IMAGE PROCESSING STRATEGIESBY USING MLAS AND DEEP LEARNING MODELS FOR ROBUST INTELLIGENT MALWARE OPTIMAL PARAMETERS

K.V. Ranga Rao, Dept of Computer Science and Engineering, Sree Venkateswara College Of Engineering, Nellore (Dt), Andhra Pradesh, India.
N. Harish , Dept of Computer Science and Engineering, Sree Venkateswara College Of Engineering, Nellore (Dt), Andhra Pradesh, India.
Chegu. Rupa Kalpana, Dept of Computer Science and Engineering, Sree Venkateswara College Of Engineering, Nellore (Dt), Andhra Pradesh, India.
N.Kesav Kumar, Dept of Electronics and Communication Engineering, Sree Venkateswara College Of Engineering, Nellore (Dt), Andhra Pradesh, India.

ABSTRACT

Polymorphic, metamorphic, and other confusing techniques are used by modern malware to quickly change its behaviour and produce a wide range of malwares. Since new malware is frequently a variation of old malware, machine learning algorithms (MLAs) are already being used to direct a successful malware investigation. This necessitates a large amount of highlight representation, feature learning, and feature engineering. The attribute building process can be completely bypassed by using advanced MLAs, such as deep learning. Despite ongoing research in this area, the training data affects how well the algorithms work. Overcoming prejudice and undertaking unbiased study of these strategies is necessary in order to discover new, more efficient techniques for zero day malware. Traditional MLAs and deep learning architectures for malware detection, classification, and organization are compared in this study using both public and private datasets. The train and test divisions of the test examination make use of separate public and private datasets that were gathered across a variety of time periods. Additionally, we suggest a novel image processing method with parameters that are ideal for deep learning models and MLAs.

Keywords: MLAs, malware, deep learning.

1. INTRODUCTION

In this computerized era of Industry 4.0, technology is advancing quickly, which has an impact on both business and daily life operations. IoT and apps have helped advance the data society's development concept. However, overcoming security concerns will be extremely difficult given that cybercriminals target specific PCs and systems in order to steal sensitive data for financial gain and set up denial-of-service systems. These cybercriminals utilise malware or malicious software to seriously compromise systems and expose their flaws. [1]. Malware (OS) is computer software that is intended to harm the operating system. Malware can go by a number of names, depending on its function and behaviour, including backdoor, adware, spyware, worm, Trojan, root kit, virus, or ransomware.

2. LITERATURE SURVEY

Rossow,C.,Dietrichetal

36 academic articles from 2006 to 2011 that rely on the execution of malware are examined for methodological care and caution. Six major academic security conferences mentioned 40% of these publications. We regularly find errors, including questionable assumptions about the usage of execution-driven datasets (25% of the papers), a general lack of descriptions of security precautions applied during tests (71% of the articles), and generally insufficient representations of the exploratory arrangement.

Top-tier venues are not exempt from deficiencies, and the network must do a better job of handling malware datasets.

Saxe,J., &Berlin, Ketal

We describe a malware identification system based on Invincea Deep Neural Networks (DNN) that scales to real-world training example volumes on commercial hardware and achieves a useful recognition rate at a very low false positive rate. By leveraging more than 400,000 software binaries that were directly obtained from our clients and internal malware databases, we demonstrate that our framework achieves a 95% detection rate at a 0.1% false positive rate (FPR).. We also offer a nonparametric approach to help the classifier's scores more precisely represent the projected accuracy in the deployment environment.

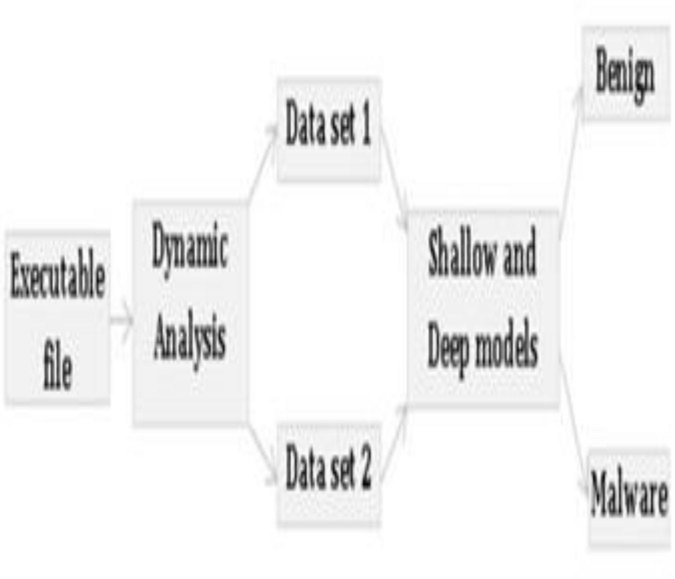
3. PROBLEM DEFINTION

To discover cyberattacks today, we analyse request data both statically and dynamically. Static analysis relies on signatures, So, in order to determine whether the packet is normal or contains an attack signature, we compare the contents of the new request packet with the most recent attack signature. In order to discover malware and attacks, dynamic analysis involves dynamic programme execution, although it is time-consuming.

