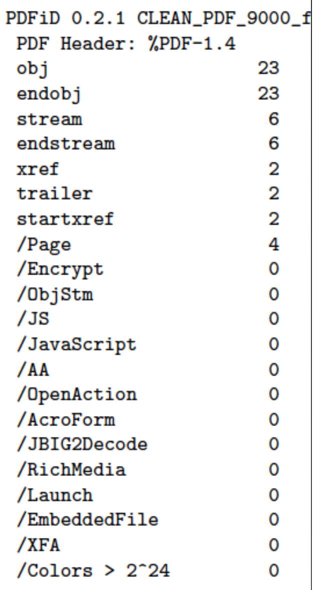
Machine Learning techniques for malicious PDF detection

Mattia Rosso, Lorenzo Ippolito, Martino Picasso

Introduction

- Classifying PDF files: malicious and benign
- ML: supervised binary classification
- Malware analysis: static keyword-based
- Benchmark: Contagio dataset
- Evasion attacks and countermeasures



```
PDFiD 0.2.1 CLEAN_PDF_9000_f
 PDF Header: %PDF-1.4
 obj                    23
 endobj                 23
 stream                  6
 endstream               6
 xref                    2
 trailer                 2
 startxref               2
 /Page                   4
 /Encrypt                0
 /ObjStm                 0
 /JS                     0
 /JavaScript             0
 /AA                     0
 /OpenAction             0
 /AcroForm               0
 /JBIG2Decode            0
 /RichMedia              0
 /Launch                 0
 /EmbeddedFile           0
 /XFA                    0
 /Colors > 2^24          0
```

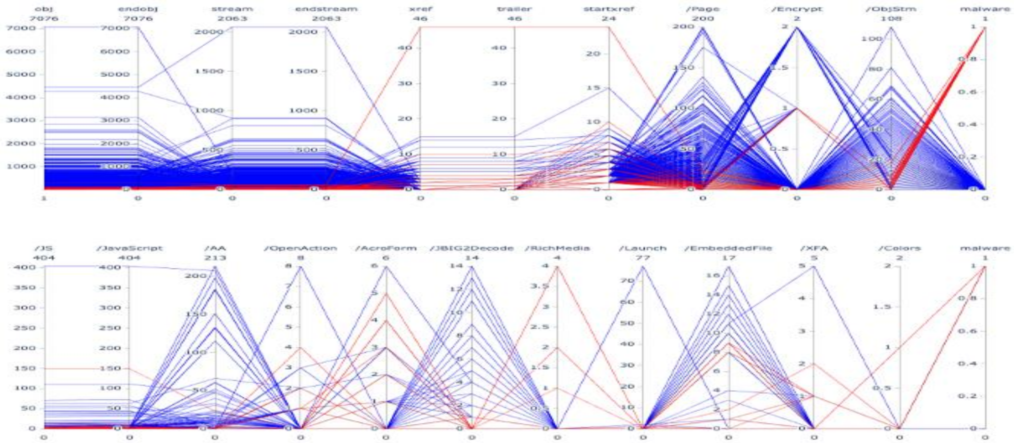
Contagio Benchmark

- 11100 Malicious and 9109 Benign PDFs
- Broadly adopted dataset

The PDFiD Tool

- Features vector: 21 keywords extracted
- Discriminative features for malware analysis

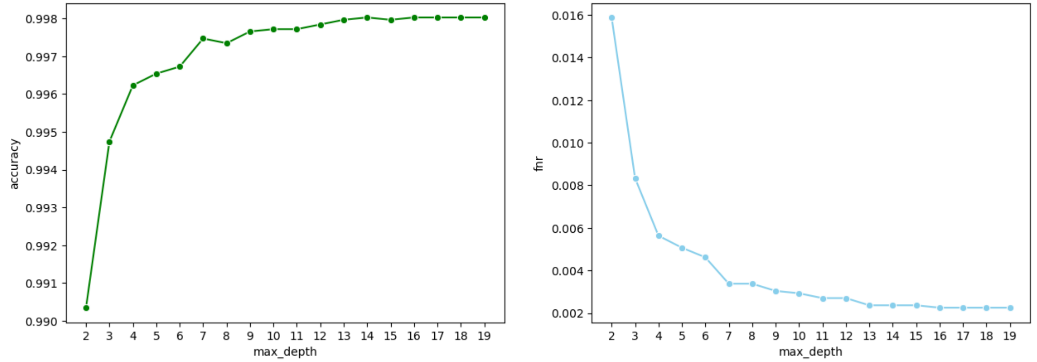
Features analysis



Model validation and metrics

- Accuracy and FNR (malicious PDFs classified benign)
- Static Train and Test split, k-fold Cross-Validation on Train

