Anomaly Detection in File Access Patterns

Ricardo Machado
NMEC : 102737

universidade de aveiro

WHY MONITOR FILE ACCESS?

- **Motivation:** Data security is crucial in corporate networks, where unauthorized file access can lead to data breaches, financial losses, and reputational damage, making anomaly detection a crucial security measure
  - **Example:** In 2022, more than 30% of companies suffered internal attacks that exposed sensitive data

- **Objective:** Implement a system to detect anomalous access patterns effectively, helping to prevent threats

universidade de aveiro

DATA SOURCES

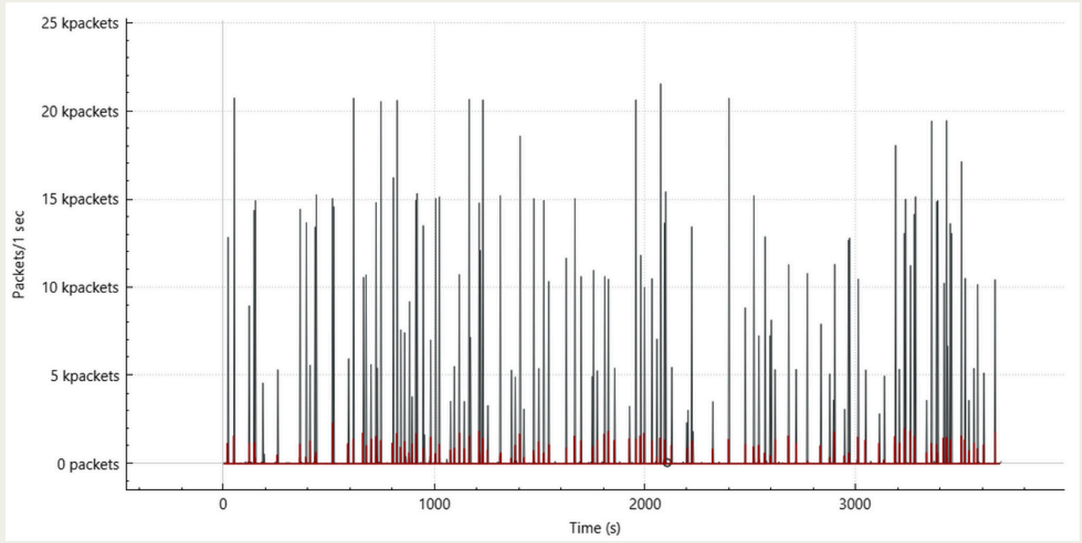
- Acquisition of Google Drive usage data captures from users with two different profiles: student and teacher
- Bot script development to access Google Drive (basic and advanced)
  - Basic: simple simulation with a fixed time interval
  - Advanced: introduces more randomness and unpredictability

universidade de aveiro

WIRESHARK I/O GRAPHS FROM BOTH BOTS

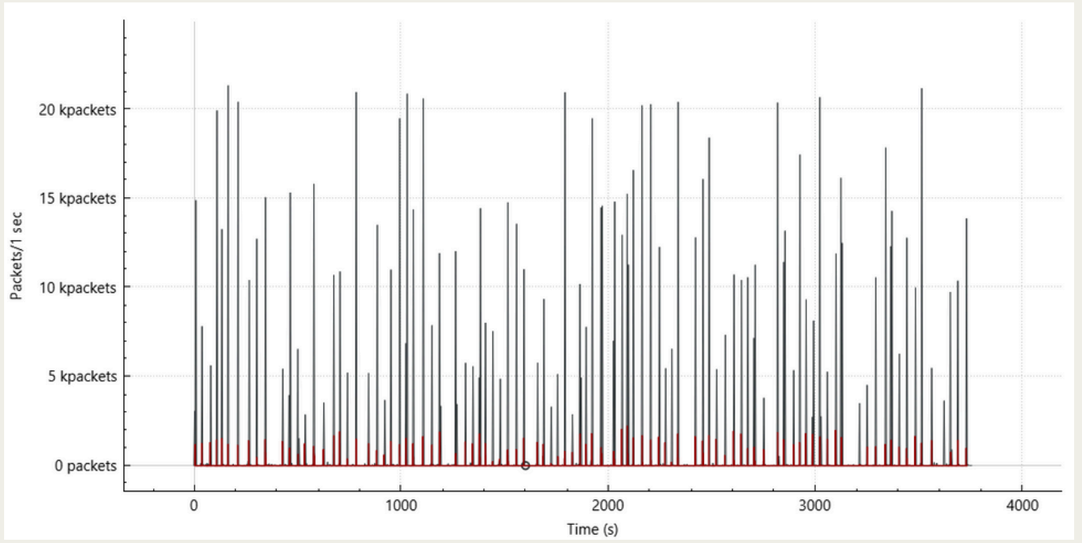
Basic Bot

- Aggressive and irregular spikes
- Easier Detection - abnormal spikes



Advanced Bot

- More balanced and controlled
- Harder Detection - disguised



METRICS EXTRACTED

- Number of packets for upload and download
- Volume of bytes for upload and download

