University of Guelph

Computer Science Department

CIS*6530 - Threat Intel & Risk Analysis

Extracting Opcodes: Researching APT group set 5

Submission 4

(20%)

Team Members

Nafis Ahmed Awsaf | nawsaf@uoguelph.ca | 1402517

Jacob Drobena | jdrobena@uoguelph.ca | 0969071

Professor: Dr. Ali Dehghantanha

Deadline: Oct 30nd, September 2025

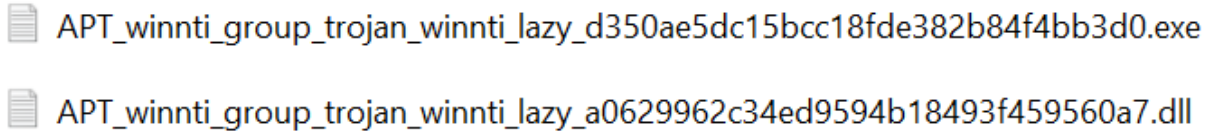
Table of Contents:

I. Introduction:

The primary goal of this project was to preprocess malware opcode files and evaluate their effectiveness in classifying different Advanced Persistent Threats (APTs) using classical machine learning models such as SVM, KNN, and Decision Trees. This phase continues the previously implemented opcode extraction pipeline pulling more opcode data from a large set of Advanced Persistent Threat (APT) malware samples. Through extending the timeout period for opcode processing, the present opcode dataset was achieved. This dataset will serve as the basis for the following steps in feature extraction and machine learning (ML) analysis.

II. Data Preparation:

Raw opcode files were structured per executable and annotated with filenames indicating the associated APT's



Opcode mnemonics were extracted, and duplicate or trivial differences across families (such as Whitefly vs TG-3390, or Machete variants being clones of DarkHotel) were removed based on hash similarity as well as notes from MITRE ATT&CK; profiles and vendor overlap (e.g., Promethium merged with Neodymium).

073cbd6533835a3df9a0c569c	ThreatGrou
073cbd6533835a3df9a0c569c	Whitefly
073cbd6533835a3df9a0c569c	Whitefly
073cbd6533835a3df9a0c569c	Whitefly
073cbd6533835a3df9a0c569c	Whitefly

Low-representation APTs (those with <5 samples) were previously merged into an 'OTHER' class. However it was imperative for the dataset to remove the under represented values for proper training.

III. Dataset & Preprocessing:

Opcode sequences were converted into n-gram features using scikit-learn's CountVectorizer. Both unigram (1-gram (MOV, PUSH, CALL)) and bigram (2 gram (PUSH MOV, MOV CALL)) features were explored. Files with extreme opcode counts or structure were flagged and removed using z-score outlier detection. A similarity threshold of 0.75 was used for identifying duplicate samples across APTs using Jaccard similarity. Files flagged as structurally identical or excessively similar (especially across different APT labels) were also excluded.

s\APT_Promethium_trojan_graf	within-APT-outlier	7.231994	APT internal outlier
s\APT_Promethium_Win32Stror	within-APT-outlier	7.231994	APT internal outlier
s\APT_Carbanak_trojan_mint_z	within-APT-outlier	4.185978	APT internal outlier
s\APT_LotusBlossom_trojan_m	within-APT-outlier	3.855278	APT internal outlier
s\APT_StealthFalcon_trojan_de	within-APT-outlier	3.739271	APT internal outlier
s\APT_APT18_9603b62268c2b	within-APT-outlier	3.479871	APT internal outlier
s\APT_APT38_trojan_alreay_do	within-APT-outlier	3.448888	APT internal outlier

The final dataset resulting in 600 opcode files as a result of refactoring.

IV. Evaluation Metrics:

Metrics included:
- Accuracy: Overall correct predictions.
- Precision: True positives over predicted positives per class.
- Recall: True positives over actual positives per class.
- F1-score: Harmonic mean of macro precision and recall.
- Confusion matrix: To visualize misclassifications across APTs. Macro-averaging was chosen due to class imbalance, giving equal weight to all APT families.

