UNIVERSITÀ DEGLI STUDI DI PALERMO

Corso di Laurea Magistrale Ingegneria Informatica

Dipartimento di Ingegneria

SISTEMI IBRIDI PER IL RILEVAMENTO DI MALWARE NEI DISPOSITIVI MOBILI

TESI DI LAUREA DI

Irene Di Prima

RELATORE

Prof. Alessandra De Paola

ANNO ACCADEMICO 2022 - 2023

Università Degli Studi Di Palermo

Dipartimento di Ingegneria

Corso di Laurea Magistrale Ingegneria Informatica

Sistemi ibridi per il rilevamento di malware nei dispositivi mobili.

Tesi di Laurea di

Irene Di Prima

Relatore

Prof. Alessandra De Paola

Sommario

Al giorno d'oggi, i malware rappresentano una tra le principali minacce per la sicurezza dei dispositivi mobili, in particolar modo per i sistemi Android i quali, oltre ad essere tra i dispositivi mobili più diffusi, risultano essere anche quelli maggiormente oggetto di attacchi informatici. La crescente complessità dei malware rende particolarmente complesso il loro riconoscimento, che richiede ormai l'utilizzo di tecniche avanzate e sofisticate, spesso basate sul *Machine Learning*. Questi metodi si basano sull'estrazione di caratteristiche significative dai malware a partire da una loro analisi statica o dinamica. Tuttavia, tra i numerosi approcci proposti in letteratura, risultano particolarmente promettenti gli approcci basati su analisi ibrida, in grado cioè di coniugare i vantaggi dei due diversi tipi di approccio. Questo lavoro di tesi offre una panoramica sui sistemi di rilevamento basati su analisi ibrida proposti negli ultimi anni, classificandoli in base al modo in cui sono state utilizzate le caratteristiche statiche e dinamiche, e rispetto alle tecniche di *Machine Learning* utilizzate. I risultati sperimentali proposti dai lavori in letteratura mostrano che l'utilizzo di caratteristiche ibride per il rilevamento dei malware consente di ottenere ottimi risultati e ad oggi resta uno degli approcci più promettenti per il rilevamento di software malevoli. Risultano tuttavia aperte numerose sfide, come discusso in questo lavoro di tesi.

INDICE

1 INTRODUZIONE

Nel corso degli anni il numero di dispositivi mobili divenuti parte integrante della vita quotidiana è aumentato esponenzialmente, tanto da sollevare delle legittime preoccupazioni per quanto riguarda l'esposizione a software malevoli e il rischio di attacchi informatici. Secondo quanto riportato sul Blog di Google Italy [1], già nella prima metà del 2022 erano circa tre miliardi i dispositivi Android attivi mensilmente in tutto il mondo, un terzo dei quali attivato solo durante l'ultimo anno. Poiché dispositivi come smartphone, tablet e smartwatch al giorno d'oggi vengono utilizzati per gestire dati sensibili di vari tipi (personali, bancari, finanziari, ecc…), è inevitabile che i malintenzionati vedano in essi degli obiettivi relativamente facili e convenienti da colpire per commettere illeciti e azioni fraudolente; da qui, la necessità di mettere a disposizione degli utenti dei meccanismi di difesa efficaci che permettano di contrastare azioni malevole, limitare eventuali danni e proteggere i dispositivi. In questo contesto s'inseriscono lo studio e la messa a punto di tecniche avanzate per il riconoscimento di malware, spesso in grado di distinguere non solo tra software malevoli e legittimi ma anche capaci di individuare la famiglia di appartenenza del malware, così da permettere l'adozione di contromisure adeguate per contrastare l'infezione ed evitarne la diffusione su altri dispositivi. Negli ultimi decenni sono state ideate numerose tecniche e sono stati testati diversi sistemi in grado di effettuare il riconoscimento di malware restituendo dei buoni risultati e delle buone performance ma, parallelamente, anche i malware si sono evoluti, diventando sempre più complessi e sofisticati, ragion per cui la rilevazione dei malware resta una sfida aperta tutt'oggi. Al giorno d'oggi, una prima distinzione dei sistemi di rilevamento malware può essere fatta tra sistemi mobili e sistemi basati su cloud. Nel primo caso, si tratta di installare sul dispositivo un programma dedicato che si occupa di analizzare le applicazioni (prima che vengano installate o dopo) per distinguere software malevoli da quelli legittimi. Questo approccio è vantaggioso perché non prevede ritardi temporali nel rilevamento, l'utente riceve immediatamente informazioni riguardo al potenziale pericolo e inoltre, in alcuni casi, i sistemi sono anche in grado di individuare delle contromisure da adottare per arginare o impedire l'infezione. L'aspetto negativo di questo tipo di sistemi riguarda l'utilizzo delle risorse del dispositivo (batteria, memoria), la scarsa

scalabilità e la necessità di costanti aggiornamenti per permettere al sistema di riconoscere correttamente anche le minacce più recenti e salvaguardare la sicurezza del dispositivo. L'utilizzo di sistemi basati su cloud risulta essere vantaggioso perché in grado di superare alcune delle limitazioni imposte dall'utilizzo di dispositivi basati su mobile: le risorse del dispositivo non verrebbero sfruttate dal sistema di rilevamento malware e qualsiasi tipo di dispositivo potrebbe accedere più facilmente al sistema, che sarebbe anche molto più semplice da aggiornare. Al tempo stesso, anche questo approccio presenta degli svantaggi: nel momento in cui l'utente carica dei file su cloud potrebbero sorgere dei problemi di sicurezza causati dalla condivisione di informazioni potenzialmente sensibili; inoltre, è difficile realizzare un sistema basato su cloud che sia in grado di effettuare il rilevamento in tempo reale poiché si deve tenere conto dei tempi necessari per lo scambio di informazioni e, infine è necessaria non solo una buona connessione ma anche un'ottimizzazione delle comunicazioni.

