Optimal Ensemble Learning for Automated Android Malware Detection in Cyber Security Applications

[1]M. Akshay Sai, [2]D. Sampath, [3]G. Sarala, [4]Dr. R. SanthoshKumar
[1,2,3]UG Scholar, Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[4]Professor and Head, Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
hodcse@smec.ac.in

Abstract:

Current technological advancement in computer systems has transformed the lives of humans from real to virtual environments. Malware is unnecessary software that is often utilized to launch cyberattacks. Malware variants are still evolving by using advanced packing and obfuscation methods. These approaches make malware classification and detection more challenging. New techniques that are different from conventional systems should be utilized for effectively combating new malware variants. Machine learning (ML) methods are ineffective in identifying all complex and new malware variants. The deep learning (DL) method can be a promising solution to detect all malware variants. This paper presents an Automated Android Malware Detection using Optimal Ensemble Learning Approach for Cybersecurity (AAMD-OELAC) technique. The major aim of the AAMD-OELAC technique lies in the automated classification and identification of Android malware. To achieve this, the AAMD-OELAC technique performs data pre processing at the preliminary stage. For the Android malware detection process, the AAMD-OELAC technique follows an ensemble learning process using three ML models, namely Least Square Support Vector Machine (LSSVM), kernel extreme learning machine (KELM), and Regularized random vector functional link neural network (RRVFLN). Finally, the hunter-prey optimization (HPO) approach is exploited for the optimal parameter tuning of the three DL models, and it helps accomplish improved malware detection results. To denote the supremacy of the AAMD-OELAC method, a comprehensive experimental analysis is conducted. The simulation results portrayed the supremacy of the AAMD-OELAC technique over other existing approaches.

Key words: Malware, cyberattacks, Machine learning, Deep learning, Optimal Ensemble Learning, data preprocessing, Least Square Support Vector Machine, kernel extreme learning machine, Regularized random vector functional link neural network, hunter-prey optimization.

1.INTRODUCTION

Cybersecurity is becoming a main area of immediate concern to network engineers and computer scientists, so satisfying The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose . solutions to several problems are in order. Consequently, the fast technological developments and their inherent integrations in every aspect of lifestyles, various malware apps, and targets become well-identified and studied. Android malware is the malware variety that gained significant interest in the web world. One common operating system is Android, which dominates the operating system market. Malware invasive methods emerge for avoiding identification, as few malware applications have more than 50 parameters that make detection a difficult one. Hence, it is essential to devise techniques that deal with the continuous growth of Android malware to find it, deactivate or remove it efficiently. All these difficulties engage scholars in the area and urge them to continue more research to find malware and manage it properly. Thus, researchers have developed three mechanisms to find Android malware such as dynamic, static, and hybrid analysis methods. Static analysis extracts the features that assist in identifying harmful performance for apps without a demanding actual application deployment. But this kind of analysis suffered from code obfuscation methods which assist help malware author to avoid static methods. Dynamic analysis can be used for determining the malware of apps in their runtime. Commonly, the static analysis feature offers the capability of locating the malware element using source code, while the dynamic analysis feature offers the capability of finding the location of malware in a runtime environment. Android developers and users can be exposed to unnecessary risks and dangers with malware [8]. This study covers malware detection methods. The detection of malware using the ML model includes Android Application Packages (APKs) for deriving an appropriate set of features. Deep learning (DL) and machine learning (ML) approaches can be used for recognizing malicious APKs. Like malware detection, vulnerability detection in software code has two stages: training ML on derived attributes to find vulnerable code segments and feature generation utilizing code analysis [10]. This paper presents an Automated Android Malware Detection using Optimal Ensemble Learning Approach for Cybersecurity (AAMD-OELAC) technique. The AAMDOELAC technique performs data preprocessing at the preliminary stage. For the Android malware detection process, the AAMD-OELAC 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 AAMD-OELAC approach, a comprehensive experimental analysis is carried out. In short, the key contributions are listed as follows.

