Escuela Politécnica Nacional

Chunk-based Malware Classifiers for Communication Networks

David Fabián Cevallos Salas

August, 2023

Schedule

1. Introduction
2. Related work
3. Background
4. CIC-MalMem-2022 Dataset
5. Methodology
6. Analysis and Results
7. Conclusion
8. Paper

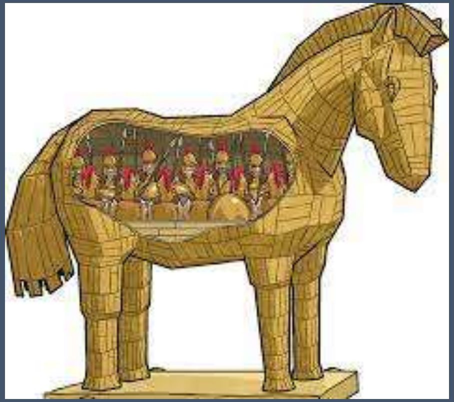
Introduction

- Malware continue to affect to individuals, organizations, countries and communities.
- Mainly three categories of chunk-based malware: Spyware, Ransomware and Trojan Horses.
- Classify malware categories is a difficult task, even more malware families and strains.

Spyware

Ransomware

Trojan Horse



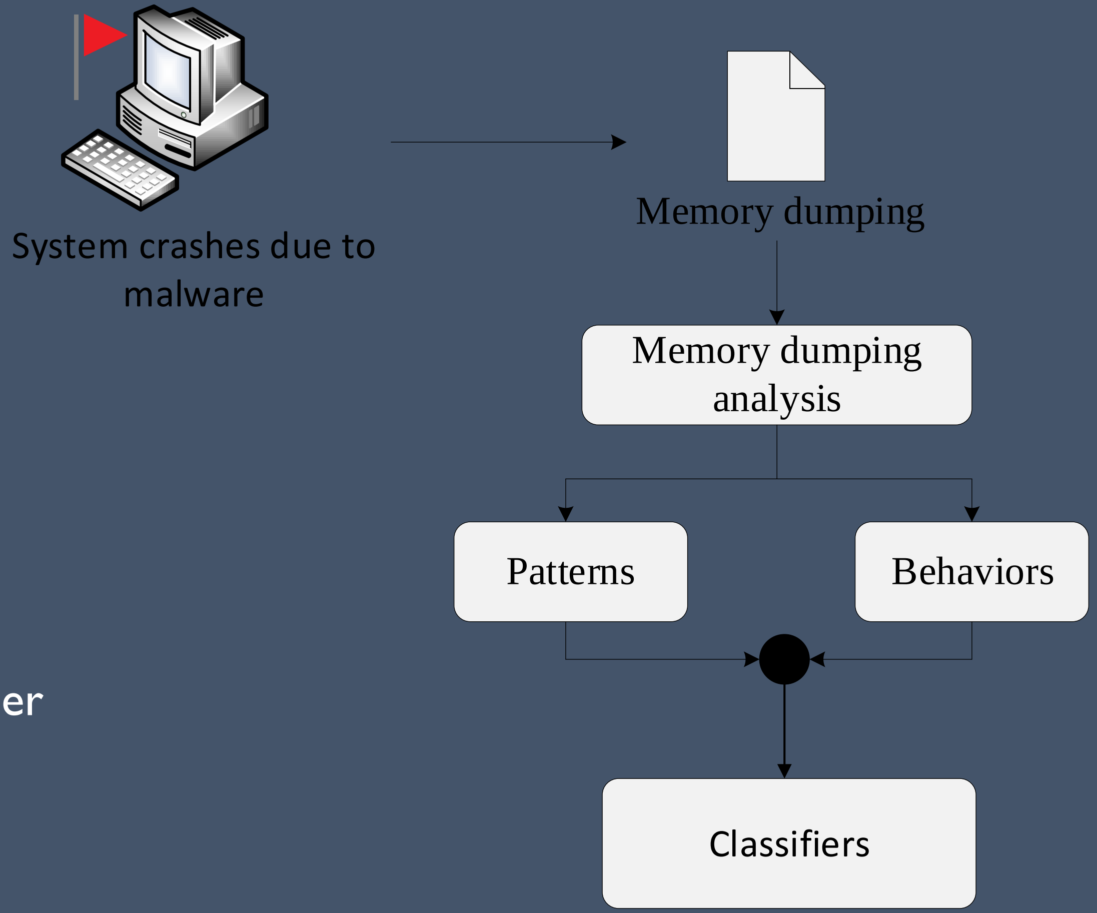
Introduction: Problem

- Security controls act mainly before runtime:
  Advanced Malware Protection (AMP)
  Antivirus
  E-mail protection,
  Educating users,
  Others.

- At runtime is likely imposible to detect and interrupt an exploit.

- This fact certainly produce data leakages.

- Thus, recognizing patterns and behaviours at runtime (such as number of thread, callbacks, handles, mutex, and semaphores) might help to classify **malware families**.



System crashes due to malware

Memory dumping

Memory dumping analysis

Patterns

Behaviors

Classifiers

Related work

AML with LSTM [1]

Chip multiprocessors [2]

Self learning DNN [3]

Heterogeneous Deep Neuronal Network [4]

Bidirectional Long Short-Term Memory NN [5]

Parallel Computing [6]

Refereces

[1] I. Yilmaz, A. Siraj, and D. Ulybyshev, "Improving DGA-Based malicious domain classifiers for malware defense with adversarial machine learning," 4th IEEE Conference on Information and Communication Technology, CICT 2020, 2020

[2] B. Kumar, S. Thakur, K. Basu, M. Fujita, and V. Singh, "A low overhead methodology for validating memory consistency models in chip multiprocessors," Proceedings - 33rd International Conference on VLSI Design, VLSID 2020 - Held concurrently with 19th International Conference on Embedded Systems, pp. 101–106, 2020

[3] J. Yang and Y. Guo, "AEFETA: Encrypted traffic classification framework based on self-learning of feature," 2021 IEEE 6th International Conference on Intelligent Computing and Signal Processing, ICSP 2021, no. Icsp, pp. 876–880, 2021.

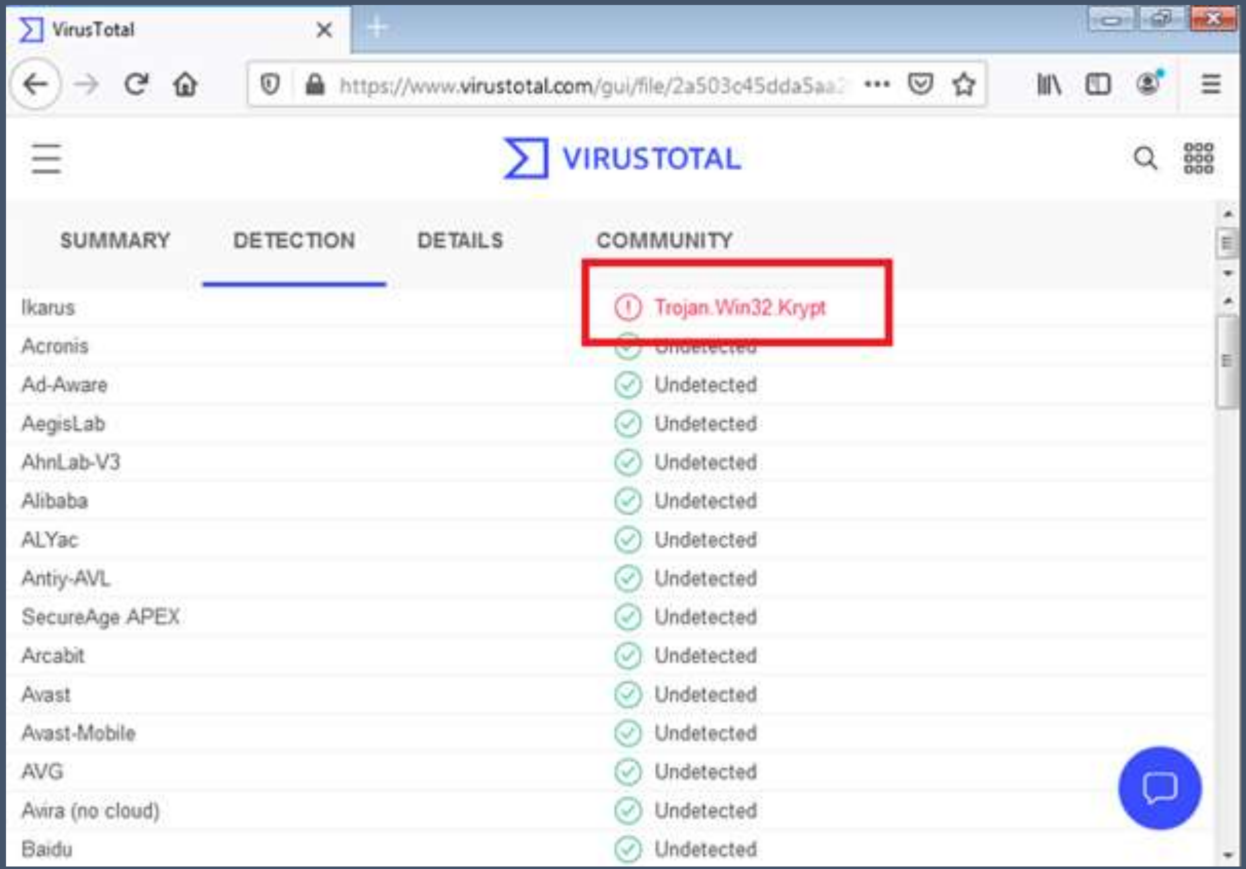
[4] L. Yang, G. Liu, Y. Dai, J. Wang, and J. Zhai, "Detecting stealthy domain generation algorithms using heterogeneous deep neural network framework," IEEE Access, vol. 8, pp. 82 876–82 889, 2020.

