows the efficiency of automatically learning from data and overcoming the challenges posed by traditional methods. This underscores the urgency for the cyber security community to stay ahead of such techniques through continuous research and development of more sophisticated detection tools. In general, a diverse set of methodologies remains crucial for comprehensive and adaptive malware detection.

As a result of the studies presented on malware detection techniques in the context of MNNs, it is apparent that a comprehensive and adaptive defence system must be based on a variety of methodologies, each contributing to its strengths. Our strategies for detecting and mitigating MNNs must evolve at the same pace as they do, necessitating a combination of traditional methods and cutting-edge advances in artificial intelligence and machine learning.

C. MITIGATION STRATEGIES

Information systems must be secure to combat evolving cyber threats, particularly malware that is becoming more sophisticated. To counter the challenges associated with malware detection, researchers have proposed several mitigation strategies. In the following sections, we explore the variety of approaches in the literature, from deep learning architectures to new frameworks and countermeasures.

Mitigating embedded malware within neural networks presents significant challenges. Addressing these challenges requires a deep understanding of the model architecture and behaviour, as well as the development of specialised techniques to protect against adversarial attacks and manipulation of model inputs. While neural model compression can play a role in reducing the complexity and size of neural networks, additional defence mechanisms are often necessary to ensure robust security against sophisticated threats. In contrast, mitigating malware in ordinary software involves actions such as patching vulnerabilities, updating software, implementing access controls, and deploying network firewalls. Traditional mitigation strategies focus on addressing known vulnerabilities and implementing security measures to prevent unauthorised access and malicious activity.

It is important to also take into account the changing nature of malware, which now includes MNNs. These evil models present specific challenges as they can mimic legitimate behaviour and avoid traditional detection methods. As a result, new strategies need to be created to effectively address this new threat.

Sewak et al. [5] advocate for the efficacy of Deep Learning architectures, specifically Auto-Encoders (AEs) for Feature Extraction and Deep Neural Networks for malware classification. These architectures eliminate the need for custom feature engineering, offering scalability and general defence against malware by automatically extracting higher conceptual features from the data.

Yang et al. [2] highlight the challenges posed by rapid developments in anti-malware technologies, highlighting the increasing use of encryption and obfuscation by malware authors. Static detection methods face difficulties in signature matching and code analysis, which require adaptive strategies.

Sewak et al. [6] stress the importance of feature engineering, class imbalance handling, and various long-short-term memory (LSTM) network configurations in mitigation strategies. Precautions during model porting and the transfer of models for malware detection are also crucial to enhance the overall resilience of detection systems.

Poornima and Subramanian [8] suggest strategies for resource conservation and accuracy maintenance, focusing on feature selection and hybrid deep learning frameworks for optimising malware file detection.

Fang et al. [1] introduce the DQEAF framework, which uses reinforcement learning to evade traditional detection engines while maintaining the structure of the malware. The framework exhibits robustness against various families of malicious software.

Wang et al. [4] propose countermeasures against EvilModels, focusing on adjustment of parameters, modification of neural network models, and safeguarding the model supply chain to improve overall security.

Kan et al. [9] offer a lightweight CNN based on deep learning principles, utilising disassembled instructions as raw data for efficient feature extraction in the identification of polymorphic and zero-day malware.

Venkatramulu et al. [10] propose a multilayered artificial neural network for mitigation, incorporating data from

various sources to gain a comprehensive understanding of the characteristics of malware.

Liu and Wang [11] combine deep learning with feature extraction, using grayscale images and OpCode 3grams to provide a multifaceted view of malware characteristics, presenting a more automated and adaptable alternative to traditional methods.

Zhan et al. [12] introduce AMGmal, a mitigation strategy focusing on adversarial attacks on DNN-based detectors. The strategy involves fooling visualisation-based detectors, reserving functionality, and minimising perturbation, contributing to enhanced robustness against adversarial attacks.

The studies reviewed in this section underscore the dynamic and ever-evolving nature of cyber security threats, especially in the context of MNNs. Together, these papers enhance the progress in detecting and mitigating malware by presenting various methods such as deep learning models, flexible frameworks, feature engineering techniques, and defences against evasion tactics and adversarial attacks. Each study focuses on different obstacles in malware detection, ultimately improving the efficiency and durability of information systems against complex cyber threats. It is clear that these threats require not only awareness but also a continuous evolution of mitigation strategies.

