A Machine Learning Technique to Detect Behavior Based Malware

Pavitra.P.Madanapalle[1], Mr. P. Firoze Khan[2], Pranathi.E[3], Prasanna.U[4], Ramya sree.K[5]
[1,2,3,4,5] *Institute of Technology & Science, Post Box N0:14, Kadiri Road, Angallu(V), Madanapalle - 517325 Annamayya District, AndhraPradesh, India*

Abstract-Malware has been a concern to enterprises for a long time, and progress in identifying malware on time has been slow. Malware may easily affect the system by executing excessive services that place strain on the system and impede its smooth operation. There are Malware has been a Malware has been a concern to enterprises for a long time, and progress in identifying malware on time has been slow. Malware may easily affect the system by executing excessive services that place strain on the system and impede its smooth operation. There are basically two approaches for identifying malware: the traditional method of detecting malware based on signatures and the new method based on behavior. The malware's behavior is characterised by the action it conducts when active in the machine, such as executing the operating system functions or downloading infected files from the internet. The suggested method identifies malware by analysing its activity. The suggested model in this study is a hybrid of Support Vector Machine and Principle Component Analysis. During validation, our suggested model attained an accuracy of 91.75% with 91% precision, 70% recall, and for real Malwares. The constant surge in malware abusing the Internet has become a severe concern. As compared to the high pace of malware dissemination, manual heuristic inspection of malware analysis is no longer regarded effective and efficient. As a result, automated behavior-based malware detection employing machine learning algorithms is seen as a comprehensive solution. The behavior of each virus in an emulated (sandbox) environment will be automatically examined, and behavior reports will be generated. Support Vector Machine (SVM) and PCA classifiers were employed in this study. Based on the analysis of the tests and experimental results of all the SVM classifier's tests and experimental results, the overall best performance was reached by 90.3%, with a precision of 96.8%. In conclusion, a proof-of-concept based on autonomous behavior-based malware analysis and the use of machine learning algorithms might identify malware effectively and efficiently.

I.INTRODUCTION

With the increased use of the internet and computer systems, data (personal and professional) security has become a serious challenge. Computers using the internet download massive amounts of data from the internet, which may potentially include viruses. Malware is known by many various names, including malicious code, harmful programmes, and malicious executable files. The increasing sophistication of malware assaults has rendered computer systems increasingly vulnerable to hacking. According to Kaspersky Laboratories, malware is "a form of computer software designed to infect a legitimate user's computer and inflict harm on it in a variety of ways". With the rising diversity of malwares, anti-virus scanners cannot guarantee the identification of every form of malware based on its signature, resulting in millions of hosts being targeted and inflicting significant harm to data and other connected systems. According to Kaspersky Lab (2016), 6,563,145 distinct machines were attacked, with around 4,000,000 new forms of malware discovered.

As a result, safeguarding the network and user machines from malware is a critical cyber security duty for a single user or a whole enterprise, because a single assault may result in considerable loss and harm. The goal of this work is to provide a malware detection system that will give an effective approach to identify malware based on the actions it may undertake on the machine on which it is installed. Malware comes in many forms however the following are the most common.

A **virus** is a little bit of code that has the power to replicate itself. It is connected to any genuine file and runs its code when the file is downloaded or processed.

Worms, like viruses, have the ability to multiply themselves. The main distinction between a worm and a virus is that a worm operates on a network while a virus does not. It duplicated itself by sending copies to the devices linked to the network.

Spyware that is frequently included with free applications. After the user installs the programme, spyware is activated and begins collecting the user's personal information from the system and passing it on to the host is software machine through a network.

Adware is a harmful piece of code that is connected to any advertisement that is playing on the screen or a 'click me' button. When a user clicks on a button or advertising, the code associated to it runs and installs a virus or bot onto the user's computer.

Trojans often mislead users by masquerading as authenticating programmes, such as any login page to a website or contact information form.

Botnets are defined as a networked collection of several bots. A single bot is a little piece of code that is assigned the duty of allowing a hacker simple access to a user's system. A hacker using a bot can launch viruses on the user's system, capture personal information, or decrease the performance of the user's machine.

Malware Detection is done in two phases.

1. malware analysis
2. malware detection

Malware Analysis is the first phase of the Detection. In this phase the data is collected of previously known malwares. Features are generated and extracted of those malwares and an algorithm is developed based on those features to detect the new incoming malwares

Malware Detection comes after the analysis is done and a proper algorithm is generated which provides a high accuracy in detecting the malware. The algorithm developed is then implemented on the incoming packets and then checked whether it is a malware or benign.

