ANDROMAK:
State of the art Android ML based detection systems at your fingertips

Matthieu Simonin (SED Rennes)
David Bromberg (WIDE Team)

Context 1/4

Interested in the experimentation practices in various contexts

- Distributed systems[1]
- Organizer / participant of meetings around experimentations
  - XUG meetings: `https://xug.gitlabpages.inria.fr/meetings/` (Check out the upcoming sessions)
  - AskTheSED: coming soon

- Recently interested in Android malware classifications methodology
  - Static analysis with scale requirements
  - Environment local machine, Grid'5000

[1] Cherrueau et al. EnosLib: A Library for Experiment-Driven Research in Distributed Computing. 10.1109/TPDS.2021.3111159

Context 2/4

$$APK : x \xrightarrow{\quad C \quad} C(x) \in \{0, 1\}$$

- Some contributions focuse on building C
  - *"Here is a new C and I'll show you how good it is"*
- Some contributions focuse on evading an existing C
  - Black/white box adversarial / poisoning
  - *"This dataset evades this C, here's how we made it."*
- Some (less) contributions focus on the overall methodology
  - Reproducibility
  - Bias identification and evaluation
  - Brings nuances: *"yes your C is better in this setting BUT. . . "*

The three kinds of contributions are complementary[2]

[2] BlackBurn et al.: The Truth, The Whole Truth, and Nothing But the Truth: A Pragmatic Guide to Assessing Empirical Evaluations `https://doi.org/10.1145/2983574`

Context 3/4

In a supervised context:

$$\mathcal{D} = \{APK\} \quad \left.\begin{array}{l} \xrightarrow{\text{fem}} X \\ \xrightarrow{\text{gt}} y \end{array}\right\} \xrightarrow{\text{train}} \mathcal{C} \xrightarrow{\text{test}} scores$$

$$\Downarrow$$

$$Eval : (\mathcal{D},\ fem,\ gt,\ \mathcal{C},\ \mathcal{S}) \rightarrow scores$$

Some baseline:

- Drebin paper:
  - $Eval(120K + 5K(drebin), fem_{drebin}, VirusTotal*, SVC, 66/33)$
  - comparison with other related fem on the same dataset
- Mamadroid paper:
  - $\mathcal{D}$: product of GW (oldbenign/new benign) + MW(drebin/ V.Share)
  - $\mathcal{C}$: Random Forest / k-NN
  - $\mathcal{S}$: 10-fold CV + mean f1 scores

Context 4/4

$$\left.\begin{array}{l} \mathcal{D} = \{APK\} \begin{array}{l} \xrightarrow{\text{fem}} X \\ \xrightarrow{\text{gt}} y \end{array} \end{array}\right\} \xrightarrow{\text{train}} \mathcal{C} \xrightarrow{\text{test}} scores$$

$$\Downarrow$$

$$Eval : (\mathcal{D},\ fem,\ gt,\ \mathcal{C},\ \mathcal{S}) \rightarrow scores$$

You need Variations[3] into the pipeline == parameter change

- $\mathcal{D}$: dataset
- fem: feature extraction method (core of many contributions)
- gt: ground truth (there isn't one true ground-truth)
- $\mathcal{C}$: model / hyperparameters (including DNN models...)
- selection procedure (5/10 fold-CV, time-aware split . . .)
- and the associated statistical analysis (hypothesis testing . . .)

[3] Feitelson, D.: From repeatability to reproducibility and corroboration. ACMSIGOPS Oper. Syst. Rev.49, 3–11 (2015)

Example of variations in time series classification[4]

2 classical dimensions to fiddle with: datasets(x85) and models(x10)

TABLE I

AVERAGE CLASSIFICATION ACCURACY OVER 10 RUNS OF STSF AND TSF FOR EACH DATASET. OTHER METHODS ARE AS REPORTED IN THEIR PAPERS.

