Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30949

Android Malicious Application Detection Model Based on Adversarial Distillation and Multi-modal Feature Weighting

Yanjie Xia

Xiangtan Institute of Technology, Xiangtan, China
xyj32456728@outlook.com

Abstract. As Android has become the most widely used mobile operating system, it has also become a primary target for cyberattacks, leading to a continuous surge in malicious applications with increasingly complex behaviors. Traditional malware detection approaches struggle to maintain stability and robustness against such dynamic threats. To address these issues, an Android malicious application detection model ADMFW-MalDet based on adversarial distillation and multimodal feature weighting is proposed. Through the multimodal feature weighting and adversarial distillation framework optimized by quantum particle swarm optimization, the teacher-student adversarial distillation fanatic is adopted to enhance the robustness of the model. Experiments were conducted on the Drebin dataset and a high evaluation rate was achieved through five evaluations, and significantly enhancing robustness and adaptability. These results not only validate the effectiveness of the adversarial distillation and multimodal feature weighting strategies but also provide a new paradigm for constructing adaptive and resilient Android malware detection systems.

Keywords: Android malicious application detection, adversarial distillation, ulti-modal feature weighting

1. Introduction

Android is a Linux kernel based open source operating system developed by Google Inc., is the most popular and largest installed base, worldwide on smart gadgets. Android is now also providing operating platform for IoT devices called Android Things [1]. While Android is the largest operating system in the mobile market, it has also become a major target of cyber attacks. McAfee found that 6.2% of the threats discovered on Google in 2022 belonged to the communication category. These threats were mainly malicious software disguised as SMS and messaging applications [2], which invaded users and had a significant impact on their privacy. Kaspersky discovered that malicious applications masquerading as legitimate VPN tools and games threaten user security through phishing and data theft [3], and the methods they employ pose a serious threat to mobile users.

The detection of malicious Android applications is becoming increasingly important, and machine learning methods are mostly adopted in domestic and foreign research. Jyotiprakash Mishra et al. evaluated the impact of duplicates in the Drebin dataset on the accuracy of malware detection models [4]. In this experiment, Yung-Feng Lu et al. adopted SVM and permission features, achieving a detection accuracy rate of up to 99% for unknown Android malware [5]. Akshay

DOI: 10.54254/2755-2721/2026.TJ30949

Agarwal et al. found that deep neural networks are vulnerable even to memorized data by adding small perturbations to the training images and conducting data augmentation [6]. Navaprakash. N et al. compared R-CNN with CNN using the Kaggle dataset. The results showed that R-CNN had higher accuracy and statistical significance in the text extraction task [7].

Although significant progress has been made in existing research, Android malware detection still faces three core challenges: First, traditional feature selection methods are statically rigid and cannot adapt to the changes in the importance of malware features. Second, malware variants and adversarial attacks lead to a decline in model detection performance. Third, the one-sidedness of single-index evaluation and the incomplete assessment of model performance.

In response to the above issues, this paper proposes the ADMFW-MalDet detection model. Through a multimodal feature weighting and adversarial distillation framework optimized by quantum particle swarm optimization, an end-to-end malware detection solution is constructed. Quantum optimization algorithm is the core idea of this study is to use the global search ability of optimize the feature space, by against the robustness and efficiency of distillation framework balance model, and establish the comprehensive evaluation index to verify the effectiveness of the method.

The core contributions of this article lie in three aspects.

The dynamic weighting mechanism of multimodal features is inspired by the first quantum. Traditional feature methods often fall into local optimization. In this study, the quantum particle swarm optimization algorithm is introduced to achieve dynamic feature weighting, which can break away from local optimum.

The second adversarial distillation framework builds a teacher-student network architecture. By transferring the robust teacher model obtained from adversarial training to a compact student network, the student model maintains a streamlined architecture while also acquiring the robustness and generalization ability of the teacher model.

The third comprehensive five-index evaluation system, which particularly introduces MCC to comprehensively consider true positives, true negatives, false positives and false negatives, providing more reliable performance evaluation in category-imbalanced datasets. The comprehensive fitness function integrates the five indicators by weighting, providing direction for model optimization.

