International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 86-91
Published Online September 2022 in IJEAST (http://www.ijeast.com)

MALWARE DETECTION USING MACHINE LEARNING TECHNIQUES

Ms.Sakshi Joshi,
Department of C.Sc & Engg.
KLS VPP, Belagavi, Karnataka, India

Mr.Santosh Mahagaonkar
Research Head, NICT Solutions &Research
Belagavi, Karnataka, India

Abstract-**Malware attacks have become serious and crucial issue now a days, as it can affect victim in many ways. Hence detecting malware at early stage is an essential aspect in the security of computer systems. Existing malware system contains a traditional antivirus detection method that depends on signature-based and behavioral methods. Traditional methods of malware detection are not that effective and cannot detect unknown malwares. In recent years machine learning is coming out as an emerging and challenging field in malware detection. Proposed method implements machine learning and deep learning technique for detecting malware. This is achieved using machine learning algorithm, Support Vector Machine and deep learning concept using Convolutional Neural Networks where in malwares are represented as images. The study compares the performance of conventional, machine learning-based, and deep learning-based malware detection techniques. Proposed method implemented for malware detection using Convolutional Neural Networks with malware images is more secure compare to dynamic based method as binary malware files are converted to images and images are never executed also it can reduce drawbacks of traditional signature based method at some extent.**

Keywords—**Machine Learning, Convolutional Neural Network, Support Vector Machine.**

I. INTRODUCTION

Malware detection techniques have received considerable attention and scope due to increasing graph of cyber-attacks day by day. Malicious Software abbreviated as Malware**,** is a program developed to infringe and damage a computer system and information or data without the possessor's knowledge and permission, which is a very serious menace to the security of systems from last so many years. The threat is increasing with alarming pace as the use of internet in our daily activities is growing expansively. Malware is classified as worms, viruses, Trojan horses, ransom ware, spyware, and root kits. etc. Malware-Family has been built and engineered to harm the victim's computer in a variety of ways, such as by causing damage to the target system, stealing information, and more. So, today it is extremely essential to invent new techniques and different approaches for detecting malware.

II. MOTIVATION

The volume, intensity, and malware strike on the global economy have been steadily increasing in recent years. It is estimated that around 1 million malware files are developed every day, based on statistics and business data, and have an impact on and cause harm to the international economy in the amount of roughly 60, 00,00,00,00,000 US Dollars by 2022. In current scenario, for users and organizations protecting computer systems and network is one of the fundamental and top priority task, because even a single cyber-attack can result in severe damages to data and grim loss. There are many recent cases of malware attacks like CovidLock-20 ransom ware, Lockergoga-19 ransomware, Emotet -18 trojan which were responsible for huge amount of damage in terms of data access, financial loss, information theft etc. Frequent cyber-attacks calls for the need of reliable and precise techniques for detection. The motive of this work is to analyze different malware detection methods and provide malware detection system which gives good results with better accuracy.

III. PROBLEM DEFINITION

Malware detection system has become a fundamental need today because it works as an early warning system for malware attacks. There are numerous approaches developed for this issue. Each has its own merits and demerits over one another. As a result of the current research, machine learning approaches have been widely employed to expedite and enhance malware detection while maintaining a high accuracy rate. This paper presents different malware detection methods based on machine learning (ML) and deep learning technique. It provides performance analyses of implemented methods which helps to demonstrate which method works better over another.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 86-91
Published Online September 2022 in IJEAST (http://www.ijeast.com)

IV. LITERATURE SURVEY

The author O Aslanetal. [1] has reviewed all the approaches and methods for malware detection with pros and cons of every approach in this paper. It explains malware techniques, with algorithms and respective method schema systematically. This paper also gives details of different malware datasets available. It gives comparison of different malware approaches.

Analysis of static and dynamic approaches of malware detection was proposed by **Muhammad Ijaz et al.[2]**.It has been demonstrated that it is feasible to do dynamic malware investigation by combining various features in a variety of ways.

The improved signature based method has been proposed by the author **Pankaj Kohli et al.[3].** Signatures are generated depending on characteristics of complete malware class instead of single malware. On the basis of the API calls made by members of a malware class, it is possible to determine the malware class behavior.

Zhao et al.[4]The author presented a novel malware detection approach that makes use of machine learning and combines dynamic and static characteristics to identify malicious code. Author worked with NB and SVM conventional ML models for finding accuracy. The method overcomes demerits of traditional methods at some extent.

