IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

Detection of Android Malware Using Genetic Algorithm based Optimized Feature Selection

Bharathi P[1], Karthik M[2], Suma Sri M[3], Sai Manoj K[4], Tapaswini K[5]
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4.5]
Raghu Institute of Technology, Visakhapatnam, AP, India

Abstract: *Android platform due to open source characteristic and Google backing has the largest global market share. Being the world's most popular operating system, it has drawn the attention of cyber criminals operating particularly through wide distribution of malicious applications. This paper proposes an effectual machine-learning based approach for Android Malware Detection making use of evolutionary Genetic algorithm for discriminatory feature selection. Selected features from Genetic algorithm are used to train machine learning classifiers and their capability in identification of Malware before and after feature selection is compared. The experimentation results validate that Genetic algorithm gives most optimized feature subset helping in reduction of feature dimension to less than half of the original feature-set. For machine learning classifiers, a classification accuracy of over 94% can be maintained with a large feature reduction, thereby improving classification accuracy computational complexity of learning classifiers.*

Keywords: Android Malware

I. INTRODUCTION

Android Apps are uninhibitedly accessible on Google Playstore, the official Android application store just as outsider application stores for clients to download. Because of its open source nature and fame, malware scholars are progressively zeroing in on creating malignant applications for Android working framework. However, despite numerous efforts from Google Playstore to prevent pernicious applications, they actually discover their approach to mass market and cause harm to clients by abusing personal data associated with their telephone directory, mail accounts, GPS locations and other data for outsiders to abuse, or more likely to assume responsibility for their telephones remotely. Consequently, it is necessary to perform malware detection on such harmful applications which are a genuine threat to Android. Android Malware investigation is of two types: Static Analysis and Dynamic Analysis. Static investigation essentially includes breaking down the code structure without executing it while dynamic examination will be assessment of the runtime conduct of Android Apps in obliged climate. Yielded to the ever-expanding variations of Android Malware presenting zero-day dangers, an effective system for recognition of Android malwares is required. Instead of the signature-based method, which requires updates to the mark information base on a regular basis.

II. LITERATURE SURVEY

2.1 D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "Drebin: Effective and Explainable Detection of Android Malware in Your Pocket," in Proceedings 2014 Network and Distributed System Security Symposium, 2014.
The Android platform is at risk of being hacked by malicious applications. Traditional protections can no longer keep up with the number and variety of these applications, so Android devices often remain unprotected from novel malware. We present DREBIN, a lightweight technique for identifying Android malware that enables legitimate recognition of pernicious applications on cell phones. DREBIN plays out a comprehensive static examination of an application in light of the restricted assets that prevent application testing at runtime. By installing these features in a joint vector space, common malware examples can be detected more naturally. distinguished and utilized for clarifying the choices of our strategy. In an assessment with 123,670 applications and 5,580 malware tests DREBIN outflanks a few related methodologies and identifies 94.01% of the malware with bogus cautions. Where the clarifications

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

accommodated every recognition uncover pertinent properties of the recognized malware. Approximately ten seconds are required for an examination on five popular cell phones, making it appropriate for verifying the legitimacy of the downloaded applications.

2.2 N. Milosevic, A. Dehghantanha, and K. K. R. Choo, "AI supported malware android characterization," Comput. Electr. Eng., vol. 61, pp. 266–274, 2017.
It has been quite a while since malware has been used to lead digital assaults. The large number of mobile devices, which store secret and private data, has made them a key target for malware engineers. In spite of the fact that Android is the most popular portable working framework, malware engineers continue to discover Android to be an intriguing stage for contaminating weak clients. Thus, manual malware examination is a nigh on impossible endeavor because of the huge amount of Android malware species available. Malware criminological agents could benefit from AI strategies in their fight against pernicious projects. A pack of words portrayal model is used in both approaches to examine the source code for flexible applications: one based on consent, and the other based on source code analysis. Source code characterization achieved 95.1% F-score, while method based on consent names only achieved 89% F-score. In order to reduce the time required for cell phone malware analysis, we use our methodology to analyze static code and recognize malware with high precision.

