MH-1M: One of The Most Comprehensive and Up-to-Date Dataset for Advanced Android Malware Detection

Hendrio Bragança
Joner Assolin*, Vanderson Rocha,
Diego Kreutz*, Eduardo Feitosa

Federal University of Amazonas
*Federal University of Pampa

Motivation

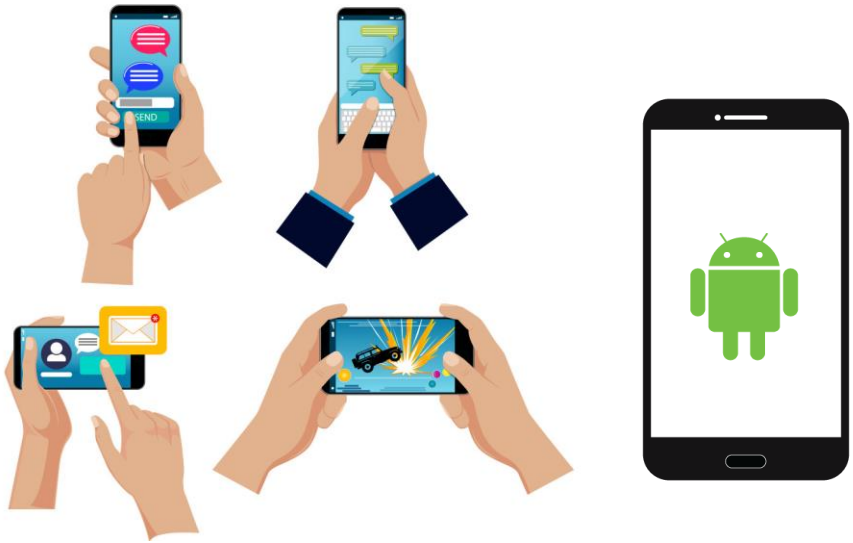
- ✓ 4 billion active smartphone users
- ✓ In 5 years, this number could grow by 50%

https://www.statista.com/

Motivation

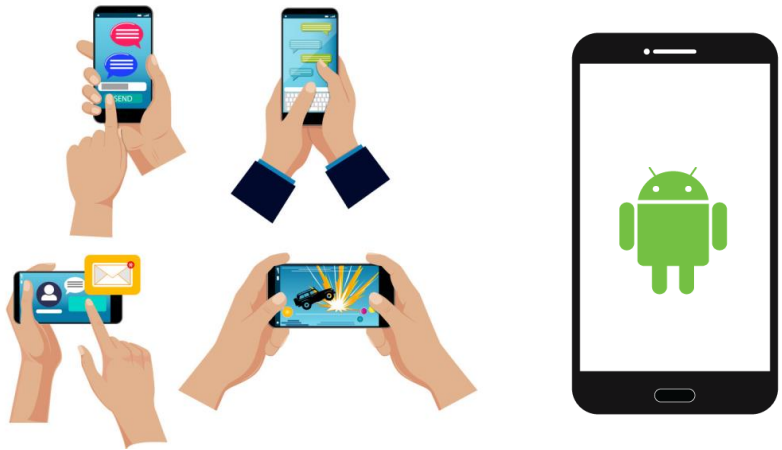
- ✓ The leading mobile operating system Android has 70 % of market share