- An intelligent AAMD-OELAC technique comprising data pre processing, ensemble learning, and HPO-based hyper parameter tuning is presented for Android malware detection. To the best of our knowledge, the AAMD-OELAC technique never existed in the literature.
- Perform ensemble learning-based classification process comprising LS-SVM, KELM, and RRVFLN models for Android malware detection.
- The combination of the HPO algorithm and ensemble learning process improves the detection accuracy of Android malware. By utilizing multiple classifiers and optimization strategies, the model can effectively identify malicious patterns and behaviours in Android applications.

2. LITERATURE SURVEY

Shaukat et al. devise a new DL-related method for detecting malware. It delivered superior outcomes to classical methods by merging dynamic and static analysis benefits. Firstly, it visualizes a portable executable (PE) file as coloured images. Secondly, it extracted deep features from colour images utilizing fine-tuned DL method. Thirdly, it finds malware related to the deep features of SVM. Geremias et al. presented a method using image-based DL called novel multi-view Android malware identification, applied threefold. Firstly, as per the many feature sets in multi-view settings, apps were assessed, thereby raising the data presented for the classification. Secondly, the derived feature set is transformed into image formats while preserving the essential elements of data distribution, keeping the data for the classifier task. Thirdly, built images are collectively depicted in one shot, all in a predefined image channel, allowing the implementation of DL structure.

Kim et al. modelled a malware detection system called MAPAS that attains higher precision and adaptable use of computational resources. MAPAS examined the performances of malicious apps based on API call graphs of them through CNN. However, the presented MAPAS technique does not utilize a classifier method produced by CNN, it uses CNN to find typical attributes of the API call graph of malware. Fallah and Bidgoly developed a technique related to LSTM for detecting malware-having the capability of differentiating benign and malware samples and identifying and detecting unseen and new types of malware. In this study, the author has executed many studies to show the abilities of the presented technique, including new malware family detection, malware identification, malware family identification, as well as assessing the minimal time needed to find malware Sihag et al. introduced DL-based Android malware identification with the use of DYnamic features (De-LADY), a resilient obfuscation method. It has used behavioural features from dynamic analysis of an application performed in the emulated setting. Wang et al. present a hybrid method related to DAE and CNN. Firstly, to enhance the precision of malware detection, the author reconstructed the high-dimensional feature of apps and used many CNN to find Android malware. Secondly, to diminish the training period, the author used DAE as a pre-training approach for CNN. With the consolidation, DAE and CNN method (DAE-CNN) can study flexible patterns quickly. Yadav et al. presented a performance comparison of 26 existing pretrained CNN methods in Android malware detection. Depending on the outcomes, to find Android malware, an EfficientNet-B4 CNN-based approach was devised with the use of an image-based malware representation of the Android DEX file. From the malware images, EfficientNet-B4 extracted relevant attributes. Masum and Shahriar devised a DL structure named Droid-NNet, for classifying an study flexible patterns quickly. Yadav et al. presented a performance comparison of 26 existing pretrained CNN methods in Android malware detection. Depending on the outcomes, to find Android malware, an EfficientNet-B4 CNN-based approach was devised with the use of an image-based malware representation of the Android DEX file. From the malware images, EfficientNet-B4 extracted relevant attributes. Masum and Shahriar devised a DL structure named Droid-NNet, for classifying malware. But this technique Droid-NNet, is a deep learner that surpasses existing cutting-edge ML approaches. Idrees et al. examine PIndroid – a new Permission and Intents-based structure to detect Android malware applications. As we know, PIndroid is the primary solution, which utilizes a group of permissions and purposes supplemented with Ensemble approaches for correct malware detection. In , the authors establish that once the concept drift was discussed, permissions create long-lasting and effectual malware detection methods. Taha and Barukab introduce a mechanism for Android malware classification utilizing optimizer ensemble learning depending on GA. The GA was utilized for optimizing the parameter settings from the RF technique for obtaining the maximum Android malware classifier accuracy. Sabanci et al. intended to categorize pepper seeds belonging to distinct cultivars with CNN techniques. Two methods are presented for classification. Initially, the CNN approaches (ResNet50 and ResNet18) are trained for pepper seeds. Secondary, diverse in the first, the features of pre-training CNN approaches are fused, and feature selection has been executed to the fused features. In , the authors examine recent algorithms utilized for Android Malware Detection. As a result, an outline of the Android system exposed the underlying processes and the problems facing its security structure.malware. But this technique Droid-NNet, is a deep learner that surpasses existing cutting-edge ML approaches

