JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MALWARE DETECTION USING MACHINE LEARNING

Chandraleka Alluri[1], Pulusuganti Sunny Manohar[1], Tarun Tej Nimmaka[1], Chirala Sai Mani Venkata Jitendara Reddy[1], Ms.Sapna[2]

Students, Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India-530045[1]

Assistant Professor, Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India[2]

Abstract: Nowadays, security has become the key element of enterprise-level applications, It is the developer's responsibility to check the threats and malicious activities that need to be verified and cross-checked at each and every level of data transfer. Intruders are trying in many ways to access the application data by spoofing, sniffing, and other most popular network-based attacks. This proposal implements analyzing a data set and checks with the transaction data processed across server to client. Different machined learning algorithms are implemented in order to check the malware attacks across the network transmission

Keywords: Machine Learning, Support vector machine(SVM),k-nearest neighbor(KNN),BR-ANN, confusion matrix,Tkinter

I. INTRODUCTION

The breakthrough in internet technology and computer networking have made high-speed shared internet possible. The effect of this development is the daily increase in the number of computer systems that have become susceptible to malware attacks. The innovation has made the internet a huge storehouse where resources are virtualized and utilized to the needs of users. Despite the immense benefits that the internet revolution has brought, there are numerous challenges that it also poses to the security of computer systems. The conventional computer system is entirely centered on a single host machine running the operating system, while several machines connected to the host are running on the guest operating system. The prevalent security threat confronting the users is the attack on a computer system by malicious programs that spread to other computers that have not been infected. The threat posed by malware infections has become a major challenge in the field of computer security over the years. The number of new malware on the internet keeps on increasing at an alarming rate even as anti-virus companies are making an effort to curtail the trend so as to make the vast number of computer user safe. Malware has evolved over time and is becoming more sophisticated than before. It is now more difficult to detect them. There is, therefore, the need to invent more efficient techniques that can detect and prevent these attacks. Malware is a malicious program that infringes on the security of a computer system in terms of privacy, reliability, and accessibility of data. This trend has made academicians and industry practitioners move from conventional static detection techniques to more dynamic, sophisticated, and spontaneous methods that apply accumulated malware behavior to detect malware attacks. A malware can simply be defined as a malicious program that the user unsuspectingly installs on their machine, and later, these programs can begin to disrupt the proper operation of the machine or might continue unnoticed and carry out malicious actions without being noticed. When the attacker gains control of the machine, he can then have access to any information stored on the machine. Some of the deceptive approaches used to install malware on the computer system through the internet include repackaging the software, update attacks or desire for download. The attacker employs any of the methods mentioned before to create malicious software by inserting a certain type of malware into it before uploading it to the internet. Malware can be described as various types of software that have the capacity to wreak havoc on a computer system or illegally make use of this information without the consent of the users. Malware can be categorized into various types, for instance, Botnet, Backdoor, Ransomware, Rootkits,

Virus, Worms, and Trojan Horse, Spyware, Adware, Scareware, and Trapdoor. They are used to attack computer systems and for performing criminal activities such as scams, phishing, service misuse, and root access.

II. LITERATURE SURVEY

Title: Evaluating Machine Learning Classifiers to detect Android Malware

AUTHORS: Prerna Agrawal, Bhushan Trivedi

ABSTRACT: Malware Detection using conventional methods is incompetent to detect new and generic malware. For the investigation of a variety of malware, there were no ready-made machine-learning datasets available for malware detection. So, we generated our dataset by downloading a variety of malware files from the world's famous malware projects. By performing unstructured data collection from the downloaded APK files and feature mining process, the final dataset was generated with 16300 records and a total of 215 features. There was a need to evaluate the performance of the generated dataset with supervised machine learning classifiers. So, in this paper, we propose a malware detection approach using different supervised machine learning classifiers. Here, supervised algorithms, Feature Reduction Techniques, and Ensembling techniques are used to evaluate the performance of the generated dataset. Machine Learning classifiers are evaluated on the evaluation parameters like AUC, FPR, TPR, Cohen Kappa Score, Precision, and Accuracy. We also represented the results of classifiers using Bar plots of Accuracy and plotting the ROC curve. From the results of machine learning classifiers, the performance of the CatBoost Classifier is highest with an Accuracy of 93.15%, having a value of ROC curve as 0.91 and Cohen Kappa Score as 81.56%.

Title: A Static Malware Detection System Using Data Mining Methods

AUTHORS: Usukhbayar Baldangombo, Nyamjav Jambaljav, and Shi-Jinn Horng

ABSTRACT: A serious threat today is malicious executables. It is designed to damage computer systems and some of them spread over the network without the knowledge of the owner using the system. Two approaches have been derived for it i.e. Signature Based Detection and Heuristic Based Detection. These approaches performed well against known malicious programs but could not catch the new malicious programs. Different researchers have proposed methods using data mining and machine learning to detect new malicious programs. The method based on data mining and machine learning has shown good results compared to other approaches. This work presents a static malware detection system using data mining techniques such as Information Gain, Principal component analysis, and three classifiers: SVM, J48, and Naive Bayes. To overcome the lack of usual anti-virus products, we use methods of static analysis to extract valuable features of Windows PE file. We extract raw features of Windows executables, which are PE header information, DLLs, and API functions inside each DLL of Windows PE file. Thereafter, Information Gain, and calling frequencies of the raw features are calculated to select valuable subset features, and then Principal Component Analysis is used for dimensionality reduction of the selected features. By adopting the concepts of machine learning and data mining, we construct a static malware detection system that has a detection rate of 99.6%.

iii. Proposed methodology

3.1 Dataset

The first step was collection of dataset which we had obtained from many websites like github,kaggle etc.we had combined many forms inorder to get the best dataset and the dataset which we obtained was a huge data which consists of more than 20 malware attacks

Table 3.1 Feature Information of Dataset

3.2 Data Pre-Processing

In the data pre-processing step, we made many changes in the dataset, which included the deletion of duplicate elements or rows in order to remove these, we used data preprocessing and for filling the null values, we used mean, mode and median to fill the null values.

