

Malware Detection and Classification using Machine Learning and Cuckoo Sandbox

Shubhankar Gaikwad
TE-31265
Seminar Guide- Prof. S. N. Girme

Introduction

- Battle between code makers and code breakers.
- Security tools- IDS, Honeypots, Sandbox.
- Need for dynamic malware analysis.
- Cuckoo Sandbox- Automated Malware Detection tool.
- Machine Learning for automating the process.

Motivation

- Importance of secure and robust codes in this information revolutionized world.
- Growing cyber attacks and threats.
- Need for stricter surveillance regimes to protect all individuals and data.

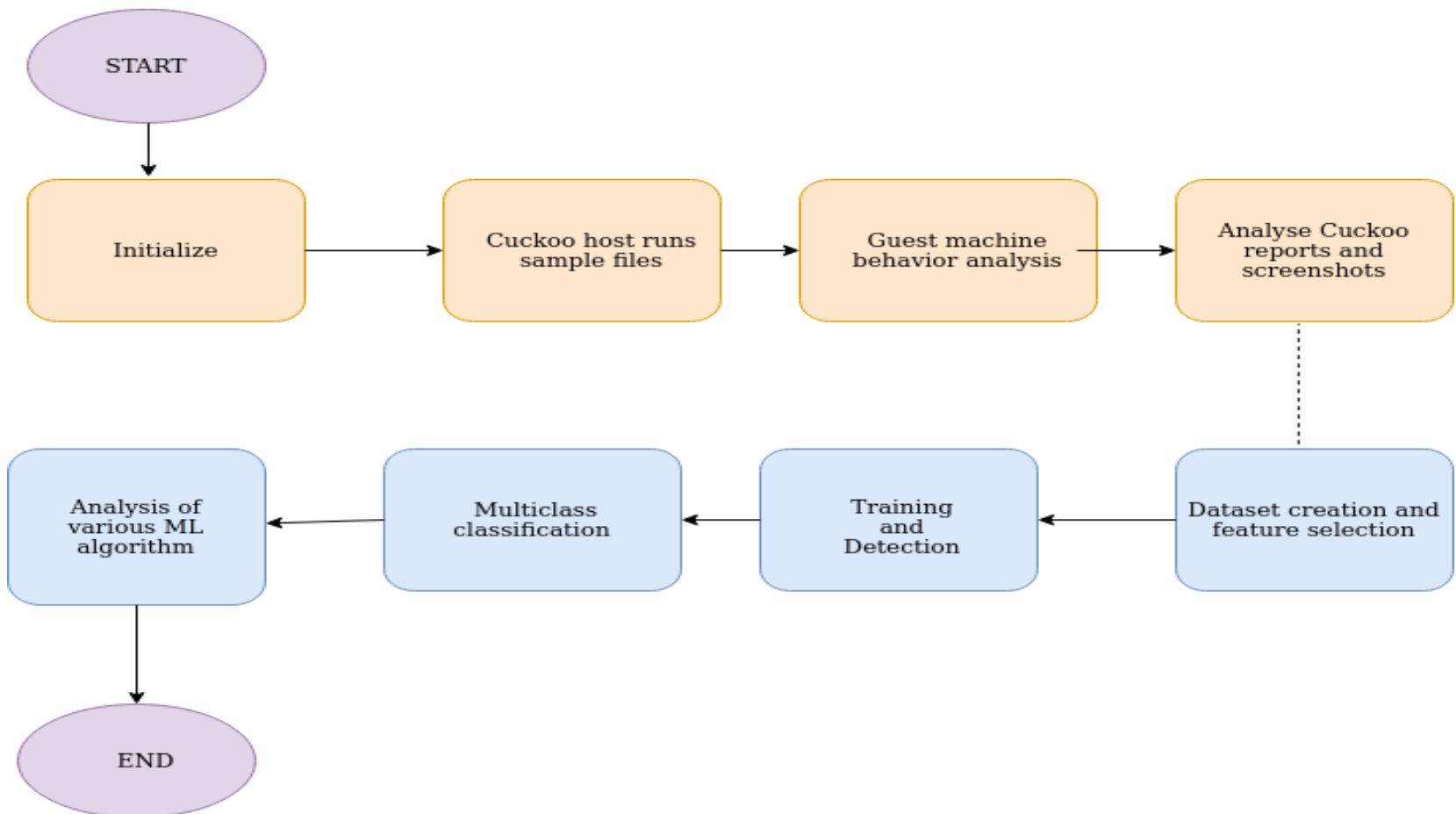
Literature Survey

Index	Paper	Authors	Methodology	Results/ Conclusion
1.	Deep Learning Approach for Intelligent Intrusion Detection System. [1]	R. Vinayakumar, Mamoun Alazab, K. P. Soman, Prabakaran Poornachandran, Ameer Al-Nemrat, Sitalakshmi Venkatraman	Analyses and studies various machine learning algorithms for the publicly available datasets and compares the intrusion detection results with Deep Learning approach	Found that minimal feature selection of the multi-class DNN worked more efficiently than traditional machine learning algorithms
2.	A Novel Machine Learning Based Malware Detection and Classification Framework. [2]	Kamalakanta Sethi, Rahul Kumar, Lingaraj Sethi, Padmalochan Bera, Prashanta Kumar Patra	New testing dataset is used by testing malicious files from VirusTotal and VirusShare sites using Cuckoo SandBox	A reduce in false positive rates is achieved and it is found that Decision Trees provide better results on analysis of the data

