MALWARE DETECTION AND ANALYSIS USING MACHINE LEARNING

at

Sathyabama Institute of Science and Technology

(Deemed to be University)

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Engineering Degree in Computer Science and Engineering

By

MANOJ SIRIGIRI

REG. NO. 39110604

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119

APRIL 2023

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —A grade by NAAC JeppiaarNagar,
Rajiv Gandhi Salai, Chennai – 600 119
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **MANOJ SIRIGIRI (39110604)** who carried out the Project Phase-2 entitled **"MALWARE DETECTION AND ANALYSIS USING MACHINE LEARNING"** under my supervision from January 2023 to April 2023.

Internal Guide

Ms. R. YOGITHA M.E., (Ph.D)

Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva-voce Examination held on 20.04.2023

Internal Examiner **External Examiner**

DECLARATION

I, **Manoj sirigiri (Reg No- 39110604),** hereby declare that the Project Phase-2 Report entitled **“MALWARE DETECTION AND ANALYSIS USING MACHINE LEARNING “done** by me under the guidance of **Ms. R. Yogitha M.E., (Ph.D)** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 20.04.2023

PLACE: Chennai

MANOJ SIRIGIRI

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D**, **Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms. R. Yogitha M.E., (Ph.D) ,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

The increasing prevalence of malware is one of the most significant threats to the security of computer systems and the Internet. Malware is any software that has the intention of performing malicious activities on a targeted system, including stealing data, disrupting operations, and causing financial losses. There are many types of malware, and attackerscan use various communication strategies to infect systems and spread malware, including Trojans, keyloggers, port forwarding, application format converters, and social engineering tactics. To address the challenge of detecting and classifying malware effectively, we conducted a study to explore various malware types and communication strategies. We then applied machine learning algorithms to the data to extract valuable insights and enable us to classify malware as malicious or non-malicious. The machine learning algorithms we used in our study were the Random Forest, K-Nearest Neighbor, and Support Vector Machine algorithms. We applied these algorithms to a dataset containing a large number of malware samples and a set of features that describe each sample's behavior and characteristics. To evaluate the performance of the machine learning models, we analyzed the resulting classification report, accuracy, and f1 score metrics. The classification report provided detailed information about the classification results, including the precision, recall, and f1- score for each class. The accuracy metric measured the percentage of correctly classified samples, while the f1 score was a measure of the classifier's accuracy that combines both precision and recall. Our analysis revealed that the Random Forest algorithm achieved the highest f1-score for the validation dataset, indicating that it was the best performing algorithm for classifying both malicious and non-malicious software. The results also showed that the K-Nearest Neighbor and Support Vector Machine algorithms performed well but did not achieve the same level of accuracy as the Random Forest algorithm. By using machine learning algorithms to classify malware, we aim to improve the identification and classification of malware, thus enhancing online privacy for individuals. Our work demonstrates the value of applying machine learning techniques to identify and prevent malware attacks, which can help to protect computer systems and the Internet from harm. In summary, this study highlights the importance of understanding various malware types and communication strategies and the potential impact of malware attacks on organizations and individuals. By using machine learning algorithms to classify malware, we can detect and prevent malware attacks effectively, thus enhancing online privacy and security.

List of figures

S.no	Figure name	Page number
1	Most common Cyber attacks	9
2	Options in a trojan	11
3	Working of keylogger	11
4	Social engineering cycle	12
5	back door in malware	19
6	NSIS	20
7	Types of scanning in an antivirus	20
8	malware found and deleted	21
9	Phase 1 architecture	22
10	Phase 2 architecture	22
11	Phase 3 architecture	23
12	Phase 4 architecture	23
13	Data set containing information About applications	24
14	Results of gradient boost algorithm	25
15	Results of logistic-regression algorithm	25
16	Results of random forest algorithm	26
17	The user interface of running witha machine learning model	27
18	final result of the proposed system	28

List of abbreviations

S.no	Abbreviations	Full forms
1	AI	Artificial Intelligence
2	SVM	Support Vector Machine
3	ML	Machine Learning
4	KNN	K-Nearest Neighbor
5	HTTPS	Hypertext Transfer Protocol Secure
6	HTTP	Hypertext Transfer Protocol
7	URL	Uniform Resource Locator

INDEX

Chapter No		TITLE	Page No.
	ABSTRACT		5
	LIST OF FIGURES		7
1	INTRODUCTION		8
	1.1	CYBERSECURITY	8
	1.2	MACHINE LEARNING	9
	1.3	MALWARE AND ITS TYPES	10
2	LITERATURE SURVEY		13
	2.1	INFERENCES FROM LITERATURE SURVEY	17
3	ARCHITECTURE DESIGN AND PROPOSED SYSTEM		18
	3.1	WHAT IS TROJAN	18
	3.2	HOW TROJAN WORKS	18
	3.3	WHAT CAN A TROJAN DO	19
	3.4	WHAT IS KEY LOGGER	19
	3.5	HOW TO CONVERT PYTHON SCRIPT TO AN EXE	20
	3.6	ANTI-VIRUS DETECTION METHODS	20
	3.7	PROPOSED SYSTEMS	21
	3.8	ARCHITECTURE DIAGRAMS	22
4	PRE-INSTALLATION DETECTION THROUGH MACHINE		24
	4.1	DATA SET	24
	4.2	ALGORITHMS USED	24
	4.3	STREAMLIT INTERFACE	26
5	RESULTS AND CONCLUSION		28
6	REFERENCES		29

CHAPTER -1

INTRODUCTION

1.1 CYBERSECURITY

In recent years, networks have evolved from a mere means of communication to a present computational infrastructure. Networks have become larger, faster, andhighly dynamic. The pervasive use of computer and network technologies in all walks of life has turned cybersecurity issues into national security issues.
sequence, cyber-attacks against apparently “non-critical” services may produce unforeseen side effects of devastating proportions. First of all, the Malware is a malicious technique which are created by any software programmer with an intention to cause damage to a computer or server. The malware does the damage after it is implemented or introduced in some way into a target’s computer and can corrupt the data of target’s computer. Generally, viruses are sent mostly through emails as attachment files. When someone opens a mail, it appears to be similar to trusted companies or friend’s mail, attachment or through link. It has become very easy to compromise an email account in modern world. Mostly the attachment would be like a document, picture etc. as soon as the clicks it the attacker gets a connection with victim. Victim notices that the media attached was of some kind of irrelevant or blank and closes the concerned window. A virus is computer program which can copy itself and infect a system. The virus is created in such a way that it sends similar mails to others nodes in network. The virus injected in to the victim’s system does the task as programmed by the attacker like compromising personal, company’s details, manipulation of system configuration etc. The problem to be examined involves the high spreading rate of computer malware (viruses, worms, Trojan horses, rootkits, botnets, backdoors, and other malicious software) and conventional signature matching-based antivirus systems fail to detect polymorphic and new, previously unseen malicious executables. Malware are spreading all over the world through the Internet and are increasing day by day, thus becoming a serious threat. The manual heuristic inspection of static malware analysis is no longer considered effective and efficient compared against the high spreading rate of malware.

