**Malware Analysis Toolkit with Machine Learning Integration**

* **Data Collection and Preparation-** Gather a dataset containing malicious samples kaggle.
* **Feature Extraction-** byte sequences, opcode frequency, file metadata
* **Feature Selection-** subset of the features to improve the model's performance
* **Data Splitting**-80 Training and 20 testing sets
* **Feature Scaling**-
* **Model Selection and Training-** Find the most feasible one by developing all possible versions of ML.
* **Hyperparameter Tuning**
* **Model Evaluation**
* **Interpretation and Analysis**
* **Iterate and Refine**
* **Deployment**

**Target Audience**

* Security Software Developers
* Law Enforcement and Government Agencies
* Malware Analysts
* Penetration Testers
* Security Operations Center (SOC) Teams
* Forensic Investigators
* Reverse Engineers
* Threat Intelligence Analysts
* Ethical Hackers
* Incident Responders
* Security Researchers

1. Problem Identification

To a preliminary study on the relevant work

To Identify Research Gap

Review Existing Work

Review Existing Technology

Review Evaluation Methods

1. Literature Review(Abstract, Introduction, Conclusion) To study Existing Work

a. To identify Available Dataset

b. To identify preprocess technique

c. To identify the available technology to build, train and test the model.

d. To identify evaluation metrics available to test the model

e. ……..

Read 5 papers (Fully)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Citation | Brief Summary of the Work | Technology | Limitations | Contribution |
| Yadav, K., Sethi, A., et al. (2022).  <https://www.igiglobal.com/gateway/chapter/292228>  ‌ | The capacity to detect possible future malware by learning from existing malware patterns makes this method very popular | Machine Learning |  |  |
| Babaagba, K. O., & Adesanya, S. O. (2019). https://doi.org/10.1145/3318396.3318448 |  | Malware Analysis using Machine Learning. |  |  |
| Alhaidari, F., Shaib, N. A., Alsafi, et al. (2022)., 1–15. https://doi.org/10.1155/2022/1615528 |  | Using Machine Learning and Sandboxing Analysis Techniques. Computational Intelligence and Neuroscience |  |  |
| Samara, M ., El-Alfy, E., (2019) |  | Open-Source Android Malware  Detection Tools |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Technological Review

|  |  |  |  |
| --- | --- | --- | --- |
| Technology | Pros | Cons | Citation |
| SVM |  |  | ABC (2020) |
| KNN |  |  |  |
| Naïve Bayes |  |  |  |
| Random Forest |  |  |  |

Evaluation 🡪 Review

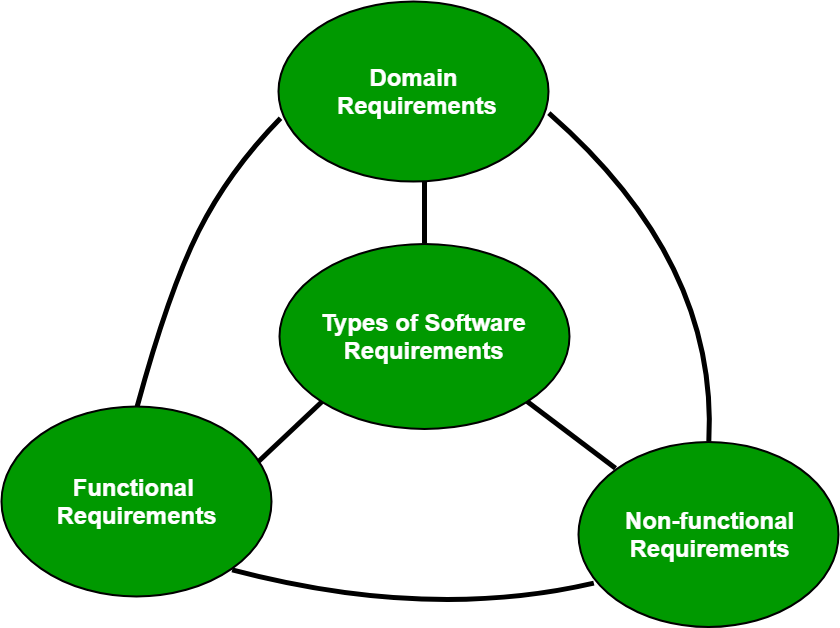
|  |  |  |  |
| --- | --- | --- | --- |
| Dataset |  |  |  |
|  |  |  |  |

1. Requirement Felicitation 🡪 SRS

a. To identify requirements elicitation methods best fits for this project.

b. To identify list function requirements.

c. To identify list of non-functional requirements.