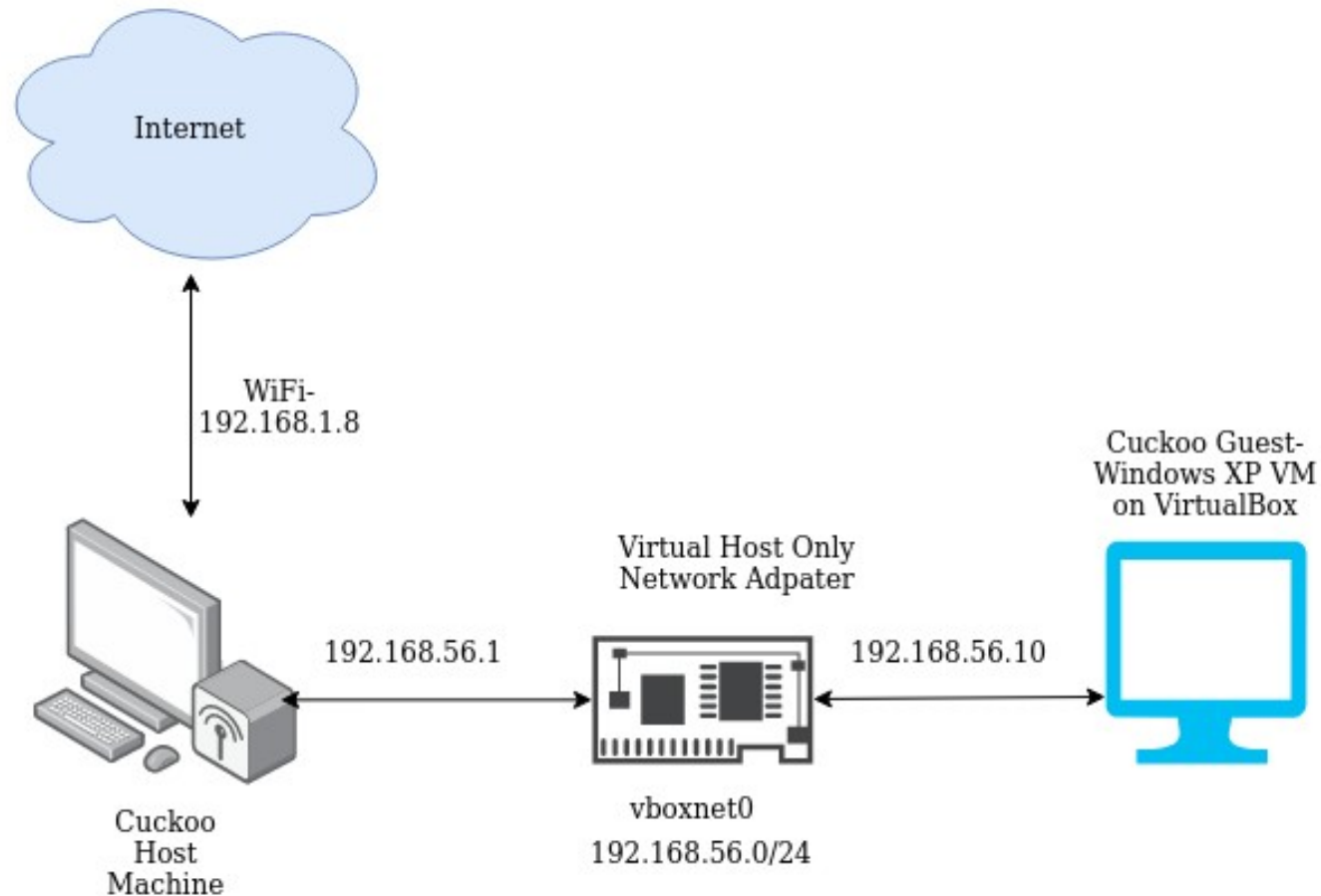
Problem Statement

- To set up a Cuckoo Sandbox environment and analyse behavior of malicious programs.
- To find the accuracy of various machine learning algorithms on the NSL-KDD dataset.

