

Humans vs. Machines in Malware Classification

2023/08/09 – USENIX Security '23
Anaheim, California, USA 🇺🇸



← *Simone Aonzo*
Yufei Han →
← Alessandro Mantovani
← Davide Balzarotti



Malware Analysis(it)s



- Malware Analysts are security experts of the Malware Analysis process
 - Examine malicious software for **classification** – Benign or Malicious?
 - An intrinsic part of their work is to read “*Sandbox Reports*”
- Problem: humans do not scale 🙈
- Machine Learning 🤖 to the rescue!
 - Vast amount of research on ML-based malware classification
 - Humans teach machines which features they should evaluate
 - Problem: models cannot reach beyond the training data



What information guides human and machine decision-making when classifying samples (by looking at sandbox reports)?

... maybe we can learn something from each other!

- *Experts* 🐒 vs. *Novices* 🐒
- *Senior Experts* 🐒 vs. *Young Experts* 🐒
- *Machines* 🤖 vs. *Machines* 🤖
- *Humans* 🐒 vs. *Machines* 🤖



Focus on Windows Malware – Portable Executable (PE) samples

Participants

- 110 humans 🐒
 - 38 Experts
 - Renowned cybersecurity companies + Academic researchers
 - 7 of them with ≥ 9 years of experience
 - 72 Novices – attended at least a malware analysis course
 - Students + Beginner CTF players
- 2 state-of-the-art Machine Learning algorithms 🤖
 - Random Forests (500 trees)
 - Convolutional Neural Network (4-layered architecture)



Experiment Setup – Humans



We designed an web-based game: “*Detect Me If You Can!*” [DMIYC]

- Design elements
 - Leaderboard: rank players according to their performances
 - Points: numerically represent a player’s outcome
- Participants have to correctly classify 20 VirusTotal reports
 - Using **as few features as possible**
 - ⇒ Players have to “buy” each feature

Time Left: 58:02

Feature added: 3

Static Details

Basic Properties

VT Labels

VT Submission History

Signature

PE File Info

Header Metadata

Sections

Imports

Resources

Strings

Dynamic Behavior

Network

Processes

Services

Registry

Mutexes

File System

Runtime DLLs

Report

Basic Properties

File Size	542.1 KiB (555120 bytes)
TrID	Win32 Executable MS Visual C++ (generic) (48.0%) Microsoft Visual C++ compiled executable (generic) (25.4%) Win32 Dynamic Link Library (generic) (10.1%) Win32 Executable (generic) (6.9%) OS/2 Executable (generic) (3.1%)
Magic	PE32 executable for MS Windows (GUI) Intel 80386 32-bit

Sections

Name	Virtual Address	Virtual Size	Raw Size	Entropy
.text	4096	105892	105984	6.03
.data	110592	40	512	0.31
.rdata	114688	6400	6656	5.52
.eh_frame	122880	11388	11776	4.93
.bss	135168	2748	0	0.00
.idata	139264	3084	3584	4.71
.CRT	143360	24	512	0.11
.tls	147456	32	512	0.22
.rsrc	151552	371280	371712	3.00