https://www.statista.com/



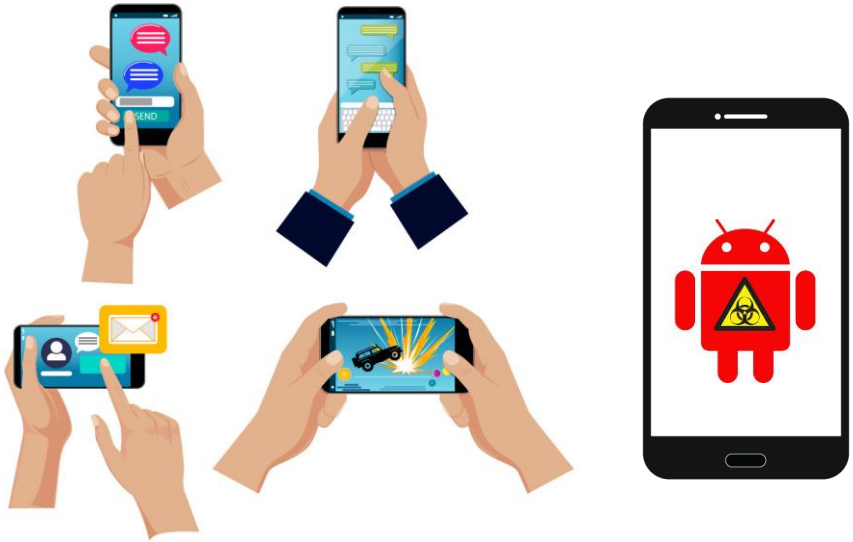
Motivation

- ✓ Open source
- ✓ Accessibility
- ✓ A large market of free applications



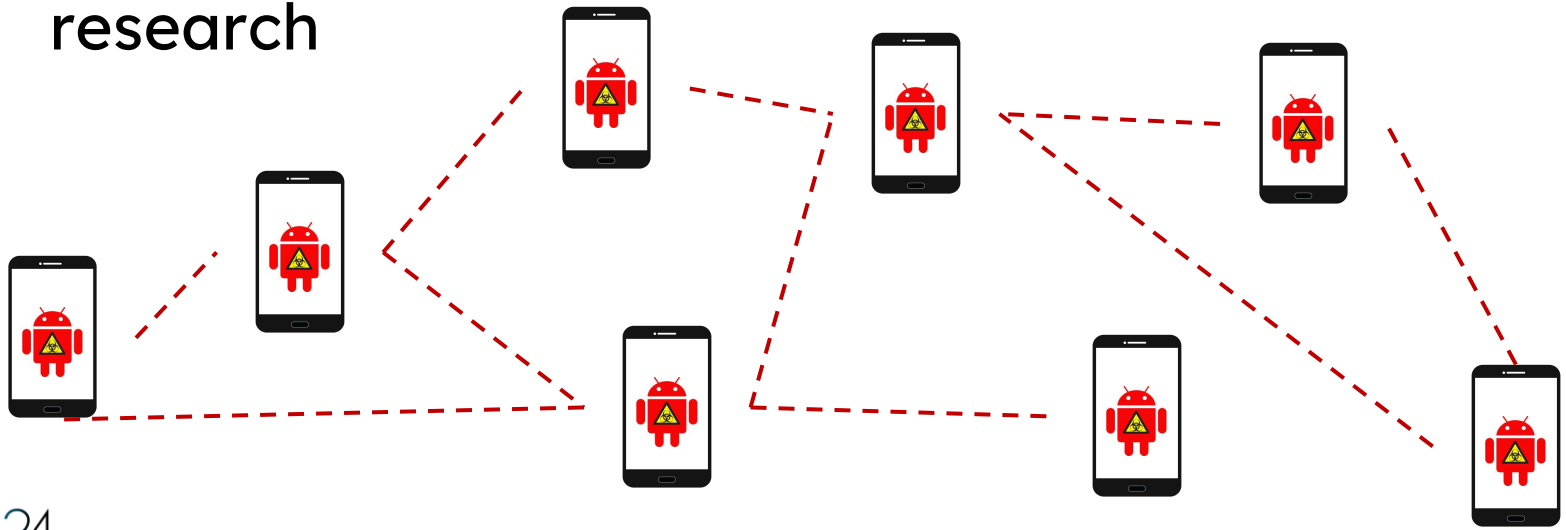
Motivation

- ✓ All of this has a cost: Malicious applications for stealing and destroying user data



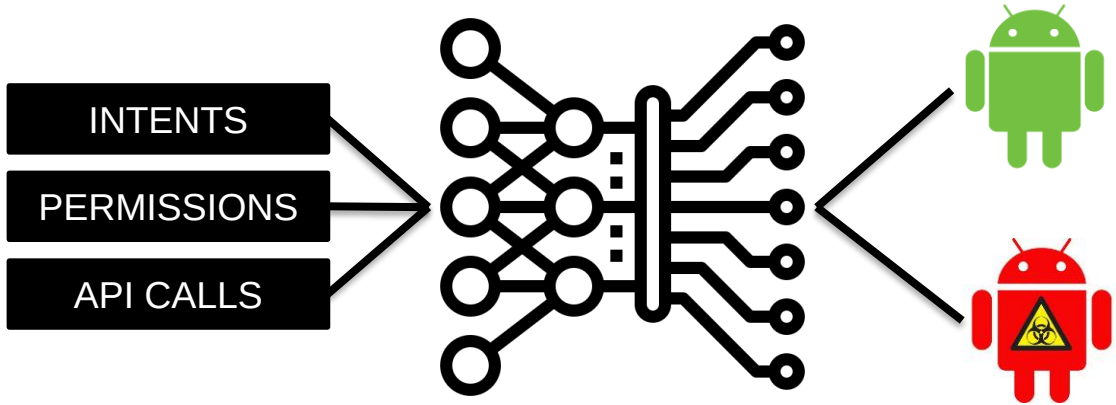
Motivation

- ✓ The spread of Android malware poses a significant challenge for cybersecurity research



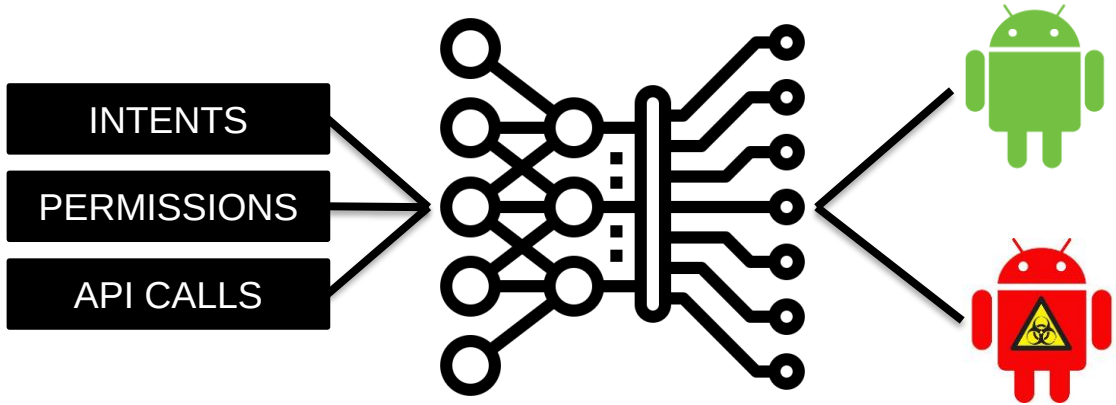
Motivation

- ✓ Machine learning (ML) algorithms have been used for uncovering malwares



Motivation

- ✓ The quality of datasets significantly impacts ML model effectiveness.



Problem

- **Quality** of the dataset used for training
  - Outdated
  - Limited number of Instances
  - Limited number of features
  - Biased

- Models that are trained on **outdated data** that does not accurately reflect the reality of malware

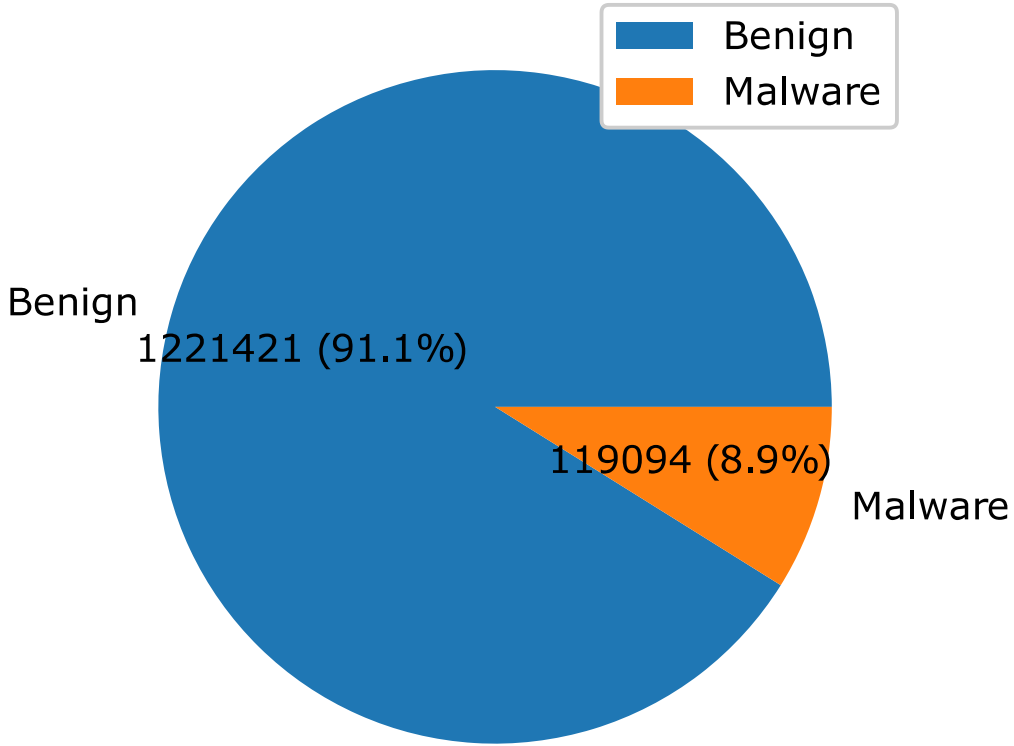
Contributions

- The **MH-1M** Malware Dataset
  - Over 1,000,000 Android samples
  - Permissions, API calls, Intents and Opcodes
- In-depth VirusTotal labeling analysis

Contributions

- Includes large metadata information with more than **400GB**
- APK's signature, file name, package name, Android's official compilation API, VirusTotal outputs
- One of the **largest public** datasets for android malware research

The MH-1M Malware Dataset



- 23247 Features
  - API Calls (22394)
  - Intents (407)
  - Permissions (214)
  - Opcodes (232)

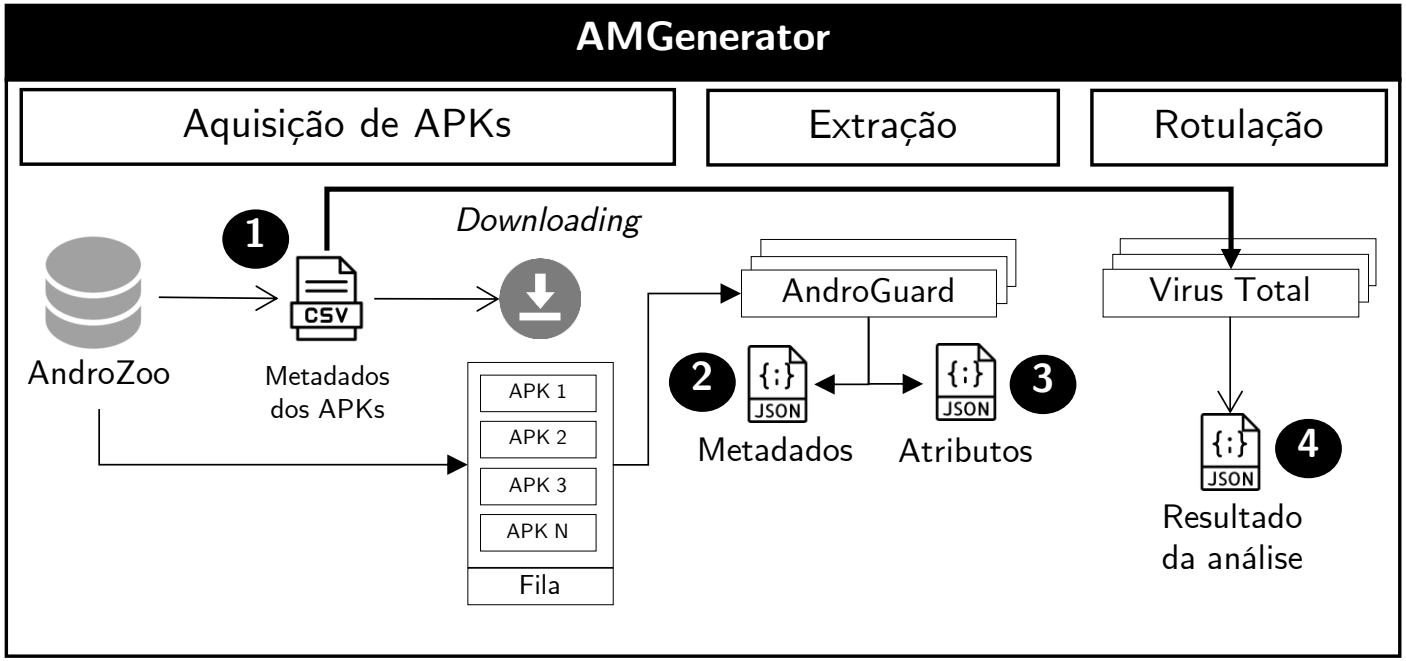
Comparison with other Datasets

Dataset	Features		Samples		
	N.	Type	Malwares	Benign	Total
AndroCrawl	81	AC (24), I (8), P (49)	10170	86562	96732
DefenseDroid	2938	P (1490) I (1448)	6000	5975	11975
DREBIN-215	215	AC (73), P (113), SC (6), I (23)	5555	9476	15036
KronoDroid (devices)	246	P (146), SC (100)	41382	36755	78137
MH-100K	24833	AC (24417), I (250), P (166)	9800	92134	101934
MH-1M	23247	AC (22394), I (407), P (214), OP (232)	119094	1221421	1340515

Comparison with other Datasets

Dataset	Features		Samples		
	N.	Type	Malwares	Benign	Total
AndroCrawl	81	AC (24), I (8), P (49)	10170	86562	96732
DefenseDroid	2938	P (1490) I (1448)	6000	5975	11975
DREBIN-215	215	AC (73), P (113), SC (6), I (23)	5555	9476	15036
KronoDroid (devices)	246	P (146), SC (100)	41382	36755	78137
MH-100K	24833	AC (24417), I (250), P (166)	9800	92134	101934
MH-1M	23247	AC (22394), I (407), P (214), OP (232)	119094	1221421	1340515

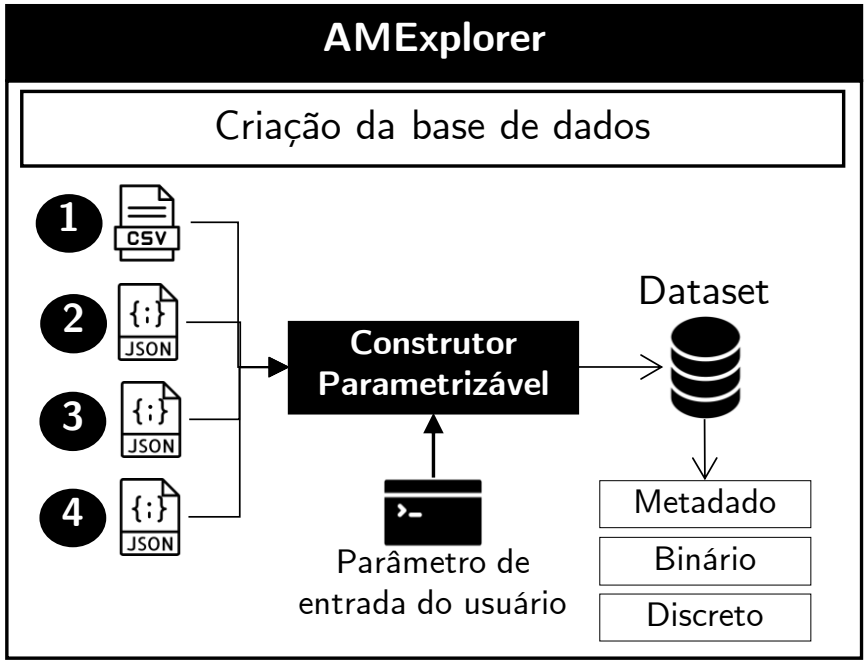
Dataset Tool: AMGenerator



AMGenerator e AMExplorer: Geração de Metadados e Construção de Datasets Android.

https://sol.sbc.org.br/index.php/sbseg_estendido/article/view/27271/27087

Dataset Tool: AMExplorer



AMGenerator e AMExplorer: Geração de Metadados e Construção de Datasets Android.

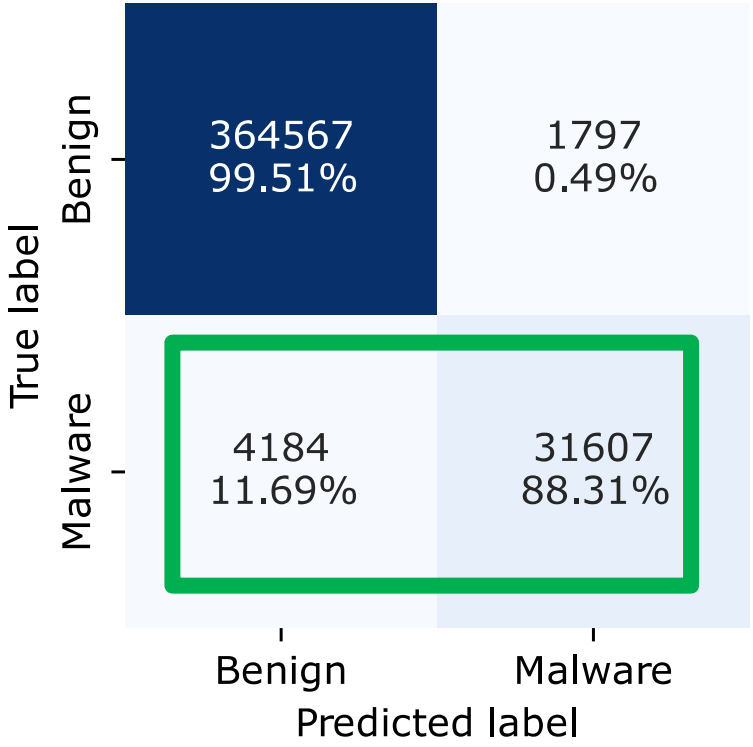
https://sol.sbc.org.br/index.php/sbseg_estendido/article/view/27271/27087

Dataset Analysis

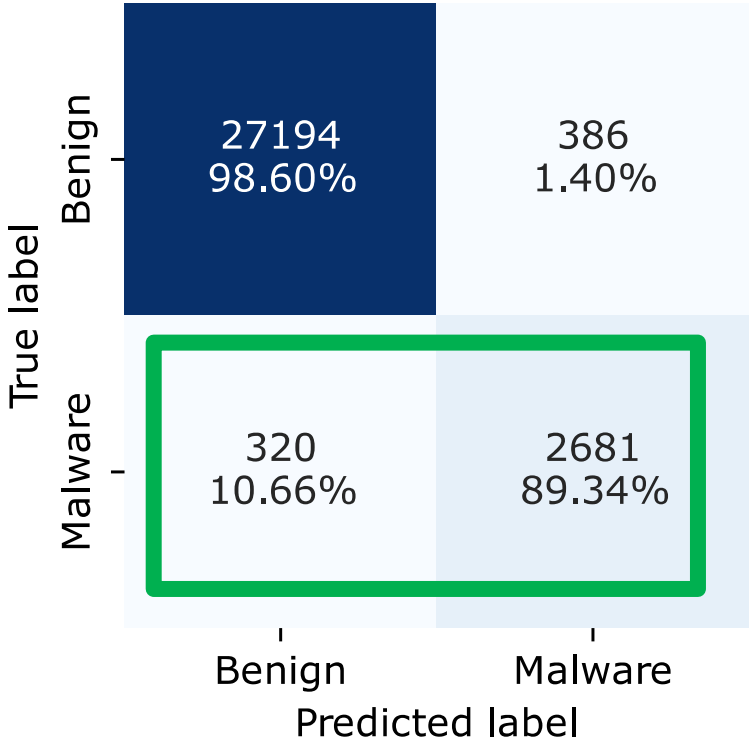
- Two experiments
  - Malware classification
  - Cross-dataset classification

- Classifier and Evaluation metrics
  - XGBOOST
  - Accuracy, precision, recall, F1-Score, and Macro-F1
  - Confusion matrix

Dataset Analysis: Classification

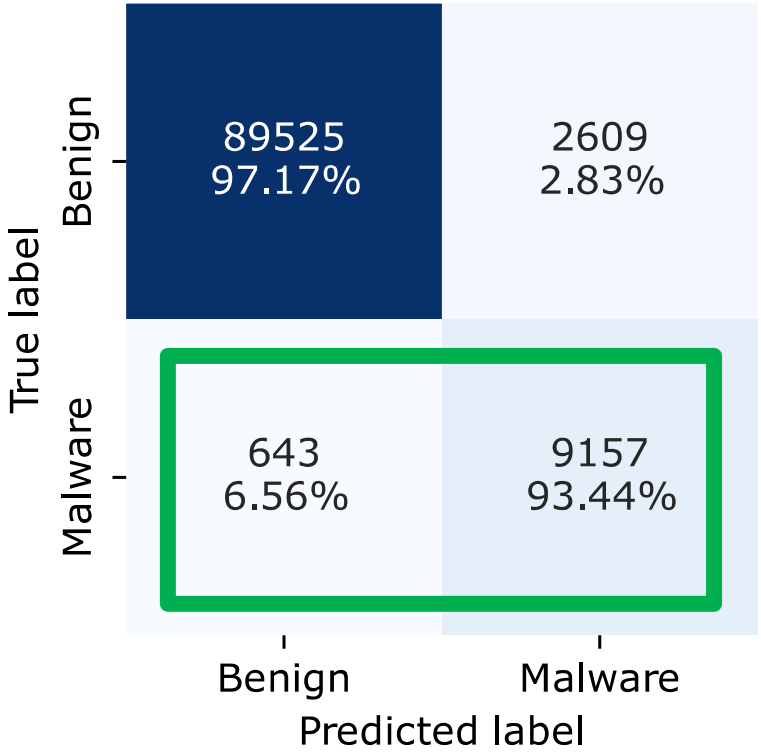


MH-1M

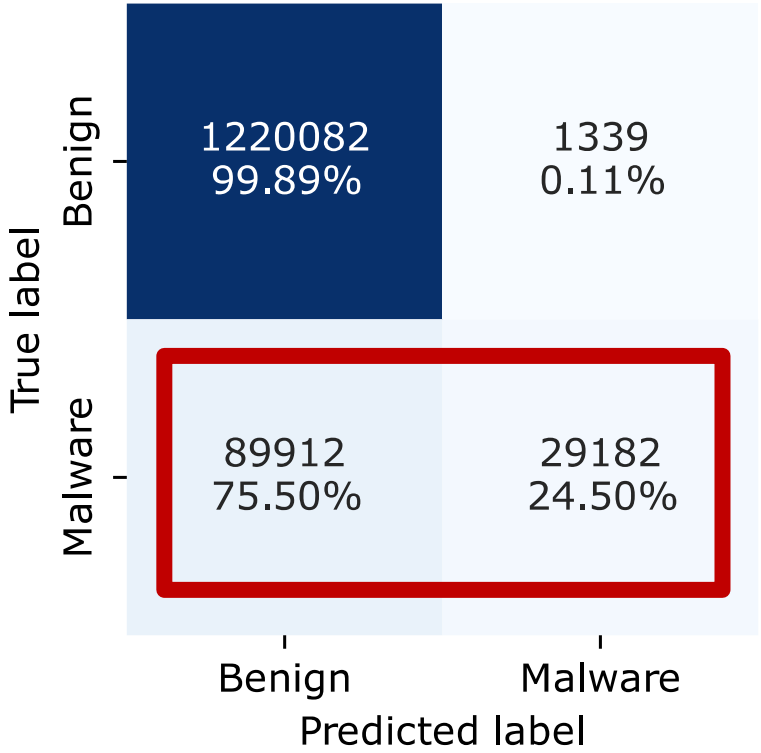


MH-100K

Dataset Analysis: Cross Classification



Training with MH-1M



Training with MH-100K

Dataset Analysis: Cross Classification

Class	Precision	Recall	F1-Score	Support
0	0.9929	0.9717	0.9822	92134
1	0.7783	0.9344	0.8492	9800
Accuracy		0.9681		
Macro (Avg)	0.8856	0.953	0.9157	101934
Weighted (Avg)	0.9722	0.9681	0.9694	101934

Training using MH-1M

Dataset Analysis: Cross Classification

Class	Precision	Recall	F1-Score	Support
0	0.9314	0.9989	0.964	1221421
1	0.9561	0.245	0.3901	119094
Accuracy		0.9319		
Macro (Avg)	0.9437	0.622	0.677	1340515
Weighted (Avg)	0.9336	0.9319	0.913	1340515

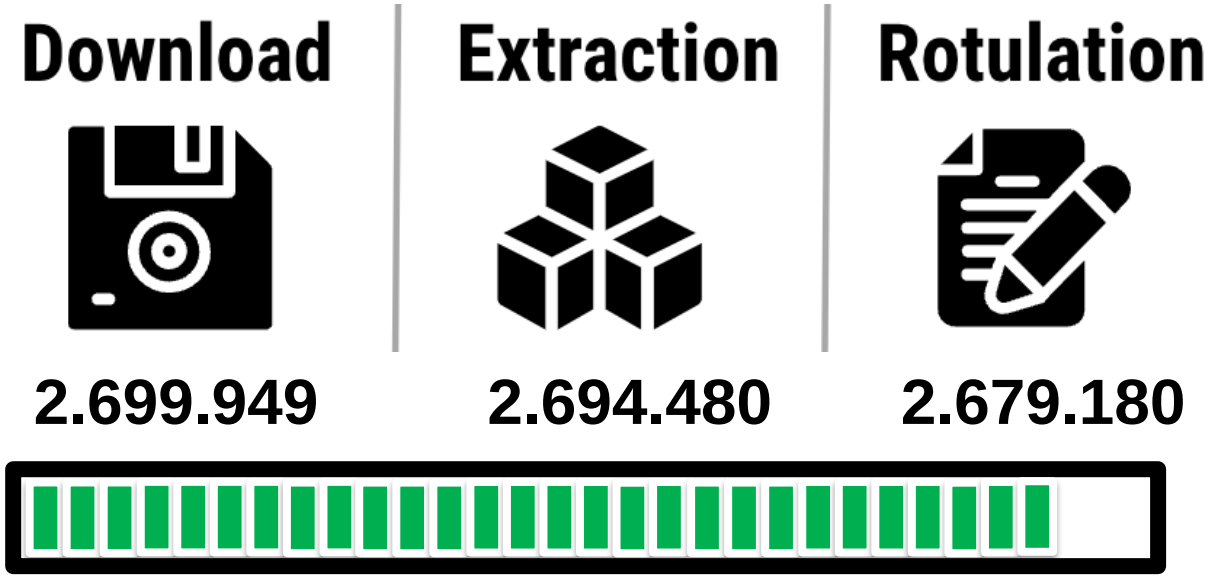
Training using MH-100K

Conclusions

- We introduced the MH-1M dataset, a large collection of more than 1M Android samples
- 3 years of continuous research
- > 400GB Metadata
- A comprehensive resource for building reliable machine learning models

Future Work

- Build a new dataset with more than 2 million samples



Data Availability

https://github.com/Malware-Hunter/MH-1M

Obrigado!

- Hendrio Bragança, Joner Assolin, Vanderson Rocha, Diego Kreutz, Eduardo Feitosa

ETSS
Emerging Technologies and Systems Security

hendrio.luis@icomp.ufam.edu.br