When a Tree Falls: Using Diversity in Ensemble Classifiers to Identify Evasion in Malware Detectors

Charles Smutz
Angelos Stavrou
George Mason University

Motivation

- Machine learning used ubiquitously to improve information security
  - SPAM
  - Malware: PEs, PDFs, Android applications, etc
  - Account misuse, fraud
- Many studies have shown that machine learning based systems are vulnerable to evasion attacks
  - Serious doubt about reliability of machine learning in adversarial environments

Problem

- If new observations differ greatly from training set, classifier is forced to extrapolate
- Classifiers often rely on features that can be mimicked
  - Features coincidental to malware
  - Many types of malware/misuse
  - Feature extractor abuse
- Proactively addressing all possible mimicry approaches not feasible

Approach

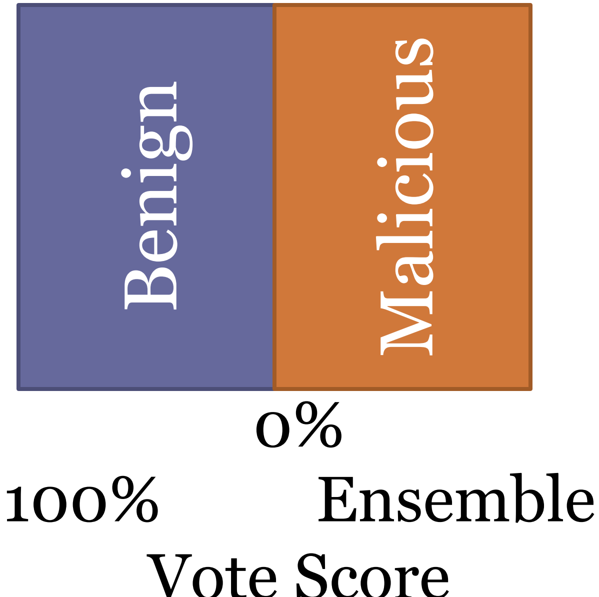
- Detect when classifiers provide poor predictions
  - Including evasion attacks
- Relies on diversity in ensemble classifiers

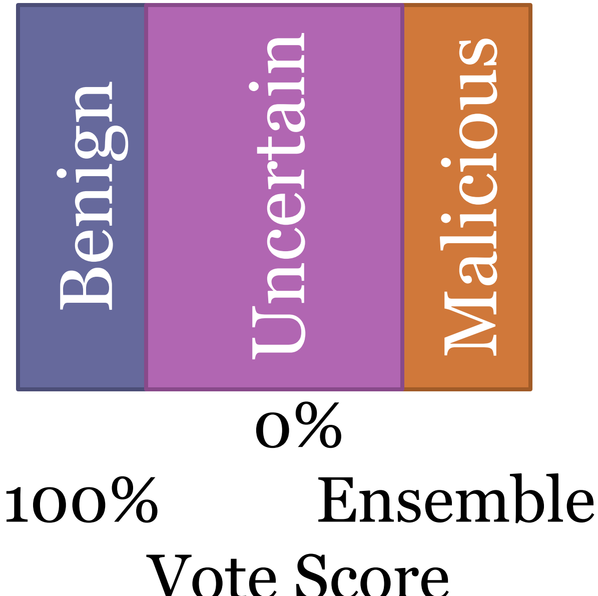
Background

- PDFrate: PDF malware detector using structural and metadata features, Random Forest classifier
  - pdfrate.com: scan with multiple classifiers
    - Contagio: 10k sample publicly known set
    - University: 100k sample training set
- PDFrate evasion attacks
  - Mimicus: Comprehensive mimicry of features (F), classifier (C), and training set (T) using replica
  - Reverse Mimicry: Scenarios that hide malicious footprint: PDFembed, EXEembed, JSinject
- Drebin: Andriod application malware detector using values from manifest and disassembly

Mutual Agreement Analysis

- When ensemble voting disagrees, prediction is unreliable
- High level of agreement on most observations

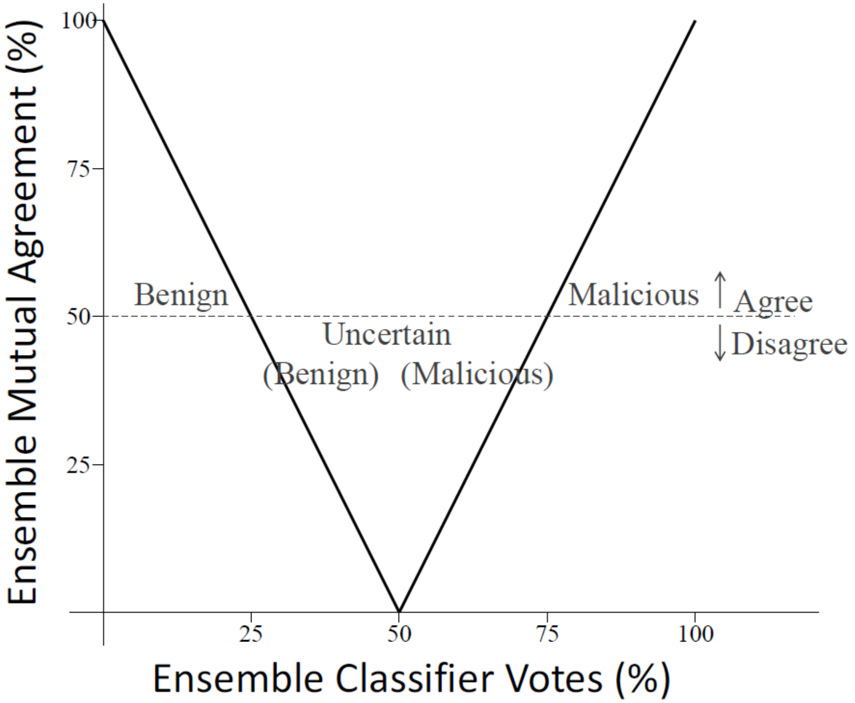




Mutual Agreement

$$A = | v - 0.5 | * 2$$

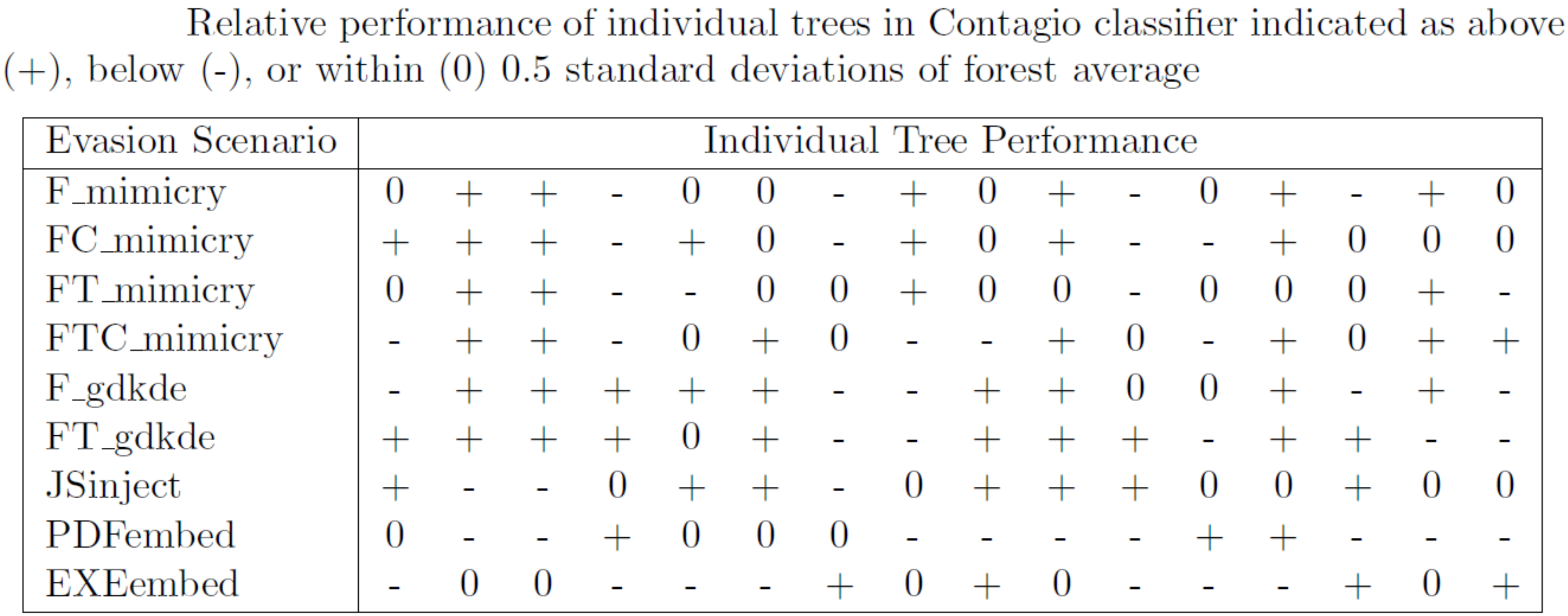
v: ensemble vote ratio
A: Mutual Agreement