In developing effective mitigation strategies for malware detection, researchers emphasise the critical importance of adaptability and comprehensive understanding, leveraging deep learning architectures such as AEs and DNNs for automated feature extraction. Novel frameworks and countermeasures against adversarial attacks, as discussed in the diverse approaches highlighted here, contribute significantly to ongoing efforts to fortify information systems against the growing challenges posed by sophisticated malware techniques.

As part of our research, we have found that adversarial attacks require new frameworks and countermeasures. Using AMGmal by Zhan et al. [12], a mitigation strategy aimed at adversarial attacks against DNN-based detectors, is an example of these forward-looking approaches required in today's cyber environment.

The following studies are notable for their diverse approaches. Each contributes uniquely to the fortification of information systems. Kan et al. [9] describe lightweight CNNs for identifying zero-day and polymorphic malware, while Venkatramulu et al. [10] describe a multilayered artificial neural network approach to malware classification. Cyber defence has many aspects, as demonstrated by these diverse methodologies.

Novel strategies as described above will be important as we continue to deal with the challenges posed by sophisticated malware such as MNNs. A thorough understanding of the malware landscape is also part of this, in addition to researching and developing advanced methods. Therefore, future research should focus on improving the adaptability, scalability, and robustness of mitigation strategies to ensure that they remain effective against this evolving new threat posed by MNN.

III. DISCUSSION AND RESULTS

A. DISCUSSION

In our exploration of malware detection and mitigation methodologies as seen in Subsections II-B and II-C, a clear trend emerges of the increasing adoption of sophisticated computational techniques, particularly advanced deep learning approaches. This trend reflects a collective effort to strengthen cyber security measures, addressing challenges such as the growing reliance on deep learning, the importance of adaptability in detection strategies, and the delicate balance between complexity and practicality in malware detection.

The literature presented in Section II reveals diverse approaches to malware detection. Signature-based methods, while effective against known threats, are vulnerable to emerging malware [2], [5]. Deep learning techniques, including AEs, DNN, CNN, and ANN [5], [9], [10], [12], show their effectiveness in automatically extracting higher conceptual features without extensive engineering. To strengthen against deliberate attacks, researchers advocate ensemble methods and frequent retraining in adversarial machine learning [7]. Adapting to the ever-evolving cyber security landscape requires a range of mitigation strategies, from feature engineering to handling class imbalances to cautious model data transfer.

Concerns arise about neural networks being misused to embed malware, prompting discussions on countermeasures and proactive defences. A holistic approach integrates detection and mitigation, focusing on efficiency, scalability, evaluation criteria, and prudent model data transfer. To effectively address evolving challenges in information security, this discussion underscores the importance of holistic and dynamic malware detection strategies.

Various related works emphasise adaptability in detection strategies. Detection systems must be able to handle new and unknown threats effectively, necessitating adaptive learning, real-time detection, and techniques to combat evolving malware threats. Malware creators continuously evolve tactics, underscoring the need for adaptive and proactive defence mechanisms.

The challenges in finding the right balance between complexity and practicability in malware detection have been discussed. Deep learning methods, while powerful, require balance to avoid overfitting. Practical implementation without compromising effectiveness is crucial, highlighting the efficiency and scalability of detection systems. Precautions and caution in model porting, as well as considering model applicability to malware detection challenges, emphasise practical implementation.

The evolution of malware detection techniques from signature-based to advanced approaches, such as deep learning and ensemble detection, is evident in the literature review. Traditional signature-based methods struggle against

new or unknown malware, serving as a starting point for this evolution. This highlights the need for adaptive, sophisticated, and interconnected strategies to combat new threats. Moreover, there is an increasing need for flexible techniques that can respond quickly without requiring a large amount of training data. This requirement highlights the significance of creating adaptable and prompt strategies that can effectively handle new challenges.

Sewak et al. [5] describe the progression of detection methods, discussing signature-based, machine learning-based, and deep learning-based detection. While signature-based methods are effective, machine learning and deep learning offer more complex capabilities. The increasing complexity of detection techniques is evident in the focus on feature engineering in machine learning and custom domain features in deep learning.

Simion et al. [7] propose ensemble methods and frequent retraining for robustness against adversarial attacks; this emphasises the need for adaptive and resilient techniques against evolving MNNs.

Deep learning architectures have significant implications for combating MNNs, as the defence can be embedded within the design of the architectures themselves. By harnessing the ability to learn representations directly from data, these architectures offer scalable and general defence mechanisms. Their scalability facilitates efficient processing of large datasets and complex models, while transfer learning capabilities enable adaptation to new defence scenarios. Additionally, ensemble methods and adversarial training techniques can be integrated seamlessly into these architectures to enhance robustness and resilience against evolving threats.