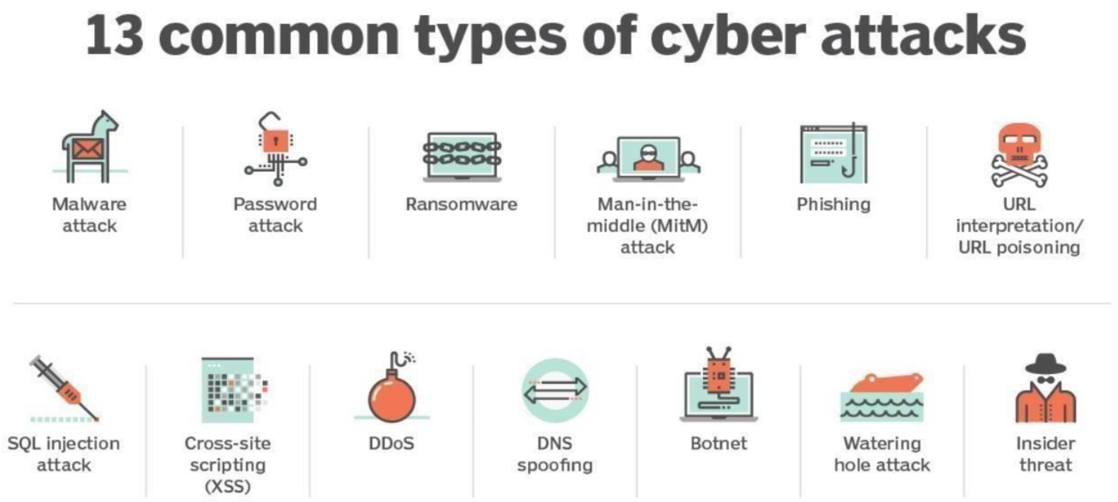


Fig 1 Most common Cyber attcaks

The manual heuristic inspection of static malware analysis is no longer considered effective and efficient compared against the high spreading rate of malware. Nevertheless, researches are trying to develop various alternative approaches in combating and detecting malware.

1.2 MACHINE LEARNING

One proposed solution is to use behaviour malware analysis combined with data mining tasks such as machine learning classification techniques to achieve effectiveness and efficiency in detecting malware. With machine learning, cybersecurity systems can analyse pattens and learn from them to help prevent similar attacks and respond to changing behaviours. It can help cybersecurity teams be more proactive in detecting threats and responding to active attacks in real time. Itcan reduce the amount of time spent on routine tasks and enable organisations to use their resources more strategically. In short, machine learning can make cybersecurity simpler, more proactive, less expensive, and far more effective. But it can only do those things if the underlying data that supports the machine learning provides a complete picture of the environment. As they say, "garbage in, garbage out. Machine learning is about developing patterns and manipulating those patterns with algorithms. In order to develop patterns, you need a lot of rich data fromeverywhere because the data needs to represent as many potential outcomes from as many potential scenarios as possible. It’s not just about the quantity of data; it’s also about the quality. In this project, I will discuss how an attacker creates and send his malware or trojans using different social engineering techniques, and how python is used to create a keyloggers and how a normal user should get protected from these types of payloads. A Trojan is a program which contains malicious or harmful code wrapped with apparently harmless programming or data in such a way that it can enter the victim’s computer undetected, providing the attacker unrestricted access to the data stored on that computer and causing immense damage to the victim. Trojans have the capability to replicate, spread and get activated upon certain predefined actions performed by the victim. With the help of a Trojan, an attacker gets access to the victim’s computer resources, stored passwords and it would enable him/her to read personal documents, delete important files or the whole drive, display pictures, and/or show messages on the screen. For example, a user downloads a music file or a video from the internet, but when he/she runs it, it triggers a dangerous program that may erase the user’s disk or send his/her credit card numbers to a stranger. In another aspect, a victim may also be used as an intermediary to launch attacks on others without letting the victim know about this.

1.3 MALWARE AND ITS TYPES

A trojan can perform many operations on attacker commands some of them are

- Monitoring user behavior through **KEYLOGGERS**
- Display monitoring
- Accessing confidential information
- Activating a system's webcam and recording video
- Taking screenshots
- Distributing viruses and other malware
- Getting full control on browser, CMD, file manager
- Formatting drives
- Deleting, downloading or altering files and file systems

Coming to keylogger Keyloggers are a type of monitoring software designed to record keystrokes made by a user. Data captured by keyloggers can be sent back to attackers via email or uploading log data to predefined websites, databases, or FTP servers.

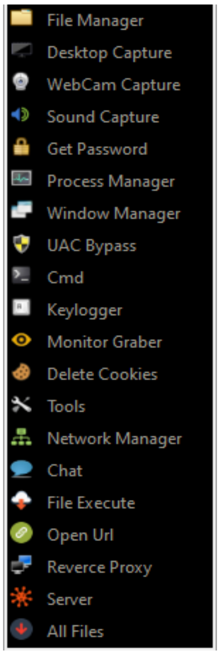
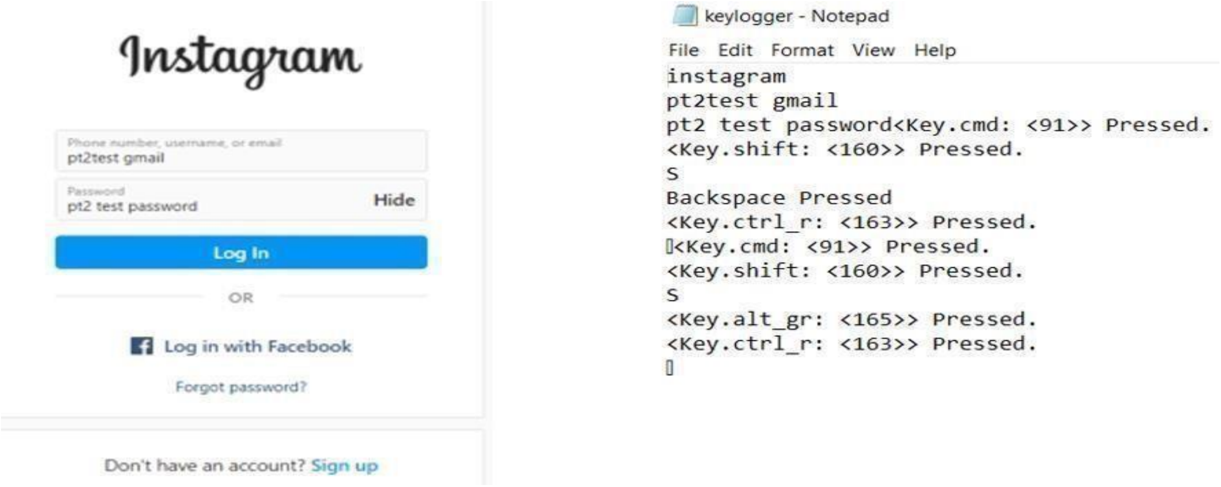


Fig 2 Options in a trojan

If the keylogger comes bundled within a large attack, attacker might simply remotely log into a machine to download keystroke data. The pynput library in python allows one to control and monitor/listen to your input devices such as they keyboard

The pynput. keyboard allows you to control and monitor the keyboard

The passwords and credit card numbers you type, the webpages you visit – all by logging your key strokes. The software is installed on your computer, and records everything you type. Then it sends this log file to a server, where the cybercriminals wait to make use of all this sensitive information



Phone number, username, or email: pt2test gmail

Password: pt2 test password Hide

```
instagram
pt2test gmail
pt2 test password<Key.cmd: <91>> Pressed.
<Key.shift: <160>> Pressed.
S
Backspace Pressed
<Key.ctrl_r: <163>> Pressed.
<Key.cmd: <91>> Pressed.
<Key.shift: <160>> Pressed.
S
<Key.alt_gr: <165>> Pressed.
<Key.ctrl_r: <163>> Pressed.
```

fig 3 Working of keylogger

Social engineering

Social engineering is the term used for a broad range of malicious activities accomplished through human interactions. It uses psychological manipulation to trick users into making security mistakes or giving away sensitive information. There are different formats an attacker can send a payload to victim.