- Ratio between 0 and 1 (or 0% and 100%)
- Proxy for Confidence on individual observations
- Threshold is tunable, 50% used in evaluations

Mutual Agreement

- Disagreement caused by extrapolation noise



Relative performance of individual trees in Contagio classifier indicated as above (+), below (-), or within (0) 0.5 standard deviations of forest average

Evasion Scenario	Individual Tree Performance															
F_mimicry	0	+	+	-	0	0	-	+	0	+	-	0	+	-	+	0
FC_mimicry	+	+	+	-	+	0	-	+	0	+	-	-	+	0	0	0
FT_mimicry	0	+	+	-	-	0	0	+	0	0	-	0	0	0	+	-
FTC_mimicry	-	+	+	-	0	+	0	-	-	+	0	-	+	0	+	+
F_gdkde	-	+	+	+	+	+	-	-	+	+	0	0	+	-	+	-
FT_gdkde	+	+	+	+	0	+	-	-	+	+	+	-	+	+	-	-
JSinject	+	-	-	0	+	+	-	0	+	+	+	0	0	+	0	0
PDFembed	0	-	-	+	0	0	0	-	-	-	-	+	+	-	-	-
EXEembed	-	0	0	-	-	-	+	0	+	0	-	-	-	+	0	+

Mutual Agreement Operation

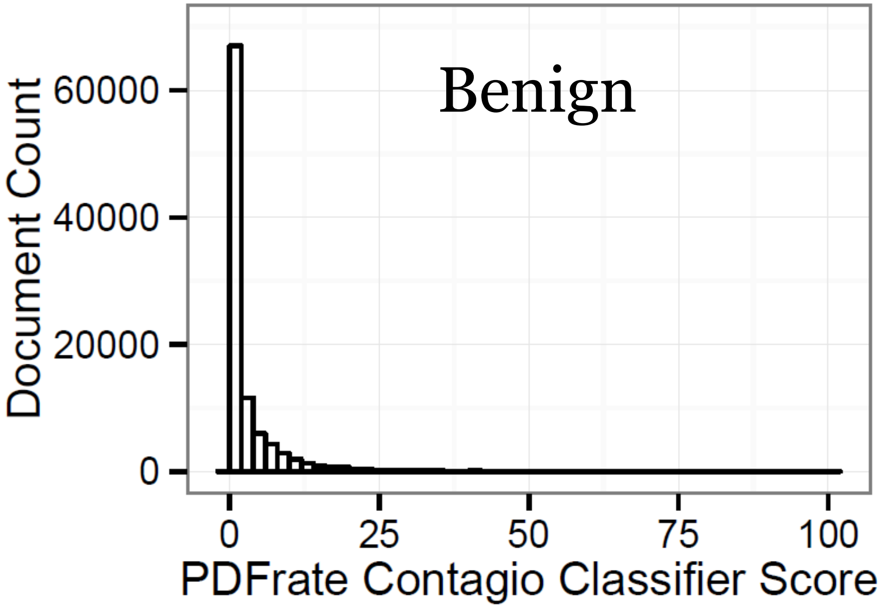
- Mutual agreement trivially calculated at classification time
- Identifies unreliable predictions
  - Identifies detector subversion as it occurs
- Uncertain observations require distinct, potentially more expensive detection mechanism
- Separates weak mimicry from strong mimicry attacks

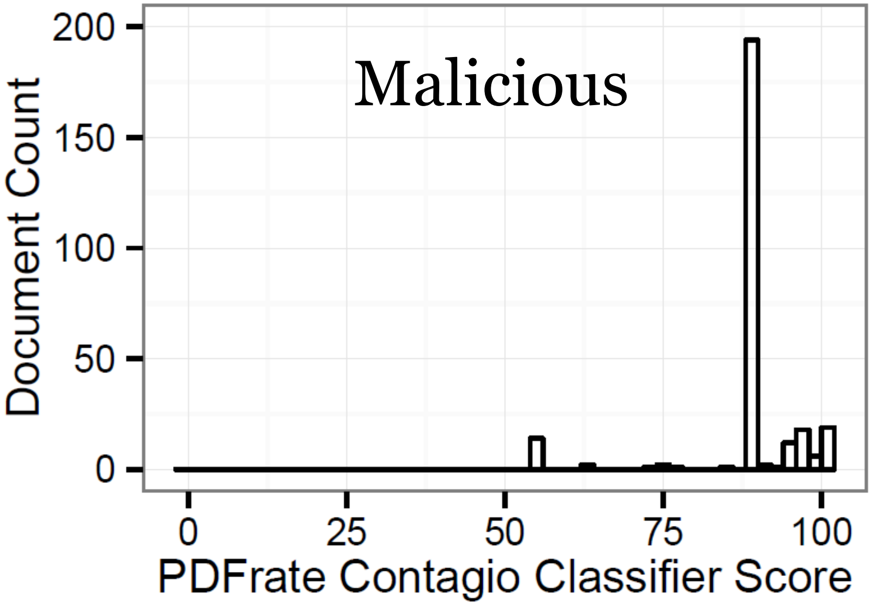
Evaluation

- Degree to which mutual agreement analysis allows separation of correct predictions from misclassification, including mimicry attacks
  - PDFrate Operational Data
  - PDFrate Evasion: Mimicus and Reverse Mimicry
  - Drebin Novel Android Malware Families
- Gradient Descent Attacks and Evasion Resistant Support Vector Machine Ensemble

Operational Data

- 100,000 PDFs (243 malicious) scanned by network sensor (web and email)





Operational Data

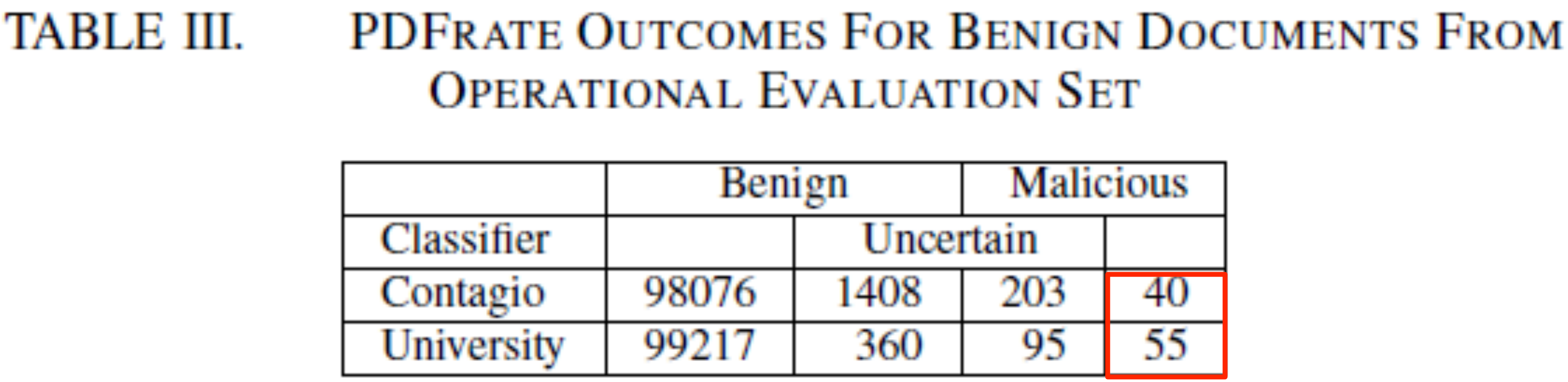


TABLE III. PDFRATE OUTCOMES FOR BENIGN DOCUMENTS FROM OPERATIONAL EVALUATION SET

	Benign		Malicious	
Classifier		Uncertain		
Contagio	98076	1408	203	40
University	99217	360	95	55