The literature review provides insight into the challenges faced by MNNs, highlighting deep learning architectures as effective defences due to their inherent adaptability and robustness. Proactive detection techniques, such as ensemble methods and frequent retraining, maintain trust and reliability despite adversarial attacks. Additionally, ethical concerns are addressed, and countermeasures against MNNs highlight the importance of adjusting parameters, modifying models, and enhancing detection and mitigation strategies.

In conclusion, the evolving challenges posed by MNNs emphasise the importance and need for advanced and adaptive defence methods that are embedded within the core of machine learning frameworks. Maintaining trust, addressing ethical considerations, and preventing potential consequences in machine learning systems underscore the necessity for continuous research and innovation to stay ahead of dynamic tactics used by malicious actors.

B. CHALLENGES IN IDENTIFYING MNNS

Identification of MNNs presents a critical challenge due to the lack of effective detection methods, tools, or techniques. This absence significantly jeopardises the security of systems and users, allowing malicious models to persist undetected for extended periods.

The risk escalates with the widespread creation and reuse of machine learning models, enabling the spread of MNNs across diverse domains. This amplifies the challenge, emphasising the persistent threat posed by undetected malicious models throughout the entire life cycle of model development and deployment.

The impact of MNNs extends beyond system vulnerabilities, it directly affects users who rely on machine learning systems for critical decision-making processes. The existence of MNNs undermines trust and reliability, potentially resulting in significant consequences across various domains, as malicious models can yield erroneous outcomes.

Moreover, the deliberate embedding of malware in neural networks raises ethical concerns, questioning the responsibility of technology creators and the potential misuse of AI technology. This not only poses challenges to the systems but also has a direct impact on user privacy, as MNNs can gain unauthorised access to sensitive information without detection.

The consequences of MNNs include financial losses for individuals and businesses, compromised system security, and privacy violations leading to identity theft. Despite research that demonstrates the possibility of embedding malware without affecting performance, the lack of effective detection mechanisms hinders the ability to prevent the potential harm caused by MNNs.

In the context of model evaluation, various metrics, such as TP, F1 score, FPR, FP, TN, FN, accuracy, and recall, are commonly used to evaluate the effectiveness of detection mechanisms. These metrics provide insight into false positive and false negative rates, accuracy, and overall balance, ensuring a comprehensive assessment of the mechanisms used to combat the threat of MNNs.

Detecting Maleficent Neural Networks poses significant challenges due to their sophisticated embedding techniques and evasion strategies. The current state-of-the-art in detecting MNNs involves a combination of traditional malware detection methods adapted to the unique characteristics of neural networks, as well as novel approaches tailored specifically to MNN detection.

One of the primary challenges in detecting MNNs lies in their covert embedding within neural network parameters, often achieved through steganography. MNNs exploit the intricate structure of neural networks, replacing segments of parameters with malicious code while preserving the model's functionality. This makes it difficult for traditional antivirus software to detect the hidden malware, as the alterations do not significantly impact the model's performance. Furthermore, the sheer complexity of deep learning architectures, with millions of interconnected parameters, and the trend towards even more complex models, amplifies the challenge of identifying these subtle changes. Furthermore, the increase in harmful models on websites like Hugging Face, where over 100 malicious AI ML models have been discovered, some of which have the ability to run code on the user's device, presents additional risks to cyber security [13]. Even though

Hugging Face has measures in place like malware detection, pickle, and secrets scanning, the existence of these harmful models emphasises the importance of being more alert and taking preventive actions to protect against such threats.

Furthermore, MNNs leverage the resilience and generalisation capabilities of neural networks to evade detection by anti-virus engines. Their ability to evade traditional detection methods and employ sophisticated delivery mechanisms underscores the need for advanced detection techniques tailored to the unique characteristics of MNNs.

Several possible approaches have been discussed to address the challenges of detecting MNNs. Behavioural Analysis involves monitoring the runtime behaviour of neural networks instead of solely relying on static analysis of model parameters. By observing deviations from normal network behaviour during inference, this technique can identify anomalies indicative of malicious activity, thereby enabling the detection of MNNs.

Adversarial Robustness techniques are aimed at enhancing the resilience of neural networks against adversarial attacks, including MNN embedding. Through adversarial training and robustness techniques, models are trained to withstand subtle parameter alterations, thus improving the detection of malicious embeddings.