3.3 Algorithms

Random Forest Classifier:

It is a classification technique that contains a collection of decision trees on the various subsets of datasets. Greater number of trees will perform with higher accuracy, and the overfitting problem is overcome. As the name suggests, multiple number of decision trees are used for classification. Hence, it is called the ensemble method.

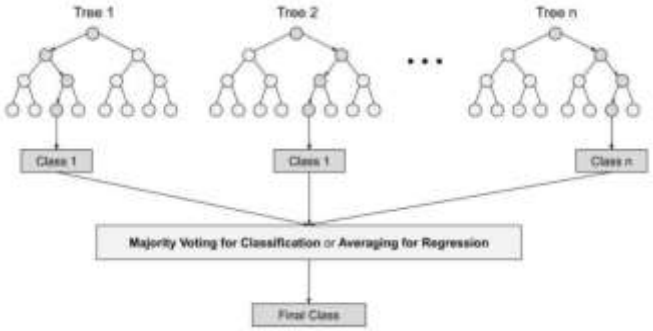


Fig 3.3.1 Random Forest Classifier

Support Vector Classifier:

Its main goal is to create a plane that segregates into classes and sub-points, which creates categories. The best decision boundary is called a hyperplane. There are two types of Support vector classifiers. They are Linear SVC, Non-Linear SVC. Linear SVC can be used for linearly separable data, whereas Non-linear SVC is used for non-linearly separable data.

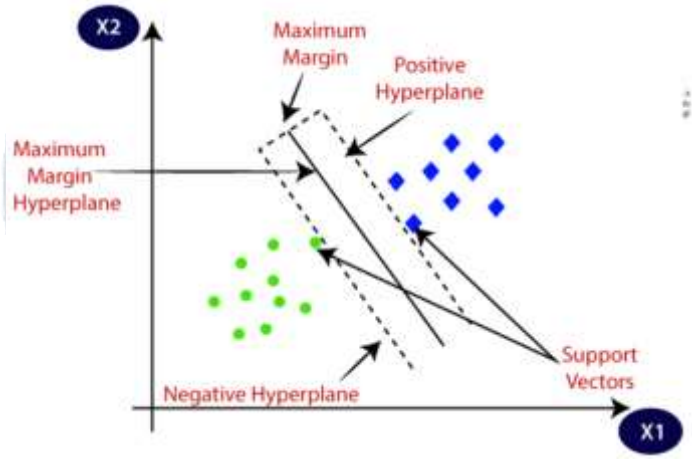


Fig 3.3.2 Support Vector Classifier Hyperplane

K-Nearest Neighbor:

The number of nearest neighbors to new input has to be classified by symbol k. It can be used as both classification and regression. It is a nonparametric algorithm, a lazy learning algorithm.

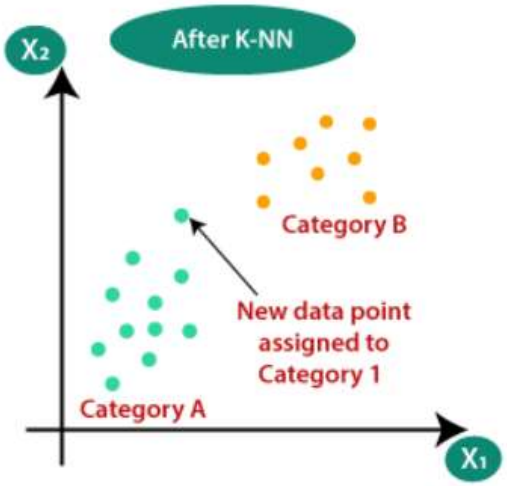


Fig 3.3.3 K-Nearest Neighbor

Decision Tree Classifier:

It is a tree-structured classification, where branches and leaf nodes are present for classification and regression. Branches are decision nodes, leaf nodes are the output of the decisions. There are various terminologies such as root node, branches or subtrees, pruning, parent node, and child node. The information gain entropy of each node is calculated. The node which has the highest information gain will be split first.

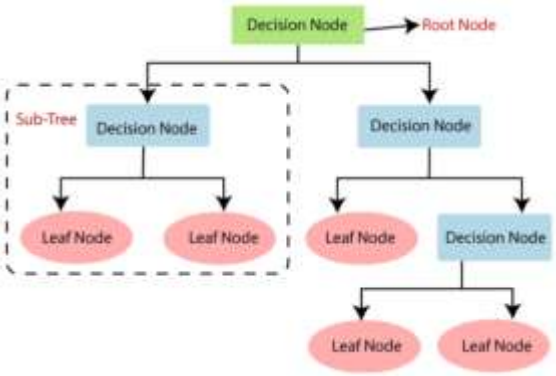


Fig 3.3.4 Decision Tree Classifier

The model is split into a training set and a testing set with a test size of 0.2 and a random state of 2 with a stratified function for sampling the dataset. It divides the dataset into the same proportion called strata so that the dataset will be divided into the same proportion. After splitting the dataset, it is standardized using a standard scaler. Thereafter, classification was performed with Decision Tree Classifier (DT), Random Forest Classifier, Support Vector Classifier (SVM), Logistic Regression (LR), Gaussian Naïve Bayes Classifier (GNB), K-Nearest Neighbours (KNN).

