2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG) | 979-8-3503-5815-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SEB4SDG60871.2024.10630194

DEVELOPMENT OF MALWARE DETECTION USING MACHINE LEARNING

1st Victoria Oluwaseyi Adedayo-Ajayi
Department of Computer Science, Landmark University Omu-Aran
Nigeria
Adedayo-ajayi.victoria@lmu.edu.ng

2nd Roseline Oluwaseun Ogundokun
Department of Multimedia Engineering,
Kaunas University of Technology, Lithuania
Department of Computer Science, Landmark University Omu Aran,
Nigeria.
rosogu@ktu.lt, ogundokun.roseline@lmu.edu.ng
SDG 11, Sustainable Cities and Communities,

3rd Hakeem Babatunde Akande
Department of Computer Science. University of Ilorin, Ilorin,
Nigeria.
ashiedu.nicole@lmu.edu.ng

4th Akinbowale Nathanial Babatunde
Department of Computer Science, Kwara State University, Malete, Nigeria
akinbowale.babatunde@kwasu.edu.ng

5nd Dubem
Department of Computer Science, Landmark University Omu Aran, Nigeria.
dubem-edozie.emeka@lmu.edu.ng

6th Aderemi Emmanuel Tunbosun
Department of Computer Science, Landmark University Omu Aran, Nigeria.
aderemi.tunbosun@lmu.edu.ng
SDG 4

7th Adekunle Kolawole Ojo
Department of Computer Science, Landmark University Omu Aran, Nigeria.
adekunle.kolawole@lmu.edu.ng

***Abstract*— With the continuous rise in computer tech, programmers and hacking occurrences are expanding and require additional security requests. Malware has been an extraordinary torment for computer clients around the world. Big organizations have gone into gigantic misfortunes due to an escape clause in security. In the current world of technology, Machine learning is considered a long-standing time and the foremost capable concept in innovation. The point is to utilize the machine learning concept and construct a show using gathering calculations that can be prepared effectively to identify malware in a framework. This extension is around the comparative consideration of standard machine learning and crossbreed calculations to determine how much effect a crossbreed machine learning calculation can have in distinguishing malware from traditional machine learning algorithms. In this study, we employ different machine-learning algorithms to detect malware. Our results show that Random Forest outperformed other machine learning algorithms with accuracy and average precision of 98% and 92%, respectively. These outcomes are crucial as malware becomes more prevalent and advanced.**

Keywords*— *Malware detection, Machine Learning, Random Forest.

I. INTRODUCTION

Cybersecurity against threats such as malware encompasses a range of safeguards and procedures intended to thwart, identify, and alleviate the risks posed by harmful software (malware) such as viruses, worms, Trojans, ransomware, and spyware [1, 2]. Its objective is to prevent the unauthorized access and compromise of computer systems, networks, and data [3]. Cybersecurity involves using several tactics, including antivirus software, intrusion detection systems, firewalls, and user education, to protect digital assets and ensure information integrity, availability, and confidentiality [4, 5]. Cybersecurity is crucial in protecting against malware as it is a vital defence mechanism in our highly interconnected digital society. Cyberattacks can disrupt operations, compromise sensitive data, and cause significant financial and reputational harm to individuals, organizations, and governments [6]. It plays a crucial role in protecting digital ecosystems by preserving privacy, mitigating data breaches, and ensuring the uninterrupted operation of vital services, making it an essential component. Malware is any software meant to harm one computer, server, or network at a time. The attack environment has expanded and become a problem, threatening users' privacy, as the Internet has increased and heterogeneous devices have been connected over the network [7]. Signature-based detection cannot detect current malware since it is either polymorphic or metamorphic [8]. Detecting and eradicating malware has gotten increasingly challenging since it is so readily available on the Internet. Anti-detection methods are freely available, making it possible for anybody to launch an attack or malicious program without prior knowledge. They're adopting practices that allow them to quickly upgrade to a newer version so they don't get caught.[9]. Machine learning has had a significant and transformative effect on several industries, including

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Authorized licensed use limited to: University of Johannesburg. Downloaded on August 17,2024 at 09:23:39 UTC from IEEE Xplore. Restrictions apply.

cybersecurity. Malware detection modules analyze the data they have gathered and been educated on to ascertain whether a particular program or network connection presents a security risk. The array contains the numbers 10 and 11. Imagine a machine learning system that can openly communicate the fundamental principles behind the patterns it has identified. Machine learning algorithms may enhance their predictive capability by receiving feedback on their performance in past tasks and adjusting appropriately. The number 12 is enclosed in square brackets.

The demand for a new detection technology is very urgent. Machine learning approaches can significantly improve system safety. Machine learning algorithms can classify malware as infected or legitimate.

