Cross-Regional Malware Detection via Model Distilling and Federated Learning

Marcus Botacin[1]

[1]Assistant Professor
Texas A&M University (TAMU), USA
botacin@tamu.edu
@MarcusBotacin

Agenda

1. Regional Malware
   - The Differences
2. FL and Distillation
   - New Architecture
   - Case Study
3. Conclusion
   - Generalization
   - Final Remarks

Agenda

1. Regional Malware
   - The Differences
2. FL and Distillation
   - New Architecture
   - Case Study
3. Conclusion
   - Generalization
   - Final Remarks

Previously


RESEARCH-ARTICLE

One Size Does Not Fit All: A Longitudinal Analysis of Brazilian Financial Malware

Authors: Marcus Botacin, Hojjat Aghakhani, Stefano Ortolani, Christopher Kruegel, Giovanni Vigna, Daniela Oliveira, Paulo Lício De Geus, André Grégio Authors Info & Affiliations

Publication: ACM Transactions on Privacy and Security • January 2021 • Article No.: 11 • https://doi.org/10.1145/3429741

Figure: **Link:** `https://dl.acm.org/doi/10.1145/3429741`

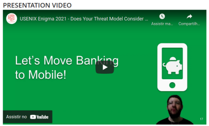

Figure: **Source:** `https://www.usenix.org/conference/enigma2021/presentation/botacin`


RESEARCH-ARTICLE

The Internet Banking [in]Security Spiral: Past, Present, and Future of Online Banking Protection Mechanisms based on a Brazilian case study

Authors: Marcus Botacin, Anatoli Kalysch, André Grégio Authors Info & Affiliations

Publication: ARES '19: Proceedings of the 14th International Conference on Availability, Reliability and Security • August 2019 • Article No.: 49 • Pages 1–10 • https://doi.org/10.1145/3339252.3340103

Figure: **Source:** `https://dl.acm.org/doi/10.1145/3339252.3340103`

Impact on AV Detection

ELSEVIER

Computers & Security

Volume 95, August 2020, 101859

We need to talk about antiviruses: challenges & pitfalls of AV evaluations

Marcus Botacin [a], Fabricio Ceschin [a], Paulo de Geus [b], André Grégio [a]

Figure: **Source:** https://www.sciencedirect.com/science/article/pii/S0167404820301310

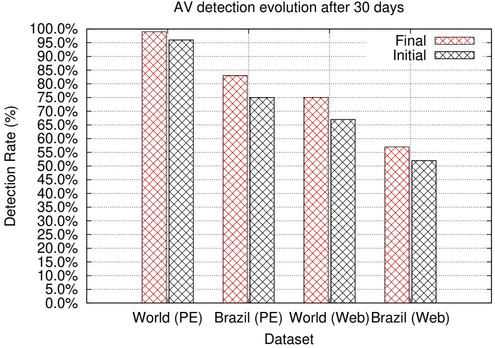

Figure: **Detection Rate:** BR samples are consistently less detected.

The Current Dataset

Table: **Dataset Differences.** Dynamic analysis events for the US, Brazil, and Japan datasets.

Behavior	**US**	**BR**	**JP**
Hosts file modification	0.04%	1.09%	0.92%
File creation	64%	24%	70%
File deletion	34%	12%	34%
File modification	63%	16%	46%
Browser modification	0%	1.03%	0.59%
Network traffic	53%	96%	52%

The Traditional Architecture

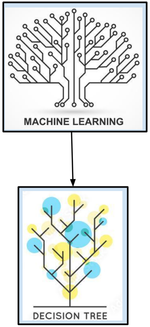

Figure: **Single Model Distillation.**

Is it enough to have global models?

US Features: Accuracy

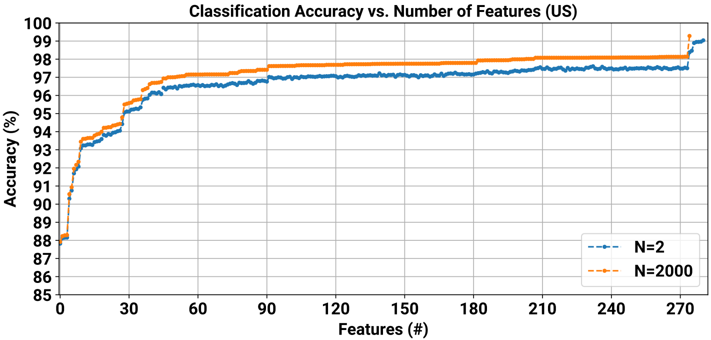

Figure: **Accuracy rates for the US dataset.** Accuracy variation with the increase of the feature set until reaching the 99% value.