Un altro criterio da considerare per suddividere in categorie diverse i sistemi di rilevamento malware è il tipo di approccio utilizzato per effettuare la classificazione. Le prime architetture, ben presto rivelatesi insufficienti, si servivano di *signature* ed euristiche per distinguere le applicazioni malevole da quelle legittime ma restituivano prestazioni insufficienti in caso di malware offuscati o dei cosiddetti malware "*zero-day*", ovvero software malevoli non ancora noti. L'esigenza di riuscire a ideare sistemi che fossero in grado di rilevare sia i software malevoli già noti, sia quelli ancora sconosciuti, ha spinto i ricercatori ad utilizzare degli approcci basati sull'Intelligenza Artificiale e, nello specifico, a servirsi delle tecniche messe a disposizione dal *machine learning*; l'implementazione di sistemi in grado di apprendere automaticamente tutt'oggi risulta essere promettente e in grado di superare le limitazioni imposte dai metodi ideati in precedenza. Per utilizzare correttamente le tecniche di apprendimento automatico è necessario analizzare i programmi da classificare ed estrarre delle opportune caratteristiche ("*features*") che permettano al sistema di rilevamento di distinguere software malevoli e software legittimi. Durante questa fase di analisi si possono utilizzare approcci diversi: l'approccio statico prevede che il programma venga analizzato senza che esso venga eseguito per poter estrarre

delle caratteristiche statiche; l'approccio dinamico richiede l'esecuzione del programma in degli ambienti protetti per poter estrarre delle opportune caratteristiche dinamiche; infine, l'approccio ibrido esegue sia l'analisi statica che quella dinamica e si serve di entrambe le tipologie di caratteristiche. Nel primo caso, l'analisi del programma viene effettuata in modo semplice e veloce ma questo approccio si è rivelato poco efficace in presenza di software malevoli che utilizzano delle tecniche di offuscamento o che scaricano del codice malevolo dopo essere stati eseguiti, ragion per cui dei sistemi basati su di esso potrebbero essere facilmente raggirati. Nel secondo caso, tramite analisi dinamica è possibile catturare e analizzare più approfonditamente il comportamento del programma potenzialmente malevolo senza che le tecniche di offuscamento possano invalidare i risultati, tuttavia alcuni malware sono in grado di riconoscere e modificare il proprio comportamento in ambienti controllati, ragion per cui l'utilizzo di sandbox o macchine virtuali potrebbe compromettere il risultato dell'analisi; in ogni caso l'estrazione di caratteristiche dinamiche richiede molto più tempo e risorse, oltre ad essere molto più complessa rispetto all'estrazione di caratteristiche statiche, e questo rappresenta un effettivo limite in diversi casi (es: l'utilizzo di dispositivi con risorse limitate). I sistemi che utilizzano l'approccio ibrido, essendo in grado di combinare in diversi modi le caratteristiche statiche e dinamiche, hanno la possibilità di definire al meglio il comportamento e le peculiarità di ogni programma e dovrebbero quindi essere in grado di individuare i software malevoli restituendo delle prestazioni migliori; difatti, i sistemi ibridi si pongono l'obiettivo di sfruttare e combinare i vantaggi di entrambi i tipi di analisi e, al tempo stesso, ridurre al minimo i rispettivi svantaggi, così da raggiungere il massimo delle prestazioni in termini di efficienza ed efficacia. Per questo motivo, questo lavoro di tesi si pone l'obiettivo di descrivere e valutare le potenzialità dell'utilizzo di dispositivi che seguono un approccio ibrido per la rilevazione dei malware.

Il resto di questo elaborato è strutturato come segue: nel capitolo 2 viene fornita una panoramica delle principali caratteristiche dei sistemi Android, mentre nei capitoli 3 e 4 viene fornita una panoramica generale sui software malevoli, con un focus specifico sui malware nei dispositivi mobili, e viene introdotto il rilevamento dei malware. Nel capitolo 5 sono presentati i sistemi di rilevazione di

malware che sfruttano delle tecniche di machine learning ed estraggono le caratteristiche tramite analisi ibrida, dei quali vengono illustrati architettura e funzionamento, mentre nel capitolo 6 si attenzionano particolarmente i dataset utilizzati in fase sperimentale. A seguire, le conclusioni di questo studio, nel capitolo 7, dove verranno riportate osservazioni e valutazioni delle architetture presentate, nonché una panoramica delle sfide ancora aperte al giorno d'oggi.

2 DIFFUSIONE E CARATTERISTICHE DEI DISPOSITIVI ANDROID

Oltre ad avere circa tre miliardi di dispositivi Android attivi mensilmente in tutto il mondo già nella prima metà del 2022, di cui un miliardo attivati nell'ultimo anno [1], stando a quanto riportato da Business of Apps [2] ad inizio 2023 il numero di dispositivi mobili che utilizzavano Android avrebbe superato i 3 miliardi, rappresentando circa i tre quarti dei dispositivi mobili nel mondo (**Figura 1**). Poiché si tratta di un sistema operativo talmente diffuso da essere il più utilizzato a livello mondiale, vien da sé che sia anche uno dei target maggiormente presi in considerazione quando si parla di attacchi informatici tramite malware. Al di là dell'enorme diffusione di Android, un'ulteriore motivazione che ha spinto diversi attaccanti a concentrarsi su questo sistema operativo risiede nelle caratteristiche specifiche dello stesso: poiché si tratta una tecnologia open source ed è possibile installare applicazioni all'interno dei dispositivi sia passando per il Google Play Store, ovvero il servizio di distribuzione digitale ufficiale, sia tramite marketplace di terze parti, i dispositivi con sistema operativo Android possono essere presi di mira e danneggiati tramite malware con più facilità. Nel tempo sono state proposte da Google delle soluzioni per cercare di impedire la diffusione di malware sui dispositivi mobili e tra queste troviamo un meccanismo di rilevazione introdotto nel 2012 di nome Bouncer [3] che aveva il compito di scansionare le applicazioni all'interno dello store (sia quelle già presenti al suo interno, sia quelle

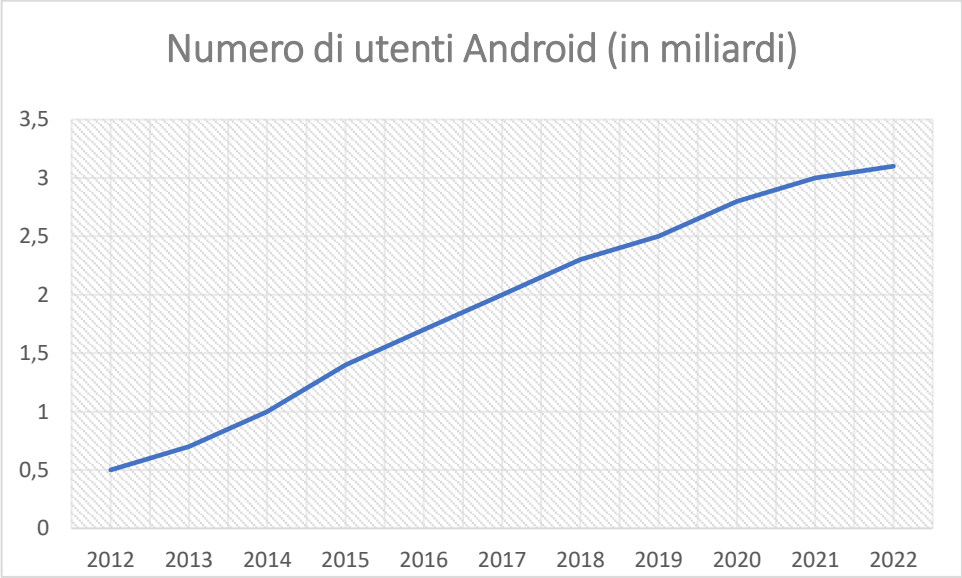

Figura 1: Numero di utenti Android nel corso degli anni. [2]

