A PROJECT REPORT

on

"MALWARE DETECTION USING MACHINE LEARNING"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN

INFORMATION TECHNOLOGY

BY

DIYA APURVA	1906327
PRIYANSH DUBEY	1906415
SHOHAM DAS	1906434
TANMAY SAHA	1906444

UNDER THE GUIDANCE OF

DR. JAGANNATH SINGH

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

April 2022

A PROJECT REPORT

on

"MALWARE DETECTION USING MACHINE LEARNING"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

DIYA APURVA	1906327
PRIYANSH DUBEY	1906415
SHOHAM DAS	1906434
TANMAY SAHA	1906444

UNDER THE GUIDANCE OF

DR. JAGANNATH SINGH

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA -75102
April 2022

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“MALWARE DETECTION USING MACHINE LEARNING“

submitted by

DIYA APURVA	1906327
PRIYANSH DUBEY	1906415
SHOHAM DAS	1906434
TANMAY SAHA	1906444

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Sci-ence & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2022-2023, under our guidance.

Date: / /

(**DR. JAGANNATH SINGH**)
Project Guide

Acknowledgements

We are profoundly grateful to **DR. JAGANNATH SINGH** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion. ....

DIYA APURVA

PRIYANSH DUBEY

SHOHAM DAS

TANMAY SAHA

ABSTRACT

Research shows that throughout the past ten years, malware have been developing dramatically, making significant monetary misfortunes different associations.Different enemies of malware organizations have been proposing answers to protect assaults from these malware. The speed, volume, and the intricacy of malware is presenting new difficulties to the counter malware local area. Present status-of-the-craftsmanship research shows that as of late, analysts and antivirus associations began applying AI and profound learning techniques for malware examination and location. We have involved opcode recurrence as a component vector and applied solo learning moreover to administer learning for malware characterization. The focal point of this instructional exercise is to introduce our work on identifying malware with different machine learning calculations and profound learning models. That's what our outcomes show the Random Forest beats the Deep Neural Network with opcode recurrence as an element. Likewise in highlight decrease, Deep Auto-Encoders are needless excess for the dataset, and rudimentary capacity like Variance Threshold perform better compared to other people. Notwithstanding the proposed systems, we will likewise talk about the unexpected issues and the one of a kind difficulties in the area, open exploration issues, restrictions, and future bearings.

Vindictive programming or malware is one of the most basic digital danger which disturb and acquire unapproved admittance to the framework. The framework can be a PC, server and PC organization. Windows working frameworks are broadly utilized working frameworks. It is simple for programmers to spread malware in these working frameworks and take advantage of its weaknesses. Malware discovery has been all the time a difficult issue and main pressing issue for information security.Numerous mark based malware discovery techniques have been introduced that work at a specific level and neglect to identify obscure malware executable records, thus the point is to examine a novel methodology which can recognize the new and inconspicuous malware.In this paper, a basic and effective malware identification model is presented which recognizes harmless and noxious executable records by removing highlights from the PE (versatile executable) headers. Different AI techniques such as Support Vector Machine, Decision Tree, Random Forest and Guileless Bayes classifiers are utilized for the characterization. Arbitrary woods classifier among the various classifiers has accomplished the most elevated precision result with the dataset of document, discretionary and area header.

Keywords: Network protection,Profound Learning,Versatile Executable,Static Analysis,Auto-Encoders

Contents

List of Figures

Chapter 1

Introduction

In the computerized age, malware has affected an enormous number of devices.The term malware comes from noxious programming which is intended to meet the destructive goal of a malignant assailant. Malware can think twice about/shrewd gadgets, take classified data, enter organizations, and disabled people's basic frameworks, and so on. These projects incorporate infections, worms, trojans, spyware, bots, rootkits, ransomware, etc.Malware is pernicious programming that has been explicitly intended to go after the frameworks.