US Features: Size

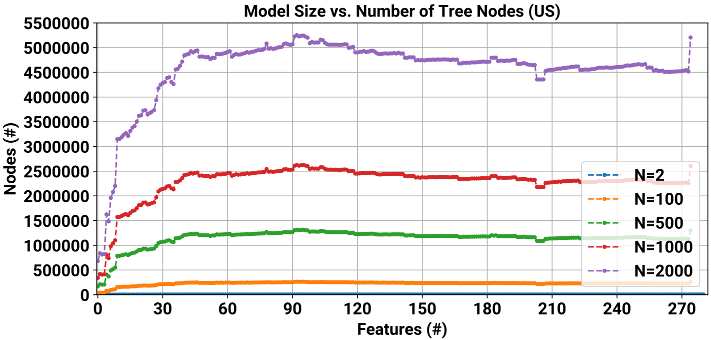

Figure: **Model size for the US dataset.** Number of nodes for an increased number of ensemble trees of increasing feature set sizes.

BR Features: Accuracy

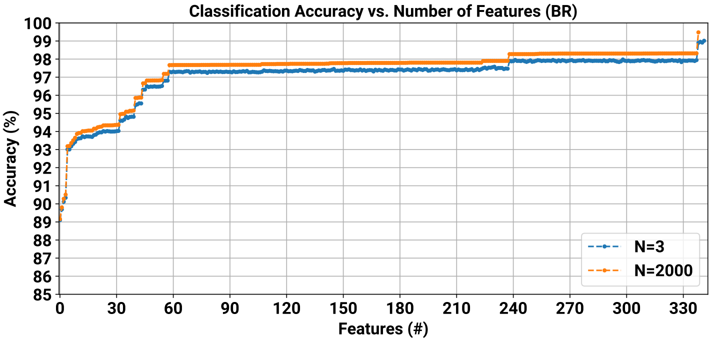

Figure: **Accuracy rates for the BR dataset.** Accuracy variation with the increase of the feature set until reaching the 99% value.

BR Features: Size

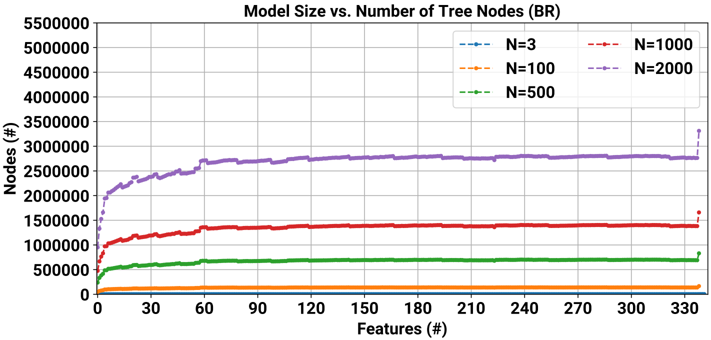

Figure: **Model size for the BR dataset.** Number of nodes for an increased number of ensemble trees of increasing feature set sizes.

JP Features: Accuracy

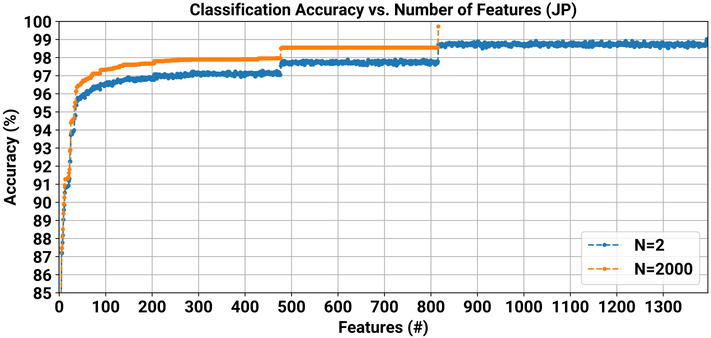

Figure: **Accuracy rates for the JP dataset.** Accuracy variation with the increase of the feature set until reaching the 99% value.