DOI: 10.54254/2755-2721/2026.TJ30949

2. Method

2.1. Overall flowchart

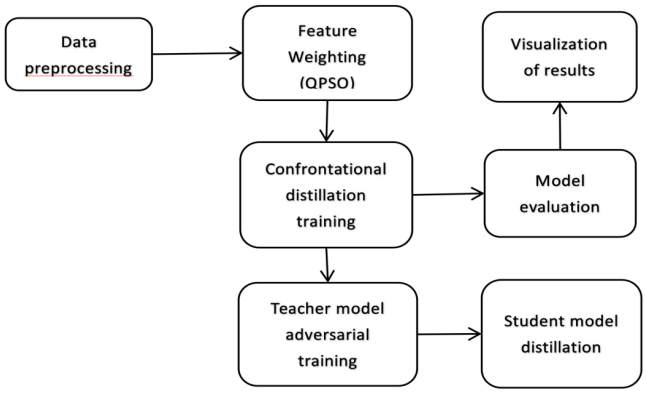

Figure 1. Flowchart (picture credit: original)

Before the experiment starts, draw the flowchart to ensure that the direction of the subsequent experiment does not deviate and to guarantee better progress of the experiment, as shown in Figure 1. Flowchart. After data acquisition, the first step is the data preprocessing stage, which processes the data, and then divides the data set into training set, verification set and test set. The second step is feature weighting, which adopts the Quantum Particle Swarm Optimization Algorithm (QPSO) to achieve dynamic allocation of feature importance. The third step is adversarial distillation training. First, the adversarial training of the teacher model is conducted, using FGSM attacks to generate adversarial samples to train the deep teacher network to enhance robustness. Knowledge of distillation process, then the teacher model of knowledge transfer to students of lightweight model.The fourth step is model evaluation. Calculation accuracy, accuracy and recall rate, F1 score and the Matthews correlation coefficient, and compared with the traditional method (random forest, support vector machine (SVM), logistic regression) comparative analysis. The fifth step is to visualize the results to verify the model's performance and interpretability.

2.2. Data acquisition and processing

This study uses the Drebin dataset, which has been used to develop and evaluate a multilevel classifier fusion method for Android malware detection. The data set by 15036 applications from Drebin project of 215 properties extracted characteristic vector. Of these, 5,560 were malicious and 9,476 were benign.

According to the flow chart in Figure 1, the data preprocessing stage is carried out. The original dataset was loaded and preliminarily analyzed and processed. In response to the possible label heterogeneity issues in the original dataset, the influence of feature scale on model training, and the unique multimodal feature structure of the dataset, the data were uniformly processed. This includes dataset standard and distribution optimization processing, intelligent label encoding and distribution optimization, type consistency conversion, robustness exception handling, and missing value filling. Data set is divided into training set (64%) (16%), validation set and test set (20%), to ensure the generalization ability of model evaluation and balanced type distribution.

DOI: 10.54254/2755-2721/2026.TJ30949

2.3. Dynamic weighting mechanism for multimodal features

Traditional particle swarm optimization (POS) algorithm is based on the position - velocity model of classical mechanics, but this study uses quantum particle swarm optimization algorithm.Based on the concept of quantum mechanics, QPSO describes the particle state by wave function, which significantly improves the global search ability and convergence performance [8,9].

$$p_{i,n}^{j} = \frac{c_1 r_{i,n}^{j} P_{i,n}^{j} + c_2 R_{i,n}^{j} G_n^{j}}{c_1 r_{i,n}^{j} + c_2 R_{i,n}^{j}} \tag{1}$$

$$\overline{C_n^{j}} = \frac{1}{M} \sum_{i=1}^{M} P_{i,n}^{j} \tag{2}$$

This innovative strategy lies in the introduction of the QPSO algorithm, such as Formulas (1) and (2), during the feature weighting process. It take advantage of all the particles, the optimal location of the average parameters by local optimal value p and p control, the global optimal value G C mean value of the optimal location for all particles. The exploration range is expanded through the quantum superposition state, which enables each particle to simultaneously explore multiple potential high-quality solution regions, significantly increasing the probability of finding the global optimal combination of feature weights.

