International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Comparison of Malware Detection Techniques for Windows and Android Devices

Dr. Nachiyappan S
Assistant Professor (Sr.) CSE
Vellore Institute Of Technology,
Chennai Campus, India

Dr. Anusha K
Associate Professor CSE
Vellore Institute Of Technology,
Chennai Campus, India

Lakshay Grover
(18mca1003)
Vellore Institute Of Technology,
Chennai Campus
Chennai,India

***Abstract*—Malware refers to malicious software perpetrators dispatch to infect individual computers or an entire organization's network. It exploits target system vulnerabilities, such as a bug in legitimate software.A malware infection can cause many problems that affect daily operation and the long-term security of a system.Malware is not a single entity but it comes in various forms and each has it's own wayof entering and corrupting the systemfor example Virus, Worms, Trojans, Ransomware are various types of malware having different families. The aim of my research is to identify the existence of malware in Windows executable files and Mobile android devices in my first research the aim is to detect a malware in an android mobile device and in the second research the aim is to detect the malware in Windows executable files. This paper compares both the techniques and pros and cons of using each technique for malware detection**

I. INTRODUCTION

The aim of my research is to detect the malware which has entered into my device, detection is necessary as it is the first step before preventing malware from entering into the device. Before detecting the malware in my device its necessary to know what malware is and where it comes from and who is the one who do all this planning and make our Android devices malware prone. We can call malware as anything which enters into your personal device without your prior knowledge and which can steal or corrupt your data. It can steal all your personal information and use for unusual purposes for which you don't have any knowledge. The question here arises is how exactly malware propagates into your personal devices for which you are the only one who have access to this device according to your personal information. A malware can come in any form or any time, even sometimes you are the only one who invites malware to enter your device without having any knowledge that how it can exactly harm your device. Network such as internet is so popular these days that a malware can enter through the internet or the network so easily into your devices, download any application from internet can also make a path for the malware to enter into your device, spam email is one of the most common path which makes an entry for the malware into your devices. Malware can also enter infected removable devices or bundled with other software. Then another questions arises is that how can someone identify the malware, most of the time it is nearly impossible to identify the malware but you get some signals, identifying them can give you an idea that something is there which is unusual and it could be a malware for example, sudden appearance of pop-ups, A puzzling increase in data usage, Bogus changes on your bill, disappearing battery charge, People on the contact list reporting strange calls and texts from our phone, a phone that heats up while performing lags, surprise apps on phone, phone turns on wifi and internet connection on its own. Hence in my first research paper aim is to find the accuracy of malware in all the android mobile devices. If we talk about a windows system The first time when something like malware came into picture was in1986. It was developed by 2 brothers in Pakistan Basit and Ajmads. They replaced boot sector of floppy disk with a copy of virus.The real boot sector is moved to another sector, hence everytime you insert a floppy it would infect the drive. The intention of this malware to was solve several problems rather than giving harm to a system. The first malware originated with the intention of harming a system was Omega virus which affected the boot sector and damage one's system. Another malware in form of virus came into picture was Walker, once the virus enter your system it displays a man walking in your system after every 30 seconds and no user input is accepted till virus is removed from the system. Another virus similar to Walker was Ambulance virus in which an ambulance was running from one side of screen to another. Virus took a big form when we noticed the new and most dangerous Casino virus coming into picture.. After malware started conquering most of the systems, it was categorized into different types on the basis of how malware enters and attacks on a particular system.Advare is one such type of malware that automatically delivers advertisements. Sometimes we see pop-ups coming in form of ads, advare can add spyware in such ads for which if user opens that add it comes with some malicious code inside that add which could track user activity and steals information. Bots are another type of special programs used to perform some specific operations for examples they can be used in botnets for spreading of Ddos attacks. Bug is another flow produces as a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

result of undesired outcome. Bugs are usually result of problems in source code or an undesired outcome. It can cause crashing or freezing of systems. Another type of malware is Ransomware that holds a system under controlled by the attacker and it ask some amount in return to make system free to use, it holds the system under its control by encrypting the system files and demands some amount from to make the system free and regain their access to the computer. Rootkit is another type of malware that once installed in the system by the attacker it can access the system remotely from any other location, and hence it can access, modify, steal information easily, once installed it is very difficult to get prevented from Rootkit because the whole control moves to the attacker of one's system and it becomes very difficult to get rid of it. Another such dangerous malware just like rootkit which one should be aware of is Trojan horse. It looks like a normal file on internet and tricks user of download the file but contains malware in actual for which user is not aware of, once downloaded it takes all system's privileges and it can destroy ones system. Malware spread in a mobile device or in a PC, while the type of malware being the same hence, the first approach is detection of malware in a Android mobile device and it checks the accuracy of malware in an Android device using different machine learning algorithms while in the second approach I detected malware in a windows system using deep learning ensemble.