Network

UDP

- <MACHINE_DNS_SERVER>:53

DNS

Hostname	Ip
71.t.online.io	212.83.161.135

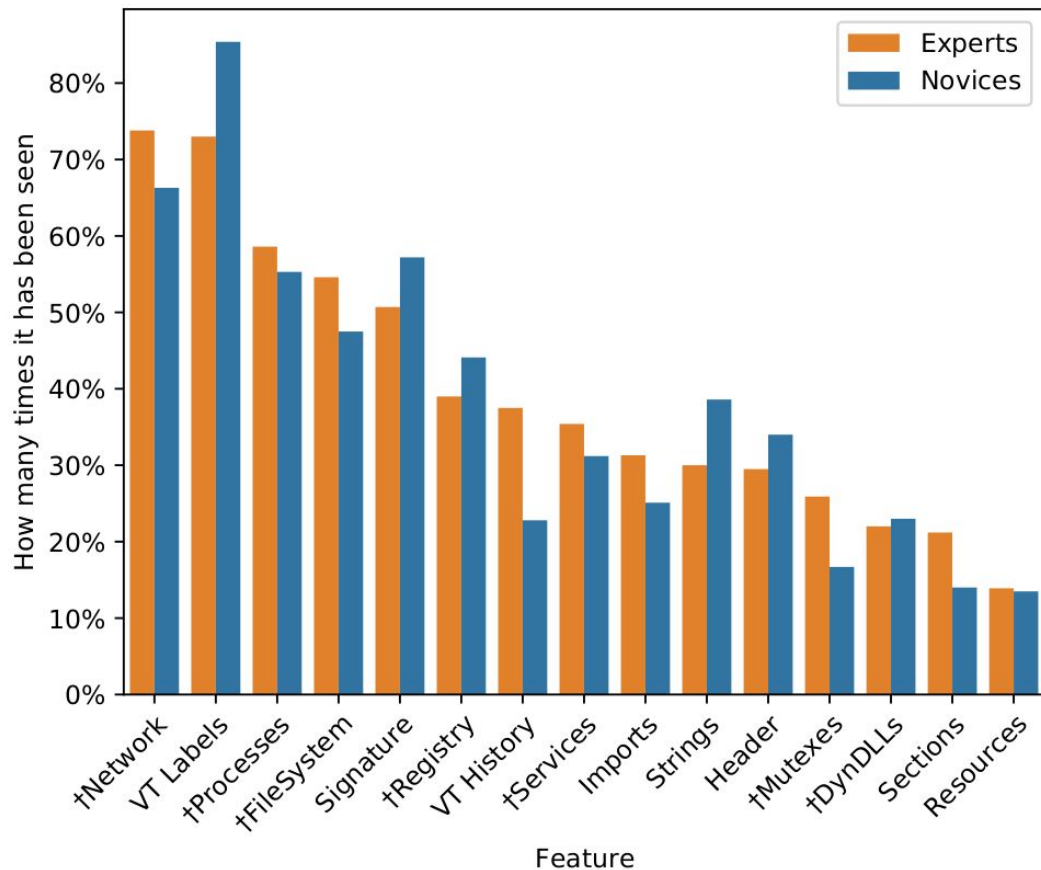
View of the results

Metric	Experts Novices	Min	Max	Avg	Std	Median
Time	E	7:48	56:48	29:04	08:53	26:51
	N	8:14	59:58	44:31	10:05	46:32
Score	E	2310	5339	4103	742	4329
	N	1072	6042	3072	1054	2991
Right Answers	E	13	19	16.1	1.4	16
	N	8	19	13.7	2.4	14
Total Used Features	E	42	165	82.0	35.1	70
	N	37	146	81.7	27.5	68.5
Unique Used Features	E	7	16	13.4	2.6	14
	N	7	16	14.1	2.1	15

Statistically-significant differences (Welch's t-test) between **Experts/Novices**

1. **Time** needed to complete the game ✓
2. Final **Score** ✓
3. Number of **Right Answers** ✓
4. ... features? 🤔

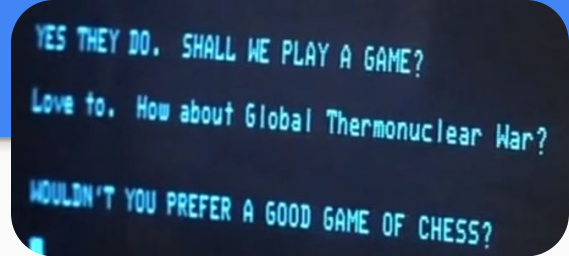
Feature Ranking by Humans




Most used top 5 features

	All	Correct	Misclassified
Experts	†Network	†Network	†Network
	VT labels	VT labels	VT labels
	†Processes	†Processes	†Processes
	†FileSystem	†FileSystem	†FileSystem
	Signature	Signature	Signature
Novices	VT labels	VT labels	VT labels
	†Network	†Network	†Network
	Signature	Signature	†Processes
	†Processes	†Processes	Signature
	†FileSystem	†FileSystem	†FileSystem

Experiment setup – Machines



- Balanced dataset of 21,944 VirusTotal reports
- “*Not Yet Another Classifier*” ⇒ State-of-the-art encoding techniques
- 5-fold cross-validation to derive averaged AUC-ROC scores
 - Training 80% – Testing 20%
 - High classification accuracy : 0.9962 for RF and 0.9950 for CNN
- ... and finally they classified the DMIYC reports, but
 - Machines had the “all feature advantage”
 - VirusTotal features excluded