D. Uppalet al.[5]the author had carried out comparison of several machine learning algorithms, such as NB and SVM models. The most significant shortcoming of machine learning-based detection systems is that they rely on a virtual platform to analyses data; the samples must be performed which not only degrades their runtime performance, but also reduces the overall system performance.

The author have carried out feature selection, feature extraction and classification for detecting malware using traditional ML approaches **Ye Y. Li et al. [9].** However, important features like file structural aspects and few dynamic features like opcode and traces of API are left out. Also deep learning and multimodal methods for malware detection, which are ongoing areas for the recent years, are not touched.

V. EXISTING SYSTEM

The existing system contains classic antivirus detection methods that rely on cryptography, adaptive and contextual methods. However, signature-based methods are not capable of detecting unknown malware variants. It identifies only those malwares whose signatures are stored in the database. To handle these issues, behavior-based detection have been proposed, In order to detect whether a file is malicious, it is examined for its properties and behavior. However, inspection and assessment can take significantly longer with this procedure.

Disadvantages of Existing System

Existing system is not capable of finding unknown and new generation polymorphic malwares. It identifies only those malware whose signatures are stored in database. Many of the times, most of the new malwares will be very similar to the known malware samples, but still signature-based method fails to detect them because they do not consider behavioral and structural properties of malwares for detection. Also processing large volume of data is not possible. Scanning of malware file characteristics would take more time. With existing system it is difficult to find accuracy in the detection method.

VI. PROPOSED SYSTEM

A subset of artificial intelligence, machine learning (ML), is a technique that allows software programmers to increase their accuracy at predicting events without having to build them explicitly from the beginning. For prediction purposes, these ML model makes use of previously collected data. Deep learning is a subclass of machine learning, which is basically a neural network with three or more layers. Neural network attempt to mimic the way human beings gain some sort of knowledge. Deep learning algorithms differ from conventional machine learning algorithms in that they are structured in levels of growing difficulty and sophistication rather than in a logical way. The system is proposed using machine learning and deep learning techniques. In the proposed system we have used machine learning algorithm SVM (Support Vector Machine) and deep learning algorithm CNN (Convolutional Neural Networks) for malware detection. Also traditional signature based method is implemented to compare performance with other two methods.

Support Vector Machine Algorithm

SVM stands for Support Vector Machine and is one of the most frequently used Supervised Learning algorithms for classification issues in machine learning. This algorithm's primary goal is to define the decision boundary that can divide n-dimensional space into classes, allowing us to place newer data points in the correct category with relative ease. Hyper plane is the name given to this most precise decision boundary.

The support vector points are the vector points that are closest to the hyper plane. This is done because these two locations are helping towards the output of the method, as well as the other vector points are not contributing to the outcome. A data point that is not contribute significantly to the overall has no influence on the model if it is removed from it. On the other hand, removing the support vectors may then alter the point of the hyperplane. When two vectors are separated by a hyper plane, this is referred to as margin. A line is separated by the points that are closest to it.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 86-91
Published Online September 2022 in IJEAST (http://www.ijeast.com)

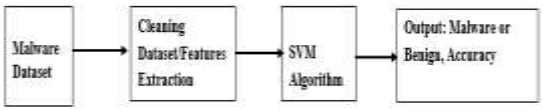

Figure 1: Proposed System for SVM Method

CNN Algorithm

When given a picture as input, a Convolutional Neural Network (ConvNet/CNN) assigns trainable weights and biases to various elements of the image, and subsequently is able to discriminate between the two images, this is referred to as deep learning. In comparison to all other classification techniques, the amount of pre-processing required by a CNN is significantly less. While simple techniques require filters to be handcrafted, ConvNets have the opportunities to succeed these filters or features with sufficient experience and training.

ConvNets

Essentially, a ConvNets is a sequence of layers, each of which turns one volume into another by the use of a partial differential equation.

ConvNets have several different sorts of layers. To illustrate, consider the following image, which was processed using a ConvNet dimension: 34*34*3.

1. Input Layer: Essentially, the layer manages of storing the image's raw input. With width 34, height 34 and depth 3.
2. Convolution Layer: It manages of calculating the output level by calculating the matrix multiplication between both the filters and the picture patch, among other things.Given that we employ a total of 14 filters for this layer, the resulting volume has the following dimensions: 34*34*14.
3. Activation Functional Layer: This component is accountable for applying the convolution layer's output to an element-by-element activation function..
4. Pool Layer: A layer is occasionally introduced into ConvNets and its major goal is to minimize volume, speed up calculation, and prevent over fitting.