Datasets	# classes	# train	# test	length	1NN-DTW	TSF	RISE	BOSS	FCN	ResNet	HIVE-COTE	PF	TS-CHIEF	STSF
Adiac	37	390	391	176	60.87	76.32	78.01	76.47	**84.40**	82.90	81.07	73.40	79.80	82.79
ArrowHead	3	36	175	251	80.00	72.57	79.43	83.43	84.30	84.50	86.29	**87.54**	83.27	67.49
Beef	5	30	30	470	66.67	86.33	83.33	80.00	69.70	75.30	**93.33**	72.00	70.61	84.00
BeetleFly	2	20	20	512	65.00	75.00	75.00	90.00	86.00	85.00	**95.00**	87.50	91.36	94.00
BirdChicken	2	20	20	512	70.00	80.00	95.00	95.00	**95.50**	88.50	85.00	86.50	90.91	90.00
Car	4	60	60	577	76.67	76.67	80.00	83.33	90.40	**92.50**	86.67	84.67	85.45	81.50
CBF	3	30	900	128	99.44	97.27	95.11	99.78	99.40	99.50	**99.89**	99.33	99.79	97.90
ChlorineConcentration	3	467	3840	166	65.00	74.93	76.85	66.09	81.40	**84.40**	71.20	63.39	71.67	78.04
CinCECGTorso	4	40	1380	1639	93.04	95.07	98.62	88.70	82.40	82.60	**99.64**	93.43	98.32	98.49
Coffee	2	28	28	286	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Computers	2	250	250	720	62.40	72.00	76.40	75.60	**82.20**	81.50	76.00	64.44	70.51	75.60
CricketX	12	390	390	300	77.95	64.97	69.74	73.59	79.20	79.10	**82.31**	80.21	81.38	68.33
CricketY	12	390	390	300	75.64	70.87	71.79	75.38	78.70	80.30	**84.87**	79.38	80.19	74.77
CricketZ	12	390	390	300	73.59	66.62	70.51	74.62	81.10	81.20	83.08	80.10	**83.40**	72.18
DiatomSizeReduction	4	16	306	345	93.46	94.80	92.81	93.14	31.30	30.10	94.12	96.57	**97.30**	96.63
DistalPhalanxOutlineAgeGroup	3	400	139	80	62.59	74.82	**76.26**	74.82	71.00	71.70	**76.26**	73.09	74.62	72.81
DistalPhalanxOutlineCorrect	2	600	276	80	72.46	77.17	77.54	72.83	76.00	77.10	77.17	**79.28**	78.23	78.84
DistalPhalanxTW	6	400	139	80	63.31	66.91	67.63	67.63	**69.00**	66.50	68.35	65.97	67.04	68.27
Earthquakes	2	322	139	512	72.66	74.82	74.82	74.82	72.70	71.20	74.82	75.40	74.82	**76.91**
ECG200	2	100	100	96	88.00	85.50	88.00	87.00	88.90	87.40	85.00	**90.90**	86.18	88.00
ECG5000	5	500	4500	140	92.51	93.89	93.69	94.13	94.00	93.40	**94.62**	93.65	94.54	94.21
ECGFiveDays	2	23	861	136	79.67	93.69	99.88	**100.00**	98.70	97.50	**100.00**	84.92	**100.00**	97.77
ElectricDevices	7	8926	7711	96	63.08	69.25	66.35	**79.92**	70.20	72.90	77.03	70.60	75.53	74.06
FaceAll	14	560	1690	131	80.77	76.83	76.15	78.17	**94.50**	83.90	80.30	89.38	84.14	78.85
FaceFour	4	24	88	350	89.77	98.41	89.77	**100.00**	92.80	95.50	95.45	97.39	**100.00**	97.73
FacesUCR	14	200	2050	131	90.78	90.01	87.51	95.71	94.60	95.50	96.29	94.59	**96.63**	88.59
FiftyWords	50	450	450	270	76.48	73.28	69.23	70.55	67.90	72.70	80.88	83.14	**84.50**	77.05
Fish	7	175	175	463	83.43	85.26	84.57	98.86	95.80	97.90	98.86	93.49	**99.43**	90.34
FordA	2	3601	1320	500	66.52	81.52	94.09	92.95	90.40	92.00	**96.44**	85.46	94.10	96.30