By using the multimodal feature fusion strategy, dynamic weight learning is carried out on the dataset through QPSO, and the original features are weighted with the obtained optimal weight coefficients to generate weighted feature vectors.

2.4. Adversarial distillation framework

This study proposes an adversarial distillation framework. The innovation of this framework lies in the organic combination of adversarial training and knowledge distillation. By constructing a teacher model with strong robustness through adversarial training, it can resist the feature variations of malicious software. Then, through the high-temperature distillation process, this robustness is transferred to a lightweight student model, significantly improving the model's efficiency. Maintain high detection performance.

To enhance the model's robustness against minimal differences in malware features, the teacher model uses the Fast Gradient Notation (FGSM) method to generate adversarial samples [10]. The Fast gradient notation (FGSM) is shown in Formula (3), where x is the input (clean) image, x^{adv} is the perturbed adversarial image, J is the classification loss function, y_{true} is true label for the input x.Calculate the gradient of the original data, then apply FGSM to generate adversarial samples, and finally update the model by combining the losses of normal samples and adversarial samples.

$$x^{adv} = x + \varepsilon \bullet sin(\nabla_x J(x.y_{true})) \tag{3}$$

During the adversarial training process, it is achieved by optimizing the loss functions of the original samples and the adversarial samples, as shown in Formula (4).

$$L_{adv} = L_{clean} + \alpha \bullet L_{adversarial} = CE(teacher(x), y) + 0.3 \bullet CE(teacher(x+\delta), y) \tag{4}$$

Transfer through high-temperature softmax. Concealed by standard softmax of rich information about the relative relations between categories from confrontational training model to the students of

DOI: 10.54254/2755-2721/2026.TJ30949

teachers.To balance the migration of knowledge and students' autonomous learning model, set the temperature (T) is 4.0, the balance coefficient (α) is 7.0, z_s and z_t , respectively the logarithm of students and teachers. From this, the distillation loss function was designed, as shown in Formula (5).

$$L_{KD} = \alpha \bullet L_{KL} + (1-\alpha) \bullet L_{CE} = \alpha \bullet KL(\sigma(\frac{z_s}{T})||\sigma(\frac{z_t}{T})) \bullet T^2 + (1-\alpha) \bullet CE(z_s, y) \quad (5)$$

2.5. Comparative analysis

By adopting a multi-dimensional performance comparison framework, in order to verify the advantages of multimodal feature weighting over traditional methods, prove that deep learning models outperform traditional kernel methods in complex pattern recognition, and demonstrate the performance improvement of nonlinear modeling over linear models, the models selected were Random Forest, SVM and Logistic Regression respectively.

3. Experiment

3.1. Evaluation index system

As shown in Figure 1, to comprehensively evaluate the model performance, this study constructed a comprehensive malware detection and evaluation system.Based on Accuracy, Precision, Recall, F1 scores and Matthews correlation coefficient (MCC), the system is complemented by balanced accuracy and average accuracy to ensure that the evaluation results are statistically more robust and balanced.At the same time, it also faces the problem that a single evaluation index often fails to comprehensively reflect the overall performance of the model. To address this issue, this experiment innovatively designed a QPSO fitness function that combines five evaluation indicators into a single optimization objective, as shown in Formula (6), where M_i is the i-th evaluation metric. W is a carefully designed weight vector that reflects the relative importance of each metric in malware detection.

$$F(w) = \sum_{i=1}^{5} \omega_i . M_I(w) \quad (6)$$

3.2. Experimental results

Table 1. Comprehensive performance comparison of malware detection models

Model	Accuracy	Precision	Recall	F1-Score	MCC	Balanced Accuracy
Random Forest	0.9867	0.9899	0.9739	0.9819	0.9715	0.9841
SVM	0.9817	0.9871	0.9631	0.9750	0.9607	0.9779
Logistic Regression	0.9784	0.9738	0.9676	0.9707	0.9536	0.9762
ADMFW-MalDet	0.9910	0.9910	0.9847	0.9878	0.9807	0.9897

