International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 20, Issue 5 (May. 2024), PP. 19-24

Malware Detection Systems on Android PlatformsUsing Genetic Algorithm

Mr. Nilesh D. Mhaiskar[*4]
(Asst. Professor)
Dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India

Veeramalla Srikanth[*1]
Dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India

Chattala Teja Varsha[*3]
Dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India

Jambula Archana Reddy[*2]
Dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India

Abstract—*This study presents an innovative approach for enhancing Android malware detection through a Genetic Algorithm (GA)-based optimized feature selection coupled with machine learning techniques. Leveraging the evolutionary principles of GA, the proposed method effectively identifies a subset of features from a large pool, maximizing the discriminative power while minimizing computational complexity. By integrating this feature selection mechanism with machine learning classifiers, the system achieves superior performance in distinguishing between, benign and malicious Android applications. Through extensive experimentation and evaluation using real-world datasets, the effectiveness of the proposed framework is demonstrated, showcasing significant improvements in detection accuracy, scalability, and efficiency compared to traditional methods. This research contributes to the advancement ofAndroid security, offering a robust and adaptable solution for combating evolving malware threats in mobile ecosystems.*
Keywords— *Genetic Algorithm, Machine Learning, Android Malware, Feature Selection Mechanism, Accuracy.*

Date of Submission: 22-04-2024 Date of Acceptance: 02-05-2024

I. INTRODUCTION

Android applications are widely available on Google Play Store and other platforms, but their open-source nature and popularity have madethem a prime target for malware developers. Despite Google's efforts to protect users, malicious apps still manage to slip through and compromise personal information, such as contacts, emails, and GPS data, for nefarious purposes.To combat this, malware analysis, which comes in two main forms- static and dynamic – is essential. Static analysis involves examining the code structure without execution, while dynamic analysis observes the runtime behavior of apps in a controlled environment. With the rise of zero-day threats, a more efficient detection mechanism is needed beyond traditional signature-based approaches, which require constant updates to the signature database.

II. EXISTING SYSTEM

The primary achievement of this research is the significant reduction of feature dimensions to less than half of the original set using Genetic Algorithms. This reduction allows for inputting into machine learning classifiers, simplifying the training process while retaining accuracy in classifying malware. Unlike exhaustive methods, which require testing numerous combinations, Genetic Algorithms employ a heuristic approach based

on fitness functions for feature selection. The optimized feature set obtained through this method is then utilized to train Support Vector Machine and Neural Network algorithms. Remarkably, a classification accuracy exceeding 79% is maintained despite the reduced feature dimensionality, effectively lowering the training complexity of the classifiers.

III. PROPOSED SYSTEM

This paper presents a Malware detection system on Android Platform using Genetic Algorithm technique. Thetechnique performs data preprocessing at the preliminary stage. For the Android malware detection process, the technique follows an ensemble learning process using three ML models, namely Least Square Support Vector Machine (LS-SVM), kernel extreme learning machine (KELM), and Regularized random vector functional link neural network (RRVFLN). Finally, the hunter-prey optimization (HPO) algorithm is exploited for the optimal parameter tuning of the three DL models, and it helps accomplish improved malware detection results. To indicate the supremacy of the malware detection system on android platform using genetic algorithm approach, a comprehensive experimental analysis is carried out.

Advantages of proposed system:

- Security
- Proposed a novel and efficient algorithm for feature selection to improve overall detection accuracy
- Machinelearning based approach in combination with static and dynamic analysis can be used to detect new variants of Android Malware posing Zero-day threats.

IV. SYSTEM DESIGN

a. System Architecture

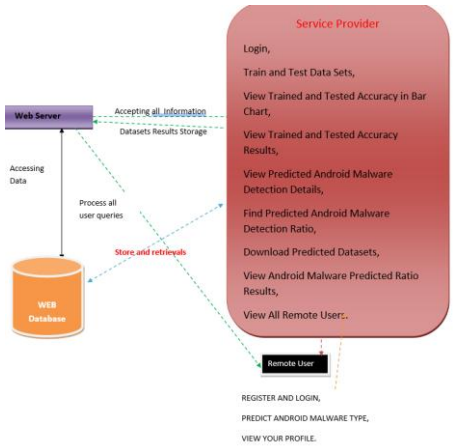

fig - 1

A web server has been developed to handle various functions related to Android malware detection. The server is designed to accept all information related to malware detection, accessing and processing data from a web database. Users can log in to the system, where service providers can train and test data sets, and view the accuracy of their trained and tested models through bar charts and detailed results. Additionally, users can view predicted Android malware detection details, including detection ratios, and download predicted datasets for further analysis. The system allows remote users to register, log in, predict Android malware types, and view their profiles, ensuring a comprehensive and user-friendly experience for all.