[5] Girinoto, H. Setiawan, P. A. W. Putro, and Y. R. Pramadi, "Comparison of LSTM Architecture for Malware Classification," Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020, pp. 93–97, 2020.

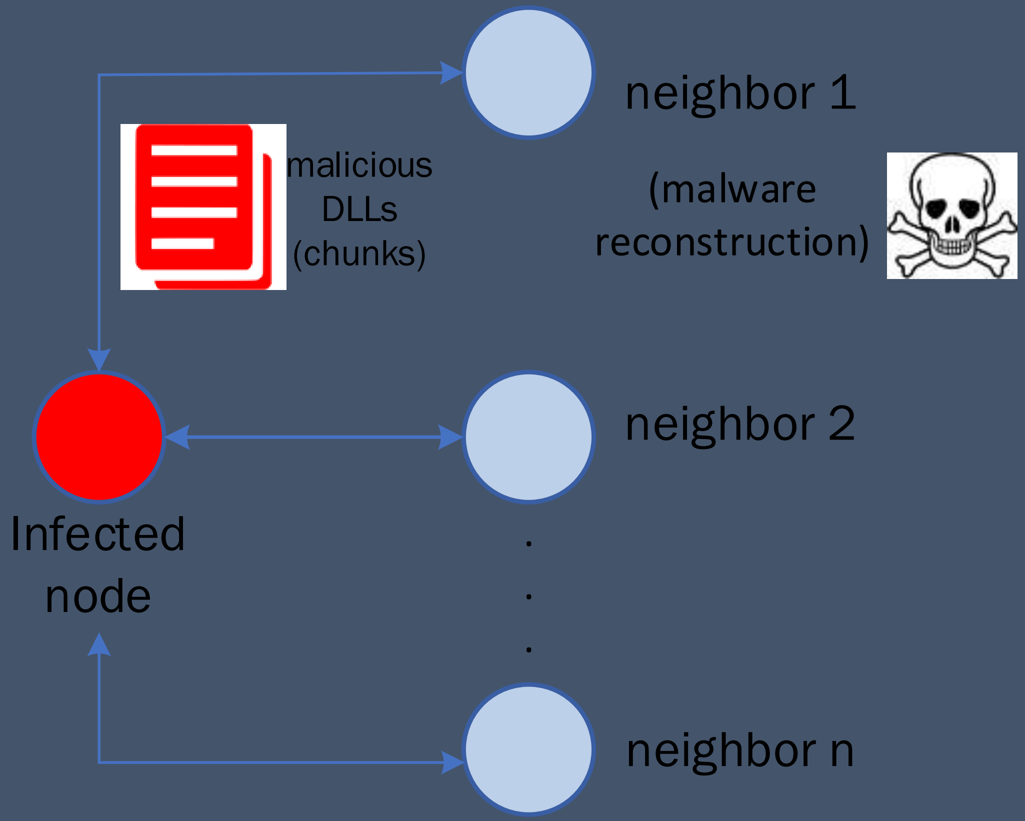
[6] D. Appello, P. Bernardi, A. Calabrese, S. Littardi, G. Pollaccia, S. Quer, V. Tancorre, and R. Ugioli, "Accelerated Analysis of Simulation Dumps through Parallelization on Multicore Architectures," Proceedings - 2021 24th International Symposium on Design and Diagnostics of Electronic Circuits and Systems, DDECS 2021, pp. 69–74, 2021

Background

- Chunk-based malware is able to transmit small pieces of information.
- Reconstruction takes place at the target.
- Once one node is infected, it tries to replicate to another node.

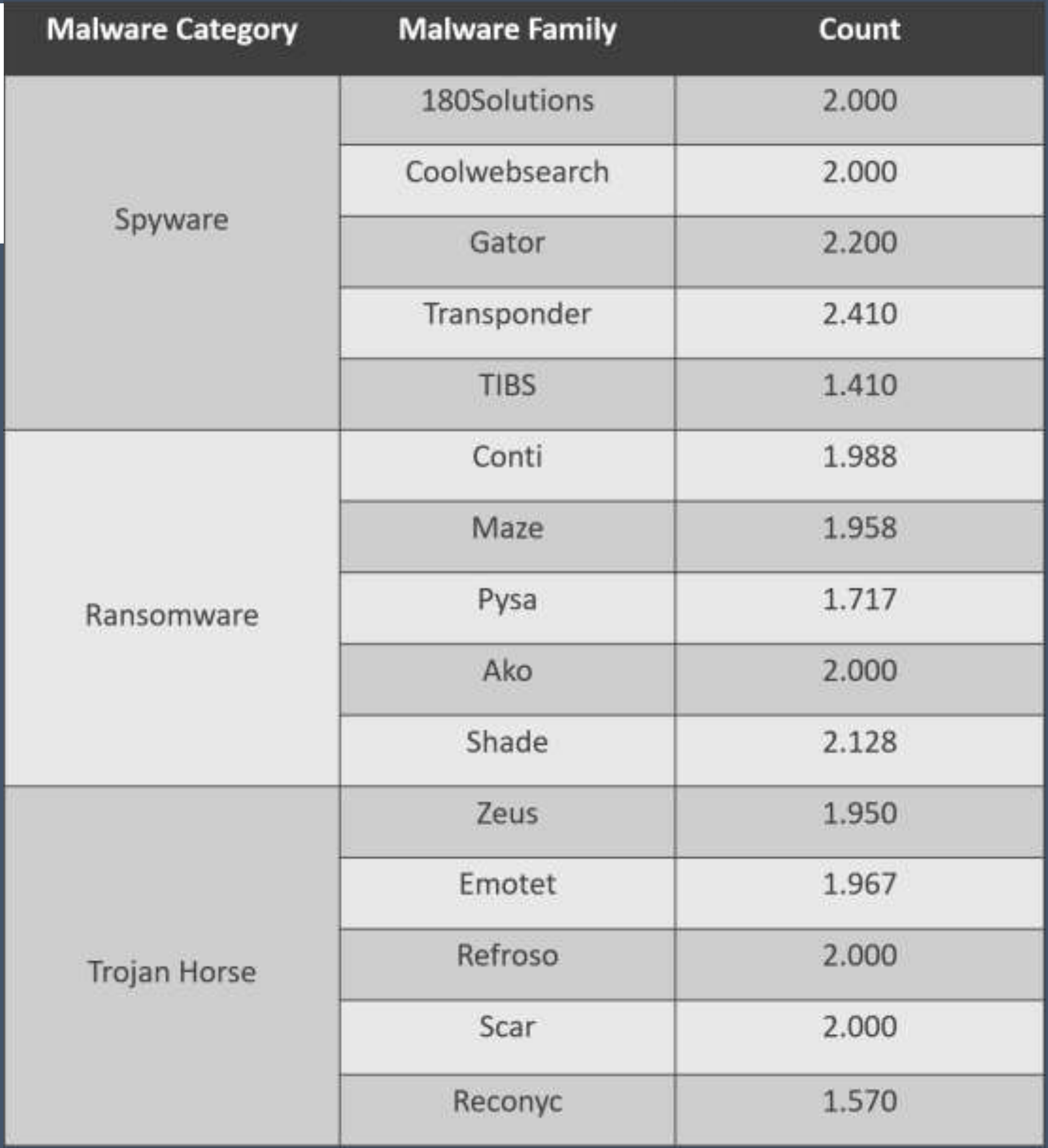


Ikarus	Trojan.Win32.Krypt
Acronis	Undetected
Ad-Aware	Undetected
AegisLab	Undetected
AhnLab-V3	Undetected
Alibaba	Undetected
ALYac	Undetected
Antiy-AVL	Undetected
SecureAge APEX	Undetected
Arcabit	Undetected
Avast	Undetected
Avast-Mobile	Undetected
AVG	Undetected
Avira (no cloud)	Undetected
Baidu	Undetected



Methodology: Dataset

- Created by the Canadian Institute for Cybersecurity of the University of New Brunswick (CICUNB).
- 29.298 benign processes and 29.298 malicious processes
- 55 features including (Number of *threads*, number of *handles*, *callbacks*, *mutex*, *semaphores*, etc)



Malware Category	Malware Family	Count
Spyware	180Solutions	2.000
	Coolwebsearch	2.000
	Gator	2.200
	Transponder	2.410
	TIBS	1.410
Ransomware	Conti	1.988
	Maze	1.958
	Pysa	1.717
	Ako	2.000
	Shade	2.128
Trojan Horse	Zeus	1.950
	Emotet	1.967
	Refroso	2.000
	Scar	2.000
	Reconyc	1.570