2.3 J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-An, and H. Ye, "Critical Permission Identification for Machine-Learning-Based Android Malware Detection," IEEE Trans. Ind. Informatics, vol. 14, no. 7, pp. 3216–3225, 2018.
Android is the most generally utilized versatile working framework (OS). The number of Android (application) markets outside of Google has increased dramatically. The nonattendance of outsider market guideline has incited research organizations to propose distinctive malware recognition methods. It is hard to plan a security recognition technique that can productively and adequately recognize malicious applications for quite a while. Then, embracing more highlights will build the unpredictability of the model and the computational expense of the framework. Consents assume a crucial role in the security of Android applications. Term Frequency—Inverse Document Frequency (TF-IDF) is utilized to survey the significance of a word for a record set in a corpus. The static examination technique doesn't have to run the application. The paper proposes a new static location technique for extracting authorizations from Android application packages using TF-IDF and Machine Learning. The authorization value (PV) of each extracted permission is computed using the TF-IDF algorithm, and the application's sensitivity value (SVOA) is also calculated. The SVOA and the number of permissions used in the application are learned and tested using AI. The proposed approach is evaluated using 6070 benign applications and 9419 malware. The investigation results show that solitary utilize perilous authorizations or the quantity of utilized consents can't precisely recognize whether an application is malevolent or amiable. For malware discovery, the proposed approach accomplish up to 99.5% exactness and the learning and preparing time just requirements 0.05s. For malware families location, the precision is 99.6%. The outcomes demonstrate that the technique for obscure/new example's recognition precision is 92.71%. Analyzed against other best in class draws near, the proposed approach is more compelling by identifying malware and malware families.

IV. PROPOSED SYSTEM

Two groups of Android applications, one containing malware and the other containing legitimate apps, are examined through reverse engineering to extract information about their features. Specifically, the analysis revolves around identifying permissions granted and types of App Components present, such as Activity, Services, Content Providers, and more. These features are used as featurevector with class labels as Malware and Goodware represented by 0 and 1 respectively in CSV format.
To reduce dimensionality of feature-set, the CSV is fed to Genetic Algorithm to select the most optimized set of features. The optimized set of features obtained is used for training two machine learning classifiers: Support Vector Machine and Neural Network.
In the proposed methodology, static features are obtained from AndroidManifest.xml which contains all the important information needed by any Android platform about the Apps. The Androguard tool has been utilized to disassemble APKs and extract their static characteristics

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

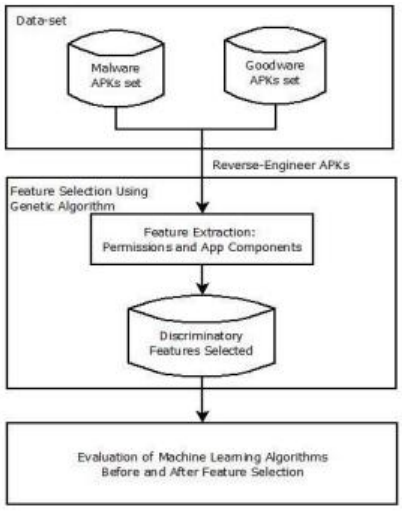

Fig. 1. Proposed Methodology

4.1 Advantages

Security

We suggested a new and effective technique for selecting features that can enhance the overall accuracy of detection. An approach that uses machine learning in conjunction with static and dynamic analysis techniques has the capability to identify novel forms of Android Malware that present zero-day risks.

V. RESULTS AND DISCUSSIONS

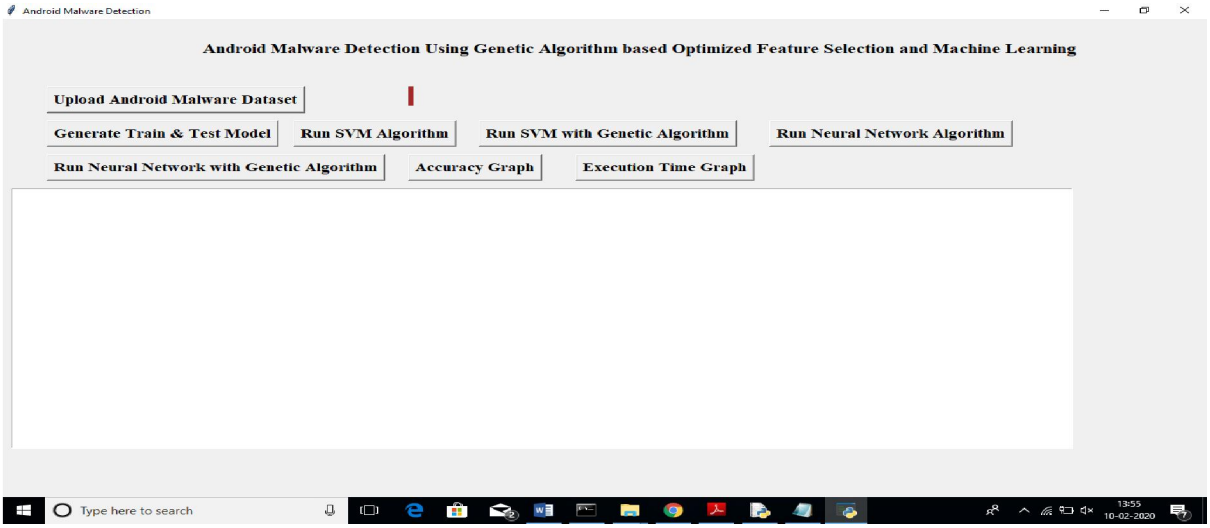

Fig. 2 Upload android Malware Dataset.

