

Predictive Analysis of Macro Malwares: A Novel approach for Macro Malware Prediction using predictive modelling and machine learning techniques



By

NS Sehrish Sajid

A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfillment of the requirements for the degree of MS in Information Security

June 2023

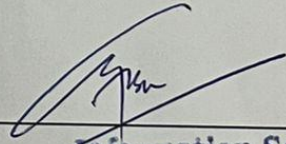
THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by **Sehrish Sajid**, Registration No. **00000320293**, of **Military College of Signals** has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations/MS Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members and local evaluators of the scholar have also been incorporated in the said thesis.

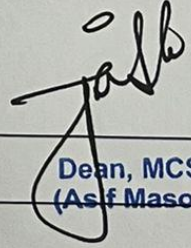
Signature: _____ 

Name of Supervisor **Assoc Prof. Dr. Muhammad Faisal Amjad**

Date: _____ 25-08-23

Signature (HOD): _____  **HoD**

Date: _____ 25-08-23
Information Security
Military College of Signals

Signature (Dean/Principal) _____  **Brig**

Date: _____ 15/9/23
Dean, MCS (NUST)
(Asf Masood, Phd)

CERTIFICATE

This is to certify that **NS Sehrish Sajid** Student of **MSIS** Reg.No **00000320293** has completed her MS Thesis title **“Predictive Analysis of Macro Malwares: A Novel approach for Macro Malware Prediction using predictive modelling and machine learning techniques”** under my supervision. I have reviewed her final thesis copy and I am satisfied with her work.

Thesis Supervisor

(Assoc Prof.Dr. Muhammad Faisal Amjad)

Dated: 23 June 2023

Declaration

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification at this either institution or elsewhere

Dedication

“In the name of Allah, the most Beneficent, the most Merciful”

I dedicate this thesis to my late father, family, and teachers who guided me during each step
of the way

Acknowledgments

All praises to Allah for the strengths and His blessing in completing this thesis.

I would like to convey my gratitude to my supervisor, Assoc Prof. Dr. Muhammad Faisal Amjad for his supervision and constant support. His invaluable help, constructive comments and suggestions throughout the experimental and thesis works are major contributions to the success of this research. Also, I would thank my committee members; Prof Dr Haider Abbas and Engr Muhammad Sohaib Khan Niazi for their support and knowledge regarding this topic.

Last, but not the least, I am highly thankful to my family. They have always stood by my dreams and aspirations and have been a great source of inspiration for me. I would like to thank them for all their care, love and support through my times of stress and excitement.

ABSTRACT

Macros are scripts used in Microsoft office documents to automate tasks written in Visual Basic for Applications (VBA). Malware authors exploit this feature and embed malicious VBA code in office documents to perform malicious activities on victim's computer.

Earlier, macros used to run automatically once an office document was opened. But in recent versions of Microsoft Office, macros are disabled by default and malware authors lure in the users to enable macros using different techniques and strategies. Once macros are enabled, the malicious code embedded within the file runs automatically and execute the malware as the malware authors intends to.

Macro malware authors use different social engineering techniques to tempt or scare users into downloading and opening them. These may be downloaded to a victim's computer by merely opening an email, an email attachment, or by performing some other usual normal operations, such as clicking a graphic to expand it in an email you receive which are usually embedded in Microsoft office files. The files often use names that entice users into opening them such as invoices, receipts, legal documents etc.

Apart from virus detection programs, different machine learning techniques have been developed in the past for detection and mitigation of macro-based malwares. But at the same time malware authors come up with more advanced evasion and obfuscation techniques to evade the detection methods created on basis of machine learning techniques.

Mostly research has been carried out on approaches to detect and mitigate malware threat. These come into play once a malware has been downloaded or executed on a system. A malware evading these techniques will infect the system, causing data theft or loss and may require substantial effort to recover data and remove the malware threat from the system. This threat becomes more pronounced depending upon the sensitivity and importance of target system; for example, a transaction server in a bank, an IT system or some Government organization.

This research focuses on Malware prediction, an emerging concept which uses AI and Machine Learning techniques to analyze an organization's web traffic and behavior to predict if and when a machine will be targeted by a Malware attack. Algorithms are used to analyze an organization's dataset that contains real samples to provide a better approach for prediction of malware attack on a system. In this research, we will mainly focus on malwares in general to develop a framework, which will later be tested on a macro malware dataset.

Table of Contents

Declaration	iv
Dedication	v
Acknowledgments	vi
ABSTRACT	vii
INTRODUCTION	1
1.1 Problem Statement	2
1.2 Motivation	3
1.4 Significance for Pakistan	4
1.5 Contributions	5
1.6 Thesis Outline	5
EXISTING RESEARCH IN MALWARES PREDICTION USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE	7
Introduction	7
2.1. Level of Research Already Carried out on the Proposed Topic	8
PRELIMINARY BACKGROUND	11
Introduction	11
History of AI	11
3.1 Machine Learning, Deep Learning and AI	13
3.2 Key Concepts in Machine Learning and AI	15
3.3 Sub-branches of Machine Learning	16
3.4 Time Series Analysis	19
3.5 How Machine Learning Works	20
3.6 Applications of Machine Learning and AI	21
Conclusion	21
MACRO MALWARE AND PREDICTION SYSTEMS	22
4.1 Malicious Macro Malwares embedded in Office Documents	22
4.3 Behavior-based detection	26
4.4 Outline of a Malware Prediction System using Machine Learning and AI	27
Conclusion	29
DESIGN OF A MALWARE PREDICTION SYSTEM USING MACHINE LEARNING AND AI	31
5.1 The Programing Environment	31
5.2 The data set	31
5.3 Sifting of Data	32

Conclusion	37
THE DESIGNED AI MODEL – EVALUATION AND PERFORMANCE ANALYSIS	38
Introduction.....	38
6.1 General Outline.....	38
6.2 Linear Regression Model.....	39
6.2 Random Forest Regressor.....	40
6.3 XGB Regressor.....	41
6.4 Support Vector Regression.....	41
6.5 Ridge Regression.....	42
6.5 Comparative Analysis.....	43
CONCLUSION	46
References.....	48

INTRODUCTION

Macros are scripts used in Microsoft office documents to automate tasks written in Visual Basic for Applications (VBA). Malware authors exploit this feature and embed malicious VBA code in office documents to perform malicious activities on victim's computer.

Earlier, macros used to run automatically upon opening an office document. New versions of Microsoft Office keep the macros disabled by default. Malware writer lure in the users to enable macros using different techniques and strategies. Once macros are enabled, the malicious code embedded within the file runs automatically and execute the malware as the malware authors intends to.

Macro malware authors use different social engineering techniques to tempt or scare users into downloading and opening them. These may be downloaded to a victim's computer by merely opening an email, an email attachment, or by performing some other usual normal operations, such as clicking a graphic to expand it in an email you receive which are usually embedded in Microsoft office files. The files often use names that entice users into opening them such as invoices, receipts, legal documents etc.

Apart from virus detection programs, different machine learning techniques have been deployed in the past for detection and mitigation of macro-based malwares. But at the same time malware authors come up with more advanced evasion and obfuscation techniques to evade the detection methods created on basis of machine learning techniques.

Mostly research has been carried out on approaches to detect and mitigate malware threat. These come into play once a malware has been downloaded or executed on a system. A malware evading these techniques will infect the system, causing data theft or loss and may require substantial effort to recover data and remove the malware threat from the system. This threat becomes more pronounced depending upon the sensitivity and importance of target system; for example, a transaction server in a bank, an IT system or some Government organization.

This research focuses on Malware prediction, an emerging concept which uses AI and Machine Learning techniques to analyze an organization's web traffic and behavior to predict if and when a machine will be targeted by a Malware attack. Algorithms are used to analyze an organization's dataset that contains real samples to provide a better approach for prediction of malware attack on a system. In this research, we will mainly focus on malwares in general to develop a framework, which will later be tested on a macro malware dataset.

1.1 Problem Statement

Organizations depend on IT and Internet to enhance productivity and output. Office documents form a backbone of an organization's business routine and are stored and shared regularly. File attachments can be conveniently shared within and outside an organization on local or web mail as a part of work routing. Malware authors use these documents to embed macros and tempt users into downloading and running these files on their computers. Once a user downloads the file and opens it, the malware author can use it to launch or deploy any kind of attack or virus on user's system. Once a

system gets infected, malware can spread to other computers on a network and effect the work efficiency apart from causing nuisance and potential threat loss or theft. Organizations, therefore, spend hefty amount to evade and mitigate Malware threat. However, if the web traffic and an organization's data is analyzed using Machine Learning and AI and the next potential malware attack can be predicted, it would substantially reduce detection and mitigation effort and would help enhance an organization's productivity. Consequently, organizations can exercise caution and take appropriate steps to thwart the threat of a Malware attack and ensure it causes minimum damage.

1.2 Motivation

Organizations make a lot of effort to ward off the Malware threat by deploying hardware and software solutions. Once an organization faces a Malware threat, there is a good chance that most of its systems are already infected. The IT team now has to make an effort to clean the systems and in the process, may have to disconnect servers, systems or LANs from the main network. This effects productivity and efficiency of an organization.

The main motivation behind selection of this topic is to propose a method, where a likely Malware attack could be predicted by using Machine Learning techniques and build a model which can analyze an organizations traffic and data and predict the likely time of a potential attack.

1.3. Objectives The main objectives of this thesis are:

- a. To carry out a study and develop an understanding of the working of malicious macros in Microsoft documents.
- b. To propose a novel technique for prediction of malicious macro on Microsoft platform using machine learning techniques based on previous original samples.
- c. To carry out validation of the proposed technique.

1.4 Significance for Pakistan

Pakistan's role in regional stability and its geopolitical significance has a target of different non-state actors and Advanced Persistent Threat (APT) Groups. Pakistan's IT infrastructure has remained a target of APT groups who launch attacks and tend to threaten Pakistan's government bodies, military, Intelligence agencies, telecoms and educational institutions.