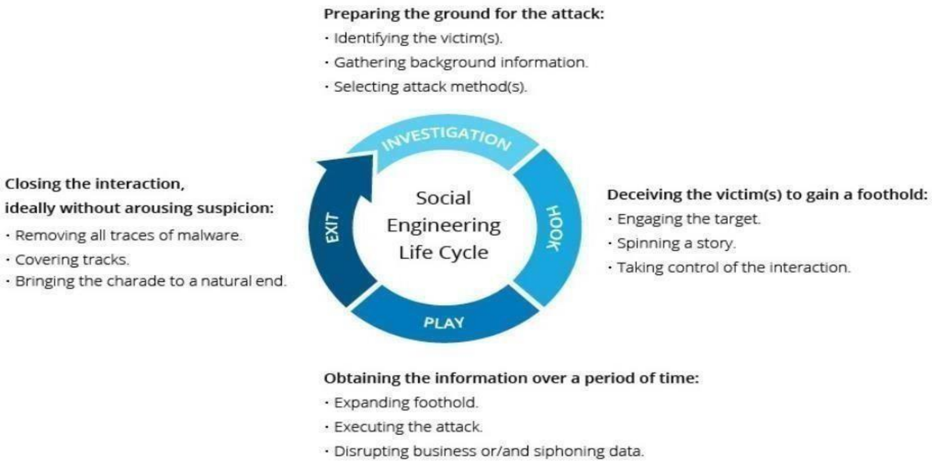

Preparing the ground for the attack:
· Identifying the victim(s).
· Gathering background information.
· Selecting attack method(s).
INVESTIGATION
Social Engineering Life Cycle
HOOK
Deceiving the victim(s) to gain a foothold:
· Engaging the target.
· Spinning a story.
· Taking control of the interaction.
PLAY
Obtaining the information over a period of time:
· Expanding foothold.
· Executing the attack.
· Disrupting business or/and siphoning data.
EXIT
Closing the interaction, ideally without arousing suspicion:
· Removing all traces of malware.
· Covering tracks.
· Bringing the charade to a natural end.

Fig 4 Social engineering cycle

Several types of detection are used to prevent these attacks mostly every antivirus system out there uses **post installation detection** data conveys information about behaviour or events caused by process activity in a system.

The idea that I am proposing is **pre installation detection** which one can use at free of cost and this method detects malware even before installation using machine learning Pre installation phase data is anything you can tell about a file without executing it. This may include executable file format descriptions, code descriptions, binary data statistics, text strings and information extracted via code emulation and other similar data.

CHAPTER – 2

LITERATURE REVIEW

1. **Research on technology of trojan horse detection**

   Yu, W., Yalin, Y., & Haodan, R. (2019, October). Research on the technology of trojan horse detection. In 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 117-119). IEEE.

- In this paper the characteristics and principles of trojans are analysed and the detection methods are compared
- This paper includes detection methods like
- Sand box testing
- Heuristic based testing

2. **Review of Signature-based Techniques in Antivirus Products**

   Al-Asli, M., & Ghaleb, T. A. (2019, April). Review of signature-based techniques in antivirus products. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-6). IEEE.

- This paper revisits existing research on virus detection using signature-based algorithms
- hybridization of cybercrime investigation models with existing antivirus products to make an extension to their benefits to the entire community.

3. **Comprehensive Review on Malware Detection Approaches**

   Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. IEEE Access, 8, 6249-6271.

- This paper briefs different malware detection techniques
- Evolution of malware detection and history of malware discussed in this paper

4. **Malware Detection Techniques**

   Idika, N., & Mathur, A. P. (2007). A survey of malware detection techniques. Purdue University, 48(2), 32-46.

- In this paper they discussed about what actually malware is followed by the categories of malware

- Paper includes future malware threats and techniques

5. **cyber security at a glance**

   Asish, M. S., & Aishwarya, R. (2019, March). Cyber security at a glance. In 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (Vol. 1, pp. 240-245). IEEE.

- This paper described various hacking techniques and its counter measures in various aspect
- Learned about different types of malwares and after effects of the installation of those malware
- Discussed all kinds of malware form old to new to describe the evolution of malware
- Explained how Classical Defence mechanism (like signature-based malware detection) used by anti-virus will fail to cope up new age malware challenges.

6. **Malware detection using machine learning and deep learning**

   Rathore, H., Agarwal, S., Sahay, S. K., & Sewak, M. (2018, December). Malware detection using machine learning and deep learning. In International Conference on Big Data Analytics (pp. 402-411). Springer, Cham.

- In this paper, they have modelled malware analysis and detection as machine learning and deep learning problem.
- They used best practices in building these models (like cross-validation, fixing class imbalance problem, etc.).

7. **The world of Malware: An overview**

   Namanya, A. P., Cullen, A., Awan, I. U., & Disso, J. P. (2018, August). The world of Malware: An overview. In 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud) (pp. 420-427). IEEE.

- This paper presents an overview of the world of malware with the intent of providing the underlying information for the intended study into developing malware detection approaches.
- This paper reviews the foundational information of malware and anti-malware systems.

- They presented summaries of works found in literature about malware evolution, malware analysis techniques, malware evasion techniques and existing malware detection methods.

8. **keyloggers: silent cyber security weapons**

   Bhardwaj, A., & Goundar, S. (2020). Keyloggers: silent cyber security weapons. Network Security, 2020(2), 14-19.

- Explained the destruction caused by keyloggers in past and how advance they became
- Keyloggers aka silent cyber weapon is deadly and undetectable in most cases
- Described the economic damage caused by keyloggers
- The privilege level at which keyloggers execute is higher than typical malware, which makes them almost impossible to detect and remove.

9. **Identification and prevention of social engineering attacks on an enterprise**

   Parthy, P. P., & Rajendran, G. (2019, October). Identification and prevention of social engineering attacks on an enterprise. In 2019 International Carnahan Conference on Security Technology (ICCST) (pp. 1-5). IEEE.

- This paper classifies the various social engineering attacks based on the perspective of an attacker.
- Explained about all types of enterprise attacks and classified them
- Many new social engineering techniques were discussed likes reverse social engineering.
- This paper helps the reader to gain insight into how social engineering can be used against enterprises.

10. **Social media: A new vector for cyber attack**

    Kunwar, R. S., & Sharma, P. (2016, April). Social media: A new vector for cyber- attack. In 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring) (pp. 1-5). IEEE.

- This paper provides an in-depth detail of threats, security risks and different types of attacks using social media.
- Discussed many thigs about spam and malicious ads in social media

- Found detailed information about the number of users who uses social media and how many are lured into these cyber-attacks gave a clear ratio about much more information.

11. The strange world of keyloggers - an overview, Part I

Creutzburg, Reiner (2017). "The strange world of keyloggers - an overview, Part I". Electronic Imaging. 2017 (6): 139–148.

- provided a summary of the relevant hard-, software, and mobile keyloggers that are available and in use, as well as a bibliographic overview of keyloggers. Keyloggers' capabilities, accessibility, and detection potential are examined and detailed.

12. A Novel Approach of Unprivileged Keylogger Detection.

Wajahat, A., Imran, A., Latif, J., Nazir, A., & Bilal, A. (2019). A Novel Approach of Unprivileged Keylogger Detection. 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies(iCoMET).doi:10.1109/icomet.2019.8673404.

- This research focused on detecting the most common indigent user space keylogger. They demonstrated code snippets in this study that allow the client to deal with keylogger spyware without jeopardizing security.

13. A Comprehensive Study on Malware Detection and Prevention Techniques used by Anti-Virus.

Rohith, C., & Kaur, G. (2021, April). A Comprehensive Study on Malware Detection and Prevention Techniques used by Anti-Virus. In 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 429-434). IEEE.

- The purpose of this study is to describe and debate the cutting-edge technologies employed by anti-virus. They talked about how malware that lives on a system might permanently destroy its hardware components.

14. CyberSecurity Analytics to Combat Cyber Crimes

Nallaperumal, K. (2018). CyberSecurity Analytics to Combat Cyber Crimes. 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). doi:10.1109/iccic.2018.8782430

- This paper gives an introduction to the cyber-crimes and their impacts.

15. Design and Analysis of Machine Learning Based Technique for Malware Identification and Classification of Portable Document Format Files

Alshamrani, S. S. (2022). Design and Analysis of Machine Learning Based Technique for Malware Identification and Classification of Portable Document Format Files. Security & Communication Networks, 2022.