3.4 Workflow of the system

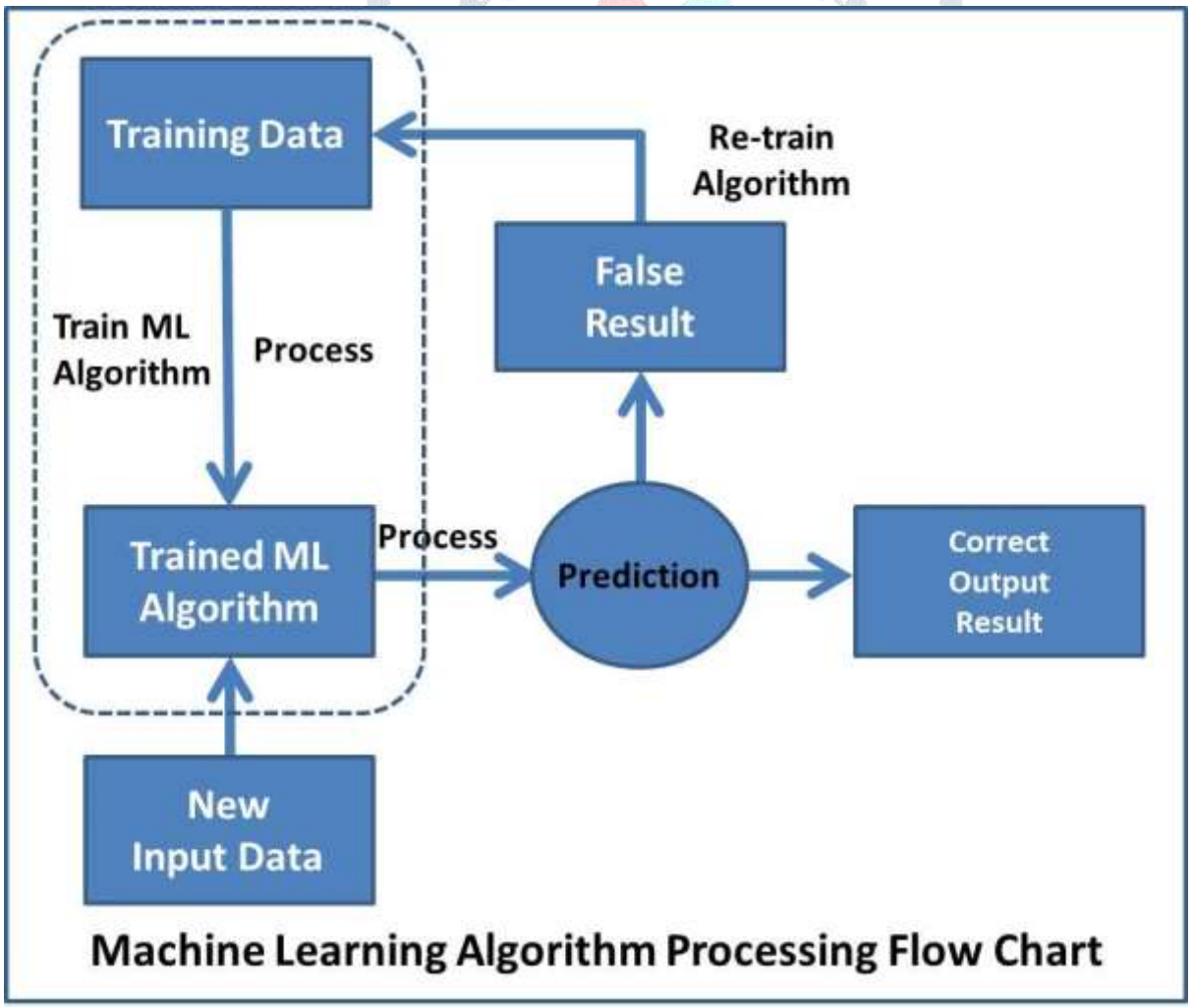
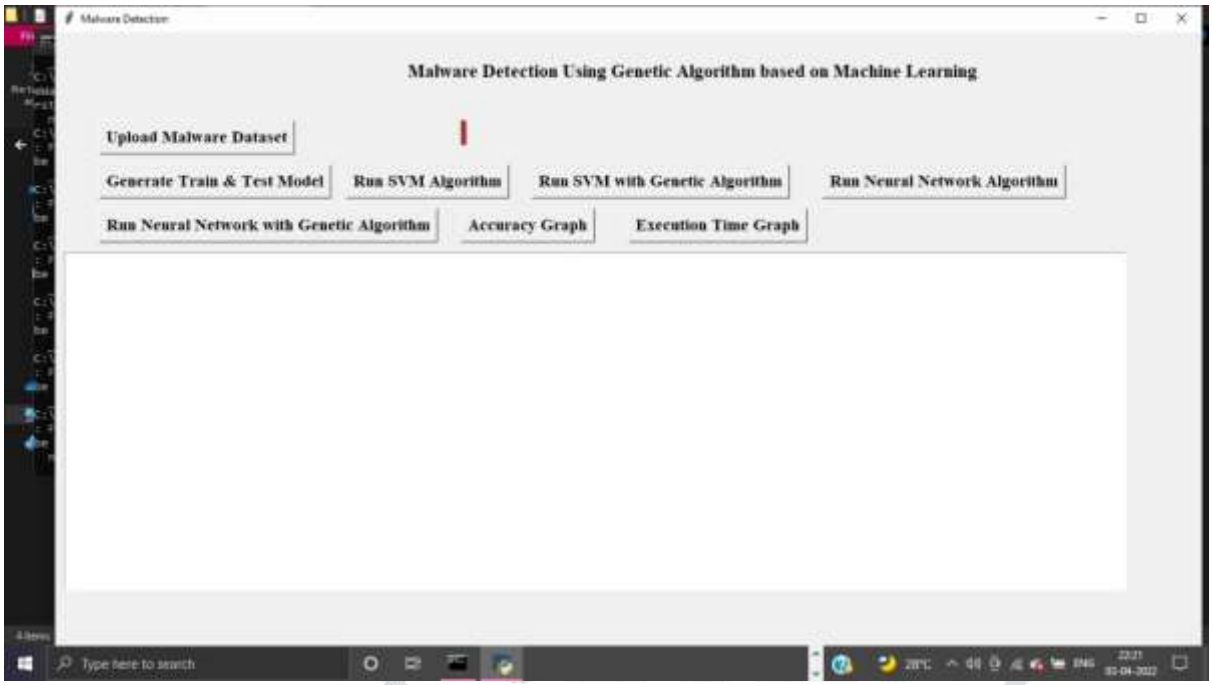