Malware Infection Index ranks Pakistan as no 1 country, with highest malware encounter rates. Its IT infrastructure remains exposed to multiple threats, specially from India, which is among top 15 VBA spam malware-sending IPs countries, employing thousands of malicious IPs to spread malware worldwide. This threat is amplified by the lower computer literacy, software piracy and poor security awareness of people in general and the IT and cyber security specialists in particular. One such example is the huge cache of Banking data of Pakistani subscribers on the dark web.

In order to bridge this gap, Malware prediction will play a pivotal role by giving a forewarning to Cyber Security professionals regarding an impending attack. Based on a Machine Learning algorithm trained using real-time data collected during the past, such a model can predict and give sufficient time to Cyber security professionals to

prepare and respond appropriately. This mode, therefore, can be deployed in various Government, military and financial organizations and will contribute towards National Cyber defense.

1.5 Contributions

Some of the advantages of this research will be:

- a. Develop an understanding of design, working and infection techniques of Microsoft Office based Macro Malwares.
- b. Understand the work done in the field of macro malware attacks and prediction by providing a detailed literature review.
- c. Propose a framework which would predict a malware attack basing on Machine Learning algorithm which analyses real data. This framework will then be applied on Macro Malwares
- d. Paving a way for future work in order to predict other malware attacks beforehand using AI & ML techniques.

1.6 Thesis Outline

The research work has been organized into following chapters:

- **Chapter 1:** Chapter 1 presents a short introduction to Malware threat to computers and networks. It also presents the problem statement, followed by motivation behind the research and enumerates research objectives. Lastly it highlights the offerings made through this research.
- **Chapter 2: Existing Research in Malwares prediction using Machine Learning and Artificial Intelligence.** Chapter 2 presents an overview of the

existing / recent research that has already been carried out in the field of Malware detection using Machine Learning.

- **Chapter 3: Preliminary Background.** This chapter gives an insight of the preliminary background knowledge that is vital in understanding Artificial Intelligence and Machine Learning, what is data and how data analysis techniques are used in Machine Learning for various purposes.
- **Chapter 4: Using Predictive Modelling and Machine Learning Techniques to design a Macro Malware Prediction System.** This chapter provides an introduction and brief description of Predictive Modeling, which serves as the foundation for our design. Furthermore, the chapter explains the basic process involved in design of a Malware Detection System based on Machine Learning and AI.
- **Chapter 5: Design of a malware prediction system using Machine Learning and AI.** This chapter explains our design of a Malware Prediction System using Machine Learning and AI. The chapter analyses the code and gives a description of its working
- **Chapter 6: The Designed AI Model – Performance Analysis.** In this chapter, we will conduct a performance analysis of the AI model we have developed. We will utilize various regression models to evaluate how our AI model performs and compare its predictions against the actual data, allowing us to gauge its accuracy in predicting the desired outcomes.
- **Chapter 7: Conclusion.** The last chapter will conclude the thesis, giving a generalised overview of the research and findings.

EXISTING RESEARCH IN MALWARES PREDICTION

USING MACHINE LEARNING AND ARTIFICIAL

INTELLIGENCE

Introduction

This chapter highlights previous significant research work carried out in the field of Malware detection and prediction using Machine Learning. The idea of making use of Artificial intelligence and Machine Learning is relatively new, however, it has been a catalyst for academia in designing new systems that can detect Malwares by making use of Machine Learning techniques.

Presently, there are two ways where Machine Learning can help in thwarting the thread of Malwares:-

- a) **Malware detection using Machine Learning.** This involves training an AI model which is trained over previous data and can detect Malwares in real-time by analyzing certain parameters, for example, its behavior.
- b) **Malware prediction using Machine Learning.** This technique involves analyzing web traffic of an organization, including frequency of Malware attacks, and predicting when the next attack is likely to occur.

The first type of technique is the commonly used type which is easy to design. The second type, however, is relatively new and will form the base of this thesis.

2.1. Level of Research Already Carried out on the Proposed Topic

Malware detection is crucial in cybersecurity since it enables for the avoidance of malware execution and download. Malicious content abounds on the internet, which may take many forms and result in data theft and financial losses. Earlier research provided a variety of approaches for predicting future hazardous content exposure [1]. Several studies have addressed the topic of predicting likelihood of infection. In this context, the paradigm considers three key attributes: the user's past infection history, their similarity to other users within a hypothetical network based on past exposures, and the growth pattern of the network. By leveraging these factors, the strategy effectively generates precise results concerning the susceptible portion of the community, surpassing conventional methods. Remarkably, it accomplishes this feat with only a fraction (1/1000) of the personal information required, utilizing browser data from over 20,000 users. [2-8]

The frequency of targeted email threats involving Microsoft (MS) document files has increased over the years. Malicious macros, in specific, have caused widespread damage in numerous businesses. An approach for detecting malicious MS Office documents has been presented in relevant work [9] where authors have presented an approach based on machine learning for identifying harmful macros. The machine learning models (Random Forest, Support Vector Machine, and Multi Layer Perceptron) are fed with parameter vectors. Subsequently, these models which have been trained on previous data, assess whether test vectors are harmful or otherwise. Through extensive testing, the proposed approach demonstrates a high F-measure of 0.93, indicating its effectiveness..

Using five differentiating traits of sophisticated malware, the authors [10] predicted complex malware like Stuxnet from a malware dataset by presenting a novel machine learning technique. After extensively studying sophisticated malware samples in the environment, they derived the characteristics or attributes associated with these malicious software. Regression models were utilized to anticipate sophisticated malware. Malware dataset was generated by combining existing datasets with real-world samples for testing reasons. Experiments indicate that if prediction characteristics are defined, our technique can anticipate Stuxnet similar sophisticated malware.

With advanced obfuscation techniques it was getting difficult to detect new malicious VBA macros. Language Model was constructed to represent VBA macros by using extracted words from VBA scripts for the machine learning techniques. LSI (Latent Semantic Indexing) was used with some Natural Language Processing (NLP) technique to construct an efficient language model which generates more precise and proficient outcomes [11]. The model is then used to train different classifier or SVM model using benign and malicious samples to detect new malicious VBA macros.

The authors [12] provided a novel model for predicting the likelihood of a malware infecting a personal computer by making use of a dataset which was chosen from Microsoft contest. The dataset included owner information, PC configuration, and softwares installed. Several classification algorithms were used in this study to determine the likelihood of a malware infecting a system. The study concluded LightGBM classifier as the best machine learning model due to its efficiency and

performance. The LightGBM approach also identifies leading factors and feature relevance.

Conclusion

In today's technology-driven society, we currently lack the capabilities to accurately forecast the likelihood of a malware infection before it occurs. However, it is crucial to address this challenge by identifying the factors that raise the probability of infection and taking necessary steps to prevent them.

PRELIMINARY BACKGROUND

Introduction

Machine learning has seen unprecedented growth in last few years and has become a major area of focus for researchers, businesses, and governments worldwide. Machine learning, a sub-domain of artificial intelligence (AI), empowers machines to learn and make forecasts from data without the need for explicit programming.

Machine learning algorithms can automatically detect patterns and make predictions based on large datasets, which is particularly useful in fields such as cybersecurity. This chapter provides a general overview of machine learning and AI, including its history, key concepts, and applications [13].

History of AI

The concept of AI dates back to the 1950s. It evolved with the development of the neural networks and expert systems. However, it was not until the advent of machine learning algorithms that AI became a practical reality. The development of machine learning can be traced back to the 1960s, with the creation of the first decision tree algorithm. Since then, machine learning has continued to evolve rapidly, with the development of numerous algorithms such as random forests, support vector and deep learning etc.

The evolution of Artificial Intelligence (AI) has been a fascinating journey marked by breakthroughs and setbacks. AI can be traced back to the 1940s when the concept of a

thinking machine was first proposed. However, it wasn't until the 1950s that the term "Artificial Intelligence" came into limelight by John McCarthy, who is considered one of the fathers of AI [14].

The early days of AI were marked by optimism, and researchers believed that it was only a matter of time before machines could match human intelligence. However, progress was impeded due to limited computing power and inadequacy of data. The development of expert systems and the rise of computer vision in 1960s marked significant progress in the field of AI. The 1970s saw the emergence of rule-based systems, and the 1980s brought about the development of neural networks.

The 1990s marked a significant boom in AI with the evolution of machine learning algorithms and the development of data-driven approaches. The advent of the internet and the availability of vast amounts of data enabled researchers to develop more sophisticated AI algorithms, leading to breakthroughs such as Deep Blue, which defeated the world chess champion, and AlphaGo, which beat the world champion at Go.

Today, AI is revolutionising almost every sphere of our lives, from IT to travel to transportation, finance and entertainment. AI-powered systems are becoming more sophisticated and intelligent, with the emergence of advanced algorithms like reinforcement learning and deep learning are considered as an alternative to human intelligence, capable of solving complex problems [15]. As we move forward, AI will continue to evolve, and we can expect even more remarkable breakthroughs in the years to come. The role of AI for decision making based on already existing data used

for supporting or replacing human decision, is considered as one of the most significant applications so far [16].

3.1 Machine Learning, Deep Learning and AI

Machine learning (ML), Deep Learning and Artificial Intelligence (AI) are co-related concepts but have different meanings. We shall try to explain these in subsequent paras.

3.1.1 Artificial Intelligence. AI is a broader concept that defines the ability of machines to accomplish tasks that would otherwise require human intelligence, such as visual perception, language processing, decision-making and problem-solving. AI, therefore, is a combination of different technologies like machine learning, deep learning and natural language processing etc. [17].

3.1.2 Machine Learning. A subset of AI that includes the process of making a machine to learn (training) and make predictions based on data is called Machine Learning. It involves using algorithms to analyze data and make predictions or decisions basing on that data. These algorithms can be supervised, unsupervised, or semi-supervised, and they can be used for complex jobs such as classification, regression and anomaly detection etc.[18].

