Ministry of Higher Education and Scientific Research
University of Mosul
College of Computer Science and Mathematics
Department of Software

Design and Implementation of a tool to detect and analyze Android malware applications

A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Software

By
Omar Emad Saied Ahmed

Supervised by

Assist.Prof.Dr. Karam Hatim Thanoon

2025 A.D. 1447 A.H.

Abstract

The fast expansion of the internet has led to the fact that Android smartphones have become widely used around the world among various users based on their work, age, and interests, because digital services are easier to access than ever before, which increases the probability to be attacked.

The main problem is that attackers have become increasingly profitable, motivating them to develop more malware to hack internet users. Antivirus and antimalware software often fail to detect new versions of malware due to the limitation in their detection methods. Furthermore, the lack of knowledge among many users makes them an easy victim for attackers.

The most important objective of this thesis is to develop a powerful tool that enables users to analyze and detect APK and XAPK files for possible threats, by examining the app permissions and monitoring app behavior during the running in a controlled emulator environment.

Two new datasets were created by using static and dynamic analysis for 400 APK samples. After feature selection, three machine learning algorithms were used: Random Forest, Support Vector Machine, and Decision Tree. Three models were trained and tested on a balanced dataset of 400 Android applications, which were collected from the internet. In classification, RF showed the best performance with 95.83% accuracy in static analysis and 92.50% in dynamic analysis. The tool was tested on a public dataset from Kaggle gives better results with 98.58% accuracy in Kaggle dataset, which ensures the proposed tool's performance.

The proposed tool was designed and implemented in Python with a simple graphical user interface (GUI) that is easy for users to operate. It has the ability to predict and detect new Android malware applications which is not included in the dataset samples and have not been trained previously. One of the future works is to upload the tool to the cloud for access scalability.

Design and Implementation of a tool to detect and analyze Android malware applications

Author: Omar Emad Saied Ahmed **Advisor: Assist.Prof.Dr. Karam Hatim Thanoon**

Publisher: University of Mosul

HIGHLITS	GRAPHICAL ABSTRACT
• Developed an intelligent tool for automatic detection and analysis of Android malware using static and dynamic analysis techniques. • Created two custom datasets (static & dynamic) from 400 APK samples including permissions and runtime behavior. • Implemented three ML classifiers (Random Forest, SVM, Decision Tree) to classify apps as benign or malicious. • Achieved 95.83% accuracy in static analysis, 92.50% accuracy in dynamic analysis, and 98.58% accuracy on Kaggle dataset. • Designed a user-friendly GUI enabling users to upload, analyze, classify, and visualize results easily. • Tool can detect previously unseen malware apps beyond the training dataset. • Proposed future deployment of the tool on the cloud for scalability and continuous updates.	

Keywords:

- Android malware
- APK / XAPK analysis
- Static analysis
- Dynamic analysis
- Machine learning
- Random Forest
- Support Vector Machine
- Decision Tree
- Feature selection
- Permissions analysis
- Malware detection tool
- Kaggle dataset

ABSTRACT

The fast expansion of the internet has led to the fact that Android smartphones have become widely used around the world among various users based on their work, age, and interests, because digital services are easier to access than ever before, which increases the probability to be attacked.

The main problem is that attackers have become increasingly profitable, motivating them to develop more malware to hack internet users. Antivirus and antimalware software often fail to detect new versions of malware due to the limitation in their detection methods. Furthermore, the lack of knowledge among many users makes them an easy victim for attackers.

The most important objective of this thesis is to develop a powerful tool that enables users to analyze and detect APK and XAPK files for possible threats, by examining the app permissions and monitoring app behavior during the running in a controlled emulator environment.

