ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

A COMPARISON STUDY TO DETECT MALWARE USING DEEP LEARNING AND MACHINE LEARNING TECHNIQUES

Dr.Sugumar, Professor, Department Of AIML, SICET, Hyderabad
Bemagoni Bhargavi, Sonaganti Harshitha, Gundu Yashwin, Ch Gopi Sai Krishna
UG Student, Department Of AIML, SICET, Hyderabad

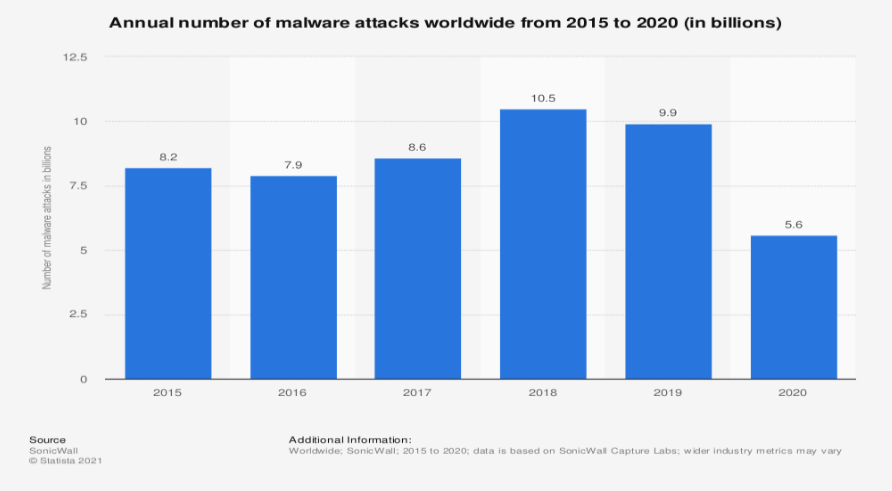
Abstract

Malware creation has evolved from simple malware that is easy to detect to complex malware that obscures and adapts quickly, becoming more difficult to detect. This study compares seven machine learning and deep learning methods to detect malware using captured bytes, opcodes, and code snippets. In this study, we aim to accurately classify malware from nine malware families. First, bytecodes, stanzas, and opcodes of different malware applications are extracted, combined, and classified using random forest, decision tree, support vector machine, K-nearest neighbor, SGD, logistic regression, naive Bayes, and deep learning. The results show that the accuracy of the deep learning model reaches 96%, which is better than other machine learning algorithm examples. Overall, this article highlights the importance of using advanced machine learning and deep learning to identify malware, especially given the complexity and evolution of today's malware. The findings suggest that deep learning techniques may be particularly effective at accurately detecting and identifying malware.Keywordsmachine learning, malware detection, bytecode, partial, opcode, random forest, decision tree, support vector machine (classifier), nearest neighbor, SGD, logistic regression, naive Bayes, deep learning model , malware classification, machine learning , Windows PC malware.

I. Introduction

Despite the success and growth of cybersecurity systems, malware is still one of the biggest threats online. Malware analysis uses techniques from various disciplines, including network analysis and activity analysis, to investigate malicious patterns to better understand their behavior and evolution over time [1]. Personal information is the most valuable asset for the majority of computer users today. Since they are very valuable to people, their safety is also very important. They must be protected at all costs. Malware was a minor problem as antivirus software protected the device. But until recently, it was used for malicious purposes that could lead to loss of life (healthy computer systems are stolen by ransomware) and loss of business and finances.Advances in malware management can be attributed to the increased use of the Internet in daily activities. It is now almost impossible to do anything without the Internet, including social, online commerce, healthcarebusiness, commerce and education. With the spread of the internet, criminals began to commit their crimes over the internet instead of the real world. Criminals often use malware to launch cyber attacks on victim systems such as phones, PCs, and laptops [2].Malware has subsequently grown over the last decade, causing huge financial losses for many businesses. -

Gifts. Figure 1 below shows statistics for millions of malware attacks each year from 2015 to 2020[3]. The report shows that there were approximately 5.6 billion malware attacks in 2020 . Therefore, it becomes important to determine whether the data contains malware. Due to the rapid development of the Internet and the resulting growth of malware, it is no longer possibl e to create search rules manually, it is necessary to create new solutions, strong protection is blocked. [4]. Because these attacks are so common, there is a growing need for technology to detect and remove malware.