3.1.3 Deep Learning. Another key concept in machine learning and AI is deep learning, which makes use of neural networks with multiple layers to detect complicated patterns in data. Deep learning has become a popular approach in recent years, particularly in applications such as image and speech recognition.

The correlation is shown in figure 3.1

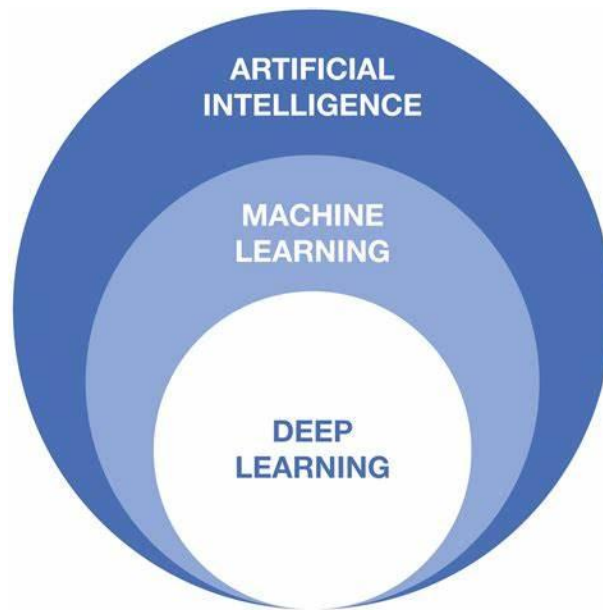


Figure 3.1: Artificial Intelligence, Machine Learning and Deep Learning

In simpler terms, machine learning is a way of achieving artificial intelligence by using data and algorithms to train machines to learn and improve their performance. Machine learning is a tool used to create intelligent systems, while AI is the broader concept of creating intelligent machines that can perform tasks that would typically require human intelligence.

In summary, machine learning is a subset of AI, and it involves using data and algorithms to train machines to learn and make predictions, while AI refers to the broader concept of creating intelligent machines that can perform tasks that would typically require human intelligence.

3.2 Key Concepts in Machine Learning and AI

Machine learning, as defined earlier, is an essential component in the realm of Artificial Intelligence. It is a method of teaching machines to learn patterns from data with or without human intervention. The data can come in different forms, including text, image, audio, and numerical data. However, how data is handled in machine learning is a critical aspect of the accuracy and efficiency of the algorithms and is identified on the base of Labeling.

Data labeling in machine learning refers to the process of assigning descriptive labels or tags to data that serve as inputs for a machine learning model. This process entails the application of labels to features or data points so that machine learning algorithms or models can recognize the relationships between the inputs and predict or classify data with a higher degree of accuracy. Data labeling is crucial in supervised learning, where labeled data is needed to train algorithms to make predictions or classifications. As the quality of labeled data directly affects the accuracy of machine learning models, data labeling is a critical step in preparing data for machine learning.

Based on the classification above, data is either classified as Labeled or Unlabeled, depending upon the purpose for which it is gathered [19].

3.2.1 Labeled Data

In machine learning, labeled data refers to data that has been manually labeled or annotated with the correct output. This means that for each input data point, the corresponding output is already known and is included in the dataset. For example, in a dataset of handwritten digit images, the labels would be the corresponding digits (0-9). Labeled data is commonly used in supervised learning, where the goal is to train a

machine learning model to predict the correct output given a new input. The algorithm is trained on the labeled data, and then it can make predictions on new, unseen data.

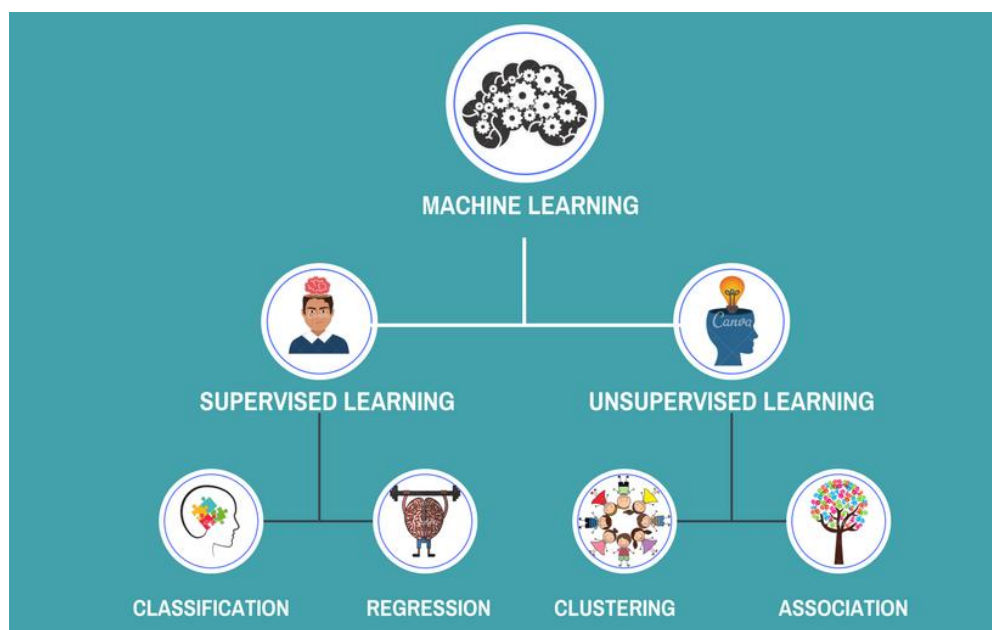
3.2.2 Unlabeled Data

On the other hand, unlabeled data, also known as unlabeled data, refers to data that does not have any known output or labels associated with it. Unlabeled data is commonly used in unsupervised learning, where the goal is to discover patterns and structures within the data without any prior knowledge of the correct output. For example, a dataset of images that does not have any corresponding labels can be used for unsupervised learning techniques such as clustering and dimensionality reduction.

3.3 Sub-branches of Machine Learning.

Machine learning and AI are based on a few key concepts, including supervised learning, unsupervised learning, and reinforcement learning.

Technically, these can be defined as :-



3.3.1 **Supervised learning.** Supervised learning is a type of machine learning in which a model is trained on a labeled dataset to make predictions about new, unseen data. The algorithm is provided with input-output pairs of data, where the input represents the features of the data, and the output represents the label or target variable. The goal of supervised learning is to learn a function that can map inputs to outputs accurately. During the training process, the algorithm adjusts its internal parameters to minimize the difference between its predicted output and the true output. Once the model has been trained on the labeled dataset, it can be used to make predictions on new, unlabeled data.

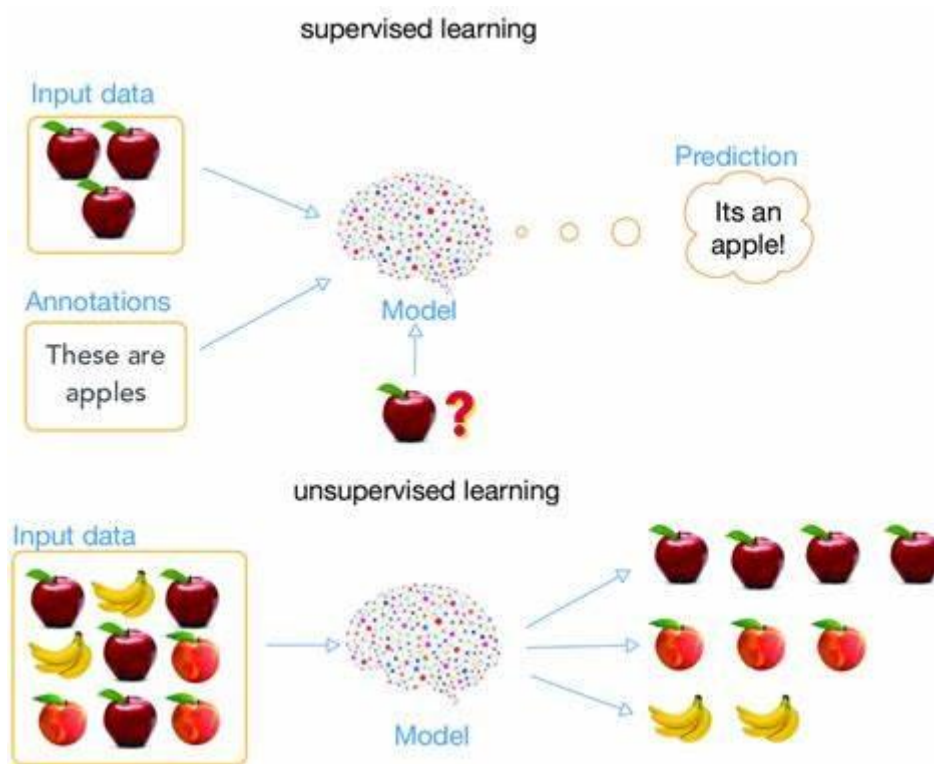
Some common algorithms used in supervised learning include linear regression, logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.

3.3.2 **Unsupervised learning.** Unsupervised learning is a type of machine learning in which the algorithm learns from data that is not labeled or classified. Unlike supervised learning, where the algorithm is provided with labeled examples, unsupervised learning algorithms must find patterns and structures within the data on their own. This makes unsupervised learning useful in scenarios where there is no labeled data available or where the data is too complex to be labeled by humans.

In unsupervised learning, the algorithm is tasked with discovering hidden patterns or relationships within the data. Common unsupervised learning techniques include clustering, dimensionality reduction, and anomaly detection. Clustering algorithms group similar data points together based on their similarity, while dimensionality reduction techniques simplify high-dimensional data into a lower-dimensional

representation. Anomaly detection algorithms identify data points that deviate significantly from the norm.