JP Features: Size

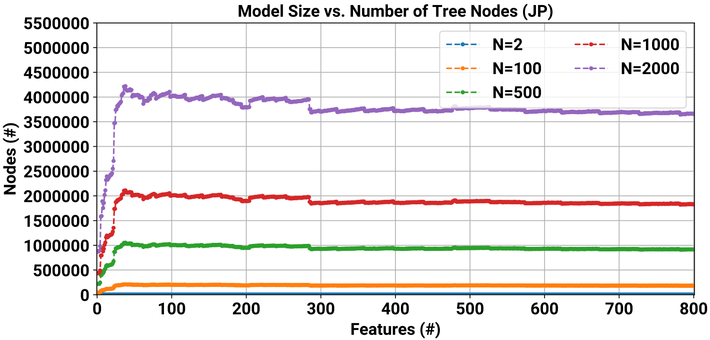

Figure: **Model size for the JP dataset.** Number of nodes for an increased number of ensemble trees of increasing feature set sizes.

Global Features: Accuracy

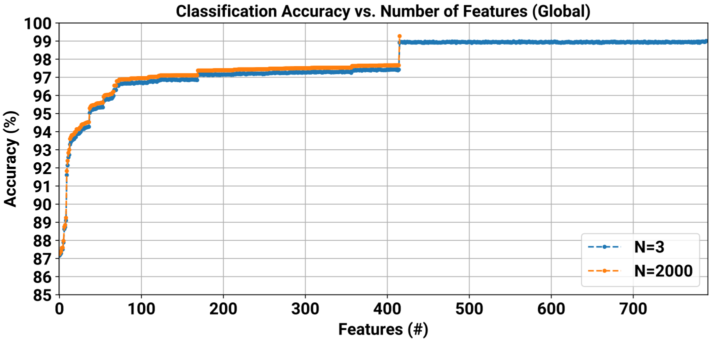

Figure: **Accuracy rates for the combined dataset.** Accuracy variation with the increase of the feature set until reaching the 99% value.

Global Features: Size

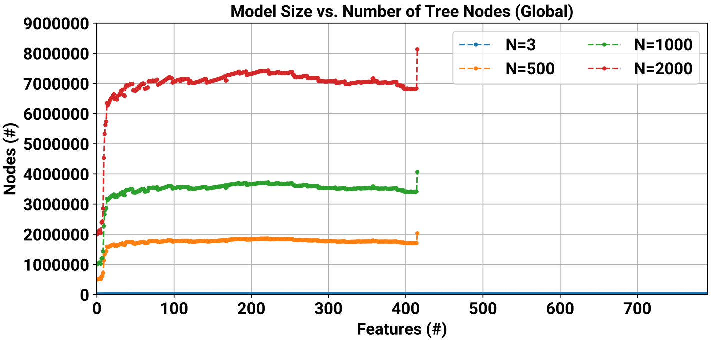

Figure: **Model size for the combined dataset.** Number of nodes for an increased number of ensemble trees of increasing feature set sizes.

Overview: Accuracy

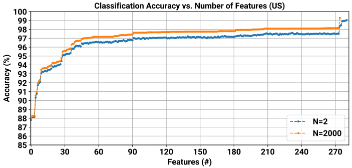

Figure: **Accuracy rates for the US dataset.**

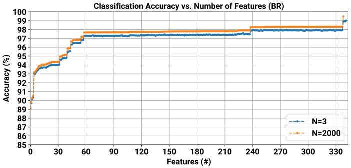

Figure: **Accuracy rates for the BR dataset.**

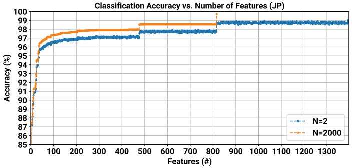

Figure: **Accuracy rates for the JP dataset.**

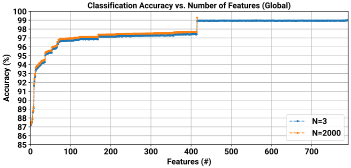

Figure: **Accuracy for the global dataset.**

Replicated Architecture

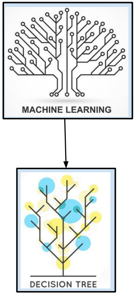

Figure: **Single Model Distillation.**

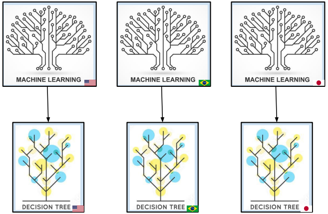

Figure: **Multiple Regional Model Distillation.**

Agenda

1. Regional Malware
   - The Differences
2. FL and Distillation
   - New Architecture
   - Case Study
3. Conclusion
   - Generalization
   - Final Remarks

Does a global model help?