caricate da poco) e analizzarle per cercare malware noti; oltre a questo, Bouncer andava ad analizzare il comportamento delle applicazioni, mettendolo a paragone con quello adottato dalle altre, alla ricerca di comportamenti malevoli, proprio per impedire la diffusione di malware. Questo metodo, tuttavia, non si è rivelato sufficientemente efficace poiché poteva essere aggirato con facilità. Per incrementare ulteriormente il livello di sicurezza, quindi, è stato introdotto Google Play Protect [4] [5] nel 2017: si tratta di un servizio che va a scansionare le applicazioni all'interno dello store e dei dispositivi con l'obiettivo di riconoscere, rilevare e impedire ulteriore diffusione di malware. Secondo quanto riportato sul sito dedicato [5] a Google Play Protect, esso sarebbe in grado di scansionare 125 miliardi di applicazioni ogni giorno, fondamentale per quanto riguarda la protezione dei dispositivi e dei dati, e sfrutterebbe un meccanismo di sicurezza basato su cloud per analizzare e testare le applicazioni messe a disposizione nello store. Queste contromisure, comunque, non si sono rivelate sufficienti per contrastare la propagazione dei malware che tutt'oggi continuano ad infettare i dispositivi e, da qui, l'esigenza di creare dei sistemi in grado di effettuare efficacemente il rilevamento dei malware, sia che si tratti di malware noti o di *zero-day*, con una buona accuratezza e senza minare all'usabilità dei dispositivi.

3 I MALWARE

I malware ("*malicious software*", letteralmente traducibile in "software malevoli") sono dei programmi contenenti del codice considerato malevolo poiché capace di danneggiare i dispositivi e i dati contenuti al loro interno. A seguire, una breve panoramica generale sui malware, prima di concentrare l'attenzione sui malware che più di frequente infettano dispositivi mobili.

3.1 GENERALITÀ SUI MALWARE

Quando si parla di malware, si fa riferimento ad un'ampia gamma di software eterogenei, cosiddetti "malevoli", che hanno come obiettivo quello di danneggiare i dispositivi e impedirne il corretto funzionamento. I malware possono essere suddivisi in famiglie e classificati secondo diversi criteri; un primo modo per distinguerli gli uni dagli altri è valutarne il comportamento e gli effetti indesiderati all'interno del dispositivo infetto. Tra le principali tipologie di malware possiamo distinguere Virus, Worm, Trojan Horses, Backdoors, Adware e Ransomware.

- **Virus**: malware che agiscono col duplice obiettivo di infettare dei file all'interno del dispositivo per renderli inutilizzabili e riprodursi per danneggiare quanti più file e dispositivi possibile;
- **Worm**: diversamente dai virus, non necessitano di altri file per infettare un dispositivo e, generalmente, vengono trasmessi tramite posta elettronica. Tendono a replicarsi e a diffondersi molto velocemente;
- **Trojan Horses**: malware che apparentemente si comportano come dei software legittimi, celando il loro comportamento malevolo;
- **Backdoors**: letteralmente traducibile come "porte sul retro", spesso vengono utilizzati in coppia con i Trojan e permettono di accedere con facilità ad un dispositivo infetto, all'insaputa dell'utente legittimo;
- **Adware**: malware che mostrano degli annunci pubblicitari agli utenti e di questi ultimi talvolta tracciano anche il comportamento;

- **Ransomware**: malware che utilizzano la cifratura per rendere inaccessibili i dati e inutilizzabili i dispositivi e richiedono un riscatto (solitamente, in bitcoin) in cambio della password o della chiave di decifratura.

La distinzione tra una famiglia e l'altra non sempre è netta e ben precisa, poiché talvolta alcuni malware assumono contemporaneamente comportamenti che possono essere associati a più tipologie di software malevoli. I malware sono in continua evoluzione, ogni anno ne vengono scoperti di nuovi e con nuove caratteristiche, ragion per cui un altro criterio di classificazione adottato è quello temporale, che vede la categoria dei malware "tradizionali", ovvero quelli più antichi, nonché i primi a diffondersi, contrapposta a quella dei malware di "nuova generazione", sviluppati più di recente [6]. Le differenze tra queste due categorie evidenziano le contromisure adottate dai malintenzionati quando hanno iniziato a diffondersi i primi sistemi di rilevazione di malware, poiché quelli di "nuova generazione" risultano essere più complessi da individuare e rimuovere, rendendo ancora più difficile il compito di costruire un framework in grado di contrastare queste minacce. I malware considerati "tradizionali", infatti, sono malware più semplici sia da individuare che da eliminare, tendono ad attaccare una sola risorsa, a utilizzare un unico processo e a non camuffarsi per nascondersi all'interno del dispositivo; al contrario, i malware di "nuova generazione" sono dei malware più complessi e persistenti, che utilizzano più processi, attaccano più risorse, e, soprattutto, utilizzano delle tecniche di offuscamento tali da ostacolare e complicare notevolmente la rilevazione e la rimozione del malware. Anche le tecniche di offuscamento possono essere di vario tipo, le principali sono presentate nella **Tabella 1**.

Tecnica	Descrizione
Cifratura	Dei blocchi di codice malevolo vengono nascosti tramite cifratura.
Oligomorfia	Il payload viene cifrato e decifrato con una chiave differente.
Polimorfismo	Vengono usate chiavi di cifrature diverse e la cifratura può essere effettuata a strati.
Metamorfismo	Ad ogni interazione, il codice del malware viene modificato per fare in modo che generi una *signature* sempre diversa.
Stealth (“Invisibilità”)	Il codice viene protetto tramite tecniche a contatore che impediscono la corretta analisi del software.
Packing	Il malware viene compresso o nascosto tramite cifratura per impedirne l’analisi.