Methodology

1. Data acquisition

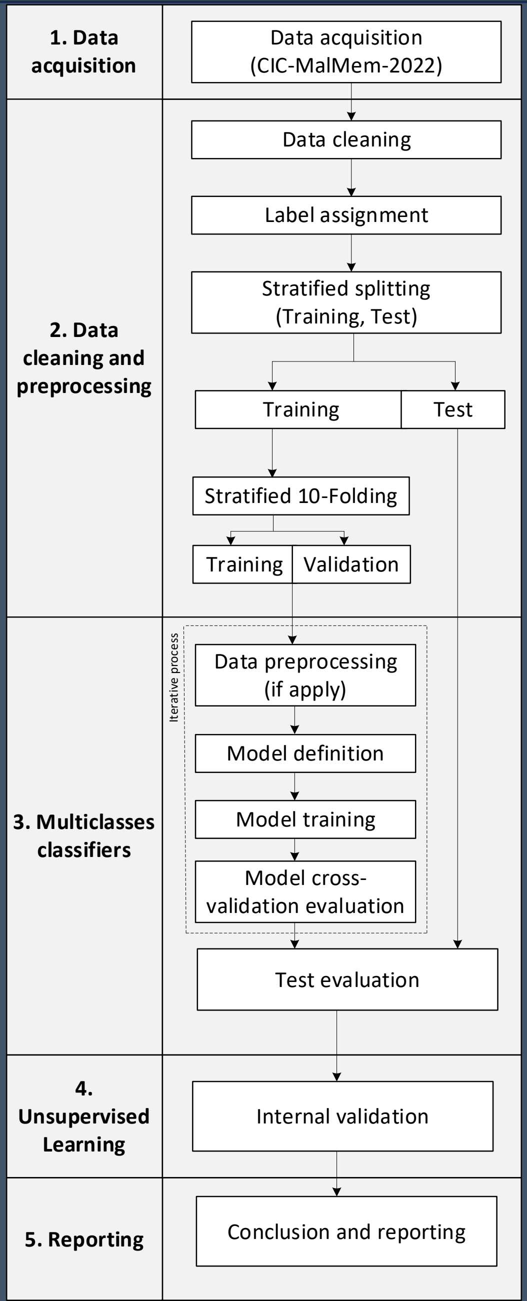
2. Data cleaning and preprocessing

3. Multiclasses classifiers

4. Unsupervised Learning

5. Conclusion and reporting

Methodology



1. Data acquisition

Data acquisition (CIC-MalMem-2022)

2. Data cleaning and preprocessing

Data cleaning

Label assignment

Stratified splitting (Training, Test)

Training

Test

Stratified 10-Folding

Training

Validation

3. Multiclasses classifiers

Iterative process

Data preprocessing (if apply)

Model definition

Model training

Model cross-validation evaluation

Test evaluation

4. Unsupervised Learning

Internal validation

5. Reporting

Conclusion and reporting

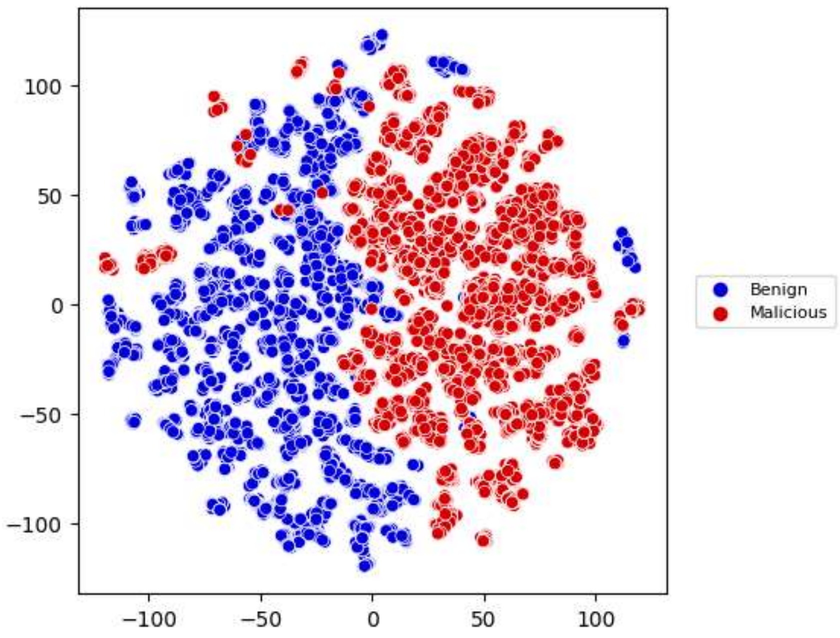
General techniques:

- Stratified splitting
- Stratified 10-Folds
- Training, Cross validation and Evaluation datasets

Machine Learning Algorithms:

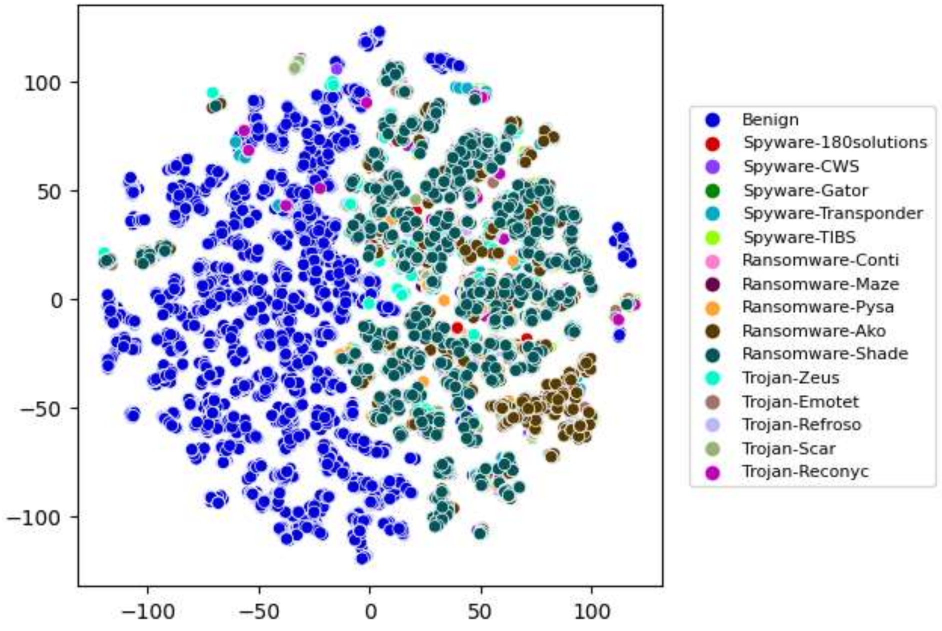
- Support Vector Machine Classifier (SVC)
- Decision tree (DT)
- Random Fores (RF)
- Linear Discriminant Analysis (LDA)
- Naive Bayes (NB)
- Deep Neuronal Network (DNN)
- Adaptive Neuro Fuzzy Inference System (ANFIS)
- SMOTED Deep Neuronal Network (S-DNN)

