|  |  |
| --- | --- |
| CSG3101 (APPLIED PROJECT) – PROJECT PROPOSAL | |
| **SEMESTER:** | 232 |
| **GROUP MEMBERS:** | Midharshan Saveen Wickramaratne (10624195) (Leader),  Mihisha Ranatunga (10627681),  Neroshan Mohan (10628000),  Chamod Kanchana Hemantha (10577115),  Ivangka Sevwandhi Liyanage (10627405) |
| **PROJECT TITLE:** | Malware Analysis Toolkit |
| **SUPERVISOR:**  **AIM:** | Ms. Dinusha Kekulawala |
|  | |

## 

1. **Background to the Problem**

**1.1 Problem**

**1.2 Significance of the problem**

**1.3 Aims**

To design, develop and evaluate a machine learning model to analyze dangerous malware in vulnerable systems of the cyber world.

**1.4 Objectives**

1. To identify the problem ….
2. To conduct literature review to identify issues with malware
3. To conduct LR to identify how ML can be used to detect and prevent attacks
4. To identify the evaluation metrics to measure the performance of the ML model
5. To identify the requirement elicitation and analysis
6. To design …
7. To design the model
8. To implement the …
9. To train and test the model using identified evaluation metrics
10. To finetune the model
11. To document

The number of malicious software variants that have been created and distributed globally has increased significantly over the past couple of years. In September 2019, it emerged that around 17.70 million new cases were created. Every day through the Internet, cybercriminals publish new malware or dangerous programs in an effort to steal or delete crucial data. The cyber community is therefore very interested in studies on data protection (Chowdhury et al., 2017). Despite the numerous features of anti-malware software, they are still unable to prevent sophisticated malware from wreaking its effects. It has developed effective ways of evading detection through methods such as self-modification, resilient mutations, and complex obfuscation techniques. As a result, the vast majority of businesses depend on their staff's training to develop their skills in malware analysis. The alternative methods in place for this cyber conundrum are still yet to be properly probed upon.

The problem has effects on a wide range of entities, involving individuals, corporations, governments, and the digital world. The scope of the problem is enormous, impacting customers all across the world “As not all systems are created or designed with the security aspect in mind, new opportunities are opening for cyber-criminals to exploit, cause damage, earn money, or steal data” (Stoleriu et al., 2023). Malware attacks have had a substantial detrimental impact on many internet users. Based on statistics, malware infects 10% to 20% of all computers connected to the internet globally. On an annual basis, attacks on cybersecurity cost billions of dollars financially as well as the resources required to repair them. Ransomware assaults increased by 118% in volume in 2019, according to McAfee Labs, while banking Trojan attacks increased by 100% between June and September of last year, making them the malware family with the fastest growth (Mat et al., 2021). Labeled data suggests that, in this instance, some of the inputs have already been mapped to the output so as to show that similar datasets selected with chosen samples of dangerous malware amounting to exceed 10,000+ records with 90% malicious data and 10% normalized data shall be preprocessed into the essence that it shall entail finding and fixing data flaws or anomalies such missing values, outliers, and duplicates. Data cleaning can be done using a variety of methods, including imputation, removal, and transformation. Thus, upon further reconciliation evaluations metrics include, for multi-class classification accuracy (proportion of correct classifications), precision (per-class), recall (per-class), F1 score (per-class), and the confusion matrix (detailed predictions breakdown). Through this regression, for predicting numeric values, metrics involve mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (variance explanation). Anomaly detection employs true positive rate (sensitivity), true negative rate (specificity), precision, recall, and F1 score. Clustering, for unsupervised grouping, is assessed through silhouette score (sample closeness), Davies-Bouldin index (cluster similarity), and adjusted Rand index (similarity adjusted for chance) for comparing true and predicted clusters' labels. “The algorithms are compared for effectiveness and efficiency using their predictive accuracy, among others, as a performance metric. From the studies, we observe that the best detection rate was attained for supervised learning with feature selection.” (Babaagba & Adesanya, 2019). The test dataset is used to determine whether the machine can predict the output after the machine has been trained using input and output in accordance with 80% training data and 20% testing data.