Unsupervised learning has applications in a wide range of fields, including finance, healthcare, and marketing. For example, in finance, unsupervised learning can be used to identify patterns in financial data that can help inform investment decisions. In healthcare, unsupervised learning can be used to identify patient subgroups based on their medical history, which can help personalize treatment plans.



3.3.3. Semi-Supervised Learning. In some cases, a dataset may have a mix of labeled and unlabeled data. This is known as semi-supervised learning and can be used in scenarios where labeled data is expensive or difficult to obtain. Semi-supervised learning algorithms use a combination of labeled and unlabeled data to make predictions.

3.3.4 **Reinforcement learning**. Reinforcement learning is a type of machine learning that involves an agent interacting with an environment in order to learn how to make decisions that maximize a cumulative reward. In reinforcement learning, the agent learns through trial and error, using feedback from the environment in the form of rewards or penalties. The agent receives input from the environment, takes an action, and then receives feedback in the form of a reward or penalty. The agent's goal is to learn a policy, or a set of rules, that tells it what action to take in each state in order to maximize the cumulative reward over time. Reinforcement learning algorithms use the concept of exploration and exploitation to learn the optimal policy. The agent explores the environment by trying different actions, and then exploits what it has learned by taking actions that have led to high rewards in the past. Reinforcement learning has applications in a wide range of fields, including robotics, game playing, and recommendation systems.

3.4 Time Series Analysis.

Time series analysis is a subfield of machine learning that focuses on analyzing and predicting time series data. Time series analysis falls under both supervised and unsupervised learning and is used to make predictions about the future behavior of the time series.

In supervised learning, time series data can be used for tasks such as predicting stock prices or weather forecasting, where the model is trained on historical data with known outcomes. In unsupervised learning, time series data can be used for tasks such

as anomaly detection or clustering, where the model is trained on the patterns and relationships in the data without prior knowledge of the outcomes.

Some common applications of time series analysis include:

- 1) Financial forecasting
- 2) Predicting sales or demand for a product
- 3) Traffic forecasting
- 4) Weather forecasting
- 5) Predicting energy demand
- 6) Health monitoring and disease outbreak detection

Machine learning algorithms such as ARIMA, LSTM, and Prophet are commonly used for time series analysis and forecasting.

3.5 How Machine Learning Works

Machine learning is a branch of artificial intelligence that teaches computers to learn from data and make accurate predictions or decisions based on what it has learnt from that data. The general process of machine learning involves the following steps:

- 1) **Data collection**: Machine learning algorithms need a huge volume of data which it can learn from. This data can come from various sources, such as sensors, databases, or web scraping.
- 2) **Data preparation**: The data, after collection, is parsed and preprocessed to remove any noise or inconsistencies. This step involves tasks like data normalization, feature selection, and data transformation.

- 3) **Training**: In this step, the machine learning algorithm is trained on the prepared data set. The algorithm learns from the data and improves its output by adjusting its parameters to close the gap between its predicted and the actual output.
- 4) **Evaluation**: After training, the algorithm's performance is assessed by testing on separate test data set. The performance evaluation depend on the type of problem being solved considering accuracy, precision, recall, and F1 score etc.
- 5) **Deployment**: After training and evaluation, the algorithm can be deployed in a real-world application. This may involve integrating it with other systems, optimizing its performance, and monitoring its behavior over time.

3.6 Applications of Machine Learning and AI

Machine learning and AI find a wide range of uses and purposes, including areas like natural language processing, computer vision, robotics, and cybersecurity. In cybersecurity, machine learning is used to identify and prevent threats such as phishing attacks, ransomware, malware and various similar cybercrime.

Conclusion

Machine learning and AI have become integral to modern society, with applications in various fields. In this chapter, we provided an overview of machine learning and AI, including its history, key concepts, and applications. In the following chapters, we will explore the deployment of machine learning techniques for malware prediction, including various algorithms and techniques used in this field.

MACRO MALWARE AND PREDICTION SYSTEMS

After having gone through the preliminary background in the preceding chapters, we shall discuss the concept and design of AI system that can be used to predict Macro Malware attacks on system using predictive modeling and machine learning techniques. In subsequent Chapter, we shall discuss its code and working in depth.

Malware attacks are a significant threat to computer systems and networks, with new malware variants emerging every day. Traditional approaches to detecting and preventing malware attacks, such as signature-based detection, are no longer sufficient, and researchers are turning to machine learning (ML) and artificial intelligence (AI) to develop better and effective solutions.

AI and ML can be used to predict malware attacks by analyzing huge volume of data and identifying sequences and trends that are indicative of malicious activity. This approach is known as *behavior-based detection*, and it involves monitoring system activity and looking for deviations from normal behavior.

4.1 Malicious Macro Malwares embedded in Office Documents

The utilization of malicious macro documents, accompanied by several threats, has grown exponentially in the last few years. The Microsoft Office document offers layout flexibility with various capabilities for the malware authors to exploit. Most of the increasing methods employed by perpetrators to spread malware has been Malicious Documents, all due to an increasing unaware audience and the lack of detection mechanisms by existing antivirus apps.

The first macro virus to spread through Microsoft Word document emerged in 1995 by the name of Concept but did not had the capability to get transmitted by means of email. That's where the advent of macro virus spread via email attachments came into being [20]. Melissa was the first macro virus which spread through an email attachment from a malicious word document. Melissa was a self-replicating worm which infected thousands of computers within a few hours.

A macro virus is basically a computer infection that can be spread with the help of different social engineering techniques. These type of macro malwares are sent to the users and transmitted to their computer by merely opening an email, an email attachment, or by performing some other usual normal operations, such as clicking a graphic to expand it in an email you receive which are usually embedded in Microsoft office files.

Macros are scripts used in Microsoft office documents to automate task scripted in VBA. The malware authors took advantage of this feature and embed malwares in office documents with the help of VBA functionality to perform malicious activities on victim's computer.

Microsoft Office documents and other document formats have evolved over time. They are not simple static files with little potential to harm any system or device. Microsoft Word and Adobe PDF have introduced macro and scripting functionality that allow documents to operate in very much the same manner as executable programs, right down to the capacity to execute processes and install other pieces of code on user systems. They spread though different mechanisms by tricking users to install malware on their system. Although malware analysts continuously try to fix the

loopholes used by malware authors to distribute their code, they are generally far behind the attackers. Due to this factor macro malwares continue to evolve with changed behavior to bypass traditional defenses and rule out previous detection mechanisms. So, it's crucial to know what different forms of macro malwares are and what we can do to protect your network, users, and sensitive business data by creating a detection mechanism for contemporary macro malwares as malware use keep increasing.

4.2 Microsoft documents file format and structure

Microsoft Office uses different file extensions for the old and new versions of Office Suite productivity software. They created two types of document file extensions having different internal formats and structure which can be helpful in proper analysis of a document.

4.2.1 OLE (Object Linking and Embedding)

A binary format classified as OLE (Object Linking and Embedding) was used primarily by Office 97-2003 to store data and information. OLE files use legacy filename extensions identified as, .pdf, .xls, .ppt or their file signature of first eight bytes i.e. D0CF11E0A1B11AE1. OLE file format is also known as OLE2 files and termed as:

- Compound Binary File Format
- Compound Document File Format
- Horrible Property File Format

Malware authors use legitimate OLE embeddings to trick users, by use of well-formatted text and different images, into allowing malicious content and installing the malicious code to their systems. So far mostly embeddings are done through Visual Basic (VB) and JavaScript (JS) malicious contents linked with the file.

4.2.2 OOXML-based File Formats

Microsoft launched a new format known as OOXML (Office Open XML) in Office 2007. The XML based file format is basically a ZIP archive consisting of an internal directory containing XML files, which are designed to store the information and content of actual documents and its metadata. The extensions used by xml file-based documents are like .docx, xlsx, pptx with various other extensions with different enabled features.

The documents we create through MS Office are saved in XML format by-default with file extensions 'x' or 'm' at the end usually. X define the xml files with no macros like .docx extensions, while 'm' means a macro comprised XML file like .docm extension. OOXML files are different from OLE binary file types. These are usually zipped files with different file structure as defined:

- Multiple XML parts describing file data, metadata, customer data
- ZIP container with compression
- Relationships define file structure
- Non-XML parts supported as native files (images, OLE objects)

The malware authors use XML file formats in different ways to hide malicious macros. The attackers spread these malwares with the help of email attachments or through URLs hosted on their C&C (Command & Control) server.

XML's versatility has contributed to its widespread use, but it includes many security vulnerabilities that can be used for multiple forms of targeted attacks at the very same time. XML file-based VBA macro attacks typically occur in following way:

- A Microsoft Word document is created, a malicious VBA macro is added by the attacker and saved it in XML format.
- The XML file is sent the victim, for example through an email attachment
- The victim downloads the file into its system, clicks on the XML file to open it and runs the VBA macro.
- The VBA macro downloads another malicious file through attacker hosted machine, drops it to victim's system and executes it.

4.3 Behavior-based detection

Behavior-based detection is a technique used to predict malware attacks by monitoring system activity and identifying deviations from normal behavior. Behavior-based detection systems are designed to detect and prevent unknown malware variants that have not been seen before. It learns from earlier malware attacks and adjusts to new threats in real-time, making them an effective tool for predicting and preventing malware attacks.

Behavior Based detection systems involves collecting data from various sources, like system logs, network traffic, user activity, and using machine learning algorithms to analyze the data. It then identifies patterns that indicate malware activity. These algorithms are trained on large datasets of existing malwares and benign samples to learn how to detect and classify different types of malware.

4.3.1 Advantages of Behavior Based Detection System

- 1) This technique can learn and adapt to new threats in real-time. As new malware variants emerge, the algorithms can analyze their behavior and identify patterns that are similar to known malware, enabling them to detect and prevent new attacks.
- 2) This technique can reduce the number of false positives and negatives. Traditional approaches to malware detection often generate a high number of false positives, which are both time-consuming and costly to investigate. AI and ML algorithms, on the other hand, can learn to distinguish between benign and malicious activity and reduce the number of false alarms.