Two new datasets were created by using static and dynamic analysis for 400 APK samples. After feature selection, three machine learning algorithms were used: Random Forest, Support Vector Machine, and Decision Tree. Three models were trained and tested on a balanced dataset of 400 Android applications, which were collected from the internet. In classification, RF showed the best performance with 95.83% accuracy in static analysis and 92.50% in dynamic analysis. The tool was tested on a public dataset from Kaggle gives better results with 98.58% accuracy in Kaggle dataset, which ensures the proposed tool's performance.

The proposed tool was designed and implemented in Python with a simple graphical user interface (GUI) that is easy for users to operate. It has the ability to predict and detect new Android malware applications which is not included in the dataset samples and have not been trained previously. One of the future works is to upload the tool to the cloud for access scalability.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم البرمجيات

تصميم وتنفيذ أداة لكشف وتحليل تطبيقات البرامج الخبيثة لنظام الاندرويد

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
البرمجيات

من قبل
عمر عماد سعيد احمد

بإشراف
ا.م.د. كرم حاتم ذنون

١٤٤٧ هـ ٢٠٢٥ م

المستخلص

بالنظر إلى الانتشار السريع للإنترنت، أصبح استخدام الهواتف الذكية التي تعمل بنظام أندرويد أكثر انتشارًا حول العالم بين مستخدمين مختلفين بناءً على عملهم وأعمارهم واهتماماتهم، وذلك لأن الخدمات الرقمية أصبحت أكثر سهولة في الوصول، مما يزيد من احتمالية التعرض للهجمات. المشكلة الرئيسية هي أن المهاجمين أصبحوا أكثر ربحية، مما يدفعهم إلى تطوير المزيد من البرمجيات الخبيثة لاختراق المستخدمين عبر الإنترنت. غالبًا ما تفشل برامج مكافحة الفيروسات والبرمجيات المضادة للبرمجيات الخبيثة في الكشف عن نسخ جديدة من البرمجيات الخبيثة بسبب طرق الكشف التي تعتمدها. بالإضافة إلى ذلك، فإن نقص الوعي بين أغلب المستخدمين يجعل منهم ضحايا سهلين للمهاجمين.

الهدف الأهم من هذا البحث هو تطوير أداة قوية تمكّن المستخدمين من تحليل وكشف ملفات XAPK/APK المحتملة التهديدات، من خلال فحص صلاحيات التطبيق ومراقبة سلوكه أثناء تشغيله في بيئة محاكاة افتراضية مضبوطّة.

تم إنشاء مجموعتين جديدتين من البيانات باستخدام التحليل الثابت والديناميكي، شملتا ٤٠٠ عينة من ملفات APK، وتم استخراج الميزات منها. بعد عملية اختيار الميزات، تم استخدام ثلاثة خوارزميات تعلم آلي: الغابة العشوائية RF، آلة الدعم الناقل SVM، وشجرة القرار DT. تم تدريب واختبار ثلاثة نماذج على مجموعة بيانات متوازنة تتكون من ٤٠٠ تطبيق أندرويد تم جمعها من الإنترنت. في عملية التصنيف، أظهرت خوارزمية الغابة العشوائية أفضل أداء بنسبة دقة بلغت ٩٥,٨٣٪ في التحليل الثابت و٩٢,٥٠٪ في التحليل الديناميكي. وتم اختبار الأداة على مجموعة بيانات عامة من منصة Kaggle ، حيث حققت نتائج أفضل بنسبة دقة تصل إلى ٩٨,٥٨٪، مما يؤكد أداء الأداة المقترحة.

تم تصميم وتنفيذ الأداة باستخدام لغة بايثون مع واجهة رسومية بسيطة وسهلة الاستخدام، وتملك القدرة على التنبؤ والكشف عن تطبيقات أندرويد الخبيثة الجديدة التي لم تُدرّب عليها خوارزميات التعلم الآلي مسبقًا، حيث يمكن أن تكون هذه التطبيقات غير متضمنة في مجموعة البيانات الأصلية. وتُقترح في المستقبل رفع الأداة إلى السحابة لزيادة القدرة على الوصول وقابلية التوسع.