The primary point of spreading the malware is to take the touchy and private information.It is a piece of code/program which is purportedly intended to disturb the working of the framework and can make advanced disorder.Malware is for the most part written to effectively gain remote admittance to the tainted framework by the attacker.Malware can be hypothetically partitioned into a few classifications as indicated by their vindictive objectives and ways of behaving. Adware,Scareware, Trojan, Spyware, Virus, Botnet and Worms are a portion of the malware types relying upon their working and host type.Antivirus programming (like Norton, McAfee, Avast, Kaspersky, AVG, Bitdefender, and so forth) is a significant line of safeguard for malware assaults.

Customarily,Antivirus programming utilized the mark based technique for malware recognition.Mark is a short succession of bytes which can be utilized to distinguish known malware. In any case, the mark based identification framework can't give protection from zero-day assaults. Likewise, malware age tool compartments like Zeus [1] can create a great many variations of the equivalent malware by utilizing different confusion strategies. Signature age is much of the time a human-driven process which is bound to become infeasible with the current malware development.Current malware identification procedures are frequently inadequate and neglect to distinguish new kinds of malware.

It is a significant worry that regardless of the advancement of hostile to malware strategy there is development of new malware and is consistently developing step by step. Subsequently, successful and effective mechanized malware identification and reaction procedures are of basic significance to the information safety.The proposed model is planned in view of a static examination approach.

The model is isolated into four primary parts: Sample Assortment, Feature Extraction, Splitting the dataset and Executable record grouping. It identifies malware before execution of the executable record. For the malware investigation different directed AI based

order techniques are utilized to group whether the source executable document has a place with harmless class or malware class. The presentation of the proposed framework is assessed by utilizing the different exhibition assessment measurements like accuracy, review, accuracy,true positive rate and bogus positive rate. Among the different order strategies, irregular woods classifier has accomplished the most noteworthy precision with joined elements of record, discretionary and segment headers. The Internet has turned into a significant piece of our daily existence and we are becoming reliant upon it. In any case, its rising use has likewise passed on the aggressors to abuse and mishandle it.

In the web world, there is an emotional expansion in the spread of malware on applications like internet browsers, media players and numerous different applications. At the point when a client introduces an application, it requests that the client acknowledge every one of the authorizations. Assailants generally request the consent through which they can access the private data. Clients know nothing about this reason for assailants and they acknowledge every one of the authorizations at the hour of use establishment and afterward they become the casualty of the assault. Subsequently, we want a viable technique to secure the secret information from any danger.

Chapter 2

Literature Review

2.1 **Evaluation of machine learning classifiers for mobile malware detection**

According to an article in statista India has 744 millions internet user,majority of which access the internet via their smartphones,it is also possible that the figure might reach 1.5 billion by the end of 2050.Therefore protection of all these users become all the more important.With the point of confronting malware,mobile gadgets have embraced customary methodologies, for example, the antivirus. This strategy isn't as productive against cell phone malware, as it requires a constant mark information base refresh. What's more, versatile malware is continually adjusted to dodge the different identification techniques.

This paper was written by Fairuz Amalina Narudin,Ali Feizollah,Nor Badrul Anuar and Abdulla Ghani where they suggest a solution to evaluate malware detection by the use of anomaly-based approach with machine learning classifiers.Among the different organization traffic includes, the four classes chose are fundamental data,content based, time based and association based. The assessment uses two datasets: public (MalGenome) and private (self-gathered). In view of the assessment results,both the Bayes organization and arbitrary timberland classifiers created more exact readings, with a 99.97% genuine positive rate instead of the multi-layer perceptron which was 93.03% on the MalGenome dataset. Nonetheless, this investigation uncovered that the k-closest neighbor classifier productively identified the most recent Android malware with a 84.57% true positive rate higher than different classifiers.

2.2 A Machine Learning Approach to Android Malware Detection

With the new development of versatile stages equipped for executing progressively complex programming and the rising omnipresence of involving portable stages in delicate applications, for example, online exchanges, there is a rising risk related with malware focused on at portable devices.In this article, the writers Justin Sash and Latifur Khan present an AI based framework for the identification of malware on Android devices.The framework removes various elements and trains a One-Class Support Vector Machine in a off-gadget way.