Deep Learning-based Detection utilises deep learning models specifically designed for MNN detection. Researchers have developed neural network architectures capable of identifying patterns indicative of malicious embeddings. These models are trained to distinguish between benign and malicious neural network parameters, thereby enhancing detection accuracy.

Anomaly Detection methods aim to identify unusual patterns or behaviours within neural networks that may indicate the presence of embedded malware. By analysing deviations from expected norms, anomaly detection approaches can detect suspicious neural network behaviour for further investigation, helping to detect MNNs.

Despite these advancements, detecting MNNs remains a formidable challenge due to the dynamic nature of malware and the evolving sophistication of embedding techniques. Furthermore, the lack of labelled datasets containing MNN samples hinders the development and evaluation of detection methods. Addressing these challenges requires ongoing research efforts to advance detection techniques and enhance the resilience of neural networks against malicious embeddings.

C. ETHICAL AND LEGAL CONSIDERATIONS

It is important to consider various ethical and legal considerations when working with MNNs after exploring ways to combat this novel threat. A deliberate attempt to embed malware in neural networks presents several technical challenges, as well as profound ethical and legal implications.

The deliberate incorporation of malware into neural networks raises profound ethical concerns and presents significant challenges at the intersection of legal, social, and professional issues. Addressing these multifaceted concerns requires a comprehensive examination of ethical frameworks and guidelines governing the responsible use of AI technology. Additionally, it underscores the importance of transparency and accountability throughout all stages of MNN development and deployment.

MNN models pose a substantial threat, which could lead to cyber-attacks, financial fraud, and social engineering, among other malicious activities. Responsible research involving MNNs requires a rigorous ethical and legal approach to evaluating risks and benefits. This includes implementing measures to prevent or mitigate abuse and misuse by malicious actors, while also prioritising the protection of user privacy, security, and autonomy. Although open-source research could enhance transparency and accountability, certain aspects of research must remain confidential to prevent misuse. Adherence to ethical principles should not be compromised, even if it entails loss of business or faces opposition from the company, ensuring that the integrity of research and its potential societal impact are upheld.

A real-world challenge lies in the potential for unintended consequences and the difficulty in regulating and controlling MNNs. In addition, attackers can employ techniques such as AML to avoid detection and bypass defences. Assessing the risks and benefits of MNNs accurately and developing appropriate governance frameworks to mitigate risks and foster responsible innovation is inherently challenging, similar to any emerging technology. Regulating MNNs should be flexible and adaptable, taking into account the quickly changing technology and evolving threats.

Additionally, the following additional factors are important to take into account to enhance the consideration regarding ethical and legal consequences:

1) It is important to explore the following elements in order to enhance the understanding of ethical and legal implications. Researching the possibility of bias in MNNs and the ethical responsibility to ensure fairness in their development and use are important tasks. This involves addressing biases that could contribute to discrimination or worsen social inequalities. It is crucial to incorporate fairness-oriented approaches in the design and utilisation of MNNs to support fair outcomes.
2) Participating in conversations regarding responsibility and culpability when MNNs cause harm or generate incorrect outcomes, as well as examining methods for holding parties accountable for MNN results. Setting distinct boundaries of responsibility and liability is crucial for addressing the ethical and legal consequences of MNNs and guaranteeing appropriate measures are taken when harm occurs.
3) Investigating the involvement of regulators in supervising the growth and implementation of MNNs, and examining the necessity for revised regulations to tackle the specific difficulties brought by MNNs and the moral factors in regulating AI technologies. Strong regulatory

structures are crucial to guarantee compliance with ethical principles and reduce the risks linked with MNN deployment.

4) Highlighting the ethical importance of safeguarding data privacy and ensuring the security of information managed by MNNs, while acknowledging the risks of data breaches and unauthorised access to sensitive data. Putting priority on protecting user data and privacy is crucial to upholding ethical standards and building trust in MNN technologies.

The importance of tackling these ethical and legal concerns in order to promote responsible innovation and reduce potential risks linked to the creation and use of MNNs is clear. By taking a proactive approach to address these issues, researchers, practitioners, and policymakers can ensure that MNNs are created and used in a way that values ethical principles, upholds individual rights, and enhances societal welfare.

D. POSSIBLE RESEARCH DIRECTIONS

To effectively combat Maleficent Neural Networks, a comprehensive strategy can be developed integrating multiple perspectives to mitigate malware risks embedded in neural networks. By addressing limitations and embracing future directions, a robust detection mechanism against these malicious entities can be achieved, making a significant contribution to the field.