3. PROPOSED METHODOLOGY

This paper presents an Automated Android Malware Detection using Optimal Ensemble Learning Approach for Cybersecurity (AAMD-OELAC) technique. The AAMDOELAC technique performs data preprocessing at the preliminary stage. For the Android malware detection process, the AAMD-OELAC 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 AAMD-OELAC approach, a comprehensive experimental analysis is carried out. The core malware detection mechanism relies on an ensemble learning approach, integrating three machine learning (ML) models: Least Square Support Vector Machine (LS-SVM), Kernel Extreme Learning Machine (KELM), and Regularized Random Vector Functional Link Neural Network (RRVFLN). These models work together to enhance classification accuracy. To further improve performance, the Hunter-Prey Optimization (HPO) algorithm is utilized for hyperparameter tuning of the ML models, refining their learning capabilities. This integration of ensemble learning and optimization strategies significantly enhances malware detection accuracy. The AAMD-OELAC technique demonstrates superior performance in detecting malicious activities in Android applications. A comprehensive experimental analysis validates its effectiveness compared to existing approaches. This method is novel, as no prior research has combined LS-SVM, KELM, and RRVFLN with HPO for Android malware detection. The ensemble learning strategy ensures that the model captures complex malicious patterns more effectively than individual classifiers. The inclusion of HPO enables optimal parameter selection, thereby refining model accuracy. By leveraging multiple classifiers and optimization algorithms, AAMD-OELAC exhibits enhanced malware detection capabilities, making it a robust solution for cybersecurity in Android devices.

Advantages

- An intelligent AAMD-OELAC technique comprising data pre processing, ensemble learning, and HPO-based hyper parameter tuning is presented for Android malware detection. To the best of our knowledge, the AAMD-OELAC technique never existed in the literature.
- Perform ensemble learning-based classification process comprising LS-SVM, KELM, and 9 RRVFLN models for Android malware detection.
- The combination of the HPO algorithm and ensemble learning process improves the detection accuracy of Android malware. By utilizing multiple classifiers and optimization strategies, the model can effectively identify malicious patterns and behaviours in Android applications.

4. EXPERIMENTAL ANALYSIS

Figure 1: Login Page

Figure 1 shows a login interface for a cybersecurity application titled Optimal Ensemble Learning for Automated Android Malware Detection in Cyber Security Applications.The interface offers login options for service providers and new users to register, indicating a multi-user platform for malware detection and analysis.

Figure 2: Train and Test DataSets

Figure 2 shows the trained and tested dataset results for different machine learning models used in the Android malware detection system. The models listed include Naive Bayes, LS-SVM, Logistic Regression, Decision Tree Classifier, and KNeighborsClassifier, with their respective accuracy scores displayed. The high accuracy values indicate effective model performance in distinguishing between benign and malicious Android applications. The structured tabs at the top suggest additional features and functionalities of the application, like dataset viewing, model evaluation, and performance analysis.

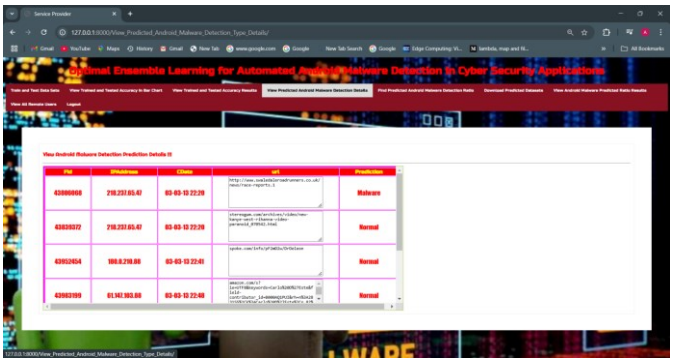

Figure 3: View Predicted android malware detection details

Figure 3 displays the View Predicted Android Malware Detection Details section of the application. It shows the prediction results for different Android applications, including details like PID, IP Address, Date, URL, and the Prediction (Malware or Normal). The highlighted "Malware" prediction indicates the detection of a malicious application, while "Normal" suggests safe applications. This feature helps in monitoring and analyzing network traffic to identify potential cybersecurity threats on Android devices.