In above screen Press the button labeled "Upload Android Malware Dataset" and proceed to upload the dataset.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

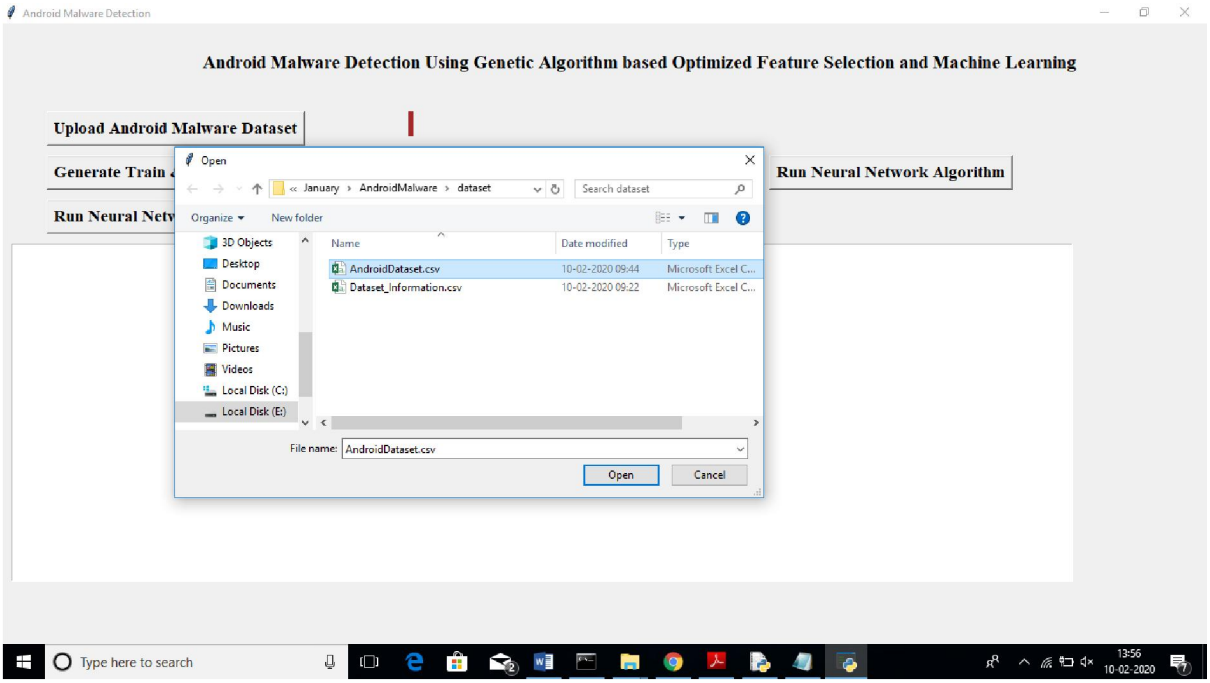

Fig. 3 AndroidDataset.csv.

In above screen I am uploading 'AndroidDataset.csv' file and after upload will get below screen

