Volume 9, Issue 4, April – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1102

Malware Detection using Machine Learning

Dilip Dalgade[1] (Professor)
Department of Computer Engineering
Rajiv Gandhi Institute of Technology,
Mumbai, India

Srushti Patyane[2]
Department of Computer Engineering
Rajiv Gandhi Institute of Technology,
Mumbai, India

Anushka Matey[3]
Department of Computer Engineering
Rajiv Gandhi Institute of Technology,
Mumbai, India

Saloni Singh[4]
Department of Computer Engineering
Rajiv Gandhi Institute of Technology,
Mumbai, India

Amey Godbole[5]
Department of Computer Engineering
Rajiv Gandhi Institute of Technology,
Mumbai, India

Abstract:- As the level of malware and viruses is on the rise, the prominence of effective detection systems is crucial. Malwares are the modern-day threats that have troubled major companies worldwide. This article explores in depth two powerful machine learning tools, Random Forest, Support Vector Machines in particular, for the detection of malware. Our study revealed the Random Forest's capacity to reach the upper detection accuracy limit of 98% by applying an analysis of a dataset of variousmalware samples. The feature selection process as well as the model improvement that we've adopted have substantially improved use of our approach for malware detection, and this is thereby highly crucial for organizations to fight against evolving cyber threats. The results of the present research support the ongoing actionsof strengthening cybersecurity security, therefore, providing invaluable information for proactive defense approach mechanisms against malicious software attacks.

Keywords:- *Malware, Machine Learning, Random Forest, Support Vector Machines (SVM), Detection Accuracy, Cybersecurity, Feature Selection, Model Optimization.*

I. INTRODUCTION

In recent years, malware attacks have become a major pain in the neck that have slowed down cybersecurity efforts and even disrupted the operations of organizations. The crimeware species always remain in the mode of evolving with sophistication of technologies and therefore the conventional signature-based detection methods are of little use against this. Therefore, novel techniques are increasingly being recognized as the most outstanding methodology that effectively recognized these threats and quench them. Machine learning algorithms which have proved to be greatly potential recently as a weapon fight against malware. Through the use of computation algorithms and big data analysis, the machine learning is instrumental and enables effective development of preventive defense mechanisms able to spot and kill all the malware variants in real time. In this research, the paper allows to make a comparison between two machine learning algorithms that are known to be the best at malware detection, namely Random Forest and Support Vector Machines (SVM). The Random Forest approach is an advanced ensemble method that involves creation of many decision trees in an attempt to boost precision of the classifier. One the other hand, SVM is efficient supervised learning method which specializes in recognizing patterns in complex datasets. This purpose will be achieved through the implementation of a dataset with a structured collection containing samples of malware.

Furthermore, the study delves into how the crucial factors like the feature selection and tuning of the models and algorithms affect their performance. Through achieving those targets, the research project at hand makes a valuable support to the efforts in intensifying cybersecurity safeguards and minimizing the risks that come with evolving cyber threats. The results of the investigation which are of a practical nature are of a great help to cybersecurity specialists toward the creation of strong detection frameworks and network securing againstmalware attacks.

II. LITERATURE SURVEY

The work of Tamás Csongor et al. [1] concentrated on utilization of malwares hashes for effect detection. They relied on both behavioral analysis andpattern recognition as the foundations of their technology generating automatic alerts in a manner that would allow for preventive threat mitigation. While the results demonstrated their approach's effectiveness, the research study was faced with challenges pertaining to the integration of false positives/negatives that could potentially affect the accuracy of the detection outcome. Though their system was highly accurate,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1102

SIMBIoTA had a true positive detection rate of 97-98% and many other capabilities, also the results proved its competence in discriminating malevolent software threats.

To the work of S. Agarkar et al. [2], they showed that by the means of a combination of machine learning methods namely Light GBM, Decision Tree, and Random Forest, the results were more robust and efficient. With a special emphasis on how to reduce false negatives and that of the security achievement better, their methodology was set. The researchers also started experimenting with using larger datasets and adding more features to help improve the discriminability of their classifiers afterward. To note, however, the accurate rates they got were impressive, with Tree Decision yielding 99.14%, Random Forest giving 99.47%, and Light GBM, with an impressive 99.50%, reasserting that their strategy worked perfectly in helping the system recognize rogue software file types while attaining high accuracy.

