International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 2, March 2020 |

Android Malware Detection Using Random Forest Learning Algorithm

Prasad Raju Sabha

UG Student, Department of Information Technology, B.K Birla college of Arts, Science and Commerce (Autonomous), Kalyan, India

ABSTRACT - Android smart phones popularity has increased in now days.in recent report 7out ofeach 10 application are malware.230,000 new malware samples are produceddaily. this is predicted to only keep growing Efficient identification of evolving malware is an urgent challenge[1]. Due to the potential for data theft that mobile phone users face, the detection of malware on Android devices has become an increasingly important issue in sector of cyber security.In this paper we use machine learning algorithm Random forest learning for detection of android malware. during this paper we also compare logistic regressor, svm and eventually Random forest learning algorithm .in logistic regression accuracy is 0.96 ,svm model accuracy is 0.92 and eventually Random forest regression accuracy is 0.98. it's highest accuracy.

KEYWORDS: android malware detection, random forest regression, logistic regression, svm learning algorithm , security issues

I.INTRODUCTION

Traditional methods like signature-based routines are unable to guard user's data. In now days rapid behaviour changes in new varieties of Android malware. Existing malware detection mechanisms are using mostly dynamic, static and hybrid analysis for detection any malware application. Recently, lots of researches have focused on signature-based detection methods, using static or dynamic analysis to get high recognizable patterns which are then used to detect malware. However, this sort of methods would decrease effective when detecting unknown malware. It's tougher to detect malicious applications using traditional methods because of code obfuscation, transformation attacks, etc. In this paper we use Random forest learning algorithm for detection of android malware. Random forest is a supervised learning algorithm. In this paper we also compare logistic regressor, svm and eventually Random forest learning algorithm.the main aim of this paper is to developed an android malware detection using Random forest regression.to detect unknown malware
Random forest regression is supervised learning algorithm ,it is easy to use and it's flexible.in random forest use multiple decision trees and a way called Bootstrap Aggregation it also called baggin. Bagging involves in training, each decision tree on a distinct data sample where sampling is finished with replacement. the concept behind this is often to mix multiple decision trees for determining the ultimate output.

II.OBJECTIVES

-To detect an unknown android malware application.
-To achieve higher accuracy in malware detection.

we find following hypothesis
Hypothesis:
1.if random forest learning algorithm is used in android malware detection then it's accuracy is increased by 0.98 . because it combine multiple decision tree and determine the final result(output).

III.LITERATURE REVIEW

(1)In this paper Authors ML-based method were used for detection of android malware.This method utilizesmore than 200 feature extracted from both static and dynamic analysis of android app for malware detection.This model achieved a high level of 96% accuracy with real world android application sets[3].(2)In this paper Authors detect Android malware, instead of using Application Programming Interface (API) calls only, In this further analyse the different

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 2, March 2020 |

relationships between them and created higher-level semantics which require more efforts for attackers to evade the detection.HinDroid which introduces a structured heterogeneous information network (HIN) representation of Android apps, and a meta-path based approach to link the apps[4].(3)In this paper Authors presented MADAM, a novel host-based malware detection system for Android devices which simultaneously analyses and correlates features at four levels: kernel, application, user and package, to detect and stop malicious behaviors MADAM is the first system which aims at detecting and stopping at run-time any kind of malware, without focusing on a specific security threat, using a behavior-based and multi-level approach[5].(4) In this paper Authors propose a novel Android malware detection framework that utilizes many static features to reflect the properties of applications in various aspects. Total seven kinds of feature extracted by analyzing files such as a manifest file, a dex file, and a .so file from an APK file, and these features enrich the extracted information to express applications' characteristics. As a result, this framework was effective enough to be used in the Android malware detection.this research is the first application of the multimodal deep learning to the Android malware detection[6].(5) In this paper Authors resent PIndroid—a novel Permissions and Intents based framework for identifying Android malware apps. The proposed approach, when applied to 1,745 real world applications, provides 99.8% accuracy (which is best reported to date). Empirical results suggest that the proposed framework is effective in detection of malware apps[7].(6)this method employs a traffic mirroring technology to collect network traffic generated by mobile apps, and the generated network traffic is transmitted to a server for data analysis. So this detection method will not subject to user's surfing habits, device resources or other device-specific factors. On the server side, this method extracts traffic features, and then uses detection models based on machine learning to detect whether the app is malicious or not[8].(7)In this method novel mechanism for detecting Android malware applications by combining static and dynamic features influencing the malicious activity by exploring their conditional dependencies. The proposed mechanism can accurately capture the malicious behavior than existing static and dynamic analysis mechanisms[9].

IV.METHODOLOGY

Random forest :In random forest regression every decision tree has high variance, but after we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on it particular sample data and hence the output doesn't rely upon one decision tree but multiple decision trees. within the case of a classification problem, the ultimate output is taken by using the majority voting classifier. within the case of a regression problem, the eventual output is that the mean of all the outputs. This part is Aggregation.