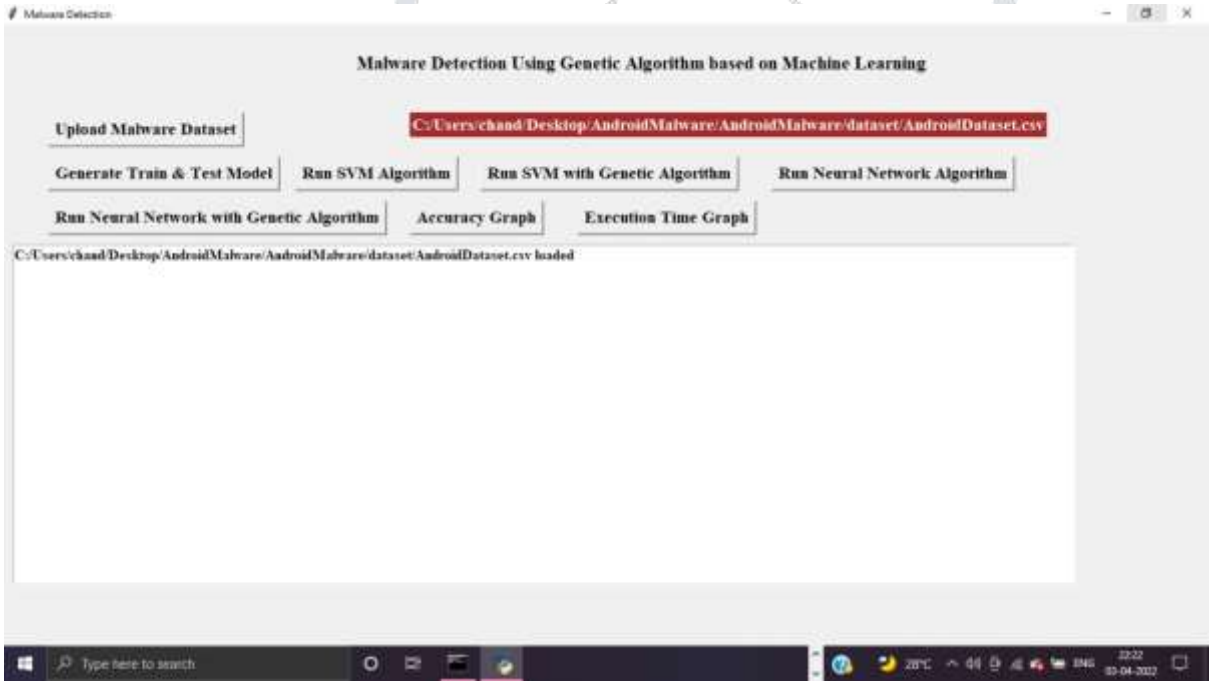
Fig 3.4 Workflow of the proposed methodology