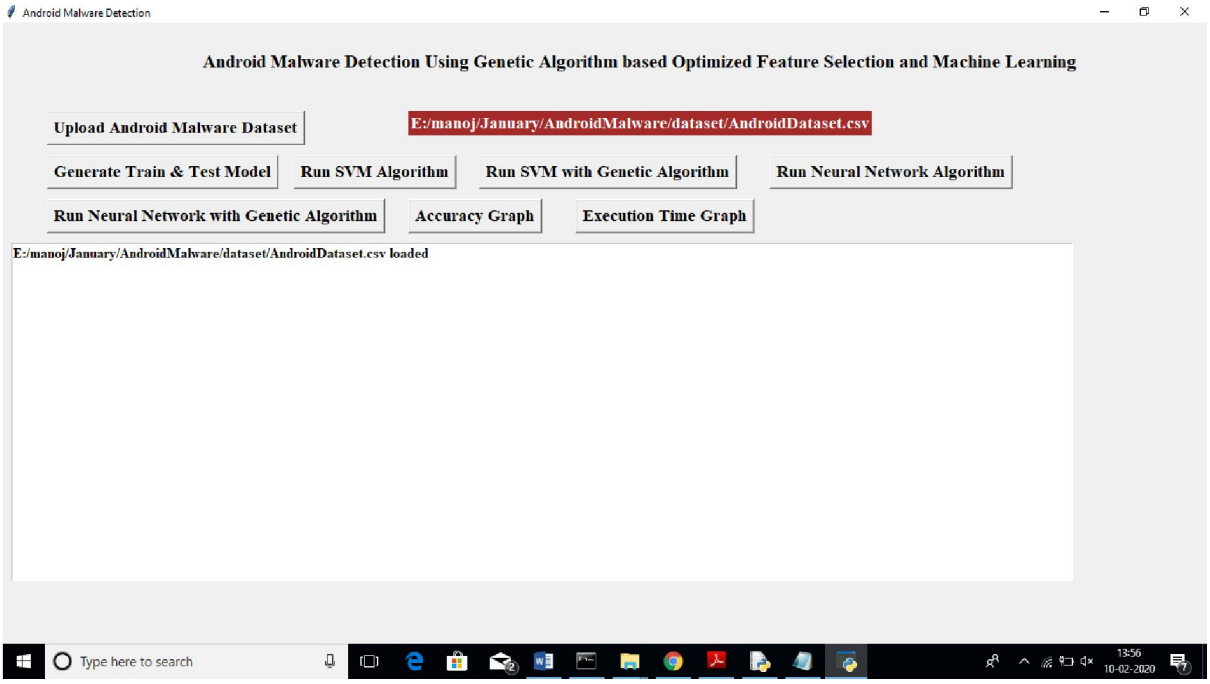

Fig. 4 Generate Train & Test Model.

Now click on 'Generate Train & Test Model' button to split dataset into train and test part. All machine learning algorithms will take 80% dataset for training and 20% dataset to test accuracy of trained model. After clicking that button will get train and test model

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

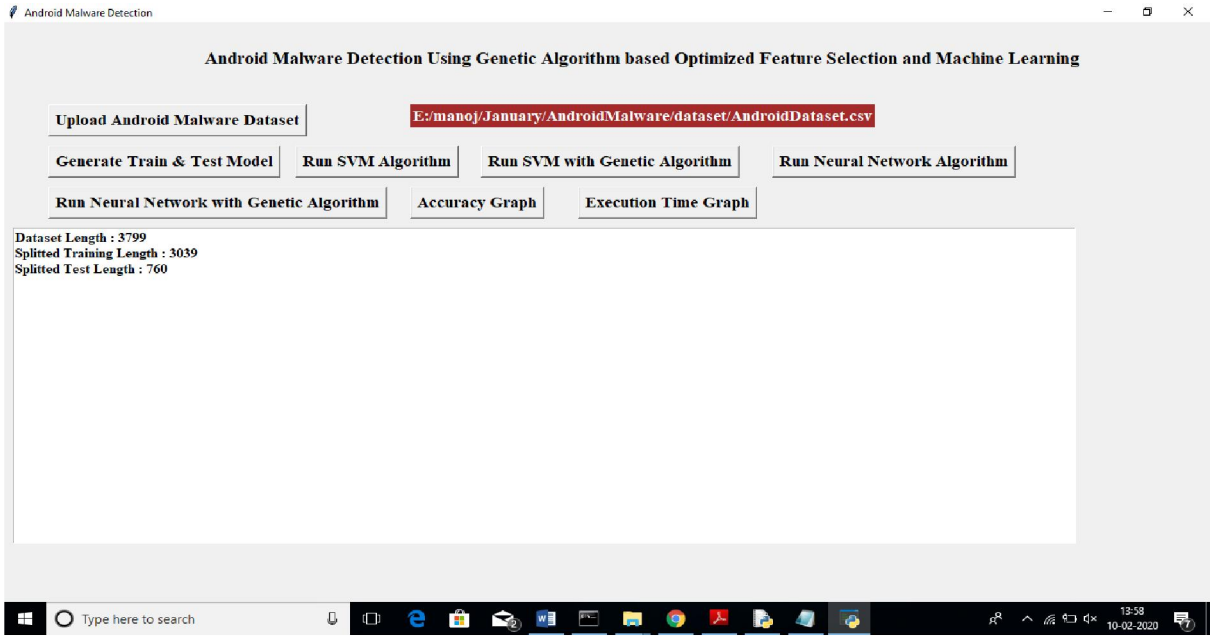

Fig. 5 Run SVM Algorithm.

