ARIEL UNIVERSITY
Department of Computer Science
SNDBOX

Malicious PDF File Detection

By: Alexey Titov, Shir Bentabou
Supervisors: Dr. Amit Dvir, Dr. Ran Dubin

Introduction:

Malware is software designed to serve any kind of cyber-attack. Phishing is one of the popular distribution methods for malware. It is the fraudulent attempt to obtain sensitive information from a target by disguising as a trustworthy entity. One of the growing ways to carry out phishing is through PDF files. Portable Document Format, is a file format used worldwide for over 20 years and has become one of the leading standards for the dissemination of textual documents.

The Problem:

PDF files have features that make them an easy-to-use attack vector. They allow pictures and hyperlinks to be embedded in them for easy use and enables the use of PDF objects and streams in the body part of the file. Moreover, since version 1.2, PDF enables JavaScript as a feature, thus helping the common user - but also opening a new attack vector. Using these properties, code can be executed without the user knowing, or with his knowledge but without the awareness that this could be unsafe for him.

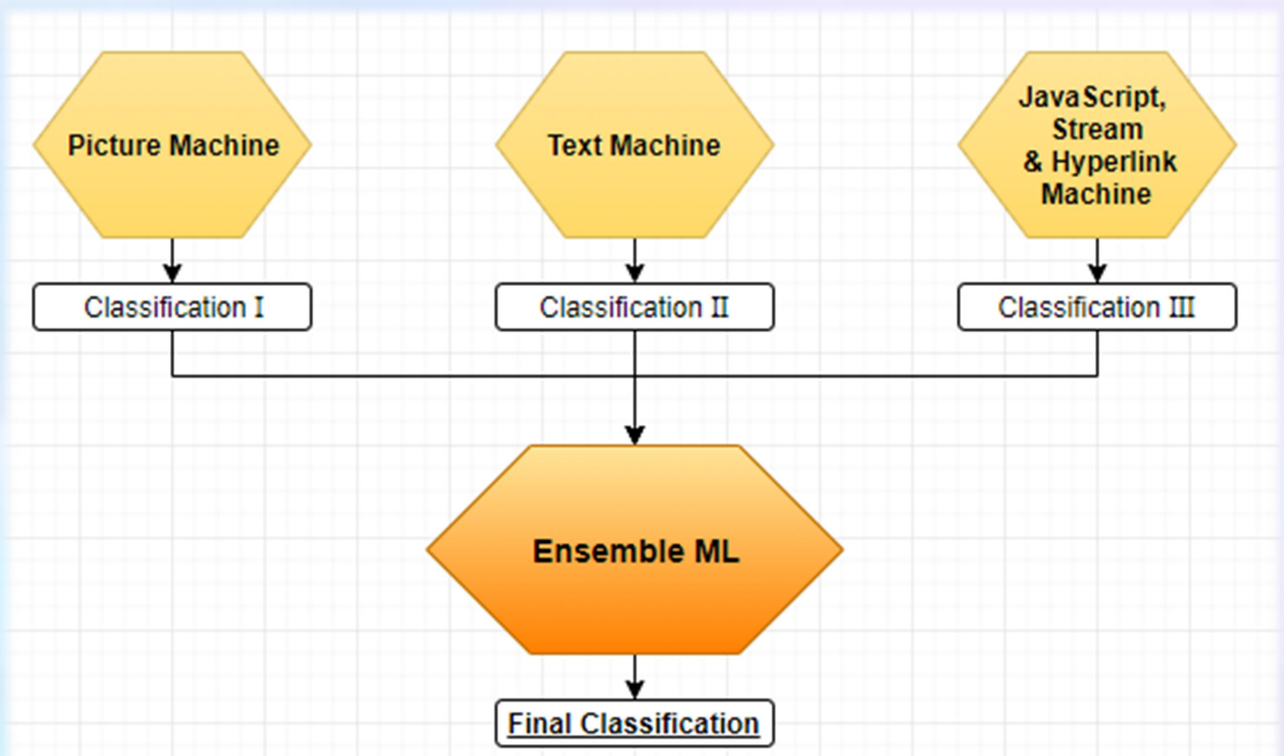
Most previous researches (e.g. [1][2]) focus on the features of the files alone, selecting the most significant ones and utilizing different machine learning methods to identify malwares in PDFs based solely on these features. This method is not ideal, since it has been proven to be unsafe against evasion attacks, in which the attacker knows the features chosen, and uses some strategy to suit himself to the classifier.

Our Solution:

Many different criteria's can be used to categorize PDF files. In our project we built a classifying machine, based on three different machines that each focuses on some criteria.

1) Information from file preview.
2) The text in the file.
3) JavaScript, objects, streams and hyperlinks.

These machines classify the files based on machine learning techniques, using a variety of algorithms.



Machine Features	
Machine Ⅰ	Picture Histogram, Blur
Machine Ⅱ	Text Vector (word2vec)
Machine Ⅲ	JS code, objects, streams, activities, URLs.

Future Improvements:

Enhancement can be reached by testing the machines on different algorithms in order to find the algorithms that provide the best results, and by improving the features that the classification is based on.

[1] Malicious PDF Documents Detection using Machine Learning Techniques A Practical Approach with Cloud Computing Applications - Jose Torres and Sergio De Los Santos
[2] Malware Detection in PDF Files Using Machine Learning - Bonan Cuan, Aliénor Damien, Claire Delaplace, Mathieu Valois