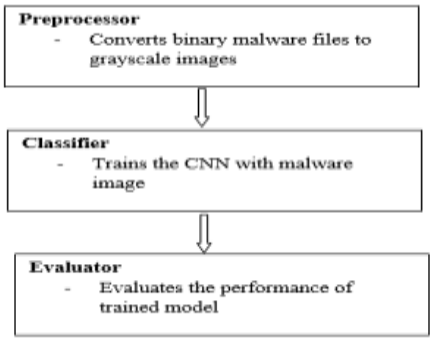

Figure 2: Proposed System for CNN Model

This system consists of three modules viz. Preprocessor, Classifier and Evaluator. Preprocessor module converts a raw input that is binary malware file into a matrix format there by converting it to gray scale image. This can be done by interpreting every byte in binary file as one pixel in an image, where in values ranges from 0 to 255. Later on, the resultant array is rearranged as a two dimensional array. Padding is done to adjust the image size. The classification module trains the CNN or evaluates the image by taking the transformed image given by preprocessor.

The evaluation module classifies images as malware or benign using the classification module and evaluates the accuracy of the model.

Proposed System for Signature Based Method

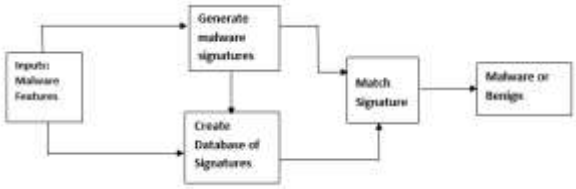

Fig 3 :Proposed System for Signature Based Method

VII. RESULTS

A. Dataset Details

Malware Dataset used for signature based method and SVM method contains 215 features.215th feature is class. '1' is indicated as Malware and '2' is indicated as Benign. In Dataset total 15036 samples data is available. From that 5560 are Malware cases and 9476 are benign cases.

Dataset used for CNN method contains around 200 files which are converted binary files of malwares.

Signature generation for malware dataset.

This part shows signatures generated for malware dataset. Signatures are generated by considering 4 bit at a time and converting it to hexadecimal numbers. This results in 52 bit hexadecimal string which is considered as malware signature.

Like this database of signatures are generated for both training and testing dataset.

Below figure shows 52bit signature in first column, second column shows category or class. 1-Malware, 2- benign. Third column shows index number for record in dataset.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 86-91
Published Online September 2022 in IJEAST (http://www.ijeast.com)

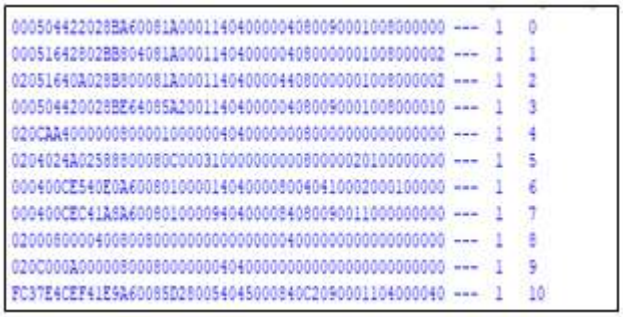

Fig 4: Signature generation for training dataset

Performance Analyses for signature based method

Below figure shows performance analysis for signature based method. Accuracy achieved for signature based method is 94%. Also confusion matrix with true positive, false positive, sensitivity, specificity values are shown.

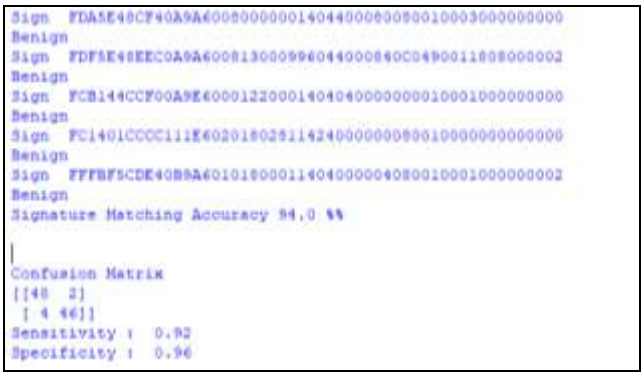

Fig 5: Performance Analyses for signature based method

B. Classification and Prediction using SVM

Classification is important approach in which program gains training data and learns from it and then uses these learnt observations on test data for classification. Here we have used SVM algorithm for classification. Below figure shows performance analysis for SVM algorithm. The accuracy achieved with this method is 95%, which is higher than signature based method. Also true positive rate is higher for SVM method.