US predicting the world

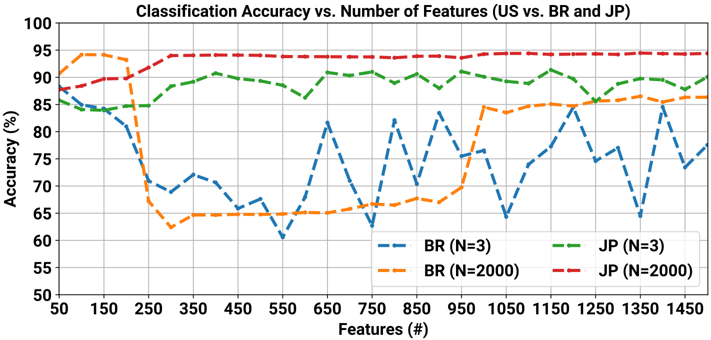

Figure: **Cross-dataset accuracy rate.** Trained US model classifying the samples from the other datasets.

BR predicting the world

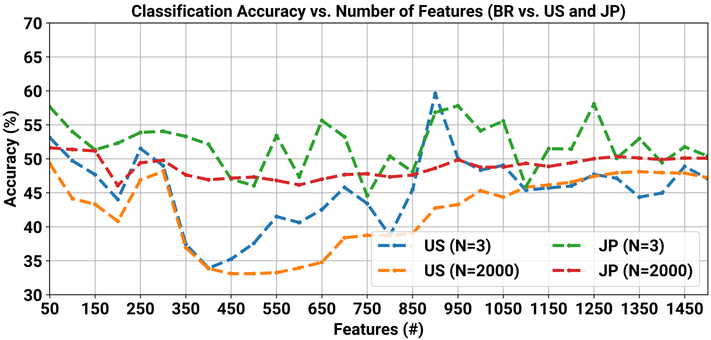

Figure: **Cross-dataset accuracy rate.** Trained BR model classifying the samples from the other datasets.

JP predicting the world

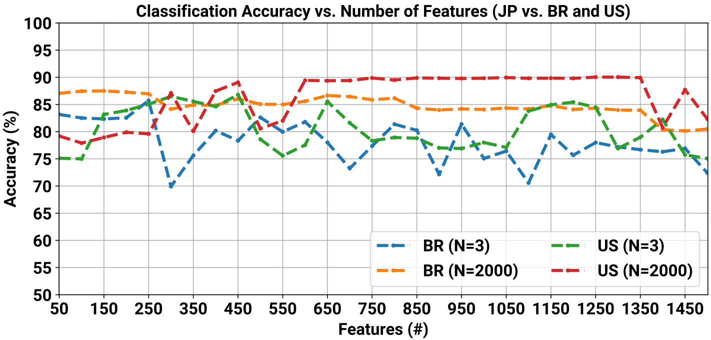

Figure: **Cross-dataset accuracy rate.** Trained JP model classifying the samples from the other datasets.

Three-Layer Architecture

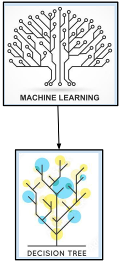

Figure: **Single Model Distillation.**

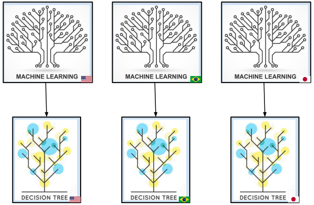

Figure: **Multiple Regional Model Distillation.**

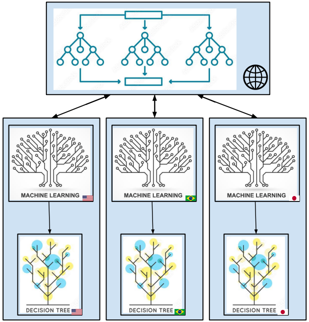

Figure: **Regional Model Distillation from Global.**

How to best build local-to-global models?

Enriching the Global Model

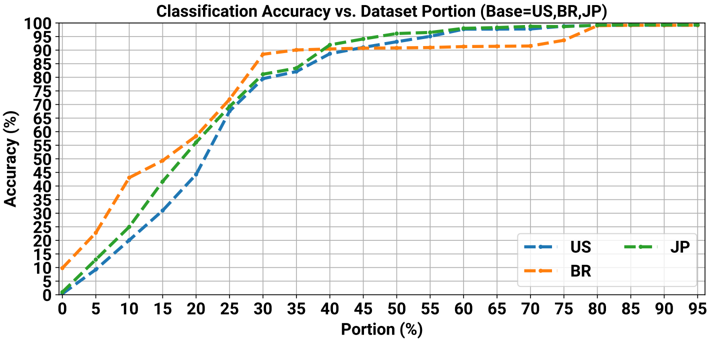

Figure: **Building a global model.** Accuracy rate for building a global model from different portions of the source datasets.