- This study presents a Machine Learning (ML) model, which can recognize JavaScript and malicious API call assaults in PDF files.

CHPAPTER - 2.1

INFERENCE FROM LITERATURE SURVEY

- From the above-mentioned literature works, it is evident that there has been effective research on history of malware and its detection methods.
- It is observed that the above-mentioned detection methods have their own pros and cons.
- Got to know how Classical Defence mechanism (like signature-based malware detection) used by anti-virus will fail to cope up new age malware challenges.
- While some of the advanced malware have no effect on using these old detection methods
- The above literature review also explained many new cyber attacks and their prevention methods
- It is observed how these malwares are used to harm the economy of country with out blood loss invading many privacy policy's and making many data breaches
- Got to know many things silent cyber-attacks like keylogger and level of damage it causes
- I got to know how expensive an efficient robust anti-virus system is and found an even affordable and effective way to prevent the advanced malware even before installation of
- Many new social engineering techniques were discussed likes reverse social engineering.
- This paper helps the reader to gain insight into how social engineering can be used against enterprises and normal windows users which make them install malicious applications

CHAPTER- 3
ARCHITECTURE DESIGN AND PROPOSED SYSTEM

Now a days an efficient, robust and scalable malware recognition module is the key component of every cybersecurity product. Machine learning Malware recognition modules decide if an application is a threat, based on the previous data that popular anti-virus collected. This data may be collected at different phases:

Pre-execution phase data is anything you can talk about a file without executing it. This may include executable file format descriptions, code descriptions, binary data statistics, text strings and information extracted via code emulation and other similar data.

Post-execution phase data conveys the information about behaviour or events caused by a process activity in a given system**.** To understand more about these detection methods first I wanted every one to know how these malwares are created and send to a particular a victim by using different social engineering techniques**.**

Creating a very deadly trojan which runs on java platform.

3.1 WHAT IS A TROJAN

A remote access Trojan (RAT) is a malware program that includes a backdoor for administrative control over the target computer.

3.2 HOW TROJAN WORKS?

- Trojan viruses work by taking advantage of a lack of security knowledge by the user and security measures on a computer, such as an antivirus and antimalware software program. A Trojan typically appears as a piece of malware attached to an email. The file, program, or application appears to come from a trusted source.
- Trojans create a virtual "backdoor" to a computer that allows hackers remote access to the computer. As such, hackers can download user data and easily steal it. **Even worse, a backdoor allows a attacker to upload additional malware to thedevice.**

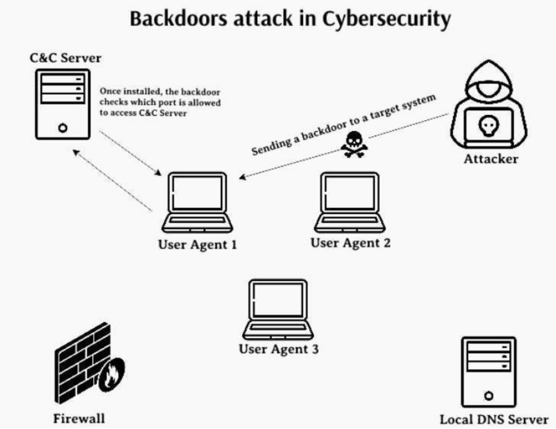

Backdoors attack in Cybersecurity
C&C Server
Once installed, the backdoor checks which port is allowed to access C&C Server
Sending a backdoor to a target system
Attacker
User Agent 1
User Agent 2
User Agent 3
Firewall
Local DNS Server

Fig 5 back door in malware

3.3 WHAT CAN A TROJAN DO?

- Monitoring user behavior through **KEYLOGGERS**
- Display monitoring
- Accessing confidential information, such as credit card and social security numbers.
- Activating a system's webcam and recording video.
- Taking screenshots.
- Distributing viruses and other malware.
- Getting full control on browser, CMD, file manager
- Formatting drives.
- Deleting, downloading or altering files and file systems.

3.4 WHAT IS A KEYLOGGER

- Keyloggers are a type of monitoring software designed to record keystrokes made by a user.
- Data captured by keyloggers can be sent back to attackers via email or uploading log data to predefined websites, databases, or FTP servers. If the keylogger comes bundled within a large attack, attacker might simply remotely log into amachine to download keystroke data.
- a keylogger which is undetectable which records key strokes of the user is quite a challenging task in the project.
- A keen explanation about the back doors and port forwarding is also discussed in the project to make people understand more about how these remote access trojans main a constant connection with the victim.

3.5 HOW TO CONVERT PYTHON SCRIPT TO AN EXE?

After that this project also discusses how the source code a malware is hidden by converting python script into exe for better delivery methods, I used null soft scriptable install system for this to happen

- Nullsoft scriptable install system converts .py files to executable application

Fig 6 NSIS

- Converting a payload script to an exe makes the malware easy to deliver to various users using different social engineering techniques
- Should convert my .py file into a zip file to use the installer
- NSIS will create an executable file. This process makes the attacker code hidden and undetectable
- Attacker can send these applications to his targets to monitor them

3.6 ANTIVIRUS DETECTION METHODS (POST INSTALLATION METHODS)

After that several post detection methods are keenly shown in the project to make people have a better understanding on how anti-virus system works and shown how these anti-virus system works.

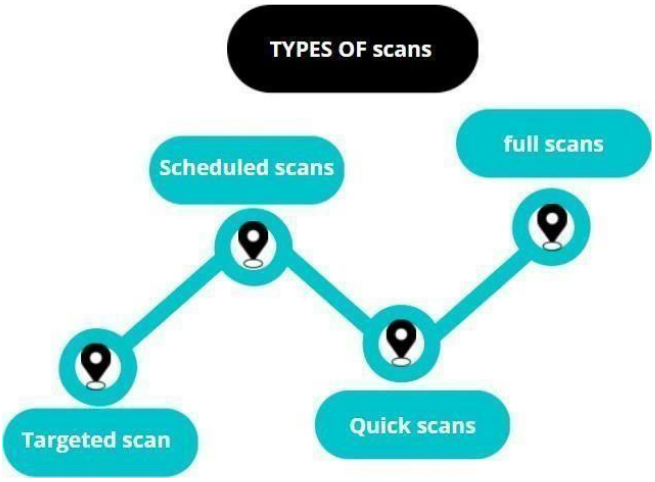
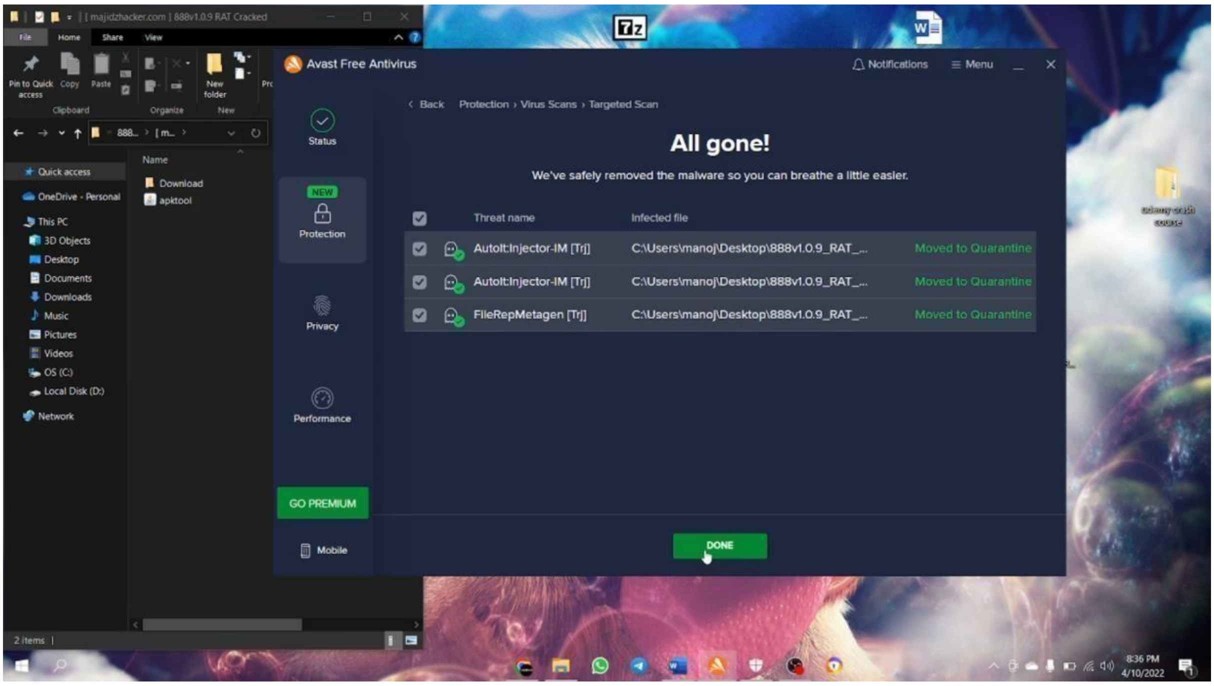