To automate various processes of malware analysis, machine learning aided malware analysis become essential and thus may employ mechanisms of logistic regression. The goal of this project is to develop a real-time malware analysis system that uses machine learning techniques to identify and categorize threats in uploaded files. This method would allow systems to more effectively prevent attacks. Thus, the need to build, refine, and assess a machine learning model for analyzing harmful malware in cyberspace's vulnerable systems. The toolbox makes it possible to draw conclusions about the intentions, behavior, and traits of the malware and thus its intent seekers as well as how cybercriminal networks are so widespread in the aspects to respond effectively. “Machine learning (ML) is among the promising techniques for such types of predictions, while the sandbox provides a safe environment for such experiments. After thorough literature review, carefully chosen ML techniques are proposed for the malware detection, under Cuckoo sandboxing (CS) environment” (Alhaidari et al., 2022). Furthermore, the malware analysis toolkit is a crucial resource in the arsenal of malware analysts, ethical hacker & incident responders teams that could be considered as target audiences of such a toolkit. This comprehensive toolkit offers a range of capabilities critical to effective cybersecurity; Identify and Classify Malware, Understand Malware Behavior, Incident Response, Mitigation Strategies, etc. Further it is extremely challenging to respond promptly due to the rapidly increasing variety and number of malware species evolving over time. At such a time of need the methods of supervised learning may show prominence in such consequential scenarios as to adopt methods that train algorithms to predict output based on training using "labeled" datasets with improvised anomaly detection based logistic regression. The model stands out from the conventional supervised learning algorithms in place and used in the recent past “Furthermore, the Neural Network (NN) algorithms including Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Full Order Radial Basis Function (RBF) algorithms are used to simulate several experiments from the literature. The simulation results show that the accuracy of MLP, CNN, and Full Order RBF are 96.95%, 98.22%, and 98.68%, respectively.”(Bawazeer et al, 2021).

In conclusion, the creation of a malware analysis toolkit is essential in the fight against cyber threats. Thus bearing in mind, the implementations so far in machine learning were shown to be prominent in various cases, the loopholes sought by the evolving malwares are to be dealt with much needed precision and such a machine learning algorithm integrated toolkit has been improvised.

## **Proposed Methodology**

2.1 Research Methodology

Research methology for this project has been covered using Saunder’s research onion (cite)

|  |  |  |
| --- | --- | --- |
| Research Philosophy | Positivism | Give the justification for selecting the philosophy |
| Research Approach | Deductive | “ |
| Research Strategy | Experiment, Self evaluation, braining storming and LR | Justify |
| Research Choice | Multi method |  |
| Time Horizon | Cross sectional? |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Research Philosophy

A positivist research philosophy is used in this project, which corresponds to a quantitative worldview. This decision is based on the problem's subjective nature, which aims to address the issues of detecting malware in files using machine learning methods. This philosophy aims to measure the malware detection efficiency and accuracy.

Research Approach

The deductive approach will be employed for this assignment. The deductive technique is founded on existing theories and information. It is used in this study to test the hypothesis that using machine learning methods to analyze files improves malware detection accuracy. Using this approach it is possible to determine the efficiency of the toolkit through quantitative evaluation and data gathering.

Research Strategy

This project involves multiple strategies, it is a combination of literature reviews, experimentation, Brainstorming, compare and contrast and supervisor's assistance. These methods are used to improve the toolkit's approach in developing and testing with existing knowledge and real-world insights.

**Literature Review:** The goal is to identify existing malware detection methods and identify gaps to develop a new toolkit with the use of machine learning. This helps to identify current methods and to improve the efficiency of the current tools.

**Experimentation:** The experimental part of the project involves data collection to create and test the malware detection model that is based on machine learning. This involves gathering various datasets, extracting features, training models, and doing real-time analysis.

**Brainstorming and Supervisor Consultation:** Brainstorming sessions are to be held throughout the project where the team and the supervisor communicate to improve the methodology, model, and toolkit features

**Compare and contrast:** With an accuracy to detect the attack, comparing the efficacy of the machine learning-based methodology which is to be built with the existing malware detection approaches.

Research Choice

The research choice methodology is multi-method, using a wide range of approaches which combine quantitative and qualitative procedures. As the project comprises a quantitative deductive approach, the literature review and advice from the supervisor are qualitative, and the comparison and contrast is both. This allows to evaluate the toolkit's performance and efficiency while gaining qualitative insights into the real-world implementation as well as potential problems.