Enriching the US Model: Random Sampling

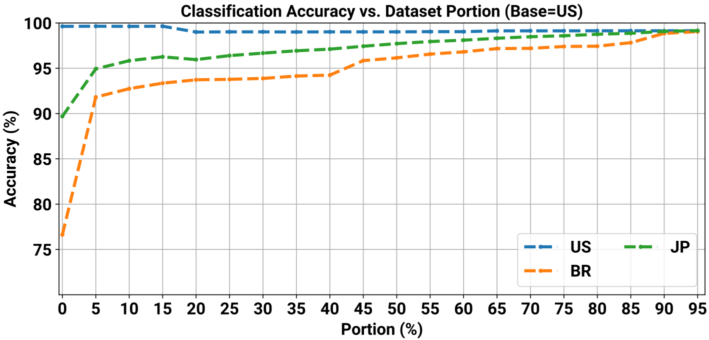

Figure: **Extending the existing US model.** Accuracy rates on the different datasets for different portions of the source datasets using random sample selection.

Enriching the BR model¿ Random Sampling

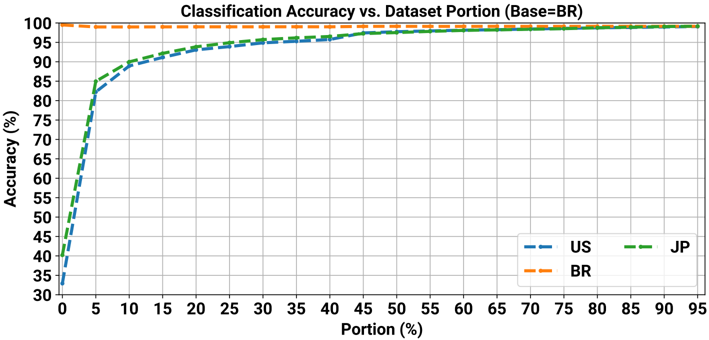

Figure: **Extending the existing BR model.** Accuracy rates on the different datasets for different portions of the source datasets using confidence-based sample selection.

Enriching the JP model: Random Sampling

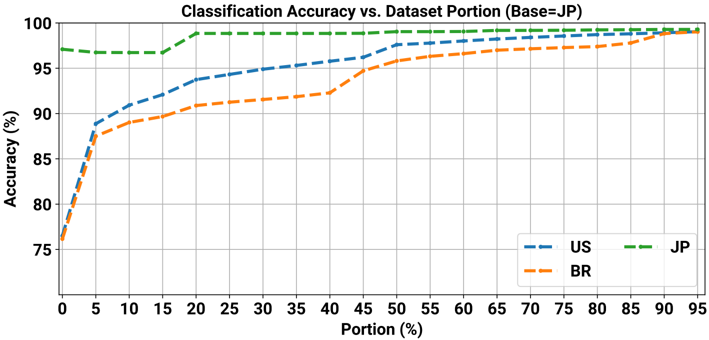

Figure: **Extending the existing JP model.** Accuracy rates on the different datasets for different portions of the source datasets using random sample selection.

Enriching the US model: Confidence-based Sampling

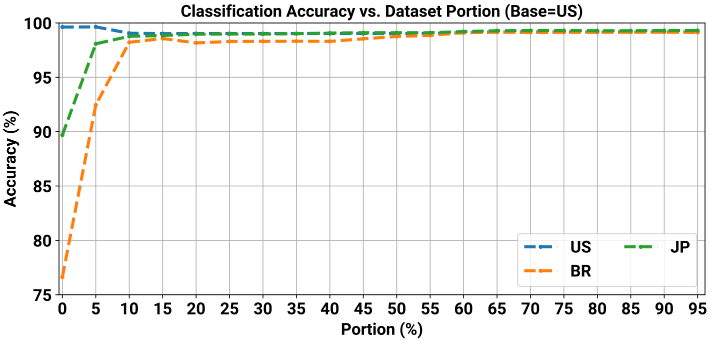

Figure: **Extending the existing US model.** Accuracy rates on the different datasets for different portions of the source datasets using confidence-based sample selection.

