Malware detection through Machine Learning, Deep Learning and Artificial Intelligence

Siddhant Kumar
Department of Computer Science
Central University of Himachal Pradesh
kumarsiddhant077@gmail.com

Mayank Chopra
Department of Computer Science
Central University of Himachal Pradesh
mayankchopra.it@gmail.com

Shubham
Department of Computer Science
Central University of Himachal Pradesh
sd9440555@gmail.com

Pradeep Chouksey
Department of Computer Science
Central University of Himachal Pradesh
dr.pradeepchouksey2@gmail.com

Shubham Sharma
Department of Computer Science
Central University of Himachal Pradesh
shubhamgharoh@gmail.com

Parveen Sadotra
Department of Computer Science
Central University of Himachal Pradesh
sadotramca2k6@gmail.com

Abstract: *To achieve accuracy in finding these malicious softwares different methods has been used such as through Machine Learning (ML) which works on supervised learning, Deep Learning (DL) which works on unsupervised learning, and Artificial Intelligence (AI)which works reinforcement learning on the basis of these techniques framework of malware detection on Operating system, malicious softwares includes viruses, worms, trojans, ransomwares, spyware, rootkits and many more softwares which interrupt's other softwares, steal information, creates blockage on processor or try to harm any user are considered as a malware. These variants of malwares have different types of objectives, different methods of breaching and for further growth. To resolve a variants of problems variants of solutions are required. On the way to solutions initially it is retaliated through ML models which works on static analysis with static feature selection which is constant and further it evolved to DL which works on dynamic analysis with no constant features defined at early stage and AI works on enhanced signature based techniques, ensemble learning, feature engineering and selection and many more. At any time whenever analysis is required there is a requirement of benchmark on the basis of which a difference can be notified. For this malware's history is considered to prepare a dataset on malware's features, availability, methods of intrusions and many more to train different models for prevention from these malwares in future.*

Keywords: *Malware, Machine Learning (ML), Deep Learning (Dl), Static analysis, Dynamic analysis, Artificial Intelligence (AI).*

Introduction

To protect user privacy and integrity in today's society, cyber security is essential. Machine learning (ML) has been crucial in the fight against cyberattacks. Although it has a poor edge in malware detection because it concentrated during compilation on pre-defined specialized traits. Deep Learning (DL) was developed to train original data without removing its features in order to overcome this drawback. Artificial intelligence significantly enhances detection and prevention capabilities. Cybersecurity experts

regularly refine AI models combining human expertise with machine intelligence to their algorithms. In this paper we discuss some techniques using artificial intelligence to detect and prevent the malware in a system or network. AI plays a significant role in malware detection and prevention. As new attack methods keep being introduced all the time.

Literature Review

1. Machine Learning Techniques :-

1.1. Through Static, Dynamic and Image classification :-

Static analysis

In the realm of static analysis for malware detection, earlier methods involved using various features such as n-grams, which consider multi-byte identifiers and strings. For classification, a more advanced approach known as Sequence Convolutional Neural Network (CNN) is used, which extracts features from the binary structure of data. “The Ember dataset, which comprises 1.1 million binary files (900,000 for training and 200,000 for testing), is used to train machine learning models for detecting Portable Executable (PE) files” (1). This dataset is valuable not just as a benchmark but also for end-to-end deep learning algorithms.

A interesting solution involves using deep neural networks to calculate a byte/entropy histogram for virus recognition. The technique aids in the extraction of useful information from binary bytes.

Dynamic analysis

In the realm of malware detection and prevention, various innovative approaches utilizing machine learning have been proposed. One strategy involves training a Recurrent Neural Network (RNN) to extract process behaviour features, followed by training a Convolutional Neural Network (CNN) for image categorization. Additionally, a natural language modelling-based on deep learning architecture for shallow multitasking is introduced for malware detection, where combining convolutional and recurrent layers enhances system call modelling and classification sequences, outperforming traditional models like SVM and Hidden Markov Models. Another method employs a malware prediction model based on RNN that anticipates malicious files before payload execution, achieving a remarkable 94% accuracy within a short execution time. The significance of dynamic analysis is underscored, with a focus on training Hidden Markov Models using both static and dynamic analyses, revealing the superiority of the dynamic approach, particularly in identifying malware families.