***Tabella 1**: Principali tecniche di offuscamento utilizzate dai malware di "nuova generazione" [6].*

L’esigenza di contrastare la crescente complessità dei malware e l’esponenziale aumento del numero di software malevoli che continuano a diffondersi da un dispositivo all’altro di anno in anno richiede che vengano impiegate tecniche di rilevamento sempre più sofisticate e capaci di ottenere risultati accettabili, affinché venga garantito un buon livello di sicurezza.

3.2 MALWARE NEI DISPOSITIVI MOBILI

Nonostante i sistemi operativi dei dispositivi mobili siano in continuo sviluppo e aggiornamento anche per ciò che riguarda la sicurezza, alcune minacce persistono e mettono costantemente a rischio i dati sensibili degli utenti. Come riportato da uno studio relativo alla sicurezza dei dispositivi mobili [7] tra le minacce principali per gli utenti dei dispositivi mobili possiamo trovare: il furto di dati, generalmente finanziari; i malware, solitamente diffusi tramite posta elettronica; phishing; furto del dispositivo; attacchi hacker; comportamento incurante degli utenti. Per quanto riguarda i malware, nello specifico, secondo questo studio quelli che principalmente sono in grado di attaccare con successo un dispositivo appartengono alle seguenti famiglie:

- Spyware: ingannano l'utente per riuscire ad ottenere le credenziali bancarie;
- Trojan e Backdoords: questi software malevoli vengono nascosti in vari tipi di applicazioni per compiere azioni fraudolente;
- Mobile Miners: generalmente distribuiti tramite posta elettronica o sms, sono dei malware che compromettono il dispositivo per utilizzarne le risorse.

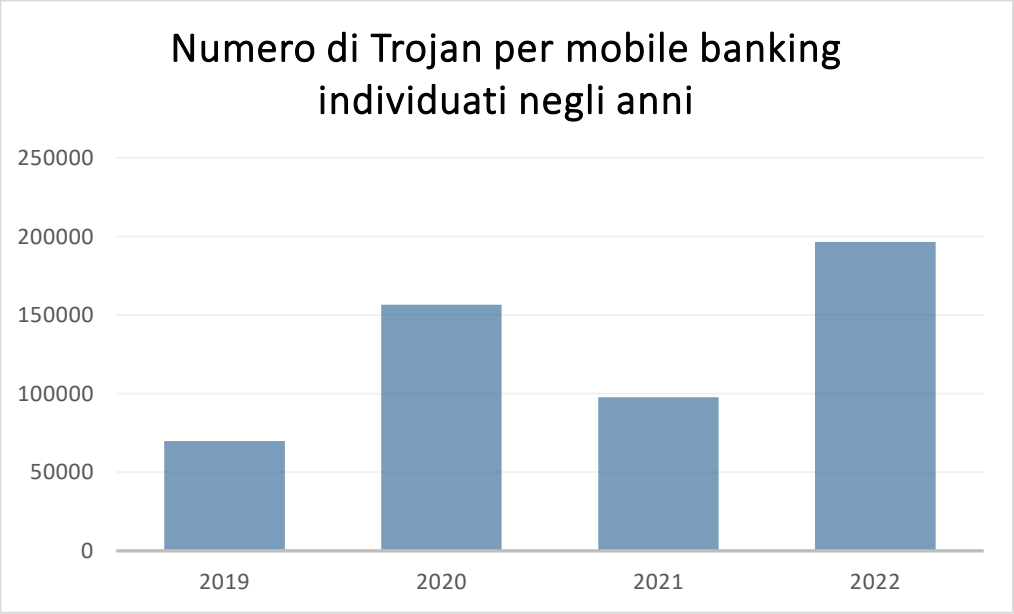

Figura 2: Crescita del numero di Trojan che attaccano le applicazioni di mobile banking. [8]

Secondo un report del 2022 realizzato da Kaspersky e pubblicato su Securelist [8], sui dispositivi mobili sono stati rilevati quasi 200.000 nuovi Trojan per le applicazioni di mobile banking e più di 10.000 nuovi Trojan Ransomware e questi dati, messi a paragone con quelli ottenuti negli anni precedenti, evidenziano un aumento del numero di nuovi malware di questo tipo attualmente in circolazione **Figura 2**.

Studi più recenti, condotti tra la fine del 2022 e l'inizio del 2023 utilizzando dati ottenuti sempre tramite Kaspersky [9] hanno evidenziato un'ulteriore crescita delle minacce sui dispositivi mobili, difatti ne sono state bloccate quasi cinque milioni tra adware, malware e riskware e circa 57.000 software malevoli erano

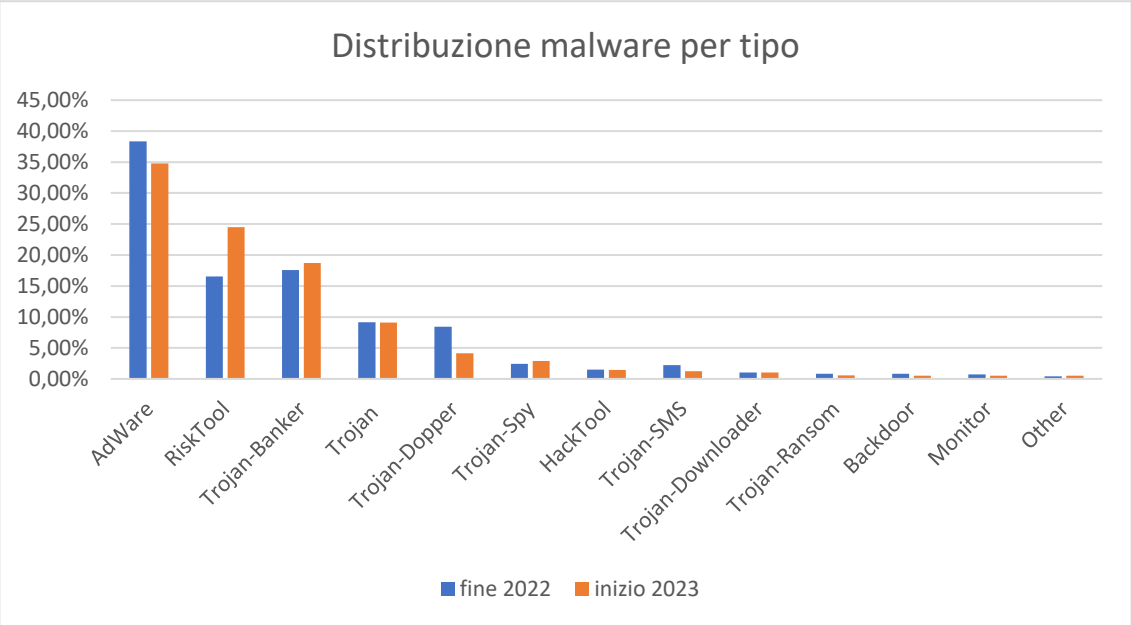

Figura 3: Distribuzione dei nuovi malware a seconda della tipologia; confronto tra la fine 2022 e l'inizio 2023. [9]