1. Design 🡪 Design Specification

a. To design architecture diagram

b. To derive a system design using OODM

c. To design suitable UI

1. Implementation 🡪 Working Prototype

a. To identify the technology stack

b. To build the machine learning model.

c. To train and test the ML model.

1. Testing 🡪 test result

a. To conduct model testing to measure the most accurate model.

b. To conduct a prototype testing (unit testing, integration testing, system testing)

c. to carryout non-functional testing

1. Finalize the Documentation and Prototype

a. Complete the project report.

8. Methodology

a. Research Methodology- using sander’s onion model

|  |  |
| --- | --- |
| **Research Philosophy** | What and Why? |
| **Research Approach** | What and Why? |
| **Research Strategy** | What and Why |
| **Research Choice** | What and Why |
| **Time Horizon** | What and Why? |

b. Development Methodology

What is the development methodology(third person perspective) The development methodology the team will work on this project is prototyping because the requirements are not clear in the beginning and evolves over the time through literature review, brainstorming and experiments.

i. Design 🡪 OOADM // SSADM (DFD)

ii. Programming paradigm 🡪 OOP or SP or FP or …

iii. Evaluation Methodology🡪 testing MML🡪model testing 🡪 evaluation metrics // Unit, integration and system 🡪 for the prototype solution methodology 🡪 what are different ways in which you can solve the problem.

1. Existing Work

a. What algorithms they have used?

b. What algorithms that can be used.

c. Project Management Methodology

i. Scope

-In scope

a. Preprocessing

b. Feature Engineering

c. Model Building

d. Training and testing

e. Deployment

f. CLI

-Out scope

a. GUI

ii. Schedule

* Gantt Chart

WBS

> Problem Identification

\*Initial LR

\* Finalizing research gap

\* Finalizing Proposal

> Literature Review

\*LR on Exiting Work

\* LR on Technology available

> Requiremnt Elicitation

\* Deriving FR

\* Deriving NFR

> Design

\* Architecture Design

\* System Desigm

\* UI Design

> Implementation

\* Model Implementation

\* UI implementation

> Testing

\* Model Testing

\* System Testing

\* Performance Testing

> Documentation

\* Finalizing the documentation

iii. Risks and Mitigations

* Risks and mitigations 🡪 related to Social, Legal, Ethical, Professional Issues (ACS)

Scope

~~Budget (not there)~~

Time

(Schedule)

9. Resource Requirements

a. Hardware Resources

b. Software Resources

c. Human Resources

i. Who are the people in your team

- What they are good at

- What role they will play

d. Skill Requirements

e. Data Requirements

Obfuscation and packing.

1. Problem Identification

2. Literature Review(Abstract, Introduction, Conclusion)

Initial Literature Review Existing published work in IEEE, ACM 🡪 Semantic Scohlar

5 research papers (last 5 year) 🡪 Abstract 🡪 Introduction 🡪 Conclusion

From this research paper you have to read 🡪 Research Gap

3. Problem background

a. Problem

b. Significance of the problem, can even bring statistics (Should be cited)

c. Aim-one sentence (To design develop and evaluate a machine learning model to identify high risk target audiences(High Risk Students) in risk area(Academic Performance) in the field targeted(Tertiary Education))

d. Objectives(Has to SMART)

i. To identify the problem related to generating tags from the product manually is time consuming and labour-intensive task there to automate using deep learning,

ii. To conduct literature review on how currently deep learning used in generating tags from the product.

iii. To identify available options to generate tags from the product image.

To design, develop and evaluate a machine learning model to identify high risk target audiences(High Risk Students) in risk area