Fig 7 Types of scanning in an antivirus

- **Signature-based detection.** Every program has a signature. A signature is nothing more than a binary pattern. When an antivirus scans a file you downloaded, it looks at its signature and compares it against an extensive database of virus signatures. If it is a known malicious signature, it will warn you immediately.

- **Heuristic-based detection.** analyses behaviour in and patterns of code to see if your file or program is infected. Any Suspicious Code is run in a runtime virtual environment to test it more. This method can find new viruses that are not in your antivirus database.
- **Behavioural-based detection.** Your antivirus is continually looking at your pc. If one of your programs behaves weirdly and is doing harmful things and asking for more read and write permissions, your antivirus will see that and let you know.
- **Sandbox detection** – If your antivirus doesn't know for sure if a program or file has a virus and it suspects something, it will run it in a sandbox. Similar to behavioural-based detection, Sandbox detection will let your file run in a virtual machine and see how it behaves.



All gone!

We've safely removed the malware so you can breathe a little easier.

☑	Threat name	Infected file	
☑	AutoIt:Injector-IM [Trj]	C:\Users\manoj\Desktop\888v1.0.9_RAT_...	Moved to Quarantine
☑	AutoIt:Injector-IM [Trj]	C:\Users\manoj\Desktop\888v1.0.9_RAT_...	Moved to Quarantine
☑	FileRepMetagen [Trj]	C:\Users\manoj\Desktop\888v1.0.9_RAT_...	Moved to Quarantine

Fig 8 Malware found and deleted

3.7 PROPOSED SYSTEM:

- To analyze different features of a malware by creating an own trojan
- To discuss how the most dangerous malware is sent to the victims
- To analyze the patterns which most of malicious malware are following.
- To develop a machine learning model which distinguish whether the application is a threat to the computer or not even **before installation of the software (pre installation detection)**

3.8 ARCHITECTURE DIAGRAMS

This phase in includes creation of trojan creation to understand more about malware and its working and how it connects to victim’s device via network

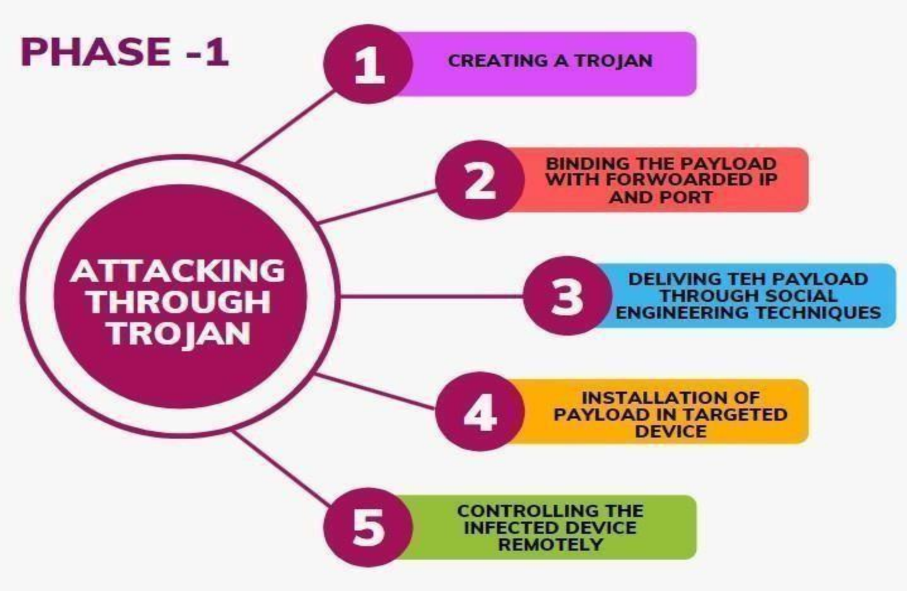

PHASE -1
ATTACKING THROUGH TROJAN
1
CREATING A TROJAN
2
BINDING THE PAYLOAD WITH FORWOARDED IP AND PORT
3
DELIVING TEH PAYLOAD THROUGH SOCIAL ENGINEERING TECHNIQUES
4
INSTALLATION OF PAYLOAD IN TARGETED DEVICE
5
CONTROLLING THE INFECTED DEVICE REMOTELY

Fig 9 Phase 1 architecture

This phase in includes creation of keylogger and converting it to exe using null soft scriptable install system for source code hiding and better delivery methods

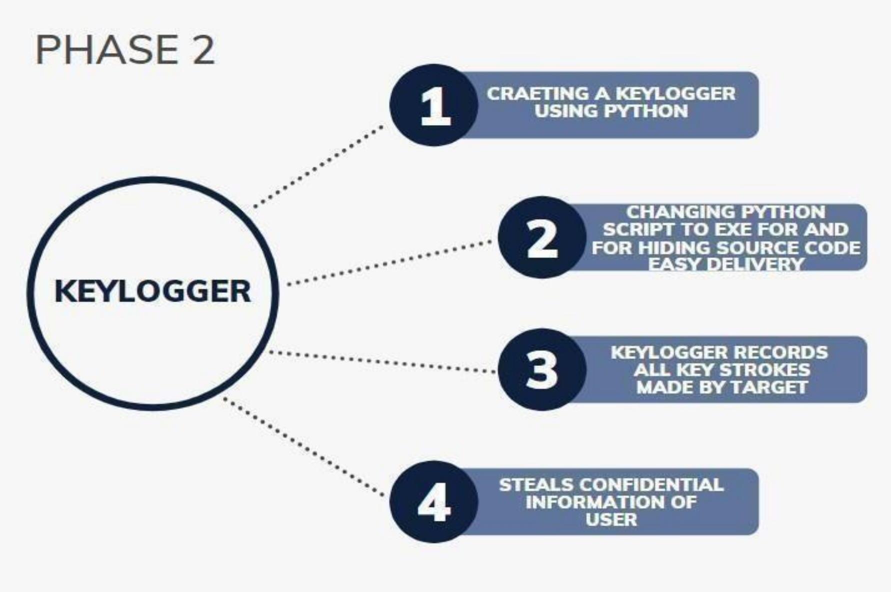

PHASE 2
KEYLOGGER
1
CRAETING A KEYLOGGER USING PYTHON
2
CHANGING PYTHON SCRIPT TO EXE FOR AND FOR HIDING SOURCE CODE EASY DELIVERY
3
KEYLOGGER RECORDS ALL KEY STROKES MADE BY TARGET
4
STEALS CONFIDENTIAL INFORMATION OF USER

Fig 10 Phase 2 architecture

This phase discusses the post installation detection methods and how they work and how these methods work in antivirus by removing malware in real time