Steps for random forest regression:

- Design or collect a dataset .
- Make sure the dataset is in an accessible format else convert it to the required format.
- Specify all noticeable anomalies and missing dataset points that may be required to achieve the required dataset.
- Create a machine learning model
- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn't satisfy your expectations, you can try improving your model accordingly or dating your data or use another data modeling technique.
- At this stage you interpret the data you have gained and report accordingly.

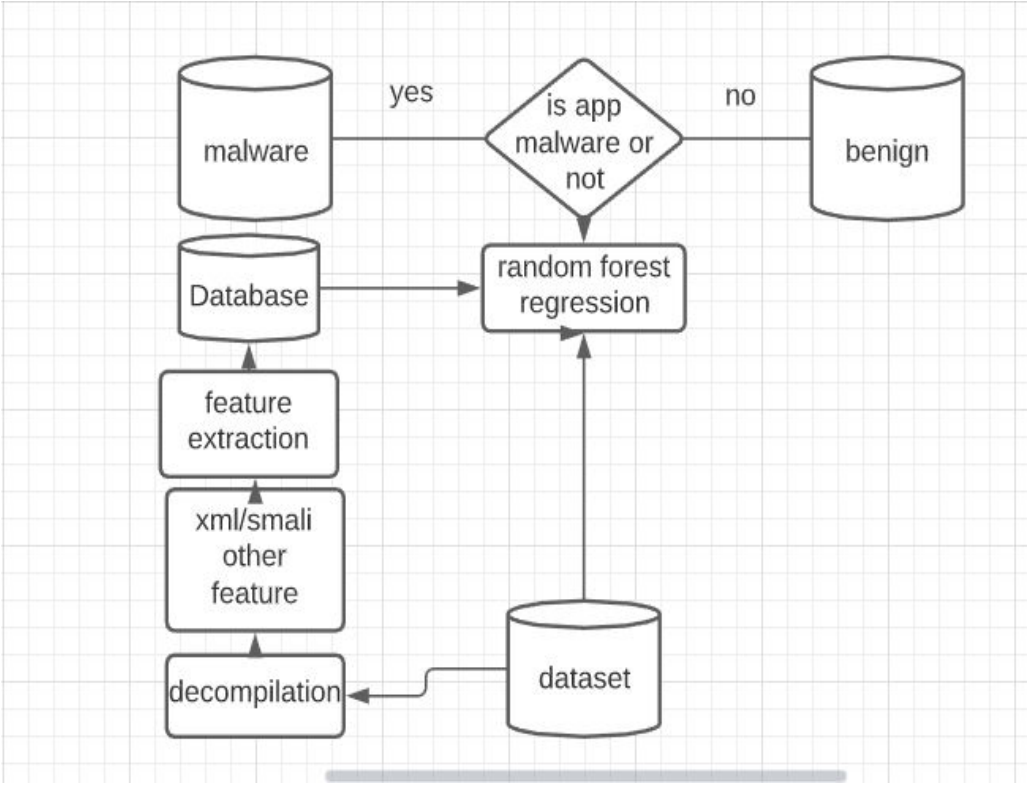
V.EXPERIMENTS

System Design:the proposed system is used to detect android malwre occuring in any android devices with help of random forest regression learning algorithm .we first extract the API calls and other features from android application by using a reverse engineering.second these feature are used random forest regression for detection of malware.

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

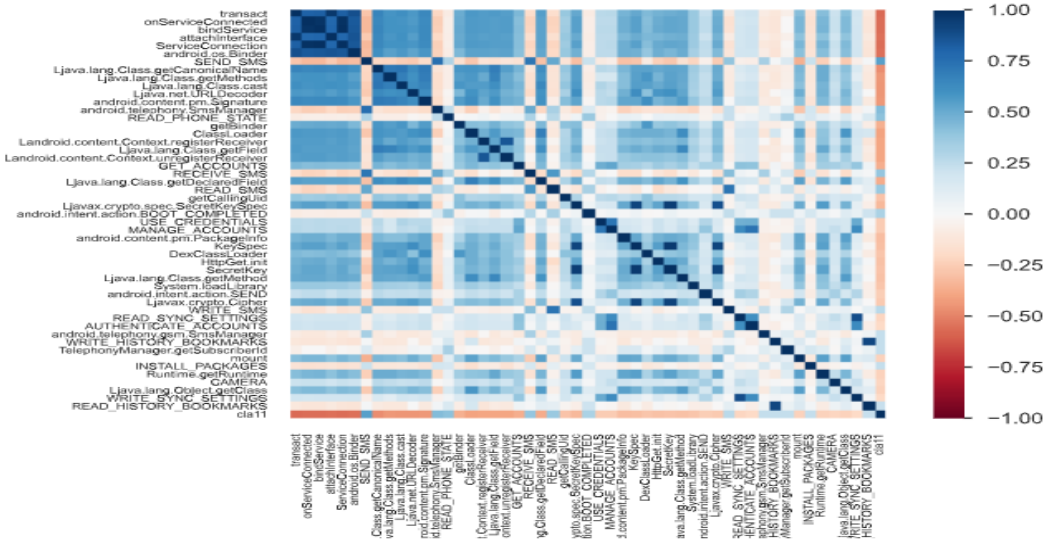
| Volume 7, Issue 2, March 2020 |

Flow diagram of a model



Dataset: The dataset is used in this model is a drebin dataset but in this dataset we selected some special features. The dataset contains 5560 application from 179 different malware families.