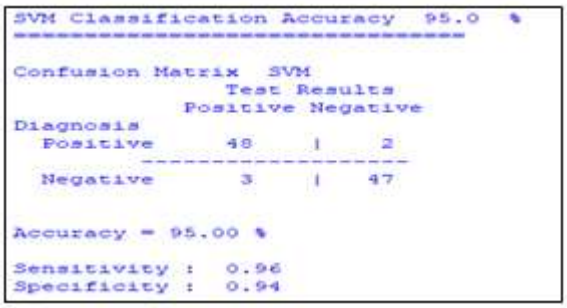

Fig 6: Performance Analyses for SVM model

C. Converting binary files of malwares to gray scale images

Figure shows converted gray scale images for two different binary files of malwares. The conversion is accomplished by expressing each byte from a binary file as a single pixel in a grey scale picture and arranging them in a two-dimensional matrix of size 300x300. When image size is small padding is done for remaining pixels to fit it in to required size.

Figure 7:Converted binary files of malwares to gray scale images

Prediction and classification using CNN model

Figure: 9 shows classification achieved using CNN model on testing dataset. It shows malware image name, its actual class and predicted class by the CNN model. Class can be 'Malware' or 'Good ware'. Output is displayed(Fig 8) by printing label as malware or good ware on image.

Fig 8:Predicted output

Fig 9:Prediction using CNN model

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 86-91
Published Online September 2022 in IJEAST (http://www.ijeast.com)

Performance Analysis of CNN model

Below figure shows performance and confusion matrix details by CNN model. Result shows accuracy achieved by CNN model is 96.59% which is greater than signature based and SVM method. Also it displays other performance parameters such as true positive, false positive rate. CNN model has higher true positive rate compare to other two methods.

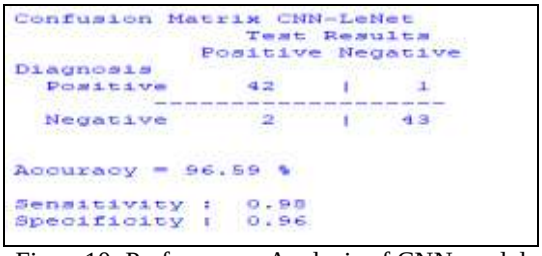

Figure10: Performance Analysis of CNN model

ROC curve graph for CNN model

Below figure shows ROC (Receiver Operating Characteristic) curve for CNN model. For a variety of various threshold values between 0.0 and 1.0, the graphic shows the FPR along the X-axis vs the TPR along the y-axis for each of the different threshold values.

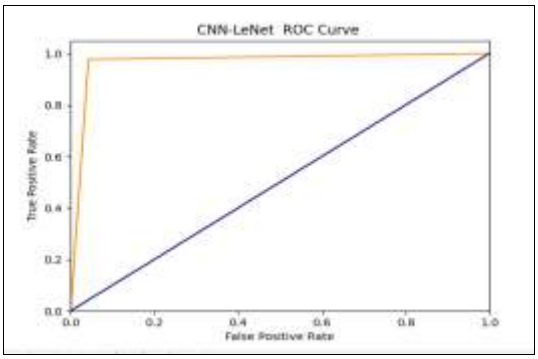

Figure 11: ROC curve for CNN model

Accuracy comparison for Signature, SVM and CNN

Figure 12 shows bar graph of accuracy comparison for signature method, SVM and CNN classifier. Result shows signature method achieved accuracy 94 %, SVM- 95% and CNN classifier 96.59 %. CNN classifier has achieved highest accuracy over other two methods.

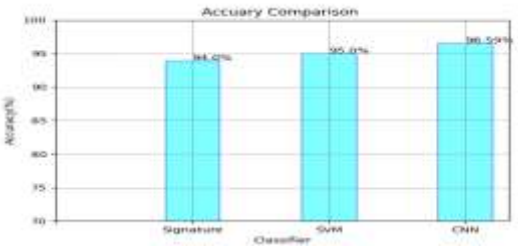

Figure 12:Accuracy comparison for Signature, SVM and CNN

VIII. CONCLUSION

In the proposed method three techniques for malware detection viz. traditional signature based, machine learning technique using SVM algorithm and deep learning through image processing using CNN algorithm have been implemented and performance is analyzed successfully. Signature based method has achieved 94% accuracy, SVM method 95% and that of CNN method is 96.59 %. We can conclude that machine learning and deep learning technique shows better accuracy over traditional method. Deep learning technique with CNN has achieved highest accuracy of 96.59% and high true positive rate.