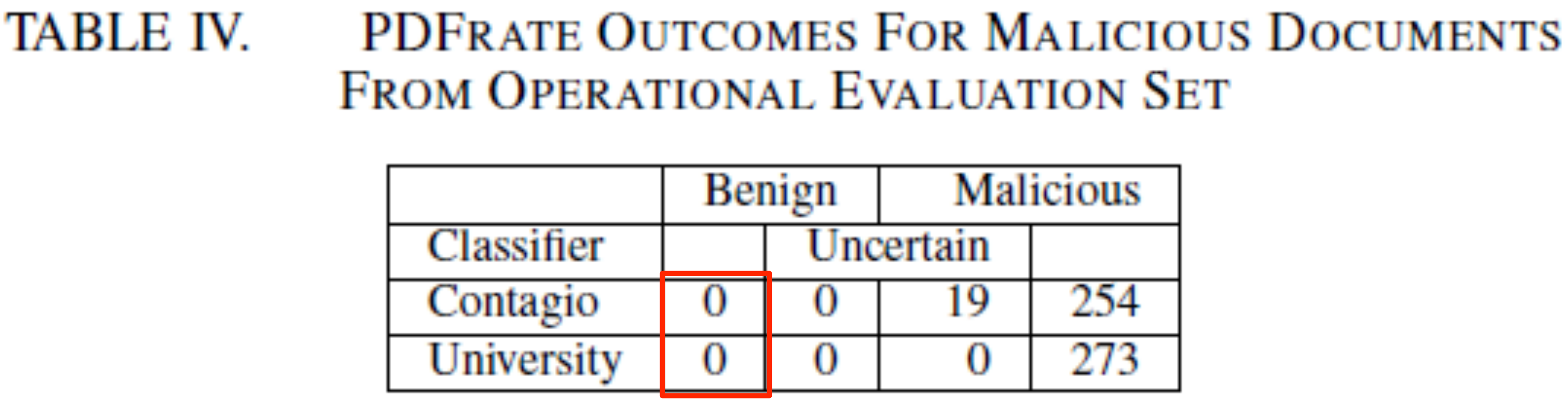


TABLE IV. PDFRATE OUTCOMES FOR MALICIOUS DOCUMENTS FROM OPERATIONAL EVALUATION SET

	Benign		Malicious	
Classifier		Uncertain		
Contagio	0	0	19	254
University	0	0	0	273

Operational Localization (Retraining)

- Update training set with portions of 10,000 documents taken from same operational source

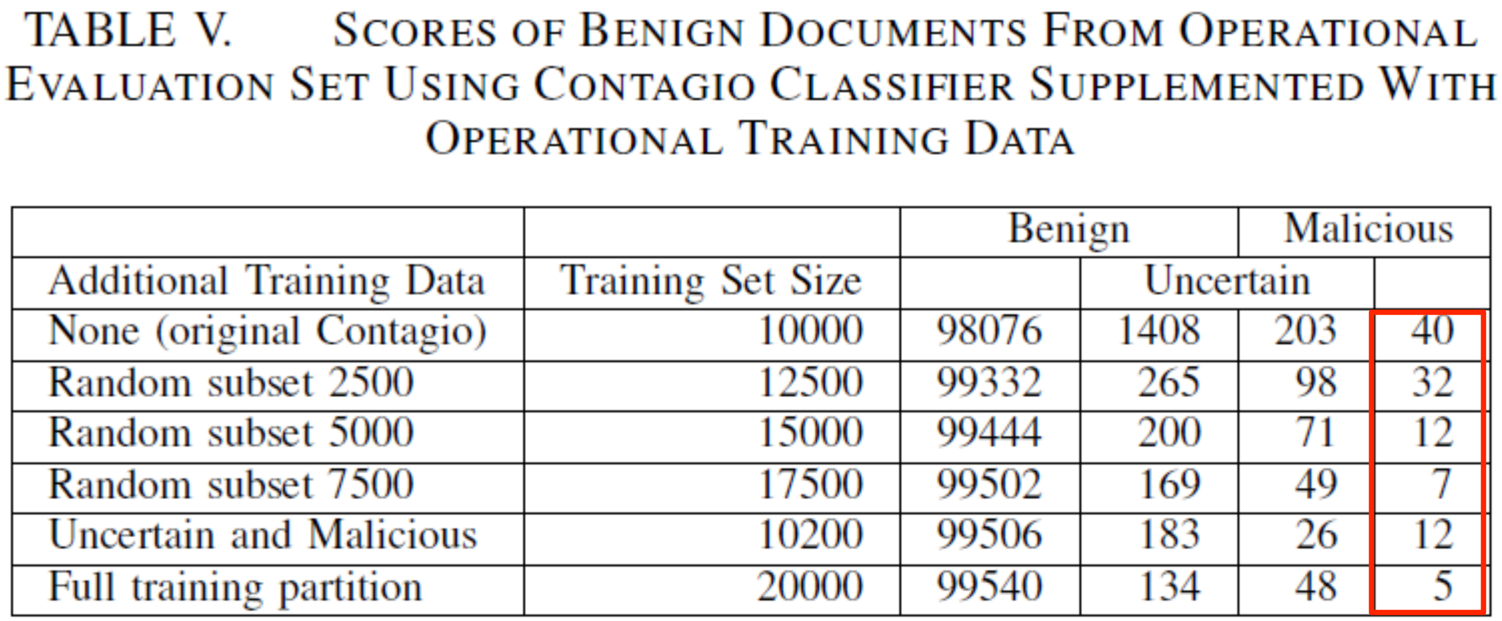
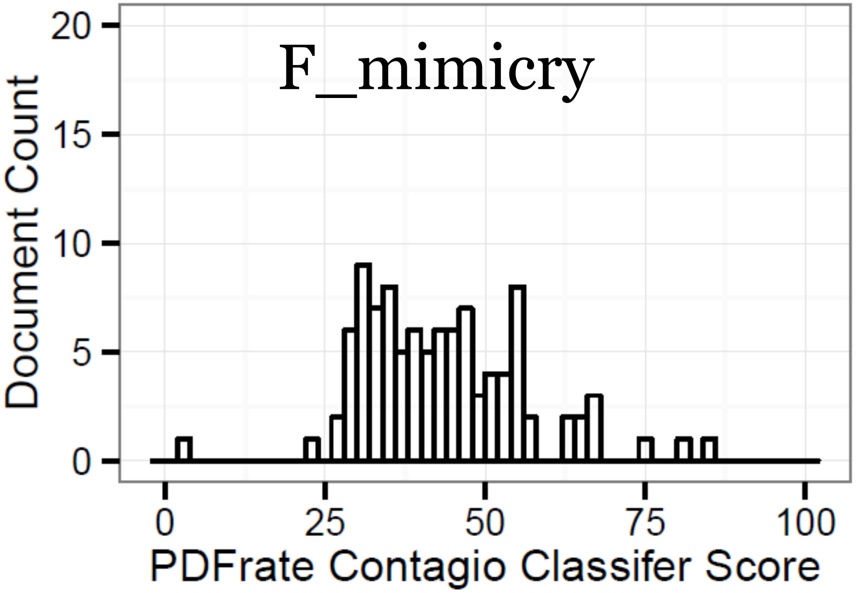
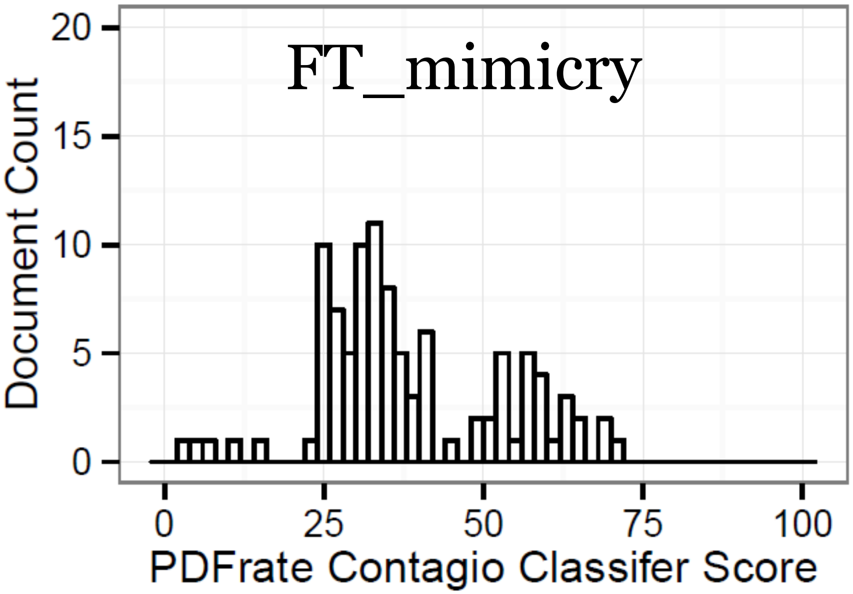