As a result, machine learning algorithms such as Decision Tree, Random Forest, XG Boost, and CatBoost were proposed in this study for malware detection.

The article is organized in the following sequence. Section 2 elaborates on previous studies in malware detection using machine learning techniques. Section 3 delineates the materials and methodologies used, whereas Section 4 elucidates the results and conclusions derived. Furthermore, Section 5 encompasses a thorough analysis of current models, followed by the formulation of findings and the presentation of suggestions for future research avenues.

II. Related Works

Various methods for detecting malware have been developed in response to the rise in malicious software activity. Scientists use various big data technologies and machine learning methods to identify dangerous code. Conventional machine-learning techniques for detecting malware take a significant amount of time to analyze, but they can accurately identify new and developing malware.

Ji et al. [10] propose SigPID, a malware detection method that utilizes permission use analysis to address the growing prevalence of Android malware. SigPID employs machine learning algorithms to categorize various types of malware and harmless applications accurately. The findings demonstrated that a support vector machine classifier attained a precision, recall, accuracy, and F-measure of 90%.

Dhalaria and Gandotra (2013) propose a hybrid approach to identify and categorize Android malware. The suggested methodology combines static and dynamic analysis approaches to detect malicious apps and organize them into distinct malware families. Machine learning models were built to identify malware and classify families by extracting information from Android applications' static and dynamic behaviours. The hybrid technique has been shown to be helpful in reliably identifying and classifying Android malware, making a valuable contribution to mobile security and helping to avoid harmful activity on Android devices.

Akhtar & Feng [14] used machine learning methods to discern and detect malware. The authors demonstrated the feasibility of identifying unauthorized traffic on computer systems and enhancing network security by using the results of malware analysis and detection via machine learning techniques to calculate the variation in correlation symmetry integral.

Guan & Li [15] provide an ensemble learning system that uses dangerous permissions as features to train a classifier to assess the presence of malware in a mobile application. The performance of the suggested model was evaluated for malware detection using a collection of real-world Android app datasets, including both harmful and benign applications. The outcomes of their methodology unequivocally demonstrate that the model can efficiently identify malware with a high level of precision.

Chowdhury, Rahman & Islam [16] provide a complete framework that utilizes data mining and machine learning classification approaches to identify and categorize dangerous software. The purpose of this framework is to safeguard sensitive data from harmful attacks. The author utilizes a strong and effective method for classifying and detecting malware by examining signature and anomaly-based characteristics.

In this paper, we focus on analyzing the performance of machine learning classifiers to achieve robust and accurate malware detection.

III. Materials and Methods

This section presents the various processes and components for malware detection in a typical machine-learning workflow. The proposed research methodology of this study is given in Fig.1.

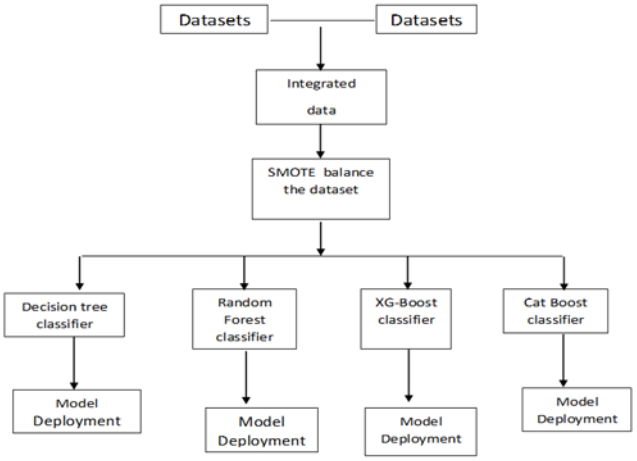


Fig. 1. Proposed Framework.

A. Data Collection and Processing

The initial stage of the model training is the collection of large datasets, which were acquired from Kaggle. Machine learning models will perform poorly for severely imbalanced data if they are fed labelled data from a minority class. Synthetic minority over-sampling technique SMOTE was used to balance the dataset.

B. Cross Validation

For machine learning models to perform correctly with actual data, they must be resilient. According to the process of determining whether numerical results characterizing hypothesized relationships between variables are acceptable as

Authorized licensed use limited to: University of Johannesburg. Downloaded on August 17,2024 at 09:23:39 UTC from IEEE Xplore. Restrictions apply.

descriptions, known as validation, the model must be free of bias and variation. Cross-validation is critical in developing a robust machine-learning model that can generalize to new datasets. We employed K-fold cross-validation to balance an imbalanced dataset and include all data points in the training phase rather than a random portion as in standard splitting methods.