The research led by S. A. Roseline [3] for assembly of hybrid multi-layered ransomware and the procedure of application of a random forest ensembling algorithm was developed for the detection of malware. This involved their approach in finding a trade- off between computation and accuracy, highlighting efficiency as the basis of the whole process and at the same time, achieving a good detection performance. Yet, the limitation of the study on using the testing set that was sized, might impact my ability to generalize the result. Regardless of that is, their proposed art technique has reached impressive success by bringing the accuracy rate of 98.91% while using up only a few resources of thecomputation as they have demonstrated.

The team of K. Sethi et al [4] applied a set of classical machine learning techniques including k-NN, SVM and Random Forest for the constituent family's malware class distribution. They classified the malware being researched by determine their characteristics, functions, and exploits. On the other hand, one of therestrictions of their research was that they used their own hand-created data source, with it being challenging to achieve real diversity and representativeness of modern malware samples like this. The fact that there is a drawback to this finding makes it more impressive that the Decision Tree algorithm, which is able to be used, performed very well maintaining a precision rate of 99.11%. This point shows the strength of their solution which helps in precise classification of malware family except constraints of dataset associated with it.

Similarly, N.A Anuar et al. [5] explored the capabilities of machine learning classifiers, including Naïve Bayes, Support Vector Machine, Random Forest, and Decision Tree, in identifying malware threats. While their focus was primarily on dynamic analysis, their findings underscored the resilience of SVM in accurately predicting malware behavior, achieving an impressive accuracy score of 95.4%. Despite the inherent limitations of dynamic analysis, the study emphasized the pivotal role of SVM in bolstering malware detection capabilities.

In contrast, P Priyadarshan et al. [6] adopted a hybrid approach combining dynamic and static analysis techniques, utilizing k-NN, Logistic Regression, and Random Forest algorithms for malware detection. Despite the absence of feature selection and reduction techniques, their approach showcased a high accuracy rate of 99.1% with Random Forest, highlighting the synergistic benefits of combining different analysis methods in enhancing detection capabilities.

Furthermore, M. Masum et al. [7] conducted a comprehensive assessment of machine learning classifiers, including Decision Tree, Random Forest, Bayesian, Logistic Regression, and Neural Network, to identify the most effective malware identifier. Their findings underscored the superiority of the Random Forest algorithm, which achieved a classification rate of 99%, highlighting the efficacy of ensemble learning techniques in improving detection efficiency.

Moreover, SH Kok et al. [8] conducted a comparative analysis of various ransomware detection techniques, encompassing state-of-the-art methods such as Bayesian Decision Tree, Dimension Reduction, Instance- Based, Clustering, Deep Learning, Ensemble, Neural Network, and Regression. While their study outlined the breadth of techniques available for ransomware detection, detailed information regarding their hybrid algorithm configuration was lacking. Nevertheless, their research contributed to advancing the field of ransomware detection, paving the way for more comprehensive and informative detection strategies.

Lastly, Ham, Hyo-Sik et al. [9] focused on Android malware detection using a Linear Support Vector Machine (SVM) classifier, aiming to ensure reliable Internet of Things (IoT) services. Despite potential biasesin the dataset, their findings demonstrated the superior performance of the Linear SVM classifier in accurately identifying malicious software instances, highlighting its effectiveness in enhancing malware detection efficiency, particularly in the context of IoT security.

III. PROPOSED SYSTEM

In order to accomplish the task of discovering and preventing malware, our structure combines different features selection methods and a number of various machine learning methods, for example, Random Forest (RF), Support Vector Machine (SVM), and cross validation. These techniques are used to pick up attribute set of population that is chosen by us intelligently. Lowered data quality and high correlation data features are detected and eliminated by variance inflation factors which are a popular feature selection technique; as a result, the effectiveness of our system isguaranteed.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1102