Implementation Workflow



Cuckoo Sandbox Architecture

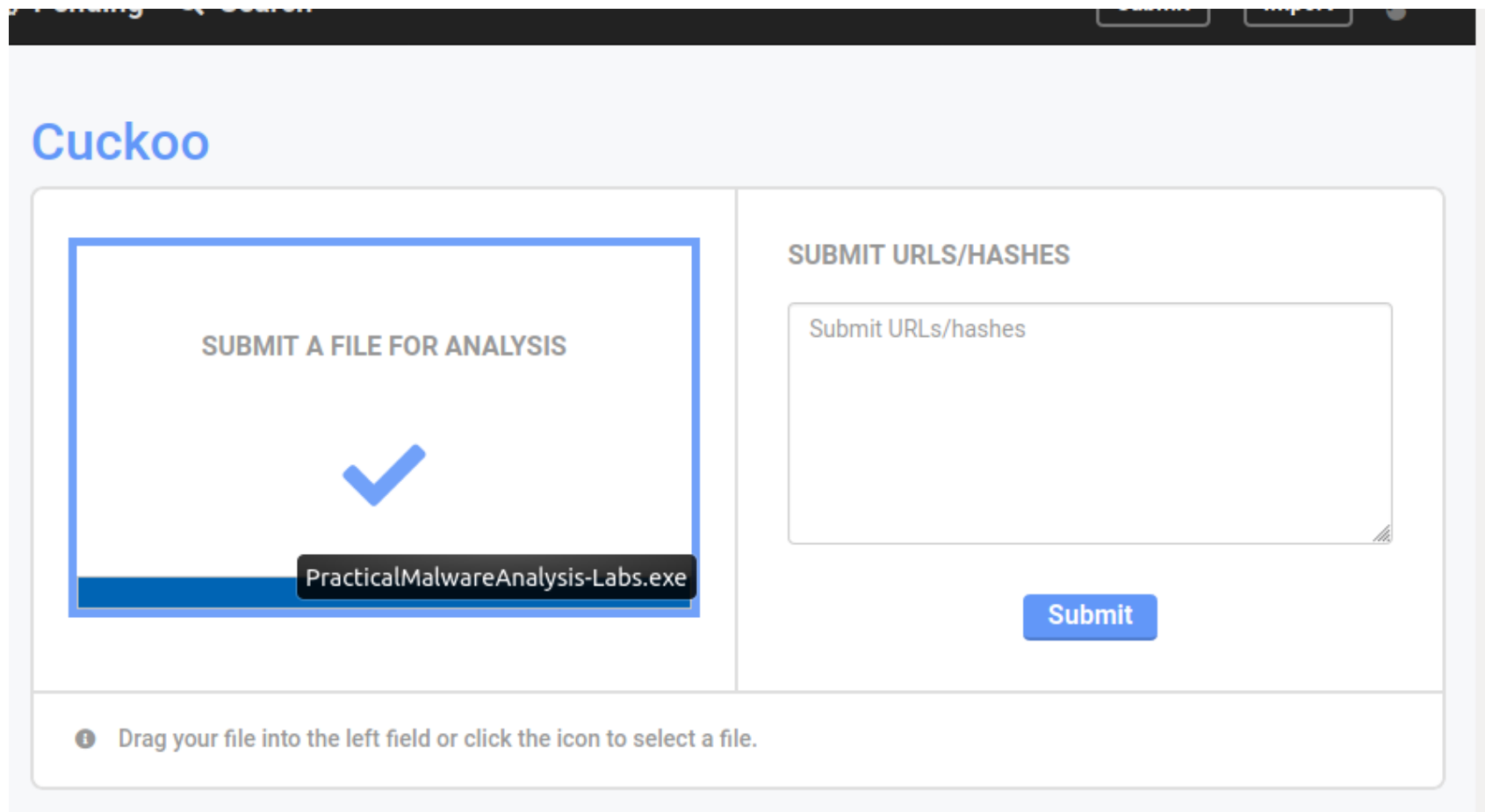


Network configurations and working

- Set up the Cuckoo installation on a different user other than main user on the Ubuntu machine.
- Create a virtual machine with Windows XP installation.
- Configure the network to create a virtual box network 192.168.56.0/24
- Setup the host ip as 192.168.56.1
- Setup the network configurations on guest virtual machine with ip 192.168.56.10
- Check if machines can ping.