In above screen we can see there are total 3799 android app records are there and application using 3039 records for training and 760 records for testing. Now we have both train and test model and now click on 'Run SVM Algorithm' button to generate SVM model on train and test and get its accuracy

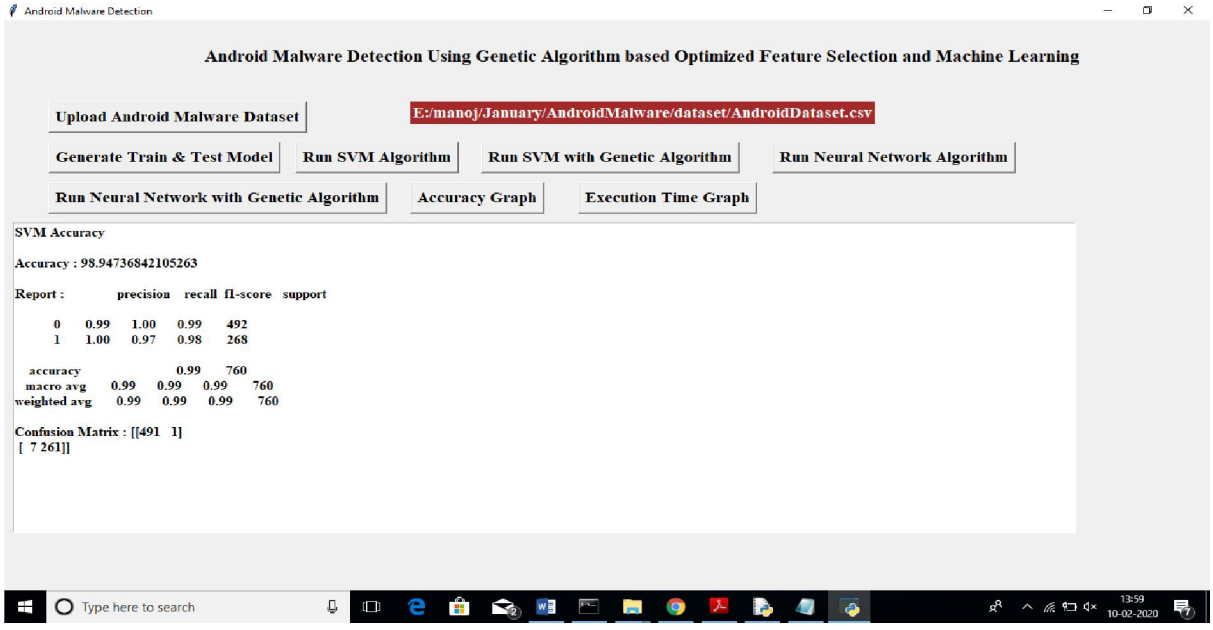

Fig. 6 Run SVM with Genetic Algorithm.

In above screen we got 98% accuracy for SVM and now click on 'Run SVM with Genetic Algorithm' button to choose optimize features and then run SVM on optimize features to get accuracy

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

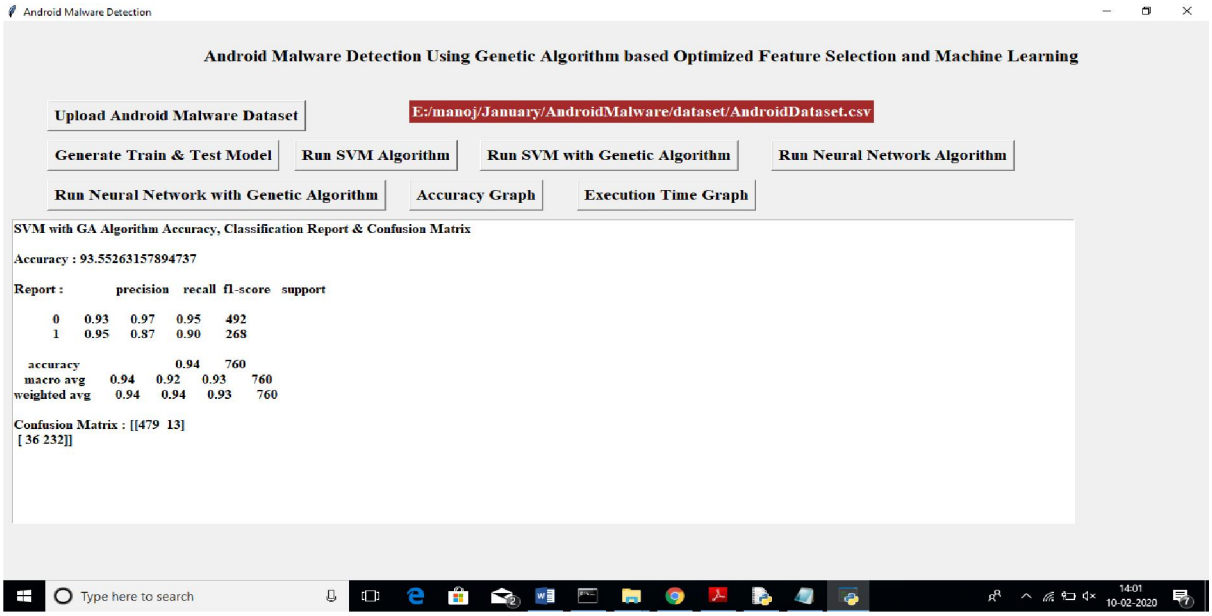

Fig. 7 Execution Time.