II. BACKGROUND AND ARCHITECTURE

The architecture of first system is generated by studying algorithm and how the system behaves when the algorithm is passed .The output of different algorithms is compared and the algorithm giving the most efficient result is selected. The first step towards detecting malware from my Android Device is collected dataset. In an Android device whenever any application is run a log data is generated. Log data is between a system and the users of that system,[2] or it is a data collection method that automatically captures the type, content, or time of transactions made by a person from a terminal with that system. Using Wire Shark application in my android device I generated a system log file for my android devices which tell me the behavior and nature of applications in my system. It tells me that when a user access any application in the system, how the applications responds. Using wireshark application, we can derive the systems logs, the output is data set having two columns as sequence and target. The sequence derived is our log data binary and target is either 0 or 1 which, 1 describes that the application is executed successfully and 0 shows that either the application is crashed or not executed properly. The architecture is shown as below.

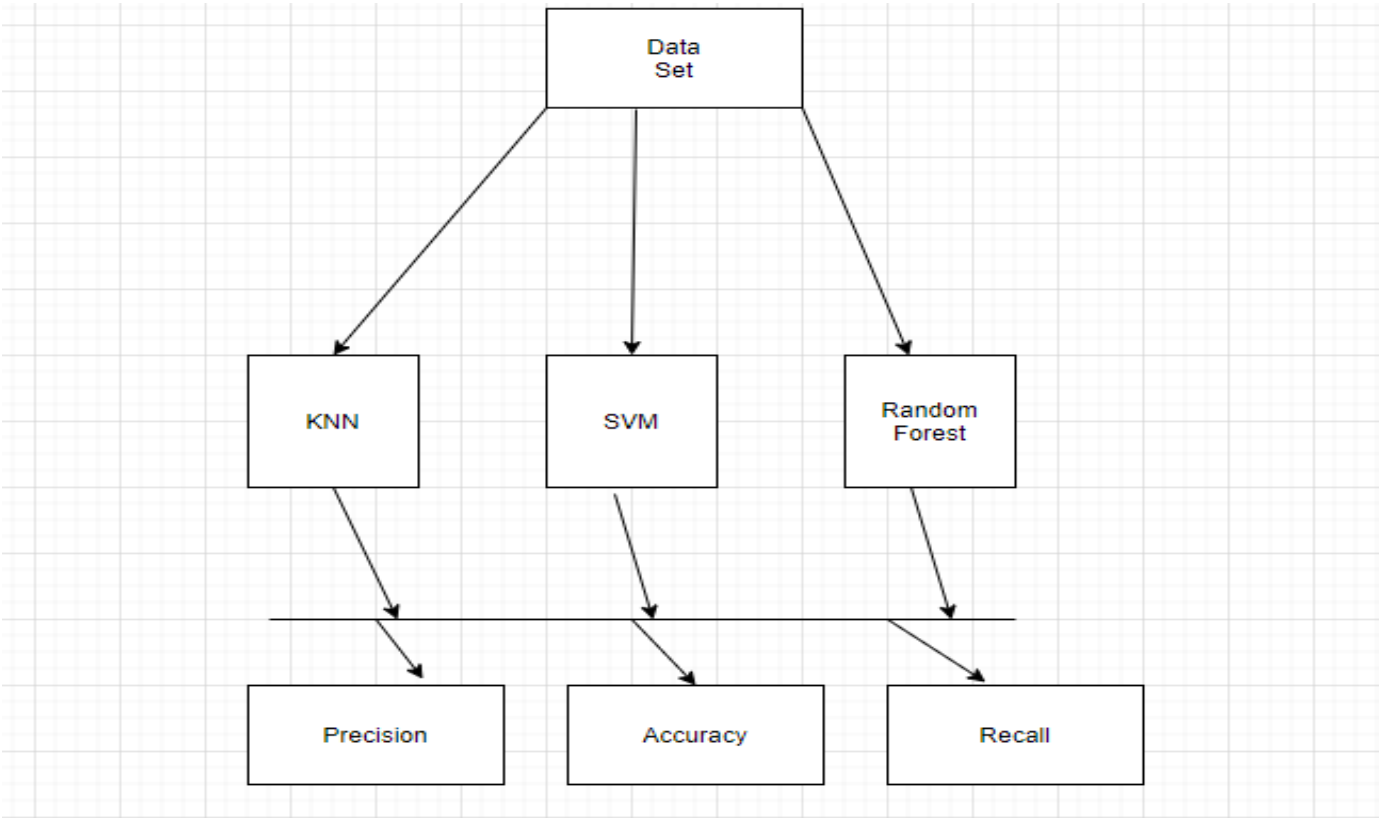

Figure 1

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

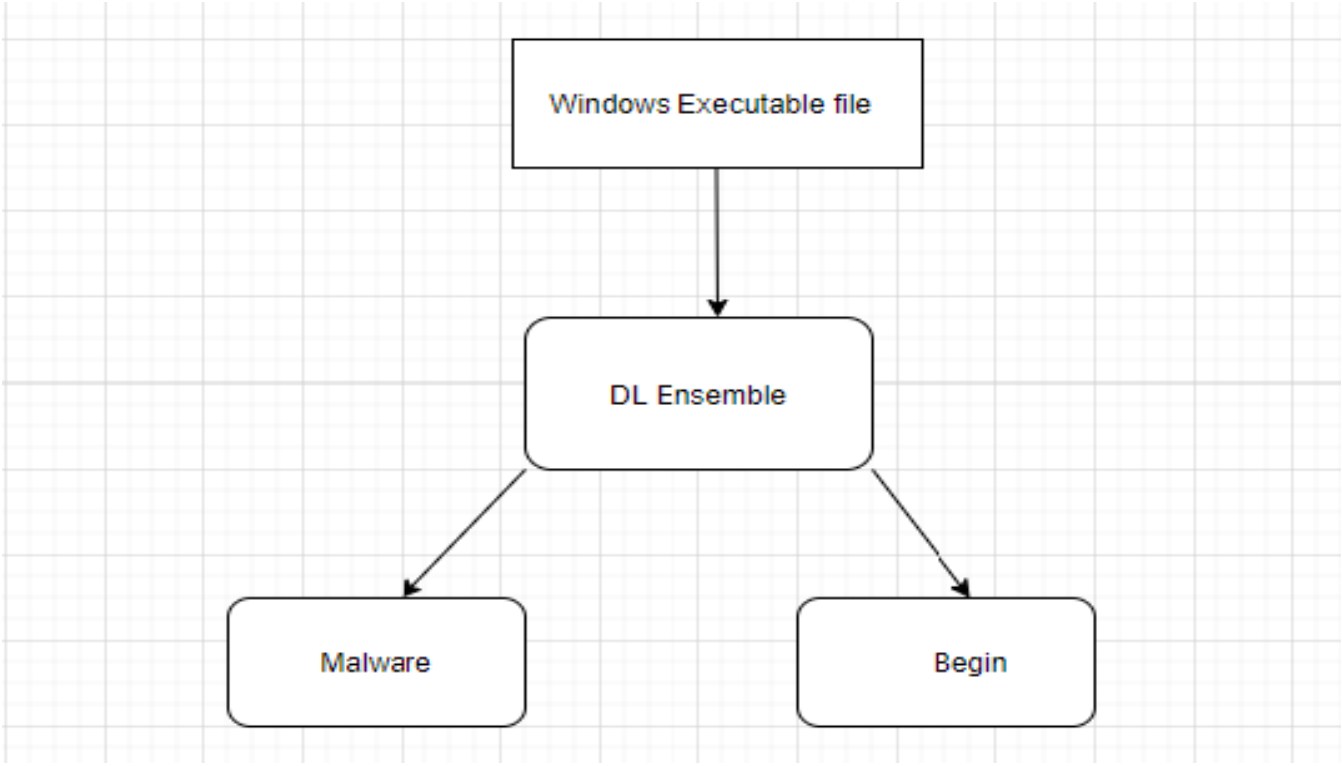

Figure 2

In first algorithm Machine learning algorithms are used on the data set while in the second algorithm DL ensemble learning is performed by building modules. The first algorithm intends to find the accuracy of malware in an android device while the next architecture is intended to classify the windows executable file as malware or begin.

III. PROPOSED WORK

In the first research machine learning algorithms were implemented in a data set which was collected using packet tracer application available in windows and different machine learning algorithms were applied to the data set inorder to generate accuracy,precision, recall,f1-support for all the algorithms in order to find the algorithm having the highest accuracy and giving the best results.