4.4 Outline of a Malware Prediction System using Machine Learning and AI

As already discussed in para 4.1, a lot of work on detection of Malware by making use of Behavior based detection system has been done. Malware has become a major threat to computer systems and networks, with new types of malware being developed every day. In order to protect against these threats, it is important to be able to forecast the likelihood of a system being infected by malware. Machine learning and AI can be used to develop a system that can identify potential threats before they can cause any harm.

The design of a malware prediction system using machine learning and AI involves several key steps. These steps can be listed as:-

4.4.2 Data Collection. A large volume of data is required to be collected from an organization that contains information about known malware attacks during last certain years. In order to predict the likelihood of attack on an organization, this data must come from a single source.

4.4.3 Data Cleansing and Normalization. Next, the data must be preprocessed and cleaned to remove any noise or inconsistencies. This step involves tasks like normalizing the data, feature selection, and data transformation. One of the challenges in designing a malware prediction system is selecting the most relevant features from the data. Features that are commonly used in malware prediction include network traffic patterns, file properties, and behavior patterns.

4.4.4 Training the Machine Learning Algorithm. Once the data is prepared, a machine learning algorithm can be employed and trained on the given data to forecast the likelihood of a system being infected by malware. Several types of machine learning algorithms like decision trees, support vector machines, and neural networks can be used for this purpose. The selection of algorithm depends on the characteristics of the dataset and the reason for which the system is being designed.

4.4.5 Additional Steps. In addition to the selection of algorithm, the quality and quantity of the training dataset are also contributing factors towards the performance of the malware prediction system. The more diverse and representative the training

data, the better the system is likely to perform. This requires ongoing data collection and updating of the system as new threats emerge.

4.4.6 System Deployment. Finally, the system must be deployed and integrated with other security systems to provide real-time predictions and alerts. This may involve optimizing the performance of the algorithm, monitoring its behavior over time, and incorporating feedback from security analysts to improve its accuracy. The system may also be integrated with other security parameters, like firewalls, intrusion detection systems and antivirus software, to provide a all-inclusive defense against malware threats.

Conclusion

Overall, a malware prediction system using machine learning and AI can help organizations to proactively protect their computer systems and networks from malware attacks. By continuously analyzing data and identifying potential threats, this system can help to minimize the risk of data breaches, financial losses, and other negative consequences of malware infections. However, it is important to recognize that no system can provide 100% protection against malware and that a layered approach to security is always recommended.

In conclusion, malware attacks are a significant threat to computer systems and networks, and traditional approaches to detecting and preventing these attacks are no longer sufficient. AI and ML can be used to predict malware attacks by analyzing huge volume of data and recognizing trends and anomalies that are indicative of malicious activity. Behavior-based detection is a promising approach to malware

prediction, and it has the potential to improve the effectiveness of malware detection and prevention.

DESIGN OF A MALWARE PREDICTION SYSTEM USING MACHINE LEARNING AND AI

5.1 The Programing Environment

Our model is programmed in Python. We used Anaconda Platform which is an Open-Source distribution of Python and R programming languages. It is a data sciences platform and contains over three hundred data science packages including Machine Learning applications, data processing and predictive analysis etc. It simplifies the programming process as it does not involve installing Python and adding Data Sciences packages separately.

5.2 The data set

The data set is the basic building block of our thesis. For obtaining accurate results, the data set was obtained from an educational institution and contained emails spreading over last 2 years. Since the objective of thesis was to predict the next malware attack, we did not require actual malware files or contents of the emails. Only the time and date of the e-mails that contained the Malwares was required for training our model. Therefore, the email data was exported in the format appended below. This was in-line with the organizational privacy policy as well. This data, however, contained those emails as well which did not contain a Malware.

Email_id	User_role	Date	Email_attachment	Email_Subject	Recipient_list	Indication_malware

The dataset is saved in a CSV file. It is read using following command and imported into the “*data*” variable

```
data=pd.read_csv('C:\\Users\\HS LAPTOP\\Email_data5000.csv')
```

5.3 Sifting of Data

5.3.1 Selecting Data that contained Malware

In the second step, data was parsed and only that email data was selected which contained a Malware. This was crucial for training our model since we need to predict when the next email containing a Malware is likely to be received. After this step, there were about 300,000 emails left which contained Malware and were sufficient to train our model.

	Email_id	user_role	Date	email_attachment	email_subject	recipient_list	indication_malware
0	1	'Database Admin'	2021-09-01 22:46:00	'foto.bt'	'Site Logo'	tking@christensen.net	1
1	2	'Network Admin'	2022-04-03 10:29:00	'foto.bt'	'Send ATM Pin'	brandon65@smith.com	0
2	3	'Database Admin'	2018-12-17 13:19:00	'Project_Code_changes.zip'	'Price Win'	richardjohnson@ali.info	1
3	4	'Data Scientist'	2021-08-30 04:43:00	'ME_Project_Timeline.xlsx'	'Paypal Account'	martinlauren@gibson.com	1
4	5	'Developer'	2022-10-02 12:59:00	'ME_Project_Timeline.xlsx'	'Site Logo'	sally22@french.biz	1
5	6	'Database Admin'	2019-06-10 11:27:00	'bogus_url'	'Send ATM Pin'	nharris@montgomery-montes.com	1

0 – indicates that the email did not contain a Malware
1 – indicates that the email contained a Malware

Figure 5.1 – Sifting of data and selection of emails containing Malware.

5.3.2 Selection of time and Malware fields for data plotting

For plotting time vs Malware, we will use the *data.drop* command to drop the fields that are not required.

```
new=data.drop(['Email_id','user_role',"email_attachment","email_subject","recipient_list"],axis=1)
```

Figure 5.2 – Dropping the fields that are not required for plotting.

The remaining two columns will be sufficient to plot time vs Malware data.

	Date	indication_malware
0	2021-09-01 22:46:00	1
1	2022-04-03 10:29:00	0
2	2018-12-17 13:19:00	1
3	2021-08-30 04:43:00	1
4	2022-10-02 12:59:00	1

Figure 5.3 – Data vs indication_malware data after dropping irrelevant fields.

We renamed the field “Indication_Malware” to “Malware”, as shown in Figure 5.4, for easy referencing and clarity. By adopting a concise and straightforward field name, we aim to streamline data retrieval and facilitate a more intuitive user experience within the database.

	Date	indication_malware		Date	Malware	
0	2021-09-01 22:46:00	1	→	0	2021-09-01 22:46:00	1
1	2022-04-03 10:29:00	0		1	2022-04-03 10:29:00	0
2	2018-12-17 13:19:00	1		2	2018-12-17 13:19:00	1
3	2021-08-30 04:43:00	1		3	2021-08-30 04:43:00	1
4	2022-10-02 12:59:00	1		4	2022-10-02 12:59:00	1

Figure 5.4 – Data vs indication_malware data after dropping irrelevant fields.

In order to achieve accurate visualization of the data, it is necessary to verify whether any of the fields within the dataset contain null values. The presence of such null values can significantly impact the reliability of the results and lead to incorrect outputs, undermining the overall accuracy of the visualization process.

Analysis of the "Date" and "Malware" columns reveals that there are a total of 4999 records present in each column, with none of them containing null values, as depicted