Source
SonicWall
© Statista 2021

Additional Information:
Worldwide; SonicWall; 2015 to 2020; data is based on SonicWall Capture Labs; wider industry metrics may vary

Fig. 1. Malware attacks (in billions) per Year from 2015-2020 [3]

can greatly benefit from advances in deep learning [6]. To improve malware detection and classification, antivirus companies are turning to machine learning, a branch of computer science that has seen success in image recognition, search, and decision making [7].

Machine learning has changed the way many people work. In many areas, including cybersecurity, over the last decade. Machine learning solutions have been frequently researched and used to aid malware detection and classification. [8].In this study, many machine learning algorithms such as random forest, decision tree, support vector machine (classifier), K nearest neighbors, SGD, logistic regression and naive Bayesian decision trees were used for malware detection. Choose the most accurate process so that the system has more measurements.This article is divided into the following sections: a literature review describing previous studies, a methodology discussing the research findings, and an overview of the test results and interpretation of all test results. This paper is structured as follows: Section 2 provides a discussion of the study, Section 3 explains the methodology adopted in this study, Section 4 presents the analysis and results, and Section 5 presents conclusions and future work

2. Related Work

Malware classification has been studied extensively due to the significant risks associated with running malware. According to existing research, malware classification methods can be divided into two groups: those using machine learning and those using nonmachine learning. Malware is code created by attackers to damage user systems. Backdoors, viruses, rootkits, ransomware, worms, and adware are all examples of malware [9].

A. Non-machine learning techniques

In the past, malware was detected using static or dynamic techniques [10]. While static scanning detects malware without executing code, scanning detects malware while it is running [2]. Another nonmachine learning technique is the use of classical knowledgebased malware dete ction methods [11] [12], which are considered some of the first research in this field and support malware researchers. Pattern detection can also be applied to bytecode or by extracting code, extracting opcodes, and finding patterns in opcodes for lines of malware. Packaging and obfuscation are often used to hide malicious code [13] [14] [15]. Data mining techniques are also used to extract malware signatures. Data mining is simply the extraction of previously unknown and useful information from big data or archived data [16]. Mehdi et al. [17] obtained N grams of calls for malware distribution [18]. It is worth noting that most of the studies do not use technology or traditional methods in static or dynamic analysis [6].

III. Methods

This section describes recommended methods for identifying malware. We detail the data, data processing, machine learning, and deep learning models used for analysis. The diagram below shows the plan.Figure 2 shows the dataset used for feature extraction after the pre-processing stage. The data is divided into two: training data and test data. This training data is then fed into machine learning and deep learning models for analysis. Finally, the model creates a distribution and produces results. For prediction, we use test data to evaluate the training model and get prediction results.

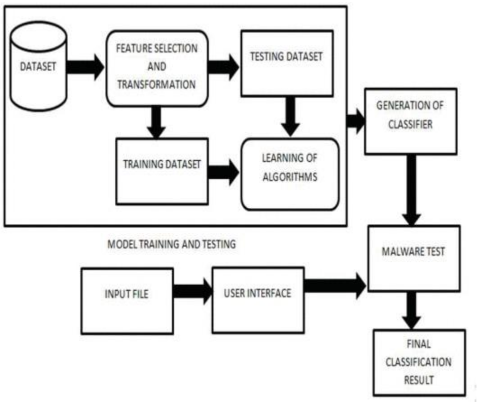


Fig. 2. Process Outline [23]

In the manual, the raw material for each file contains a hexadecimal representation of the binary content of the file, excluding the PE header (to be sterile). Also calls, strings etc. There is also a metadata directory, which is an engine that contains various metadata information extracted from binaries such as This was created using the IDA teardown tool. The binary contents of the file are stored as .bytes files, and the opcodes are stored as .asm file types. In this study, we reduced 21,000 files from 9 different malware families to 10,000 malicious files (5000 .asm and 5000 .byte files containing segments, bytes and opcode). This file [24] is approximately 0.5 TB in size without links.
B. Feature extraction and preprocessingThe malware dataset must have a feature vector or space map. Algorithms can use feature vectors for classification. Features extracted from the malware dataset are:
· Bytecode frequency
· Opcode frequency
· Segment frequency.