DOI: 10.54254/2755-2721/2026.TJ30949

Table 2. Performance improvement analysis: ADMFW-MalDet vs baseline methods

Comparison	Accuracy	Precision	Recall	F1-Score	MCC	Balanced Accuracy
vs Random Forest	+0.44%	+0.10%	+1.11%	+0.61%	+0.95%	+0.58%
vs SVM	+0.95%	+0.39%	+2.24%	+1.32%	+2.08%	+1.21%
Vs Logistic Regression	+1.29%	+1.77%	+1.77%	+1.77%	+2.85%	+1.39%
Average lmprovement	+0.89%	+0.75%	+1.70%	+1.23%	+1.96%	+1.06%

The ADMFW-MalDet model, by integrating adversarial distillation and dynamic feature weighting, significantly outperforms traditional methods in Android malware detection. In Table 1 and Table 2, the accuracy of the ADMFW-MalDet model reaches 99.10%, which is 0.43 percentage points higher than that of the optimal baseline method (Random Forest, 98.67%). The performance of the ADMFW-MalDet model is particularly outstanding in terms of the Recall metric, achieving a rate of 98.47%, an improvement of 1.08 percentage points over Random Forest (97.39%). This fully validates the model's significant advantage in terms of malicious sample detection coverage. The Matthews Correlation Coefficient (MCC) reached 98.07%, demonstrating superior generalization ability and robustness overall. It effectively balances detection accuracy and coverage, and the charts indicate that the ADMFW-MalDet model exhibits high discriminative power in malicious detection with imbalanced categories. This performance improvement is mainly attributed to the multi-modal feature dynamic weighting mechanism implemented by the quantum particle swarm optimization algorithm. This mechanism effectively identifies and enhances the role of key features in decision-making through the quantum position update formula.

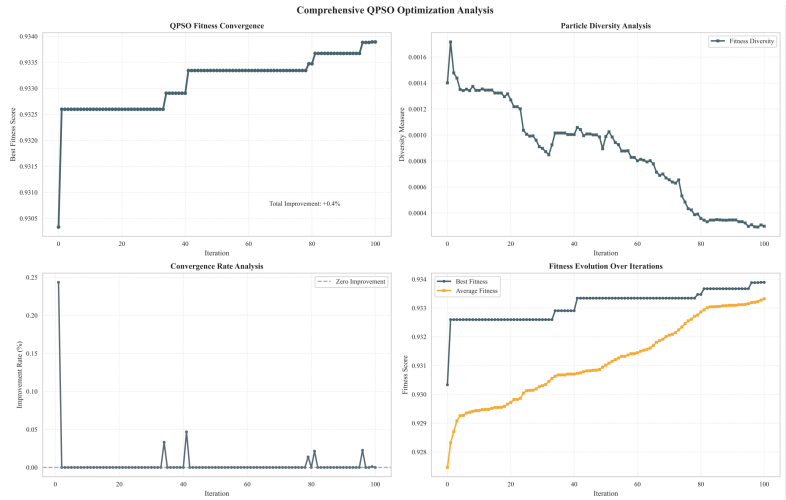

Figure 2. Comprehensive QPSO optimization analysis (picture credit: original)

In the comprehensive QPSO optimization analysis shown in Figure 2, the convergence characteristics of the quantum particle swarm optimization algorithm during the feature weight learning process are demonstrated. The convergence curve rises rapidly in the initial stage, improves steadily in the middle stage, and converges gradually in the later stage. Specifically, in the first 34 iterations, the fitness score rapidly increased from the initial 0.9303 to 0.9329, with a growth rate of 0.26%. During the 34-81 iterations, the algorithm entered the fine tuning stage, and the fitness score steadily increased to 0.9335. Finally, the convergence platform was reached in the 81-100 iteration stages, and the optimal fitness was stable at around 0.9339. The upward trend of the fitness curve