in Figure 5.5. This indicates a complete dataset with robust information, ensuring reliable and comprehensive data for further analysis and interpretation.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4999 entries, 0 to 4998
Data columns (total 2 columns):
#   Column   Non-Null Count  Dtype
---  ---      -
0   Date     4999 non-null   datetime64[ns]
1   Malware  4999 non-null   int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 78.2 KB

```

Figure 5.5 – Results after checking if any of the fields contained a null-value.

5.3.3 Converting “Date” Column into “Date” Type

To ensure precise training of the model, we have performed a conversion on the data stored in the "Date" column. Previously stored as an object, the data in this column has now been transformed into the appropriate data type, specifically "Date." This conversion enables more accurate and effective processing and analysis of the temporal information contained within the dataset.

5.3.4 Sorting Malware Attack data month wise

Afterwards, we proceeded to group the data related to malware attacks on daily basis. Utilizing the *matplotlib* library, we created a visual representation of the grouped data, allowing us to effectively visualize and analyze the patterns and trends associated with the occurrences of malware attacks over time. The result obtained after plotting Malware against Y-axis and dates against X-axis is shown in Figure 5.6 below.

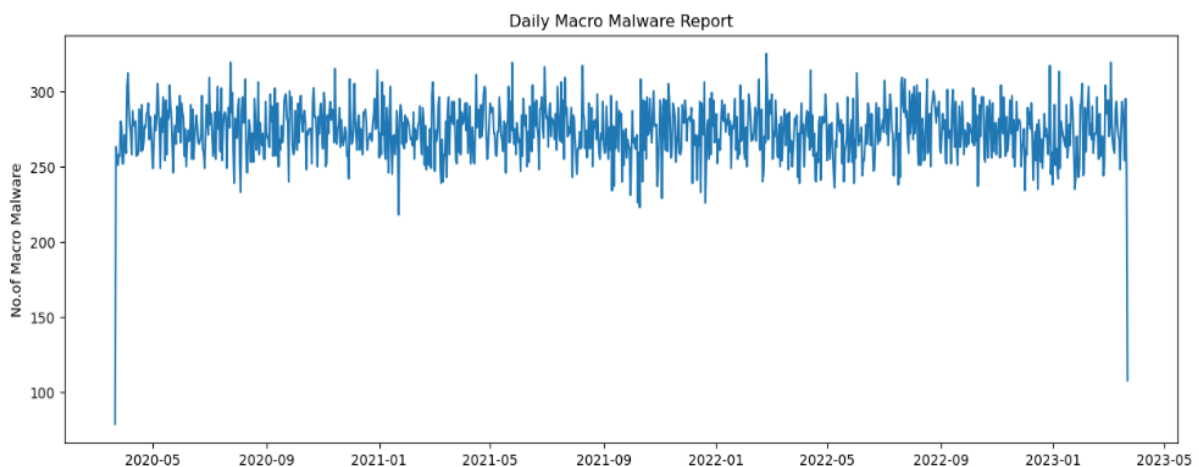


Figure 5.6 – Plotting Number of Malware VS Date data.

5.3.5 Differencing Data and preparing training set.

Differencing data is a technique used to enhance the properties of the data, remove unwanted components, and make it more suitable for AI models to effectively learn and make accurate predictions or inferences. The specific choice and application of differencing depend on the characteristics of the data and the objectives of the modeling task.

Many AI models, particularly those used for time series analysis, assume that the data being analyzed is stationary, or that the statistical properties of the data, such as the mean, variance, and covariance, are static with respect to time. However, real-world data often exhibits trends and other time-dependent patterns that violate the assumption of stationarity. By taking the difference between consecutive data points, we can eliminate the trend component and make the data stationary.

Next, we organized the malware data on a daily basis, as depicted in Figure 5.7. The X-axis denotes the differences in malware occurrences, while the Y-axis represents the corresponding month data. This visualization allows us to examine the relationship between the changes in malware incidents and their distribution across different months.

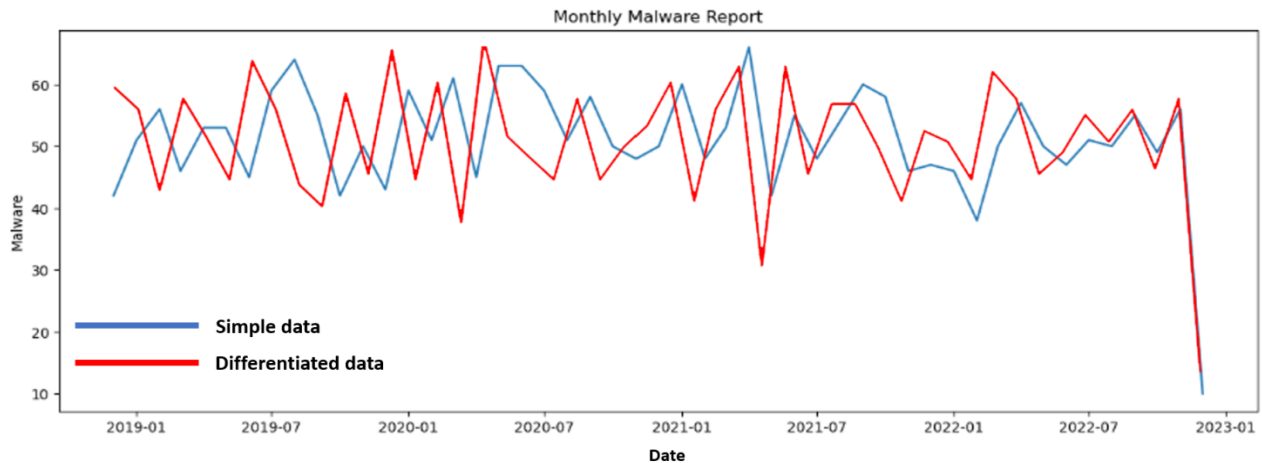


Figure 5.7 – Plotting Malware VS Date data with Simple and differentiated data

5.3.6 Slicing the training set into training and testing datasets.

In the next step, we will utilize the training set that was prepared in the previous step for both training and testing purposes. For this purpose we will divide the dataset into two segments:

1. **Training dataset:** The initial entries of the dataset will be used exclusively for training the model. We will make use of this segment of the dataset to teach the AI model to learn and capture patterns and relationships within the dataset.
2. **Prediction dataset:** The last 12 rows with multiple data entries, which were not used in the training process, will be designated as the prediction dataset. Since we have the actual values for these entries, we can compare them with the predicted values generated by our AI model using different prediction techniques. This evaluation will enable us to assess the accuracy of the model's predictions.

5.3.7 Prediction Models.

To evaluate the accuracy of our AI model, we will employ a series of prediction models. These models will generate predicted values based on the trained data, allowing us to compare them against the actual values from the prediction dataset. By

assessing the performance of various prediction models, we can determine the effectiveness and accuracy of our AI model in making accurate predictions for the given dataset. We have made use of following models:-

1. Linear Regression Model
2. XGB Regression Model
3. Random Forest Regression Model
4. Support Vector Regression Model
5. Ridge Regression Model

Conclusion

This chapter presented the design of a malware prediction system using machine learning and AI. The objective was to develop a robust and efficient system that can accurately predict future malware occurrences. Through the utilization of machine learning algorithms and AI techniques, we constructed a model capable of analyzing various features and patterns associated with malware attacks. The system was designed to leverage historical data and employ prediction models to forecast future malware incidents.

The successful implementation of this system holds great promise in enhancing cybersecurity measures and enabling proactive mitigation strategies. Future chapter will focus on the evaluation and performance analysis of this system to validate its effectiveness and explore potential areas for refinement and enhancement.

THE DESIGNED AI MODEL – EVALUATION AND PERFORMANCE ANALYSIS

Introduction

This chapter delves into the evaluation and performance analysis of the designed AI model. This crucial step aims to assess the effectiveness and accuracy of the model in achieving its intended objectives. By subjecting the AI model to rigorous testing and analysis, we can measure its performance against established metrics and benchmarks. Through various evaluation techniques, such as cross-validation and error analysis, we will examine the model's predictive capabilities and identify areas of improvement. Additionally, this chapter will explore the model's efficiency, scalability, and overall computational performance. The insights gained from this evaluation will guide us in refining and optimizing the AI model for enhanced performance and real-world applicability.

6.1 General Outline

Within the dataset, there are a total of 1065 x 32 entries, each representing a distinct month. For training purposes, a subset of 10 x 32 entries will be utilized, enabling the model to learn and capture patterns from the available data. The 12 x 32 entries are set aside to serve as a validation set, which will be used to compare the already available values with the predicted values of the Regression Model. By withholding this portion from the training process, we can assess the model's performance and accuracy on unseen data, thus allowing us to evaluate the model's ability to generalize and make reliable predictions beyond the data it was trained on.

6.2 Linear Regression Model.

This section of the chapter focuses on the evaluation and analysis of our model using the linear regression model. We will examine the model's performance in terms of accuracy and predictability.

Linear regression is an essential and extensively utilized technique in predictive modeling. It is a commonly employed statistical modeling approach that predicts a continuous dependent variable by considering one or multiple independent variables. The method assumes a linear association between the variables and calculates coefficients to minimize the sum of squared errors.