C. Classification - Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science that focuses on using data and methods to simulate human learning to improve reality [25] . Machine learning is an important part of the rapidly growing discipline of data science. Training algorithms can use statistical techniques to create classifications or predictions that can reveal important insights in data science projects [26].

In the malware classification phase, we used the following machine learning algorithms:

Random Forest
→Decision Tree
→Support Vector Machine
→K Most Neighbors
→SGD
→ Logistic Regression&
→Naive Bayes

Machine learning algorithms and neural networks use the algorithm as an input to classify unknown malware (from the file directory) into generated malware families. Our DNN model was built using TensorFlow's Keras framework.

a) Random Forest: Random Forest is a supervised machine learning algorithm that is frequently used to solve classification and regression problems. It creates decision trees from multiple models and uses majority voting and regression tools for classification. Random forest randomly selects data, builds decision trees, and averages the results [27]. It is not based on any standard [28]. It is stagnant in comparison. This means that a random forest is created from one of the data and the final output is based on the average or the most measured; overfitting is avoided [29].

b) Decision tree: decision tree is a type of tracking machine learning in which data is repeatedly segmented based on parameters. The tree can be described using two parts: the decision of the leaf and the decision of the leaf [30]. Pages indicate a decision or conclusion. Additionally, the data is classified in the decision [31]. Overfitting occurs when the decision tree evolves carelessly. [32]. Computationally, the decision tree is faster. When a decision tree combines data sets with features, it creates a process for prediction [29].

c) Support Vector Machine (Classifier): SVM (Support Vector Machine) is a supervised machine learning that can be used to solve classification and retrieval problems. We represent each data item as a location in an ndimensional space (where n is the number of features it has), and each feature is the value of a particular location in the SVM algorithm [33].

d) Neighborhood:: KNN algorithm is a supervised machine learning algorithm that can be used to solve classification and retrieval problems. The KNN algorithm determines whether similar objects are close to each other [34]. The KNN method assumes that new events/data are similar to existing events and assigns new data to categories similar to existing categories. The KNN method processes all existing data and classifies new data based on their similarity to existing data. Use the KNN method to classify new objects into appropriate groups. Although the KNN method can be used for both regression and classification, it is mainly used for classification. The K-NN algorithm is a non parametric algorithm, meaning it has no assumptions about the data [35].

e) Stochastic Gradient Descent (SGD): Stochastic Gradient Descent (SGD) is ideal for distributions and regressors for convex losses such as (linear) support vector machines and logistic r egression. Although SGD has long existed in machine learning, it has only recently begun to gain widespread attention in the context of largescale learning. SGD has been used to solve largescale machine learning problems common in text classification and natural language processing. Due to the unique data, classes in this model can be easily connected to patients with more than 105 training models and 105 features. [36].

f) Logistic Regression: Logistic regression is another tool used in machine learning. It is the preferred method for binary classification problems (problems with two sets of values). This paper [37] describes a logistic regression algorithm for machine learning. Use objective variable probability, supervised learning classification algorithms, and logistic regression. There are only two distributions because the target or variable is always bifurcated. Briefly, the variable was binary and data were expressed as 1 (pass/yes) or 0 (fail/no) [38] .

g) Naive Bayes:: Naive Bayes classifier is a simple and effective classifier that helps create fast learning models that can be implemented quickly. It is a probabilistic classifier, meaning it makes predictions based on the probability of an object occurring [39]. Naive Bayes demonstrates the advantages of text classification as a simple but powerful model of Bayes' theorem that can produce good results [40].