Humans 🐒 vs. 🤖 Machines



- Human Experts: 16/20 (avg = median)
- Machines
 - Random Forest: 17/20
 - Convolution Neural Network: 16/20
- Both ML algorithms misclassified the same two samples
 - They were not among the most difficult samples for humans
- In general, the misclassified samples by machines and humans are different

Humans vs. Machines – Feature Ranking

We adopted SHAP as a model-agnostic explanation tool

#	RF	CNN	Expert Humans
1	Resources	Resources	†Network
2	†Services	Sections	†Processes
3	Header Metadata	†Network	†FileSystem
4	†Network	†Runtime DLLs	Signature
5	Signature	Header Metadata	†Registry
6	†Runtime DLLs	Signature	†Services
7	Strings	†Services	Imports
8	Sections	†FileSystem	Strings
9	Imports	Strings	Header Metadata
10	†Mutexes	†Registry	†Mutexes
11	†Registry	†Mutexes	†Runtime DLLs
12	†FileSystem	Imports	Sections
13	†Processes	†Processes	Resources

Takeaways (1/2)



- Experts and Novices base their decisions on the same set of features
 - But also Senior Experts
- During goodwill classification
 - Experts used more features and Novices make the majority of mistakes
 - ⇒ We must teach to rule out any possible signs of bad intentions
- Humans and Machines agree on the importance of two features
 - “*Network traffic*” and a “*valid signature*”
- Machines rank top “*resources*”, Humans last ⇒ always take a look at it analysts!

Takeaways (2/2)



- Experts correctly classify samples by using less than 1/3 of the available features
 - With a preference for dynamic behaviour
- Machines prefer static features because dynamic ones are often missing
 - Research idea: semantically meaningful reconstruction of missing features
- Impact on the human-computer interaction; “modern” sandboxes must show:
 - OSINT data (e.g., IPs and domains)
 - What are the most significant features that helped classify the sample
 - ⇒ The analyst can bridge the cognitive gap

Backup slides...

Data required for the registration

		Job			
		Student	Researcher	Industry	Other
Experts	-	7	27	4	
Novices	72	-	-	-	

		Age			
		[20-25]	[26-30]	[31-40]	[40+]
Experts	-	7	21	10	
Novices	61	13	-	-	

		Years of experience				
		[0]	[1-3]	[4-6]	[7-9]	[10+]
Experts	-	13	11	9	5	
Novices	72	-	-	-	-	

Purchasable Features

Static

1. VT Labels
2. VT Submission History
3. Signature
4. Header Metadata
5. Sections
6. Imports
7. Resources
8. Strings

Dynamic

1. Network
2. Processes
3. Registry
4. Mutexes
5. File System
6. Runtime DLLs

Scoring Mechanism



Players start with a blank report

- Adds new features to the report by choosing them from a pre-defined catalog of 15 features
- Until she has gained enough information to make a **confident** binary classification
- 20 samples \rightarrow 20 rounds
- 20 **potential** points for each round
 - When she buys a new feature \rightarrow potential_points -= 1
 - “Empty feature” \rightarrow potential_points -= 0
- If the sample is correctly classified \rightarrow the player gets the remaining potential points
 - Otherwise zero 😞
- Final score = sum of all points obtained in each round * number of correct answers
 - \Rightarrow Highest possible score in DMIYC is $19 \cdot 20 \cdot 20 = 7600$

Dataset

- Benchmark Dataset: 21,944 reports from VirusTotal
 - 50% (10,972) malware
 - [2018, 2020]
 - Detection \geq 21 antivirus engines
 - No malware families were over-represented (AVClass2)
 - Most frequent family had 125/10,972 occurrences (1.1%)
 - 50% (10,972) goodware
 - Clean Windows 10 machine
 - Installed all community-maintained Chocolatey software
 - Extracted all the executable files present on the hard disk
 - Filtered by detection $<$ 3 (e.g., hacking/scanning tools)



Machine Learning Players – Classification Models

- Random Forest (RF)
 - Greedy tree-branch split strategy to divide the feature space and locate the classification boundary
 - 500 trees can provide stable classification accuracy
- Convolution Neural Network (CNN)
 - Inclines to directly fit the classification boundary in high-dimensional feature space by minimizing the correntropy loss
 - We compress the categorical attributes into low-dimensional numerical embedding vectors, i.e., word2vec
 - Applying one convolution layer to the embedding vector
 - Followed by 2 fully connected layers before filled into the softmax output