Model's Accuracy and False Negative Rate

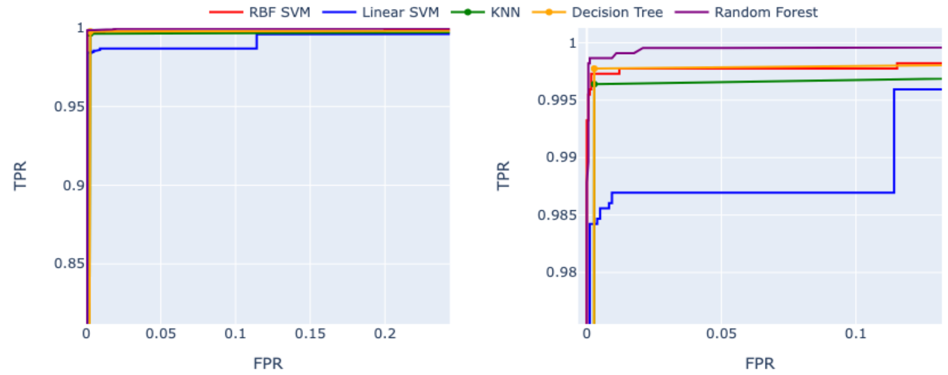


Performances achieved by the Random Forests (100 trees) on Training

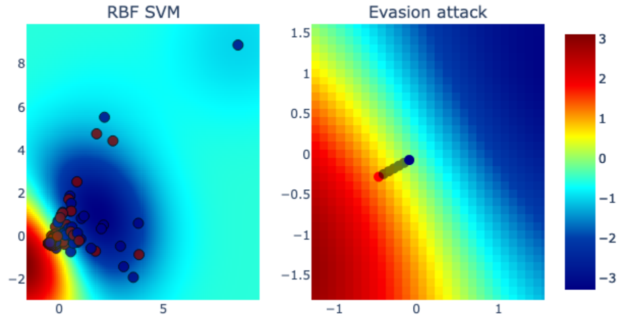
Model	Hyperparameters	Accuracy (%)	FNR (%)
Random Forests	depth=16	99.90	0.13
Decision Trees	depth=15	99.75	0.22
RBF SVM	C=5.56 γ=0.11	99.73	0.36
Linear SVM	C=6.67	99.06	1.58
KNN	k=1	99.68	0.36

Results on the Test set

ROC curves

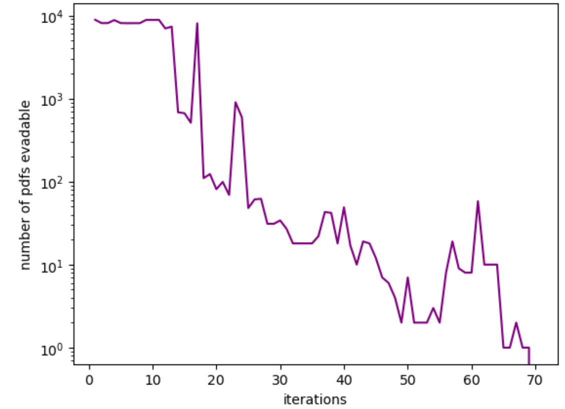


Evasion attack on RBF SVM

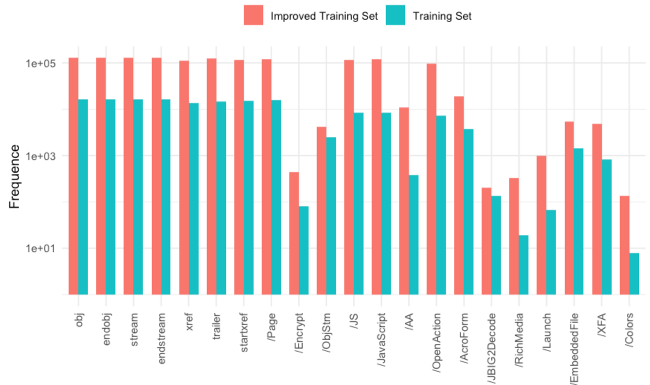


Gradient Descent based evasion attack on RBF SVM

Adversarial Learning on Decision Trees



Adversarial Learning against evasion attack of the Decision Trees


Improved Training Set
Training Set
1e+05
1e+03
1e+01
Frequence
obj
endobj
stream
endstream
xref
trailer
startxref
/Page
/Encrypt
/ObjStm
/JS
/JavaScript
/AA
/OpenAction
/AcroForm
/JBIG2Decode
/RichMedia
/Launch
/EmbeddedFile
/XFA
/Colors