Based on the analysed literature, six key future directions have been identified in the context of MNNs.

(I) A comprehensive strategy to combat MNNs effectively involves a holistic approach that incorporates diverse perspectives. Exploring metamorphic malware detection techniques and utilising advances in deep learning is key. Identification of complex malware embeddings requires the combination of different neural network architectures, such as RNNs, LSTMs, and ESNs. Researchers can enhance the effectiveness of combating MNNs by improving the detection system against sophisticated malware embeddings through the integration of these neural network architectures.

(II) As part of future research that extends the current literature, it is imperative to refine saliency detection techniques, placing a specific emphasis on precision, particularly in the domains of intricate malware datasets. These techniques play a crucial role in identifying sophisticated malware threats embedded in neural networks. This effort is dedicated to advancing AML and GANs, thus facilitating the development of robust cyber defences against progressively sophisticated threats. To increase the efficacy of saliency detection, exploring the integration of explainable AI holds promise. This not only contributes to improved precision but also fosters transparency and trust in detection mechanisms. The resolution of challenges such as imbalanced datasets is vital for the ongoing progress of research in this field. By providing insights into how decisions are made, explainable AI ensures that cyber security professionals and end-users can understand and trust the outcomes of AI-powered security measures. Further, addressing challenges such as imbalanced datasets is crucial for advancing research in cyber security, as it improves the reliability and effectiveness of AI-driven threat detection and mitigation strategies, thereby bolstering defences against evolving cyber threats [14].

(III) The development of adaptive malware detection systems could involve the integration of reinforcement learning and transfer learning techniques, as suggested by Yang et al. [2]. By incorporating Explainable AI, transparency, trustworthiness, and effective understanding of cyber security decisions can be improved.

(IV) Collaboration and real-world applications, as suggested by Sewak et al. [6] and others. Future research should also focus on refining detection models based on real-world data, such as PCAP files, and validating these models in different scenarios. To develop effective strategies against MNNs, cyber security professionals, AI experts, and machine learning specialists must collaborate.

(V) Researchers must extend their research to a variety of platforms, including ELF and Android, as suggested by Fang et al. [1], as well as explore new embedding methods and defence strategies, as highlighted by Wang et al. [4]. Feature extraction techniques must also be improved and malware detection methods must be diversified to keep up with the rapid evolution of cyber threats.

(VI) Addressing specific challenges and expanding research domains, as Kan et al. [9] have emphasised. The emphasis on deepening neural networks and Zhan et al. [12] focus on visual adversarial attacks demonstrate the need to expand research domains and address specific malware detection challenges. Moreover, as MNNs continue to evolve, researchers could investigate and develop advanced adversarial defence mechanisms, aligning these suggested future directions with established guidelines and recommendations to create an integrated framework.

Taking into account the dynamic nature of malware evolution, future efforts should focus on improving the efficiency, scalability, and accuracy of detection methods.

As we can see from the current research, several promising future directions are possible. Furthermore, as MNNs continue to evolve, researchers need to investigate and develop advanced adversarial defence mechanisms, aligning these future suggested directions with established guidelines and recommendations to create an integrated framework.

In addition, the incorporation of reinforcement learning and transfer learning techniques could enhance the real-time adaptation of malware detection mechanisms. To ensure

robustness in various scenarios and practical applicability, these techniques should be evaluated for their effectiveness against adversarial evasion attacks.

To identify potential detection methods, researchers need to experiment to gain insight into how malware embeds itself in neural networks. To ensure responsible research practices, it is important to consider legal and ethical issues during experimentation.

To counter MNNs, cyber security professionals, machine learning experts, and policymakers must collaborate to formulate ethical guidelines.

To ensure the effectiveness of detection mechanisms, researchers and industry professionals should collaborate to assess research findings in real-world scenarios. For effective malware detection within neural networks, experimentation and an understanding of the embedding process are important.

Finally, to ensure responsible and impactful research, legal and ethical considerations must be implemented and considered throughout the entire experimentation process, from data to the final application and impact.

IV. CONCLUSION

The goal of this work was to analyse harmful characteristics and immediate challenges posed by MNNs. This was done through a literature survey and an analysis of related work. We conducted a comprehensive literature review and analysis of the state-of-the-art. This includes the analysis of existing approaches to embed malicious code into neural networks and possible countermeasures. We also provide and propose future research directions and guidelines. The analysis showed that there is little awareness about the topic in general and that there is a lack of reliable detection methods. The lack of general understanding and counter methods is highlighting the need for cooperation among researchers and practitioners in AI and cyber security to address connected risks.