Figure 6.1 shows the output of Linear Regression Model on our dataset. The actual data is represented by blue line whereas the orange line represents the Predicted data.

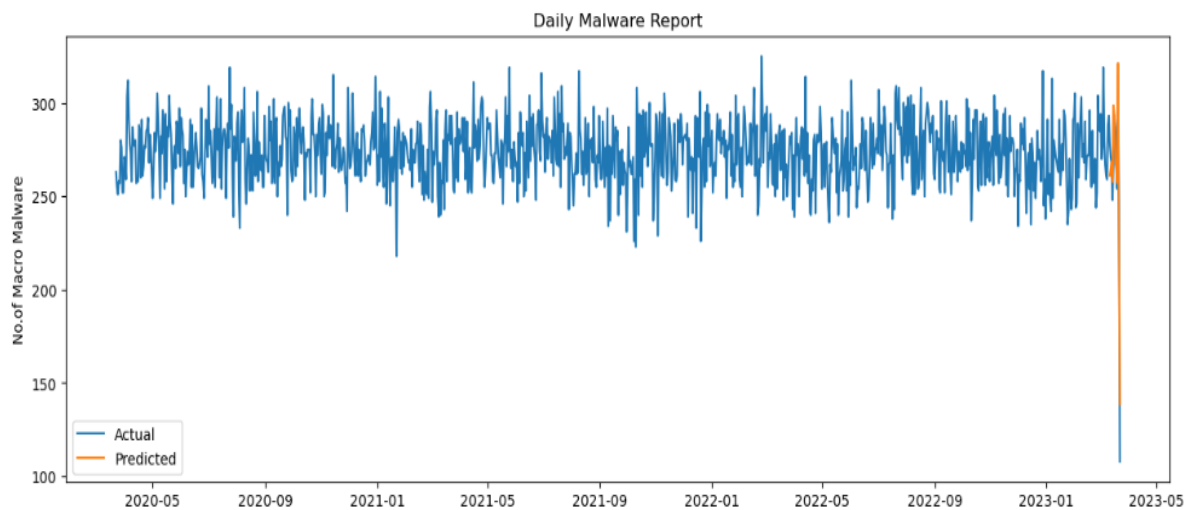


Figure 6.1: Malware prediction using Linear Regression Model

As seen in Figure 6.1 above, the Linear Regression Model was able to predict the output with an accuracy of 0.8387 or 83.8 %.

6.2 Random Forest Regressor.

Random Forest Regressor is a powerful and versatile algorithm for regression tasks. Random Forest is a popular ensemble learning method that combines multiple decision trees to make accurate predictions. It leverages the strength of individual trees while mitigating overfitting and improving generalization. By randomly selecting features and creating diverse trees, the Random Forest Regressor excels in handling complex datasets and capturing non-linear relationships. Its robustness, flexibility, and ability to handle large datasets make it a valuable tool in regression analysis.

Figure 6.2 shows the output of Random Forest Regression Model on our dataset. The actual data is represented by blue line whereas the orange line represents the Predicted data.

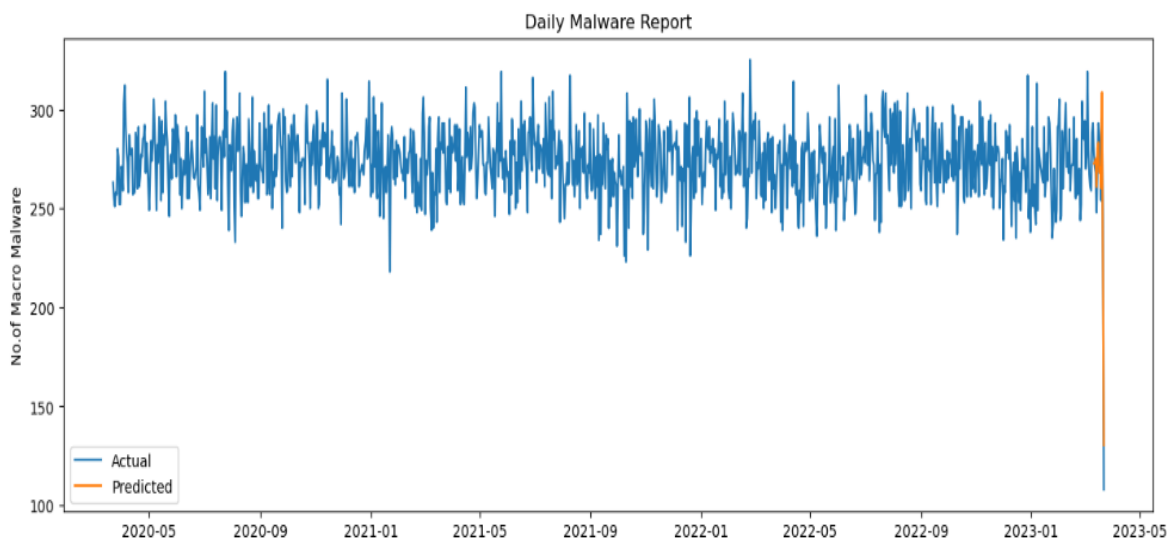


Figure 6.2: Malware prediction using Random Forest Regression Model

As seen in Figure 6.2 above, the Random Forest Regression Model was able to predict the output with an accuracy of 0.838 or 83.8 %.

6.3 XGB Regressor.

In this section, we exposed our dataset to the XGB Regressor, a highly effective algorithm for regression tasks. XGB, short for Extreme Gradient Boosting, is an optimized implementation of gradient boosting that excels in predictive accuracy and computational efficiency. By iteratively building an ensemble of weak learners, XGB Regressor leverages gradient descent and regularization techniques to enhance performance. With its ability to handle complex datasets, handle missing values, and capture non-linear relationships, XGB Regressor has gained popularity as a powerful tool for regression analysis.

Figure 6.3 shows the output of XGB Regressor Model on our dataset. The actual data is represented by a blue line whereas the orange line represents the Predicted data.

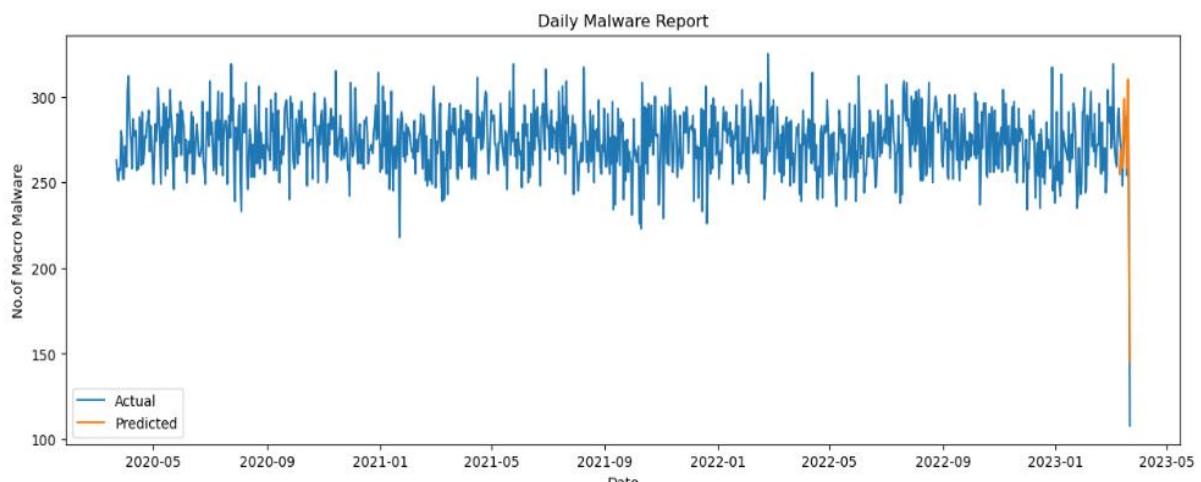


Figure 6.3: Malware prediction using XGB Regression Model

As seen in Figure 6.3 above, the XGB Regression Model was able to predict the output with an accuracy of 0.8554 or 85.54 %.

6.4 Support Vector Regression.

The Support Vector Regressor (SVR) is a powerful algorithm used for regression tasks. It is based on the Support Vector Machine (SVM) framework, known for its

effectiveness in classification problems. SVR extends SVM to handle regression by finding the optimal hyperplane that maximizes the margin while minimizing the error between predicted and actual values. By transforming the data into higher-dimensional spaces, SVR captures complex relationships and offers robustness against outliers. With its flexibility and ability to handle various types of data, SVR has emerged as a valuable tool in regression analysis.

Figure 6.4 shows the output of Support Vector Regression Model on our dataset. The actual data is represented by blue line whereas the orange line represents the Predicted data.

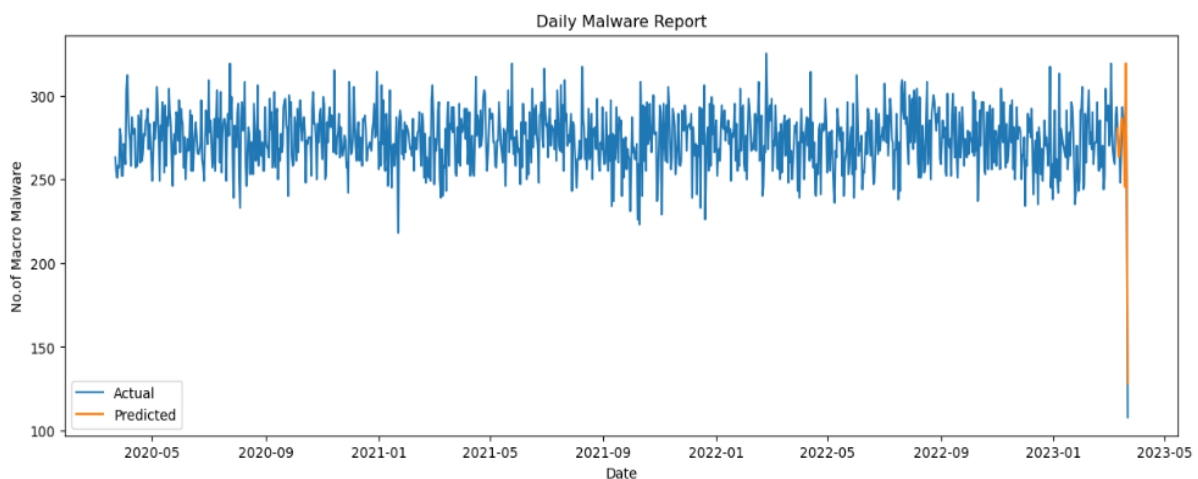


Figure 6.4: Malware prediction using Support Vector Regression Model

As seen in Figure 6.4 above, the Support Vector Regression Model was able to predict the output with an accuracy of 0.903 or 90.3 %.

6.5 Ridge Regression.

Ridge regression is a regression technique that addresses the issue of multicollinearity in linear regression models. It introduces a regularization term, known as the ridge penalty, which helps prevent overfitting and stabilizes the model's coefficients. By shrinking the coefficients towards zero, Ridge regression reduces the impact of highly

correlated predictors, improving the model's generalization ability. This technique strikes a balance between model complexity and simplicity, making it particularly useful when dealing with datasets with high-dimensional features or collinear predictors.

Figure 6.5 shows the output of Ridge Regression Model on our dataset. The actual data is represented by blue line whereas the orange line represents the Predicted data.

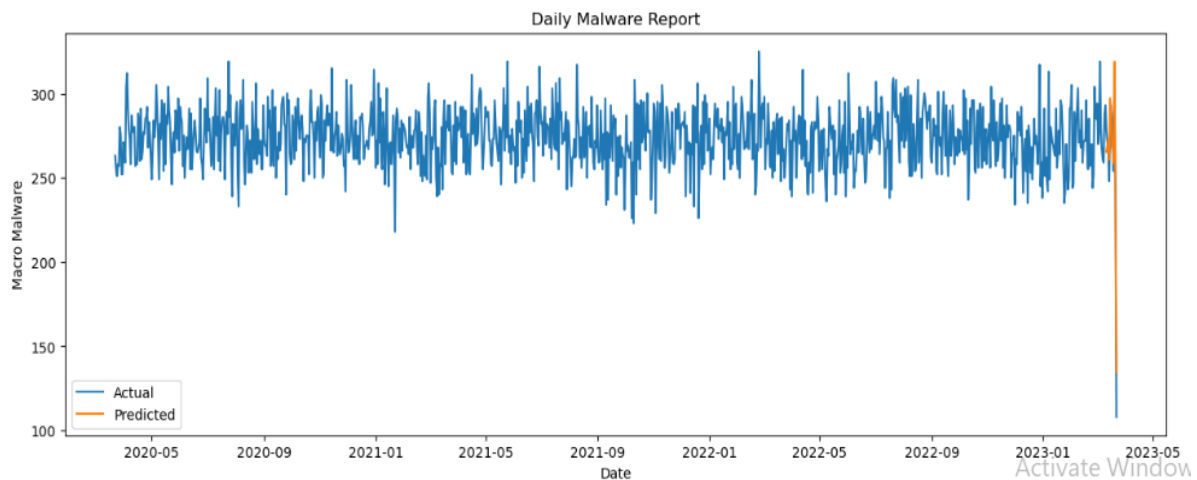


Figure 6.5: Malware prediction using Ridge Regression Model

As seen in Figure 6.5 above, the Ridge Regression Model was able to predict the output with an accuracy of 0.8675 or 86.75 %.

6.5 Comparative Analysis

The results reveal compelling insights into their respective predictive capabilities:

1. Linear Regression Model achieved a notable prediction accuracy of 0.8438. While providing a solid baseline, it demonstrated the potential for improvement by other more complex models.

2. The Random Forest Model displayed a competitive performance with an accuracy of 0.838. Its ensemble nature allowed it to capture intricate relationships within the data, making it a robust choice for this prediction task.
3. XGB Regression emerged as one of the top performers, achieving an accuracy of 0.855. Its boosted decision-making approach enabled it to excel in capturing nuanced patterns within the email dataset.
4. Support Vector Regression demonstrated remarkable predictive power with an accuracy of 0.903, showcasing its ability to handle complex relationships and outperform several other models.
5. The Ridge Regression model achieved an accuracy of 0.867, reflecting its effectiveness in mitigating multicollinearity and producing reliable predictions.

	Date	Macro Malware	Linear Prediction	Random Forest	XGB	SVR	Ridge
0	2023-03-11	278	261.123009	272.66	254.792980	280.218284	265.355828
1	2023-03-12	271	264.614152	272.76	269.337692	263.320460	268.912507
2	2023-03-13	268	268.376120	275.07	263.378576	263.361949	272.624807
3	2023-03-14	248	257.734448	260.79	258.468591	271.126054	260.623884
4	2023-03-15	271	298.592734	280.42	284.028919	284.251509	296.988459
5	2023-03-16	293	293.474376	283.19	298.772310	286.424621	291.158406
6	2023-03-17	287	268.410667	268.12	278.289716	282.251396	269.338800
7	2023-03-18	276	264.661460	270.60	271.818649	263.614346	266.414202
8	2023-03-19	254	257.265214	260.13	258.294884	245.449115	258.589814
9	2023-03-20	295	321.323313	308.52	310.068359	319.098240	318.921692
10	2023-03-21	245	227.410982	224.18	232.608221	228.244847	227.054639
11	2023-03-22	108	138.611322	130.53	145.307675	128.141831	134.569721

Figure 6.6: Comparison of Model

As seen in Figure 6.6 above, the Support Vector Regression model demonstrated the highest accuracy among the algorithms tested, showcasing its suitability for accurately predicting macro malwares within the email dataset. Nevertheless, each algorithm

exhibited distinct strengths and limitations, suggesting that a combination or ensemble approach might further enhance predictive performance. The insights gained from this comparison lay the groundwork for more effective macro malware detection strategies in email security.

CONCLUSION

This thesis has explored the use of artificial intelligence and machine learning techniques for malware prediction. The objective was to develop an effective system that can anticipate and mitigate the impact of malware attacks. Throughout the research, various machine learning algorithms and AI models were employed, including decision trees, random forests and support vector machines, among others.