ricollegabili a dei Trojan per il mobile banking. Nella **Figura 3** viene evidenziata la distribuzione dei nuovi malware per dispositivi mobili rilevati tra la fine del 2022 e l'inizio del 2023: il primato in entrambi i casi è degli Adware, anche se è comunque presente una buona percentuale di software malevoli che fa chiaramente parte della famiglia dei Trojan.

4 RILEVAMENTO MALWARE

I primi studi riguardanti il rilevamento di malware coincidono con la diffusione dei malware stessi sui dispositivi informatici e principalmente si concentravano sulla rilevazione di virus. Si trattava per lo più di sistemi semplici, progettati per funzionare su una varietà di dispositivi limitata e dove le caratteristiche base erano sempre le stesse. L'esigenza di progettare sistemi di rilevamento dei malware più complessi è diventata sempre più impellente come diretta conseguenza dell'evoluzione tecnologica, che ha permesso la diffusione di dispositivi molto più variegati ed eterogenei sia nelle capacità che nelle caratteristiche. I dispositivi mobili, infatti, possono essere di vari tipi ed esibire diverse caratteristiche ma ognuno di essi deve garantire un buon livello di sicurezza e privacy agli utenti, ragion per cui i sistemi di rilevamento malware devono essere pensati e testati in modo da funzionare correttamente e restituire una buona accuratezza anche su dispositivi con risorse limitate senza minarne l'usabilità; questo, chiaramente, complica notevolmente il problema. Le tecniche per il rilevamento di malware possono avere due diversi obiettivi: nel primo caso, possono andare ad esaminare un dato software prima che questo venga installato ed infetti il dispositivo; nel secondo caso, possono andare ad effettuare dei controlli sui software già installati all'interno dei dispositivi ed esaminarne il comportamento, così da impedire un'eventuale e ulteriore propagazione di un malware. In più, oltre a distinguere tra malware e software non malevoli, alcuni sistemi di rilevamento malware vengono progettati anche per essere in grado di individuare la famiglia di appartenenza dell'eventuale malware, così da permettere l'adozione delle giuste contromisure, qualora necessario.

4.1 DAI PRIMI APPROCCI AL MACHINE LEARNING

Le prime tecniche per il rilevamento dei malware, adottate anche dai software antivirus, prevedevano l'utilizzo delle cosiddette "signature" [6], ovvero delle caratteristiche estratte dai software malevoli che andavano a riconoscere e a rappresentare univocamente un malware. Per implementare il meccanismo di rilevamento, queste signature venivano calcolate a partire da ogni malware noto e venivano memorizzate all'interno di un database, in modo tale da dover semplicemente effettuare una ricerca al suo interno ogni volta che veniva analizzato un software ed era necessario capire se si trattasse di un malware o meno. I pro di questo approccio stanno senz'altro nella semplicità e nella buona accuratezza nel riconoscere i malware già noti; da contro, vi è non solo l'incapacità di questo metodo nell'individuare malware zero-day e quelli che si nascondevano usando delle tecniche di offuscamento, ma anche la necessità di inserire manualmente i nuovi dati all'interno del database ogni volta che veniva scoperto un nuovo malware. Altri approcci si basavano sull'analisi del comportamento del malware o su particolari euristiche formulate su specifiche caratteristiche [6], tuttavia, anche in questo tipo di approcci si riscontravano non poche difficoltà nel rilevare malware sconosciuti o offuscati. Poiché questi metodi si rivelano chiaramente insufficienti per rilevare malware più moderni, che sono attualmente anche quelli più diffusi, è stato necessario cambiare approccio e trovare dei meccanismi in grado di funzionare bene sia con malware più "classici", sia coi cosiddetti malware di "nuova generazione". In questo contesto si inserisce il Machine Learning, una branca dell'Intelligenza Artificiale che permette ai dispositivi di svolgere delle attività senza il bisogno di programmarle esplicitamente. Le tecniche che sfruttano degli algoritmi di apprendimento appaiono utili nell'ambito del rilevamento software perché addestrando opportunamente un sistema, quest'ultimo dovrebbe essere in grado, alla fine, di rilevare anche malware mai visti durante la fase di *training*. In generale, nell'ambito del Machine Learning, le tipologie di apprendimento che possono essere implementate sono principalmente tre:

- **Apprendimento Non Supervisionato**: nel caso dell'apprendimento non supervisionato i dati utilizzati per l'addestramento sono privi di etichette,

spetta al sistema il compito di riconoscere e suddividere in gruppi i dati a seconda delle caratteristiche disponibili;

- **Apprendimento Supervisionato**: nell'apprendimento supervisionato i dati di training sono provvisti di etichette e il compito del sistema è apprendere per poter predire l'etichetta di appartenenza dei dati futuri (classificazione);
- **Apprendimento per Rinforzo**: nell'apprendimento per rinforzo viene implementato un meccanismo di premi e punizioni per permettere al sistema di capire come agire.

Un'alternativa, ovvero una variante delle prime due, talvolta utilizzata nella messa a punto di sistemi per il rilevamento malware, prende il nome di apprendimento **semi-supervisionato**, perché in parte si usa un approccio non supervisionato e in parte si usa quello supervisionato. Indipendentemente dal tipo di apprendimento che viene implementato all'interno del sistema, comunque, risulta essere particolarmente critica la scelta del tipo di analisi che deve essere effettuata durante la fase di estrazione delle caratteristiche. L'obiettivo principale del sistema di rilevamento malware, in ogni caso, è quello di restituire delle buone performance, che vengono valutate utilizzando delle metriche e degli indicatori opportuni.

