Weighted Ensemble Method on Multiple N-Gram Opcode Probability Output for Mobile Malware Detection

Noor Azleen Anuar, Mohd Zaki Mas'ud*, Nor Azman Mat Ariff, Nazrulazhar Bahaman and Erman Hamid

Information Security, Digital Forensic and Computer Networking (INSFORNET),
Faculty of Information Technology and Communication,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

*Corresponding e-mail: zaki.masud@utem.edu.my

Keywords: Mobile Malware Detection, Opcode, weighted ensemble

ABSTRACT – Recent trend of working from home due to pandemic has caused internet users to use mobile devices to fulfil their everyday activities. With the current Mobile device's capability, users are easier to engage in a variety of online activities such as internet browsing, online banking transactions, social networking, and others. However, this scenario has expose users to mobile devices threat such as mobile malware. Various detection approaches for mobile malware have been established, however the virus creator has their own tactics for evading detection. As a result, an improved mobile malware detection approach is required to protect smartphone users from malicious threats. Based on this reason, this study enhances the Mobile Malware Detection method by introducing multiple N-Gram opcode features with weighted ensemble classification method which can improved the detection performance in terms of detection accuracy and minimizing false alerts.

1. INTRODUCTION

During the pandemic of covid-19, a Threat Evolution report issues by Kaspersky [1] has shown that in In Quarter 1 of 2020, the number of malware attacks were 14,446,496, followed by 14,203,865, in Quarter 2, 16,440,099 in Quarter 3 and the highest number of malware attacks were in the Quarter 4 of 2020 with 18,085,657 of attacks. These data shown even in just one year the number of malware attacks are very high. As more users are now connected to the internet through mobile devices, malware author is now turning their target towards mobile users.

The ability of malware evade detection are also more advance, a report by McAfee Labs [2], several evasion techniques were used by mobile malware in their journey to bypass Android protection mechanisms. The most common evasion techniques are anti-sandbox, anti-security tools, code injection, anti-monitoring, and anti-debugging. Hence, resulting in a more complex process to detect mobile malware static analysis and signature-based detection can address issues but it can be very difficult to detect zero-day or obfuscated code because it relies on a unique signature. Whereas Dynamic analysis and anomaly-based detection can curb the problem, however, it can result in a relatively high rate of false alerts [3].

Based on this reason, this study explores Mobile Malware Detection (MMD) method by incorporated a static analysis approach and anomaly-based detection method to compensate between the advantages and drawbacks of previous detection method. This paper proposed a weighted ensemble method for mobile malware detection based on the N-Gram opcode probability output to improve the detection accuracy and minimizing the false alert. Multiple N-gram opcode is extracted through static analysis from the binary of android applications. Opcodes represent the binary instruction on an application. Whereas weighted ensemble method is used for improving the probability output generated from the classification phase.

The remainder of the paper will be structured as follows. Section 2 explains the research methodology used in this study. Section 3 evaluate the experimental analysis and results. Finally, Section 4 summarizes the research.

2. METHODOLOGY

A total of 2,000 malicious and benign samples are tested in this experiment to evaluate the proposed MMD method. The malicious samples were collected from ArgusLab Android Malware Dataset [4], whereas the benign samples were retrieved from Google PlayStore.

Opcode Androguard was used during the Data Collection and Extraction phase to extract the opcode from a collection of malicious and benign *.apk* files. Opcode sequence features were chosen for this investigation because they can reveal malicious properties even when the source codes were encrypted or obfuscated. The opcode sequence is subjected to n-gram opcode generation, in which the opcode was classified into n-length. The N-gram was used to express the relationship between mobile malware behavior and its opcode sequence during execution. For this experiment, the lengths of 'n' extracted were n=1, n=2, and n=3. However, due to the complexity of machine learning algorithms and the large number of features [5] [6], no further opcode sequence lengths were extracted.

The filter method is then applied in the Feature Selection and Classification step. Bag of word (BOW) features use all available features. Whereas the Chi-Square (CS) and Information Gain (IG) selection method works by rating all features before selecting the ones that are relevant [7]. The N-Gram opcode features were ranked and filtered by 80% of the total amount of features for the CS and IG feature selection methods. 60% of the total instances were trained, whereas the other 40% served as test data. The selected N-Gram opcode features were then analyzed and evaluated using

a linear SVM classifier.

The probability output for each linear SVM classification phase was also referred to as a single classifier. Averaging will be applied to all of the single classifiers.