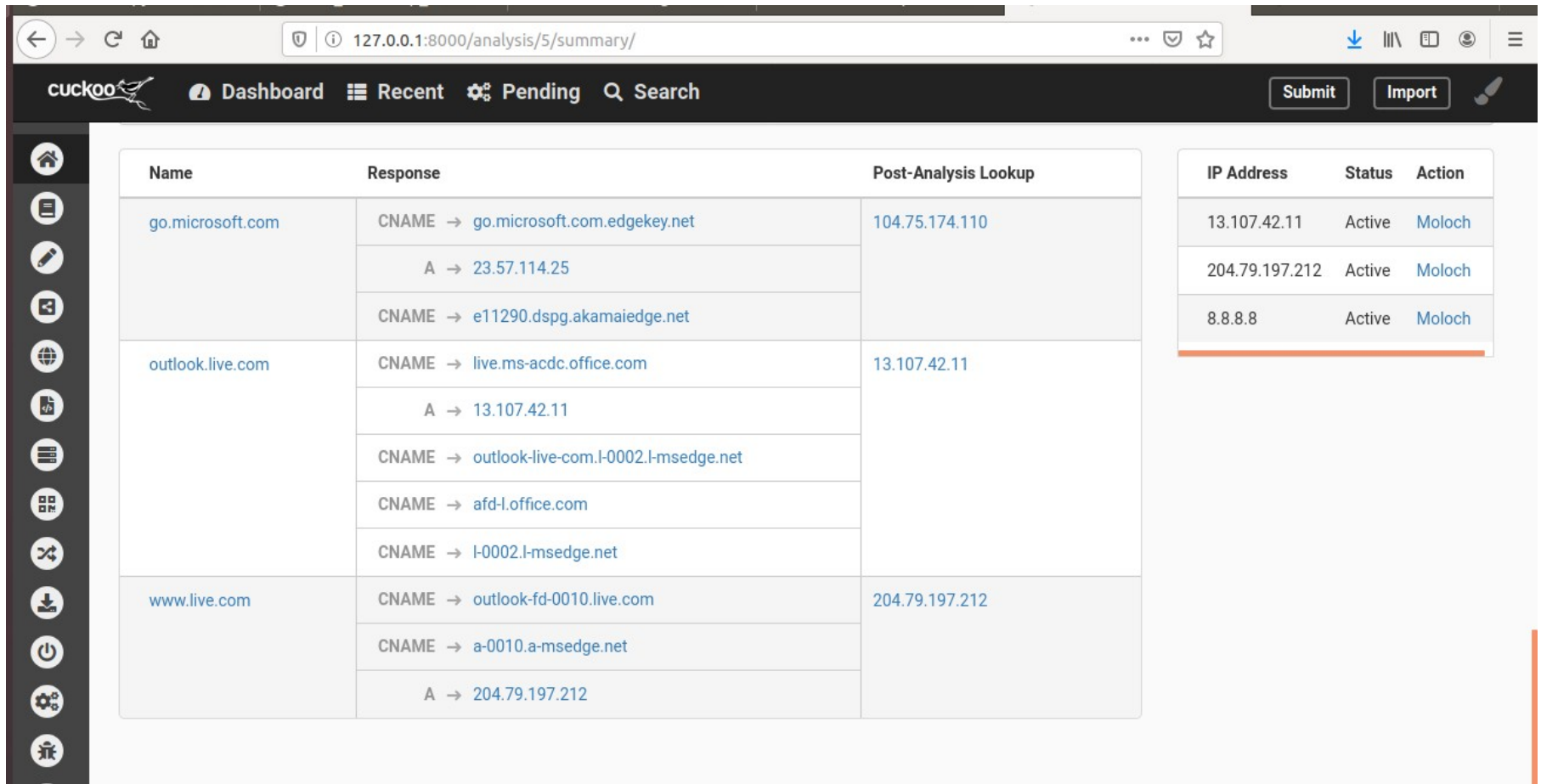
- Now install suitable softwares on VM and install the python agent file from host to guest.
- Take the snapshot of stable VM
- Try installing malicious files from host to guest through command line or Cuckoo web server.
- Wait for analysis report.
- Check the report and screenshots to analyze the malicious activity.

Working of Cuckoo Sandbox



The screenshot displays the Cuckoo Sandbox web interface. On the left, a box titled "SUBMIT A FILE FOR ANALYSIS" contains a large blue checkmark icon and a file named "PracticalMalwareAnalysis-Labs.exe" being dragged into the submission area. On the right, a box titled "SUBMIT URLS/HASHES" features a text input field labeled "Submit URLs/hashes" and a blue "Submit" button. At the bottom, a footer message states: "Drag your file into the left field or click the icon to select a file."

