ON THE USE OF MACHINE LEARNING TECHNIQUES TO DETECT MALWARE IN MOBILE APPLICATIONS

Catarina Palma[1] A45241@alunos.isel.pt

Artur Ferreira[1,3] artur.ferreira@isel.pt

Mário Figueiredo[2,3] mario.figueiredo@tecnico.ulisboa.pt

[1] ISEL, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa
[2] IST, Instituto Superior Técnico, Universidade de Lisboa
[3] IT, Instituto de Telecomunicações, Lisboa

The Problem – Malware in Mobile Apps

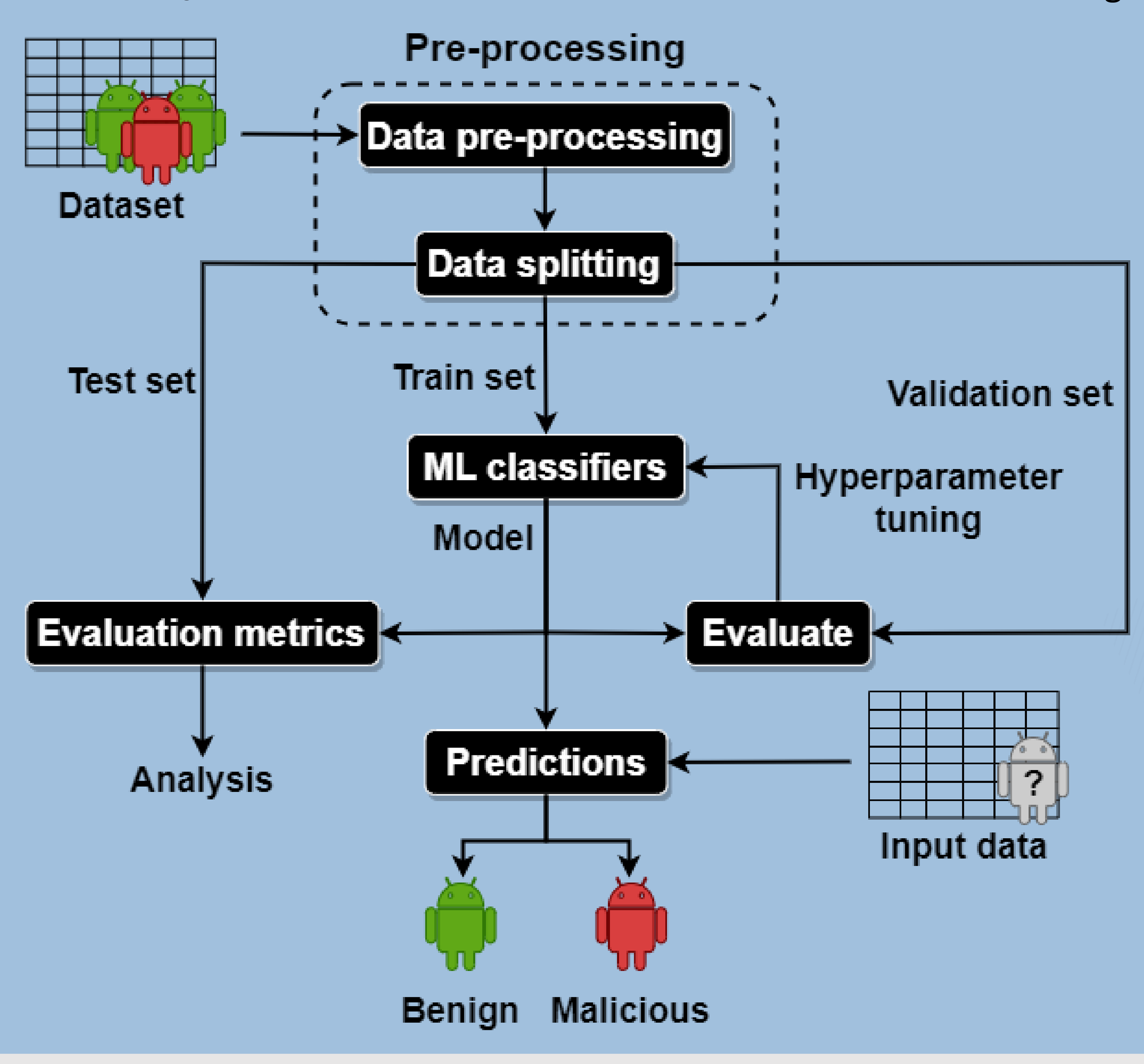
- 70% of mobile phones use Android
- In Q3 2022, Google Play Store hosted around **3.5 million apps**
- **Android applications** are a **prized target for malware** developers
- Existing security measures to mitigate malware are, to some extent, successful
- However, malware keeps growing in both sophistication and diffusion
- In 2020, **5.7 million** Android malware packages were detected, tripling 2019's 2.1 million

Goals

- Explore **machine learning (ML) and feature selection (FS)** approaches to detect malware in Android apps
- Check the importance and impact of:
  - data pre-processing
  - feature selection
  - different classification techniques

Proposed Approach

- **Supervised** ML approach
- **Binary classification problem**
- Two target classes: benign, malicious



Techniques and Evaluation Metrics

Data pre-processing

- Categorical features → numerical features, through label encoding
- Different methods to impute missing values
- Min-Max normalisation

Feature Selection

- Relevance-redundancy FS (RRFS)
- Fisher ratio relevance measure (supervised)
- Absolute cosine redundancy measure

Data splitting

- Random split
- 70/30 ratio for train/test

ML classifiers

Random Forest (RF) | Support Vector Machine (SVM) | K-Nearest Neighbours (KNN) | Naïve Bayes (NB)

Evaluation Metrics

- **Confusion Matrix**
  - True positive (TP) → malicious app as malicious
  - True negative (TN) → benign app as benign
  - False positive (FP) → benign app as malicious
  - False negative (FN) → malicious app as benign

Predicted values \ Actual values	Pos. (+)	Neg. (-)
Pos. (+)	TP	FP
Neg. (-)	FN	TN

- **Accuracy** (Acc) = $\frac{TN+TP}{TN+TP+FN+FP}$
- **Recall** (Rec) = $\frac{TP}{TP+FN}$, (true positive rate or sensitivity)

Public Domain Datasets

	Drebin	CICAndMal2017
'n' instances	15036	29999
'd' features	215	183
Release year	2014	2018
Categorical features	✗	✓
Numerical features	✓	✓
Missing values	✗	✓
Class label ratio	1/3	1/3
Class label majority	benign	malicious

Experimental Results and Evaluation

Baseline

Classifier	Dataset	Acc (%)	TN	FP	FN	TP	Rec (%)
RF	Drebin	98.60	2814	13	50	1634	97.03
RF	CICAndMal2017	80.49	2060	930	781	5001	86.49
SVM	Drebin	97.94	2805	22	71	1613	95.78
SVM	CICAndMal2017	65.82	6	2984	14	5768	99.76
KNN	Drebin	97.58	2782	45	64	1620	96.20
KNN	CICAndMal2017	64.00	940	2050	1108	4672	80.84
NB	Drebin	93.08	2611	216	96	1588	94.30
NB	CICAndMal2017	65.50	461	2529	497	5285	91.40

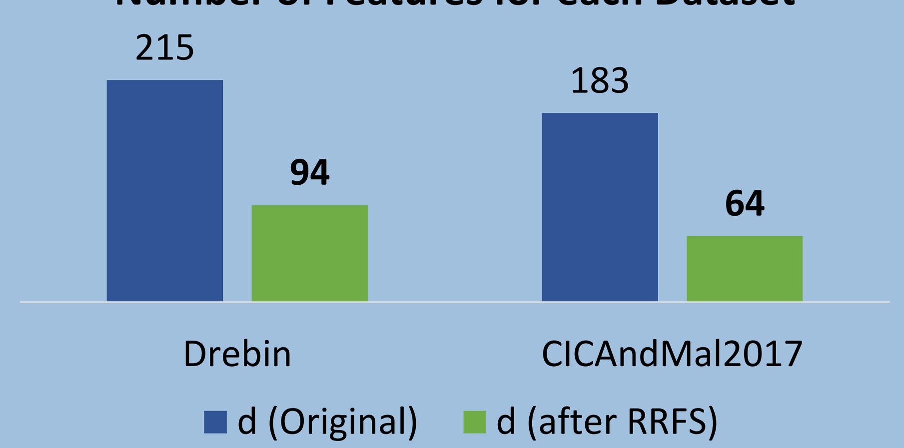
Handling Missing Values

Classifier	Method	Acc (%)	TN	FP	FN	TP	Rec (%)
RF	Remove instances with missing values	80.55	2074	916	790	4992	86.33
RF	Remove features with missing values	80.88	2089	883	838	5190	86.10
RF	Replace missing values with the mean	81.06	2088	884	821	5207	86.38
SVM	Remove instances with missing values	65.74	219	2771	234	5548	95.95
SVM	Remove features with missing values	67.28	419	2553	392	5636	93.50
SVM	Replace missing values with the mean	67.07	373	2599	365	5663	93.94

Feature Selection

Classifier	Dataset	Acc (%) Baseline	Acc (%) RRFS
RF	Drebin	98.60	96.92
RF	CICAndMal2017	80.49	81.42
SVM	Drebin	97.94	96.36
SVM	CICAndMal2017	65.82	70.42

Number of Features for each Dataset



Conclusions

- **ML and FS approaches effectively mitigate this problem**
- RF and SVM classifiers present the best results
- The baseline and dimensionality-reduced datasets exhibit similar metrics
- Results arguably compensated by dimensionality reduction
- A reduction of **56%** in the Drebin dataset and **65%** in the CICAndMal2017 dataset
- No ideal solution was found

Future Work

- Further investigation with different FS techniques
- Additional experiments with different datasets
- More evaluation metrics should be considered

INFORUM SIMPÓSIO DE INFORMÁTICA