Enriching the BR model: Confidence-based Sampling

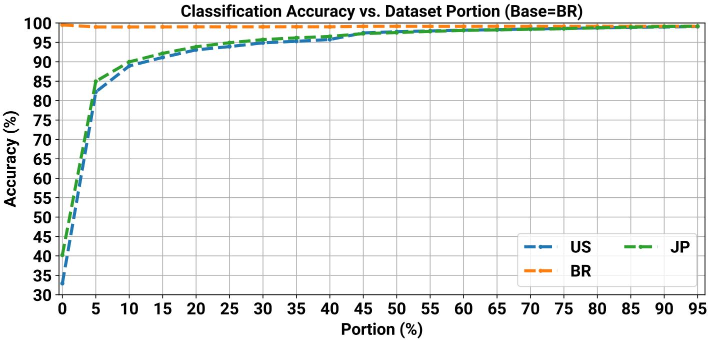

Figure: **Extending the existing BR model.** Accuracy rates on the different datasets for different portions of the source datasets using confidence-based sample selection.

Enriching the JP model: Confidence-based Sampling

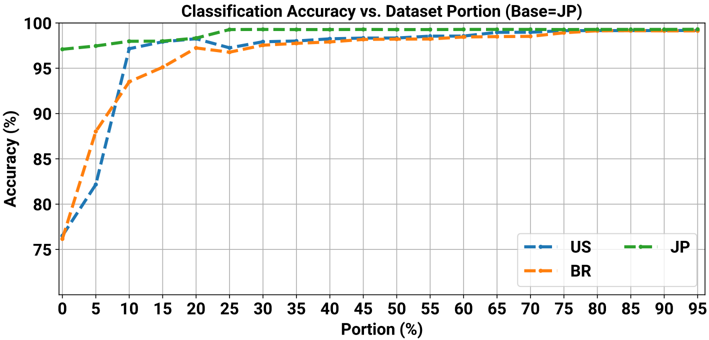

Figure: **Extending the existing JP model.** Accuracy rates on the different datasets for different portions of the source datasets using confidence-based sample selection.

Are real models trained from scratch?

Self-Distillation

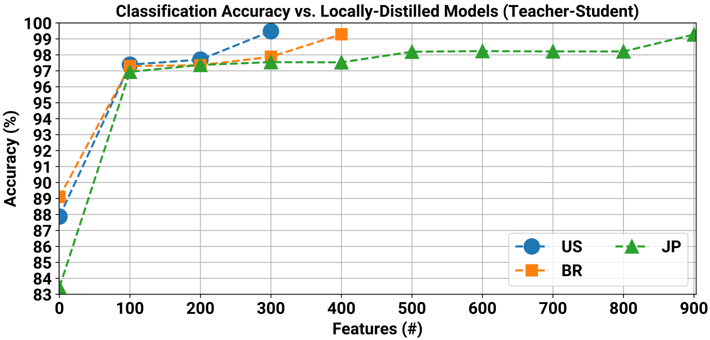

Figure: **Self-Model Distilling.** Number of features required to achieve the maximum accuracy rate for the different datasets.

Global-to-Local Distillation

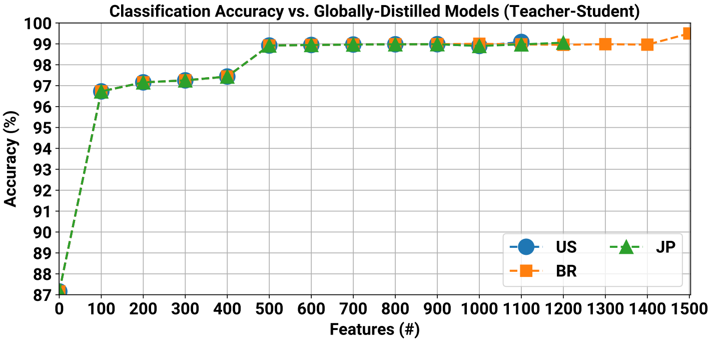

Figure: **Global to Local Model Distilling.** Number of features required to achieve the maximum accuracy rate for the different datasets.