4. PROPOSED APPROACH

To address this issue and enhance detection accuracy with both classic and contemporary malware attacks, the author employs a number of machine learning algorithms, such as Support Vector Machine, Random Forest, Decision Tree, Naive Bayes, Logistic Regression, K Nearest Neighbours, and Deep Learning Algorithms like (CNN) and LSTM. CNN and LSTM are the two most effective algorithms.

SYSTEM ARCHITECTURE



Dynamic Analysis-based Deep Learning Architecture

5. PROPOSED METHODOLOGY

DATASET

The author uses a binary malware dataset named "MALIMG" to carry out this study and assess the efficacy of machine learning methods. There are 25 malware families in this dataset.

FAMILIES OF MALWARE

There are 25 different malware families in the dataset, and their names are shown below: 'DialerAdialer.C', 'BackdoorAgent.FYI', and 'WormAllaple "WormAllaple.A', 'WormAllaple.L', 'Trojan Alueron.gen', 'Worm:AutoITAutorun.K', 'Trojan C2Lop.P', 'Trojan C2Lop.gen', 'Dialer Dialplatform.B', 'Trojan Downloader Dontovo.A', and 'Worm 'RogueFakerean', 'DialerInstantaccess', TrojanMalex.gen, Trojan Downloader Obfuscator.AD, Backdoor Rbot!gen, Trojan Skintrim.N, Trojan Downloader Swizzor.gen!E, Trojan Downloader Swizzor.gen!I, Worm VB.AT, Trojan Downloader Wintrim.BX, and WormYuner.A are some examples of malicious software.

MALWARE CLASSIFICATION

Using portable executables (PE), a few security scientists have leveraged their domain knowledge to recognise static malware. The two most widely used methods for discovering static malware without domain-level knowledge are now the analysis of byte n-grams and strings. In any event, the n-gram technique has poor performance and is expensive to compute. When developing an ML model to distinguish between harmful and beneficial files, it can be challenging to apply domain-level knowledge to omit the essential features.

The Windows operating system's inability to constantly uphold its own requirements and standards is the cause of this. The malware identification system needs to be changed in order to comply with upcoming security requirements because specifications and standards are always evolving. As a result, MLAs have been applied with features derived from the parsed data of the PE file.

CNN(convolutional neural network) The feed forward network (FFN) of the past, mostly used in the area of image processing, is supplemented by CNN. All connections, hidden layers, and its units are located there. For categorization, the CNN network has a totally linked layer. Every neuron in the totally linked layer has a connection to every other neuron.

ALGORITHM

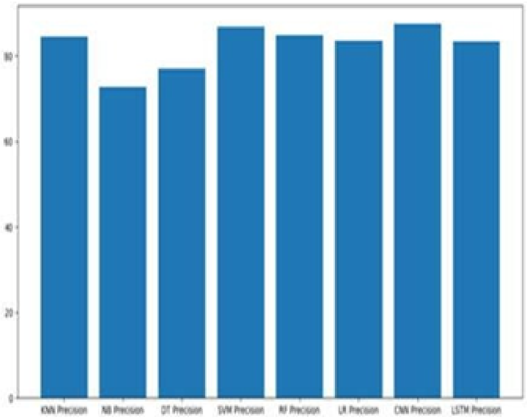
ML ALGORITHMS

Decision tree, Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, and K-Nearest Neighbours

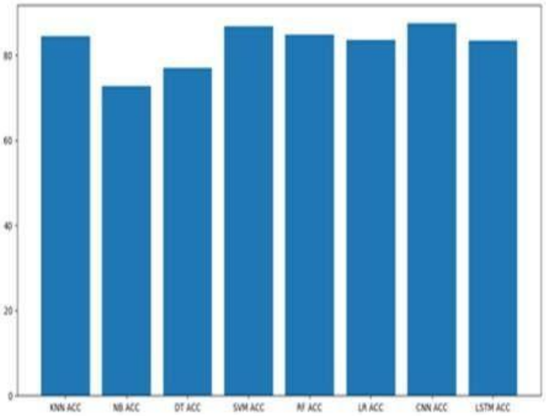
DEEP LEARNING ALGORITHMS

Convolution neural networks (CNN) and LSTM applications will convert this binary dataset into grayscale images in order to train and test machine learning models. MalConv CNN and MalConv LSTM are the names of these algorithms, and EMBER is the name of another algorithm. They transform binary input to pictures before building models. Datasets are converted into binary pictures by the application, which then uses 80% of them to train a model and 20% of them to test it. Every time we upload new test malware binary data, the programme updates the trained model to forecast the class of malware.