II.LITERATURE SURVEY

Introduction:

A literature review is a comprehensive summary of previous research on a topic. The literature review surveys scholarly articles, books, and other sources relevant to a particular area of research. The review should enumerate, describe, summarize, objectively evaluate and clarify this previous research. It should give a theoretical base for the research and help you (the author) determine the nature of your research. The literature review acknowledges the work of previous researchers, and in so doing, assures the reader that your work has been well conceived. It is assumed that by mentioning a previous work in the field of study, that the author has read, evaluated, and assimilated that work into the work at hand.

A literature review creates a "landscape" for the reader, giving her or him a full understanding of the developments in the field.

[1] M. A. Jerlin and K. Marimuthu, "A New Malware Detection System Using Machine Learning Techniques for API Call Sequences," *J. Appl. Secur. Res.*, vol. 13, no. 1, pp. 45–62, 2018.

In this system, M. A. Jerlin and K. Marimuthu are using a novel context-sensitive and keyword density-based method for classifying webpages by using three supervised machine learning techniques, support vector machine, maximum entropy, and extreme learning machine. The performance is evaluated by using a benchmark data set which consists of one hundred thousand webpages. Experimental results show the accuracy of 98.24%.

[2] B. Sanz, I. Santos, C. Laorden, X. Ugarte-Pedrero, P.G. Bringas, and G.Álvarez, "PUMA: Permission usage to detect malware in android," *Adv. Intell. Syst. Comput.*, vol. 189 AISC, pp. 289–298, 2013.

In this research paper, we present PUMA, which is used for predicting malicious Android applications through machine-learning techniques by analysing the extracted permissions from the application itself.

[3] MY. Fan, Y. Ye, and L. Chen, "Malicious sequential pattern mining for automatic malware detection," ***Expert Syst. Appl.,*** **vol. 52, pp. 16–25, 2016**.

The authors tested their method on a dataset including 1,800 malware samples and 1,800 benign traces. Their technique had a detection rate of 97.9% and a false positive rate of 1.1%, according to the data.

Overall, the research provides a unique method for automatically detecting malware using sequential pattern mining. With high detection rates and low false positive rates, the method is a viable tool for malware detection in practise.

[4] U. Baldangombo, N. Jambaljav, and S.-J. Horng, "A Static Malware Detection System Using Data Mining Methods," 2013

The paper's methodology employs five data mining techniques: Decision Tree, Naive Bayes, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and Random Forest. The datasets from the Malware Genome Project were utilised by the authors to train and test the models. The databases included both benign and malicious samples.

[5] Y. Saint Yen and H. M. Sun, "An Android mutation malware detection based on deep learning using visualization of importance from codes," ***Microelectron. Reliab.,*** **vol. 93, no. October 2018, pp. 109–114, 2019.**

The authors tested their technique on a dataset of Android applications that comprised both benign and malicious applications. They tested their deep learning model on the remaining apps after training it on a portion of the dataset. They discovered that their method was efficient in identifying malware in Android applications, with an accuracy rate of more than 97%.

[6] S. Vanjire and M. Lakshmi, "Behavior-Based Malware Detection System Approach For Mobile Security Using Machine Learning," ***2021 International Conference on Artificial Intelligence and Machine Vision (AIMV),*** **Gandhinagar, India, 2021, pp. 1-4, doi: 10.1109/AIMV53313.2021.9671009.**

To train and evaluate their machine learning models, the authors employ a dataset of over 800 Android applications, including both benign and dangerous apps. They analyse the performance of several methods, such as decision trees, random forests, and support vector machines (SVMs), and discover that SVMs deliver the greatest results in terms of accuracy, precision, and recall.

[7] W. Liu, P. Ren, K. Liu and H. -x. Duan, "Behavior-Based Malware Analysis and Detection," 2011 First International Workshop on Complexity and Data Mining, Nanjing, China, 2011, pp. 39-42, doi: 10.1109/IWCDM.2011.17.

The authors test their approach using a dataset of known malware samples and demonstrate that it detects a substantial majority of the malware samples. They also compare their technique to other behavior-based detection methods, demonstrating that it beats the others in terms of detection rate and false positive rate.

