CERIAS

The Center for Education and Research in Information Assurance and Security

Malware Language Processing "MLP": Developing a new paradigm for malware analysis and classification using Machine Learning and Artificial Intelligence

Graduate Student : Solomon Sonya PI: Professor Dongyan Xu 2024 Department of Computer Science, Purdue University

1. Problem Statement

- Malware continues to increase in prevalence and sophistication.
- Traditional detection mechanisms including antivirus software fail to adequately detect new and varied malware.
- Artificial Intelligence and Machine Learning models (AI/ML) provide advanced capabilities that can be applied to the cybersecurity domain to enhance detection and classification of malware and malware families. Building a **robust and automated artificially intelligent malware analysis framework** and **producing new, standardized malware datasets for future classification in AI/ML** however, **are not trivial**.

2. Research Objective

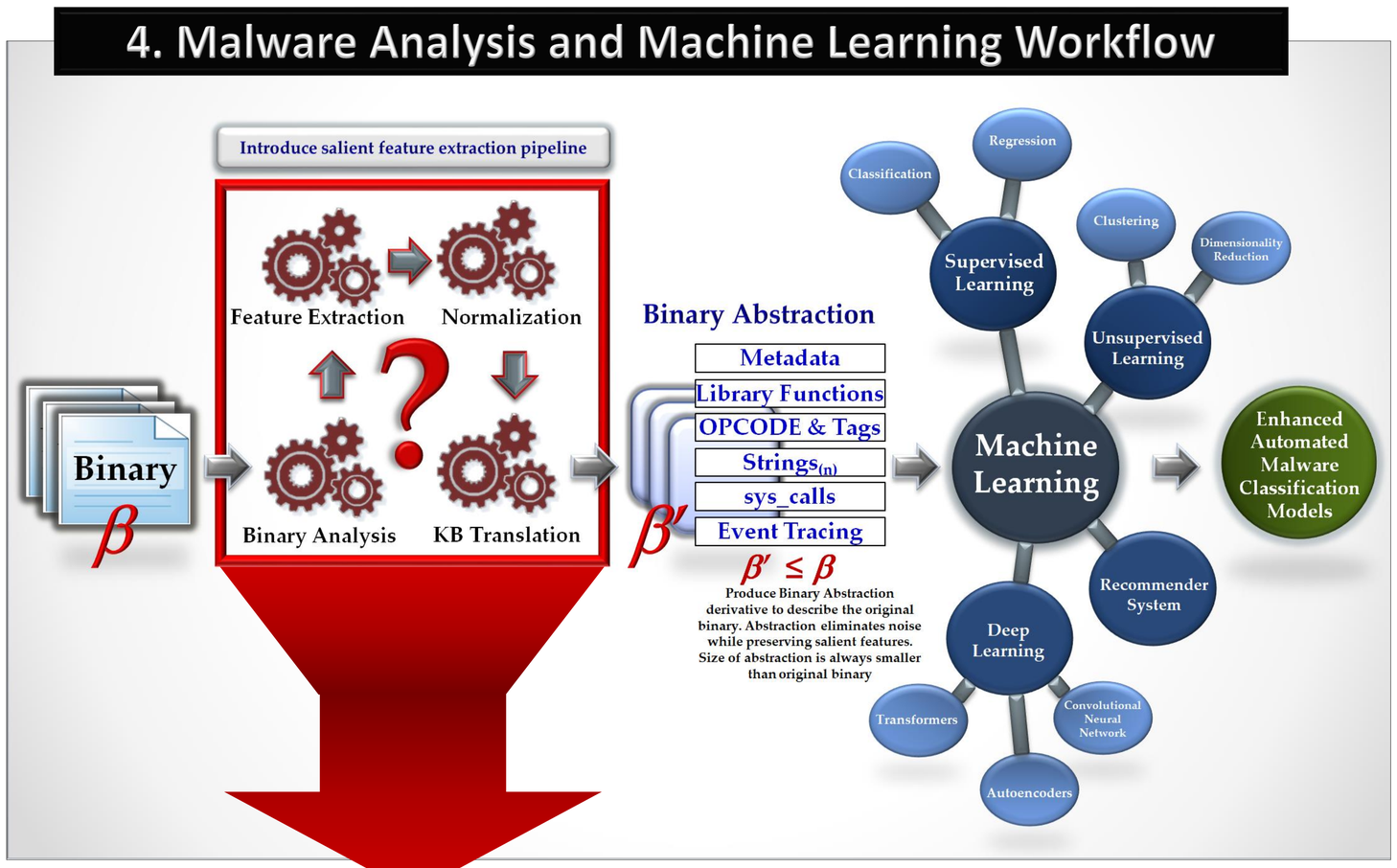
The principal objectives of this research are to deliver a new malware analysis framework, create a new ML heterogeneous classification model to analyze malware, produce uniform datasets for additional AI/ML analysis, and increase classification accuracy across complete and highly diverse malware corpora.