Heterogeneous Distillation

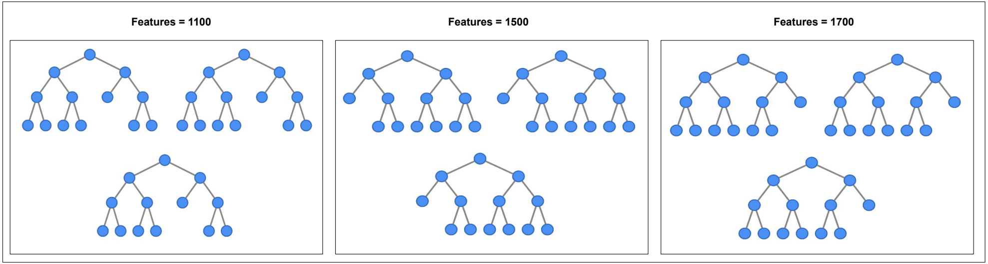

Figure: **RF's ensemble of different features set sizes.**

Heterogeneous Distillation

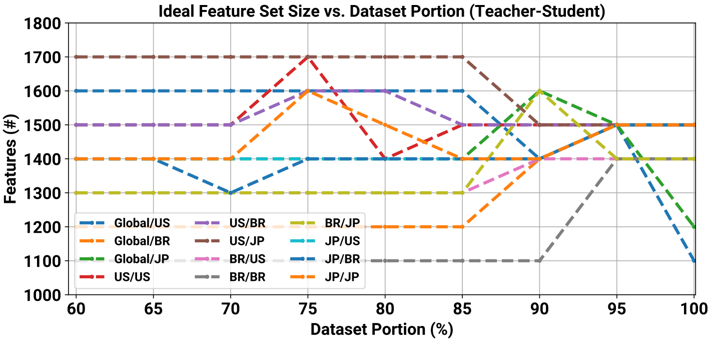

Figure: **Global to Local Model Distilling.** Variation on the number of features required to achieve the maximum accuracy rate for different portions of the source datasets.

What is the real impact of ML on AVs?

ML-derived YARA Rules

```
import "pe"

rule rule_from_ml_0 {
 condition:
  pe.imports(/(.).dll/i, /closehandle/i)
  and
  pe.characteristics & pe.EXECUTABLE_IMAGE
  and
  pe.exports(/dllunregisterserver/i)
}
```

Code 1: Yara rule generated from the ML model.

Matching Time

Table: **Matching performance.** Wall time (s) for matching Yara rules derived from ML models of different feature sets sizes against a real, infected filesystem.

Features	1100	1200	1300	
Time	13m57s	14m00s (+0.3%)	14m05s (+1%)	
Features	1400	1500	1600	1700
Time	14m50s (+6%)	15m57s (+14%)	17m58 (+29%)	19m33s (+40%)

Explaining Rule's Performance

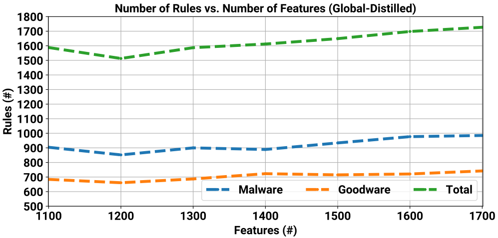

Figure: **Number of rules vs. feature size.** The number of generated rules moderately increases with the number of features.

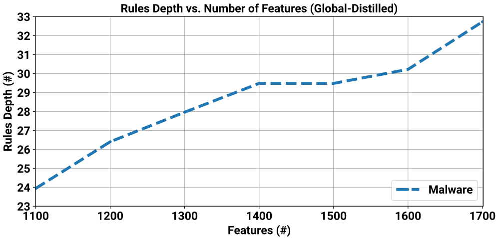

Figure: **Rules depth vs. feature size.** The average depth of the rules increases with the number of features.

Agenda

1 Regional Malware
- The Differences

2 FL and Distillation
- New Architecture
- Case Study

3 Conclusion
- Generalization
- Final Remarks

The Original Scenario

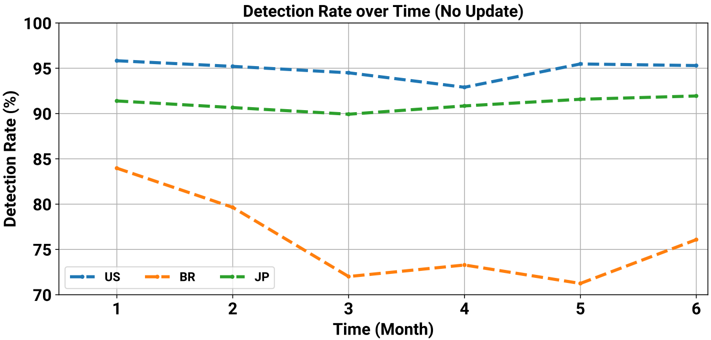

Figure: **Detection rate as a time-series for the individual static models.** Previously trained classifiers attempt to detect new threats. Performance degradation due to concept drift is observed.