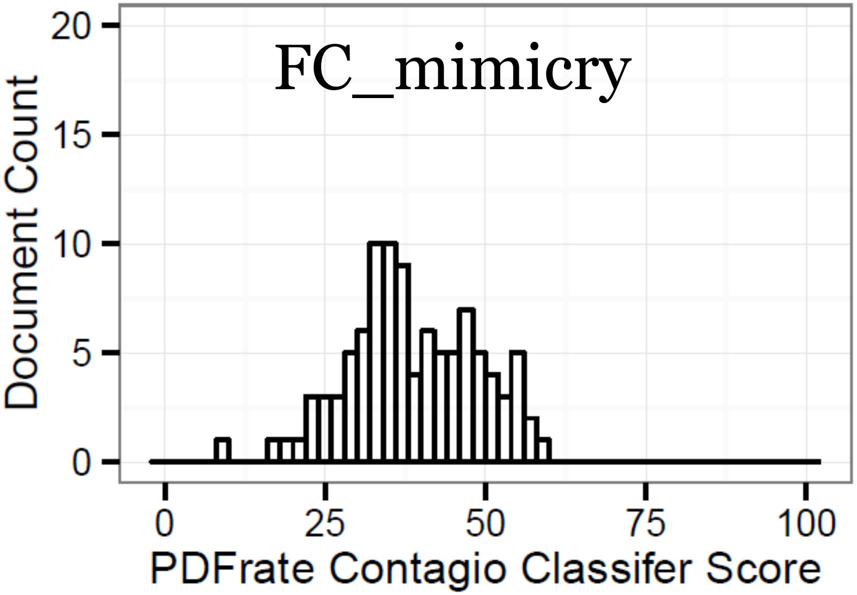
TABLE V. SCORES OF BENIGN DOCUMENTS FROM OPERATIONAL EVALUATION SET USING CONTAGIO CLASSIFIER SUPPLEMENTED WITH OPERATIONAL TRAINING DATA

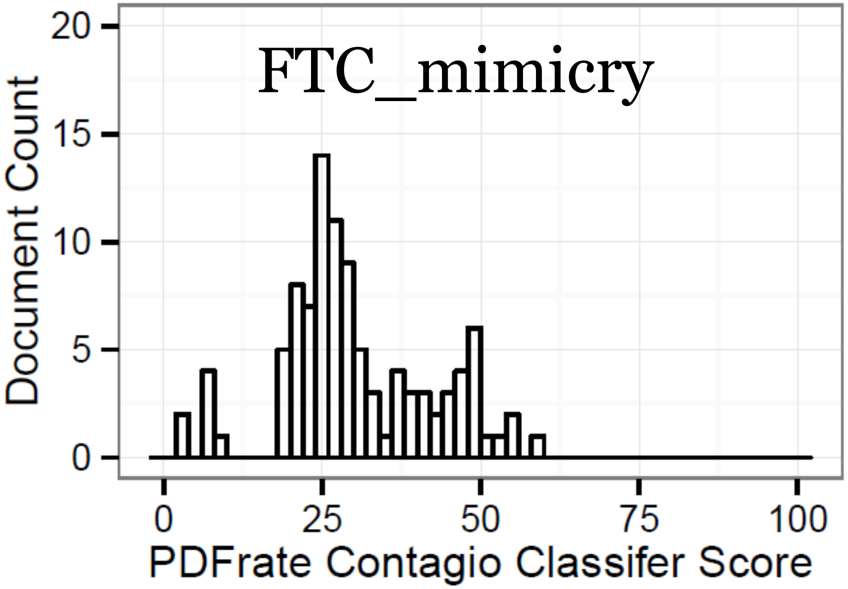
		Benign			Malicious
Additional Training Data	Training Set Size		Uncertain		
None (original Contagio)	10000	98076	1408	203	40
Random subset 2500	12500	99332	265	98	32
Random subset 5000	15000	99444	200	71	12
Random subset 7500	17500	99502	169	49	7
Uncertain and Malicious	10200	99506	183	26	12
Full training partition	20000	99540	134	48	5

Mimicus Results


F_mimicry
Document Count
20
15
10
5
0
0
25
50
75
100
PDFrate Contagio Classifer Score


FT_mimicry
Document Count
20
15
10
5
0
0
25
50
75
100
PDFrate Contagio Classifer Score


FC_mimicry
Document Count
20
15
10
5
0
0
25
50
75
100
PDFrate Contagio Classifer Score


FTC_mimicry
Document Count
20
15
10
5
0
0
25
50
75
100
PDFrate Contagio Classifer Score

Mimicus Results

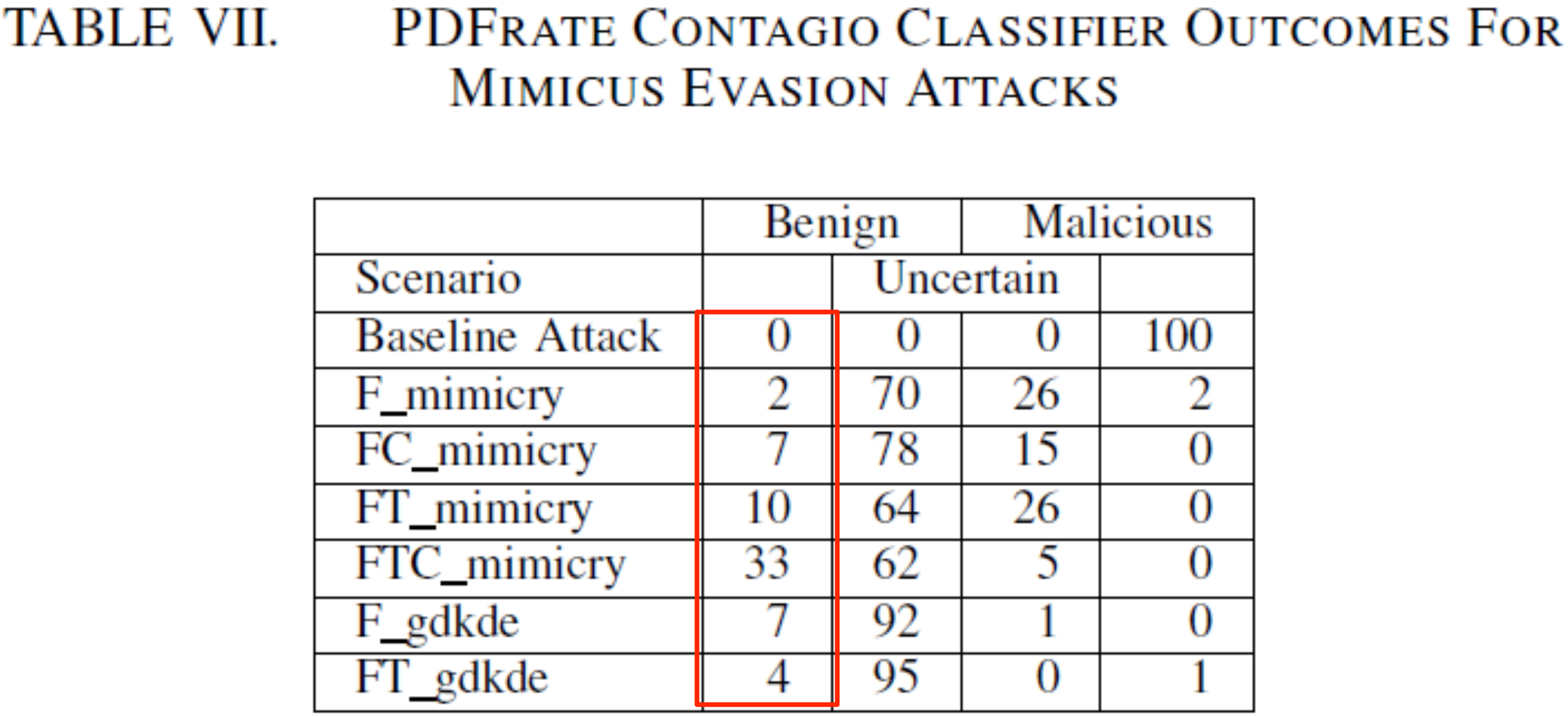
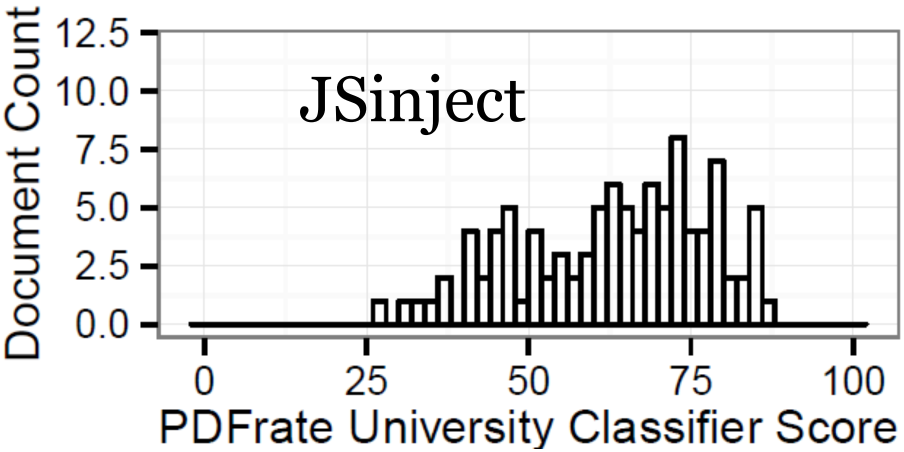
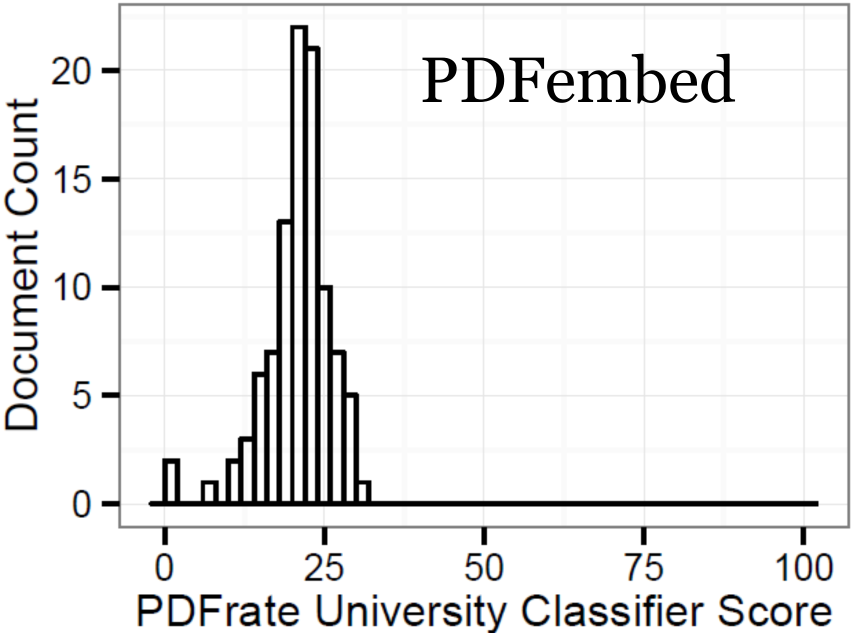