FordB	2	3636	810	500	59.88	68.77	81.11	82.00	87.80	**91.30**	82.35	71.49	82.96	79.42
GunPoint	2	50	150	150	91.33	95.07	98.00	**100.00**	**100.00**	99.10	**100.00**	99.73	**100.00**	92.00
Ham	2	109	105	431	60.00	74.29	68.57	66.67	71.80	**75.70**	66.67	66.00	71.52	73.81
HandOutlines	2	1000	370	2709	87.84	91.89	88.38	91.10	80.60	91.10	**93.24**	92.14	93.22	92.03
Haptics	5	155	308	1092	41.56	43.57	45.78	46.10	48.00	51.90	**51.95**	44.45	51.68	50.75
Herring	2	64	64	512	53.12	60.94	64.06	54.69	60.80	61.90	**68.75**	57.97	58.81	62.97
InlineSkate	7	100	550	1884	38.73	32.24	34.91	51.64	33.90	37.30	50.00	54.18	52.69	**55.47**
InsectWingbeatSound	11	220	1980	256	57.37	63.28	65.51	52.32	39.30	50.70	65.51	61.87	64.29	**66.56**
ItalyPowerDemand	2	67	1029	24	95.53	97.00	95.34	90.86	96.10	96.30	96.31	96.71	**97.06**	**97.06**
LargeKitchenAppliances	3	375	375	720	79.47	57.07	63.73	76.53	**90.20**	90.00	86.40	78.19	80.68	79.39
Lightning2	2	60	61	637	**86.89**	79.51	70.49	83.61	73.90	77.00	81.97	86.56	74.81	72.46
Lightning7	7	70	73	319	71.23	74.11	69.86	68.49	82.70	**84.50**	73.97	82.19	76.34	76.99
Mallat	8	55	2345	1024	91.43	96.46	92.15	93.82	96.70	97.20	96.20	95.76	**97.50**	96.88
Meat	3	60	60	448	93.33	93.33	93.33	90.00	85.30	**96.80**	93.33	93.33	88.79	93.17
MedicalImages	10	381	760	99	74.74	78.00	66.18	71.84	77.90	77.00	77.76	75.82	**79.58**	78.59
MiddlePhalanxOutlineAgeGroup	3	400	154	80	51.95	57.79	**59.74**	54.55	55.30	56.90	**59.74**	56.23	58.32	56.82
MiddlePhalanxOutlineCorrect	2	600	291	80	76.63	82.82	82.13	78.01	80.10	80.90	83.16	83.64	**85.35**	82.27
MiddlePhalanxTW	6	399	154	80	50.65	56.49	**59.09**	54.55	51.20	48.40	57.14	52.92	55.02	58.90
MoteStrain	2	20	1252	84	88.63	86.90	87.22	87.86	93.70	92.80	93.29	90.24	**94.75**	92.36
NonInvasiveFetalECGThorax1	42	1800	1965	750	82.90	89.97	90.13	83.82	**95.60**	94.50	93.03	90.66	91.13	93.27
NonInvasiveFetalECGThorax2	42	1800	1965	750	87.02	91.13	91.55	90.08	**95.30**	94.60	94.45	93.99	94.50	94.06
OliveOil	4	30	30	570	86.67	90.67	90.00	86.67	72.30	83.00	90.00	86.67	88.79	**93.33**
OSULeaf	6	200	242	427	59.92	58.39	64.88	95.45	97.70	97.90	97.93	82.73	**99.14**	79.83
PhalangesOutlinesCorrect	2	1800	858	80	76.11	80.30	81.35	77.16	82.00	83.90	80.65	82.35	**84.50**	83.17
Phoneme	39	214	1896	1024	22.68	21.20	35.55	26.48	32.50	33.40	**38.24**	32.01	36.91	32.52
Plane	7	105	105	144	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
ProximalPhalanxOutlineAgeGroup	3	400	205	80	78.54	84.88	85.37	83.41	83.10	85.30	**85.85**	84.63	84.97	84.44
ProximalPhalanxOutlineCorrect	2	600	291	80	79.04	82.82	87.63	84.88	90.30	**92.10**	87.97	87.32	88.82	90.52
ProximalPhalanxTW	6	400	205	80	76.10	81.46	80.98	80.00	76.70	78.00	81.46	77.90	**81.86**	76.49