4.2 ANALISI TRAMITE CARATTERISTICHE STATICHE O DINAMICHE E APPROCCIO IBRIDO

La scelta della tipologia di caratteristiche da analizzare è molto importante perché, a seconda del tipo di caratteristiche prese in considerazione, il sistema di rilevamento malware può restituire risultati diversi. Sono possibili tre tipi di analisi: l'analisi statica, l'analisi dinamica e l'analisi ibrida. L'analisi statica viene effettuata prendendo in considerazione il codice dell'applicazione senza che questa venga mandata in esecuzione, generalmente prevede che venga esaminato il file .apk dell'applicazione o il codice binario per andare ad estrarre delle caratteristiche cosiddette "statiche". L'analisi dinamica, al contrario, ha come obiettivo l'estrazione delle caratteristiche a partire dal comportamento dell'applicazione, perciò quest'ultima viene mandata in esecuzione e solo dopo vengono estratte delle *features*, esaminando gli effetti della messa in esecuzione e le interazioni col sistema operativo; tipicamente vengono utilizzate delle sandbox o degli ambienti controllati per eseguire il file in modo da poterlo esaminare senza correre dei rischi, tuttavia alcuni malware sono in grado di nascondersi in presenza di ambienti creati appositamente, e questo può limitare l'utilità di questo tipo di analisi. In questi casi, l'alternativa è quella di mettere a disposizione dei dispositivi reali utilizzati esclusivamente con l'obiettivo di raccogliere dati e simulare il comportamento di un utente medio per rilevare degli eventi tipici del mondo reale. L'analisi ibrida, infine, prevede che vengano estratte delle caratteristiche sia tramite analisi statica, sia tramite analisi dinamica, in modo da trarre vantaggio da entrambe le tipologie di analisi e riuscire a mettere insieme i lati positivi di entrambi gli approcci. È su quest'ultima tipologia di analisi che si concentra questo lavoro di tesi, poiché l'analisi ibrida rappresenta una frontiera particolarmente interessante per quanto riguarda il rilevamento di malware e i sistemi in grado di utilizzare questo tipo di approccio spesso si rivelano più utili e performanti rispetto a quelli che utilizzano solo un'analisi statica o solamente quella dinamica.

OMISSIS

7 Conclusioni

L'evoluzione costante dei malware e la diffusione crescente di software che implementano diversi tipi di comportamenti malevoli, in concomitanza con la maggiore accessibilità di dispositivi mobili di vari tipi, soprattutto Android, alimenta la necessità che vengano messi a punto dei sistemi di rilevamento malware performanti e in grado di arginare efficacemente queste minacce. Questo lavoro di tesi si è posto l'obiettivo di mettere in evidenza problematiche quali la crescita esponenziale dei malware e della loro complessità nel corso del tempo e la diffusione di dispositivi mobili che esibiscono caratteristiche diverse, per poi concentrarsi su sistemi che utilizzano un approccio ibrido per il rilevamento di malware, una categoria di architetture che risulta essere particolarmente promettente e performante ma che richiede di essere ulteriormente approfondita e sviluppata. I vantaggi dell'utilizzo di un approccio ibrido sono evidenti sperimentalmente, difatti molti dei sistemi presentati hanno messo a confronto le prestazioni ottenute utilizzando solo caratteristiche statiche o solo caratteristiche dinamiche con quelle ottenute quando queste caratteristiche vengono sfruttate insieme e i risultati ottenuti utilizzando contemporaneamente sia le caratteristiche statiche che quelle dinamiche sono sempre stati superiori in termini di correttezza. Le modalità di utilizzo delle caratteristiche statiche e dinamiche possono essere diverse e se da un lato i sistemi che utilizzano separatamente il risultato dell'analisi statica e dinamica sono in grado di ridurre i costi computazionali mandando in esecuzione solamente un sottoinsieme delle applicazioni esaminate, dall'altro lato i sistemi che utilizzano delle caratteristiche integrate riescono a descrivere in modo più specifico e particolareggiato le applicazioni malevole, restituendo quindi dei risultati più affidabili: ideale potrebbe essere l'individuazione di un approccio doppiamente ibrido, sia nell'utilizzo di caratteristiche ibride, sia nella modalità di utilizzo delle stesse, poiché costruire un'architettura in grado di valutare se utilizzare le caratteristiche separatamente o no potrebbe permettere di mettere assieme gli aspetti positivi di entrambe le modalità individuate. Alcuni dei sistemi proposti, inoltre, hanno utilizzato degli approcci basati su cloud proprio con l'intenzione di limitare il più possibile il consumo delle risorse sui dispositivi, tenendo conto dell'eventualità che alcuni di essi siano dotati di una quantità

limitata di energia o di scarsa potenza di calcolo; una possibilità per studi futuri potrebbe essere quella di rivalutare e rielaborare degli approcci che sfruttano eccessivamente le risorse dei dispositivi, affinché possano essere considerati applicabili in dei casi reali e attuali, possibilmente individuando il giusto compromesso tra operazioni da effettuare sul dispositivo e operazioni eseguibili su cloud. Particolarmente d'interesse potrebbe essere l'idea di fondo presentata da MADAM [12], poiché integrare un meccanismo per il rilevamento dei malware direttamente a livello di sistema operativo potrebbe essere un buon modo per fornire un buon livello di sicurezza anche ad utenti inconsapevoli o inesperti, e quindi poco capaci di prendere in autonomia le giuste contromisure e difendersi in caso di attacchi informatici tramite malware; inoltre, così facendo potrebbe essere anche più semplice aggiornare e rendere più robusto il meccanismo di difesa tramite dei periodici aggiornamenti del sistema operativo che agiscono anche su quella funzionalità. Affinché i sistemi per il rilevamento dei malware siano in grado di adottare le giuste contromisure in presenza di minacce, potrebbe essere utile aggiornare e sviluppare ulteriormente gli approcci che restituiscono una classificazione esclusivamente binaria per fare in modo che siano in grado di individuare anche la famiglia o la categoria di appartenenza del malware; inoltre, sia per poter porre rimedio ad eventuali infezioni da malware, sia per permettere un miglioramento futuro dei sistemi di rilevamento malware, una sfida aperta è quella riguardante l'interpretabilità dei risultati. Tra gli approcci presentati, ben pochi hanno affrontato e preso in considerazione il problema relativo all'interpretabilità delle classificazioni ottenute, nonostante si tratti di una questione particolarmente importante e degna di essere attenzionata: comprendere meglio quali parti del codice sono responsabili dei comportamenti malevoli potrebbe permettere di individuare nuovi pattern o dei meccanismi per il rilevamento più efficienti ed efficaci, a prescindere dalla tipologia di malware, inoltre potrebbero essere evidenziate delle differenze tra i malware attuali e quelli del passato per esaminare l'evoluzione dei software malevoli nel corso del tempo e riuscire a prevederne delle possibili evoluzioni future, in modo da irrobustire e rendere più difficile aggirare i sistemi di rilevamento malware attuali. Un approccio che senz'altro va indagato maggiormente è quello riguardante il trasferimento dell'apprendimento, poiché questo potrebbe essere un ottimo espediente per ideare dei sistemi per il rilevamento di malware efficaci e, al tempo stesso,

particolarmente efficienti, in grado di rendere meno onerosi in termini sia di risorse che di tempistiche gli aggiornamenti dei sistemi. Un altro problema che va sicuramente tenuto da conto è quello riguardante l'*adversarial machine learning,* ovvero una categoria di attacchi che ha come obiettivo quello di rendere meno efficaci i sistemi informatici basati su algoritmi di apprendimento automatico, poiché dei potenziali attaccanti potrebbero essere in grado di rendere meno attendibili le classificazioni restituite dai sistemi di rilevamento malware; studi futuri potrebbero concentrarsi non solo sullo sviluppo di architetture resistenti a questi attacchi ma anche sulla revisione e l'ampliamento di quelli già esistenti, individuando delle modifiche da apportare affinché questi approcci risultino più robusti senza che sia necessario sostituirli nella loro interezza. Un aspetto particolarmente importante da tenere in considerazione, poi, riguarda i dataset da utilizzare in fase di apprendimento e valutazione delle prestazioni dei sistemi ideati: come mostrato nel capitolo dedicato, la maggior parte dei dataset in uso risulta essere parecchio datata. Se prendiamo in considerazione il Genome Project (2012) [30] o anche il dataset CICAndMal2017 (2018-2019) [35] [14], la maggior parte di queste raccolte contengono delle applicazioni (malevole e non) che non risultano essere abbastanza attuali e attendibili. Come mostrato nei primi capitoli della tesi, il numero di minacce è in costante aumento e i software malevoli sono in costante evoluzione e cambiano nel corso del tempo, per questo motivo valutare le performance di un sistema basandosi su campioni troppo datati potrebbe restituire dei risultati poco attendibili e rendere i sistemi obsoleti in poco tempo o difficilmente applicabili in situazioni reali. A tal fine, non solo è fondamentale che vengano messi a punto dei dataset che siano sufficientemente vasti, vari, aggiornati con maggiore frequenza e che stiano realmente al passo coi tempi, ma potrebbe essere anche utile, a livello sperimentale, effettuare più spesso dei test che mettano a confronto le prestazioni ottenute utilizzando campioni che fanno riferimento a periodi diversi per valutare in che modo essi rispondono al cambiamento e all'evoluzione dei malware.

8 Bibliografia

[1] S. Samat, «Vivere in un mondo multi-dispositivo con Android,» 2022. [Online]. Available: https://blog.google/intl/it-it/vivere-in-un-mondo-multi-dispositivo-con-android/#:~:text=Ci%20sono%20oltre%20tre%20miliardi,miliardo%20di%20nuovi%20telefoni%20Android..

[2] D. Curry, «Business Of Apps,» 2023. [Online]. Available: https://www.businessofapps.com/data/android-statistics/.

[3] H. Lockheimer, «Google Mobile Blog,» 2012. [Online]. Available: https://googlemobile.blogspot.com/2012/02/android-and-security.html.

[4] Android, «Android Security & Privacy,» 2019. [Online]. Available: https://source.android.com/docs/security/overview/reports/Google_Android_Security_2018_Report_Final.pdf.

[5] G. P. Protect, «Google Play Protect,» [Online]. Available: https://developers.google.com/android/play-protect?hl=it.

[6] O. Aslan e R. Samet, «A Comprehensive Review on Malware Detection,» *IEEE Access,* vol. 8, pp. 6249-6271, 2020.

[7] P. Weichbroth e Ł. Łysik, «Mobile Security: Threats and Best Practices,» *Mobile Information Systems,* vol. 2020, 2020.

[8] T. Shishkova, «Securelist,» 2022. [Online]. Available: https://securelist.com/mobile-threat-report-2022/108844/?fbclid=IwAR3dtme_HnKnPgl391zyzB48uXEDRhlJzKlo6WB0nkojOnGsXe5CU7t1rHQ.

[9] A. Kivva, «Securelist,» June 2023. [Online]. Available: https://securelist.com/it-threat-evolution-q1-2023-mobile-statistics/109893/?fbclid=IwAR2Jf1fMcjoIORFrm-vJ7ITtJgvXdhqavn_utieytTaVXMCbOyPPb4z5kk8.

[10] A. Developers, «Android Developers,» [Online]. Available: https://developer.android.com/guide/topics/manifest/manifest-intro?hl=en.

[11] Y. Liu, Y. Zhang, H. Li e X. Chen, «A Hybrid Malware Detecting Scheme for Mobile Android Applications,» in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, USA, 2016.

[12] A. Saracino, D. Sgandurra, G. Dini e F. Martinelli, «MADAM: Effective and Efficient Behavior-based Android Malware Detection and Prevention,» *IEEE Transactions on Dependable and Secure Computing,* vol. 15, n. 1, pp. 83-97, 2018.

[13] S. Arshad, M. Shah A., A. Wahid, A. Mehmood, H. Song e H. Yu, «SAMADroid: A Novel 3-Level Hybrid Malware Detection Model for Android Operating System,» *IEEE Access,* vol. 6, pp. 4321-4339, 2018.

[14] L. Taheri, A. F. A. Kadir e A. H. Lashkari, «Extensible Android Malware Detection and Family Classification Using Network-Flows and API-Calls,» in *2019 International Carnahan Conference on Security Technology (ICCST)*, Chennai, India, 2019.

[15] J. Feng, L. Shen, Z. Chen, W. Yuying e H. Li, «A Two-Layer Deep Learning Method for Android Malware Detection Using Network Traffic,» *IEEE Access,* vol. 8, pp. 125786-125796, 2020.

[16] A. S. Oliveira e R. J. Sassi, «Chimera: An Android Malware Detection Method Based on Multimodal Deep Learning and Hybrid Analysis.,» *TechRxiv,* 2020.

[17] C. Ding, N. Luktarhan, B. Lu e W. Zhang, «A Hybrid Analysis-Based Approach to Android Malware Family Classification.,» *Entropy,* vol. 23, n. 8, 2021.

[18] F. Faghihi, M. Zulkernine e S. Ding, «AIM: An Android Interpretable Malware detector based on application class modeling,» *Journal of Information Security and Applications,* vol. 75, 2023.

[19] Z. Yuan, Y. Lu, Z. Wang e Y. Xue, «Droid-Sec: Deep Learning in Android Malware Detection,» *SIGCOMM Computer Communication Review,* vol. 44, n. 4, pp. 371-372, 2014.

[20] Shijo e Salim, «Integrated static and dynamic analysis for malware detection,» *Procedia Computer Science,* vol. 46, pp. 804-811, 2015.

[21] M. Lindorfer, M. Neugschwandtner e C. Platzer, «MARVIN: Efficient and Comprehensive Mobile App Classification Through Static and Dynamic Analysis,» in *2015 IEEE 39th Annual Computer Software and Applications Conference*, Taichung, Taiwan, 2015.

[22] Z. Yuan, Y. Lu e Y. Xue, «DroidDetector: Android Malware Characterization and Detection Using Deep Learning,» *Tsinghua Science and Technology,* vol. 21, n. 1, pp. 114-123, 2016.

[23] M. K. Alzaylaee, S. Y. Yerima e S. Sezer, «DL-Droid: Deep learning based android malware detection using real devices,» *Elsevier: Computers & Security,* vol. 89, 2019.

[24] S. J. Hussain, U. Ahmed, H. Liaquat, S. Mir, N. Jhanjhi e M. Humayun, «IMIAD: Intelligent Malware Identification for Android Platform,» in *International Conference on Computer & Information Science (ICCIS)*, Sakaka, Arabia Saudita, 2019.

[25] T. Lu, Y. Du, L. Ouyang, Q. Chen e X. Wang, «Android Malware Detection Based on a Hybrid Deep Learning Model,» *Security and Communication Networks,* vol. 2020, 2020.

[26] R. Surendran, T. Thomas e S. Emmanuel, «A TAN based hybrid model for android malware detection,» *Journal of Information Security and Applications,* vol. 54, 2020.

[27] W. Wang, C. Ren, H. Song, S. Zhang e P. Liu, «FGL_Droid: An Efficient Android Malware Detection Method Based on Hybrid Analysis,» *Security and Communication Networks,* vol. 2022, 2022.

[28] A. K. T. L. Yam, J. M. R. Ballesta, J. A. H. Lanceta, M. K. T. Mogol e R. Labanan, «Hybrid Android Malware Detection Model using Machine learning Algorithms,» in *2022 2nd International Conference in Information and Computing Research (iCORE)*, Cebu, Philippines, 2022.

[29] S. Acharya, U. Rawat e R. Bhatnagar, «A Low Computational Cost Method for Mobile Malware Detection Using Transfer Learning and Familial Classification Using Topic Modelling,» *Applied Computational Intelligence and Soft Computing,* vol. 2022, 2022.

[30] «Android Malware Genome Project,» [Online]. Available: http://www.malgenomeproject.org/.

[31] «The Drebin Dataset,» [Online]. Available: https://drebin.mlsec.org/.

[32] «VirusShare,» [Online]. Available: https://virusshare.com/.

[33] «Contagio Malware Dump,» [Online]. Available: https://contagiodump.blogspot.com/.

[34] «VirusTotal,» [Online]. Available: https://docs.virustotal.com/docs/how-it-works.

[35] A. H. Lashkari, A. F. A. Kadir, L. Taheri e A. A. Ghorbani, «Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification,» in *2018 International Carnahan Conference on Security Technology (ICCST)*, Montreal, Canada, 2018.

[36] «Android Adware and General Malware Dataset (CIC-AAGM2017),» [Online]. Available: https://www.unb.ca/cic/datasets/android-adware.html.

[37] A. Martína, R. Lara-Cabrera e D. Camachoa, «Android malware detection through hybrid features fusion and ensemble classifiers: The AndroPyTool framework and the OmniDroid dataset,» *Information Fusion,* vol. 52, pp. 128-142, 2019.

[38] J. Jang, H. Kang, J. Woo, A. Mohaisen e H. K. Kim, «Andro-Dumpsys: Anti-malware system based on the similarity of malware creator and malware centric information,» *Computers & Security,* vol. 58, pp. 125-138, 2016.

[39] «AndroZoo,» [Online]. Available: https://androzoo.uni.lu/.

[40] D. Maiorca, D. Ariu, I. Corona, M. Aresu e G. Giacinto, «Stealth attacks: An extended insight into the obfuscation effects on Android malware,» *Computers & Security,* vol. 51, pp. 16-31, 2015

[41] V. Agate, P. Ferraro, G. Lo Re, Sajal K. Das, BLIND: A privacy preserving truth discovery system for mobile crowdsensing, Journal of Network and Computer Applications, Volume 223, 2024, 103811, ISSN 1084-8045.

[42] V. Agate, S. Drago, P. Ferraro and G. L. Re, "Anomaly Detection for Reoccurring Concept Drift in Smart Environments," *2022 18th International Conference on Mobility, Sensing and Networking (MSN)*, Guangzhou, China, 2022, pp. 113-120, doi: 10.1109/MSN57253.2022.00031.

[43] A. De Paola, S. Gaglio, G. L. Re and M. Morana, "A hybrid system for malware detection on big data," *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Honolulu, HI, USA, 2018, pp. 45-50, doi: 10.1109/INFCOMW.2018.8406963.

[44] F. Concone, A. De Paola, G. L. Re and M. Morana, "Twitter analysis for real-time malware discovery," *2017 AEIT International Annual*

Conference, Cagliari, Italy, 2017, pp. 1-6, doi: 10.23919/AEIT.2017.8240551.

[45] SMCP: a Secure Mobile Crowdsensing Protocol for fog-based applications F. Concone, G. Lo Re, M. Morana. In Journal of Human-centric Computing and Information Sciences (HCIS 2020)

[46] A. Augello, G. Falzone and G. L. Re, "DCFL: Dynamic Clustered Federated Learning under Differential Privacy Settings," *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Atlanta, GA, USA, 2023, pp. 614-619, doi: 10.1109/PerComWorkshops56833.2023.10150285.