Drift Detection Scenario

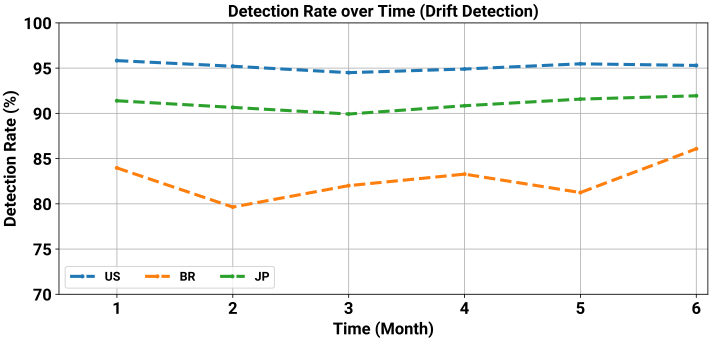

Figure: **Detection rate as a time-series for the individual, drift-aware models.** The retraining of models when concept drift is detected takes the detection rate back to its original level.

Drift Detection + Federated Learning Scenario

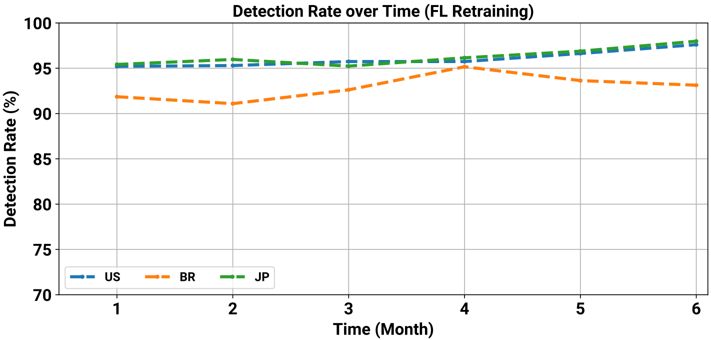

Figure: **Detection rate as a time-series for the globally-distilled models.** The use of data from a global model not only mitigated the drift effects but also increased the detection rate for all datasets.

Overview

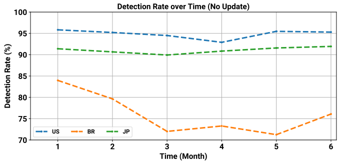

Figure: **Original**

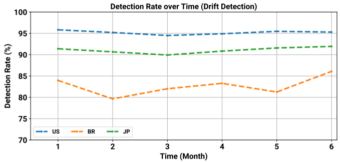

Figure: **Drift Detection**

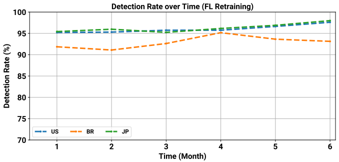

Figure: **Drift Detection + FL**

Agenda

1. Regional Malware
   - The Differences
2. FL and Distillation
   - New Architecture
   - Case Study
3. Conclusion
   - Generalization
   - Final Remarks

Extending to other feature selectors

Table: **Feature Selection Method.** Ideal feature set size for the multiple regional malware datasets.

	US	**BR**	**JP**
F-Score	290	340	800
Chi2	292	342	803
Mutual Info	294	345	812

Extending to other classifiers

Table: **Classifier Influence** on the detection of different regional malware datasets. Feature set sizes.

	95%			**99%**		
	US	**BR**	**JP**	**US**	**BR**	**JP**
RF	35	40	45	290	340	800
SGD	35	40	45	292	342	805
AdaBoost	35	40	45	292	342	805
SVM	36	41	46	295	345	813

Extending to other distillation techniques

Table: **Distillation Technique Influence** on the detection of different regional malware datasets. Feature set sizes.

	US	**BR**	**JP**
TS	300 (+3%)	400 (+17%)	900 (+12.5%)
FMF	299 (+3%)	402 (+18%)	902 (+12.5%)

Agenda

1. Regional Malware
   - The Differences
2. FL and Distillation
   - New Architecture
   - Case Study
3. Conclusion
   - Generalization
   - Final Remarks

Thanks!

Questions? Comments?

botacin@tamu.edu
@MarcusBotacin