RefrigerationDevices	3	375	375	720	44.00	**58.93**	54.40	49.87	50.80	52.50	55.73	53.23	55.83	58.03
ScreenType	3	375	375	720	41.07	45.60	52.80	46.40	**62.50**	62.20	58.93	45.52	50.81	53.33
ShapeletSim	2	20	180	500	69.44	47.78	78.33	**100.00**	72.40	77.90	**100.00**	77.61	**100.00**	98.33
ShapesAll	60	600	600	512	80.17	79.17	83.33	90.83	89.50	92.10	90.50	88.58	**93.00**	85.22
SmallKitchenAppliances	3	375	375	720	67.20	81.07	81.07	72.53	78.30	78.60	**85.33**	74.43	82.21	83.44
SonyAIBORobotSurface1	2	20	601	70	69.55	75.64	82.20	63.23	**96.00**	95.80	76.54	84.58	82.64	90.67
SonyAIBORobotSurface2	2	27	953	65	85.94	81.86	91.08	85.94	**97.90**	97.80	92.76	89.63	92.48	83.25
StarLightCurves	3	1000	8236	1024	89.83	96.40	97.50	97.78	96.10	97.20	98.15	98.13	**98.24**	97.84
Strawberry	2	613	370	235	94.59	96.49	96.49	97.57	97.20	**98.10**	97.03	96.84	96.63	96.38
SwedishLeaf	15	500	625	128	84.64	89.57	93.60	92.16	**96.90**	95.60	95.36	94.66	96.55	94.29
Symbols	6	25	995	398	93.77	88.56	93.27	96.68	95.50	90.60	97.39	96.16	**97.66**	88.39
SyntheticControl	6	300	300	60	98.33	97.57	66.67	96.67	99.00	**100.00**	99.67	99.53	99.79	99.03
ToeSegmentation1	2	40	228	277	75.00	74.12	90.79	93.86	96.10	96.30	**98.25**	92.46	96.53	84.43
ToeSegmentation2	2	36	130	343	90.77	81.54	90.00	**96.15**	88.00	90.60	95.38	86.23	95.38	88.46
Trace	4	100	100	275	99.00	97.80	96.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.00
TwoLeadECG	2	23	1139	82	86.83	90.39	88.76	98.07	**100.00**	**100.00**	99.65	98.86	99.46	98.72
TwoPatterns	4	1000	4000	128	99.85	94.67	43.50	99.30	87.10	**100.00**	**100.00**	99.96	**100.00**	99.77
UWaveGestureLibraryAll	8	896	3582	945	96.23	95.73	92.13	93.89	81.70	86.00	96.85	**97.23**	96.89	95.48
UWaveGestureLibraryX	8	896	3582	315	77.44	78.96	61.86	76.21	75.40	78.00	83.98	82.86	**84.11**	81.10
UWaveGestureLibraryY	8	896	3582	315	70.18	71.15	66.69	68.51	63.90	67.00	76.55	76.15	**77.23**	74.16
UWaveGestureLibraryZ	8	896	3582	315	67.50	73.58	64.99	69.49	72.60	75.00	78.31	76.40	**78.44**	75.86
Wafer	2	1000	6164	152	99.59	99.50	99.55	99.48	99.70	99.90	99.94	99.55	99.91	**99.98**
Wine	2	57	54	234	61.11	62.96	64.81	74.07	58.70	74.40	77.78	56.85	**89.06**	66.85
WordSynonyms	25	267	638	270	74.92	62.43	58.93	63.79	56.40	62.20	73.82	77.87	**78.74**	63.64
Worms	5	181	77	900	53.25	61.04	66.23	55.84	76.50	79.10	55.84	71.82	**80.17**	76.75
WormsTwoClass	2	181	77	900	58.44	62.34	**83.12**	**83.12**	72.60	74.70	77.92	78.44	81.58	79.09
Yoga	2	300	3000	426	84.30	84.14	81.83	**91.83**	83.90	87.00	91.77	87.86	83.47	82.80
Average accuracy					75.91	78.19	78.84	81.16	80.92	82.48	**84.71**	81.94	**84.64**	82.60
Average rank					8.16	6.90	6.55	6.10	5.57	**4.70**	**3.36**	5.42	**3.28**	4.95