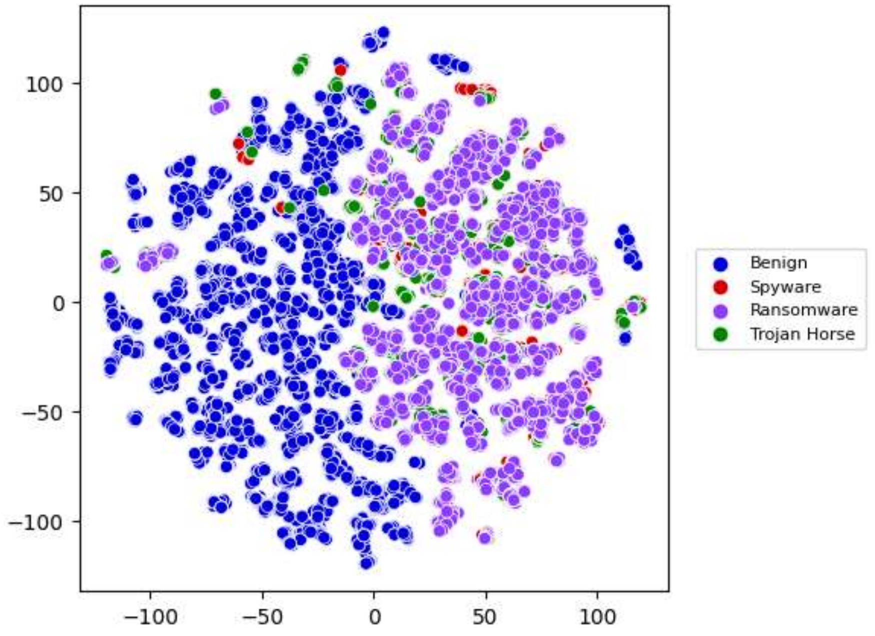
Analysis and Results



Binary classification

Families classification





Categories classification

Analysis and Results

Cross-validation average metrics

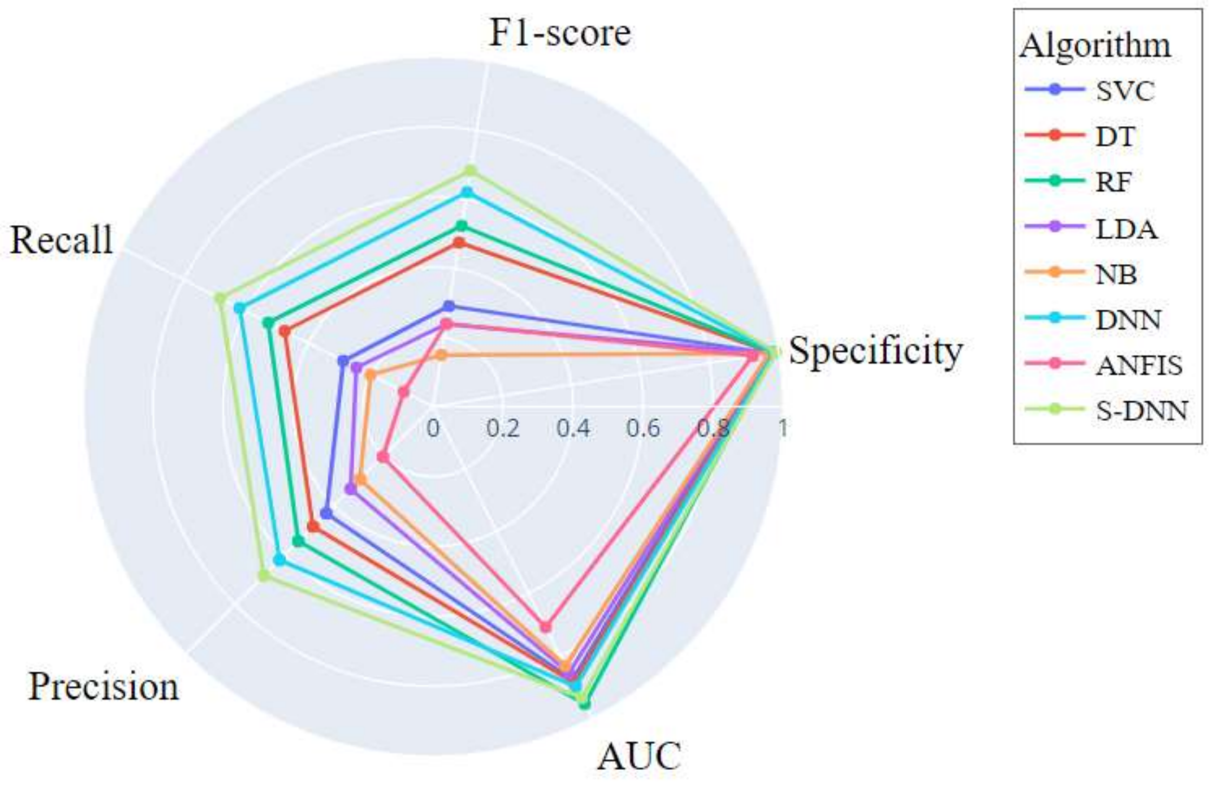
Algorithm	Precision	Recall	F1-score	Accuracy	Error	Specificity	False Positive Rate	AUC
SVC	0,4315	0,2879	0,2908	0,6215	0,3785	0,9755	0,0245	0,8760
DT	0,4850	0,4759	0,4749	0,7200	0,2800	0,9819	0,0181	0,8855
RF	0,5443	0,5284	0,5229	0,7484	0,2516	0,9837	0,0163	0,9541
LDA	0,3325	0,2457	0,2387	0,5970	0,4030	0,9738	0,00262	0,8547
NB	0,2944	0,2006	0,1492	0,5647	0,4352	0,9718	0,0282	0,8321
DNN	0,6202	0,6202	0,6202	0,6202	0,3798	0,9867	0,0133	0,8981
ANFIS	0,2031	0,0941	0,2406	0,3115	0,6885	0,9246	0,0754	0,7078
S-DNN	0,6830	0,6830	0,6830	0,6830	0,3170	0,9921	0,0079	0,9344

Test metrics

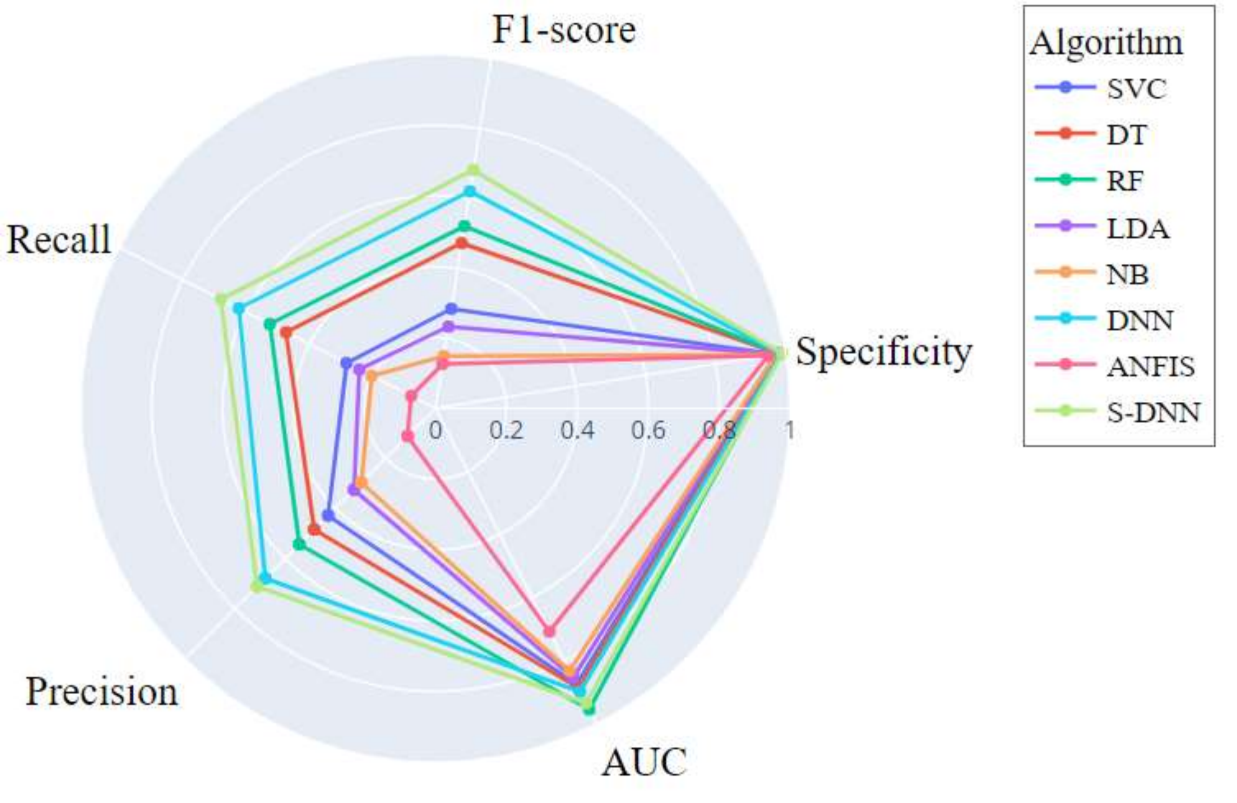
Algorithm	Precision	Recall	F1-score	Accuracy	Error	Specificity	False Positive Rate	AUC
SVC	0,4277	0,2814	0,2845	0,6184	0,3819	0,9752	0,0248	0,8752
DT	0,4836	0,4729	0,4723	0,7189	0,2811	0,9818	0,0182	0,8856
RF	0,5429	0,5241	0,5209	0,7464	0,2536	0,9836	0,0164	0,9546
LDA	0,3251	0,2411	0,2337	0,5949	0,4051	0,9737	0,0263	0,8541
NB	0,2956	0,2016	0,1491	0,5660	0,4340	0,9719	0,0281	0,8319
DNN	0,6788	0,6220	0,6201	0,6220	0,3780	0,9853	0,0147	0,8981
ANFIS	0,1110	0,0766	0,1266	0,1918	0,8082	0,9494	0,0506	0,7077
S-DNN	0,7115	0,6788	0,6815	0,6788	0,3212	0,9901	0,0099	0,9344