Time horizon

A longitudinal time frame is used as the data is collected before and after implementing. This makes it easy to assess the toolkit's efficiency and flexibility in different situations, as well as how it responds to malware attacks.

Development Methodology

Main approach for the development of this project is prototyping. Prototype is chosen as the project's requirements are unclear at the start and are likely to grow as it progresses and a prototype-based approach lets iterative development and modification of the toolkit as new insights and requirements develop.

Design

The project will apply the Object-Oriented Analysis and Design Methodology (OOADM), which encourages modularity, scalability, and adaptability as the malware analysis toolkit requires an extensible design.

Programming Paradigm

The programming paradigm chosen is Object Oriented Programming as the malware analysis toolkit requires a modular and organized development.

Evaluation Methodology

Solution methodology

What are the steps to build a ML model and how you are planning to do each of them

Project Management Methodology

What is the project management methodology to use

Scope (Initial set of functional and nonfunctional requirements)

In-Scope

Out-Scope

**Deliverables & Dates**

1. Primary Malware Classifying Model

* Create a straightforward machine learning model to sort malware samples into a small number of specified classes.
* Due to time constraints, use a relatively limited dataset for training.

1. Basic Feature Selection

* Use fundamental feature extraction methods to record the key traits of the malware samples.
* Concentrate on a subset of salient characteristics pertinent to the selected malware types.

1. User Interface (UI) Prototype:

* Design a simple user interface that enables users to upload files for analysis.
* Show a high-level description of the file's attributes along with a basic classification result.

1. Core Visualization

* Create simple visualizations highlighting key findings from the investigation, like a pie chart or Gantt chart displaying the distribution of various malware classifications.

1. Information Guides and Operating Steps

* Produce clear user manuals that explain how to use the product in detail.
* Describe the classification strategy used, the attributes chosen, and the tool's limitations.

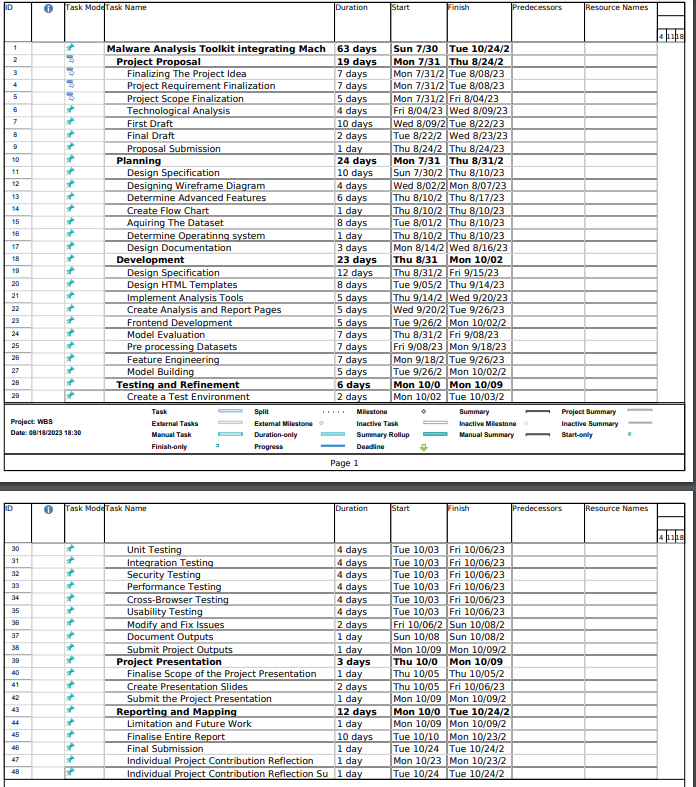
1. Analysis and Attestation

* To make sure the tool produces accurate findings, test it extensively with fictitious malware files.
* Use a small validation dataset to verify the tool's performance.

1. Presentation to Supervisor:

* Make a presentation outlining the tool's features, constraints, and possibilities in the future.
* Demonstrate the tool to your manager, emphasizing its features and the actions you've made to address the project's goals.

1. Malware Identification Reports - Deliver reports that accurately identify and classify different types of malware based on the behaviors, characteristics, and patterns. These reports aid in understanding the nature of the threats and inform appropriate response strategies.
2. Behavioral Analysis Findings - Provide detailed insights into the behavior of analyzed malware, including the actions it takes within a system, interactions with network resources, and potential impact on data integrity and confidentiality.
3. Dynamic Analysis Reports - Offer observations and data collected during the environment. These reports shed light on runtime behaviors, system modifications and network communications initiated by the malware.