FEATURES TO EXTRACT

- For each metric defined:
  - Mean, median, standard deviation
  - Percentiles(75%/80%/90%/98%)

universidade de aveiro

TRAINING AND TEST FEATURES

Initial approach

- Train using 70% of each user's features (student & teacher)
- Test on the remaining 30% of user features (same users)
- Test on bot features(Basic and Advanced bots)

Second approach

- Train using 50% of each user's features (student & teacher)
- Test on the remaining 50% of user features (same users)
- Test on bot features(Basic and Advanced bots)

universidade de aveiro

ANOMALY DETECTION TECHNIQUES USED

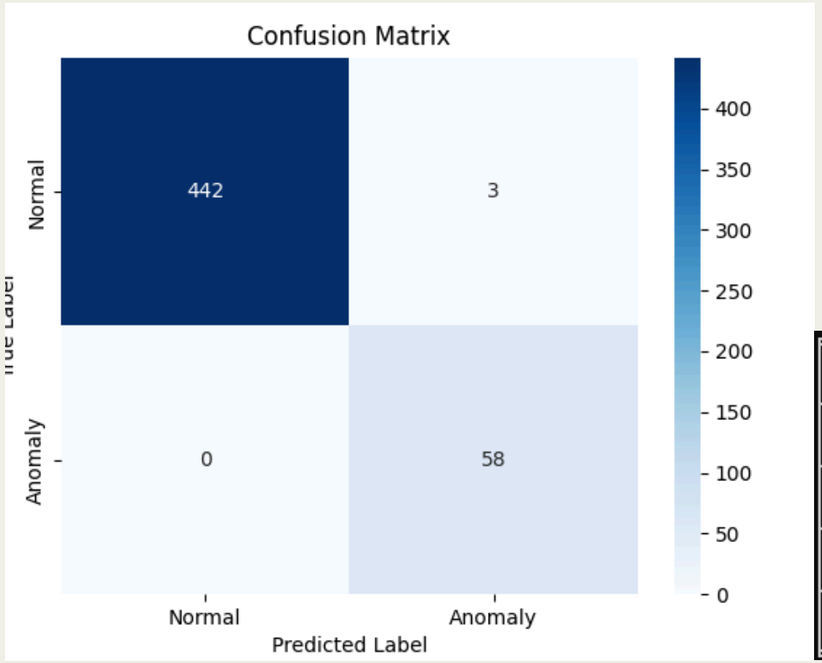
- **Statistical Analyses:**
  - Centroids distances
  - Centroids distances with PCA features
  - Multivariate with PCA features
- **Machine Learning:**
  - One Class Support Vector Machine (Linear, RBF and Poly Kernels)
  - One Class Support Vector Machine with PCA features (Linear, RBF and Poly Kernels)
- **Anomaly Detection:**
  - Isolation Forest with PCA features
  - Isolation Forest without PCA features