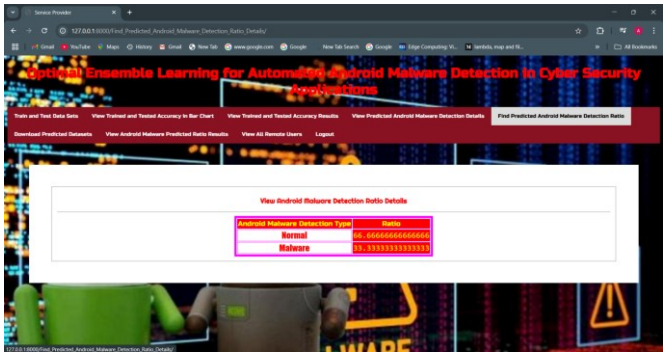

Figure 4: Find predicted android malware detection ratio

Figure 4 displays the Find Predicted Android Malware Detection Ratio section of the application. It shows the distribution of predicted malware detection results, indicating that 66.67% of the analyzed Android applications are labeled as Normal, while 33.33% are identified as Malware. This ratio provides insights into the prevalence of malicious apps in the dataset, helping assess the effectiveness and coverage of the malware detection system.

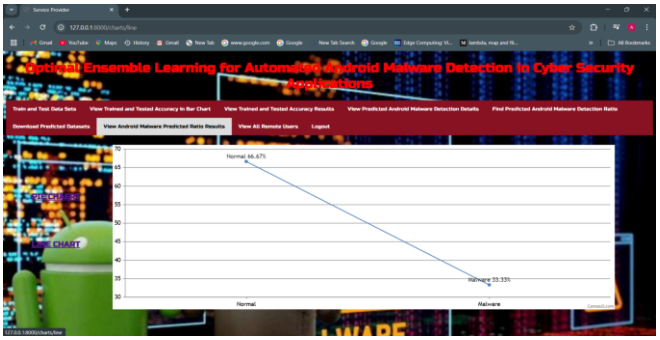

Figure 5: View Android malware predicted ratio results

Figure 5 presents the View Android Malware Predicted Ratio Results through a line chart, displaying the proportion of Normal and Malware predictions. The chart indicates that 66.67% of the analyzed Android applications are classified as Normal, while 33.33% are identified as Malware, aligning with the previously shown detection ratio data. The visualization helps in easily understanding the distribution of malware and benign applications, emphasizing the need for robust malware detection techniques in Android cybersecurity.

5. CONCLUSION

In this study, we have developed the design of the AAMD-OELAC technique for an accurate and automated Android malware detection process. The intention of the AAMD-OELAC approach focused on the automatic recognition and classification of Android malware. To achieve this, the AAMD-OELAC technique encompasses data preprocessing, ensemble classification, and HPO-based parameter tuning. For the Android malware detection process, the AAMD-OELAC technique follows an ensemble learning process using three ML models namely LS-SVM, KELM, and RRVFLN. Finally, the HPO algorithm is exploited for the optimal parameter tuning of the three DL

models and it helps in accomplishing improved malware detection results. To portray the supremacy of the AAMD-OELAC method, a wide-ranging experimental analysis is conducted. The simulation results portrayed the supremacy of the AAMDOELAC technique over other existing approaches. Future work could focus on developing more advanced techniques to capture and analyze fine-grained behaviours, enabling better detection of sophisticated malware. In addition, future work could explore privacy-preserving approaches such as secure multi-party computation or federated learning, which enable collaborative malware detection without compromising user privacy.

REFERENCES

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

[2] A. Albakri, F. Alhayan, N. Alturki, S. Ahamed, and S. Shamsudheen,‘‘Metaheuristics with deep learning model for cybersecurity and Android malware detection and classification,’’ Appl. Sci., vol. 13, no. 4, p. 2172, 2023.

[3] L. Hammood, İ. A. Doğru, and K. Kılıç, ‘‘Machine learning-based adaptive genetic algorithm for Android malware detection in auto-driving vehicles,’’ Appl. Sci., vol. 13, no. 9, p. 5403, 2023

[4] D.Wang, T. Chen, Z. Zhang, and N. Zhang, ‘‘A survey of Android malware detection based on deep learning,’’ in Proc. Int. Conf. Mach. Learn. Cyber Secur. Cham, Switzerland: Springer, 2023, pp. 228–242.