b. Data Flow Diagram: -

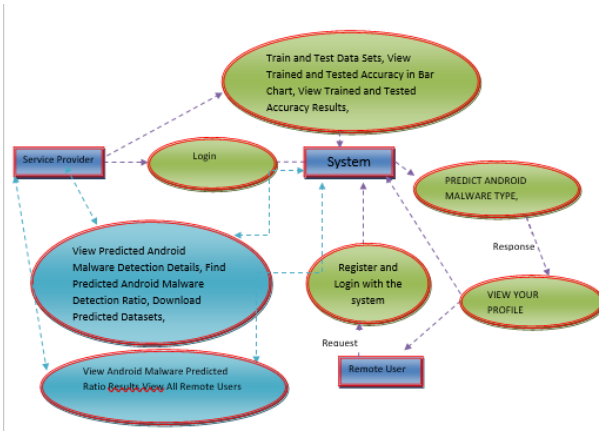

fig 2

c. Sequence Diagram: -

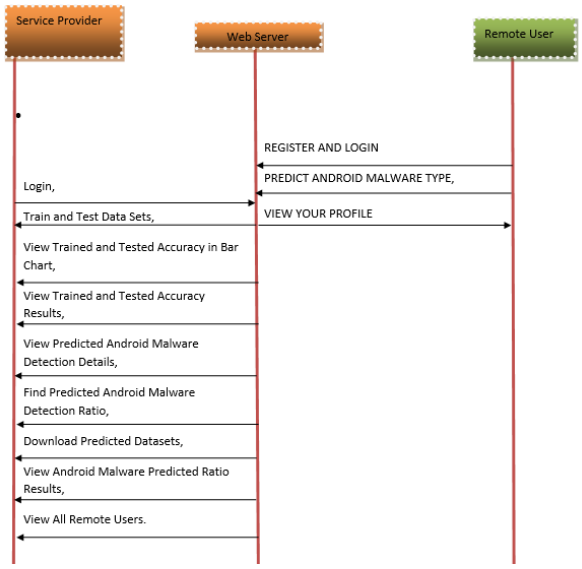

fig 3

V. METHODOLOGY

Modules

a. Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Predicted Android Malware Detection Details, Find Predicted Android Malware Detection Ratio, Download Predicted Datasets, View Android Malware Predicted Ratio Results, View All Remote Users.

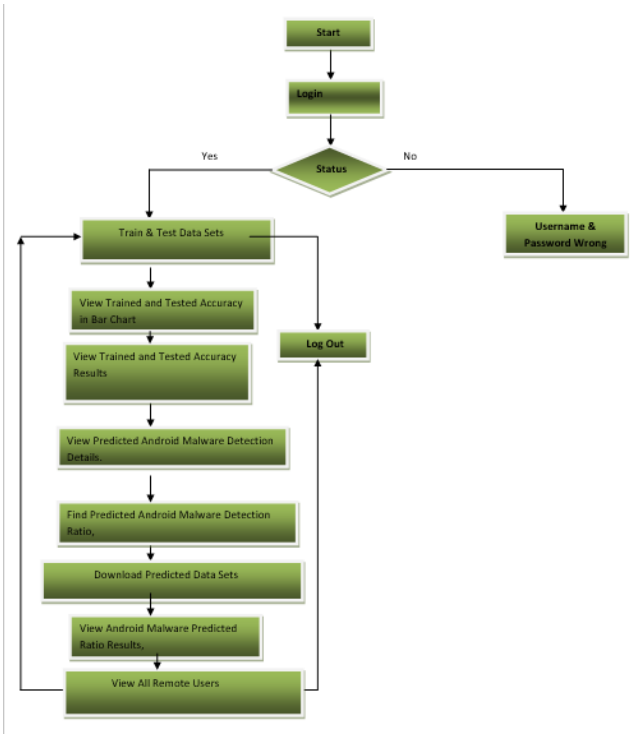

fig 4

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT ANDROID MALWARE TYPE, VIEW YOUR PROFILE.