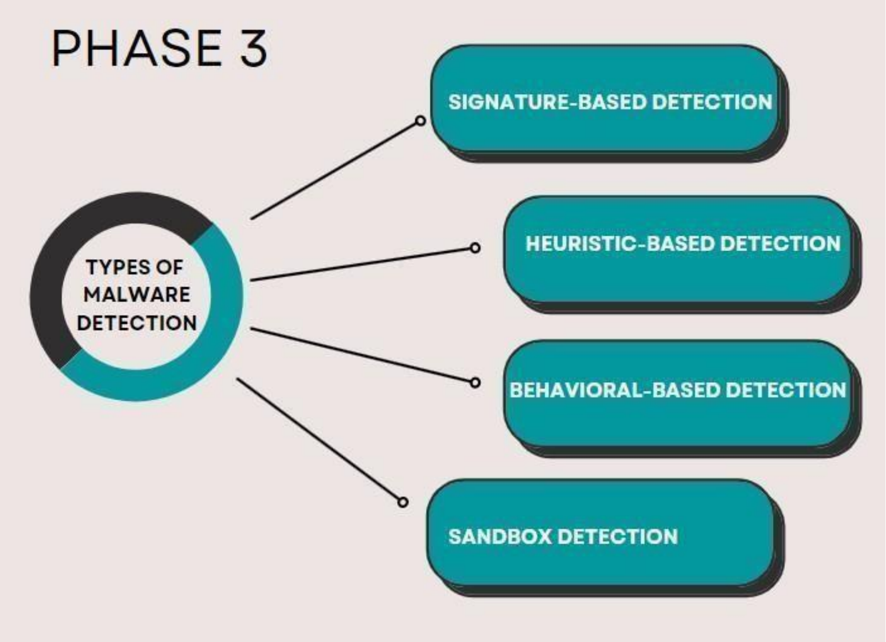

PHASE 3
TYPES OF MALWARE DETECTION
SIGNATURE-BASED DETECTION
HEURISTIC-BASED DETECTION
BEHAVIORAL-BASED DETECTION
SANDBOX DETECTION

Fig 11 Phase 3 architecture

This is the future work of the project to collect data about the features of the malware out there and build a machine learning model to distinguish malware applications even before the installation of the application

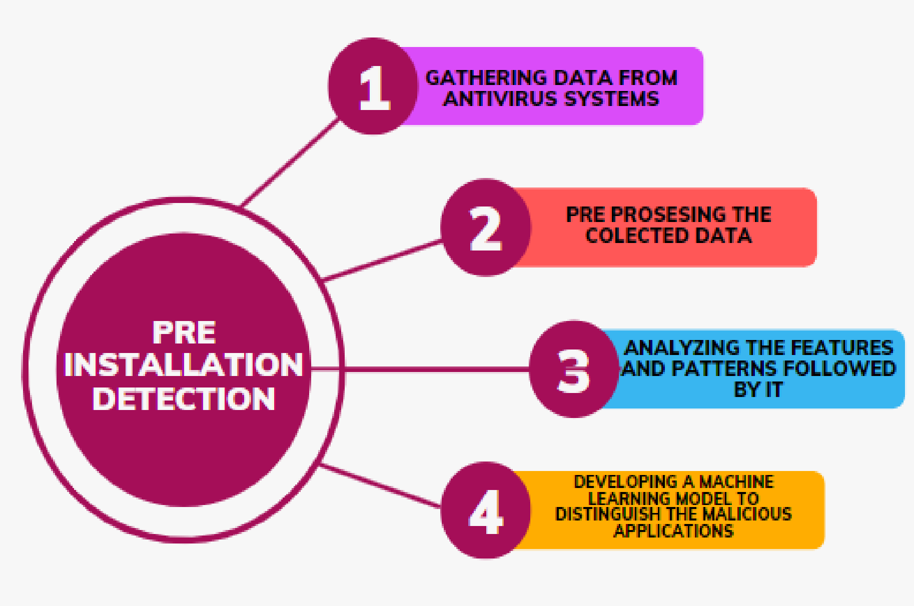

PRE INSTALLATION DETECTION
1
GATHERING DATA FROM ANTIVIRUS SYSTEMS
2
PRE PROSESING THE COLECTED DATA
3
ANALYZING THE FEATURES AND PATTERNS FOLLOWED BY IT
4
DEVELOPING A MACHINE LEARNING MODEL TO DISTINGUISH THE MALICIOUS APPLICATIONS

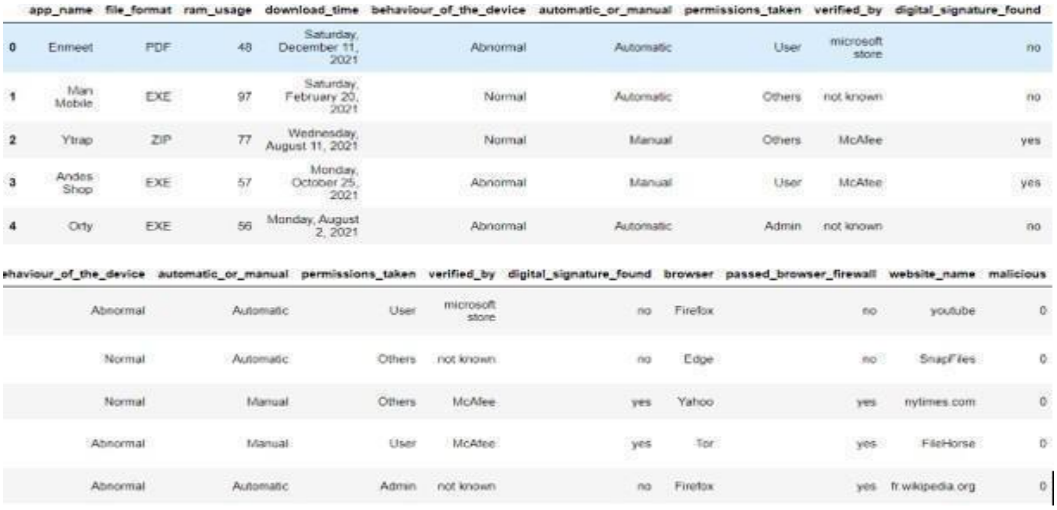
Fig 12 Phase 4 architecture

CHAPTER 4

PRE-INSTALLATION DETECTION THROUGH MACHINE LEARNINIG

4.1 DATA SET

The data set contains a wealth of information about the applications that are understandable to the average user and contains a large number of malware samples. Machine-learning algorithms will differentiate between safe and malicious applications based on these attributes and the features that assist us in the process include the Application name, downloaded file type, RAM use, downloaded time, the behaviour of the device after download, automated or manual installation,permissions requested during installation confirmed by and sources, whether or nota digital signature was detected, the browser, passes or fails the browser firewall, website name, or source of the downloaded program, and, malicious or not being the target variable..



	app_name	file_format	ram_usage	download_time	behaviour_of_the_device	automatic_or_manual	permissions_taken	verified_by	digital_signature_found
0	Enmeet	PDF	48	Saturday, December 11, 2021	Abnormal	Automatic	User	microsoft store	no
1	Man Mobile	EXE	97	Saturday, February 20, 2021	Normal	Automatic	Others	not known	no
2	Ytrap	ZIP	77	Wednesday, August 11, 2021	Normal	Manual	Others	McAfee	yes
3	Andes Shop	EXE	57	Monday, October 25, 2021	Abnormal	Manual	User	McAfee	yes
4	Orty	EXE	56	Monday, August 2, 2021	Abnormal	Automatic	Admin	not known	no

ehaviour_of_the_device	automatic_or_manual	permissions_taken	verified_by	digital_signature_found	browser	passed_browser_firewall	website_name	malicious
Abnormal	Automatic	User	microsoft store	no	Firefox	no	youtube	0
Normal	Automatic	Others	not known	no	Edge	no	SnapFiles	0
Normal	Manual	Others	McAfee	yes	Yahoo	yes	nytimes.com	0
Abnormal	Manual	User	McAfee	yes	Tor	yes	FileHorse	0
Abnormal	Automatic	Admin	not known	no	Firefox	yes	fr.wikipedia.org	0

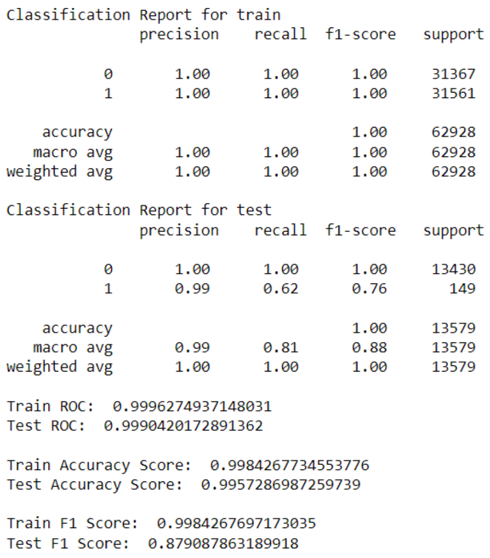
Fig 13 Data set containing information about apps.