Overall, the article offers a thorough introduction of behavior-based malware analysis and detection, as well as a practical way for identifying malware based on its behaviour. Through experimentation and assessment, the authors establish the usefulness of their technique, making this study an important contribution to the field of malware research and detection.

[8] W. Liu, P. Ren, K. Liu and H. -x. Duan, "Behavior-Based Malware Analysis and Detection," 2011 First International Workshop on Complexity and Data Mining, Nanjing, China, 2011, pp. 39-42, doi: 10.1109/IWCDM.2011.17.

The authors tested their approach on a dataset of 100 Android apps, ten of which were known to contain malware. The findings indicated that the suggested approach has a 90% detection rate with a 10% false positive rate.

Overall, the study proposes a unique way to identifying malware on mobile phones that may be effective in combating the rising problem of malware on mobile platforms. The proposed technique is simple and effective, and it may be included into current mobile security systems.

[9] I. Firdausi, C. lim, A. Erwin and A. S. Nugroho, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection," 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, Jakarta, Indonesia, 2010, pp. 201-203, doi: 10.1109/ACT.2010.33.

The authors go over the different methodologies for detecting behavior-based malware, such as static analysis, dynamic analysis, and hybrid analysis. They then give a thorough examination of machine learning algorithms for behavior-based malware detection, such as artificial neural networks, support vector machines, decision trees, and Bayesian networks.

Finally, the article emphasises the significance of behavior-based malware detection in today's digital environment and provides a thorough examination of machine learning approaches that may be utilised for this purpose. The study provides significant information for cybersecurity and malware detection academics and practitioners.

[10] Wei-Ling Chang, Hung-Min Sun and Wei Wu, "An Android Behavior-Based Malware Detection Method using Machine Learning," 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Hong Kong, 2016, pp. 1-4, doi: 10.1109/ICSPCC.2016.7753624.

The suggested technique is divided into two parts: feature extraction and classification. The authors extract a collection of behavioural characteristics from the Android application during the feature extraction step by observing its runtime behaviour. System calls, network activity, file access, and process information are examples of these functionalities. In the classification step, the authors use the retrieved information to build a machine learning system that classifies the programme as malware or benign.

[11] T. -H. Nguyen and M. Yoo, "A behavior-based mobile malware detection model in software-defined networking," 2017 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2017, pp. 1-3, doi: 10.1109/ICISCT.2017.8188590.

The suggested model is made up of three primary parts: a mobile device, an SDN controller, and a detecting module. The application to be examined for malicious activity is executed on the mobile device, while the SDN controller acts as a bridge between the mobile device and the detection module. The detection module is in charge of studying the application's activity and deciding if it is malicious or not.

[12] J. Hegedus, Y. Miche, A. Ilin and A. Lendasse, "Methodology for Behavioral-based Malware Analysis and Detection Using Random Projections and K-Nearest Neighbors Classifiers," 2011 Seventh International Conference on Computational Intelligence and Security, Sanya, China, 2011, pp. 1016-1023, doi: 10.1109/CIS.2011.227.

The research finishes by emphasising the suggested methodology's potential for usage in real-world applications such as antivirus software and intrusion detection systems. The authors also propose numerous future research directions, including researching alternate feature selection and dimensionality reduction strategies, as well as investigating the usage of ensemble classifiers to increase detection performance.

[13] S. Chaudhary and A. Garg, "A Machine Learning Technique to Detect Behavior Based Malware," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 655-659, doi: 10.1109/Confluence47617.2020.9058173.

The authors assess their technique using a dataset of system activity records. The collection is made up of activity logs from several workstations, some of which include malware and others do not. The authors categorise the activity logs using a range of machine learning methods, including decision trees, random forests, and support vector machines.

[14] T. -H. Nguyen and M. Yoo, "A behavior-based mobile malware detection model in software-defined networking," 2017 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2017, pp. 1-3, doi: 10.1109/ICISCT.2017.8188590.