VIII. Results & Observations:

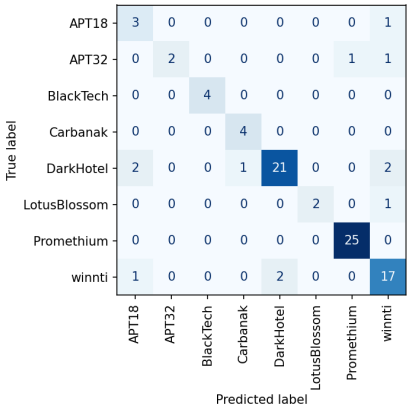
After completing the preprocessing pipeline, performing outlier removal, deleting duplicated opcode hashes, and rebalancing rare APT families, the refined dataset was used to train and evaluate three classical machine-learning models. Decision Tree, KNN (k=4), and SVM across both 1-gram and 2-gram opcode feature spaces.

The revised dataset was significantly cleaner and more representative than previous iterations. This led to a substantial increase in classification performance, especially for Decision Trees, which benefited from the reduced cross-APT opcode similarity and the removal of high-variance outlier files.

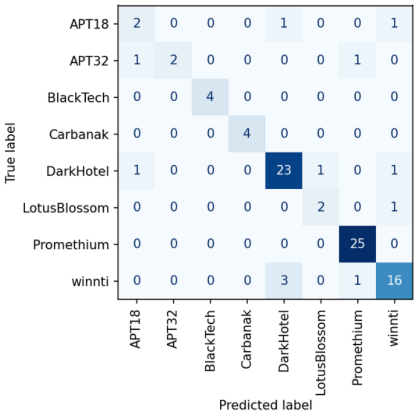
```
*** Overall Classifier Comparison ***
           Model Feature  Accuracy  Precision  Recall  F1 Score
0  Decision Tree  1-gram     86.67      88.26   86.67     86.68
1            KNN  1-gram     71.11      70.97   71.11     70.89
2            SVM  1-gram     47.78      29.90   47.78     35.80
3  Decision Tree  2-gram     86.67      86.82   86.67     86.31
4            KNN  2-gram     71.11      70.97   71.11     70.89
5            SVM  2-gram     47.78      29.90   47.78     35.80
```

VIIII. Appendix and other images:

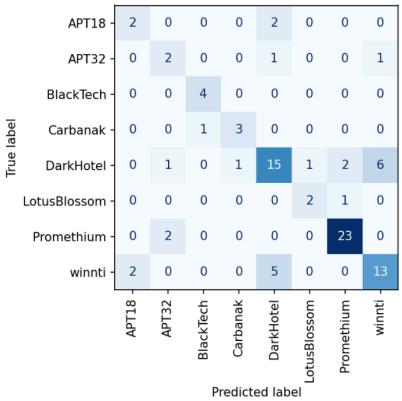
Decision tree 1 gram



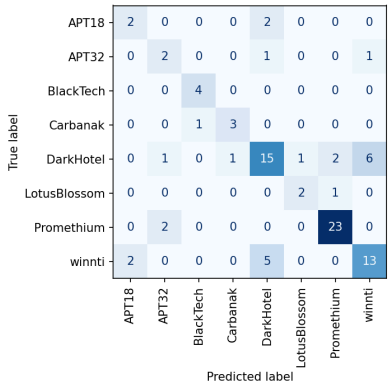
Decision tree 2 gram



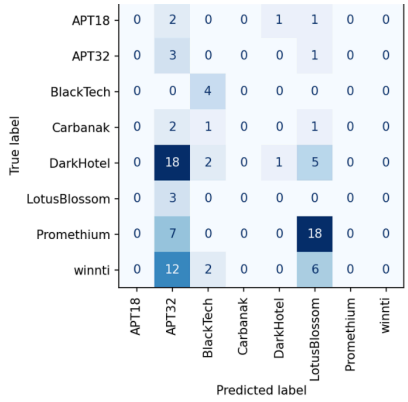
KNN 1 gram



KNN 2 gram



SVM 1 gram



SVM 2 gram