Deep Learning methods

A Framework using Deep Learning:
Due to the failure of IDSs against contemporary or zero-day attacks, a new distributed ledger-based framework has been developed. This framework extracts samples from datasets, pre-processes them, and then does additional analysis.

It is necessary to select the dataset, model, algorithm, and comparison preference appropriately. The processed data set will be used to train the deep learning model. In fog computing, pre-processed data is transmitted to the fog node, where malware analysis is carried out and updated to the cloud layer.. Better security for new device connections over the cloud is ensured by doing this.

Data collection and processing are carried out locally before being sent to a cloud environment.

Table 1 :- "FFNN and LSTM DL models to detect cyber-attacks in IOT networks"(2)

Purpose	DL model	Dataset	Hidden layers	Units	Dropouts (%)	Epochs
Detection	FFNN	BoT-IoT NSL-KDD	3 4	100/50/100 48/128/88/188	- 25/45/25/35	15 10
Detection	LSTM	BoT-IoT NSL-KDD	3 5	100/64/32 50/100/200/100/50	- -	30 20

Table 1 shows information about detection methods and findings with FFNN and LSSTM models on different datasets. The models are trained and tested using a centralized DL model. In broader networks, it has a stronger advantage in intrusion detection. "With both datasets, only FFNN is used in a distributed framework" (2).

1.1. Hybrid Deep Learning Model :-

Since the analysis is carried out in a coordinated statically and dynamically manner, it follows the progression from machine learning (ML) to deep learning algorithms (DL). Because dynamic analysis tracks overall behavior, it can address issues where static analysis falls short, such as polymorphic deformation and code obfuscation. Three methods were first proposed in the dynamic analysis criteria: TaintDroid (3), which works well with sensitive data, real-time analysis with fewer resources, and so on. The other is ntLeakSemaic (4), which deals with erroneous network transmission. Third, DroidScraper (5) is employed in postmortem and forensic analysis as well as runtime data structure recovery. Natural language processing, speech recognition, picture recognition, and nonlinear relationship recognition are areas where deep learning excels.

Traditional algorithms are being replaced by Deep Belief Networks (DBN) and Gate Recurrent Units (GRU). This is because DBN has a faster learning rate for static data and better performance features, while GRU handles longer-running operations, making their combination effective. Contrastive divergence or CD is used to train DBN in order to improve optimization. Batch size, epochs, and learning rate are used when pre-training parameters are being set. We need to identify the point with the least amount of loss based on these parameters, so values are taken ranging from small to large to identify the analytical loss point at which it begins to increase following a gradual decrease. Large batch sizes and epochs can reduce the model's performance, so they must be selected carefully.

It is not a good idea to include a second layer in the network layer because it will contain a large number of samples that are inappropriate for this model. The initialization of network training parameters must not be too large to prevent saturation of the activation value, nor can it be too small to prevent the network from disappearing over multiple layers. The Gaussian distribution is the answer for this. Algorithms are used to optimize the model, which is necessary. Adam stands for Adaptive Moment

Estimation, a hybrid of the Momentum, Adagrad (6) and RMSProp (6) algorithms.

1. Malware detection techniques through Artificial Intelligence:-

The evolution of malware detection systems will involve the processing of malicious files and the execution of analysis to comprehend its properties through the integration of AI technologies. Twenty major features will be identified by applying techniques such as “Fisher Score (FS), Chi-Square (CS), Information Gain (IG), Gain Ratio (GR), and Uncertainty Symmetric (US)” (7). The system analyses several classes in order to train the classifier.

“Malware detection techniques using artificial intelligence is classified into three types which includes signature-based detection techniques,anomaly-based detection techniques, heuristic-based detection techniques” (7), “specification-based detection techniques” (8).