DOI: 10.54254/2755-2721/2026.TJ30949

proves the continuous effectiveness of the feature weighting strategy, while the formation of the final convergence platform indicates that the algorithm has found a feature weight combination close to the optimal one. The particle diversity curve shows the opposite trend, gradually decreasing from the initial 0.0014 to 0.0003 at convergence, reflecting the natural evolution of the particle swarm from extensive exploration to concentrated development during the optimization process. Both the excellent optimization efficiency and stability have been demonstrated during the convergence process of the quantum particle swarm optimization algorithm. The enhancement of the global exploration ability of the algorithm and the avoidance of the problem of premature convergence of the algorithm are both achieved under the introduction of quantum behavior mechanisms, which are all compared with the traditional particle swarm optimization algorithmhis optimization process has laid a solid foundation for the outstanding performance of ADMFW-MalDet, verifying the practical value and innovative significance of the multimodal feature dynamic weighting mechanism in the Android malware detection task.

4. Conclusions

This study proposes an innovative Android malware detection model, ADMFW-MalDet. This model combines the fusion of QPSO feature weighting and adversarial distillation to verify its detection effectiveness on the Drebin dataset.In various studies, dynamic feature addition, salient features and suppression of irrelevant features are implemented using quantum particle swarm optimization algorithm.Moreover, the adversarial distillation framework has been employed to transfer the teacher model to the student network, fostering a powerful detection capability and achieving better results compared to traditional static features. Despite these advancements, there are still some limitations worth noting in future research, including the need for further experiments on emerging malware in this model and further improvements to the QPSO algorithm. Additionally, the use of datasets needs to be expanded to further refine the QPSO algorithm and enhance its exploration capabilities.

Future work will move towards a new paradigm of intelligent security defense, aiming to establish a new generation of intelligent security defense paradigms that can cover more complex attack scenarios.

References

[1] Umasankar. (2017, September). Analysis of latest vulnerabilities in Android. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1236–1241). IEEE.
[2] McAfee. (2023, February 27). McAfee 2023 consumer mobile threat report. McAfee Blog. Retrieved from https: //www.mcafee.com/blogs/internet-security/mcafee-2023-consumer-mobile-threat-report/
[3] Kaspersky. (2024). Consumer and privacy predictions for 2025 [Report]. Retrieved from https: //securelist.com/ksb-consumer-and-privacy-predictions-2025/114620/
[4] Mishra, J., Sahay, S. K., Rathore, H., & Kumar, L. (2021). Duplicates in the Drebin dataset and reduction in the accuracy of the malware detection models. In 2021 26th IEEE Asia-Pacific Conference on Communications (APCC) (pp. 161–165). IEEE.
[5] Lu, Y.-F., Kuo, C.-F., Chen, H.-Y., Chen, C.-W., & Chou, S.-C. (2018). A SVM-based malware detection mechanism for Android devices. In 2018 International Conference on System Science and Engineering (ICSSE) (pp. 1–6). IEEE.
[6] Agarwal, A., Vatsa, M., Singh, R., & Ratha, N. (2021). Intelligent and adaptive mixup technique for adversarial robustness. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 824–828). IEEE.
[7] N., N., Reddy, S. V., & Dakshinesh, S. (2025). Improving accuracy in text extraction from images using region-based convolutional neural networks algorithm compared to convolutional neural network algorithm. In 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE) (pp. 706–710). IEEE.

DOI: 10.54254/2755-2721/2026.TJ30949

[8] Sun, J., Xu, W., & Feng, B. (2004). A global search strategy of quantum-behaved particle swarm optimization. In 2004 IEEE Conference on Cybernetics and Intelligent Systems (Vol. 1, pp. 111–116). IEEE.
[9] Sun, J., Fang, W., Wu, X., Palade, V., & Xu, W. (2012). Quantum-behaved particle swarm optimization: Analysis of individual particle behavior and parameter selection. Evolutionary Computation, 20(3), 349–393.
[10] Lad, A., Bhale, R., & Belgamwar, S. (2024). Fast gradient sign method (FGSM) variants in white box settings: A comparative study. In 2024 International Conference on Inventive Computation Technologies (ICICT) (pp. 382–386). IEEE.