MNNs are complex threats that rapidly expand and evolve, and more research must be done. The analysis presented in this work can be a starting point for this. Additionally, our research highlights the importance of continuous collaboration and investigation in creating adaptable and efficient defence strategies against the changing danger posed by MNNs which will also help to safeguard the security and reliability of AI and machine learning systems from malicious actions.

In future work, we plan to investigate potential defence methods and extend the survey to the real world. This will help us understand the distribution of potentially infected neural networks and the extent of harm they may cause to AI ecosystem

REFERENCES

[1] Z. Fang, J. Wang, B. Li, S. Wu, Y. Zhou, and H. Huang, "Evading anti-malware engines with deep reinforcement learning," *IEEE Access*, vol. 7, pp. 48867–48879, 2019.

[2] S. Yang, S. Li, W. Chen, and Y. Liu, "A real-time and adaptive-learning malware detection method based on API-pair graph," *IEEE Access*, vol. 8, pp. 208120–208135, 2020.

[3] Z. Wang, C. Liu, and X. Cui, "EvilModel: Hiding malware inside of neural network models," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Sep. 2021, pp. 1–7.

[4] Z. Wang, C. Liu, X. Cui, J. Yin, and X. Wang, "EvilModel 2.0: Bringing neural network models into malware attacks," *Comput. Secur.*, vol. 120, Sep. 2022, Art. no. 102807.

[5] M. Sewak, S. K. Sahay, and H. Rathore, "An investigation of a deep learning based malware detection system," in *Proc. 13th Int. Conf. Availability, Rel. Secur.*, Aug. 2018, pp. 1–5.

[6] M. Sewak, S. K. Sahay, and H. Rathore, "Assessment of the relative importance of different hyper-parameters of LSTM for an IDS," in *Proc. IEEE REGION 10 Conf. (TENCON)*, Nov. 2020, pp. 414–419.

[7] C. Simion, G. Balan, and D. T. Gavrilut, "Improving detection of malicious samples by using state-of-the-art adversarial machine learning algorithms," in *Proc. 15th Int. Conf. Secur. Inf. Netw. (SIN)*, Nov. 2022, pp. 1–8.

[8] S. Poornima and T. Subramanian, "Effective feature extraction via N-skip Gram instruction embedding model using deep neural network for designing anti-malware application," in *Proc. 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2023, pp. 2118–2123.

[9] Z. Kan, H. Wang, G. Xu, Y. Guo, and X. Chen, "Towards light-weight deep learning based malware detection," in *Proc. IEEE 42nd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, vol. 1, Jul. 2018, pp. 600–609.

[10] S. Venkatramulu, M. S. B. Phridviraj, S. Pratapagiri, S. Madugula, S. Kiran, and V. C. S. Rao, "Usage patterns and implementation of machine learning for malware detection and predictive evaluation," in *Proc. 2nd Int. Conf. Artif. Intell. Smart Energy (ICAIS)*, Feb. 2022, pp. 244–247.

[11] L. Liu andB. Wang, "Automatic malware detection using deep learning based on static analysis," in *Data Science*, B. Zou, M. Li, H. Wang, X. Song, W. Xie, and Z. Lu, Eds. Singapore: Springer, 2017, pp. 500–507.

[12] D. Zhan, Y. Duan, Y. Hu, L. Yin, Z. Pan, and S. Guo, "AMGmal: Adaptive mask-guided adversarial attack against malware detection with minimal perturbation," *Comput. Secur.*, vol. 127, Apr. 2023, Art. no. 103103.

[13] *Data Scientists Targeted by Malicious Hugging Face ML Models With Silent Backdoor*. Accessed: May 15, 2024. [Online]. Available: https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/

[14] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. D. Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805.

STEPHANIE ZUBICUETA PORTALES received the Bachelor of Science degree (Hons.) in computing and IT from The Open University, in 2023, and the Bachelor of Science degree in information science from the University of Bergen, in 2024. She is currently involved in various research work, one of them being on maleficent neural networks. Her research interests include artificial intelligence and cyber security.

MICHAEL ALEXANDER RIEGLER received the Ph.D. degree from the University of Oslo, Oslo, Norway, in 2017. He is currently a Research Scientist with Simulamet, Oslo. His research interests include machine learning, video analysis and understanding, image processing, image retrieval, social computing, and applications of artificial intelligence in medicine.