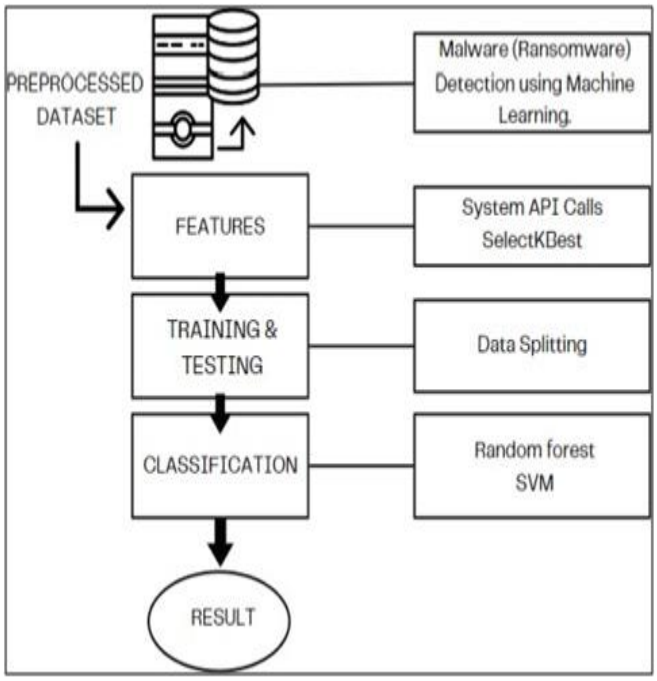

Fig 1 System Architecture

➢ *For the Proposed System, we will Adapt the Steps Outlined in the Recommended Approach for Malware Detection:*

- *Data Collection:*

At the initial stage, we gather a dataset which is composed not only of some ransomware samples but also legitimate software files. The dataset shall have enough size as well as a variety in order to be able to draw accurate conclusions for the machine learning model. The system relies on a dataset where each sample comprises 138,047 labeled instances of ransomware and the remaining 30% of the dataset contain regular observations.

- *Feature Extraction:*

With data being collected and the features to be extracted, we then develop statistical models that can combine historical data with the data collected. These metadata features are something like file size, file type, entropy, and moving to the most significantone and common for both economic and industrial espionage methods, System API Calls. System API Calls are the main source of information that expose the secret of software behavior and should be the main criterion the detection system considers to differ between harmless andmalicious software.

- *Machine Learning Framework:*

Earlier, we got the features by extraction that forms the basis of the machine learning platform. We can do this by using Random Forest, SVM and cross validation depending on the machine learning technique that gives the best performance. It is crucial to set up a balanced dataset that yields both clean and bad patterns, without dataset bias tendency, to counter the model.

- *Evaluation:*

Further, an assessment is carried out to monitor the accuracy of the model after training. The evaluation of this quality may contain an estimation of the following metrics: precision, recall, accuracy, etc. These indicators can demonstrate how good the model is in classifying ransom ware and that which of them are legitimate software.

- *Implementation:*

Subsequently, the model is evaluated for the purposes of training, and then the instrumented model is deployed on a real-time environment for malware detection. Monitoring round the clock is necessary to detect any activities that are suspect to aid ascertain malwares. When it is found out, the conduct which is considered as IT security violations can be reported to users or system administrators to rectify threats.

- *Dataset Splitting:*

The feature extraction is the first step that prior to splitting the data set is set into the training and, the testing data. This is accomplished being that the ratio of the split is 80-20; where 80% of the data is set aside to train the model and 20% is employed in testing its performance. This ensures that the model learn on a large amount of data for training to avoid over training but also a subset of unseen data for evaluation.

- *Algorithm Selection:*

The random forest and SVM algorithms are selected concerning the accuracy of the model. Based on this work, these algorithms show great ability to cope with the high-dimensional data and, naturally, they are the best option for malware detection task that is the classification since they are able to deal with the aforementioned problems.

➢ *Random Forest uses Multiple Decision Tress. Here are Some Basic Formulas and Concepts Related to Decision Trees:*

- *Entropy (H(S)):*

Entropy is a measure of impurity or disorder in a dataset. For a binary classification problem (two classes, typically 0 and 1), the entropy of a dataset S is calculated as:

$$H(S) = -p(0) * \log2(p(0)) - p(1) * \log2(p(1))$$

Where:

✓ p(0) is the proportion of class 0 instances in S.
✓ p(1) is the proportion of class 1 instances in S.

The goal is to minimize entropy by splitting the dataset into subsets that are as pure as possible.

- *Information Gain (IG):*

Information gain measures the reduction in entropy achieved by splitting a dataset based on a particular feature. For a feature F and a dataset S, the information gain is calculated as:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1102

IG(S, F) = H(S) - Σ((|S_v| / |S|) * H(S_v))

Where:

- ✓ H(S) is the entropy of the original dataset S.
- ✓ S_v represents the subset of S when the feature F has value v.
- ✓ |S| is the size of the dataset S.

A higher information gain indicates a better feature for splitting the dataset.

- *Gini Impurity (Gini(S)):*

Gini impurity is another measure of impurity used in decision trees. For a dataset S with multiple classes, the Gini impurity is calculated as:

Gini(S) = 1 - Σ(p_i^2)

Where:

- ✓ p_i is the proportion of instances belonging to class i in S.