Regardless of the fast development of the Android stage, there are as of now very much archived instances of Android malware, for example, DroidDream, which was found in north of 50 applications on the authority Android market in March 2011.Furthermore, it was found that Android's built in security highlights are generally inadequate, and that even non malignant projects can uncover classified data. An investigation of 204,040 Android applications observed 211 malignant applications on the authority android market and option marketplaces.

The issue of utilizing an AI based classifier to recognize malware presents two principle challenges: first, given an application, it is must to extricate some kind of component portrayal of the application; second, there is an informational collection that is practically protected, so we should pick a classifier that can be prepared on just a single class.To address the main issue, a heterogeneous list of capabilities is separated and process each element freely utilizing different kernels.To address the subsequent issue, a One-Class Support Vector Machine is utilized, which can be prepared utilizing just harmless applications.

2.3 OPEM: Malware Detection using a Static-Dynamic Supervised Malware Detectors

The most common commercial way of detecting malware is Signature-Based Detection, but it constantly fails to detect malwares.The method adopted to mitigate these issues is supervised malware detection.The two popular types of supervised malware detection are static and dynamic detectors.

The way machine learning malware detections work is that they rely on the datasets containing previously samples and elements of potentially harmful malwares and compare the present suspected malware to those patterns.Static malware detectors extracts and analyzes the suitable information and parts of the executables without actually running it. It is a faster and safer way of malware detection as no execution of the malicious software is done.However,Static detectors fail to grasp the malicious components of executables that are compressed or encrypted.

Thats where Dynamic Malware Detectors come in, dynamic detectors executed the executables in a safe controlled environment called the ‘Sandbox’.Dynamic detectors are very efficient as there can simulate the intended result of a executable and deem it harmful or harmless.But, they have their own shortcomings as well, dynamic detectors can facilitate the executable to learn about the environment they are ran in and become resilient to it.Moreover, they also consume a lot of resources and time to execute the program.So, there was a attempt to combine the features of both the methods of malware detections which is know as, OPEM.

In OPEM approach both Static and Dynamic employes the set of features of both the methods of malware detection.In Static detection the malware's code is divided in number of codes of fixed length and frequency of their occurrence .These divisions are called as opcode sequences.In the dynamic approach the executable in run inside a controlled and secure environment called a sandbox.Once inside a Sandbox one can proceed in two ways, Firstly the system can take a snapshot of the system before the execution of the code and one after the execution of the code.Then you can compare the two states, find the difference in the states and estimate whether it was positive or negative.Secondly, one can use special programs to analyze the executable during the running of the code and find the discrepancies.

2.4 Malware detection using Machine Learning

This paper was written by Dragos¸Gavrilut ,Mihai Cimpoes , Dan Anton, Liviu Ciortuz1 and here they propose a versatile framework wherein one can utilize different AI calculations to effectively recognize malware records and clean documents, while pointing to limit the quantity of misleading positives.This paper presents the thoughts behind our system by working initially with overflow uneven perceptrons and besides with overflow kernelized uneven perceptrons. In the wake of having been effectively tried on medium-size datasets of malware and clean records, the thoughts behind this system were submitted to an increasing process that empower us to work with exceptionally enormous datasets of malware and clean documents.

This approach was utilized as malware recognition through norm, signature based strategies were getting increasingly more troublesome since all current malware applications will quite often have numerous polymorphic layers to keep away from discovery or to utilize side systems to consequently update themselves to a more current form at brief timeframes to stay away from location by any antivirus programming.

Chapter 3

Malware (brief for “malicious software”) could be a file or code, regularly conveyed over an arrangement, that contaminates, investigates, takes or conducts virtually any behavior an assailant needs. As a result, malware comes in so numerous variations, there are various strategies to contaminate computational frameworks.

3.1 Types of Malware

Virus -

Viruses are a subgroup of malware. A virus is a malevolent program connected to a document or file that underpins macros to execute its code and spread from host to host. Once downloaded, the virus will lay torpid until the document is opened and in utilize.