IV. RESULTS AND DISCUSSION

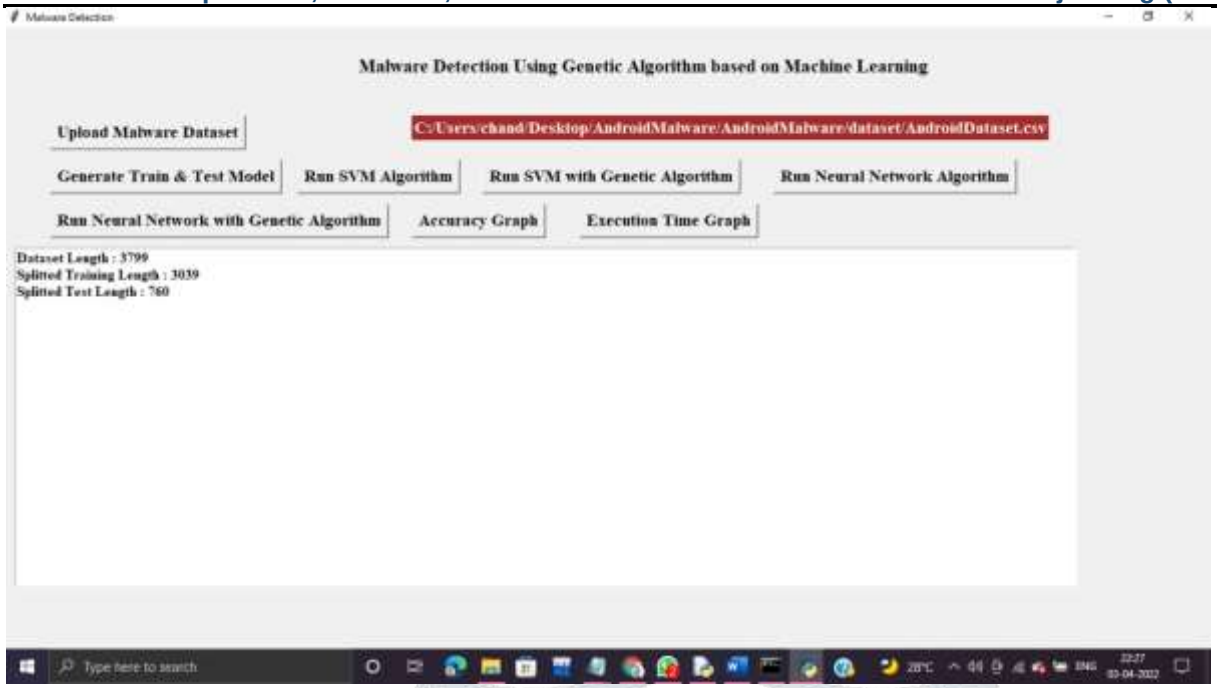
4.2.1After we run the code we will get the display of the page:



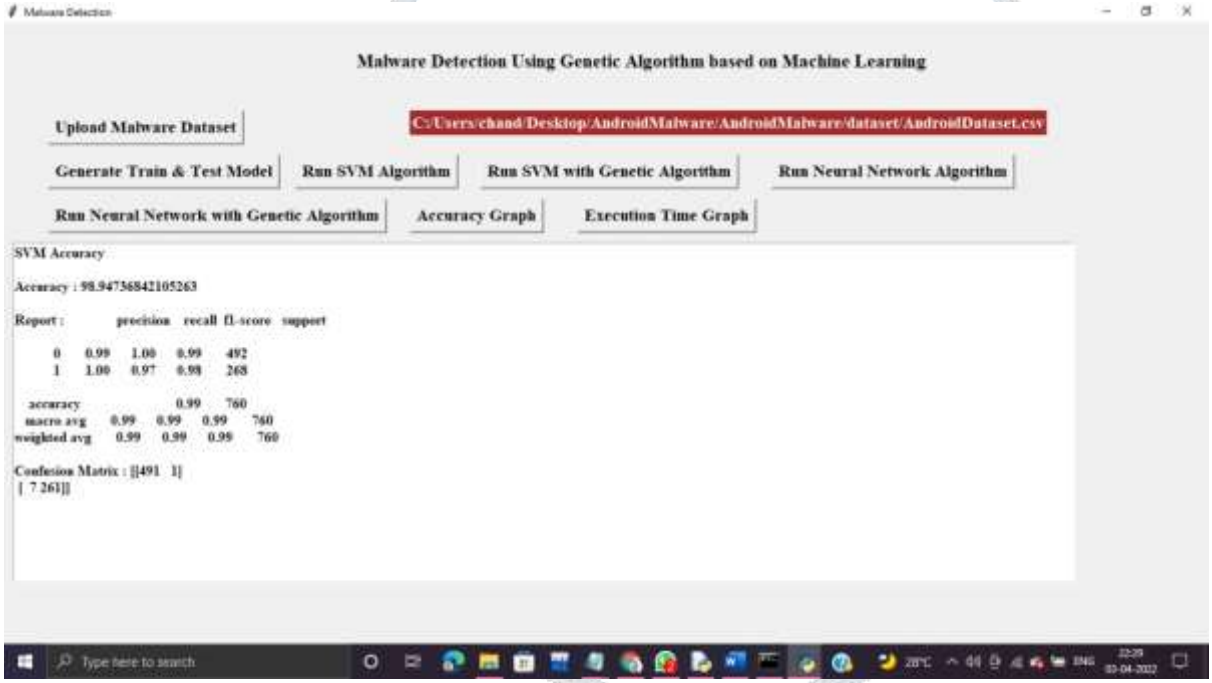
4.2.2 Then click on "upload malware dataset" to give the input:



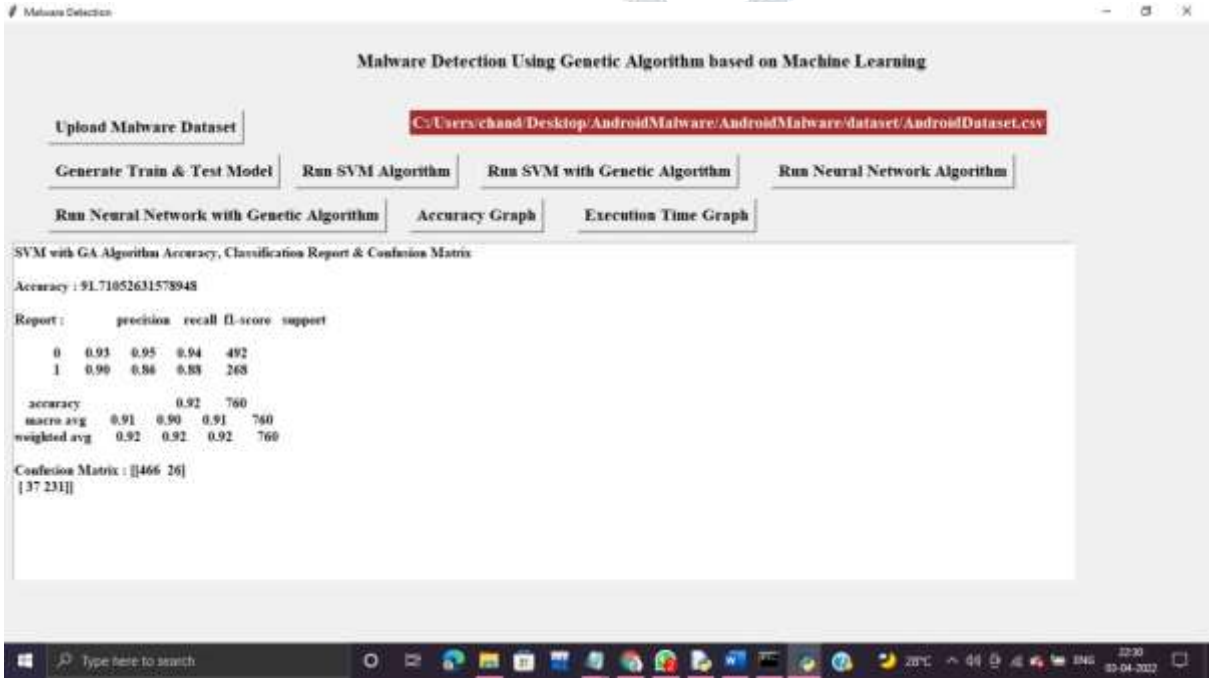
Next, we 4.2.3. can click on "Generate Train and Test Model"



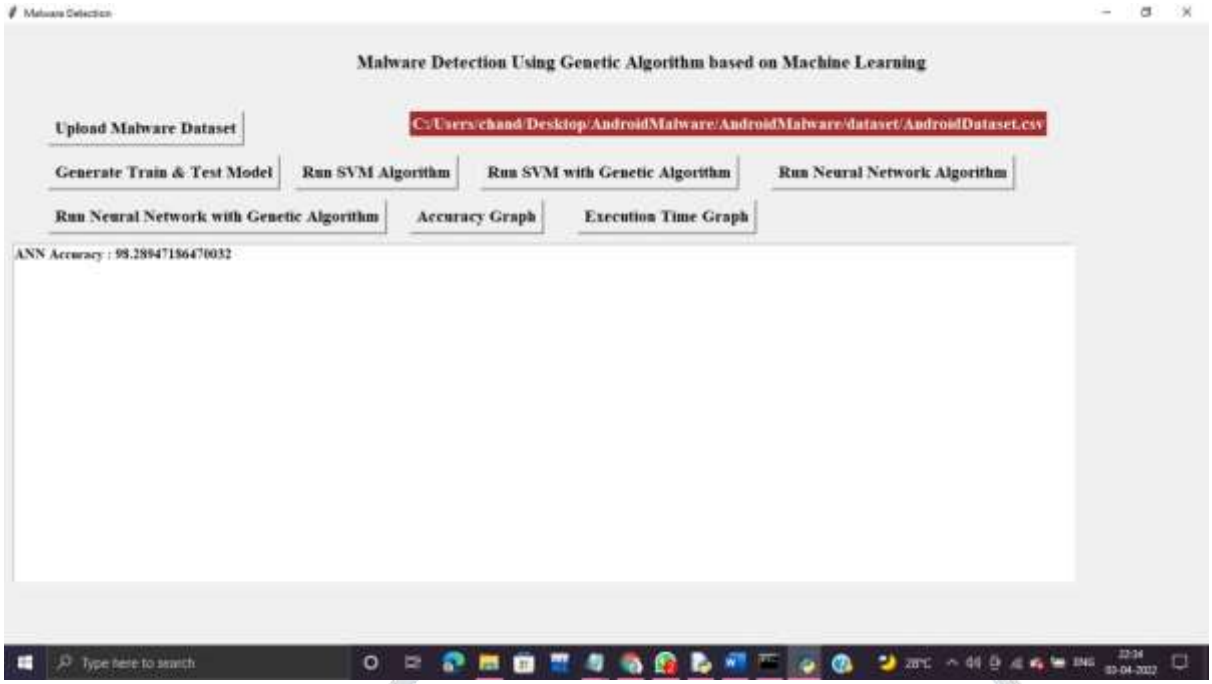
4.2.4. Click on “Run SVM Algorithm”:



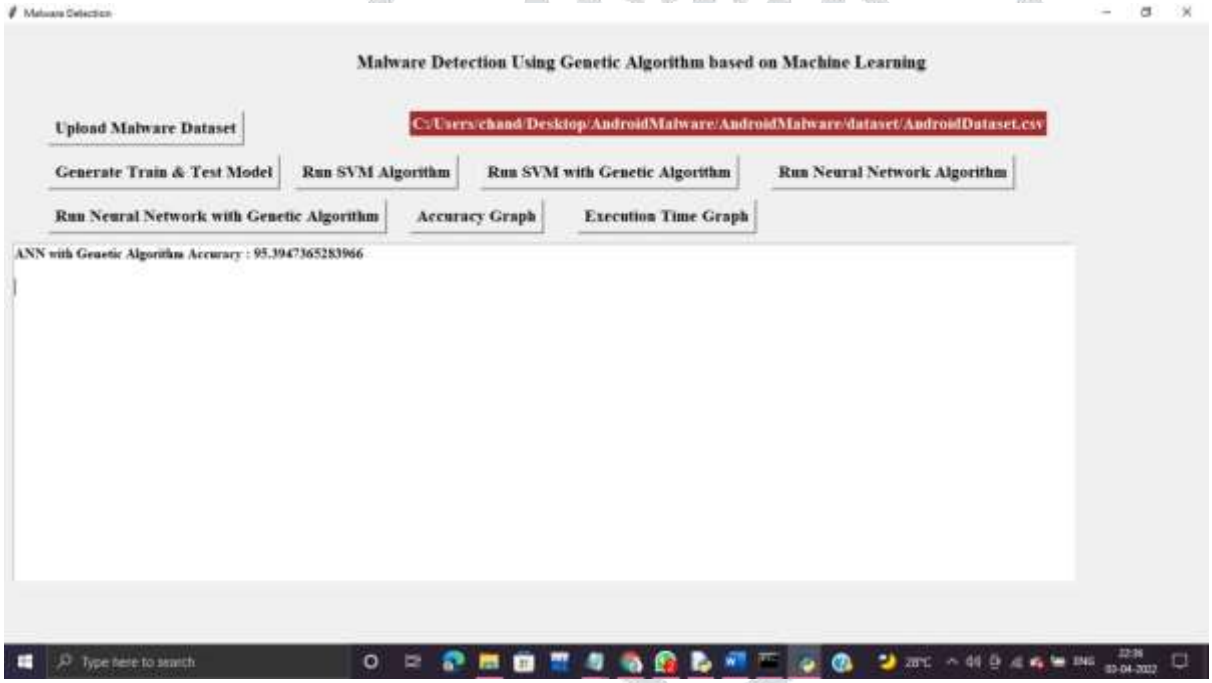
4.2.5. Click on “Run SVM with Genetic Algorithm”:



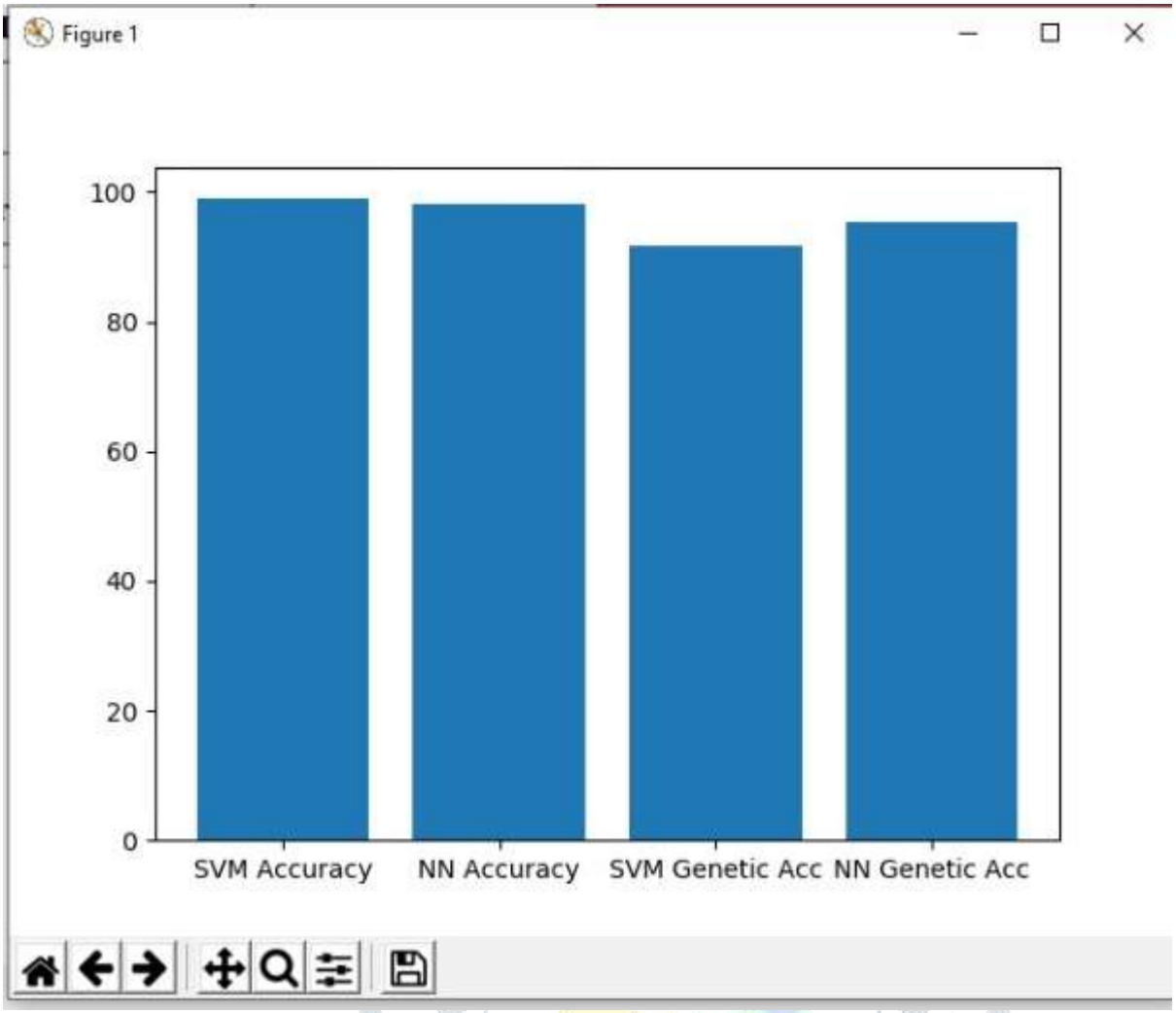
4.2.6. Click on “Run Neural Network Algorithm”:



4.2.7. Click on “Run Neural Network with Genetic Algorithm”:

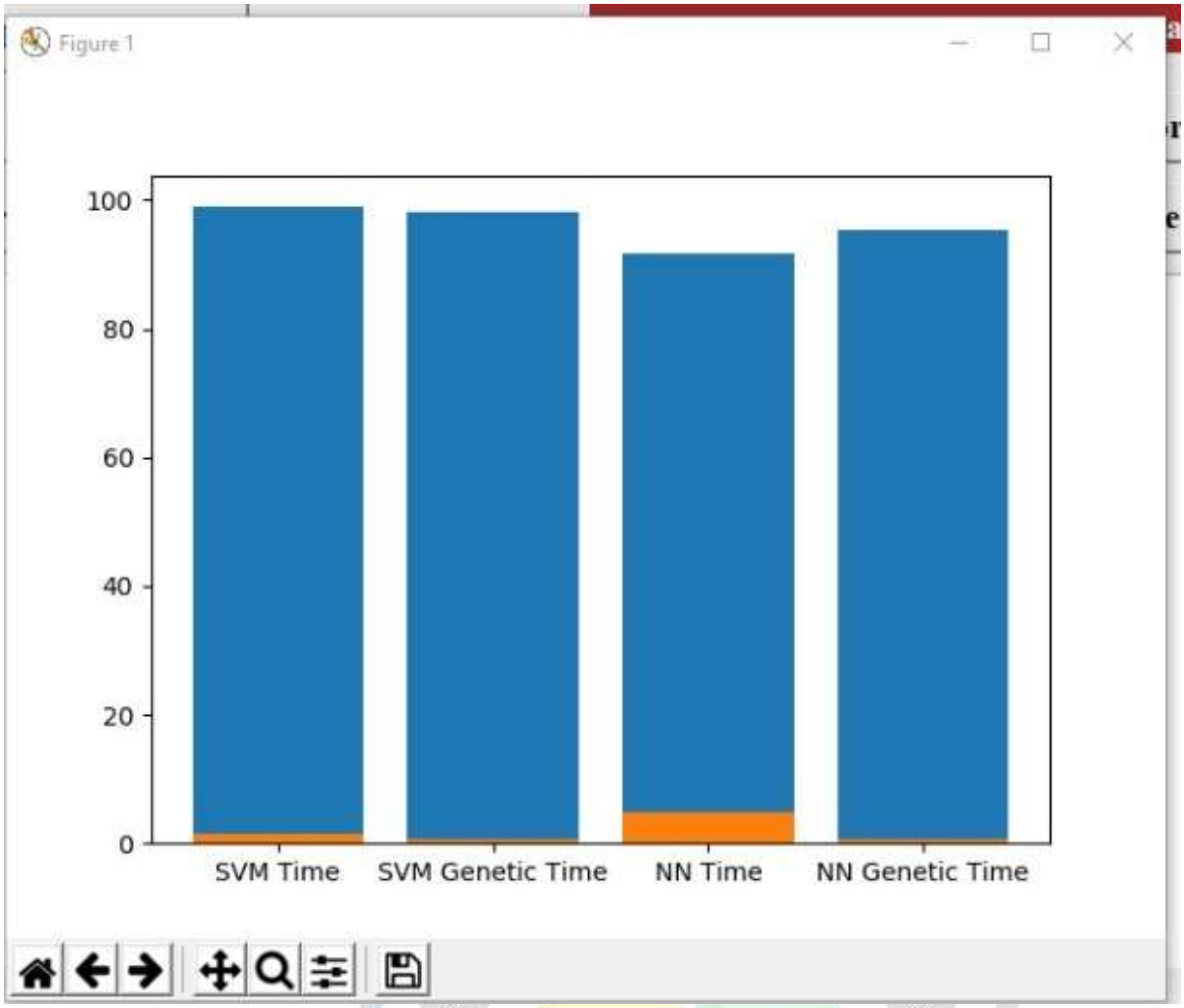


4.2.8. ACCURACY GRAPH:



4.2.9. Click on "Execution Graph"

Fitting the model:



V. CONCLUSION

In this model, we got an accuracy of 98.9% in the SVM algorithm and an accuracy of 92.5% by using SVM with the Genetic algorithm, 98.8%% with the Neural Network Algorithm and 94.2% with the Neural Network with the Generic Algorithm.

Future scope:

In the future process, detecting malware is not only the process of providing good security for the devices. In order to maintain a level of security in the present situation, the user or the device needs to maintain a special level of topologies and make sure that they are using a good security level band with not every malware can be detected by the algorithm itself some can enter through the firewalls and the wireless network of the specified user. Therefore, detecting the strength and security of all networks can lead to security for the devices.

REFERENCES

[1] Baban U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare and Manisha Darade, "Translation Of Languages using Deep Learning", Internation Journal of Advanced Research in Science", Communication and Technology, 2021

[2] C. Beluah Christalin Latha and S. Carolin Jeeva, "Translation of Languages Using Machine RNN", Informatics in Medicine Unlocked, 2019

[3] Karna Vishnu Vardhan Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Shiva Jothi Paramasivam, Hui Na Chua, S.Pranavanand, "Translation Of Languages Using NLP", MDPI,2021