IV. RESULTS

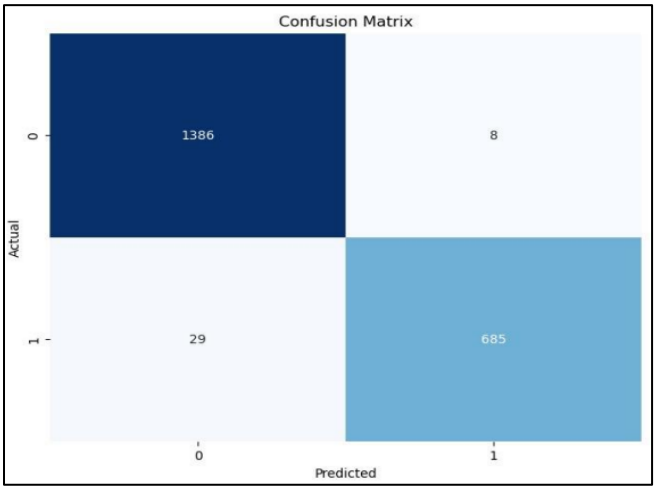

Fig 2 Confusion Matrix for RF

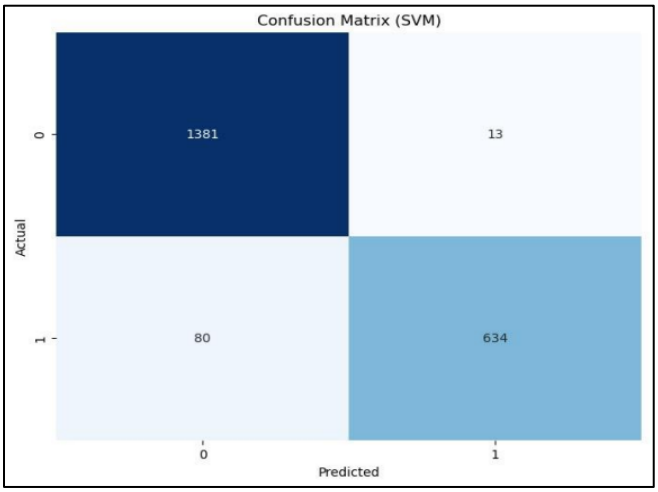

Fig 3 Confusion Matrix for SVM

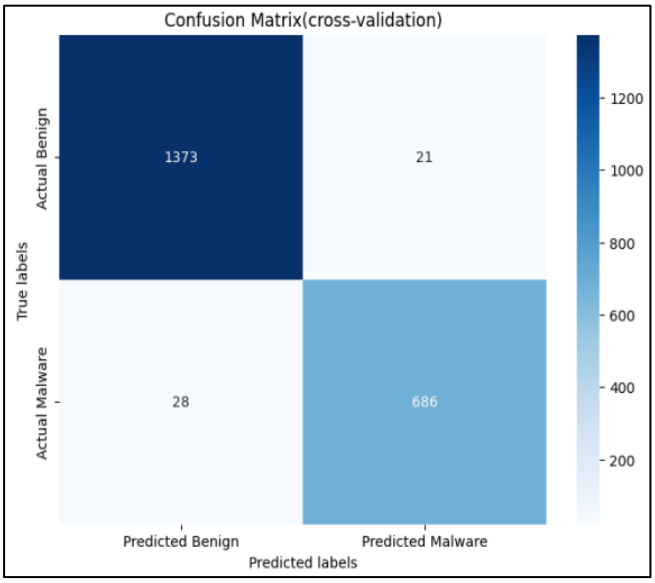

Fig 4 Confusion Matrix for Cross Validation

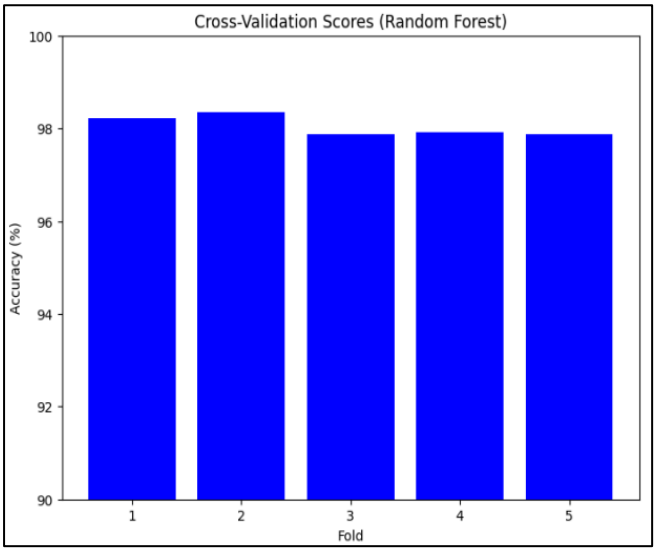

Fig 5 Cross Validation Scores Graph

V. SCOPE

The future prospects of implementing Support Vector Machine (SVM), Random Forest, and cross-validation methodologies in cybersecurity present a promising and diverse landscape. These sophisticated algorithms offer heightened capabilities in detecting intricate malware and ransomware threats, thereby ensuring more resilient security measures. Their integration into existing frameworks holds the promise of fortifying defenses against evolving cyber threats, while their scalability and adaptability ensure efficient performance in managing extensive data processing demands. Collaborative research endeavors and educational initiatives are poised to benefit from these techniques, fostering innovation andknowledge dissemination within the cybersecurity domain. In summary, the utilization of SVM, Random Forest, and cross-validation methodologies holds significant potential for advancing cybersecurity resilience and staying ahead of emerging threats.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1102

VI. CONCLUSION

Our findings reveal that the best set of parameters for our dataset consisted of 'max_depth' = 20, 'min_samples_split' = 5, and 'n_estimators' = 50. Employing these optimized parameters resulted in a significantly improved accuracy of 98.20% on our classification task. This highlights the importance of parameter tuning in maximizing the performance of Random Forest models. Furthermore, we discussed the implications of our findings and emphasized the significance of hyperparameter optimization in machine learning model development. By fine-tuning the parameters of Random Forest, practitioners can achieve superior performance in classification tasks across various domains.

In conclusion, this research underscores the effectiveness of optimizing Random Forest parameters for enhancing classification accuracy. Our results provide valuable insights for researchers and practitioners seeking to leverage Random Forest efficiently in their machine learning applications.