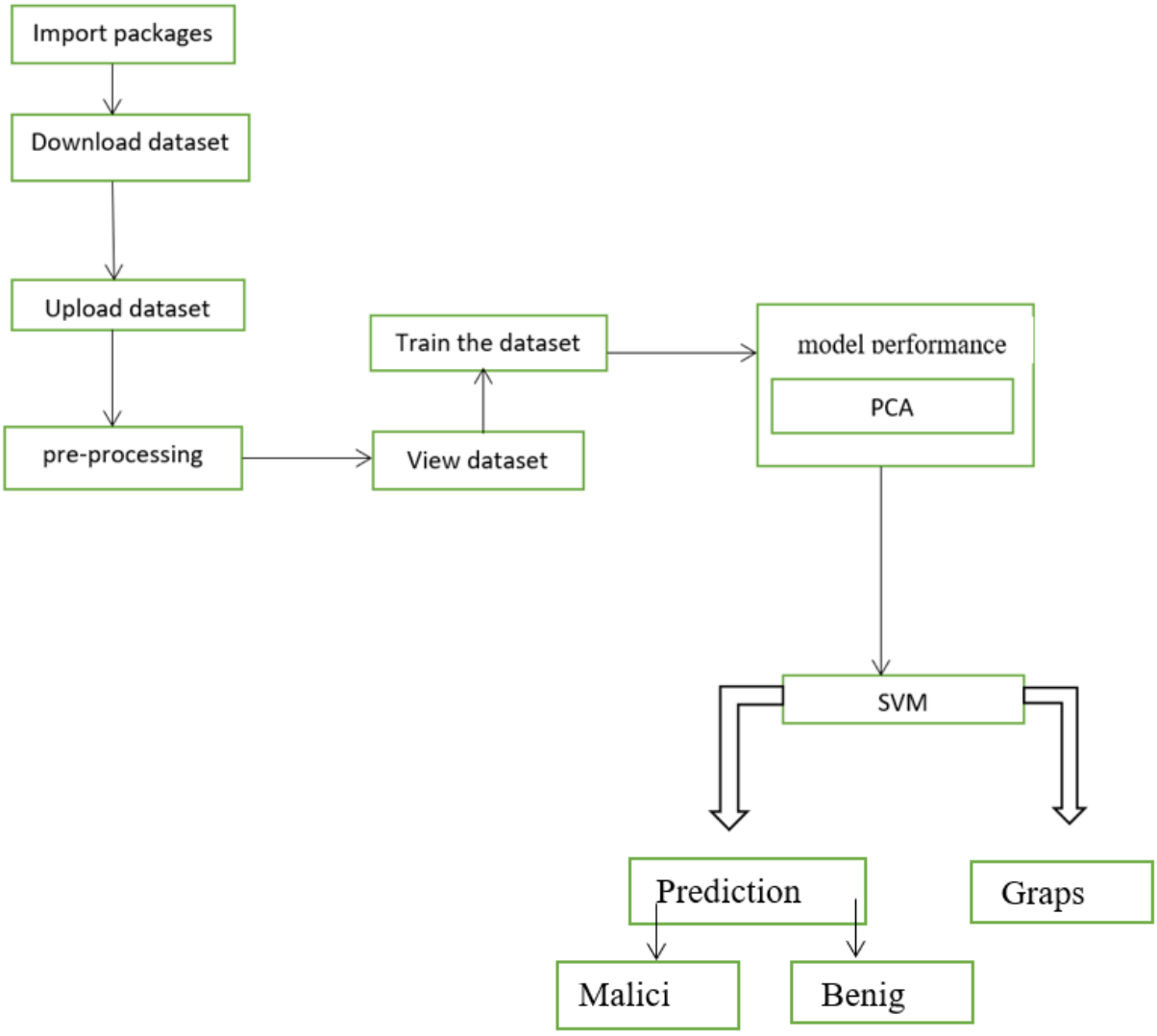
Using real-world malware samples, the authors performed tests to assess the effectiveness of their approach. The findings demonstrate that the suggested model may achieve high detection rates while producing few false positives. In addition, the authors compare their model to other cutting-edge algorithms and show that it beats them in terms of detection accuracy.

III. PROPOSED WORK WITH ARCHITECTURE

This section will describe the detailed description of the proposed work done for the detection of malware.
Dataset: - Different malware and benign that were previously detected by various sources were collected and subjected to feature extraction. On completion of the feature extraction a total of 77 features were generated which will be used for training the model.

Advantages:

1. Accuracy is maximum
2. Prediction is accurate



IV. NOVELTY

This section will describe the detailed description of the proposed work done for the detection of malware.
Behaviors of any file is defined as the task performed by that file when it implemented on a machine. Example of these behaviors can be payload persistence, Stealth Techniques, Environment Mapping etc. Behavior-based malware detection interprets an incoming file based on its predefined tasks and activities before it can implement that action. A file actions, or in some cases its potential-threat, is investigated for malicious activities.