Analysis and Results

CROSS-VALIDATION METRICS

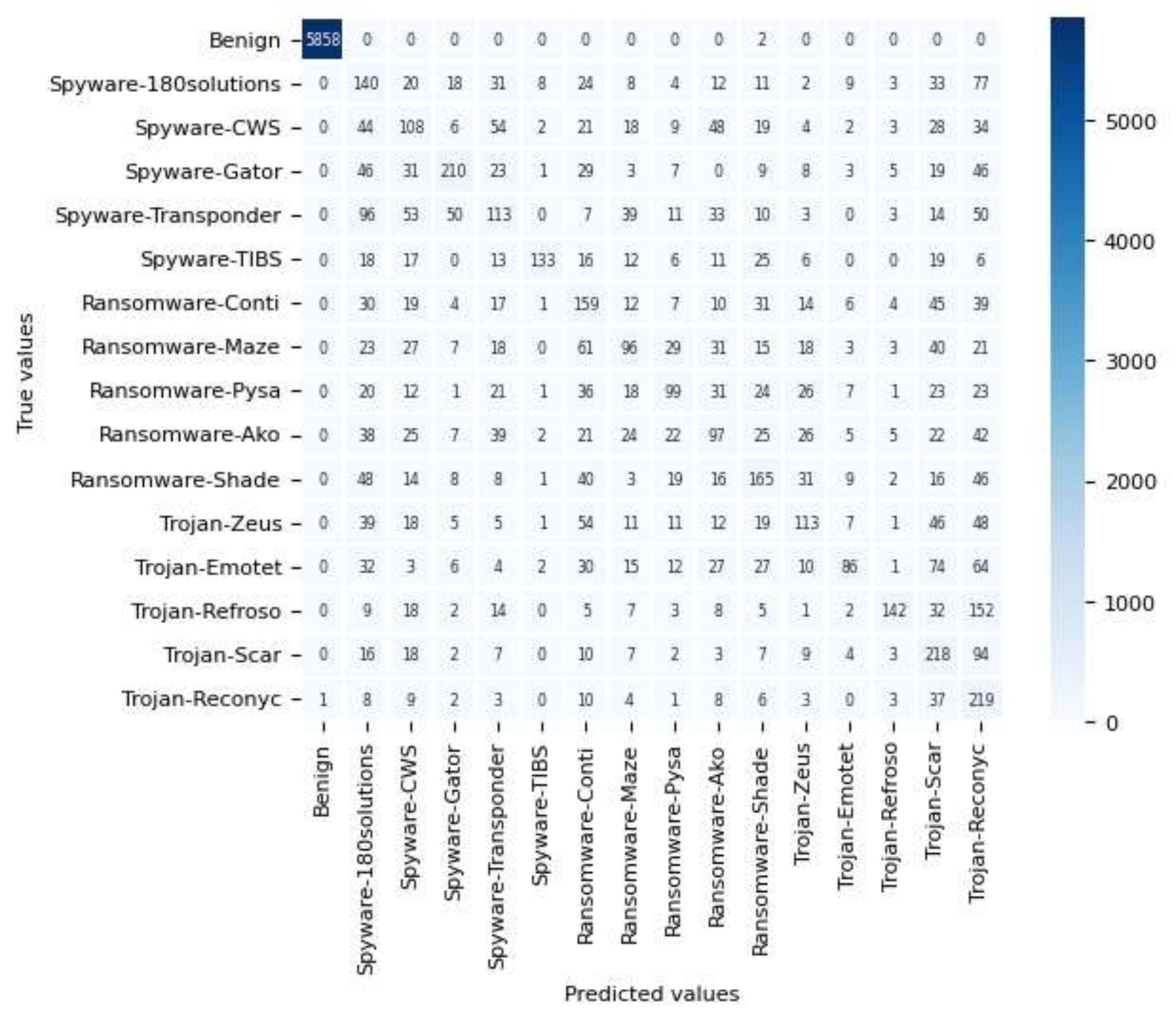


TEST METRICS



Analysis and Results

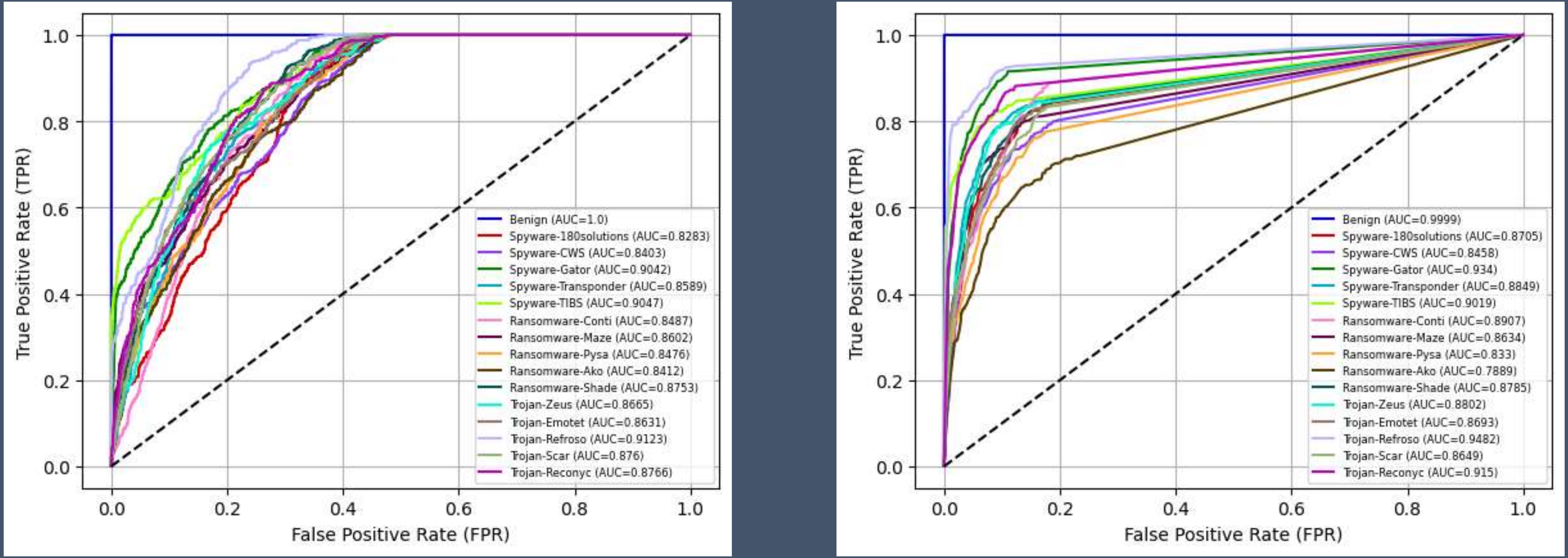
TEST S-DNN CONFUSION MATRIX



	Benign	Spyware-180solutions	Spyware-CWS	Spyware-Gator	Spyware-Transponder	Spyware-TIBS	Ransomware-Conti	Ransomware-Maze	Ransomware-Pysa	Ransomware-Ako	Ransomware-Shade	Trojan-Zeus	Trojan-Emotet	Trojan-Refroso	Trojan-Scar	Trojan-Reconyc
Benign	5858	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Spyware-180solutions	0	140	20	18	31	8	24	8	4	12	11	2	9	3	33	77
Spyware-CWS	0	44	108	6	54	2	21	18	9	48	19	4	2	3	28	34
Spyware-Gator	0	46	31	210	23	1	29	3	7	0	9	8	3	5	19	46
Spyware-Transponder	0	96	53	50	113	0	7	39	11	33	10	3	0	3	14	50
Spyware-TIBS	0	18	17	0	13	133	16	12	6	11	25	6	0	0	19	6
Ransomware-Conti	0	30	19	4	17	1	159	12	7	10	31	14	6	4	45	39
Ransomware-Maze	0	23	27	7	18	0	61	96	29	31	15	18	3	3	40	21
Ransomware-Pysa	0	20	12	1	21	1	36	18	99	31	24	26	7	1	23	23
Ransomware-Ako	0	38	25	7	39	2	21	24	22	97	25	26	5	5	22	42
Ransomware-Shade	0	48	14	8	8	1	40	3	19	16	165	31	9	2	16	46
Trojan-Zeus	0	39	18	5	5	1	54	11	11	12	19	113	7	1	46	48
Trojan-Emotet	0	32	3	6	4	2	30	15	12	27	27	10	86	1	74	64
Trojan-Refroso	0	9	18	2	14	0	5	7	3	8	5	1	2	142	32	152
Trojan-Scar	0	16	18	2	7	0	10	7	2	3	7	9	4	3	218	94
Trojan-Reconyc	1	8	9	2	3	0	10	4	1	8	6	3	0	3	37	219