The following algorithms are applied to the data set in order to get efficient results from the given data set:

Naïve Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

With relation to our data set it could be understood as:

No pair of features in our data set is dependent on any cause. Packets transferred from our system to the device all are independent with each other and data transmitted is also independent.

Secondly, each part of data set extracted is given the same weight or importance.

After executing on the data set following output is retrieved.

Table 1

	precision	recall	f1-score	support
0	0.97	0.97	0.97	104
1	0.97	0.97	0.97	97
micro avg	0.97	0.97	0.97	201
macro avg	0.97	0.97	0.97	201
weighted avg	0.97	0.97	0.97	201

Precision attempts to answer the following question:What proportion of positive identifications was actually correct?0 indicates positive identifiers and 1 indicates negative identifiers. If the outcome is 0 it indicates the occurrence of malware which is not present in the device and It indicates the percentage of malware present in the device.Recall attempts to answer the following question:What proportion of actual positives was identified correctly?F1 is an overall measure of a model's accuracy that combines precision andrecall.The support of a rule indicates how frequently the items in the rule occur together. Average is calculated to get more accurate results..

K-nearest neighbour: To evaluate any technique we generally look at 3 important aspects:Ease to interpret output,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Calculation time, Predictive Power, the learning is based "how similar" is a data (a vector) from other. After executing the algorithm following outcomes were retrieved:

Table 2

	precision	recall	f1-score	support
0	0.94	0.99	0.97	98
1	0.99	0.94	0.97	103
micro avg	0.97	0.97	0.97	201
macro avg	0.97	0.97	0.97	201
weighted avg	0.97	0.97	0.97	201

The percentage of malware is less as compared to percentage of begin, after execution of knn classifier. It gives more accurate results than Naïve Bayes classifier since difference between malware and begin apps is more than other classifier. Confusion matrix for above algorithm is above:

SVM: SVM is a machine learning algorithm used for classification and regression analysis. SVM analyze the large amount of data to identify the patterns from the data set. By using SVM algorithm we have generated the parallel partitions by generating two parallel lines. SVM is used to design a hyperplane that classifies all training vectors in two classes

Random Forest: Random Forest is a supervised learning algorithm used for classification and regression. It is most flexible and easy to use algorithm. Random forest algorithm creates the decision trees on randomly selected data sample and then it gets the prediction from each tree and selects the best solution. It has various applications. It can be used to identify and predict student performance as we are doing in this paper.

As the above algorithms were used in the first research in the second research different DL models were build and the output was predicted based on the DL models.The first model was trained on a numerical vector representation of words extracted from the rawbytes of the PE files. Words were obtained by decoding every byte of a file as a utf-8 encoded character, regardless of its intended encoding, and combining characters to form words, delimiting words with whitespace characters, such as spaces and tabs.

Opcode Model-The PE file package was used to find the entry point of the code. The frequency of opcodes was found out to find the top 50 most frequent opcodes.The principal model was prepared on a numerical vector port of words taken from the rawbytes of PE files.

The subsequent model was trained on a far more modest number of features which is a likelihood presence of opcodes inside the assembly code of filetabs.The frequency of every one of these main 50 opcodes was estimated and standardized into the likelihood of it's appearance by separating it by total number of opcodes.

III. CONCLUSION

The conclusion I draw from the above research and reading my literature survey papers is that how can I identify a system being Malware prone or not. Before identifying a malware in an application its necessary to differentiate an application, weather it is a malware or a begin app. Identification of an application is necessary in this sense because our aim is to detect weather the malware exist or not and begin apps are those apps which are totally malware free apps, and finally different classification algorithms are applied to my data set and results are compared to calculate the accuracy, precision, f1-score and support for each of the algorithms.

IV. REFERENCES

[1] Ki-Hyeon Kim, Mi-Jung Choi*,"Android Malware DetectionusingMultivariate Time-Series Technique"
[2] MayankJaiswal, Yasir Malik, FehmiJaafar, "Android Gaming Malware Detection Using System Call Analysis"
[3] XIONG Ping1, WANG Xiaofeng2,5, NIU Wenjia3, ZHU Tianqing4, LI Gang, **"**Android Malware Detection with Contrasting Permission Patterns"
[4] SHAILA SHARMEEN1, HUDA 1, (Member, IEEE), JEMAL H. ABAWAJY1, (Senior Member, IEEE), WALAA NAG ISMAIL2, AND MOHAMMAD MEHEDI HASSAN 2, (Member, IEEE),"Malware Threats and Detection for Industrial Mobile-IoT Networks"