[4] Cabello et al. "Fast and Accurate Time Series Classification Through Supervised Interval Search," 2020 IEEE International Conference on Data Mining (ICDM), 2020

Question

Are we there in the Android Malware classification context ?

Not quite[5]

$Eval$: $(\mathcal{D}, \ fem, \ gt, \ \mathcal{C}, \ \mathcal{S}) \rightarrow scores$

hard to vary

- No reference datasets
- Unavailability / deprecated
- Tight coupling between $\mathcal{D}$ and fem/gt

easy to vary

- Nice APIs (sklearn / tensorflow . . .)

[5] Daoudi et al. Lessons Learnt on Reproducibility in Machine Learning Based Android Malware Detection. Empir Software Eng 26, 74 (2021). https://doi.org/10.1007/s10664-021-09955-7

Walkthrough Andromak

A library to play with: "$Eval : (\mathcal{D}, fem, gt, \mathcal{C}, \mathcal{S}) \rightarrow scores$" in the Android context

- Goal 1: Allow experimenters to build easily various evaluations (write their own *Eval* functions)
- Goal 2: Knowledge base around classical techniques around Android malware classification

Design principles:

- Provide an implementation compatible for minimalistic infrastructure (single machine or cluster of machines with a minimal shared storage)
- Favor(abuse of?) functionnal aspects of the language

Knowledge base (datasets)

The library has built-in support for

Dataset	#sample	MW%	Ref
CICMalAnal17	4K	50%	10.1016/j.cose.2011.12.012
GM19	10K	50%	hal-02288116
Maldroid20	17K	75%*	10.1109/DASC-PICom-...
Malscan19	30K	50%	10.1109/ASE.2019.00023
Tess19	120K	10%	usenix/Pendlebury19

How ? `Store` abstraction

- Generic API for getting/storing APK, metadata, any transformation of APK/metadata
- Can be populated from an existing dataset (on disk or dl from androzoo)
- Can be backed in a FS (or a DB – theorically)

Knowledge base (fem)

Various *fem* (decoupled from the datasets):

fem	Réf
Drebin	ndss/ArpSHGR14
Mamadroid	10.1145/3313391
Malscan	10.1109/ASE.2019.00023
Revealdroid	10.1145/3162625
Hindroid	0.1145/3097983.3098026

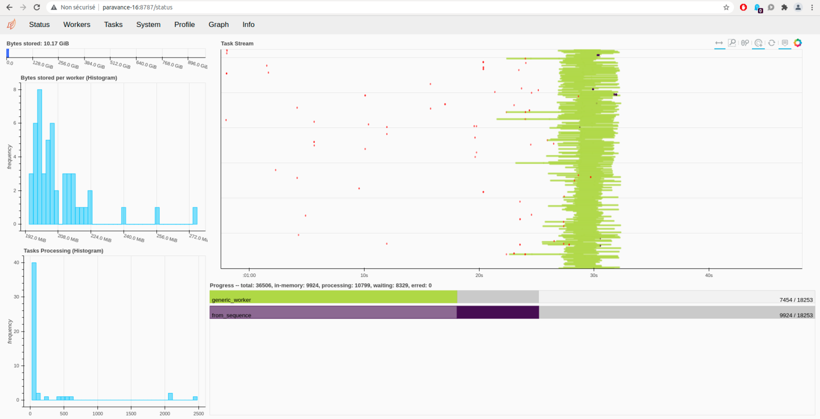
How ? Rely on the definition of

- *extract* : $APK \rightarrow Feature$
  - can be full python or wrap a docker execution
  - serializable $\Rightarrow$ implicit parallelism/distribution on a cluster
- *dump* : $Feature \rightarrow bytes$
  - compression / sparsification
  - storage optimization

Example: massively parallel fem execution

extract serializable $\Rightarrow$ implicit parallelism/distribution on a cluster

- screenshot of fem_{drebin} execution on 1M apk
- Andromak uses Dask[6] transparently (here on Grid5000 using 50 nodes / 250 workers)