Analysis Reports



Name	Response	Post-Analysis Lookup
go.microsoft.com	CNAME → go.microsoft.com.edgekey.net	104.75.174.110
	A → 23.57.114.25	
	CNAME → e11290.dspg.akamaiedge.net	
outlook.live.com	CNAME → live.ms-acdc.office.com	13.107.42.11
	A → 13.107.42.11	
	CNAME → outlook-live-com.l-0002.l-msedge.net	
	CNAME → afd-l.office.com	
	CNAME → l-0002.l-msedge.net	
www.live.com	CNAME → outlook-fd-0010.live.com	204.79.197.212
	CNAME → a-0010.a-msedge.net	
	A → 204.79.197.212	

IP Address	Status	Action
13.107.42.11	Active	Moloch
204.79.197.212	Active	Moloch
8.8.8.8	Active	Moloch

PCAP Generation and Scoring

The screenshot displays the Cuckoo Sandbox web interface. The browser address bar shows the URL `127.0.0.1:8000/analysis/5/summary`. The interface includes a top navigation bar with links for Dashboard, Recent, Pending, and Search, along with Submit and Import buttons. A left sidebar contains various analysis tools, with counts for Behavioral Analysis (3), Dropped Files (2), and Process Memory (4). The main content area shows the summary for the archive `2019-12-25-traffic-analysis-exercise.pcap @ 2019-12-25-traffic-analysis-exercise.pcap.zip`. A table lists various hashes and analysis results. To the right, a 'Score' section indicates a score of 3.0 out of 10, and a 'Feedback' section provides a link for reporting issues. The bottom of the interface shows a table for 'Information on Execution' with columns for Category, Started, Completed, Duration, Routing, and Logs.

Summary

Size	633.5KB
Type	tcpdump capture file (little-endian) - version 2.4 (Ethernet, capture length 65535)
MD5	a71de91b1ea15c5331be0391c72c9d50
SHA1	beb5cec852e7bab883f7122fe745ab00cf0948e2
SHA256	a86c6a31ed04ed05571f997d296f8e7c7be7f262f6e602bd72d9e5de656945c5
SHA512	Show SHA512
CRC32	4C64E8A1
ssdeep	None
Yara	None matched

Score

This archive shows numerous signs of malicious behavior.

The score of this archive is **3.0 out of 10**.

Please notice: The scoring system is currently still in development and should be considered an *alpha* feature.