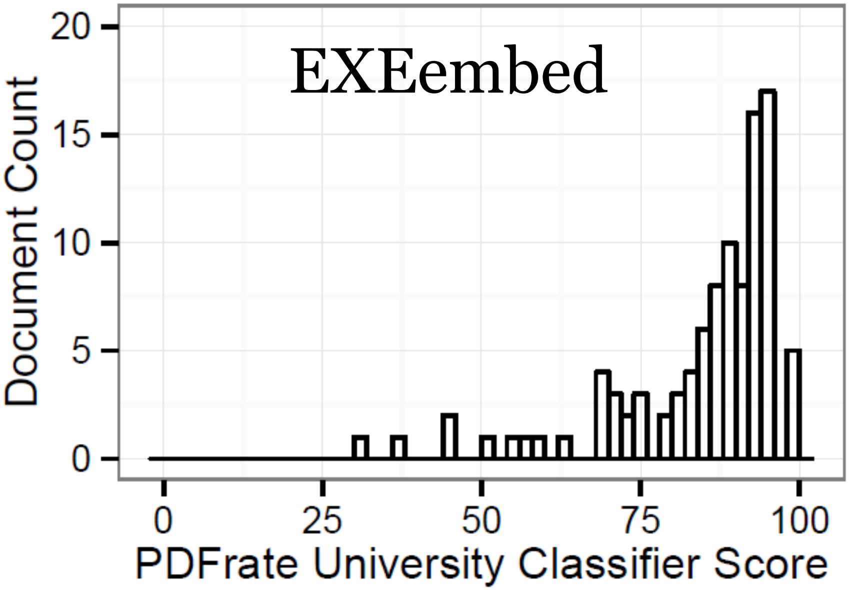
TABLE VII. PDFRATE CONTAGIO CLASSIFIER OUTCOMES FOR MIMICUS EVASION ATTACKS

	Benign		Malicious	
Scenario		Uncertain		
Baseline Attack	0	0	0	100
F_mimicry	2	70	26	2
FC_mimicry	7	78	15	0
FT_mimicry	10	64	26	0
FTC_mimicry	33	62	5	0
F_gdkde	7	92	1	0
FT_gdkde	4	95	0	1

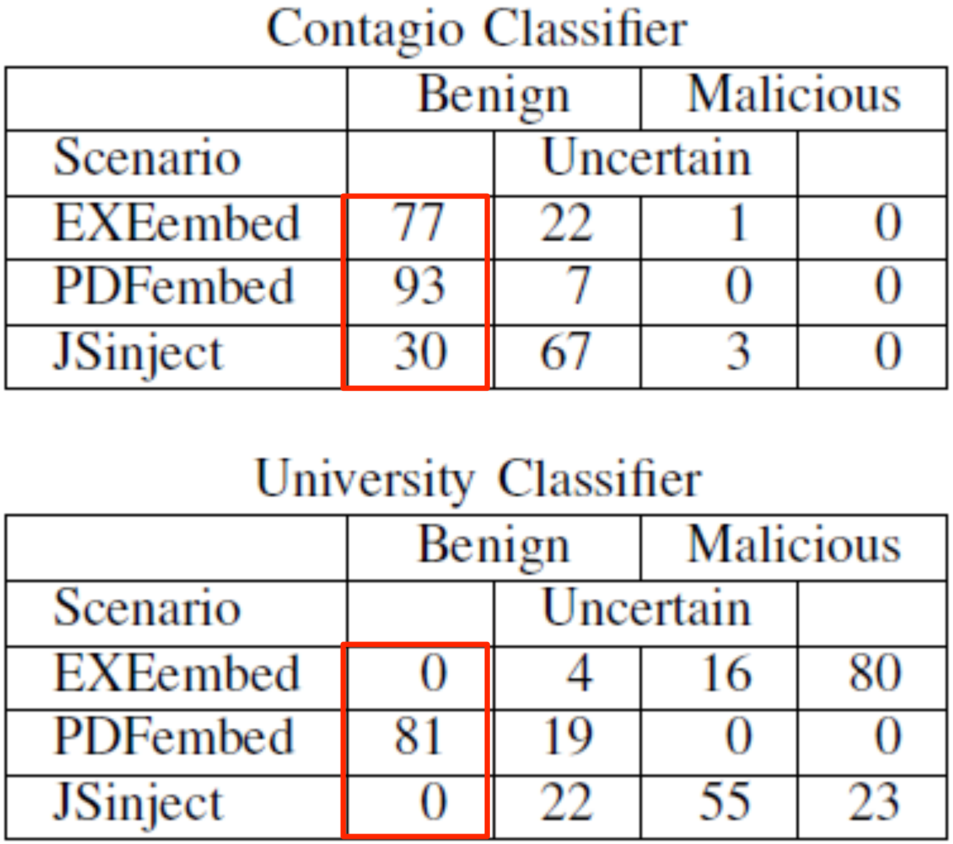
Reverse Mimicry Results


JSinject
Document Count
12.5
10.0
7.5
5.0
2.5
0.0
0
25
50
75
100
PDFrate University Classifier Score


PDFembed
Document Count
20
15
10
5
0
0
25
50
75
100
PDFrate University Classifier Score


EXEembed
Document Count
20
15
10
5
0
0
25
50
75
100
PDFrate University Classifier Score