Worms -

Worms are a noxious software that quickly reproduces and spreads to any gadget inside the arrangement. Not at all like viruses, worms don't require programs to spread. A worm contaminates a gadget by means of a downloaded file or a network connection some time recently it duplicates and scatters at an exponential rate.

Trojan Horses -

Trojan viruses are masked as supportive software programs. But once the client downloads it, the Trojan virus can pick up to get sensitive information and after that modify, block, or erase the information. This could be greatly hurtful to the execution of the gadget. Not at all like ordinary infections and worms, Trojan viruses are not outlined to self-replicate.

Spyware -

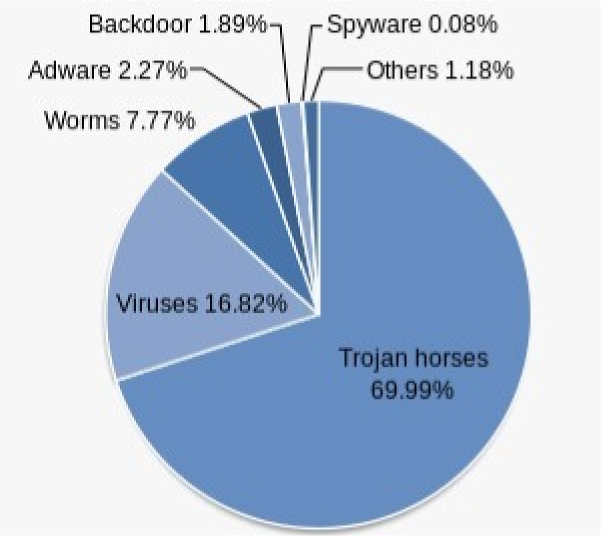
Spyware is malevolent software that runs subtly on a computer and reports back to an inaccessible client. Instead of basically disturbing a device’s operations, spyware targets sensitive data and can give inaccessible access to predators. Spyware is frequently utilized to take money related or individual data. A specific type of spyware may be a keylogger, which records your keystrokes to uncover passwords and individual data.

Adware -

Adware is pernicious software utilized to gather information on your computer utilization and give suitable notices to you. Whereas adware isn't continuously unsafe, in a few cases, adware can cause issues for your framework. Adware can divert your browser to hazardous destinations, and it can indeed contain Trojan steeds and spyware. Moreover, critical levels of adware can moderate down your framework recognizably. Since not all adware is noxious, it is critical to have assurance that continually and intellectuals look at these programs.

Backdoor -

A backdoor may be a malware sort that invalidates typical confirmation strategies to get a framework. As a result, farther get to is allowed to assets inside an application, such as databases and record servers, giving culprits the capacity to remotely issue framework commands and overhaul malware.



3.2 Propagation Of Malware

> Email attachments

Malware is commonly conveyed by means of emails that energize the beneficiary to open a malevolent attachment. The record can be conveyed in various formats, including a ZIP record, PDF, Word report, Excel spreadsheet and others. Once the attachment is opened, the Malware may be sent quickly in other circumstances, attackers may hold up days, weeks or indeed months after contamination to encrypt the victim's records, as was the case within the Emotet/Trickbot assaults.

>Malicious URLs

Attackers moreover utilize email and social media stages to disperse ransomware by inserting malevolent links into messages.
To encourage you to tap on the malevolent links, the messages are ordinarily worded in a way that brings out a sense of urgency or interest. Clicking on the link triggers the download of ransomware, which scrambles your framework and holds your information for ransom

.

>Remote desktop protocol

RDP is a communications protocol that permits you to put through to another computer over an organized association, is another prevalent assault vector. Cybercriminals take advantage of this by utilizing port-scanners to scour the Web for computers with uncovered ports. They at that point attempt to get to the machine by misusing security vulnerabilities or utilizing brute drive assaults to break the machine's login credentials.
Once the attacker has picked up the machine, they can do more or less anything they wish. Regularly this includes disabling your antivirus program and other security arrangements, erasing available backups and conveying the ransomware. They may also leave a backdoor they can utilize within the future.