6. RESULTS



The accuracy graph for all algorithms is shown above, and CNN performs better. The x-axis in the graph above denotes the algorithm name, while the y-axis is the precision value.



As seen in the precision graph above for all methods, CNN performs better. In the graph above, the y-axis shows the accuracy value and the x-axis the name of the algorithm.

CONCLUSION

In this study, the effectiveness of the traditional MLAs and deep learning architectures for malware detection was evaluated using static analysis, dynamic analysis, and image processing techniques. Scale Mal Net was also recommended as an incredibly flexible framework for categorising and organising zero-day malware. The malware analysis process in this structure is split into two stages and uses deep learning on end user-provided malware samples. A combination of static and dynamic inspection was employed in the first stage to categorise malware. In the second step, malware samples were gathered, and malware classifications were compared using picture-preparation methods. Many test examinations in this investigation were conducted by using a variety of models on both publicly accessible benchmark datasets.

REFERENCES

1. Anderson,R.,Barton,C.,Böhme,R.,Clayton,R.,VanEeten,M.J.,Levi,M., &Savage,S.(2013).Measuring the cost of cybercrime. In The economics of information security and privacy (pp. 265300).Springer,Berlin,Heidelberg.

2. Li,B., Roundy,K., Gates,C., &Vorobeychik,Y.(2017,March).LargeScaleIdentificationofMaliciousSingletonFiles.InProceedings of the Seventh ACM onConference on Data andApplication Securityand Privacy(pp.227-238).ACM.

Alazab,M.,Venkataraman,S.,&Watters,P.(2010,July).Towardsunderstandingmalwarebehaviourbytheextraction of API calls. In 2010 SecondCybercrimeandTrustworthyComputingWorkshop(pp. 52-59).IEEE.

3. Tang,M.,Alazab,M.,&Luo,Y.(2017).Big data for cybersecurity: vulnerabilitydisclosuretrendsanddependencies.IEEETransactions onBigData.
4. Alazab, M., Venkatraman, S., Watters,P.,&Alazab,M.(2011,December).Zerodaymalwaredetectionbasedon supervised learning algorithms of API callsignatures.InProceedingsoftheNinthAustralasianDataMiningConference-Volume121(pp.171-182).AustralianComputerSociety,
5. Alazab,M.,Venkatraman,S.,Watters, P.,Alazab,M.,&Alazab,A.(2011,January).Cybercrime:thecaseofobfuscatedmalware. In 7th ICGS3/4th eDemocracy JointConferences2011:
6. ProceedingsoftheInternationalConferenceinGlobal Security, Safety andSustainability/InternationalConferenceone-Democracy(pp.1-8). [Springer]. Alazab,M.(2015).Profilingandclassifyingthebehavior ofmalicious codes. JournalofSystemsandSoftware,100,91102.
7. Huda,S.,Abawajy,J.,Alazab,M.,Abdollalihian,M.,Islam,R.,&Yearwood, J.(2016).Hybridsofsupportvectormachinewrapperandfilterbasedframeworkfor malwaredetection. Future GenerationComputerSystems,55,376-390.
8. Raff, E., Sylvester, J., & Nicholas, C.(2017, November). Learning the PE Header,Malware Detection with Minimal DomainKnowledge.InProceedingsofthe10thACMWorkshoponArtificialIntelligenceandSecurity(pp .121-132).ACM.
9. Rossow, C., Dietrich, C. J., Grier, C.,Kreibich,C.,Paxson,V.,Pohlmann,N., &Van Steen,M. (2012,May). Prudentpractices for designing malwareexperiments:Status quo and outlook. In Security andPrivacy (SP), 2012 IEEE Symposiumon(pp.65-79).IEEE.
10. Raff, E., Barker, J., Sylvester, J., Brandon, R.,Catanzaro, B., & Nicholas, C. (2017).Malware detection by eating a wholeexe.arXivpreprintarXiv:1710.09435
11. Krcál,M.,ˇSvec,O.,Bálek,M.,&Jaˇsek,O.(2018).DeepConvolutional MalwareClassifiersCanLearnfromRawExecutablesandLabelsOnly.
12. Rhode, M., Burnap, P., & Jones, K.(2018). Early-stage malware predictionusing recurrent neural networks. Computers&Security, 77, 578-594.
13. Anderson, H. S., Kharkar, A., Filar,B., & Roth, P. (2017). Evading machinelearning malware detection. Black Hat.
14. Verma, R. (2018, March). SecurityAnalytics: Adapting Data Science forSecurity Challenges. In Proceedings of theFourth ACM International Workshop onSecurity and Privacy Analytics (pp. 40-41).ACM.and privately acquired datasets showed that deep learning-based techniques beat traditional MLAs.