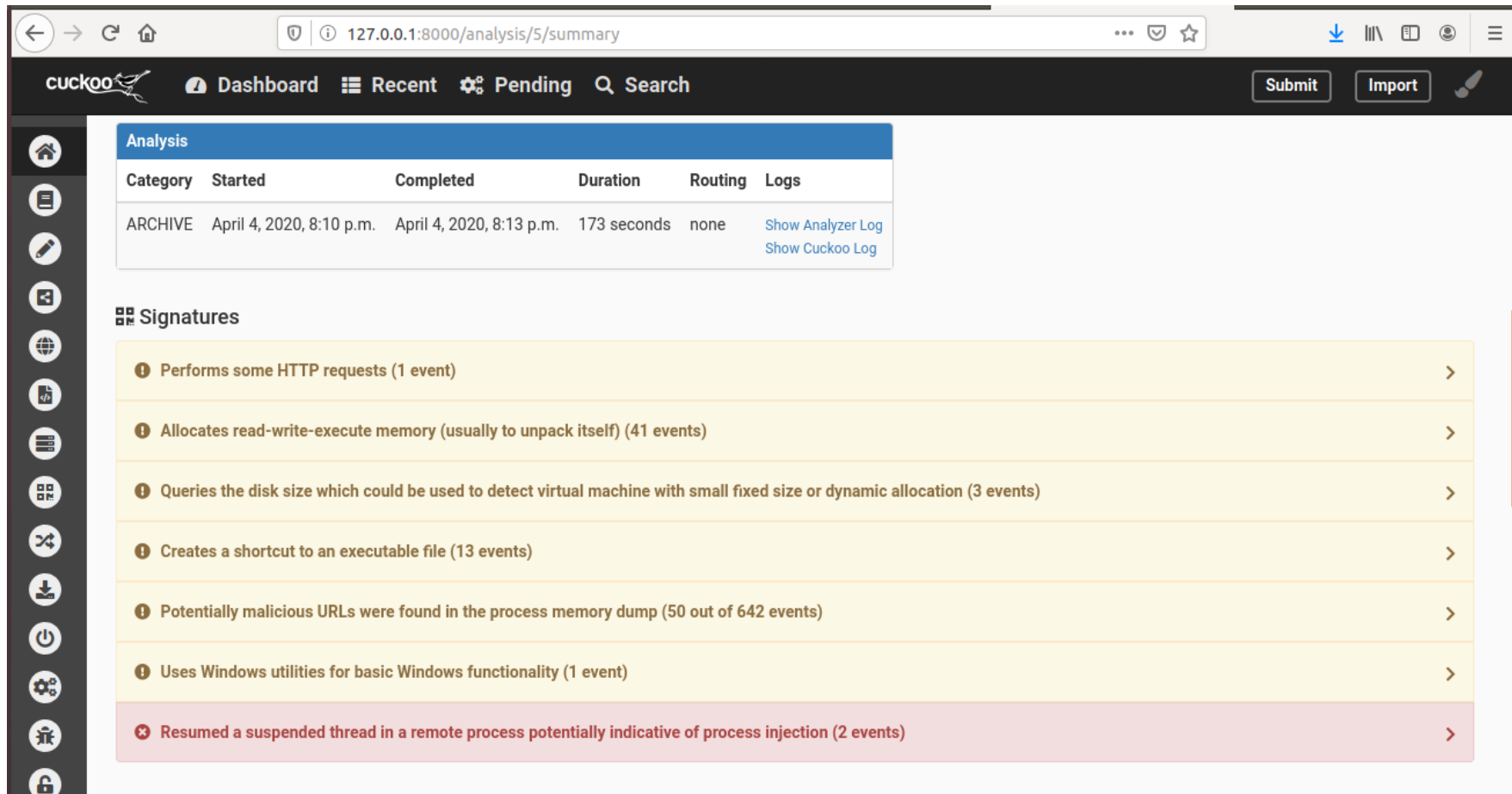
Feedback

Expecting different results? Send us this analysis and we will inspect it. [Click here](#)

Information on Execution

Category	Started	Completed	Duration	Routing	Logs
----------	---------	-----------	----------	---------	------

Signatures Captured



The screenshot displays the Cuckoo Sandbox web interface. The browser address bar shows the URL `127.0.0.1:8000/analysis/5/summary`. The interface includes a navigation bar with links to Dashboard, Recent, Pending, and Search, along with Submit and Import buttons. A sidebar on the left contains icons for various functions. The main content area is divided into two sections: Analysis and Signatures.