The SVM combined with Genetic algorithm achieved a 93% accuracy rating on the above screen. Although the accuracy rate of Genetic with SVM is lower, its execution time is faster, which is visible in the comparison graph.

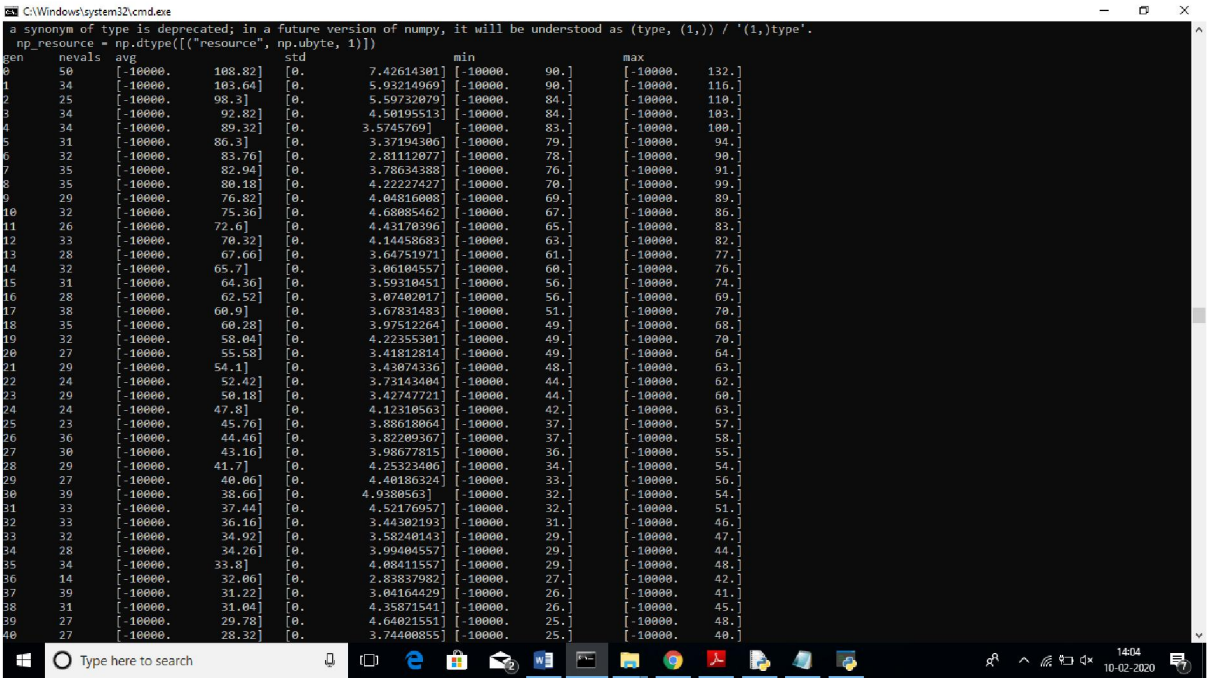

Fig. 8 Run Neural Network Algorithm.

In above console we can see genetic algorithm chooses 40 features from all dataset features.

Now click on 'Run Neural Network Algorithm' button to test neural network accuracy.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