1. Signature-based detection technique:

The term signature-based detection method, which helps identify and detect attacks by searching for particular patterns, is made up of four parts, “using a signature-based technique, programmers scan a file and compare information with the database”, and uses the signatures of viruses to identify malware inside the database. If the information match with the data in the database, it indicates that there are viruses in the file. This technique works well for identifying known malware, it is not as effective at identifying unknown malware. Supervised machine learning algorithms are trained on malware behaviours in order to detect malware based on signatures. This procedure helps identify and classify unknown actions as either malicious or harmless (7).

2. Anomaly-based detection techniques:

The use of anomaly-based detection techniques is important for handling security issues and safeguarding against different types of harmful attacks. The limitation of signature-based detection methods is addressed by anomaly-based detection methods, which use classification techniques to identify known and unknown malware. “Advantage of this change in detection methods from pattern-based to classification-based increases the ability to detect known and unknown malware” (7).

3. Heuristic-based detection techniques

Malware detection efficiency is increased through artificial intelligence with the help ofsignature-based and anomaly-based detection systems. To enhance the categorization process, a neural network and a machine learning technique known as genetic algorithm along with neural network were introduced to improve the classification method. “Without any prior system information, this method uses features like inheritance, selection, and combination that provide the advantage of obtaining optimal solutions from various sources” (7).

4. Specification-based detection techniques

Cloud-based malware detection approaches use cloud server mode to detect malware, whereas specification-based malware detection techniques use specific criteria to identify malware. These methods try to identify any irregularities, errors, or unexpected system behaviour. These methods are crucial for maintaining the accuracy and

safety of the system. Early error, vulnerability, or malicious activity detection can be improved by specification-detection approaches (8).

Conclusion

The contributions of Deep Learning (DL) and Machine Learning (ML) models to malware detection have been compiled. In this analysis we have addressed key features like static and dynamic analysis, algorithms with different condition of folds and split to train and analyse algorithm's accuracy. Image processing techniques, representing a novel frontier in malware detection, leverage deep learning for analyzing malware datasets. Approaches based on image texture analysis, signal processing, and even representing binary bytes as audio signals were discussed. The use of artificial intelligence (AI) for detecting malware involves different effective methods such as signature-based, anomaly-based, heuristic-based, and specification-based techniques, each with its significance in strengthening security against known and unknown threats. While signature-based methods be outstanding in recognizing known threats but struggle with unknown ones, anomaly-based approaches overcome this limitation by using classification methods.

References

1. A comprehensive survey on deep learning based malware detection techniques. Gopinath, M., and Sibi Chakkaravarthy Sethuraman. s.l. : Computer Science Review 47, 2023.
2. Deep Learning Based Detection for Cyber Attacks in IoT Networks: A Distributed Attack Detection Framework. Olivia Jullian, Beatriz Otero, Eva Rodriguez, Norma Gutierrez, Héctor Antona, Ramon Canal. 2023, Journal of Network and Systems Management, p. 24.
3. "TaintDroid," ACM Transactions on Computer Systems. W. Enck, P. Gilbert, S. Han. 2014, p. 29.
4. "Leaksemantic: identifying abnormal sensitive network transmissions in mobile applications. H. Fu, Z. Zheng, S. Bose, M. Bishop, and P. Mohapatra. 2017, IEEE Conference on Computer Communications (INFOCOM), pp. 1-9.
5. "DroidScraper: a tool for Android in-memory object recovery and reconstruction,". A. Ali-Gombe, S. Sudhakaran, A. Case, and G. G. Richard,. 2019, International Symposium on Research in Attacks, Intrusions and Defenses, pp. 547-559.
6. Android Malware Detection Based on a Hybrid Deep Laerning Model. Tianliang Lu, Yanhui Du , Li Ouyang, Qiuyu Chen, and Xirui Wang. 2020, Hindawi Security and Communication Networks, pp. 1-11.
7. Malware Detection and Prevention using Artificial Intelligence Techniques. Faruk, M. J. Hossain. Orlando, FL, USA, : IEEE International Conference on Big Data, 2021.
8. IoT malware classification based on lightweight convolutional neural networks. Yuan, Baoguo. s.l. : IEEE Internet of Things Journal, 2021.