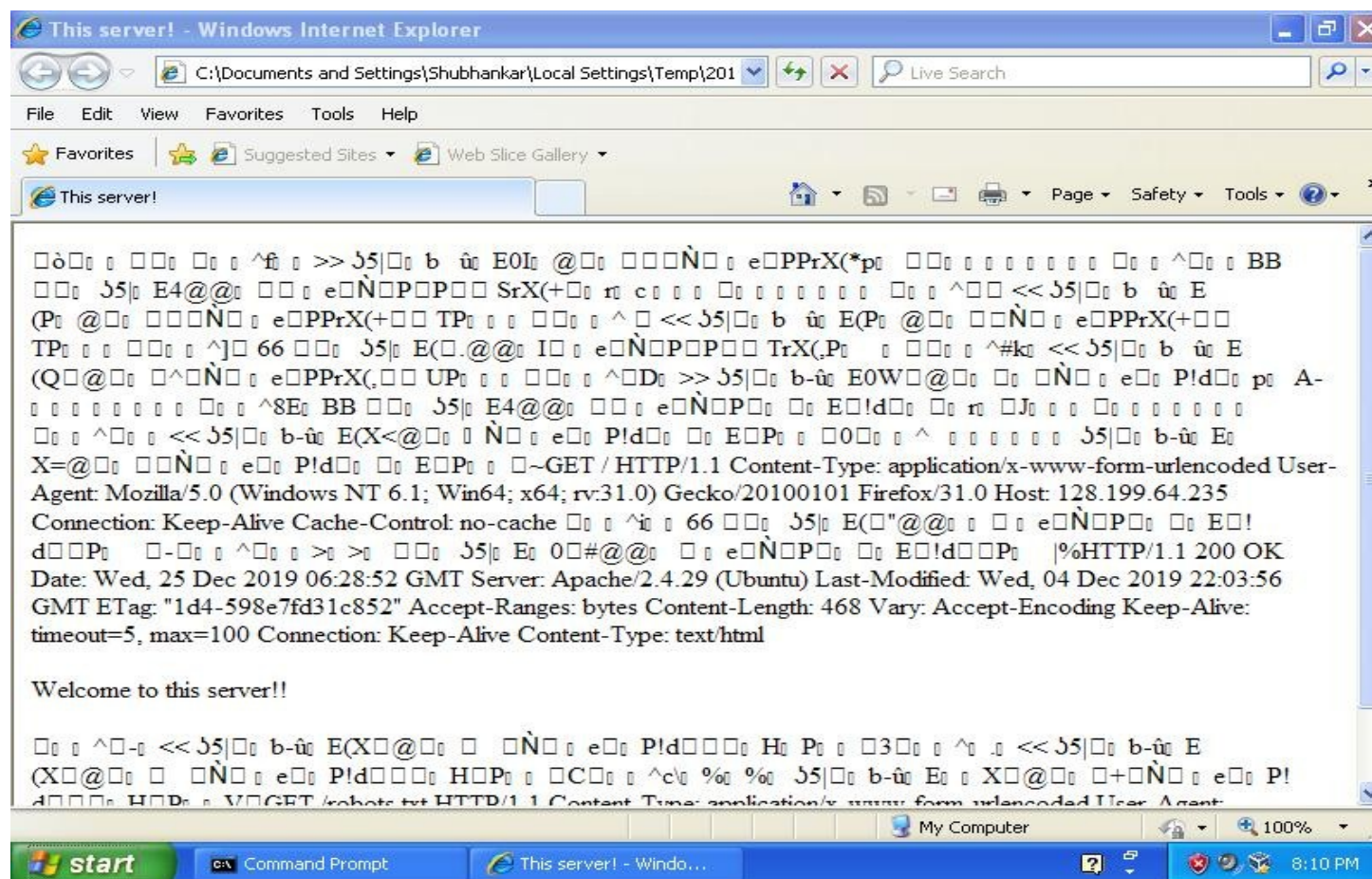
Analysis

Category	Started	Completed	Duration	Routing	Logs
ARCHIVE	April 4, 2020, 8:10 p.m.	April 4, 2020, 8:13 p.m.	173 seconds	none	Show Analyzer Log Show Cuckoo Log

Signatures

- Performs some HTTP requests (1 event)
- Allocates read-write-execute memory (usually to unpack itself) (41 events)
- Queries the disk size which could be used to detect virtual machine with small fixed size or dynamic allocation (3 events)
- Creates a shortcut to an executable file (13 events)
- Potentially malicious URLs were found in the process memory dump (50 out of 642 events)
- Uses Windows utilities for basic Windows functionality (1 event)
- Resumed a suspended thread in a remote process potentially indicative of process injection (2 events)

Screenshot from the Cuckoo Guest



Machine Learning for Malware Detection

K-Nearest Neighbors-

- Discrete classification of a point by comparing to the majority of nearest k training points.
- Can be used to detect abnormal behaviors over wireless sensor networks.

Decision Trees-

- Leaf nodes contain labels and intermediate nodes are conditions.
- Repetitive checking is done against conditions to reach down to a label.
- Useful for signature based malware detection.

Random Forests

- It is similar to bagging.
- It creates large collection of de-correlated trees and then averages them out.

Neural Networks

- Inspired by the biological neural networks.
- The input layer is activated and data is passed to hidden layers with hyper parameters.
- Multiclass output layer can be achieved.