Analysis and Results

TEST ROC CURVES

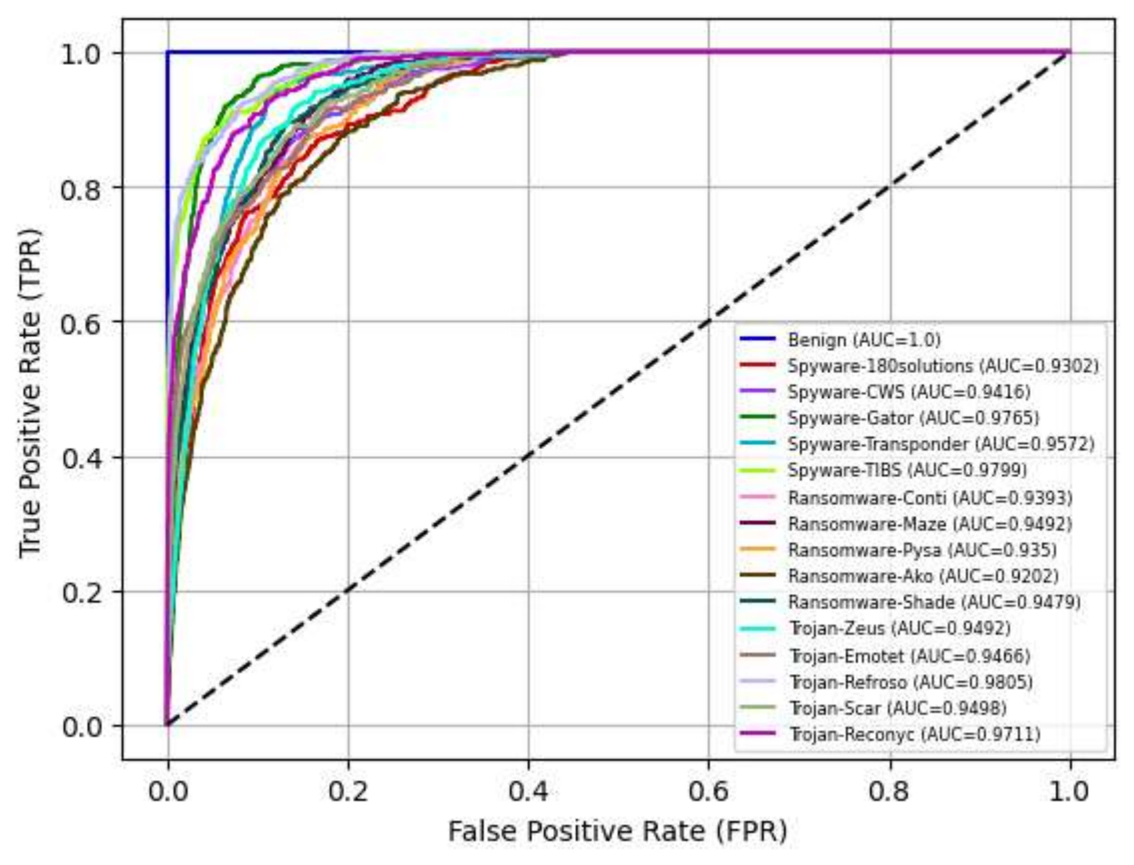


SVC

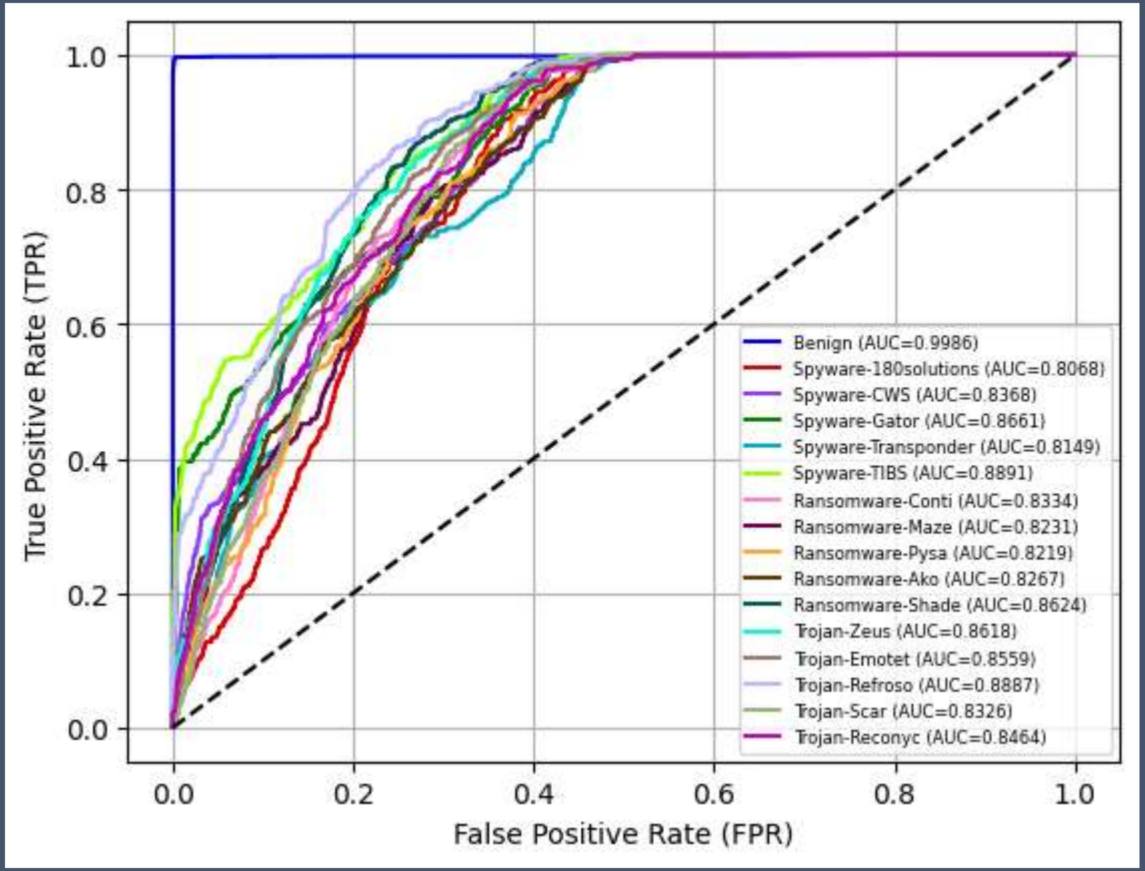
DT

Analysis and Results

RF

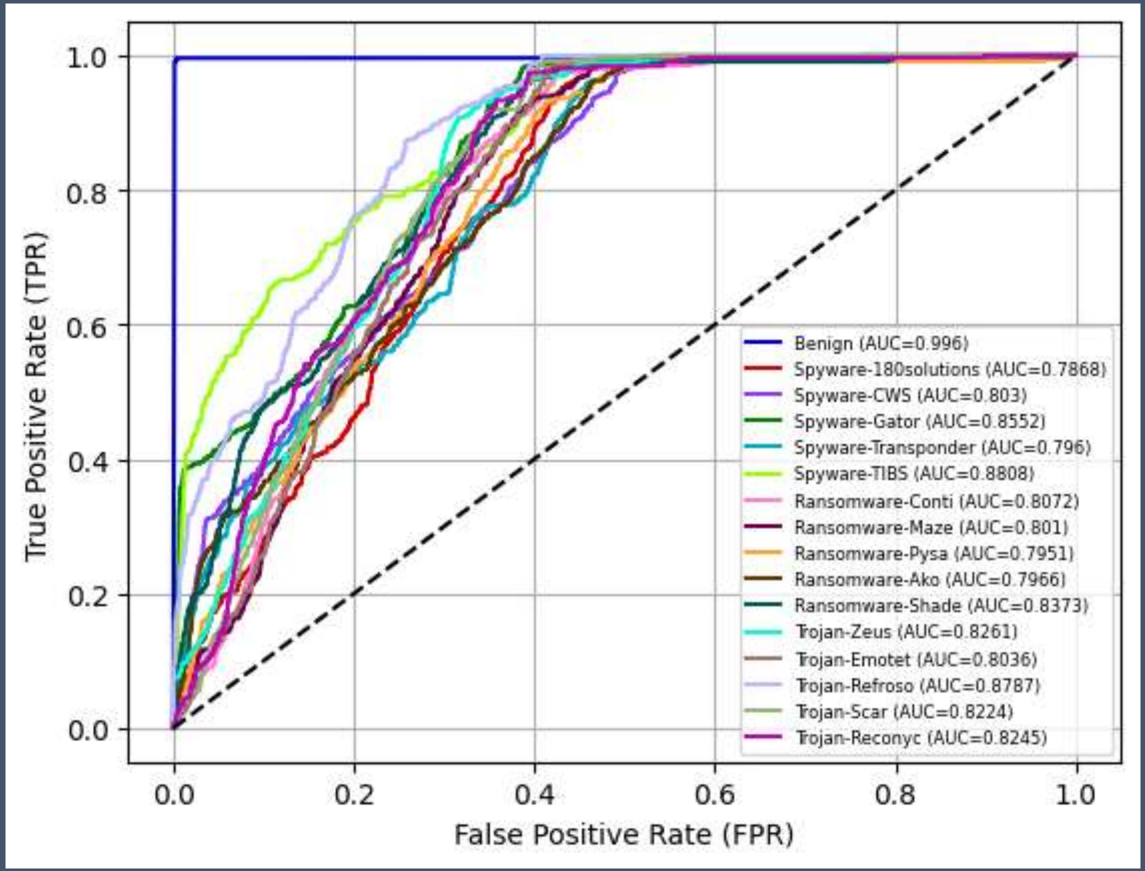


LDA

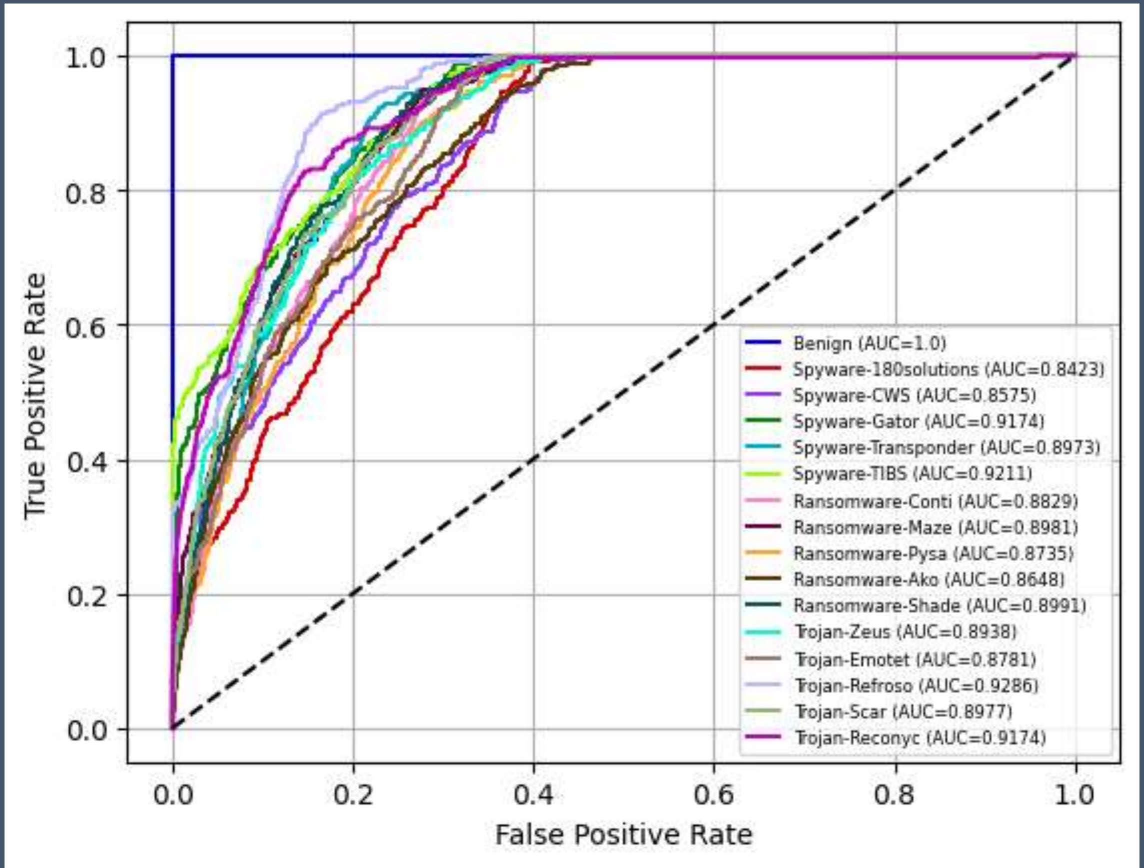


Analysis and Results

NB

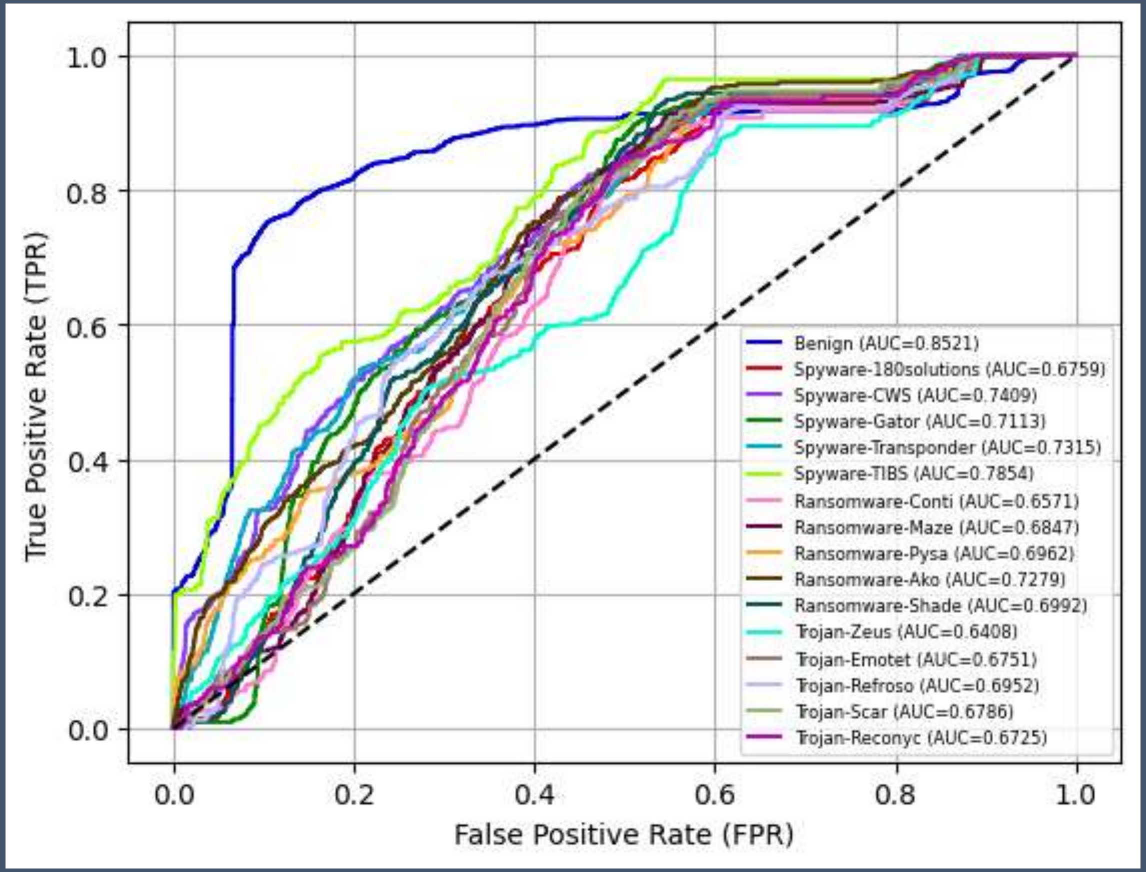


DNN

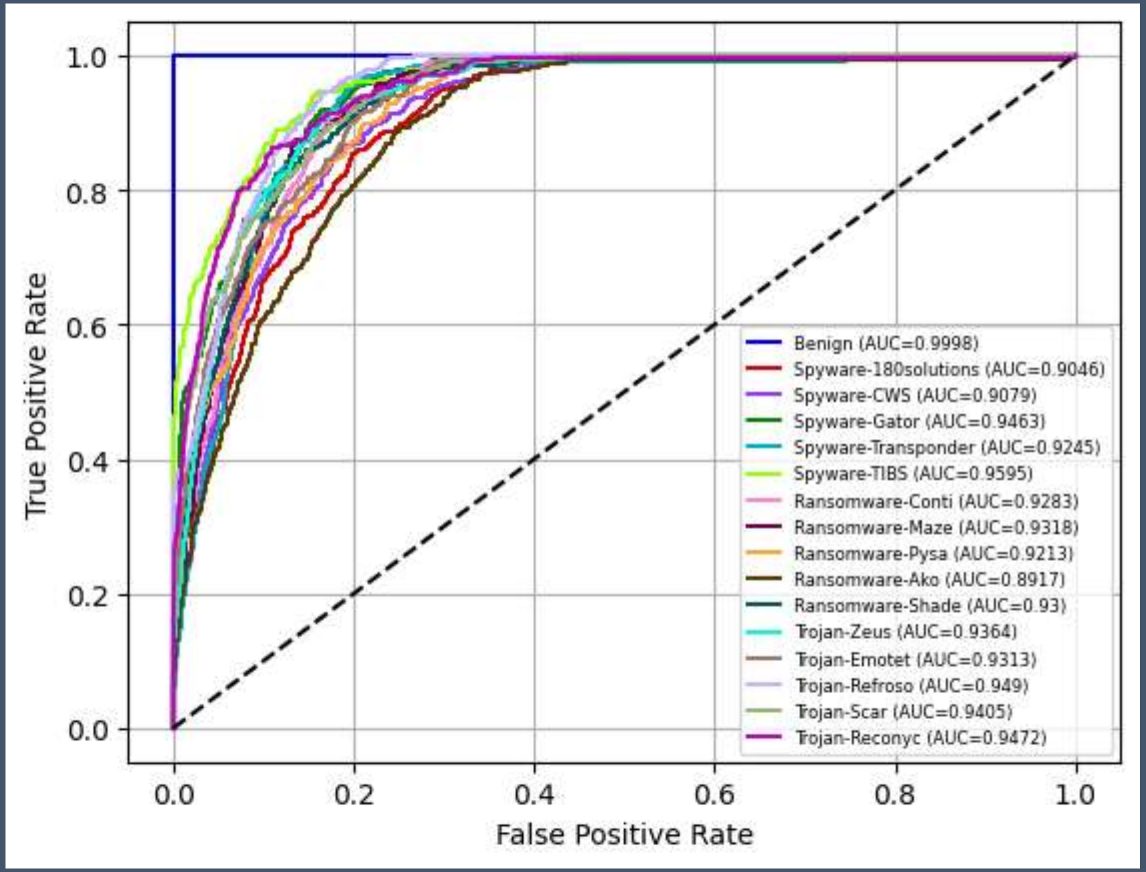


Analysis and Results

ANFIS



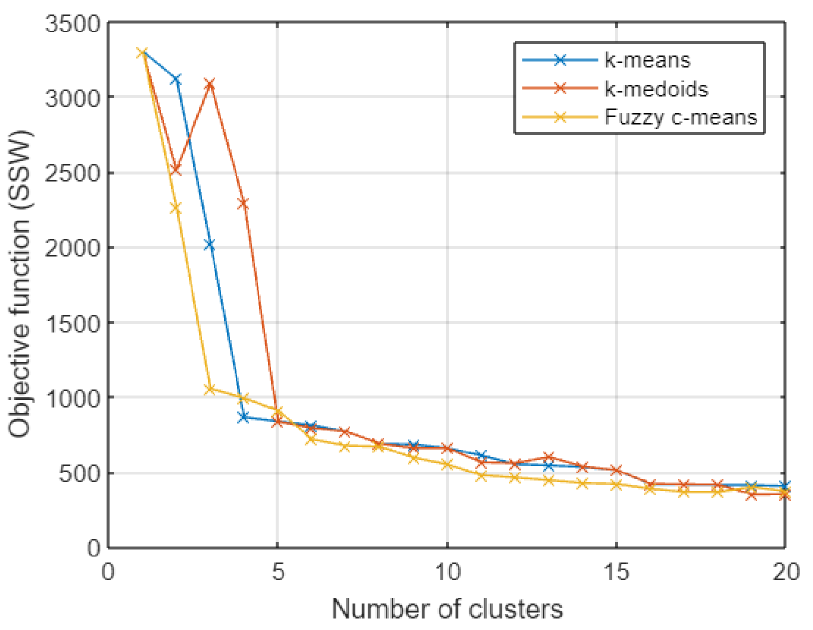
S-DNN



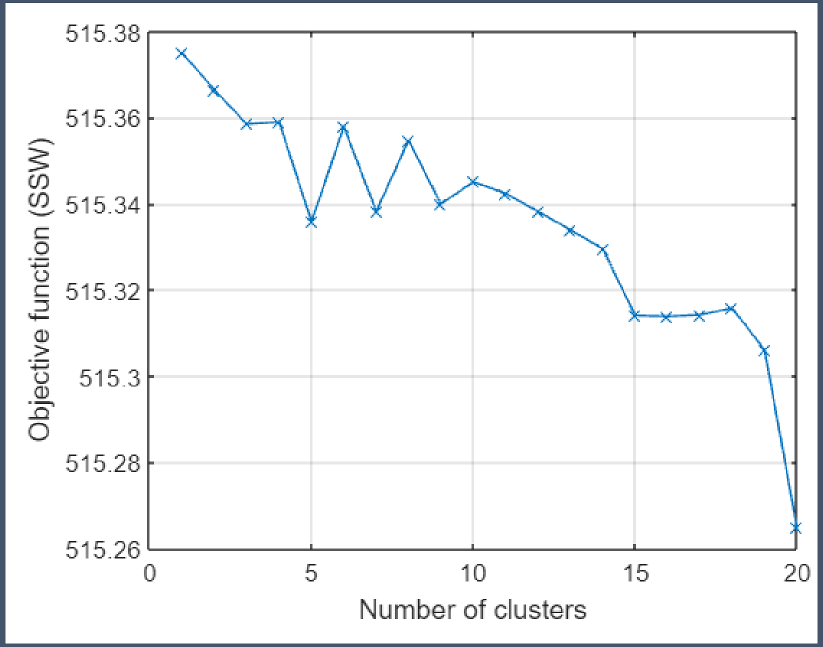
Analysis and Results

Unsupervised Learning

Elbow diagram



Hierarchical clustering



Parameters values

k-means:
k = 16

K-medoids:
k = 16

DBSCAN:
min_neighbors = 3
epsilon = 2.049,1

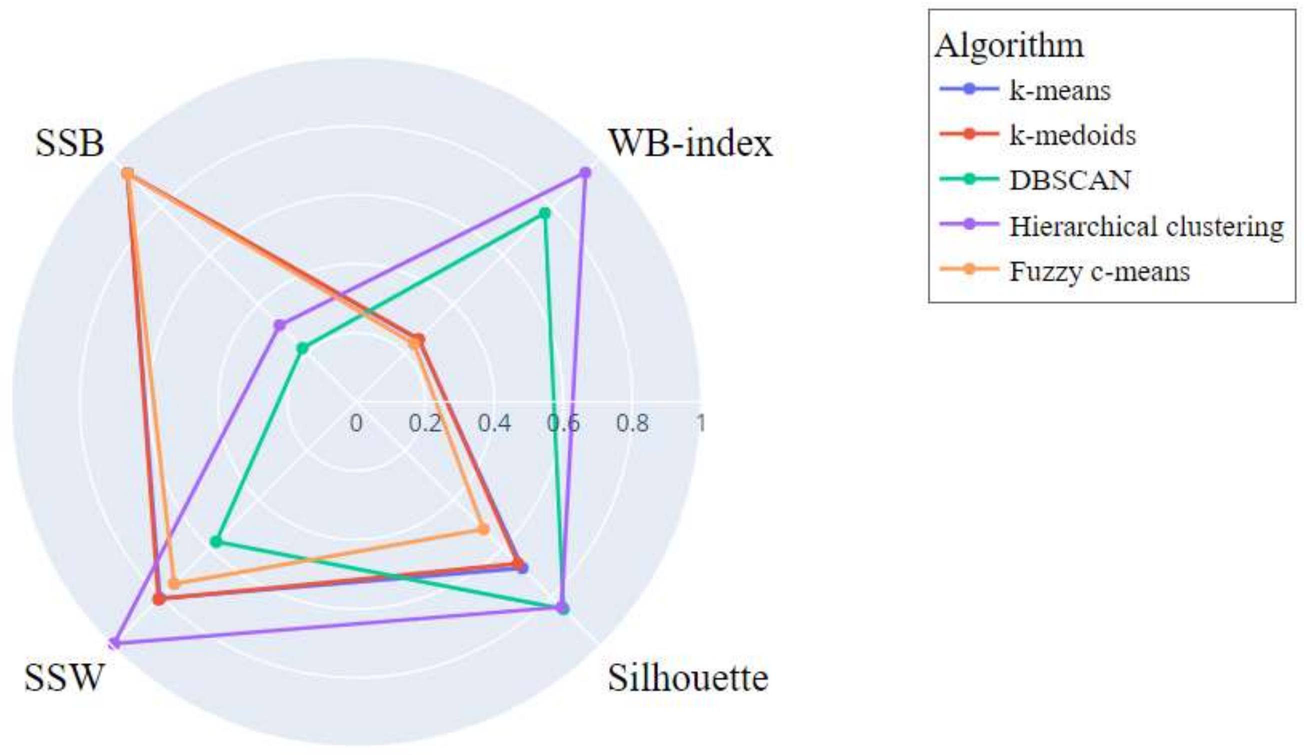
Hierarchical clustering:
Nro. clusters = 16
Single linkage

Fuzzy c-means:
c = 16

Analysis and Results

Unsupervised Learning

Algorithm	SSW	SSB	WB-index	Silhouette
k-means	418,5362	3.276,6000	2,0437	0,6815
k-medoids	420,7788	3.275,8000	2,0552	0,6631
DBSCAN	297,8611	770,2884	6,1870	0,8945
Hierarchical clustering	515,3139	1.097,8000	7,5107	0,8410
Fuzzy c-means	387,7675	3.273,2000	1,8955	0,5233



Conclusion

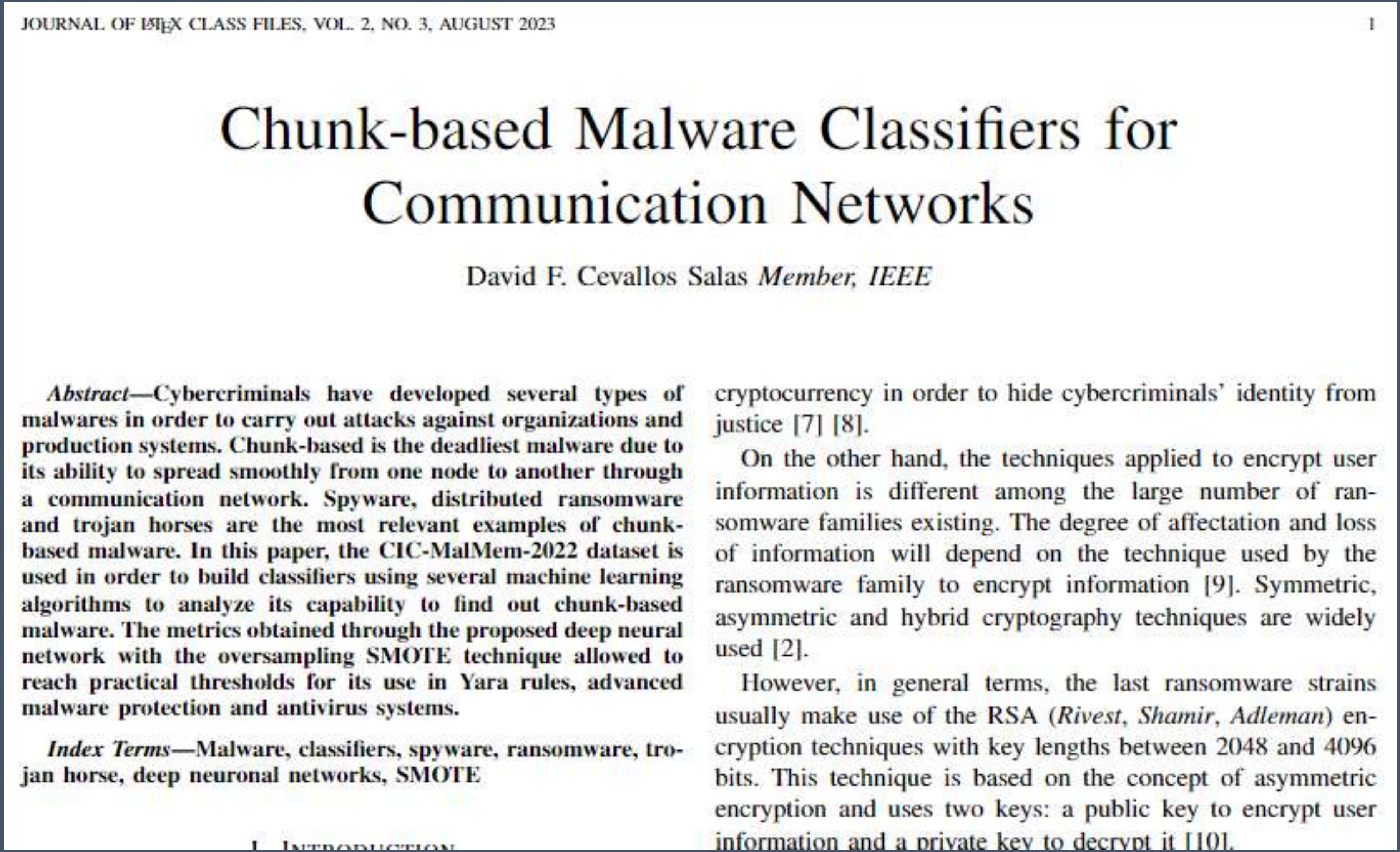