4.2 ALGORITHMS USED

I implemented several machine algorithms in order to improve efficiency and accuracy like

- **Gradient boost**

The machine learning boosting system known as gradient boosting represents a decision tree for large and complex data. It is predicated on the idea that the next model will lower the overall prediction error when combined with the previous set of models. Decision trees are used to make the most accurate predictions. The gradient boosting method is also known as the statistical predictive algorithm. Even though it allows for the generalization and optimization of divergent loss functions, it still functions largely in the same way as earlier boosting methods. Processes for classification and regression frequently use gradient boosting.

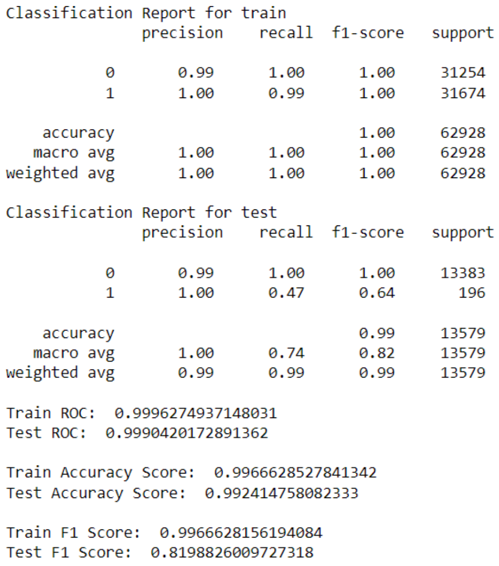


```
Classification Report for train
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     31367
           1       1.00      1.00      1.00     31561

    accuracy                           1.00     62928
   macro avg       1.00      1.00      1.00     62928
weighted avg       1.00      1.00      1.00     62928

Classification Report for test
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     13430
           1       0.99      0.62      0.76       149

    accuracy                           1.00     13579
   macro avg       0.99      0.81      0.88     13579
weighted avg       1.00      1.00      1.00     13579

Train ROC:  0.9996274937148031
Test ROC:  0.9990420172891362

Train Accuracy Score:  0.9984267734553776
Test Accuracy Score:  0.9957286987259739

Train F1 Score:  0.9984267697173035
Test F1 Score:  0.879087863189918
```

Figure 14: Results of gradient boost algorithm

- **Logistic regression**

 Logistic regression is a well-known Machine Learning algorithm from the Supervised Learning method. It forecasts the categorical variables based on a set of unconventional variables. A categorical dependent variable's output is predictedusing logistic regression. As a result, the outcome must be categorical or discrete. It can be zero or one, true or False, and so on, but rather than presenting the actual values like 0 and 1, it presents the probability values that fall between zero and one.

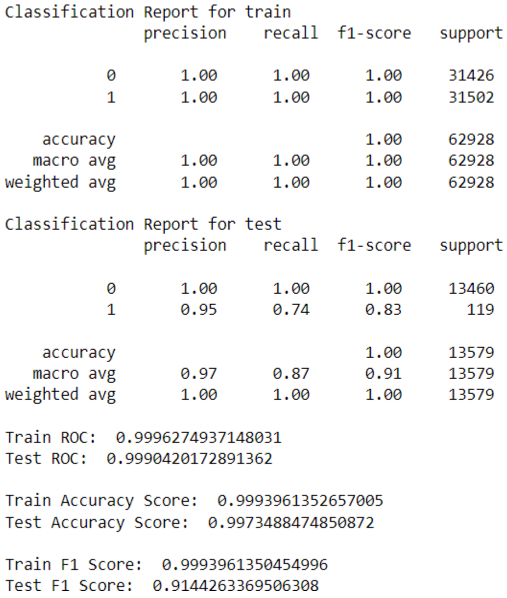


```
Classification Report for train
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     31254
           1       1.00      0.99      1.00     31674

    accuracy                           1.00     62928
   macro avg       1.00      1.00      1.00     62928
weighted avg       1.00      1.00      1.00     62928

Classification Report for test
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     13383
           1       1.00      0.47      0.64       196

    accuracy                           0.99     13579
   macro avg       1.00      0.74      0.82     13579
weighted avg       0.99      0.99      0.99     13579

Train ROC:  0.9996274937148031
Test ROC:  0.9990420172891362

Train Accuracy Score:  0.9966628527841342
Test Accuracy Score:  0.992414758082333

Train F1 Score:  0.9966628156194084
Test F1 Score:  0.8198826009727318
```

Figure 15: Results of logistic-regression algorithm

- **Random Forest**

  Random Forest is a popular supervised machine learning technique. It can be used in machine learning to solve classification and regression problems. It is based on the concept of ensemble learning, which is also the method of combining multiple classifiers to address a complex problem and improve the model's performance. Random Forest is a classifier that averages different decision trees on different subsets of a given dataset to improve the dataset's projected accuracy. Rather than relying solely on one decision tree, the random forest forecasts the correct outcome by taking into account the predictions made by each tree.



```
Classification Report for train
               precision    recall  f1-score   support

           0       1.00      1.00      1.00     31426
           1       1.00      1.00      1.00     31502

    accuracy                           1.00     62928
   macro avg       1.00      1.00      1.00     62928
weighted avg       1.00      1.00      1.00     62928

Classification Report for test
               precision    recall  f1-score   support

           0       1.00      1.00      1.00     13460
           1       0.95      0.74      0.83       119

    accuracy                           1.00     13579
   macro avg       0.97      0.87      0.91     13579
weighted avg       1.00      1.00      1.00     13579

Train ROC:  0.9996274937148031
Test ROC:  0.9990420172891362

Train Accuracy Score:  0.9993961352657005
Test Accuracy Score:  0.9973488474850872

Train F1 Score:  0.9993961350454996
Test F1 Score:  0.9144263369506308
```

Figure 16: Results of random forest algorithm

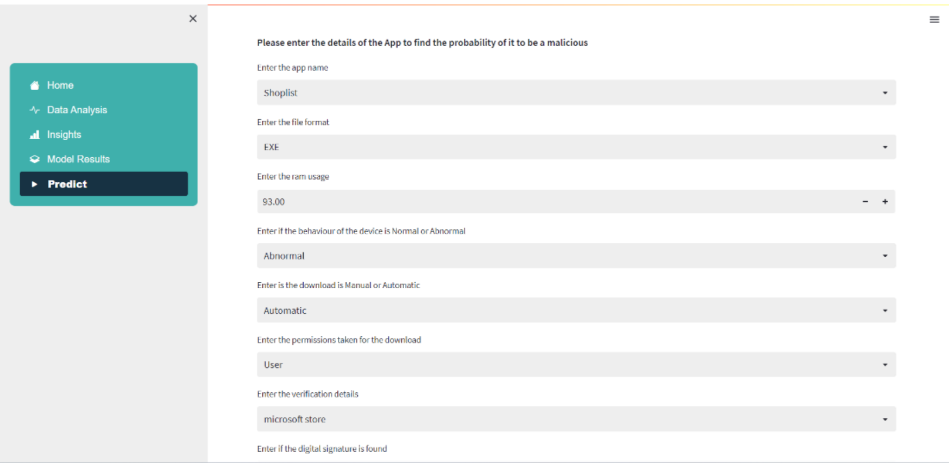
4.3 USER INTERFACE WITH STREAMLIT

Streamlit is an open-source framework used to create user interfaces for machine learning models designed to detect malware before installation. It allows users to input information about the downloaded application, such as file size, name, format, and whether it has been verified by a trusted source. The user interfaces also allow for inputting the source of the downloads, browsers used to download the applications, whether the downloads passed the Chrome firewall or not, and whether the applications have a digital signature or not.