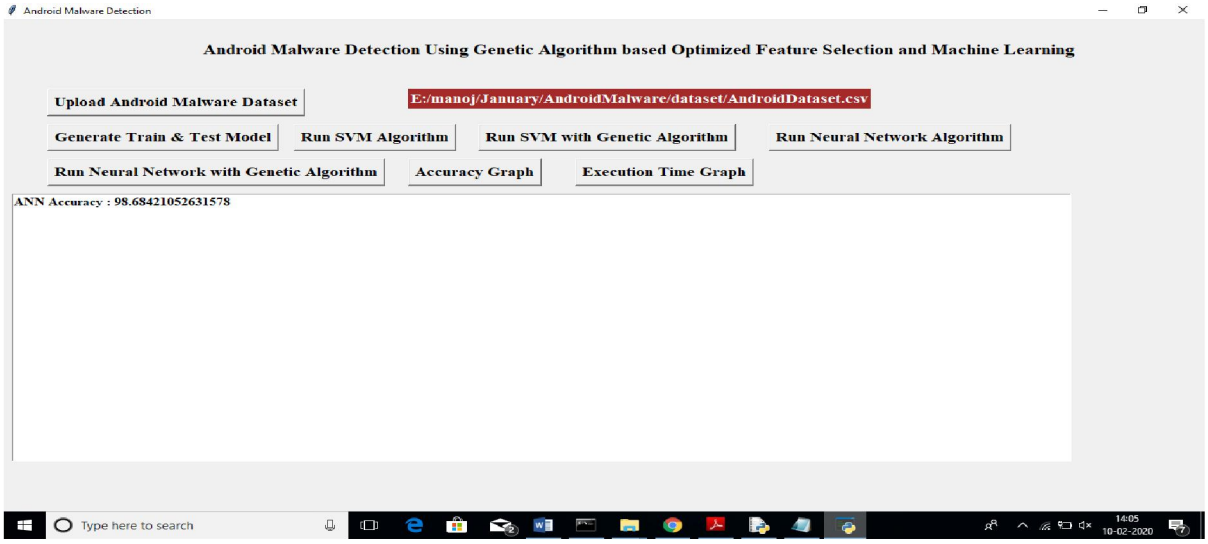

Fig. 9 Run Neural Network with Genetic Algorithm.

In the screen above, the neural network produced an accuracy of 98.64%. To obtain the neural network accuracy using a genetic algorithm, simply click on the 'Run Neural Network with Genetic Algorithm' button.

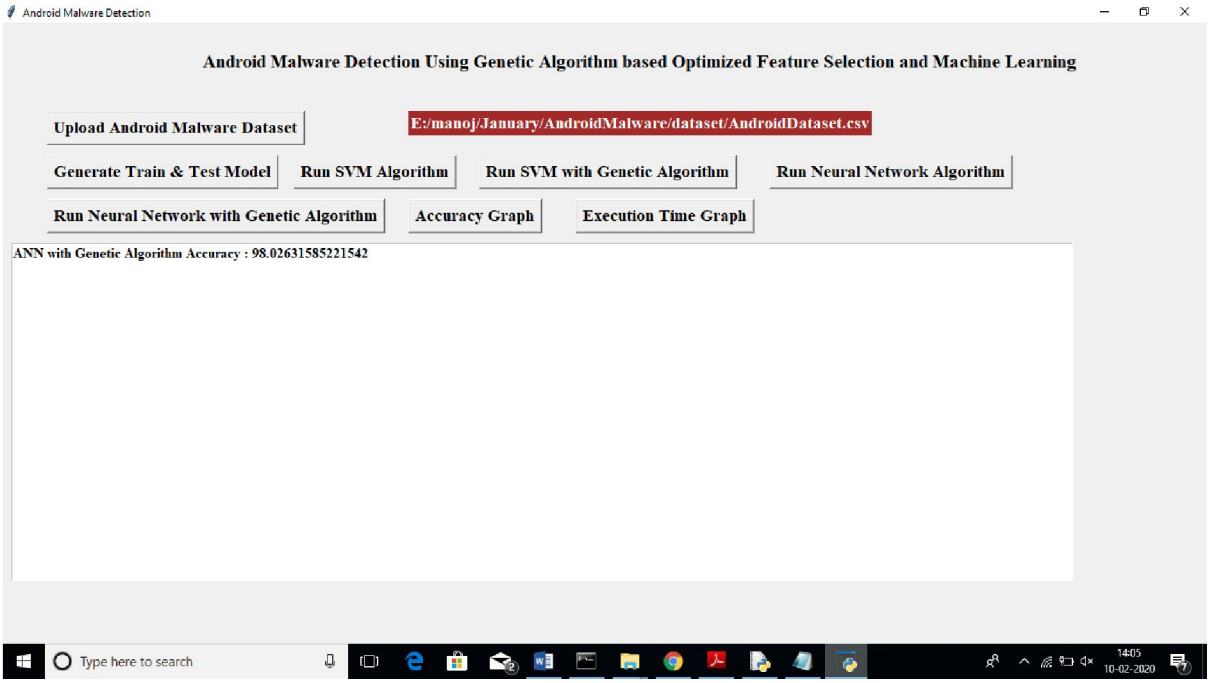

Fig. 10 Accuracy Graph.

On the screen above, the genetic algorithm associated with NN achieved an accuracy rate of 98.02%. To view the accuracy of all the algorithms in the form of a graph, please click on the 'Accuracy Graph' button

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

Fig. 11 Execution Time Graph.