[5] S. S. Sammen, M. Ehteram, Z. Sheikh Khozani, and L. M. Sidek,‘‘Binary coati optimization algorithm- multi- kernel least square support vector machine-extreme learning machine model (BCOAMKLSSVM-ELM): A new hybrid machine learning model for predicting reservoir water level,’’ Water, vol. 15, no. 8, p. 1593, 2023.

[6] J. Geremias, E. K. Viegas, A. O. Santin, A. Britto, and P. Horchulhack, ‘‘Towards multiview Android malware detection through image-based deep learning,’’ in Proc. Int.Wireless Commun. Mobile Comput. (IWCMC), May 2022, pp. 572–577. 72516 VOLUME 11, 2023 IEEE Transaction on Machine Learning,Volume:11, 2023.

[7] K. Shaukat, S. Luo, and V. Varadharajan, ‘‘A novel deep learning-based approach for malware detection,’’ Eng. Appl. Artif. Intell., vol. 122, 2023, Art. no. 106030.

[8] H.-J. Zhu, W. Gu, L.-M. Wang, Z.-C. Xu, and V. S. Sheng, ‘‘Android malware detection based on multi-head squeeze-and-excitation residual network,’’ Expert Syst. Appl., vol. 212, 2023, Art. no. 118705.

[9] E. C. Bayazit, O. K. Sahingoz, and B. Dogan, ‘‘Deep learning based malware detection for Android systems: A comparative analysis,’’ Tehnički vjesnik, vol. 30, no. 3, pp. 787–796, 2023.

[10] M. Ibrahim, B. Issa, and M. B. Jasser, ‘‘A method for automatic Android malware detection based on static analysis and deep learning,’’ IEEE Access, vol. 10, pp. 117334–117352, 2022

[11] 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. 66,2022, Art. no. 103159.

[12] P. Bhat and K. Dutta, ‘‘A multi-tiered feature selection model for Android malware detection based on feature discrimination and information gain,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 34, no. 10, pp. 9464–9477, 2022. 49

[13] 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, 2022.

[14] S. Fallah and A. J. Bidgoly, ‘‘Android malware detection using network traffic based on sequential deep learning models,’’ Softw., Pract. Exper., vol. 52, no. 9, pp. 1987–2004, 2022.

[15] A. Guerra-Manzanares, H. Bahsi, and M. Luckner, ‘‘Leveraging the first line of defense: A study on the evolution and usage of Android security permissions for enhanced Android malware detection,’’ J. Comput. Virol. Hacking Techn., vol. 19, no. 1, pp. 65–96, 2022

[16] A. Taha and O. Barukab, ‘‘Android malware classification using optimized ensemble learning based on genetic algorithms,’’ Sustainability, vol. 14, no. 21, p. 14406, 2022.

[17] K. Sabanci, M. F. Aslan, E. Ropelewska, and M. F. Unlersen, ‘‘A convolutional neural network-based comparative study for pepper seed classification: Analysis of selected deep features with support vector machine,’’ J. Food Process Eng., vol. 45, no. 6, 2022, Art. no. e13955.

[18] P. Yadav, N. Menon, V. Ravi, S. Vishvanathan, and T. D. Pham, ‘‘Efficient- Net convolutional neural networks-based Android malware detection,’’ Comput. Secur., vol. 115, 2022, Art. no. 102622.

[19] Y. Zhao, L. Li, H. Wang, H. Cai, T. F. Bissyandé, J. Klein, and J. Grundy, ‘‘On the impact of sample duplication in machine-learning-based Android malware detection,’’ ACM Trans. Softw. Eng. Methodol., vol. 30, no. 3, pp. 1–38, 2021.

[20] V. Sihag, M. Vardhan, P. Singh, G. Choudhary, and S. Son, ‘‘De-LADY: Deep learningbased Android malware detection using dynamic features,’’ J. Internet Serv. Inf. Secur., vol. 11, no. 2, p. 34, 2021.