•Generate high-quality detection rules automatically
•Support faster and more consistent response to new threats

ÓBUDA UNIVERSITY
JOHN VON NEUMANN FACULTY OF INFORMATICS

Adaptive Malware Classification and Automatic YARA Rule Generation

Ferenc Leitold
Óbuda University, Budapest, HUNGARY
leitold.ferenc@nik.uni-obuda.hu

Eszter Kail
Óbuda University, Budapest, HUNGARY
kail.eszter@nik.uni-obuda.hu

INTRODUCTION

The growing number and diversity of malicious software challenge traditional classification systems, which rely on predefined labels and struggle to keep pace with emerging threats.

YARA rules are powerful tools for malware identification and threat hunting, but manual creation is slow and expert-dependent. This makes it difficult to respond quickly to new malware families.

Our research proposes a **self-adaptive malware classification framework** combined with **automatic YARA rule generation**. By integrating **hybrid static and dynamic features**, applying unsupervised clustering, and turning results into **actionable YARA signatures**, we aim to:

- Group malware into meaningful families at scale
- Generate high-quality detection rules automatically
- Support faster and more consistent response to new threats

This approach provides a scalable foundation for organizing the growing malware ecosystem and improving operational threat intelligence.

EXPECTED OUTCOME

Improved clustering: More consistent grouping of malware variants, including unknown or emerging families, enabling better threat tracking.

Automated YARA rule creation: Faster generation of accurate, reproducible rules directly from data, reducing manual effort and human error.

Operational benefits: Better detection coverage, fewer false positives, and faster incident response for security teams.

Prototype readiness: Provides a prototype that can be integrated into existing malware analysis and threat intelligence workflows.

METHODOLOGY (planned)

Our approach combines static and dynamic analysis, unsupervised clustering, and automated rule generation in a modular pipeline:

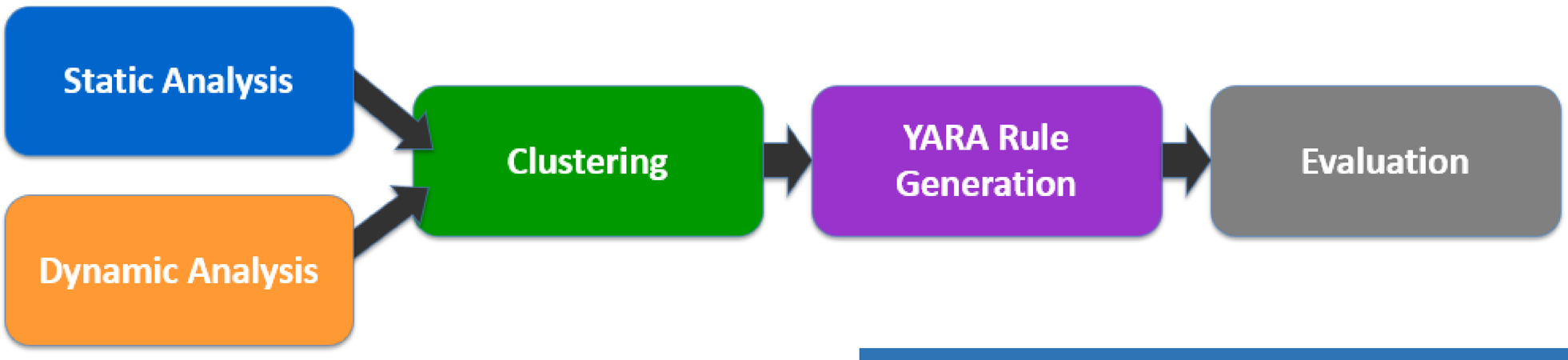
1. Static Analysis
- Disassembly of Windows PE files to extract instruction patterns
- Extraction of PE structural indicators using YARA's PE module
- Statistical selection of discriminative byte patterns (n-grams, basic blocks)

2. Dynamic Analysis
- Sandbox execution to capture behavioral traces:
  - API calls
  - File and registry modifications
  - Network indicators (domains, IPs, URLs)
- Conversion of runtime events into feature vectors for clustering

3. Clustering & Weak Supervision
- Application of Spectral Clustering, DBSCAN, and biclustering
- Comparison of clusters with AV engine family labels to improve group coherence
- Assignment of new samples to existing clusters or creation of new ones

4. Automatic YARA Rule Generation
- Translation of selected static and dynamic features into human-readable YARA rules
- Integration of packer-specific signatures to detect protected samples
- Use of genetic algorithms to refine rules and reduce false positives

5. Evaluation
- Metrics: Detection rate, False Positive rate, Precision, Recall, F1-score
- Testing on unseen samples to measure zero-day detection capability.



CONTACT

Ferenc Leitold
Óbuda University, Budapest, HUNGARY
leitold.ferenc@nik.uni-obuda.hu

Eszter Kail
Óbuda University, Budapest, HUNGARY
kail.eszter@nik.uni-obuda.hu

REFERENCES

[1] Z. Mansour, W. Ou, S. H. H. Ding, M. Zulkernine, and P. Charland, **"NeuroYara: Learning to Rank for Yara Rules Generation Through Deep Language Modeling and Discriminative N-Gram Encoding,"** IEEE Trans. Dependable Secure Comput., vol. 22, no. 2, pp. 1747–1762, Mar. 2025, doi: 10.1109/TDSC.2024.3449641.

[2] Edward Raff, Richard Zak, Gary Lopez Munoz, William, Fleming, Hyrum S. Anderson, Bobby Filar, Charles, and Nicholas, and James Holt**, "Automatic Yara Rule Generation Using Biclustering,"** 13th ACM Workshop Artif. Intell. Secur. AISec'20, [Online]. Available: arXiv: 2009.03779. 2020.

[3] A. Zhdanov**, "Generation of Static YARA-Signatures Using Genetic Algorithm,"** in 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), June 2019, pp. 220–228. doi: 10.1109/EuroSPW.2019.00031.