D. Deep Learning:

Deep learning allows multilayer computational models to learn different levels of abstraction of object representations. These technologies have improved greatly cuttingedge technologies in speech recognition, object detection, object detection, and many other fields such as drug development and genomics [41].

TABLE I BYTES

Algorithm	Accuracy	Precision	fscore	recall
DecisionTree	0.791	0.715	0.742	0.791
SVC	0.922	0.927	0.921	0.922
KNearest	0.905	0.914	0.905	0.905
NaiveBayes	0.705	0.805	0.710	0.705
SGD	0.835	0.848	0.816	0.8352
Logistic Regression	0.904	0.848	0.817	0.835
RandomForest	0.654	0.509	0.565	0.654
DNN	0.959	0.96	0.96	0.96

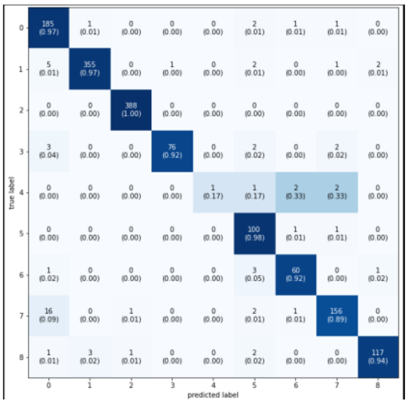
a) DNN Preprocessing and Model Creation: For this feature, we prepare the data in the desired format with a deep neural network. Since our labels are numbers, we use label encoding to convert the labels into an encoded field. In the model, we create a neural network model using the training model x, y test x, y, the encoder of the epochs and the preprocessing stage. The input method uses the relu activation function, eight nodes, and the input length of the training data. The model then adds three more layers with ten nodes using the same relu activation f unction. There is also a relu activation function in the layers below, but we use the softmax activation function in the last layer.We wrote our model using Adam optimizer and categorical cross entropy loss. We train the model using timeless and batch size 2. The smaller the batch size, the slower the network train, but the loss is small (all else being equal). Then we evaluate our model and prepare the analysis results. IV. Performance evaluation/Results

In this section, we focus on the results of the classification tests we conducted. We first show the distribution of malware family types in the dataset. This gives us an understanding of how data is distributed.Machine learning and deep learning neural networks are used to calculate accuracy, precision, F1 score, and recall. A method called calculation tpr fpr uses actual y and estimated y values to create a confusion matrix Figure 4. Confusion generated by deep learning models calculates true good, bad, bad, bad bad with TPR and FPR. Thus, the accuracy is calculated using the F1 score and the model.

A. Experimental results

In the process of combining the extracted results for analysis, SVC using all machine learning algorithms achieves 92% accuracy and 92% accuracy after DNN. Deep learning, on the other hand, results in 95% accuracy and 95% accuracy.

The analysis proves that deep learning outperforms machine learning due to backpropagation and gradient descent techniques.Furthermore, the performance of the system is determined by calculating the negative value and negative value using the confusion matrix (deep learning) in Figure 4.



	0	1	2	3	4	5	6	7	8
0	185 (0.97)	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.01)	1 (0.01)	1 (0.01)	0 (0.00)
1	5 (0.01)	355 (0.97)	0 (0.00)	1 (0.00)	0 (0.00)	2 (0.01)	0 (0.00)	1 (0.00)	2 (0.01)
2	0 (0.00)	0 (0.00)	388 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
3	3 (0.04)	0 (0.00)	0 (0.00)	76 (0.92)	0 (0.00)	2 (0.02)	0 (0.00)	2 (0.02)	0 (0.00)
4	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.17)	1 (0.17)	2 (0.33)	2 (0.33)	0 (0.00)
5	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	100 (0.98)	1 (0.01)	1 (0.01)	0 (0.00)
6	1 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	3 (0.05)	60 (0.92)	0 (0.00)	1 (0.02)
7	16 (0.09)	0 (0.00)	1 (0.01)	0 (0.00)	0 (0.00)	2 (0.01)	1 (0.01)	156 (0.89)	0 (0.00)
8	1 (0.01)	3 (0.02)	1 (0.01)	0 (0.00)	0 (0.00)	2 (0.02)	0 (0.00)	0 (0.00)	117 (0.94)