In the given graph, the algorithm name is represented on the x-axis and accuracy on the y-axis, where the Support Vector Machine (SVM) algorithm achieved the highest accuracy compared to other algorithms. Click the 'Execution Time Graph' button to view the execution time of each algorithm.

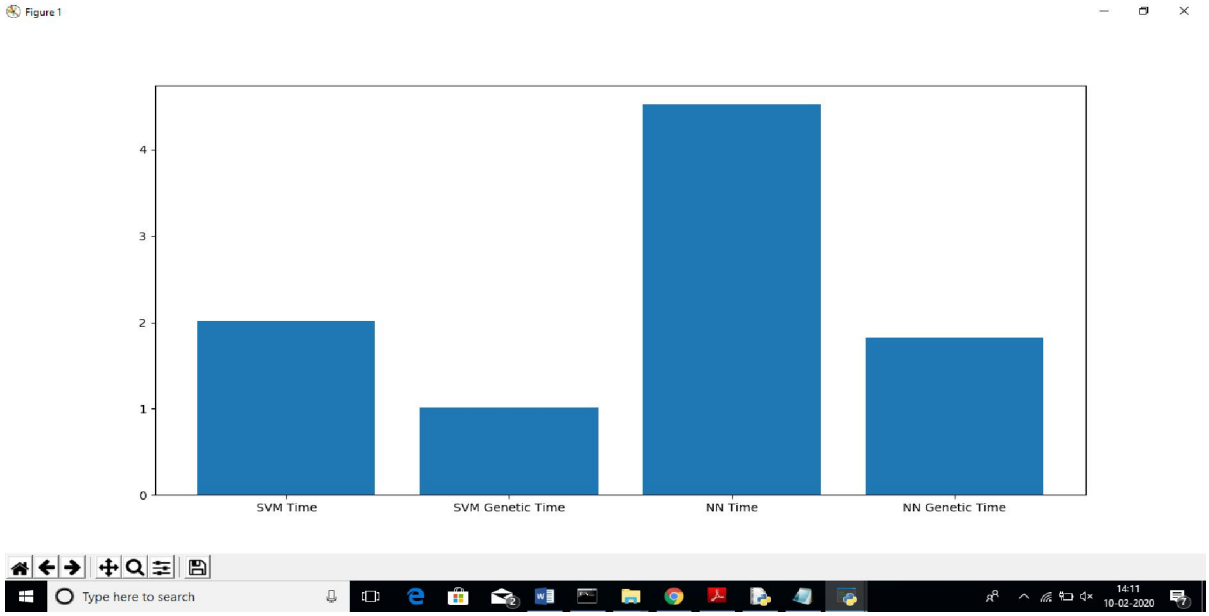

Fig. 12 Genetic Algorithm Graph.

On the graphical representation above, the algorithm name is given along the x-axis, and the execution time is listed on the y-axis. It can be inferred from the graph that the genetic algorithm leads to a faster development of machine learning models.

VI. CONCLUSION

With the rising number of threats on Android platforms, mainly through malicious applications and malware, it is crucial to create a system capable of accurately detecting such malware. While traditional signature-based methods

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, April 2023

cannot identify new variations of malware that pose zero-day threats, AI-based techniques are being developed. Our proposed system utilizes a Genetic Algorithm to obtain the most optimized feature subset that can be used to train AI algorithms most effectively. Based on experimental results, it has been observed that Support Vector Machine and Neural Network classifiers can achieve a fair accuracy level of over 94% while working with lower dimensional feature sets, which reduces the training complexity of the classifiers. To produce better results, further research could look at using larger datasets and analyzing how the Genetic Algorithm impacts other AI algorithms when used in combination

REFERENCES

[1]. N. Milosevic, A. Dehghantanha, and K. K. R. Choo, "Machine learning aided Android malware classification," Comput. Electr. Eng., vol. 61, pp. 266–274, 2017.

[2]. J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-An, and H. Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," IEEE Trans. Ind. Informatics, vol. 14, no. 7, pp. 3216–3225, 2018.

[3]. A. Saracino, D. Sgandurra, G. Dini, and F. Martinelli, "MADAM: Effective and Efficient Behavior-based Android Malware Detection and Prevention," IEEE Trans. Dependable Secur. Comput., vol. 15, no. 1, pp. 83–97, 2018.

[4]. S. Arshad, M. A. Shah, A. Wahid, A. Mehmood, H. Song, and H. Yu, "SAMADroid: A Novel 3-Level Hybrid Malware Detection Model for Android Operating System," IEEE Access, vol. 6, pp. 4321–4339, 2018.