>Drive-by downloads

A drive-by download is any download that happens without your information. Malware attackers make utilize of drive-by downloads by either facilitating the malevolent content on their possess location or, more commonly, infusing it into authentic websites by misusing known vulnerabilities.
When you visit the contaminated site, the pernicious substance analyzes your gadget for particular vulnerabilities and naturally executes the ransomware within the background.
Unlike numerous other assault vectors, drive-by downloads don't require any input from the user. You don't need to tap on anything, you don't ought to introduce anything and you don't need to open a noxious connection– going to a tainted site is all it takes to ended up contaminated.

3.3 Types of Algorithm methods used

> Random Forest Model

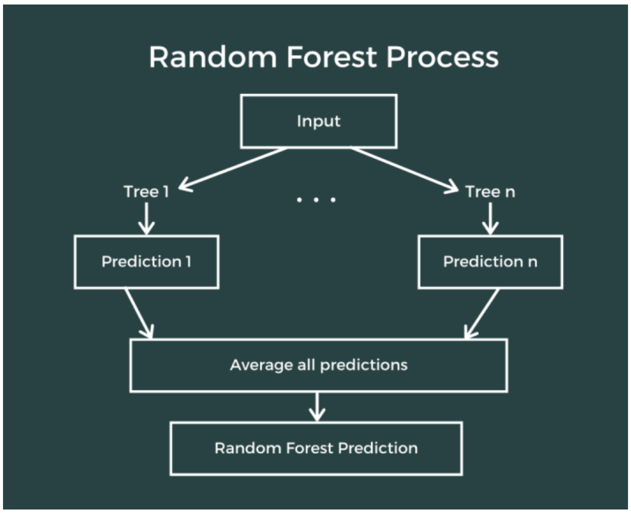
Random forest could be a supervised learning algorithm which uses a gathering learning method for classification and regression. Random woodland may be a sacking method and not a boosting strategy. The trees in random forests is running at the same time. There's no interaction between these trees, whereas building the trees. It works by building a huge number of choice trees at preparing time and yielding the lesson that's the mode of the classes(Classification) or cruel expectation (regression) of the person trees.