[6] https://docs.dask.org/en/latest/

Example: dataset / fem decoupling

FEM		MalAnal17	GM19	MalDroid20	Malscan17	Tess19
reference	apps	2,126	10,000	17,187	30,715	129,728
	mw	426	5,000	13,150	15,430	12,737
	gw	1,700	5,000	4,037	15,285	116,991
common	apps	1,940	9,158	15,781	28,856	124,979
	mw	407	4,323	12,172	14,108	12,027
	gw	1,533	4,835	3,609	14,748	112,952
drebin	apps	2,116	10,000	17,186	30,715	129,728
	mw	426	5,000	13,149	15,430	12,737
	gw	1,690	5,000	4,037	15,285	116,991
hindroid	apps	2,111	9,930	16,661	30,690	129,658
	mw	425	4,941	12,630	15,430	12,706
	gw	1,686	4,989	4,031	15,260	116,952
malscan	apps	2,121	9,947	16,702	30,715	129,344
	mw	425	4,947	12,671	15,430	12,679
	gw	1,696	5,000	4,031	15,285	116,665
mamadroguard	apps	2,121	9,945	16,661	30,715	129,677
	mw	425	4,945	12,630	15,430	12,706
	gw	1,696	5,000	4,031	15,285	116,971
mamadroid	apps	1,978	9,288	15,881	29,333	125,660
	mw	408	4,430	12,211	14,527	12,145
	gw	1,570	4,858	3,670	14,806	113,515
revealdroid	apps	2,100	9,764	16,325	30,142	129,258
	mw	420	4,781	12,355	14,897	12,633
	gw	1,680	4,983	3,970	15,245	116,625

Table 3. Extraction effectiveness. number of applications (resp. malware, resp. goodware) that are initially in the various datasets (reference), or successfully extracted for each extraction method. Applications that are successfully extracted by all the extraction techniques are accounted in the common bloc.

Knowledge base (Domain specific ML)

- Some default $\mathcal{C}$ (for reproducibility purpose)
  - Related classifiers (best ones according to the related papers)
    - e.g. $\mathcal{C}_{drebin} = SVC$
  - Multilayer Percepton [7]
- Specific selection method / Hypothesis testing
  - Time aware splits (usenix/Pendlebury19)
  - Wilcoxon, Ranks ...

[7] Grosse et al. (2017) Adversarial Examples for Malware Detection. ESORICS 2017. https://doi.org/10.1007/978-3-319-66399-9_4

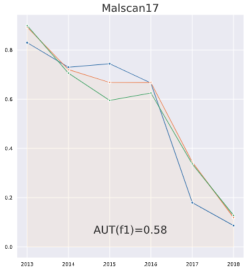
Example: comparison of 3 variants

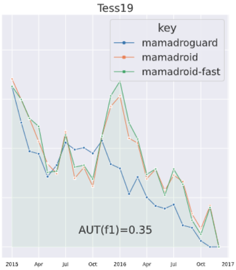
Eval: $(\mathcal{D},\ fem,\ gt,\ \mathcal{C},\ \mathcal{S}) \to scores$ with 3 different *fem* and 5 different $\mathcal{C}$

- $\mathcal{S}$: 5x2-fold CV (mean) f1-scores

	MalAnal17		GM19		MalDroid20		Malscan17		Tess19	
	coupled	free	coupled	free	coupled	free	coupled	free	coupled	free
mamadroguard	0.82	0.82	0.92	0.92	0.97	0.97	0.95	0.95	0.74	0.75
mamadroid	0.82	0.81	0.92	0.92	0.97	0.97	0.96	0.96	0.80	0.80
mamadroid-fast	0.82	0.80	0.93	0.93	0.96	0.97	0.96	0.96	0.81	0.81

- $\mathcal{S}$: time-aware model selection with AUT(f1) score





Conclusion

- Andromak
  - Let you introduce in your evaluation pipeline
    - variations
    - scale
  - Future:
    - research report / paper + release
    - sandboxed analysis, new datasets, visualisations ...
- Talk to SED people about your experimentations / problems
  - XUG / AskTheSED events