Uuniversity of Mosul
College of Computer Sciences and Mathematics
Departement of Computer Sciences

Detection and Classification of Android Malware using Machine Learning

A Thesis Submitted
By

Omar Shamil Ahmed

M.Sc. / Thesis

Computer Science

Supervised by

Dr. Omar Abdulmunem Ibrahim Al-Dabbagh

2021 A. D. **1443 A. H.**

Abstract

The connected digital information network is an integral part of our daily activities. Every day we add, store, process and share a large amount of electronic data. This information is always very important, sensitive and secret, and therefore disclosure of this information represents a great risk. Cybersecurity is the comprehensive security system to protect the digital information network (in communication networks, operating systems and databases) from all kinds of cyber-attacks. To support and improve the cybersecurity system, Machine Learning techniques are used, which can learn on its own to detect and prevent cyber threats and attacks.

In this thesis, ML algorithms and DL models are used to support cybersecurity to detect and classify Android malware by monitoring network traffic. The proposed techniques were designed in six phases. In the first phase, data acquisition and feature extraction are done from real-world dataset CICAndMal2017. In the second phase, these data were analyzed and data preprocessing techniques were applied, which includes handling samples with null values, deleting features with low variance values, and data normalization.

In the third phase, two feature selection techniques (Univariant selection and Select from model) were used to select the best number of features within the total feature set. In the fourth phase, the data was shuffled and split into training, validation, and testing samples through two different experiments. Then, six ML algorithms were used for malware detection, which were trained on samples extracted from previous processes and on different numbers of features.

In the fifth phase, after the malware detection process, two DL models (MLP and LSTM) were designed and used to classify these malwares into four classes. In the last phase, the proposed detection and classification techniques were evaluated using several performance metrics. Malware detection techniques have achieved a detection accuracy about 99%, and malware

classification techniques have achieved a classification accuracy about 96% in detecting and classifying four types of malwares.

جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

كشف وتصنيف برمجيات اندرويد الضارة باستخدام التعلم الالي

عمر شامل احمد

رسالة ماجستير
علوم الحاسوب

بإشـراف

د. عمر عبدالمنعم إبراهيم الدباغ

1443 هـ 2021 م

الخلاصة

تعتبر شبكة المعلومات الرقمية المتصلة جزءًا لا يتجزأ من نشاطاتنا اليومية. حيث نقوم يوميا بإضافة كمية كبيرة من البيانات الالكترونية وتخزينها ومعالجتها ومشاركتها. دائما ما تكون هذه المعلومات مهمة جداً وحساسة وسرية وبالتالي يمثل كشف هذه المعلومات خطراً كبيراً. يعد الامن السيبراني النظام الأمني الشامل لحماية شبكة المعلومات الرقمية (في شبكات الاتصالات وأنظمة التشغيل وقواعد البيانات) من كل أنواع الهجمات السيبرانية. لدعم وتحسين نظام الامن السيبراني، يتم استخدام تقنيات التعلم الالي، والتي يمكن ان تتعلم من تلقاء نفسها لاكتشاف التهديدات والهجمات السيبرانية ومنعها.

في هذه الرسالة، تم استخدام خوارزميات التعلم الالي ونماذج التعلم العميق لدعم الامن السيبراني في كشف وتصنيف برامج اندرويد الضارة من خلال مراقبة حركة مرور الشبكة. تم تصميم التقنيات المقترحة في ستة اطوار. في الطور الأول يتم الحصول على البيانات واستخراج الميزات باستخدام مجموعة بيانات واقعية وهي CICAndMal2017. في الطور الثاني، تم تحليل هذه البيانات وتطبيق تقنيات معالجة مسبقة للبيانات والتي تتضمن معالجة العناصر التي تحتوي على قيم خالية، وحذف الميزات المتضمنة قيم ذات تباين قليل، ومن ثم تطبيع البيانات.

في الطور الثالث، تم استخدام تقنيتي اختيار الميزات (Univariant selection و Select from model) لانتقاء أفضل عدد من الميزات ضمن مجموعة الميزات المتكاملة. في الطور الرابع، تم خلط البيانات وتقسيمها الى عينات للتدريب والتحقق والاختبار من خلال تجربتين مختلفتين، بعد ذلك تم استخدام ست خوارزميات تعلم الي لاكتشاف البرامج الضارة، حيث تم تدريبها على العينات المستخرجة من العمليات السابقة وعلى اعداد مختلفة من الميزات.

في الطور الخامس وبعد عملية الكشف عن البرامج الضارة، تم تصميم واستخدام نموذجين تعلم عميق لتصنيف هذه البرامج الضارة الى أربعة أصناف باستخدام نموذجي التعلم العميق MLP و LSYM.