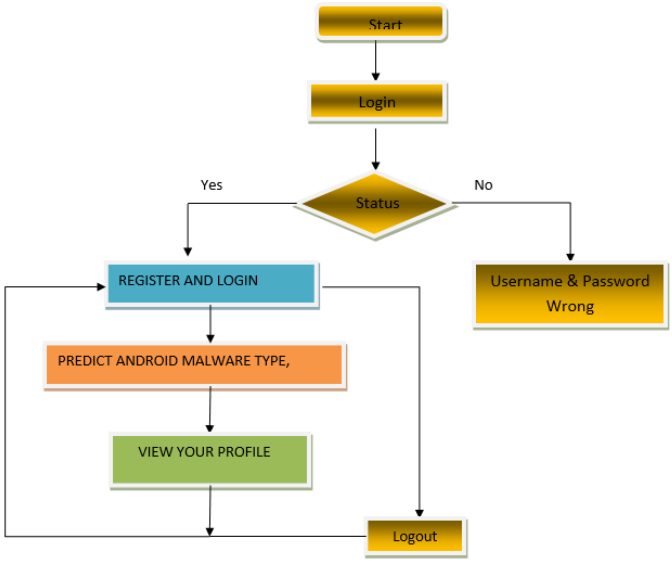

fig 5

VI. EVALUATION

a) Algorithms: -

i. Decision Tree: -

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision-making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labelled with this class
Step 2. Otherwise, let T be some test with possible outcomes O1, O2, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2, Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi, we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

ii. SVM: -

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels fornewly acquired instances. Unlike generative machine learning approaches, which require computations ofconditional probability distributions, a discriminant classification function takes a data point *x* and assignsit to one of the different classes that are a part of the classification task. Less powerful than generativeapproaches, which are mostly used when prediction involves outlier detection, discriminant approachesrequire fewer computational resources and less training data, especially for a multidimensional featurespace and when only posterior probabilities are needed. From a geometric perspective, learning a classifieris equivalent to finding the equation for a multidimensional surface that best separates the different classesin the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptron's*, both of which are widely used for classification in machine learning. For perceptron's, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptron's is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

iii. Logistic Regression Classifiers: -

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

VII. RESULT

The overall outcomes of the Malware detection system on Android platform using Genetic Algorithm method with other models. The outcomes identified that the NB approach has poor performance, whereas the AdaBoostM1 model gains slightly enhanced results. Along with that, the Machine Learning models accomplish moderately closer performance. However, the AAMD-OELAC technique offers better results with increased accu_y of 98.93%, prec_n of 99.15%, reca_l of 98.93%, and F_score of 99.04%. Finally, the computational time (CT) analysis of the Malware detection system on Android platform using Genetic Algorithm technique is compared with recent approaches in T The outcomes exhibited that the Malware detection system on Android platform using Genetic Algorithm technique reaches the least CT value of 8s. At the same time, the existing models have reached increased CT values. These results highlighted that the Malware detection system on Android platform using Genetic Algorithm technique shows maximum performance over other models on malware classification.

VIII. CONCLUSION

The Malware detection system on Android platform using Genetic Algorithm technique for accurate and automated Android malware detection. This approach focuses on automatic recognition and classification of Android malware through data preprocessing, ensemble classification, and HPO-based parameter tuning. Using Machine Learning such as DT, SVM, LSmodels, the technique achieves superior malware detection results. Experimental analysis demonstrates the effectiveness of Malware detection system on Android platform using Genetic Algorithm over existing methods.

REFERENCE

[1]. H. Rathore, A. Nandanwar, S. K. Sahay, and M. Sewak, ‘‘Adversarialsuperiority in Android malware detection: Lessons from reinforcementlearning based evasion attacks and defenses,’’ Forensic Sci. Int., Digit. Invest., vol. 44, Mar. 2023, Art. no. 301511.

[2]. H. Wang, W. Zhang, and H. He, ‘‘You are what the permissions told me! Android malware detection based on hybrid tactics,’’ J. Inf. Secur. Appl., vol. 66, May 2022, Art. no. 103159.

[3]. A. Taha and O. Barukab, ‘‘Android malwareclassification using optimizedensemble learning based on genetic algorithms,’’ Sustainability, vol. 14,no. 21, p. 14406, Nov. 2022.

[4]. O. N. Elayan and A. M. Mustafa, ‘‘Android malware detectionusing deep learning,’’ Proc. Comput. Sci., vol. 184, pp. 847–852,Jan. 2021.

[5]. J. Kim, Y. Ban, E. Ko, H. Cho, and J. H. Yi, ‘‘MAPAS: A practical deep learning-based Android malware detection system,’’ Int. J. Inf. Secur. Vol. 21, no. 4, pp. 725–738, Aug. 2022