REFERENCES

[1]. Tamás, Csongor, Dorottya Papp and Levente Buttyán. “SIMBIoTA: Similarity-based Malware Detection on IoT Devices.” International Conference on Internet of Things, Big Data and Security, doi:10.5220/0010441500580069

[2]. S. Agarkar and S. Ghosh, "Malware Detection & Classification using Machine Learning," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), Gunupur Odisha,India, 2020, pp. 1-6, doi: 10.1109/iSSSC50941.2020.9358835.

[3]. S. A. Roseline, A. D. Sasisri, S. Geetha and C. Balasubramanian, "Towards Efficient Malware Detection and Classification using Multilayered Random Forest Ensemble Technique," 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 2019, pp. 1-6, doi: 10.1109/CCST.2019.8888406.

[4]. K. Sethi, R. Kumar, L. Sethi, P. Bera and P. K. Patra, "A Novel Machine Learning Based Malware Detection and Classification Framework," 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 2019, pp. 1-4, doi: 10.1109/CyberSecPODS.2019.8885196.

[5]. N. A. Anuar, M. Zaki Mas’ud, N. Bahaman and N. A. Mat Ariff, "Analysis of Machine Learning Classifier in Android Malware Detection Through Opcode," 2020 IEEE Conference on Application, Information and Network Security (AINS), Kota Kinabalu, Malaysia, 2020, pp. 7-11, doi: 10.1109/AINS50155.2020.9315060.

[6]. P. Priyadarshan, P. Sarangi, A. Rath and G. Panda, "Machine Learning Based Improved Malware Detection Schemes," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 925-931, doi: 10.1109/Confluence51648.2021.9377123.

[7]. M. Masum, M. J. Hossain Faruk, H. Shahriar, K. Qian, D. Lo and M. I. Adnan, "Ransomware Classification and Detection with Machine Learning Algorithms," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0316-0322,doi:10.1109/CCWC54503.2022.9720869.

[8]. SH Kok, Azween Abdullah, NZ Jhanjhi and Mahadevan Supramaniam, "Ransomware, Threat and Detection Techniques: A Review" 2019 International Journal of Computer Science and Network Security, VOL.19 No.2, Februar 2019

[9]. Ham, Hyo-Sik & Kim, Hwan-Hee & Kim, Myung-Sup & Choi, Mi-Jung. (2014). Linear SVM- Based Android Malware Detection for Reliable IoT Services. Journal of Applied Mathematics. 2014. 1-10. 10.1155/2014/594501.