Furthermore, the Weight of each SVM probability output is then assigned with weight derived using Particle Swarm Optimization (PSO) Algorithm. PSO was employed in this stage to obtain the appropriate weight for the ensemble method [8]. The resulting weight works as a jury, with the best group of multiple N-Gram opcode probability output receiving a greater weight. Unlike feature selection, the least relevant set of classifiers was not removed because when numerous classifiers are integrated or ensembled together, even the one paired with the least relevant classifier may provide significant accuracy in identifying mobile malware applications.

Finally, the outcomes of the Weighted Ensemble method on multiple N-Gram opcode probability output were analysed, and the evaluation measured Accuracy, True Positive Rate (TPR), and False Positive Rate (FPR). Figure 1 depicts the overview of the proposed method.

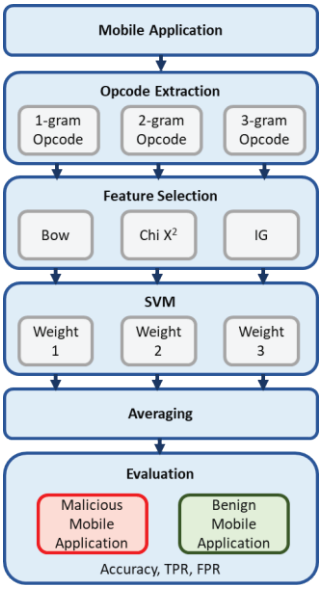

Figure 1 Mobile Malware Detection through multiple N-gram Opcode Sequence using Weighted Ensemble

3. RESULT AND DISCUSSION

This segment addresses the findings of an analysis on the weighted ensemble method for N-Gram opcode probability output. The analytical findings of the experiment are shown in Table 1, the table shows BOW achieved the lowest performance of all the classifier algorithms. Ultimately, the optimum accuracy is achieved by doing a weighted ensemble method on IG N-Gram opcode probability output using PSO, with an accuracy if 96.55%. Although CS and IG were having the same accuracy, IG is proven to be better as it has the highest TPR of 99.10% and the lowest FPR of only 0.90%. The highest Accuracy, TPR, and lowest FPR are highlighted in Table 1.

Table 1 Performance Metrics Results

Classifier	TPR	FPR	Accuracy
BOW	98.70	1.30	96.50
CS	98.80	1.20	96.55
IG	99.10	0.90	96.55

Therefore, based on the result in Table 1 Mobile Malware Detection method through weighted ensemble method for multiple N-Gram opcode probability output using IG feature selection is able to classify between malicious and benign mobile application with the best Detection Accuracy, TPR, and FPR.

4. CONCLUSIONS

This research paper proposed a mobile malware detection system using multiple N-Gram opcode features with weighted ensemble classification method. The experimental result shows that weighted ensemble method for mobile malware detection based on the N-Gram opcode probability output using IG feature selection and SVM are able to classify between malicious and benign mobile application with an accuracy of 96.55%, TPR of 99.10 and of 0.9%. In the future, this research will explore different classification algorithm and ensemble combination method.

ACKNOWLEDGEMENT

The authors are grateful to INSFORNET Research Group of Universiti Teknikal Malaysia Melaka (UTeM) for the support and special acknowledgement to Ministry of Education Malaysia for providing financial support through the Fundamental Research Grant Scheme (FRGS/2018/FTMK-CACT/F00391).

REFERENCES

[1] Tenenboim-Chekina, L., O. Barad, A. Shabtai, D. Mimran, L. Rokach, B. Shapira, and Y. Elovici. Detecting application update attack on mobile devices through network featur. *In 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 91-92. IEEE, 2013.

[2] Symantec Corporation, *2018* Internet Security Trend Report, [online] Available at: https://docs.broadcom.com/doc/istr-23-2018-executive-summary-en-aa [Accessed 7 Sep. 2019].

[3] V. P., A. Zemmari, and M. Conti, A machine learning based approach to detect malicious android apps using discriminant system calls ,*Futur. Gener. Comput. Syst.*, vol. 94, pp. 333–350, May 2019.

[5] I. Santos, F. Brezo, X. Ugarte-Pedrero, and P. G. Bringas,.Opcode sequences as representation of executables for data-mining-based unknown malware detection,. *Inf. Sci. (Ny).*, vol. 231, pp. 64–82, May 2013, doi: 10.1016/j.ins.2011.08.020.

[6] J. Zhang, Z. Qin, H. Yin, L. Ou, and K. Zhang, .A feature-hybrid malware variants detection using CNN based opcode embedding and BPNN based API embedding,. *Computer Security*, vol. 84, pp. 376–392, 2019, doi: 10.1016/j.cose.2019.04.005.