Malware Detection

Aprendizagem Aplicada à Segurança

Mário Antunes

November 28, 2025

Malware Classification using Dynamic and Hybrid Features

Objective

In this lab, you will explore two different approaches to malware detection using Machine Learning. You will move from analyzing how a program behaves (Dynamic Analysis) to a more complex view that combines behavior with file structure (Hybrid Analysis).

Part 1: Post-Execution Analysis (Dynamic)

Dataset:

API Call Sequences (Kaggle)

Concept

Dynamic analysis involves running the malware in a secure sandbox (like Cuckoo Sandbox) and recording what it does. The most common "behavioral signature" is the sequence of API calls (e.g., CreateFile, RegOpenKey, SocketConnect).

Your Task

1. Load the Dataset: Import the API Call dataset.
2. Explore the Data:
   - Look at the api_calls column. Notice how it is a sequence of text strings?
   - Machine Learning models need numbers, not strings.
3. Preprocessing Challenge:
   - You need to convert these text sequences into numerical data.
   - Hint: Look into Label Encoding (assigning a number to every unique API name) or TF-IDF (counting frequency).
4. Modeling:
   - Train a model to predict the malware family (or binary classification).
   - Since this is sequential data (Time Series), simple models like Random Forest work, but models like LSTMs (Long Short-Term Memory) or GRUs are theoretically better.

Part 2: Hybrid Analysis (Static + Dynamic)

Dataset:

Malware Static and Dynamic Features (UCI)

Concept

Static analysis looks at the file on the disk (PE Headers, file size, imports) without running it. Dynamic analysis looks at execution. A Hybrid model uses both to cover the weaknesses of each approach.

Your Task

1. Load the Dataset: Import the UCI dataset.
2. Compare Features:
    - Identify which columns are Static (usually starting with check_sum, section_, or header_).
    - Identify which columns are Dynamic (usually specific API names or counts).
3. Correlation Analysis:
    - Do static features correlate with dynamic features?
4. Modeling:
    - Train a Classifier (e.g., Random Forest or XGBoost) using all features.
    - Experiment: Try training a model using only static features, then only dynamic features. Compare the accuracy to the combined model.