The findings of this study demonstrate the potential of machine learning and AI in predicting malware occurrences. The models developed in this thesis exhibited promising results, achieving high accuracy rates in identifying and classifying malware samples. By leveraging features such as file properties, behavior analysis, and network traffic patterns, the models successfully captured the intricate characteristics and patterns of malicious software.

Furthermore, the research highlighted the importance of feature engineering, data preprocessing, and model selection in optimizing the performance of malware prediction systems. The choice of appropriate features and careful data preparation significantly influenced the models' accuracy and robustness. Additionally, the selection of suitable algorithms and techniques played a pivotal role in achieving reliable and efficient predictions.

The outcomes of this thesis contribute to the field of cybersecurity by exploring the potential of machine learning and AI in combating malware threats. The developed models can aid in real-time monitoring, threat detection, and proactive defense mechanisms. However, it is crucial to acknowledge that the field of malware

is continuously evolving, necessitating ongoing research and adaptation of the models to stay ahead of emerging threats.

In conclusion, this thesis establishes a solid foundation for future advancements in malware prediction using machine learning and AI. By harnessing the power of these technologies, we can strengthen our defenses against malicious software, safeguard critical systems and data, and mitigate the devastating consequences of cyber attacks.

References

- [1] Yavneh, Amir, Roy Lothan, and Dan Yamin. "Co-similar malware infection patterns as a predictor of future risk." *PloS one* 16.3 (2021): e0249273.
- [2] Sharif, Mahmood, et al. "Predicting impending exposure to malicious content from user behavior." *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018.
- [3] Canali, Davide, Leyla Bilge, and Davide Balzarotti. "On the effectiveness of risk prediction based on users browsing behavior." *Proceedings of the 9th ACM symposium on Information, computer and communications security*. 2014.
- [4] Gratian, Margaret, et al. "Identifying infected users via network traffic." *Computers & Security* 80 (2019): 306-316.
- [5] Kang, Chanhyun, et al. "Ensemble models for data-driven prediction of malware infections." *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016.
- [6] Bilge, Leyla, Yufei Han, and Matteo Dell'Amico. "Riskteller: Predicting the risk of cyber incidents." *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017.
- [7] Lévesque, Fanny Lalonde, José M. Fernandez, and Anil Somayaji. "Risk prediction of malware victimization based on user behavior." *2014 9th international conference on malicious and unwanted software: The Americas (MALWARE)*. IEEE, 2014.
- [8] Liu, Yang, et al. "Cloudy with a chance of breach: Forecasting cyber security incidents." *24th USENIX Security Symposium (USENIX Security 15)*. 2015.

- [9] Miura, Hiroya, Mamoru Mimura, and Hidema Tanaka. "Macros finder: Do you remember loveletter?." International Conference on Information Security Practice and Experience. Springer, Cham, 2018.
- [10] Bahtiyar, Şerif, Mehmet Barış Yaman, and Can Yılmaz Altıniğne. "A multi-dimensional machine learning approach to predict advanced malware." Computer networks 160 (2019): 118-129.
- [11] Mimura, Mamoru, and Taro Ohminami. "Towards efficient detection of malicious VBA macros with LSI." International Workshop on Security. Springer, Cham, 2019.
- [12] Shahini, Maryam, Ramin Farhanian, and Marcus Ellis. "Machine Learning to Predict the Likelihood of a Personal Computer to Be Infected with Malware." SMU Data Science Review 2.2 (2019)
- [13] K. Murphy, Machine Learning: A Probabilistic Perspective (MIT Press, Cambridge, MA, 2012)
- [14] Winston, Patrick Henry. Artificial intelligence. Addison-Wesley Longman Publishing Co., Inc., 1984.
- [15] R. Colom, S. Karama, E. Jung, and R. J. Haier, "Human Intelligence and Brain Networks," Dialogues in Clinical Neuroscience, vol. 12, no. 4, pp. 489–501, 2010
- [16] Duan, Yanqing, John S. Edwards, and Yogesh K. Dwivedi. "Artificial intelligence for decision making in the era of Big Data—evolution, challenges and research agenda." International journal of information management 48 (2019): 63-71.
- [17] C. M. Signorelli, "Can Computers Become Conscious and Overcome Humans?," Frontiers in Robotics and AI, vol. 5, pp. 1–20, 2018.

- [18] N. Kühn, M. Goutier, R. Hirt, and G. Satzger, "Machine Learning in Artificial Intelligence: Towards a Common Understanding," in Proceedings of International Conference on DSystem Sciences (HICSS-52), pp. 1–11, 2019.
- [19] Seeger, Matthias. Learning with labeled and unlabeled data. No. REP_WORK. 2000.
- [20] Threat Intelligence – Macro malware'. Microsoft Protection Centre. Accessed May 2017. www.microsoft.com/security/portal/enterprise/threatreports_july_2015.aspx.
- [21] Edmond and O'Brien, Darragh, "Detection of malicious VBA macros using Machine Learning methods," (AICS 2018), 6-7 Dec 2018.
- [22] S. Kim, Seokmyung Hong, Jaesang Oh, H. Lee, " Obfuscated VBA Macro Detection Using Machine Learning," 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Luxembourg City, Luxembourg, June 2018.
- [23] Mamoru Mimura, Taro Ohminami, "Towards Efficient Detection of Malicious VBA Macros with LSI" Advances in Information and Computer Security, IWSEC, June 2019
- [24] Mamoru Mimura, "Using fake text vectors to improve the sensitivity of minority class for macro malware detection", National Defense Academy, Hashirimizu, Yokosuka, Kanagawa, Japan, October 2020
- [25] Sharif, Mahmood, et al. "Predicting impending exposure to malicious content from user behavior." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.
- [26] Canali, Davide, Leyla Bilge, and Davide Balzarotti. "On the effectiveness of risk prediction based on users browsing behavior." Proceedings of the 9th ACM symposium on Information, computer and communications security. 2014.

- [27] Gratian, Margaret, et al. "Identifying infected users via network traffic." *Computers & Security* 80 (2019): 306-316.
- [28] Kang, Chanhyun, et al. "Ensemble models for data-driven prediction of malware infections." *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016.
- [29] Bilge, Leyla, Yufei Han, and Matteo Dell'Amico. "Riskteller: Predicting the risk of cyber incidents." *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017.
- [30] Lévesque, Fanny Lalonde, José M. Fernandez, and Anil Somayaji. "Risk prediction of malware victimization based on user behavior." *2014 9th international conference on malicious and unwanted software: The Americas (MALWARE)*. IEEE, 2014.
- [31] Liu, Yang, et al. "Cloudy with a chance of breach: Forecasting cyber security incidents." *24th USENIX Security Symposium (USENIX Security 15)*. 2015.
- [32] Miura, Hiroya, Mamoru Mimura, and Hidema Tanaka. "Macros finder: Do you remember loveletter?." *International Conference on Information Security Practice and Experience*. Springer, Cham, 2018.