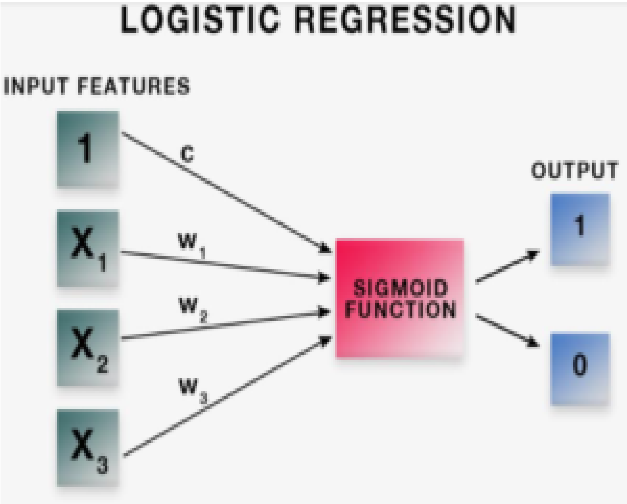
> Logistic Regression

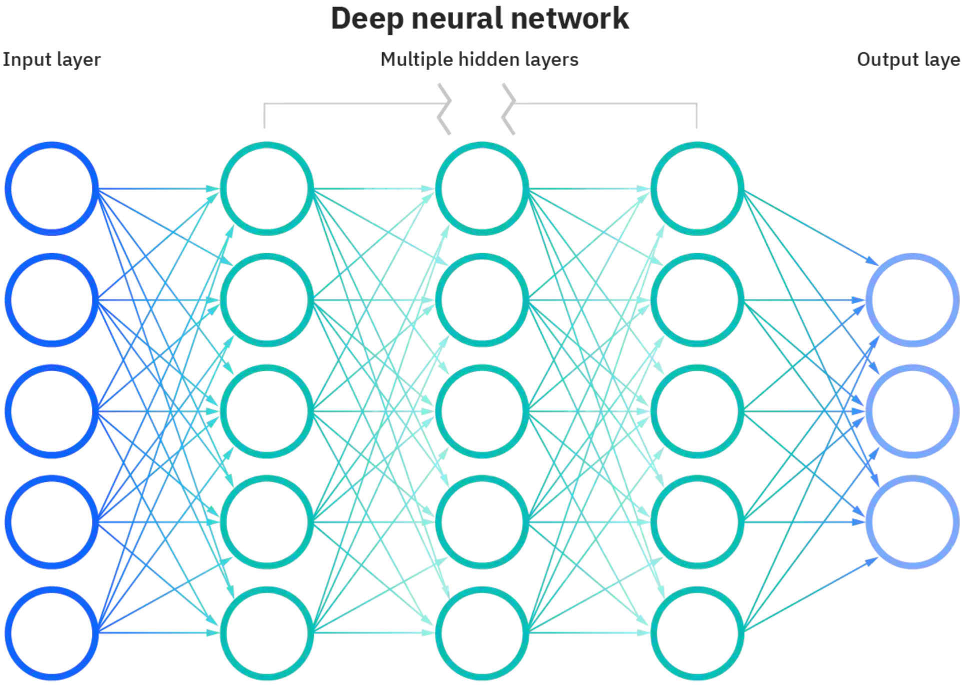
Logistic Regression is a Machine Learning method that's utilized to solve classification issues. It could be a prescient expository method that's based on the likelihood thought. The classification algorithm Logistic Regression is utilized to anticipate the probability of a categorical subordinate variable. The dependant variable in logistic regression may be a binary variable with data coded as 1 (yes,True, normal, success, etc.) or (no, False, abnormal, failure, etc.).

> Neural Networks

A neural network is an arrangement of calculations that endeavors to recognize fundamental connections in a set of information through a prepare that mirrors the way the human brain works.In this sense, neural systems imply to frameworks of neurons, either normal or manufactured in nature.







Chapter 4

Implementation

• Built machine learning and deep learning models to detect malware.

• Make a comparison of all models and select the best model.

4.1 Methodology

First we have imported some relevant libraries to build the machine learning models , then imported the data-set. For this project we have taken a data-set from GitHub: https://github.com/PacktPublishing/Mastering-Machine-Learning-for-Penetration-Testing/blob/master/Chapter03/MalwareData.csv.gz
then we have cleaned the data-set , then we have divided the data-set into train and test data-set then further we have evaluated data-set by Random Forest, Logistic Regression and Neural Network methods to know which method is best.

4. 2 Testing

From testing we see that in data-set has
> 41,323 binaries (exe ,dll) - legitimate
> 96,724 malware files
And in further testing i.e. evaluation by different models we get the accuracy by each model. It is mentioned in the table below

Test ID	Test Case Models	Train Data Accuracy	Test Data Accuracy	F1-Score
T01	Random Forest	0.9828%	0.9858%	0.9730%
T02	Logistic Regression	0.7015%	0.6972%	0.0%
T03	Neural Network	0.9687%	0.9703%	0.9498%

4.3 Result Analysis

From the above result by comparing the overall performance of the model we can see that ***Random Forest Model*** is best for our project.
We can see that Logistic Regression has an F1 score of 0 that is why we reject it.
Neural Networks is the second best model for us but it didn't work efficiently because it needs a lot of data-set it doesn't work on the small type of data-set .For example it will not work on a data-set of few MBs it needs at-least 100 MB or GBs of datasets.