Reverse Mimicry Results



Contagio Classifier

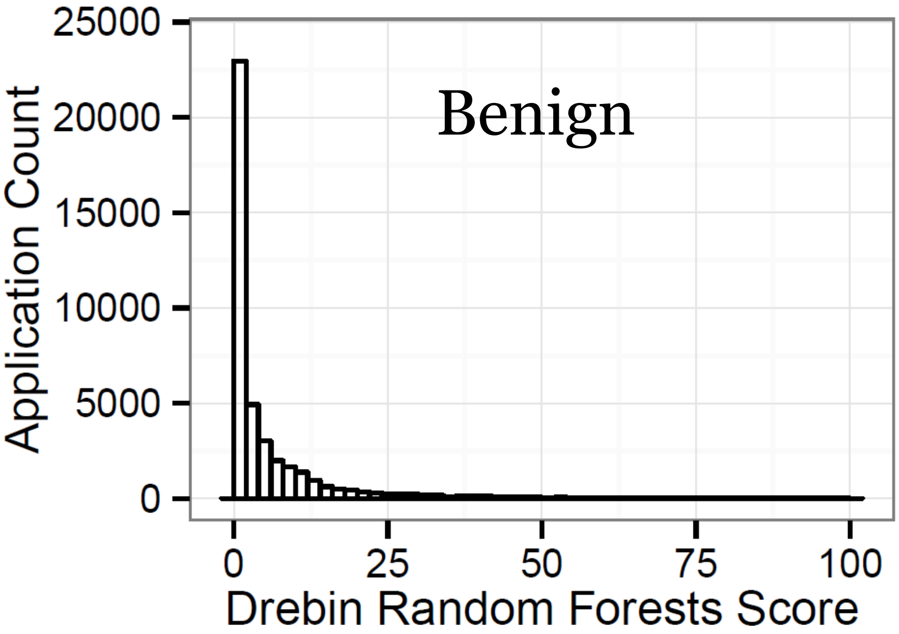
	Benign		Malicious	
Scenario		Uncertain		
EXEembed	77	22	1	0
PDFembed	93	7	0	0
JSinject	30	67	3	0

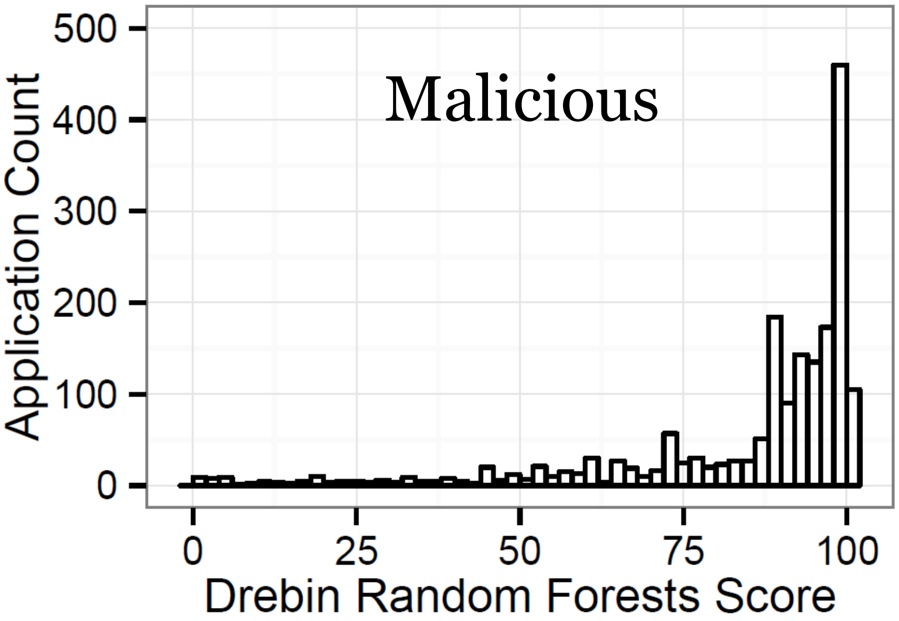
University Classifier

	Benign		Malicious	
Scenario		Uncertain		
EXEembed	0	4	16	80
PDFembed	81	19	0	0
JSinject	0	22	55	23

Drebin Android Malware Detector

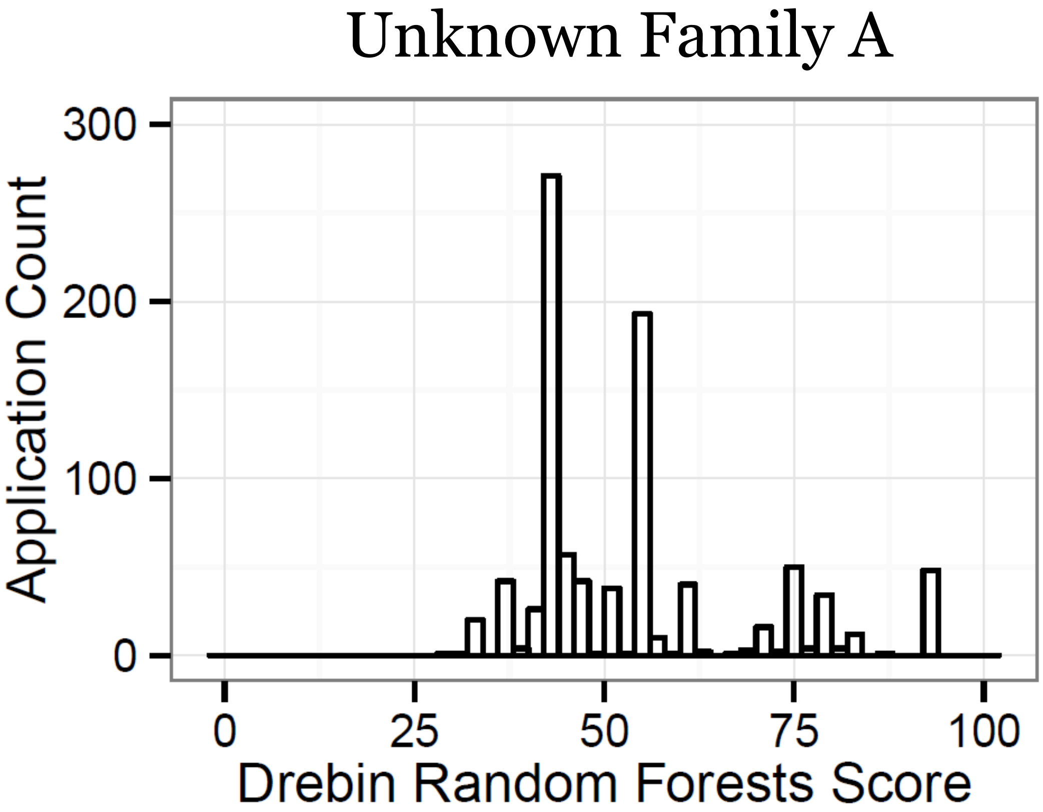
- Modified from original linear SVM to use Random Forests



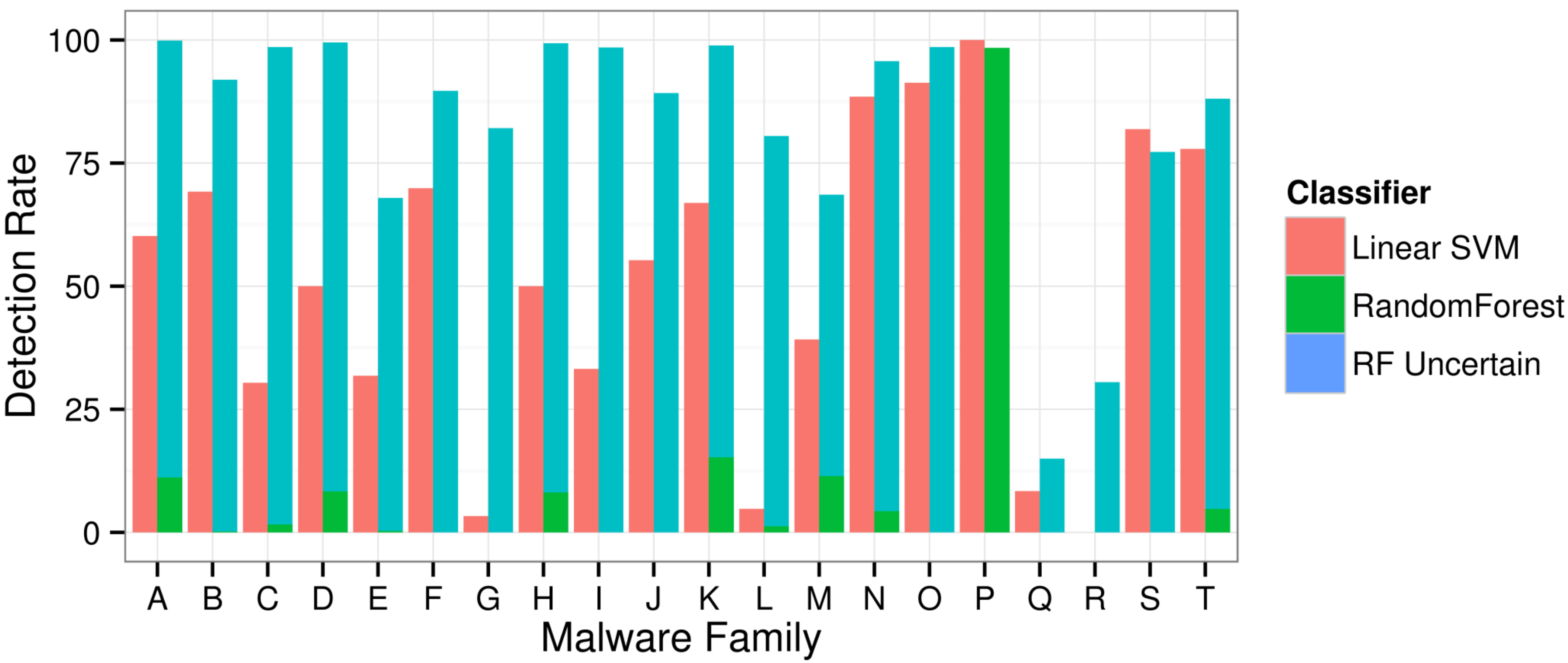


Drebin Unknown Family Detection

- Malware samples labeled by family
- Each family withheld from training set, included in evaluation