3. Research Contribution

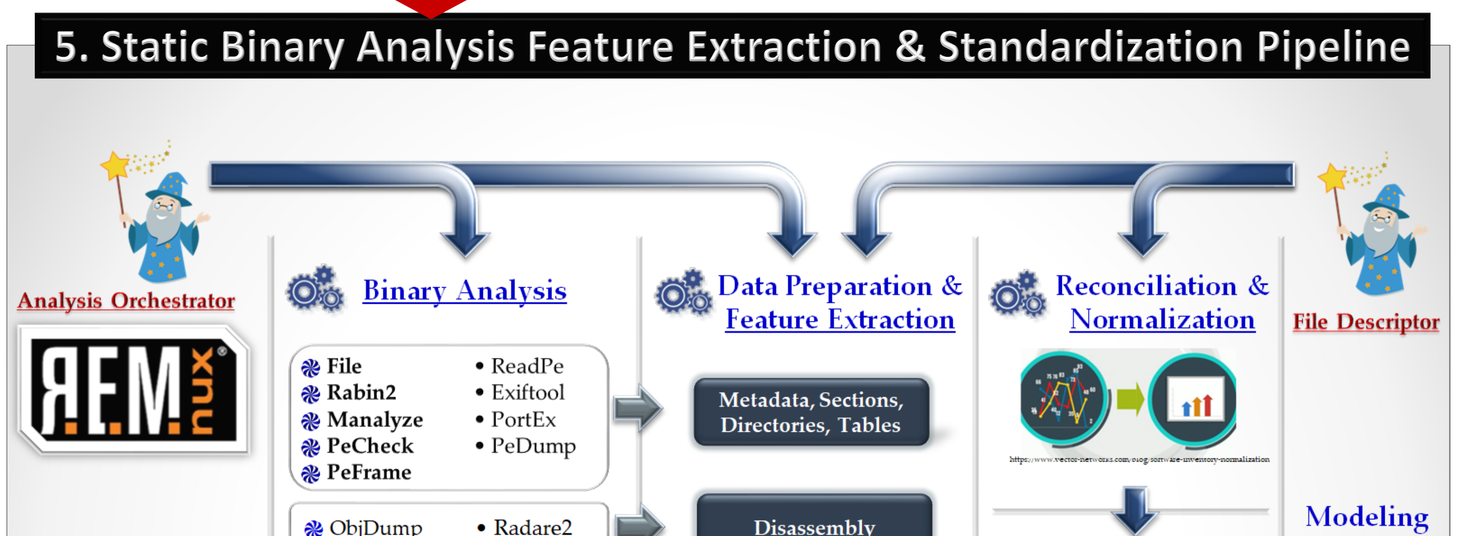
To our knowledge, this is the first research that:

- Releases a complete framework that automates malware analysis and produces new and standardized malware description datasets (ready for AI/ML analysis)
- Develops a classification ensemble applied to malware analysis (named the Malware Ensemble Classification Facility) that improves traditional ML model selection.
- **Releases trained ensemble classifier providing enhanced classification results of an entire 200+ gigabyte, malware family corpus consisting of 80K+ unique malware samples and 70+ malware families**.