One of the benefits of using Streamlit is its simplicity, allowing for easy creation of interactive web-based interfaces for machine learning models without the need for extensive web development experience. The interfaces allow users to input data and see the results of the models' predictions in real-time, making it easy for them to understand and use. Additionally, Streamlit offers a wide range of built-in components such as sliders, dropdown menus, and text fields, which can be easily incorporated into the interfaces, making them more user-friendly.

To ensure a smooth experience, a help section in the interface is created to provide detailed information about the information that needs to be provided and how to provide it in case the user gets stuck or confused. This feature helps to resolve any confusion and guides the user through the process of providing the necessary information to the model.

In summary, Streamlit is a powerful tool used to build user-friendly interfaces for machine learning models to detect malware before installation. Its simplicity and built-in components make it an ideal choice for projects, allowing for the creation of functional and easy-to-use interfaces that make it easy for users to input information and see the results of the models' predictions. And with the added feature of a separate space to guide the user, the experience is made as seamless as possible.



Please enter the details of the App to find the probability of it to be a malicious

Enter the app name: Shoplist

Enter the file format: EXE

Enter the ram usage: 93.00

Enter if the behaviour of the device is Normal or Abnormal: Abnormal

Enter is the download is Manual or Automatic: Automatic

Enter the permissions taken for the download: User

Enter the verification details: microsoft store

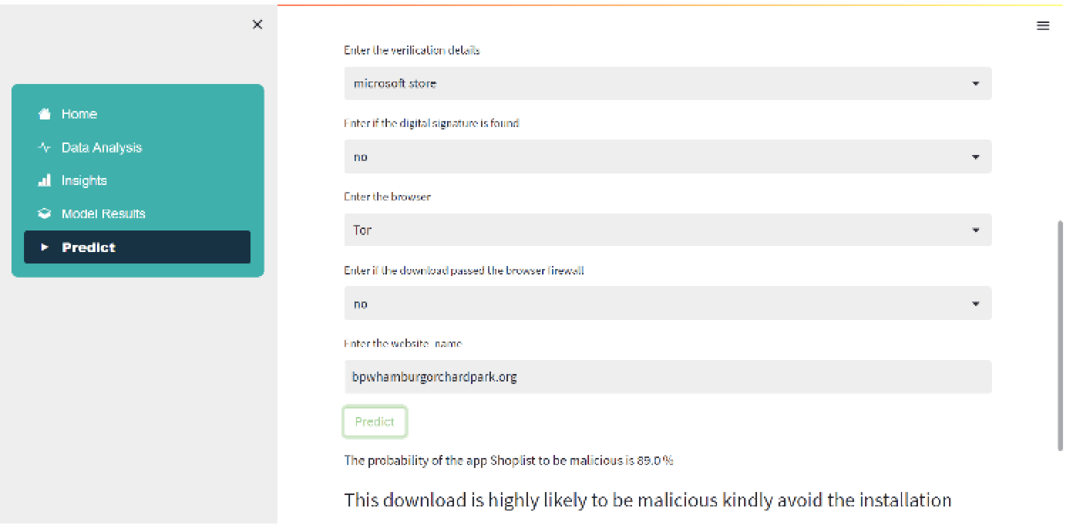
Enter if the digital signature is found

Figure 17: the user interface of running with a machine learning model

CHAPTER 5

RESULTS AND CONCLUSION

Developed a machine learning model that predicts malicious applications with an accuracy of 99%. The random forest algorithm is chosen as the champion model based on the highest f1-score for the test dataset. To determine how safe the downloaded application is, the user must provide specific information on the streamlit interface built for the same.



Enter the verilication details

microsoft store

Enter if the digital signature is found

no

Enter the browser

Tor

Enter if the download passed the browser firewall

no

Enter the website name

bpwhamburgorchardpark.org

Predict

The probability of the app Shoplist to be malicious is 89.0 %

This download is highly likely to be malicious kindly avoid the installation

Figure 18: final result of the proposed system

This project described numerous hacking methods and their defenses from diverse angles. Hackers must be kept out of the network in order to secure sensitive data. The existing system is expensive and time-consuming compared to the proposed system. Machine learning plays a key role in this project for predicting the threat levelof an application without even installing the application. I believe this project helps readers enhance their understanding of cybersecurity and address security vulnerabilities in their computer operations. It also assists in the transmission of knowledge about emerging security concerns. Prevention is preferable

CHAPTER 6

REFERENCES

[1] Yu, W., Yalin, Y., & Haodan, R. (2019, October). Research on the technology of trojan horse detection. In 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 117-119). IEEE.

[2] Al-Asli, M., & Ghaleb, T. A. (2019, April). Review of signature-based techniques in antivirus products. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-6). IEEE.

[3] Asish, M. S., & Aishwarya, R. (2019, March). Cyber security at a glance. In 2019 Fifth International]Conference on Science Technology Engineering and Mathematics (ICONSTEM) (Vol. 1, pp. 240-245). IEEE.

[4] Rathore, H., Agarwal, S., Sahay, S. K., & Sewak, M. (2018, December). Malware detection using machine learning and deep learning. In International Conference on Big Data Analytics (pp. 402-411). Springer, Cham.

[5] Namanya, A. P., Cullen, A., Awan, I. U., & Disso, J. P. (2018, August). The world of Malware: An overview. In 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud) (pp. 420-427). IEEE.

[6] Bhardwaj, A., & Goundar, S. (2020). Keyloggers: silent cyber security weapons. Network Security, 2020(2), 14-19.

[7] Parthy, P. P., & Rajendran, G. (2019, October). Identification and prevention of social engineering attacks on an enterprise. In 2019 International Carnahan Conference on Security Technology (ICCST) (pp. 1-5). IEEE.

[8] Kunwar, R. S., & Sharma, P. (2016, April). Social media: A new vector for cyber- attack. In 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring) (pp. 1-5). IEEE.

[9] Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. IEEE Access, 8, 6249-6271.

[10] Idika, N., & Mathur, A. P. (2007). A survey of malware detection Purdue University, 48(2), 32-46.

[11] Creutzburg, Reiner (2017). "The strange world of keyloggers - an overview, Part I". Electronic Imaging. 2017 (6): 139–148.

[12] Wajahat, A., Imran, A., Latif, J., Nazir, A., & Bilal, A. (2019). A Novel Approach of Unprivileged Keylogger Detection. 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET).doi:10.1109/icomet.2019.8673404.

[13] Rohith, C., & Kaur, G. (2021, April). A Comprehensive Study on Malware Detection and Prevention Techniques used by Anti-Virus. In 2021 2nd InternationalConference on Intelligent Engineering and Management (ICIEM) (pp. 429-434). IEEE.

[14] Nallaperumal, K. (2018). CyberSecurity Analytics to Combat Cyber Crimes. 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). doi:10.1109/iccic.2018.8782430

[15] Alshamrani, S. S. (2022). Design and Analysis of Machine Learning Based Technique for Malware Identification and Classification of Portable Document Format Files. Security & Communication Networks, 2022.