Fig. 4. Confusion Matrix

I. Discussion

cuttingedge technologies in speech recognition, object detection, object detection, and many other fields such as drug development and genomics [41].

a) DNN Preprocessing and Model Creation: For this feature, we prepare the data in the desired format with a deep neural network. Since our labels are numbers, we use label encoding to convert the labels into an encoded field. In the model, we create a neural network model using the training model x, y test x, y, the encoder of the epochs and the preprocessing stage. The input method uses the relu activation function, eight nodes, and the input length of the training data. The model then adds three more layers with ten nodes using the same relu activation function. There is also a relu activation function in the layers below, but we use the softmax ac tivation function in the last layer.We wrote our model using Adam optimizer and categorical cross entropy loss. We train the model using timeless and batch size 2. The smaller the batch size, the slower the network train, but the loss is small (all else being equal). Then we evaluat e our model and prepare the analysis results. IV. Performance evaluation/Results

In this section, we focus on the results of the classification tests we conducted. We first show the distribution of malware family types in the dataset. This gives us an understanding of how data is distributed.Machine learning and deep learning neural networks are used to calculate accuracy, precision, F1 score, and recall. A method called calculation tpr fpr uses actual y an d estimated y values to create a confusion matrix Figure 4. Confusion generated by deep learn ing models calculates true good, bad, bad, bad bad with TPR and FPR. Thus, the accuracy is calculated using the F1 score and the model.

A. Experimental results

In the process of combining the extracted results for analysis, SVC using all machine learning algorithms achieves 92% accuracy and 92% accuracy after DNN. Deep learning, on the other hand, results in 95% accuracy and 95% accuracy.

The analysis proves that deep learning outperforms machine learning due to backpropagation and gradient descent techniques.Furthermore, the performance of the system is determined by calculating the negative value and negative value using the confusion matrix (deep learning) in Figure 4.

This may be particularly important due to the increasing complexity and adaptability of mode rn malware; This makes it difficult to find using traditional methods. Overall, the study demo nstrates the potential of deep learning for detecting and classifying malware and suggests that further research in this area will contribute to cybersecurity.

REFERENCES

[1] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations”, J. Mod. Transport. 24(1):62–72..

[2] Tessa K. Anderson, “Kernel density estimation and Kmeans clustering to profile road accident hotspots”, Accident Analysis and Prevention 41,359–364.

[3] Shristi Sonal and Saumya Suman “A Framework for Analysis of Road Accidents” Proceedings of International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR)

[4] Analysis of road accidents in India using data mining classification algorithms- E. Suganya,S. Vijayrani.

[5] Hao, W., Kamga, C., Yang, X., Ma, J., Thorson, E., Zhong, M., & Wu, C., (2016), Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States, Transportation research part F: traffic psychology and behavior, 43, 379-386.

[6] Li, L., Shrestha, S., & Hu, G., (2017), Analysis of road traffic fatal accidents using data mining techniques, In Software Engineering Research, Management and Applications (SERA), IEEE 15th International Conference on (pp. 363-370). IEEE.

[7] El Tayeb, A. A., Pareek, V., & Araar, A. (2015). Applying association rules mining algorithms for traffic accidents in Dubai. International Journal of Soft Computing and Engineering.

[8] Bahram Sadeghi Bigham ,(2014),ROAD ACCIDENT DATA ANALYSIS: A DATA MINING APPROACH, Indian Journal Of Scientific Research 3(3):437-443.

[9] Divya Bansal, Lekha Bhambhu, "Execution of Apriori algorithm of data mining directed towards tumultuous crimes concerning women", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 9, September 2013.

[10]S. Krishnaveni, M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques", International Journal of Computer Applications, vol. 23, no. 7, pp. 40-48, June 2011.