universidade de aveiro

BEST RESULTS FROM INITIAL APPROACH

Isolation Forest with PCA

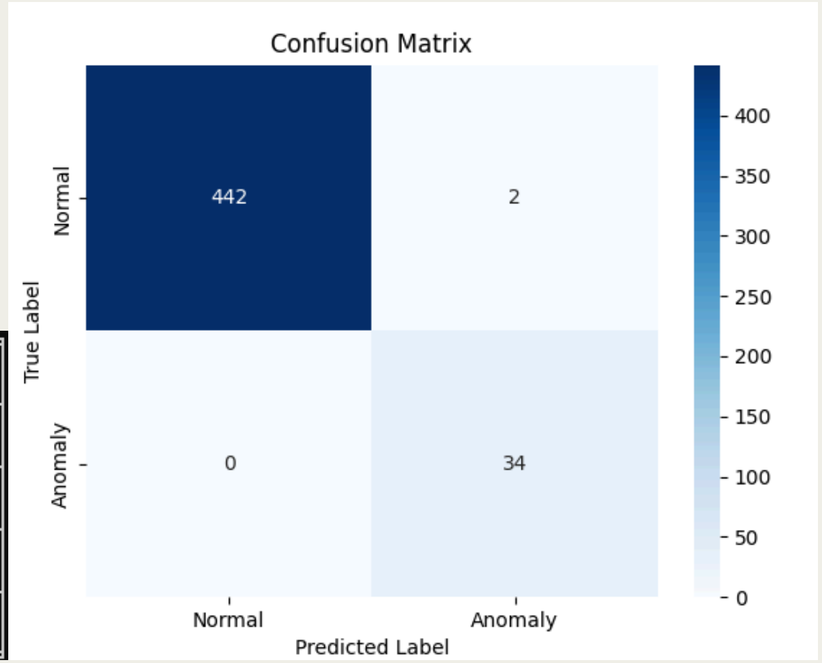
Basic Bot



Metric	Value
Accuracy	99.40%
Precision	95.08%
Recall	100.00%
F1-Score	0.97

Advanced Bot

Metric	Value
Accuracy	99.58%
Precision	94.44%
Recall	100.00%
F1-Score	0.97



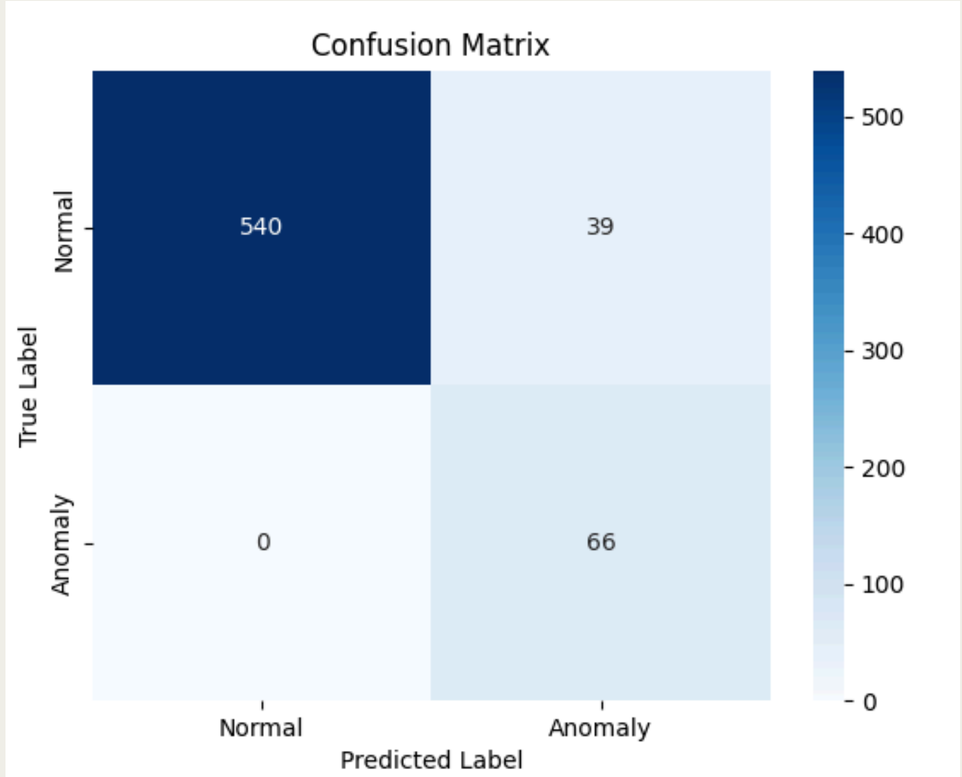
CONCLUSION FOR INITIAL APPROACH

The Isolation Forest model, when using PCA-transformed features, proves to be the most effective in identifying anomalies across all bot traffic in the dataset

universidade de aveiro

BEST RESULTS FROM SECOND APPROACH

Multivariate with PCA



Metric	Value
Accuracy	93.95%
Precision	62.86%
Recall	100.00%
F1-Score	0.77

CONCLUSION FOR SECOND APPROACH

The most effective models is Multivariate with

PCA achieving an F1-score of 77%.

universidade de aveiro

CONCLUSION

- The Isolation Forest model with PCA features proves to be the most effective in detecting anomalies across different bot behaviors

- Multivariate analysis perform well when analyzing activity-based features

- The user-based approach yields more consistent results, suggesting that user similarity plays a crucial role in anomaly detection

universidade de aveiro

Thank you!

universidade de aveiro