C. Performance Evaluation

1. Accuracy: Accuracy measures the overall correctness of our model's predictions. It is the ratio of correctly classified instances to the total cases.

$$\text{Accuracy} = \frac{Number\ of\ correctly\ classified\ instances}{Total\ Number\ of\ instances} \times 100$$

2. Precision: Precision quantifies the proportion of correctly predicted positive instances (phishing emails) out of all the cases indicated as positive.

$$\text{Precision} = \frac{\text{True Positives}}{True\ Positives + \text{False Positives}} \times 100$$

IV. Results and Discussion

The study is implemented using a Python programming language in the Jupiter Notebook environment. Four machine learning models were used in this study. The models are decision tree (DT), random forest (RF), XGBoost and CatBoost. After the experiment, the proposed models, DT, RF, XG Boost and CatBoost, achieved an impressive accuracy of 0.97, 0.98, 0.96 and 0.96, respectively, as shown in Table 1.

The ML models also demonstrated a high precision of 0.87, 0.92, 0.86, and 0.89, indicating that they were highly likely to be correct when they predicted any malware.

TABLE I. Models Performance Evaluation

Model	Accuracy	Precision	PR Value
DT	0.97	0.87	0.85
RF	**0.98**	**0.92**	0.92
XGBoost	0.96	0.86	**0.98**
Catboost	0.96	0.89	0.90

Table 1 compares four machine learning models - Decision Tree (DT), Random Forest (RF), XGBoost, and Catboost - concerning their performance metrics: accuracy, precision, and PR value.

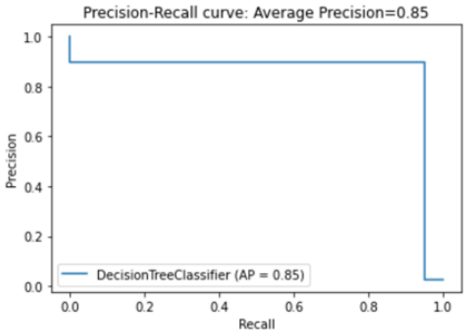
The Decision Tree (DT) model has a high level of accuracy, achieving a score of 0.97. It also exhibits strong precision, with a value of 0.87. However, the PR (Precision-Recall) value is somewhat lower at 0.85. This implies that while DT demonstrates reliability and inaccurate predictions, it may have some constraints in its ability to retain information.

The Random Forest (RF) model has exceptional performance, achieving the most remarkable accuracy (0.98) and precision (0.92) compared to other models. It also performs a PR value of 0.92, which is consistent with the precision score. This suggests a robust equilibrium between accurately forecasting positive instances and the ratio of real positives.

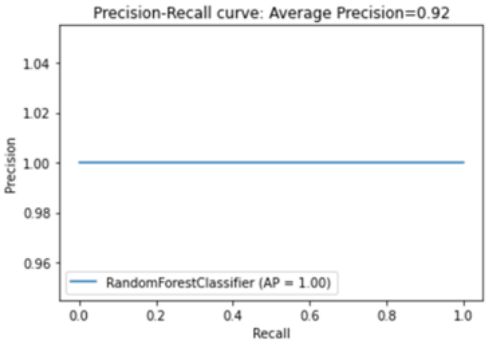
XGBoost has somewhat lesser accuracy (0.96) and precision (0.86) but distinguishes itself with the most outstanding PR value (0.98). The high PR score indicates exceptional accuracy in detecting positive occurrences among the real positive situations.

Catboost demonstrates constant performance, achieving an accuracy of 0.96, a precision of 0.89, and a PR value of 0.90. This presents a reasonable approach in terms of both overall accuracy and precision in forecasts.

The study results were also evaluated using precision-recall curves in Figures 2 to 5.

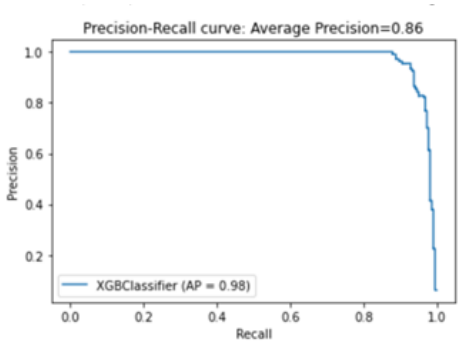


(a)

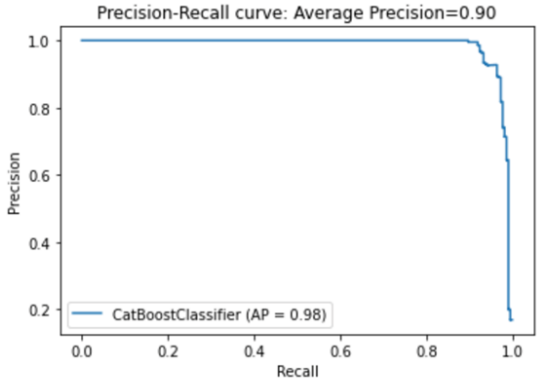


(b)

Authorized licensed use limited to: University of Johannesburg. Downloaded on August 17,2024 at 09:23:39 UTC from IEEE Xplore. Restrictions apply.