Analysis of dataset:we analys data on jupyter notebook for finding a dependent and independent variables .
Pearson correlation



| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 2, March 2020 |

VI.COMPARING A THREE MODELS RESULT

We are comparing three machine learning models using same dataset that is drebin dataset. When all of these model give below given accuracy.

1)A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problemswe use a svm learning algorithm to find android malware .but this algorithm doesn't give expected result. It give only 0.92 accuracy .

```
confusion matrix

[[8529  847]
 [ 416 5094]]


classification report
              precision    recall  f1-score   support

           0       0.95      0.91      0.93      9376
           1       0.86      0.92      0.89      5510

    accuracy                           0.92     14886
   macro avg       0.91      0.92      0.91     14886
weighted avg       0.92      0.92      0.92     14886
```

2) Logistic regression is a statistical model. Logistic regression uses a logistic function to model a binary dependent variable. Also the logistic regression is used to train the model to calculate the probability of a certain class such as pass/fail,alive/dead or yes/no.

When we using logistic regressionin android malware detection then ,our modelgive accuracy of 0.96.

```
confusion matrix
[[1888   52]
 [  57 1011]]

classification report
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      1940
           1       0.95      0.95      0.95      1068

    accuracy                           0.96      3008
   macro avg       0.96      0.96      0.96      3008
weighted avg       0.96      0.96      0.96      3008

>>>
```

3)**random forest regression:** when we using a random forest regression ,then our model give highest accuracy which is 0.98.

```
confusion matrix
[[1914   26]
 [  34 1034]]


classification report
              precision    recall  f1-score   support

           0       0.98      0.99      0.98      1940
           1       0.98      0.97      0.97      1068

    accuracy                           0.98      3008
   macro avg       0.98      0.98      0.98      3008
weighted avg       0.98      0.98      0.98      3008

accuracy  0.9800531914893617
>>>
```

VII.CONCLUSION

Android malware detection is major problem in today's life .new malware are introduce every minutes in the world.it's important issue for the field of cyber security.malware detection application is already available in world but this type

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 2, March 2020 |

of android malware detection application unable to detect unknown malware.In this paper we use random forest regression to detect an android malware , our model give highest accuracy which is 0.98. we also compare svm and logistic regression to find a best result. Svm give 0.92 accuracy in malware detection.logistic regression give 0.96 accuracy, which is more than thesvm. In future we use genetic algorithm with random forest regression . because genetic algorithm provide best and optimized result for unknown malware detection.we also use reinforcement algorithm in future for detection of android malware.

ACKNOWLEDGMENT

A special gratitude is conveyed to our prof. SwapnaAugustine Nikale department of Information technology of B.K Birla college of Arts,Science& Commerce(Autonomous) Kalyan Thane Maharashtra India.

REFERENCES

[1https://purplesec.us/resources/cyber-security-statistics /#:~:text=Trojans%20make%20up%2051.45%25%20of,US%20%242.4%20million%20in%20defense.
[2]:https://en.wikipedia.org/wiki/Reinforcement_learning
[3]:"Deep Android malware detection"by Niall mclaughlin, Jesus Martinez, del rincon,Boojoongkang,suleimanyerima,paul miller..https://dl.acm.org/doi/abs/10.1145/3029806.3029823
[4]:ShifuHou, Yanfang Ye ,Yangqiu song melihabdulhayoglu,"HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network"https://dl.acm.org/doi/abs/10.1145/3097983.3098026
[5]: Andrea Saracino, Daniele Sgandurra, Gianluca Dini and Fabio Martinelli developed ."MADAM: Effective and Efficient Behavior-based Android Malware Detection and Prevention" https://arpi.unipi.it/bitstream/11568/789315/1/2016-tdsc-ssdm-preprint.pdf
[6]:TaeGuen Kim, BooJoong Kang, Mina Rho, SakirSezer and EulGyuIm"A Multimodal Deep Learning Method for Android Malware Detection using Various Features" https://core.ac.uk/download/pdf/160471635.pdf
[7]FauziaIdreesa, MuttukrishnanRajarajana , Mauro Contib , Thomas M. Chena , YogachandranRahulamathavanc,"PIndroid: A novel Android malware detection system using ensemble learning methods"
https://openaccess.city.ac.uk/id/eprint/17316/1/
[8]http://www.download-paper.com/wp-content/uploads/2019/11/2018-elsevier-A-mobile-malware-detection-using-behavior-features-in-network-traffic.pdf
[9]:https://www.researchgate.net/profile/Tony_Thomas_Kallivayalil/publication/341592340_A_TAN_based_hybrid_model_for_android_malware_detection/links/5ed8a2cc92851c9c5e7ba07a/A-TAN-based-hybrid-model-for-android-malware-detection.pdf