References

1. A Machine Learning Approach to Android Malware Detection - Justin Sahs and Latifur Khan(University of Texas at Dallas) (2012 European Intelligence and Security Informatics Conference)
2. OPEM: A Static-Dynamic Approach for Machine-Learning-Based Malware Detection (S3Lab, DeustoTech - Computing, Deusto Institute of Technology ,University of Deusto) (Published in CISIS/ICEUTE/SOCO Special…CISIS/ICEUTE/SOCO Special Sessions 2012)
3. DRAGOS¸GAVRILUT¸ET. AL: Malware Detection Using Machine Learning (Proceedings of the International Multiconference on Computer Science and Information Technology 2009)
4. Evaluation of machine learning classifiers for mobile malware detection (Springer-Verlag Berlin Heidelberg 2014)
5. arXiv:1904.02441v1 [cs.CR] 4 April 2019
6. Review of Machine Learning Methods for Windows Malware Detection (IEEE 45670)(10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India)
7. https://blog.emsisoft.com/en/35083/how-ransomware-spreads-9-most-common-infection-methods-and-how-to-stop-them/

INDIVIDUAL CONTRIBUTION REPORT

MALWARE DETECTION USING MACHINE LEARNING

PRIYANSH DUBEY
1906415

Abstract:

Built machine learning and deep learning models to detect .

Make a comparison of all models and select the best mode

Individual contribution to project report preparation:

Contributed to Chapter 1 & Chapter 2 Of this Record.
Read Paper

- DRAGOS¸GAVRILUT¸ET. AL: Malware Detection Using Machine Learning (Proceedings of the International Multiconference on Computer Science and Information Technology 2009)
- arXiv:1904.02441v1 [cs.CR] 4 April 2019
- Review of Machine Learning Methods for Windows Malware Detection (IEEE 45670)(10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India)

For Execution did Random Forest REgression.

Priyansh

Full Signature of the Student

Full Signature of the Supervisor

INDIVIDUAL CONTRIBUTION REPORT

MALWARE DETECTION USING MACHINE LEARNING

DIYA APURVA
1906327

Abstract:

Built machine learning and deep learning models to detect .

Make a comparison of all models and select the best mode

Individual contribution and findings:

Contributed to Chapter 2 & Chapter 4 of this Record.
Read Paper

- OPEM: A Static-Dynamic Approach for Machine-Learning-Based Malware Detection (S3Lab, DeustoTech - Computing, Deusto Institute of Technology ,University of Deusto) (Published in CISIS/ICEUTE/SOCO Special…CISIS/ICEUTE/SOCO Special Sessions 2012)
- arXiv:1904.02441v1 [cs.CR] 4 April 2019
- Review of Machine Learning Methods for Windows Malware Detection (IEEE 45670)(10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India)

For Execution found the dataset , cleaned the dataset and further divided the dataset into test and train dataset and at last compared the accuracy rate of all the three models.

Diya Apurva.

Full Signature of the Student

Full Signature of the Supervisor

INDIVIDUAL CONTRIBUTION REPORT

MALWARE DETECTION USING MACHINE LEARNING

SHOHAM DAS
1906434

Abstract:

Built machine learning and deep learning models to detect .

Make a comparison of all models and select the best mode

Individual contribution and findings:

Contributed to Chapter 2 & Chapter 3 Of this Record.
Read Paper

- A Machine Learning Approach to Android Malware Detection - Justin Sahs and Latifur Khan(University of Texas at Dallas) (2012 European Intelligence and Security Informatics Conference)
- arXiv:1904.02441v1 [cs.CR] 4 April 2019
- Review of Machine Learning Methods for Windows Malware Detection (IEEE 45670)(10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India)

For Execution did Logistic Regression.

shoham das

Full Signature of the Student

Full Signature of the Supervisor

INDIVIDUAL CONTRIBUTION REPORT

MALWARE DETECTION USING MACHINE LEARNING

TANMAY SAHA

1906444

Abstract:

Built machine learning and deep learning models to detect .

Make a comparison of all models and select the best mode

Individual contribution and findings:

Contributed to Chapter 1 & Chapter 3 Of this Record.
Read Paper

- Evaluation of machine learning classifiers for mobile malware detection (Springer-Verlag Berlin Heidelberg 2014)
- arXiv:1904.02441v1 [cs.CR] 4 April 2019
- Review of Machine Learning Methods for Windows Malware Detection (IEEE 45670)(10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India)

For Execution did Neural Networks.

Tanmay Saha

Full Signature of the Student

Full Signature of the Supervisor

TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)