Drebin Classifier Comparison



Mimicus GD-KDE Attacks

- Gradient Decent and Kernel Density Estimation
  - Exploits known decision boundary of SVM
- Extremely effective against SVM based replica of PDFrate
  - Average score of 8.9%
- Classifier score spectrum is not enough

Evasion Resistant SVM Ensemble

- Construct Ensemble of multiple SVM
- Bagging of training data
  - Does not improve evasion resistance
- Feature Bagging (random sampling of features)
  - Critical for evasion resistance
- Ensemble SVM not susceptible to GD-KDE attacks

Conclusions

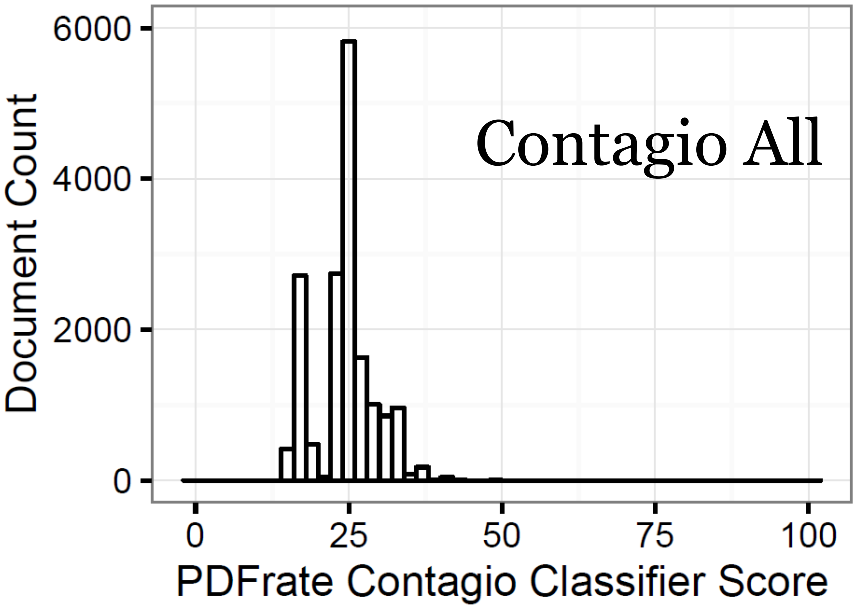
- Mutual agreement provides per observation confidence estimate
- no additional computation
- Feature bagging is critical to creating diversity required for mutual agreement analysis
- Strong (and private) training set improves evasion resistance
- Operators can detect most classifier failures
  - Perform complimentary detection, update classifier
- Mutual agreement analysis raises bar for mimicry attacks

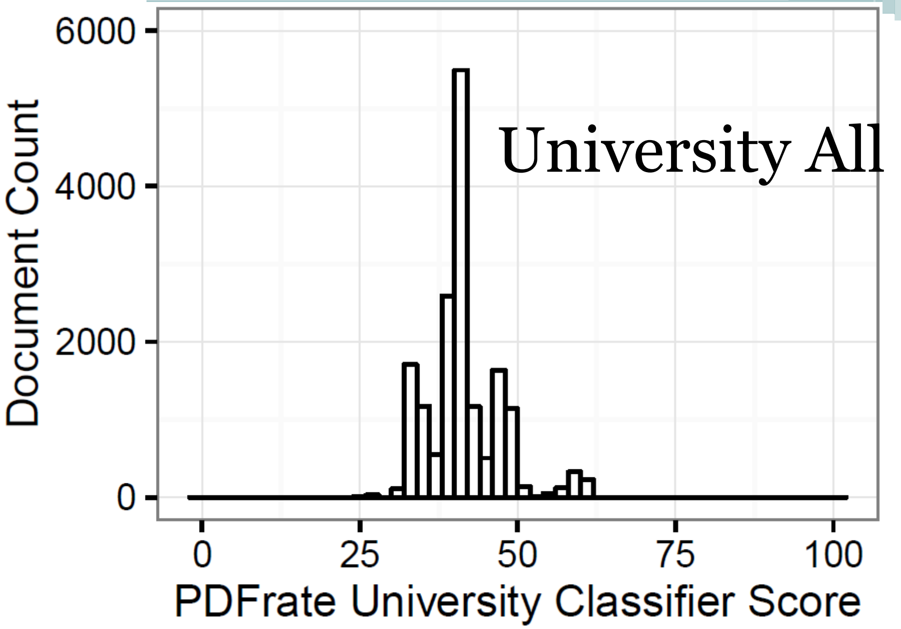
Questions

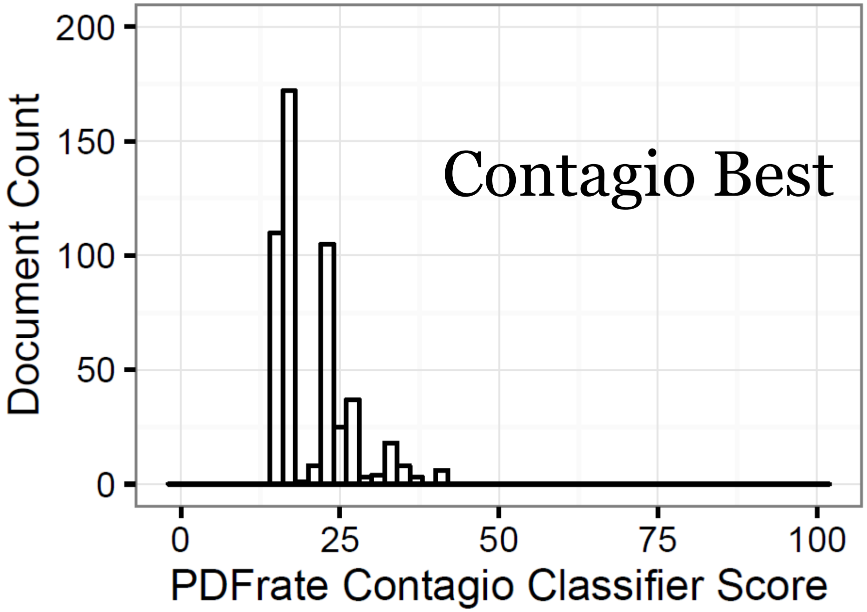
Charles Smutz, Angelos Stavrou
csmutz@gmu.edu, astavrou@gmu.edu