- Deep neural networks comprise a much more robust technique for the identification and classification of chunk-based malware families in comparison to traditional machine learning algorithms such as SVC, decision trees, random forest, among others

- Deep neural networks achieved precision and recall values balanced and almost equal.

- Oversampling helps substantially to reach better metrics in chunk-based malware classification.

- CIC-MalMem-2022 is able to classify malware, but can be improved.

Paper



JOURNAL OF LaTeX CLASS FILES, VOL. 2, NO. 3, AUGUST 2023 1

Chunk-based Malware Classifiers for Communication Networks

David F. Cevallos Salas *Member, IEEE*

Abstract—**Cybercriminals have developed several types of malwares in order to carry out attacks against organizations and production systems. Chunk-based is the deadliest malware due to its ability to spread smoothly from one node to another through a communication network. Spyware, distributed ransomware and trojan horses are the most relevant examples of chunk-based malware. In this paper, the CIC-MalMem-2022 dataset is used in order to build classifiers using several machine learning algorithms to analyze its capability to find out chunk-based malware. The metrics obtained through the proposed deep neural network with the oversampling SMOTE technique allowed to reach practical thresholds for its use in Yara rules, advanced malware protection and antivirus systems.**

Index Terms—**Malware, classifiers, spyware, ransomware, trojan horse, deep neuronal networks, SMOTE**

I. Introduction

cryptocurrency in order to hide cybercriminals' identity from justice [7] [8].

On the other hand, the techniques applied to encrypt user information is different among the large number of ransomware families existing. The degree of affectation and loss of information will depend on the technique used by the ransomware family to encrypt information [9]. Symmetric, asymmetric and hybrid cryptography techniques are widely used [2].

However, in general terms, the last ransomware strains usually make use of the RSA (*Rivest*, *Shamir*, *Adleman*) encryption techniques with key lengths between 2048 and 4096 bits. This technique is based on the concept of asymmetric encryption and uses two keys: a public key to encrypt user information and a private key to decrypt it [10].

Thanks…

David Fabián Cevallos Salas
david.cevallos03@epn.edu.ec