(c)



(d)

Fig. 2. Precision-recall Curve for the Models implemented.

V. CONCLUSION

In conclusion, this article comprehensively explores the development of a malware detection model using machine learning techniques. Amidst the escalating malware threats, the research delves into various machine learning algorithms – Decision Tree, Random Forest, XGBoost, and Catboost – to discern their effectiveness in identifying malware. The study stands out in its systematic approach, employing data collection, preprocessing, and cross-validation to enhance the model's robustness. The results reveal that Random Forest outshines others in accuracy and precision, underscoring its potential in cybersecurity applications. This work contributes significantly to malware detection and paves the way for future research in the domain, particularly in refining and advancing machine learning methodologies for cybersecurity.

REFERENCES

[1] Ogundokun, R. O., Awotunde, J. B., Sadiku, P., Adeniyi, E. A., Abiodun, M., & Dauda, O. I. (2021). An enhanced intrusion detection system using particle swarm optimization feature extraction technique. Procedia Computer Science, 193, 504-512.

[2] Kumar, I. (2023). Emerging threats in cybersecurity: a review article. *International Journal of Applied and Natural Sciences*, *1*(1), 01-08.

[3] Ogundokun, R. O., Odusami, M., Sisodia, D. S., Awotunde, J. B., & Tiwari, D. P. (2022, October). A Novel PCA-Logistic Regression for Intrusion Detection System. In International Conference on Information Systems and Management Science (pp. 575-588). Cham: Springer International Publishing

[4] Ogundokun, R. O., Misra, S., Babatunde, A. N., & Chockalingam, S. (2022, May). Cyber intrusion detection system based on machine learning classification approaches. In 2022 International Conference on Applied Artificial Intelligence (ICAPAI) (pp. 1-6). IEEE.

[5] Perwej, Y., Abbas, S. Q., Dixit, J. P., Akhtar, N., & Jaiswal, A. K. (2021). A systematic literature review on the cyber security. *International Journal of scientific research and management*, *9*(12), 669-710.

[6] Ogundokun, R. O., Basil, U., Babatunde, A. N., Abdulahi, A. T., Adenike, A. R., & Adebiyi, A. A. (2023, April). Intrusion Detection Systems Based on Machine Learning Approaches: A Systematic Review. In 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG) (Vol. 1, pp. 01-04). IEEE.

[7] Faruk, M. J. H., Shahriar, H., Valero, M., Barsha, F. L., Sobhan, S., Khan, M. A., ... & Wu, F. (2021, December). Malware detection and prevention using artificial intelligence techniques. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 5369-5377). IEEE.

[8] Bulazel, A., & Yener, B. (2017). A Survey On Automated Dynamic Malware Analysis Evasion and Counter-Evasion: PC, Mobile, and Web. Reversing and Offensive-Oriented Trends Symposium.

[9] Agarkar, S.; Ghosh, S. Malware detection & classification using machine learning. In Proceedings of the 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), Gunupur Odisha, India, 16–17 December 2020

[10] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an and H. Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3216-3225, July 2018, doi: 10.1109/TII.2017.2789219.

[11] 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.

[12] Sharma, S.; Krishna, C.R.; Sahay, S.K. Detection of advanced malware by machine learning techniques. In Proceedings of the SoCTA 2017, Jhansi, India, 22–24 December 2017.

[13] Dhalaria, M. & Gandotra, E. (2020). A Hybrid Approach for Android Malware Detection and Family Classification. International Journal of Interactive Multimedia and Artificial Intelligence. In Press. 1. 10.9781/ijimai.2020.09.001.

[14] Akhtar, M.S.; Feng, T. Malware Analysis and Detection Using Machine Learning Algorithms. Symmetry 2022, 14, 2304. https://doi.org/10.3390/sym14112304

[15] Guan, S., & Li, W. (2022). EnsembleDroid: A Malware Detection Approach for Android System based on Ensemble Learning. 2022 IEEE MIT Undergraduate Research Technology Conference (URTC), 1-5.

[16] Chowdhury, M.; Rahman, A.; Islam, R. Malware Analysis and Detection Using Data Mining and Machine Learning Classification; AISC: Chicago, IL, USA, 2017; pp. 266–274.

Authorized licensed use limited to: University of Johannesburg. Downloaded on August 17,2024 at 09:23:39 UTC from IEEE Xplore. Restrictions apply.