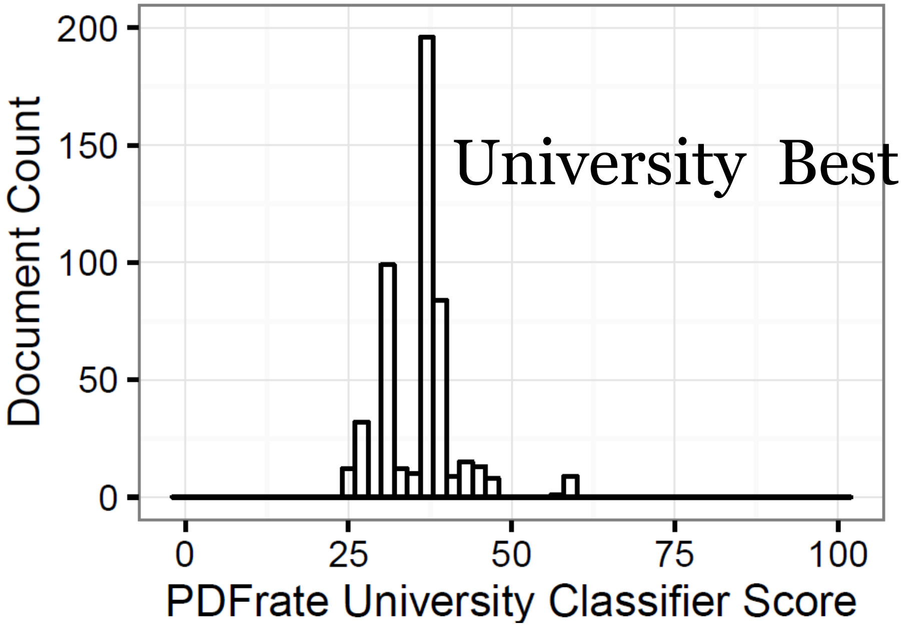
http://pdfrate.com

EvadeML Results

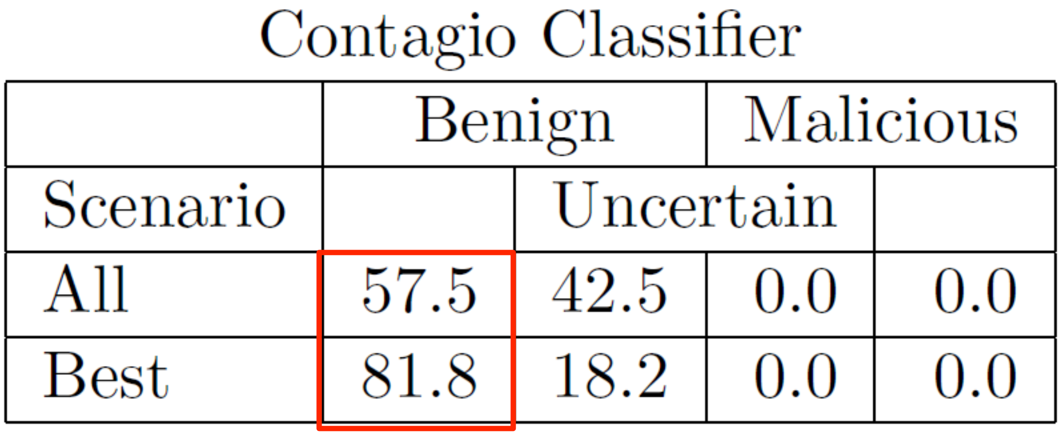

Contagio All
Document Count
6000
4000
2000
0
0
25
50
75
100
PDFrate Contagio Classifier Score


University All
Document Count
6000
4000
2000
0
0
25
50
75
100
PDFrate University Classifier Score


Contagio Best
Document Count
200
150
100
50
0
0
25
50
75
100
PDFrate Contagio Classifier Score

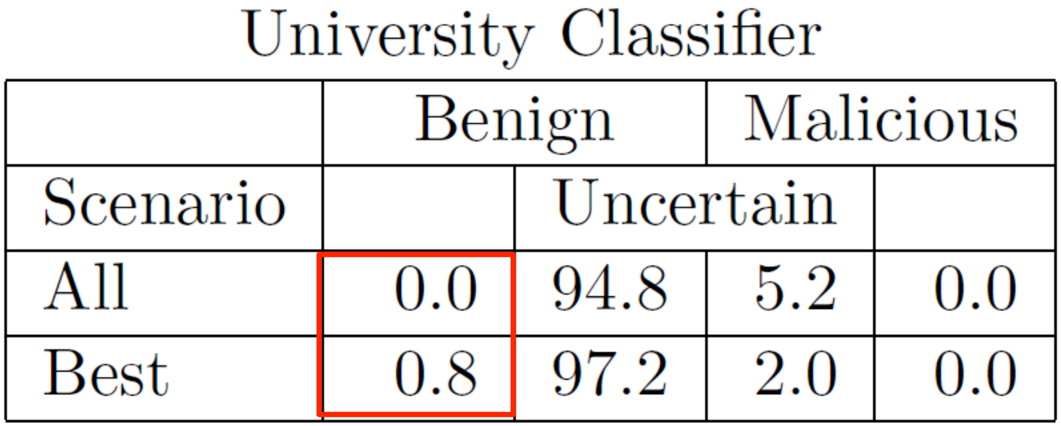

University Best
Document Count
200
150
100
50
0
0
25
50
75
100
PDFrate University Classifier Score

EvadeML Results



Contagio Classifier

	Benign		Malicious	
Scenario		Uncertain		
All	57.5	42.5	0.0	0.0
Best	81.8	18.2	0.0	0.0



University Classifier

	Benign		Malicious	
Scenario		Uncertain		
All	0.0	94.8	5.2	0.0
Best	0.8	97.2	2.0	0.0

Mutual Agreement Threshold Tuning

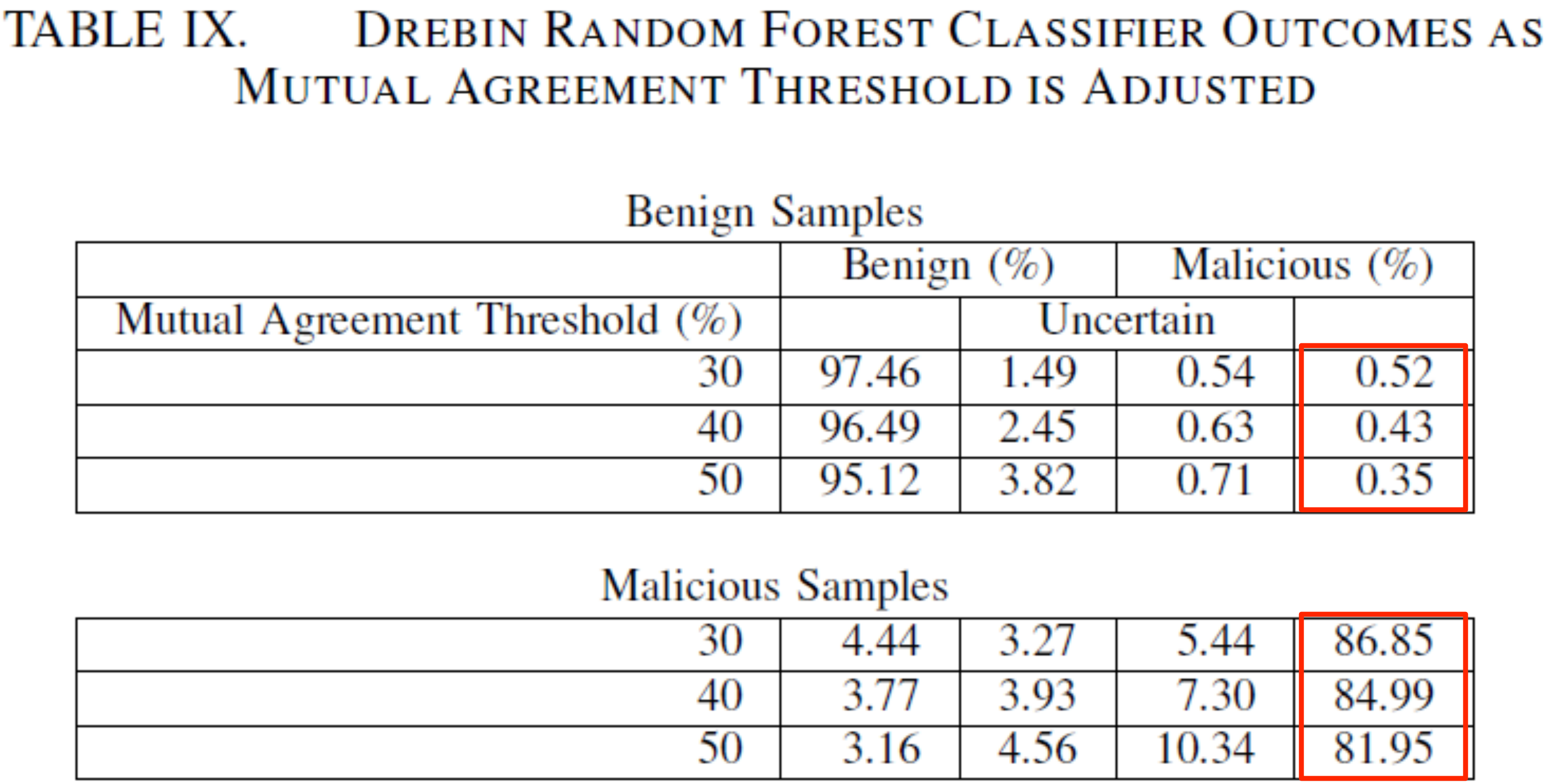


TABLE IX. Drebin Random Forest Classifier Outcomes as Mutual Agreement Threshold is Adjusted

Benign Samples

	Benign (%)		Malicious (%)	
Mutual Agreement Threshold (%)		Uncertain		
30	97.46	1.49	0.54	0.52
40	96.49	2.45	0.63	0.43
50	95.12	3.82	0.71	0.35

Malicious Samples

30	4.44	3.27	5.44	86.85
40	3.77	3.93	7.30	84.99
50	3.16	4.56	10.34	81.95