NSL-KDD Dataset

Activities LibreOffice Calc Mon 21:28

KDDTrainWL.csv - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help

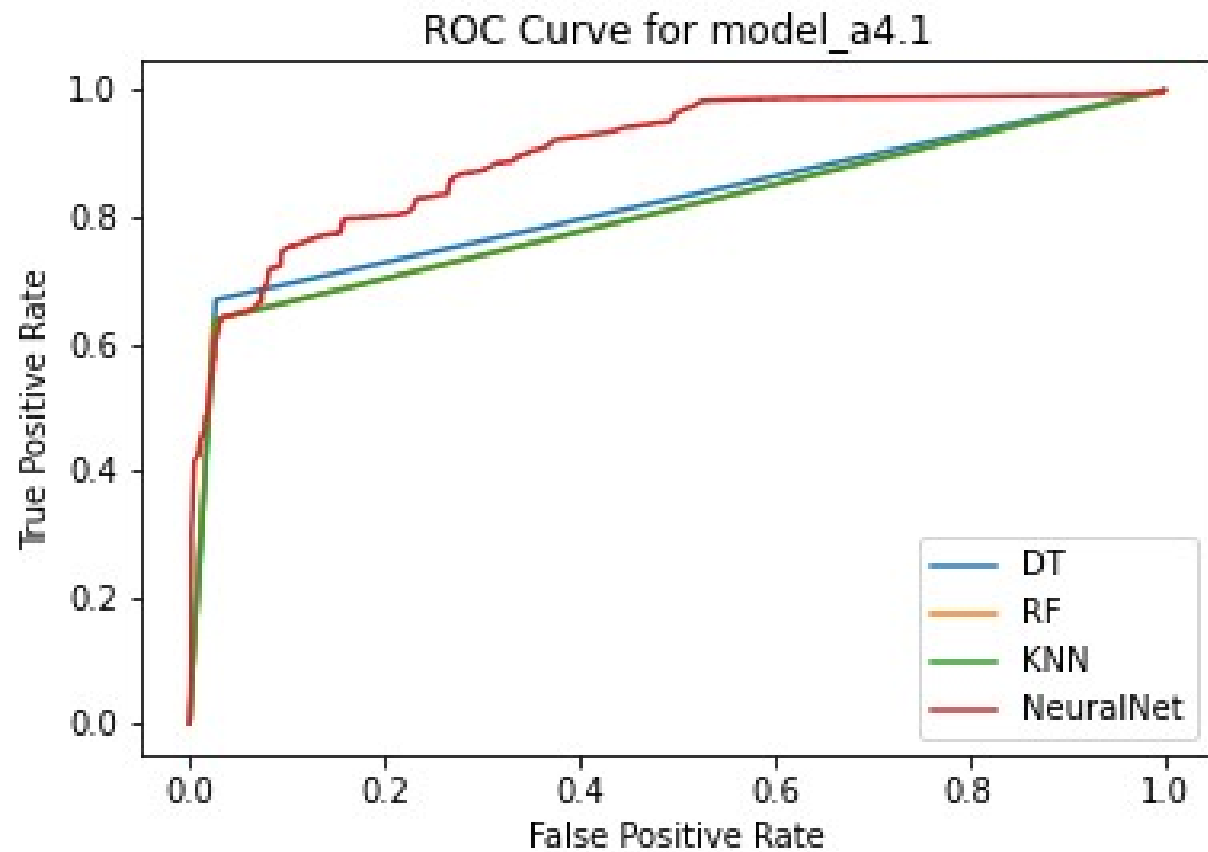
Liberation Sans 10

A1 duration

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	root_shell	su_attempted	num_root	nt
2	0	tcp	ftp_data	SF	491	0	0	0	0	0	0	0	0	0	0	0	0
3	0	udp	other	SF	146	0	0	0	0	0	0	0	0	0	0	0	0
4	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	tcp	http	SF	232	8153	0	0	0	0	0	1	0	0	0	0	0
6	0	tcp	http	SF	199	420	0	0	0	0	0	1	0	0	0	0	0
7	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	tcp	remote_job	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	tcp	http	SF	287	2251	0	0	0	0	0	1	0	0	0	0	0
15	0	tcp	ftp_data	SF	334	0	0	0	0	0	0	1	0	0	0	0	0
16	0	tcp	name	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	tcp	netbios_ns	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	tcp	http	SF	300	13788	0	0	0	0	0	1	0	0	0	0	0
19	0	icmp	eco_i	SF	18	0	0	0	0	0	0	0	0	0	0	0	0
20	0	tcp	http	SF	233	616	0	0	0	0	0	1	0	0	0	0	0
21	0	tcp	http	SF	343	1178	0	0	0	0	0	1	0	0	0	0	0
22	0	tcp	mtp	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	tcp	http	SF	253	11905	0	0	0	0	0	1	0	0	0	0	0
25	5607	udp	other	SF	147	105	0	0	0	0	0	0	0	0	0	0	0
26	0	tcp	mtp	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	507	tcp	telnet	SF	437	14421	0	0	0	0	0	1	3	0	0	0	0
28	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	tcp	http	SF	227	6588	0	0	0	0	0	1	0	0	0	0	0
30	0	tcp	http	SF	215	10499	0	0	0	0	0	1	0	0	0	0	0

Sheet 1 of 1 KDDTrainWL Default English (India) Average: ; Sum: 0 100%

Result



▼ performance matrix

```
[ ] classifier_results_df
```

	model	fpr	acc	tpr	auc	TP	FP	TN	FN
0	RF	0.024697	0.784284	0.636857	0.806080	8453	253	9991	4820
1	KNN	0.026064	0.784029	0.637460	0.805698	8461	267	9977	4812
2	DT	0.026650	0.801548	0.668952	0.821151	8879	273	9971	4394
3	NN	1.000000	0.923077	1.000000	0.901289	24	2	0	0

Future Scope

- Creation of a more valid dataset.
- Testing the pcaps generated from Cuckoo Sandbox tool with ML algorithms.
- Development of an Intelligent Intrusion Detection System.

Conclusion

- Malware analysis and classification is of importance for securing information systems and networks.
- Cuckoo Sandbox tool helps in analysing the malwares effectively and quickly and provides vectors with new labels and features.
- These vector datasets are useful in classifying malwares with different machine learning algorithms to give better efficiencies.

References

- [1] R. Vinayakumar, Mamoun Alazab, K. P. Soman, Prabaharan Poornachandran, Ameer Al-Nemrat, Sitalakshmi Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, pp.41525-41550, 2019.
- [2] Kamalakanta Sethi, Rahul Kumar, Lingaraj Sethi, Padmalochan Bera, Prashanta Kumar Patra, "A Novel Machine Learning Based Malware Detection and Classification Framework," International Conference on Cyber Security and Protection of Digital Services, Oxford, United Kingdom, pp. 1-4, 2019.
- [3] Cuckoo Sandbox-Automated Malware Analysis, "Cuckoo Sandbox book" Available at: <https://cuckoo.readthedocs.io/en/latest/> [Online]
- [4] Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, "Elements of Statistical Learning- Data Mining, Inference, and Prediction" Springer Series in Statistics Available at: <https://link.springer.com/book/10.1007/978-0-387-84858-7> [Online]