**Project Schedule**

**Risk Factors**

Following are ethical, legal, professional and social issues that may arise and need full precedence of all involved entities.

* **Privacy Violation:** Analyzing malware may involve accessing potentially confidential sensitive data. Assurance that the toolkit and practices respect user privacy and comply with relevant data protection regulations.
* **Misuse of Results:** The analysis results could potentially be used for malicious purposes. Implementing safeguards to prevent your toolkit from being misused and educate users about responsible and legal accountability use of malicious intent seekers.
* **False Positives/Negatives:** Regularly updating, patching and validating the toolkit's accuracy to minimize false results.
* **Intellectual Property Violation:** Assurance that the legal permission to analyze the software and adhere to software licensing agreements.
* **Unauthorized Access:** Gaining unauthorized access to systems or networks during analysis could violate cybersecurity laws. Conducting an analysis in controlled environments and following legal and ethical guidelines as per terms of conditions.

**Resources**

**Hardware Requirements**

* Core i5 processor – being able to use Python and develop code while conducting research.
* 8 GB RAM – to compile code as well as to Run machine learning algorithms.
* 10 GB storage – to download relevant applications in order to run the program. (Python IDE, Etc.)

**Software Requirements**

* Windows 10 or 11 64bit Operating System - An operating system that can manage the huge computing activity of machine learning is necessary. The most recent version of a 64-bit operating system will be able to manage such a resource-intensive task.
* Python 3.11 – In order to develop malware analysis toolkit using ML python should be integrated.
* **Python flask -**
* **html/css/javascript -**
* GitHub – In order for the whole team to collaborate and work as a team GitHub is used as a hub for the team to collaborate.
* Python Libraries- For coding in machine learning.
* MS Office / Google Docs - a program to generate reports and documentation.
* MS Project – to plan, allocate resources to tasks, monitor development, and assess workload.
* MS Excel – To store the Dataset.
* Oracle VM (Virtual Machine) Box- As a secure environment with Windows 10/11 64bit Integrated VM Image.
* Google Drive – To Backup Important Documents/files and extra storage use.

• Ivangka and Chamod Fill out your Requirements down here

**Data Requirements**

* A fully fledged comprehensive dataset .csv file with 10,000+ datasets.
* 90% Malicious Data, 10% Normalised Data.

**Skills Requirements**

| **Capabilities** | Ivangka Liyanage  (Software Development) | Chamod Thanippulige  (Software Development) | Mihisha Ranatunga  (Cyber Security) | Neroshan Mohan  (Cyber Security) | Saveen Wickramaratne  (Cyber Security) |
| --- | --- | --- | --- | --- | --- |
| ***Leadership and Role Acquisition*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Researching*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Planning*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Python Knowledge*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Punctuality*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Machine Learning*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Decent Problem Solving*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Resource Allocation*** | ✅ | ✅ | ✅ | ✅ | ✅ |
| ***Report Skills*** | ✅ | ✅ | ✅ | ✅ | ✅ |

**References**

1. Alhaidari, F.,et al., (2022). ZeVigilante: Detecting zero-day malware using machine learning and sandboxing analysis techniques. *Computational Intelligence and Neuroscience*, 2022, 1–15. Retrieved from- <https://doi.org/10.1155/2022/1615528>
2. Babaagba, K. O., & Adesanya, S. O. (2019). A study on the effect of feature selection on malware analysis using machine learning. *Proceedings of the 2019 8th International Conference on Educational and Information Technolog*y. Retrieved from- <https://doi.org/10.1145/3318396.3318448>
3. Chowdhury, M., Rahman, A., & Islam, R.M. (2017). Protecting data from malware threats using machine learning technique. *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 1691-1694.
4. Mat, S. R. T., et al., (2021). Towards a systematic description of the field using bibliometric analysis: malware evolution. *Scientometrics*, 126(3), 2013–2055. Retrieved from- <https://doi.org/10.1007/s11192-020-03834-6>
5. Stoleriu, R., Negru, C., & Radulescu, D. (2023). Modern Cyber Security Attacks, Detection Strategies, and Countermeasures Procedures. *2023 24th International Conference on Control Systems and Computer Science (CSCS)*, 198-205.