When there is small variation between samples which belongs to same family, signature based method cannot identify such malware samples as its signature changes. But there is no much difference between gray scales images formed when there are small variations in samples. Such image distortion CNN can identify or it can be trained to identify thereby overcoming drawback of signature based method.

Image based classification implemented using CNN technique has been shown to be extremely successful since it makes use of the structural similarities between known and fresh malware samples to identify threats. It is secure compare to dynamic based method as binary malware files are converted to images and images are never executed. Since binary malware files are converted into image representation format, we have made our analysis independent of any tool.

IX. FUTURE SCOPE

The work can be extended to analyze performance with different machine learning algorithms for various datasets. Size of the datasets can be improved to check performance which is not done here due to computational limitations. To work with deep learning techniques different image resize/compression techniques can be explored. Flexible malware detection framework can be designed which should work on different platforms.

X. REFERENCE

[1]. Omer Aslan; Samet, Refik.(2020) "A Comprehensive Review on Malware Detection Approaches". IEEE Access.

[2]. Muhammad Ijaz , Muhammad HanifDurad , Maliha Ismail (2019),"Static and Dynamic Malware Analysis Using Machine Learning" in © IEEE

[3]. Pankaj Kohli ,Bruhadeshwar Bezawad in ResearchGate , 2008 - "Signature Generation and Detection of Malware Families"

[4]. Zhao, Jingling; Zhang, Suoxing; Liu, Bohan; Cui, Baojiang.(2018) -"Malware Detection Using Machine Learning Based on the Combination of Dynamic and Static Features."in 27th International Conference on Computer Communication and Networks (ICCCN) .

[5]. D. Uppal, R. Sinha, V. Mehra, and V. Jain,(2014)- "Malware Detection and Classification Based on Extraction of API Sequences,at Conf. Advances in Computing, Communications and Informatics.

[6]. Choi,Sunoh&Jang,Sungwook & Kim,Youngsoo & Kim, Jonghyun.(2017). -"Malware detection using malware image and deep learning

[7]. Long Wen, and Haiyang Yu,((2017)-" An Android malware detection system based on machine learning" AIP Conference Proceedings 1864, 020136

[8]. Irina Baptista , Stavros Shiaeles∗ and Nicholas Kolokotronis in IEEE 2019 -"A Novel Malware Detection System Based On Machine Learning and Binary Visualization"

[9]. Ye, Y., Li, T., Adjeroh, D., Iyengar, S.S. Jun 2017., A survey on malware detection using data mining techniques. ACM Comput. Surv.

[10]. Sathyanarayan, V.Kohli, Pankaj,Bezawada, Bruhadeshwar, (2008) "Signature Generation and Detection of Malware Families",

[11]. Prof. M. P. Wankhade, Jyoti Landage,(2013) "Malware and Malware Detection Techniques: A Survey" in IJERT Vol. 2 ,December

[12]. Chumachenko, K. (2017). "Machine Learning Methods for Malware Detection and Classification."

[13]. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. (2015), - "Deep learning". In: nature 521.7553 p. 436

[14]. Razvan Pascanu et al. (2015) "Malware classification with recurrent networks". In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. IEEE, pp. 1916–1920.

[15]. Daniel Gibert Llaurado. (2016) "Convolutional neural networks ´ for malware classification". MA thesis. Universitat Politecnica de Catalunya,

[16]. Yanfang Ye et al. "A Survey on Malware Detection Using Data Mining Techniques". (June 2017),In: ACM Comput. Surv. 50.3 41:1–41:40. ISSN: 0360-0300. DOI: 10.1145/3073559

[17]. Kaiming He et al. (2016) "Deep residual learning for image recognition". In: Proceedings of the IEEE conference on computer vision and pattern recognition.

[18]. B. Sanjaa and E. Chuluun,(2013)- "Malware detection using linear SVM," Ifost, pp. 136-138, doi: 10.1109/IFOST.2013.6616872.

[19]. Kruczkowski and E. N. Szynkiewicz, (2014)- "Support Vector Machine for Malware Analysis and Classification," International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

[20]. V. J. Raymond and R. J. R. Raj(2021)-"Android malware detection and build model using support vector machine," 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT),