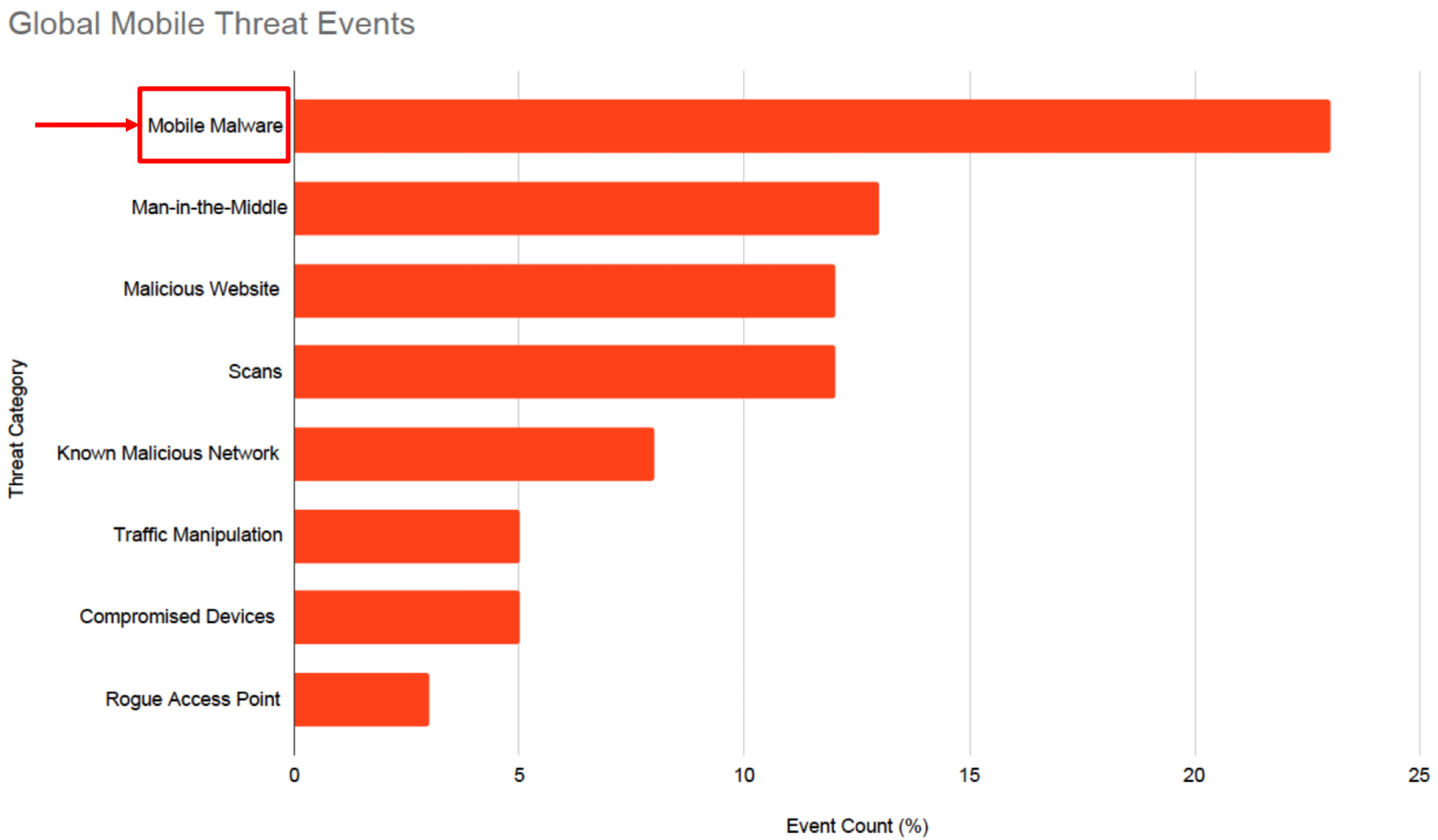
Evaluating Rule-based Global XAI Malware Detection Methods

Rui Li, Olga Gadyatskaya

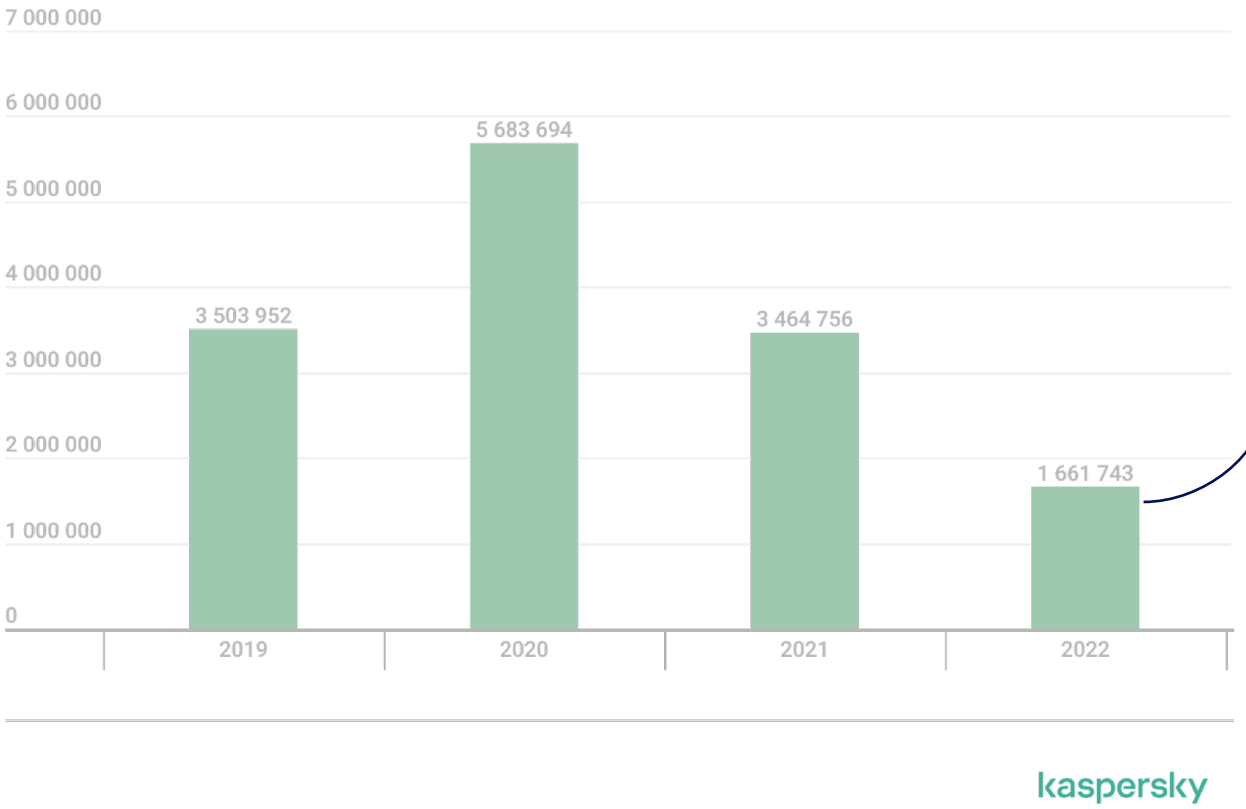
Universiteit Leiden
The Netherlands

Expected Events per Year per Device| Global Average



Zimperium 2022 Mobile Threat Report

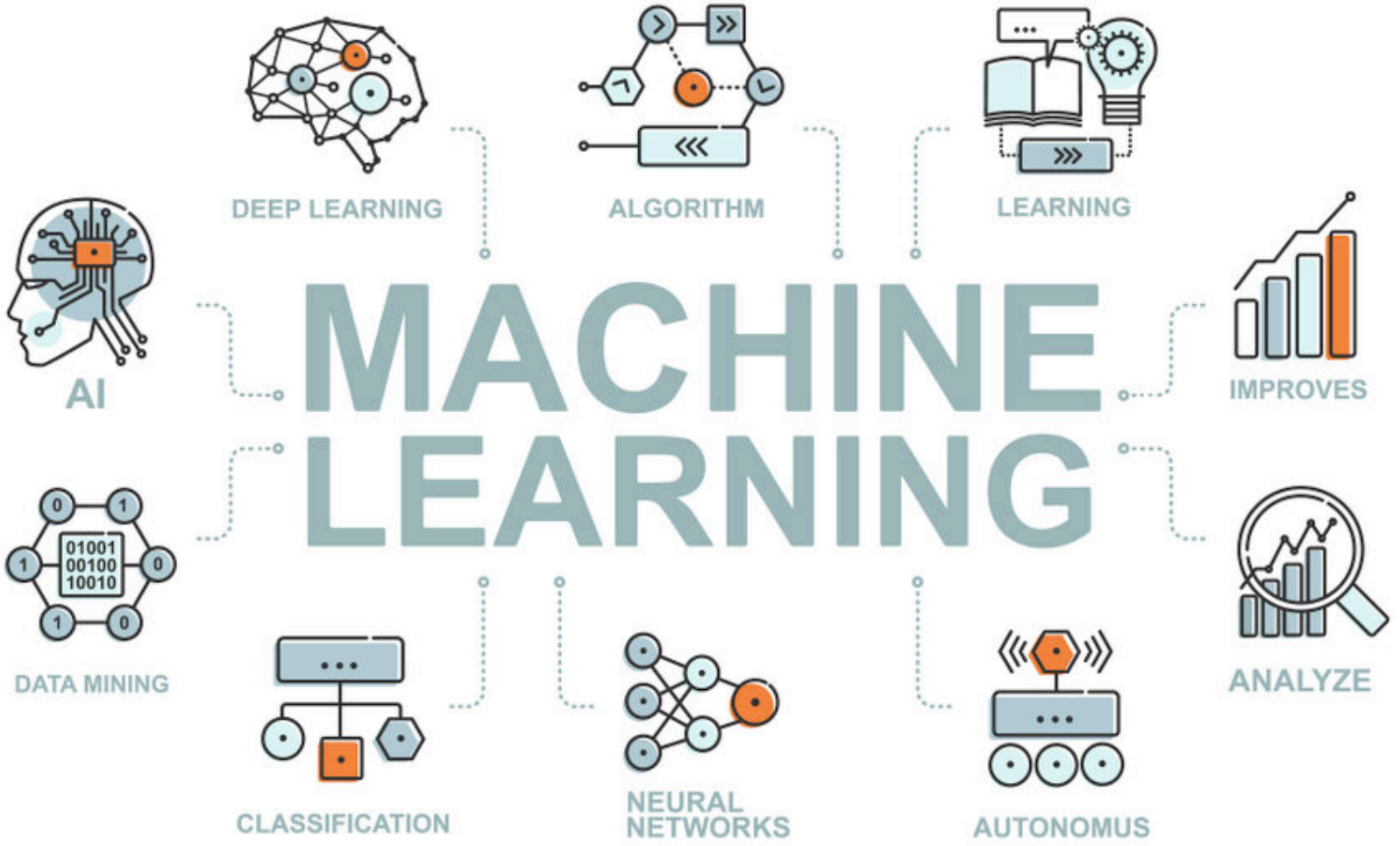
Number of detected mobile malicious installation packages in 2019–2022



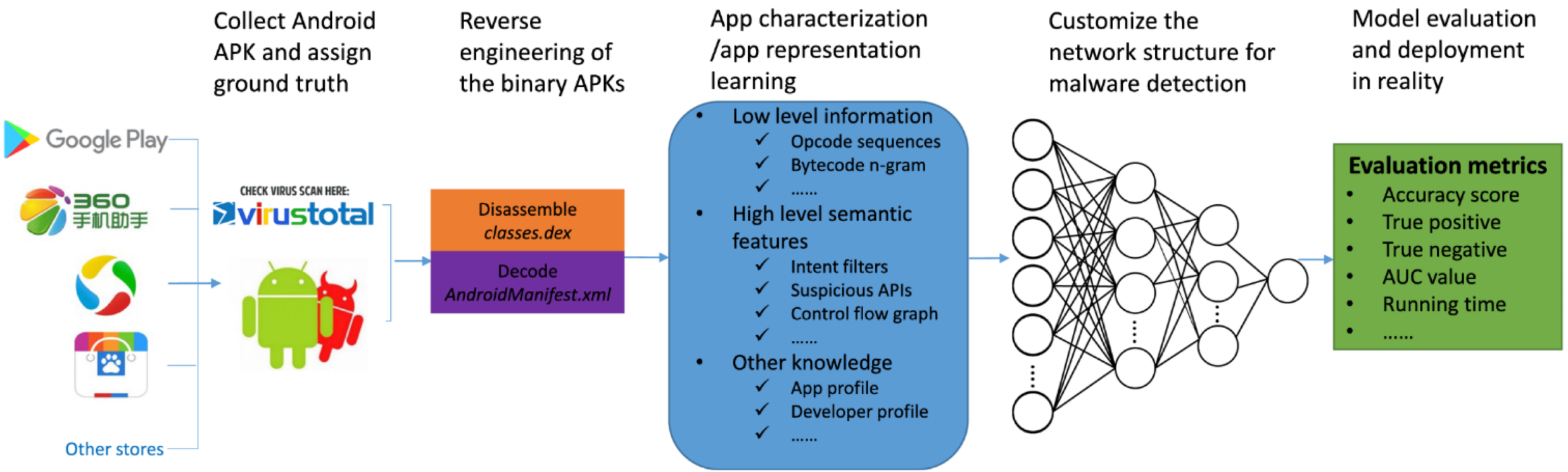
The mobile malware threat landscape in 2022

G DATA security Mobile malware report 2022

How to spot so many malicious apps?

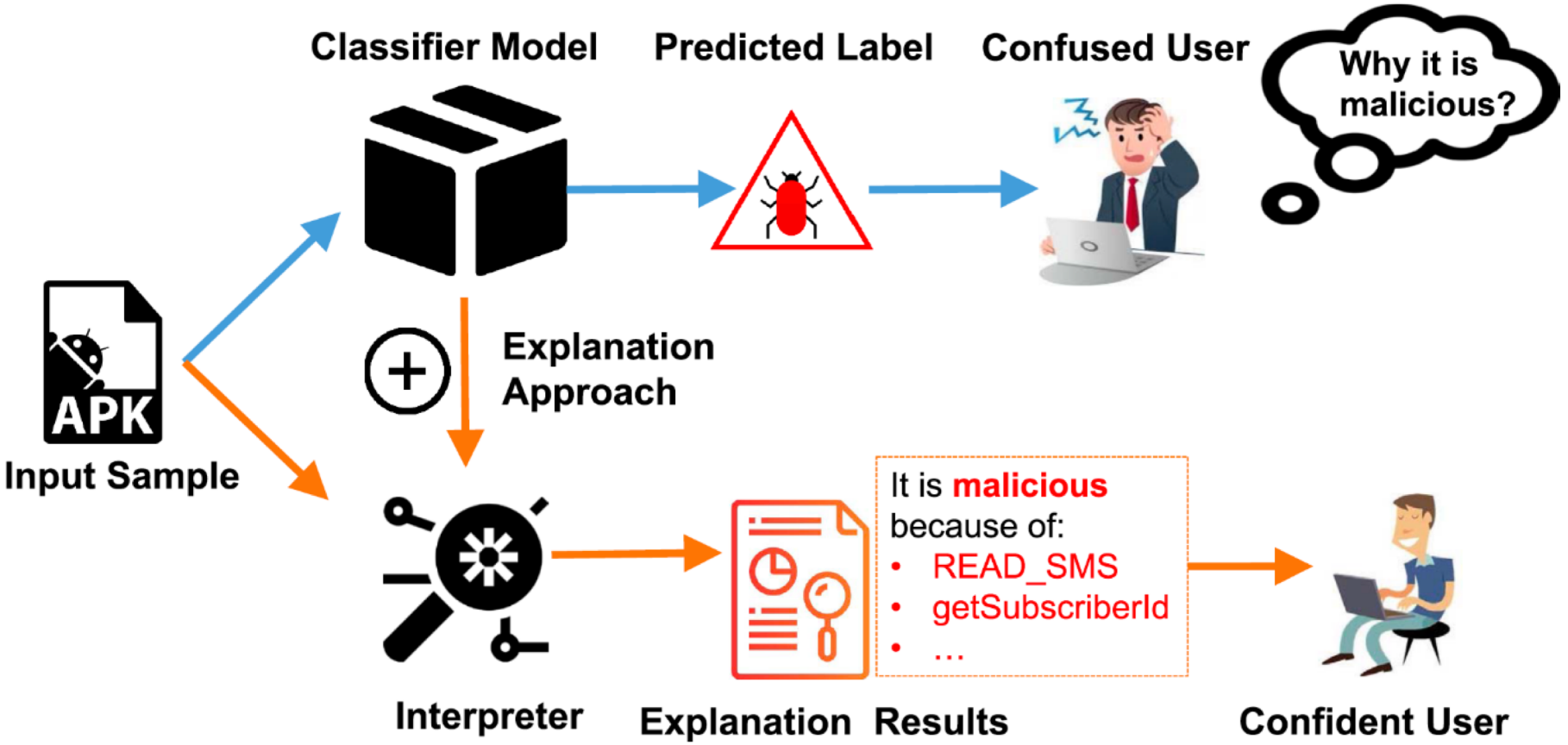

MACHINE LEARNING
DEEP LEARNING
ALGORITHM
LEARNING
AI
IMPROVES
01001
00100
10010
DATA MINING
ANALYZE
CLASSIFICATION
NEURAL NETWORKS
AUTONOMUS

Malware detection with DNNs



J. Qiu et al. "A Survey of Android Malware Detection with Deep Neural Models" in ACM CSUR 2020

Explainable AI



M. Fan et al. “Can We Trust Your Explanations? Sanity Checks for Interpreters in Android Malware Analysis” in IEEE TIFS 2021

Explanation approach

Local explanation

Why a certain app was classified as malware or benign

Prediction result

Reason

Intent:ACTION_BOOT_COMPLETED

Permission:READ_SMS

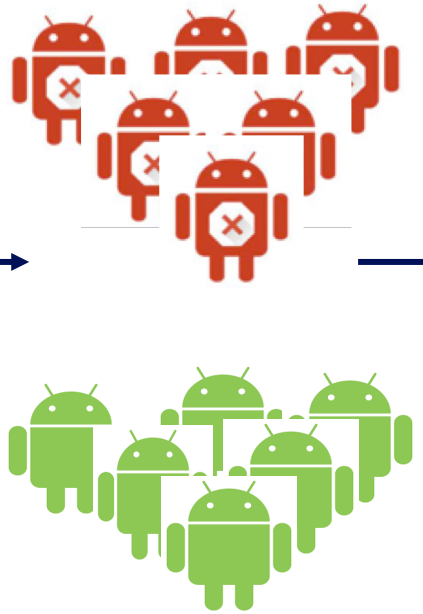
...

Evaluation

Stability
Robustness
Effectiveness

Global explanation

Why a set of apps was classified as malware or benign
(explain the detection model's decision)



If TelephonyManager.getCellocation < 3
& TelephonyManager.getSubscriberID <1
Then 0.05 else 0.88
...

Idea

Formulate three domain-specific properties for ***global*** XAI rule-based malware detection methods

Property	Idea	How ?
Stability	Generated explanations result do not vary between multiple runs	Compute the similarity of different run's rules (explanation result)
Robustness	Remain unaffected when slight variations are applied	Compute the similarity of rules that generated by slight variation samples
Effectiveness	Whether the explanation results are important to the decision-making	Mutate the "and" condition rules to "or" condition rules. Compute the mutate rules' accuracy.

$[X_0, X_1, \dots X_{2000}]$ } sim_1

$[X_{10}, X_{11}, \dots X_{2010}]$ } sim_2

....

$[X_{0+(t-1)*u}, X_1, \dots X_{2000+(t-1)*u}]$ } sim_n

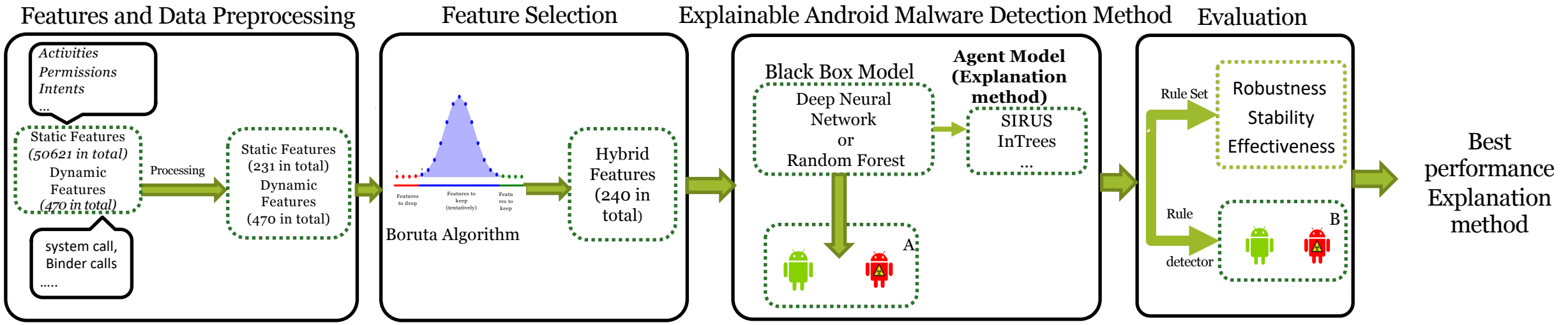
IF f1<1 & f2<1 then 0 else 1

↓

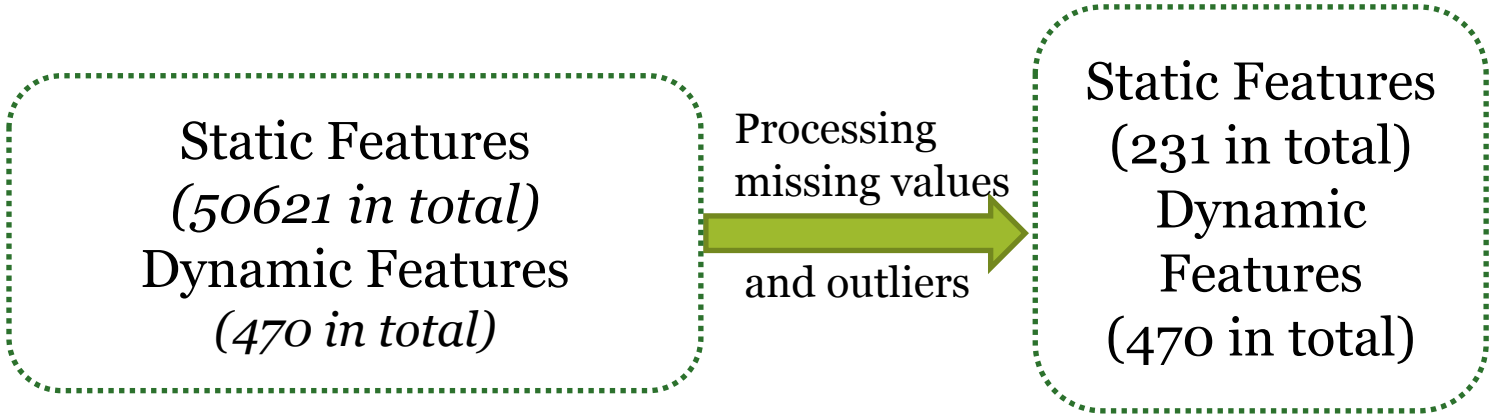
IF f1>=1 then 0 else 1

IF f2>=1 then 0 else 1

Framework


Features and Data Preprocessing
Feature Selection
Explainable Android Malware Detection Method
Evaluation
Activities
Permissions
Intents
...
Static Features
(50621 in total)
Dynamic
Features
(470 in total)
system call,
Binder calls
.....
Processing
Static Features
(231 in total)
Dynamic
Features
(470 in total)
Features
to drop
Features to
keep
(tentatively)
Featu
res to
keep
Boruta Algorithm
Hybrid
Features
(240 in
total)
Black Box Model
Deep Neural
Network
or
Random Forest
Agent Model
(Explanation
method)
SIRUS
InTrees
...
A
Rule Set
Robustness
Stability
Effectiveness
Rule
detector
B
Best
performance
Explanation
method

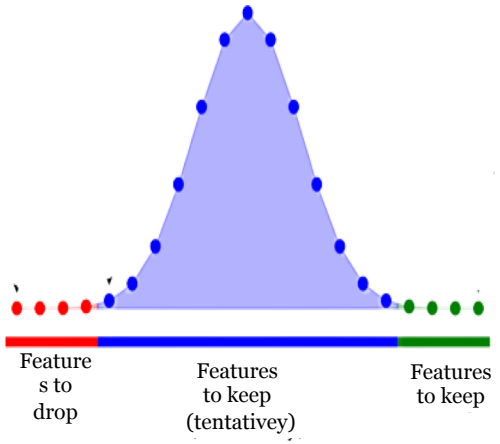
Features and Data preprocessing



- Static featues
  - sensitive permissions, activities, intents, sensitive API calls
- Dynamic features
  - system calls and Binder calls
- Data pre-processing
  - remove the features with missing values
  - transform categorical values into numeric values, etc.

Feature selection

- Boruta algorithm
  - minimizing the impact of random fluctuations and correlations during feature selection

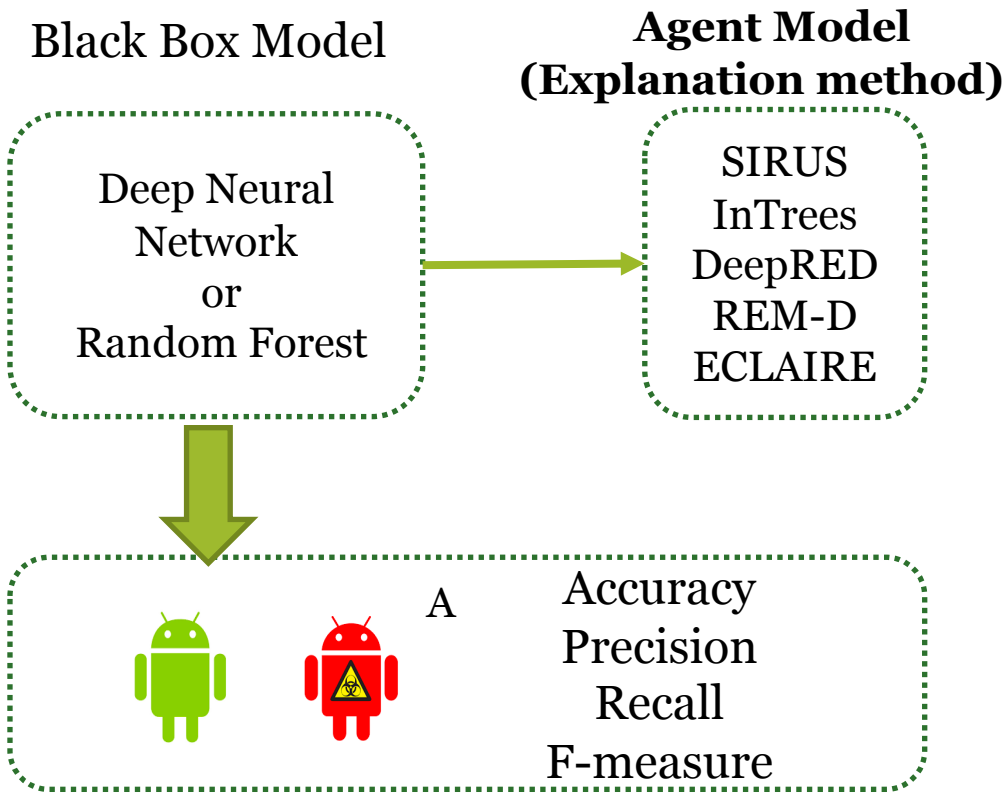


Boruta Algorithm

Hybrid
Features
(240 in total)

Explainable Android Malware Detection Method

- Training black-box malware detection models
- Training agent models



Evaluation

Agent Model (Explanation method)

SIRUS
InTrees
DeepRED
REM-D
ECLAIRE

Rule Set

Stability
Robustness
Effectiveness

Rule Detector

Accuracy
Precision
Recall
F-measure

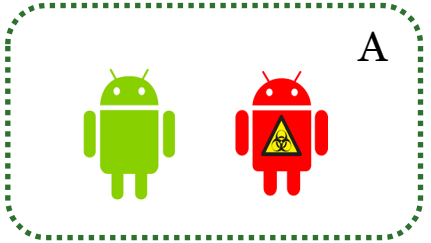
Best performance Explanation method

Data:CICMalDroid 2020

Category	Description	# of samples
Adware	Adware can infect and root-infect a device, forcing it to download specific Adware types and allowing attackers to steal personal information.	1253
Banking	Mobile Banking malware is a specialized malware designed to gain access to the user's online banking accounts by mimicking the original banking applications or banking web interface.	2100
SMS malware	SMS malware exploits the SMS service as its medium of operation to intercept SMS payload for conducting attacks. They control attack instructions by sending malicious SMS, intercepting SMS, and stealing data.	3940
Riskware	Riskware refers to legitimate programs that can cause damage if malicious users exploit them. Consequently, it can turn into any other form of malware such as Adware or Ransomware, which extends functionalities by installing newly infected applications.	2546

Results

Black-box model

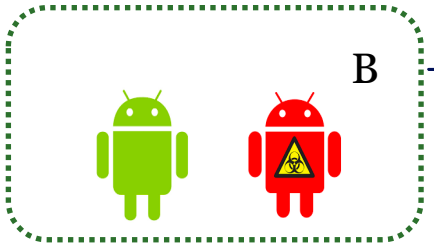


Performance of black-box models

Metric	RF	DNN
Accuracy	98.97%	95.54%
Precision	99.22%	96.73%
Recall	98.72%	93.94%
F-measure	98.74%	95.32%

Rule Set

Rule detector

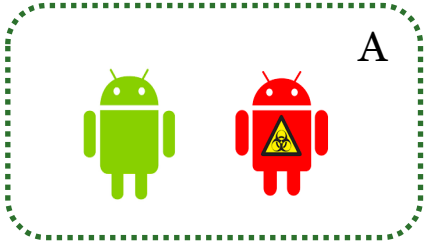


Performance of the chosen explanation methods

Metric	SIRUS (RF)	inTrees (RF)	deepRED (DNN)	REM-D (DNN)	ECLAIRE (DNN)
# of rules	55	12	3	2	2
Stability	96.15%	0%	0%	0%	0%
Robustness	95.56%	0%	0%	0%	0%
Effectiveness	91.65%	86.64%	-	-	-
Accuracy	92.47%	88.19%	88.99%	86.35%	92.34%
Precision	87.20%	91.70%	88.29%	87.85%	86.87%
Recall	99.82%	87.11%	89.05%	76.08%	93.95%
F-measure	93.09%	87.75%	88.67%	81.54%	91.16%

Results

Black-box model



Performance of black-box models

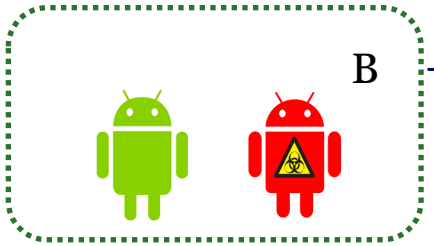
Metric	RF	DNN
Accuracy	98.97%	95.54%
Precision	99.22%	96.73%
Recall	98.72%	93.94%
F-measure	98.74%	95.32%

Performance of the chosen explanation methods

Metric	SIRUS (RF)	inTrees (RF)	deepRED (DNN)	REM-D (DNN)	ECLAIRE (DNN)
# of rules	55	12	3	2	2
Stability	96.15%	0%	0%	0%	0%
Robustness	95.56%	0%	0%	0%	0%
Effectiveness	91.65%	86.64%	-	-	-
Accuracy	92.47%	88.19%	88.99%	86.35%	92.34%
Precision	87.20%	91.70%	88.29%	87.85%	86.87%
Recall	99.82%	87.11%	89.05%	76.08%	93.95%
F-measure	93.09%	87.75%	88.67%	81.54%	91.16%

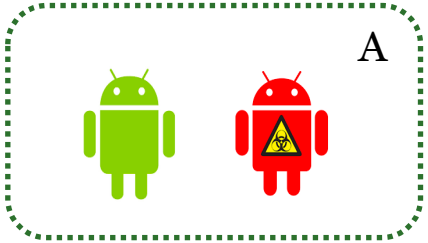
Rule Set

Rule detector



Results

Black-box model



Performance of black-box models

Metric	RF	DNN
Accuracy	98.97%	95.54%
Precision	99.22%	96.73%
Recall	98.72%	93.94%
F-measure	98.74%	95.32%

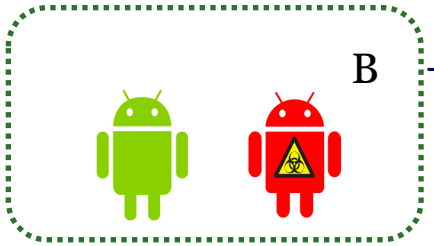
*every run the produced rules are different
*any small change will change the explanation results
*still can provide valuable malware-related information but might confused

Performance of the chosen explanation methods

Metric	SIRUS (RF)	inTrees (RF)	deepRED (DNN)	REM-D (DNN)	ECLAIRE (DNN)
# of rules	55	12	3	2	2
Stability	96.15%	0%	0%	0%	0%
Robustness	95.56%	0%	0%	0%	0%
Effectiveness	91.65%	86.64%	-	-	-
Accuracy	92.47%	88.19%	88.99%	86.35%	92.34%
Precision	87.20%	91.70%	88.29%	87.85%	86.87%
Recall	99.82%	87.11%	89.05%	76.08%	93.95%
F-measure	93.09%	87.75%	88.67%	81.54%	91.16%

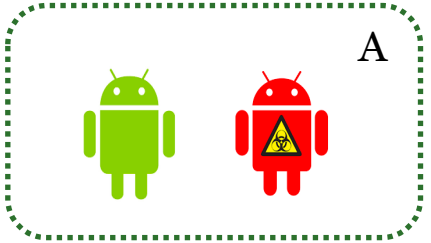
Rule Set

Rule detector



Results

Black-box model



Performance of black-box models

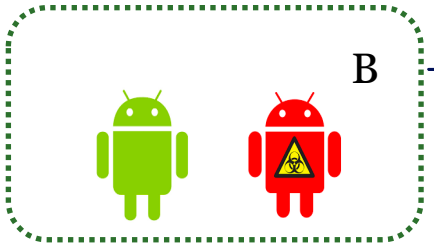
Metric	RF	DNN
Accuracy	98.97%	95.54%
Precision	99.22%	96.73%
Recall	98.72%	93.94%
F-measure	98.74%	95.32%

Performance of the chosen explanation methods

Metric	SIRUS (RF)	inTrees (RF)	deepRED (DNN)	REM-D (DNN)	ECLAIRE (DNN)
# of rules	55	12	3	2	2
Stability	96.15%	0%	0%	0%	0%
Robustness	95.56%	0%	0%	0%	0%
Effectiveness	91.65%	86.64%	-	-	-
Accuracy	92.47%	88.19%	88.99%	86.35%	92.34%
Precision	87.20%	91.70%	88.29%	87.85%	86.87%
Recall	99.82%	87.11%	89.05%	76.08%	93.95%
F-measure	93.09%	87.75%	88.67%	81.54%	91.16%

Rule Set

Rule detector



Conclusion

- Evaluate the quality of rule-based global XAI methods for Android malware detection
- Provide useful insights for malware analysts
- Formulate stability, robustness, and effectiveness to measure the quality of the detection methods

Future work

- Improve the proposed metrics

  being computable for the vast majority of available XAI methods.

- Explore the impact of the number of rules

- Extend our metrics definition

  to cover global XAI methods relying on interpretations in terms of significant features

Thank you!

Universiteit
Leiden
The Netherlands

Rules

ID	Rules
1	if TelephonyManager.getLine1Number < 2 & TelephonyManager.getSubscriberId < 1 then *0.04* else *0.87*
2	if Android.permission.SEND_SMS < 1 & removeAccessibilityInteractionCon nection < 3 then *0.012* else *0.97*
3	If TelephonyManager.getCellLocation < 3 & TelephonyManager.getSubscriberId < 1 then *0.05* else *0.88*
4	if Android.intent.action.PACKAGE_ADDED < 1 & getInstallerPackageName ≥ 1 then *0.0089* else *0.76*
5	if Android.permission.READ_PHONE_STATE < 1 & target_sdk < 19 then *0.24* else *0.52*

- Explanation of rule1
  - TelephonyManager.getLine1Number : obtains a phone number
  - TelephonyManager.getSubscriberId : gets device information.
  - If an application tries to access the phone number at least 2 times or calls for device information, then there is a 87% possibility that it belongs to malware.