4. Malware Analysis and Machine Learning Workflow

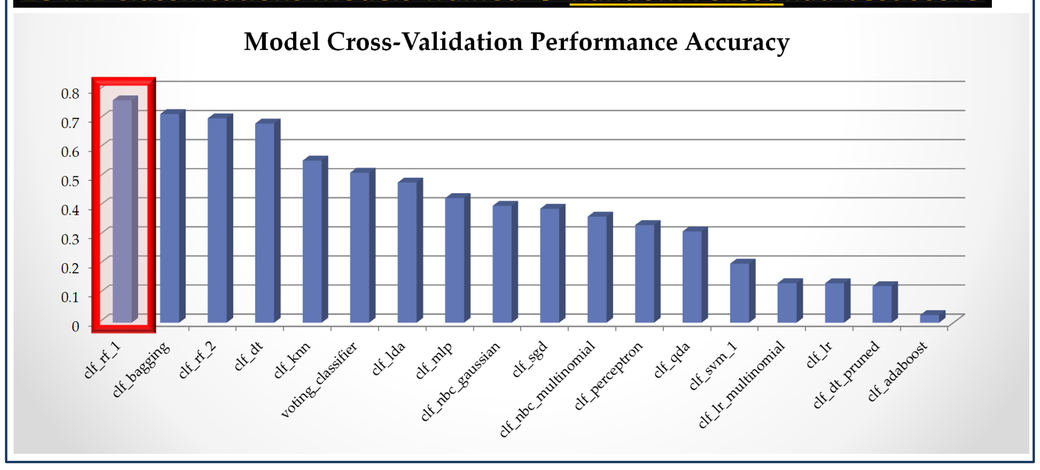


5. Static Binary Analysis Feature Extraction & Standardization Pipeline



6. Preliminary Analysis of Classification Models

18 ML Classifications Models Trained → Random Forest had best score



7. Classification Report of each Classifier Model

Ensemble classifier can improve results of overall classification performance

FYI: ClamAV detection rate: ~47%*

Precision Chart	7ev3n	9002Rat	Agent Tesla	Aria-Body	Bankshot	Bazarbac kdoor	Blister Loader	Cobalt Strike	Collector Goomba	Conti	CoreShell	Cryptolocker	Darkside	Dexbia	Dyre
clf_rf_1	0.81	0.05	0.76	0.68	0.19	0.56	0.31	0.30	0.10	0.90	0.86	0.50	0.32	0.08	0.07
clf_bagging	0.01	0.97	0.06	0.22	0.89	0.32	0.03	0.87	0.71	0.60	0.56	0.25	0.45	0.72	0.28
clf_rf_2	0.79	0.69	0.85	0.49	0.28	0.04	0.14	0.16	0.40	0.64	0.65	0.27	0.91	0.02	0.86
clf_dt	0.03	0.74	0.66	0.26	0.13	0.96	0.41	0.33	0.05	0.89	0.97	0.74	0.17	0.44	0.97
clf_knn	0.06	0.02	0.30	0.91	0.13	0.03	0.92	0.24	0.23	0.88	0.53	0.26	0.77	0.89
voting_classifier	1.00	0.97	0.32	0.49	0.32	0.06	0.24	0.42	0.21	0.61	0.29	0.89	0.21
clf_lda	0.63	0.58	0.37	0.03	0.21	0.06	0.87	0.42	0.69	0.20	0.84	0.68	0.21	0.65
clf_mlp	0.57	0.59	0.13	0.79	0.47	0.04	0.40	0.61	0.50	0.43	0.33	0.65	0.38	0.23
clf_sgd	0.50	0.56	0.69	0.04	0.79	0.62	0.38	0.60	0.43	0.88	0.26	0.41	0.13	0.57	0.53
clf_nbc_gaussian	0.80	0.53	0.07	0.97	0.22	0.92	0.08	0.57	0.09	0.52	0.83	0.62	0.90	0.29	0.28
clf_perceptron	0.26	0.44	0.73	0.85	0.63	0.37	0.60	0.45	0.33	0.46	0.22	0.32	0.71	0.95	0.57
clf_nbc_multinomial	0.99	0.63	0.58	0.94	0.99	0.92	0.81	0.75	0.54	0.67	0.96	0.70	0.55	0.49	0.94
clf_qda	0.17	0.92	0.30	0.74	0.26	0.49	0.01	0.87	0.97	0.26	0.53	0.13	0.21	0.58	0.03
clf_svm_1	0.58	0.20	0.71	0.34	0.22	0.89	0.84	0.76	0.42	0.44	0.96	0.83	0.02	0.14	0.80
clf_dt_pruned	0.31	0.24	0.88	0.80	0.23	0.73	0.76	0.36	0.87	0.08	0.10	0.87	0.25	0.40	0.11
clf_lr_multinomial	0.90	0.97	0.63	0.92	0.98	0.53	0.07	0.50	0.27	0.78	0.77	0.10	0.34	0.57	0.55
clf_lr	0.91	0.52	0.93	0.46	0.74	0.23	0.60	0.64	0.76	0.07	0.18	0.84	0.56	0.59	0.02
clf_adaboost	0.35	0.88	0.22	0.36	0.48	0.91	0.67	0.24	0.11	0.98	0.97	0.68	0.61	0.14	0.89

Best overall model

Superior Model for Specific Class

8. Conclusions & Future Work

A new framework has been developed that currently automates static malware analysis, performs feature extraction, and standardizes malware dataset production for AI/ML analysis. 18 ML classification models were trained on a sampling (3100 binaries representing 111 unique malware families) of our entire malware corpus of 80K+ binaries. Random Forests produced the best performance, however, we identify models with superior classification performance across discrete malware families. Future work includes enriching dataset production and AI/ML analysis and completing development of our new ensemble classifier named the **Malware Ensemble Classification Facility**.

PURDUE UNIVERSITY

CERIAS