Software Requirement Specification Document for Deep learning models for network anomaly detection

Eslam Khaled Barakat[1], Hassan Hesham Bassiouny[2],
Hazem Mohamed Hamza[3], Yusuf Mohammed Salem[4]
Supervised by: Dr Sahar Abdelrahman, Eng Mohamed Khaled

April 29, 2024

Table 1: Document version history

Version	Date	Reason for Change
1.0	7-JAN-2024	SRS First version's specifications are defined.
1.1	9-JAN-2024	Function Requirements, Data Design
1.2	12-JAN-2024	Class Diagram, Use case Diagram, Context Diagram

GitHub: `https://github.com/Eslamkb/GradProject`

Contents

Abstract

As the complexity and scale of present day computer networks proceed to extend, the challenge of detecting anomalous activities within these networks becomes increasingly critical for ensuring cybersecurity. Traditional methods of network anomaly detection often struggle to adapt to the dynamic and advanced nature of modern cyber threats. This project aims to implement deep neural networks models for detecting several attacks and classifying them based on learning complex patterns using rare anomalies detected from the traffic data. Many feature selection and re-sampling methods will also be tested to get the highest accuracy possible. the chosen models will be tested on UNSW-NB15 [1][Accessed 13FEB2024]. and KDD-CUP99 in order to cover as many types of attacks and protocols as possible and to be ready to be implemented later[2][Accessed 16FEB2024].

1 Introduction

1.1 Purpose of this document

The purpose behind this Software Requirements Specification (SRS) document is to frame the improvement of a deep learning models for network anomaly detection system. This document provides determinations of the system's functionalities, including model preparation and anomaly detection algorithms. It ensures that stakeholders and developers have a common understanding of the system's goals, capabilities, and limitations by acting as a road map. The SRS will direct the improvement of efficient anomaly detection system that leverages advanced deep learning techniques to identify and respond to network irregularities, in this way upgrading network security and dependability.

1.2 Scope of this document

The SRS_document presents similar_systems to this project, displays the OverView, Scope and Context of project's system_design. Moreover, this document entirely explains this project's functional and non_functional requirements, the design_limitation, the DataDesign and the Class-Diagram. Finally, this document discusses the possible operational_scenarios and presents the project's TimePlan.

1.3 Business Context

This project centers around the development of a sophisticated deep learning model for network anomaly detection, designed to enhance cybersecurity protection in organizational network systems. The need for sophisticated, automated solutions to identify and eliminate potential network breaches is critical currently when digital threats are becoming increasingly complex and frequent. This project aims to use deep learning techniques to analyze network traffic patterns and identify anomalies indicative of cybersecurity threats. It aims to make the industry have powerful tool with good security detection models to minimize the risk of data breaches and the integrity of the assets. The success of this project will not only enhance network security but also position the industry as a leader in deploying AI-driven solutions for cybersecurity, potentially opening avenues for future innovations and collaborations in this rapidly evolving field[3][Accessed 7JAN2024].

2 Similar Systems

2.1 Academic

NetworkAnomalyDetectionUsingDeepLearningTechniques:[4] [Accessed 7JAN2024]This paper proposes a deep learning-based approach for network anomaly detection in cyber-security using a one-dimensional CNN architecture. Based on an evaluation using the UNSW-NB15 dataset, the suggested model performs better than two recent studies in terms of f-score, accuracy, precision, and recall. The authors also note that their approach works much better for the attack classes when comparing it to a prior study that was done for each class. As shown in Figure 1, The class imbalance issue in the dataset is addressed by the authors through the use of the synthetic minority over-sampling technique. The study comes to the conclusion that real-time network security applications can benefit from the suggested method's effectiveness in identifying anomalies in the network. Future research, according to the authors, ought to concentrate on enhancing the model's functionality for the "other" class and evaluating it on various datasets. In general, by examining the use of 1-D CNN in anomaly detection, the research significantly advances the field of network security. Its methodical approach, careful data preparation, and in-depth research are its main advantages. However, the research would be more thorough and applicable to real-world circumstances if it addressed issues like generalization, larger comparison analysis, and practical deployment considerations.

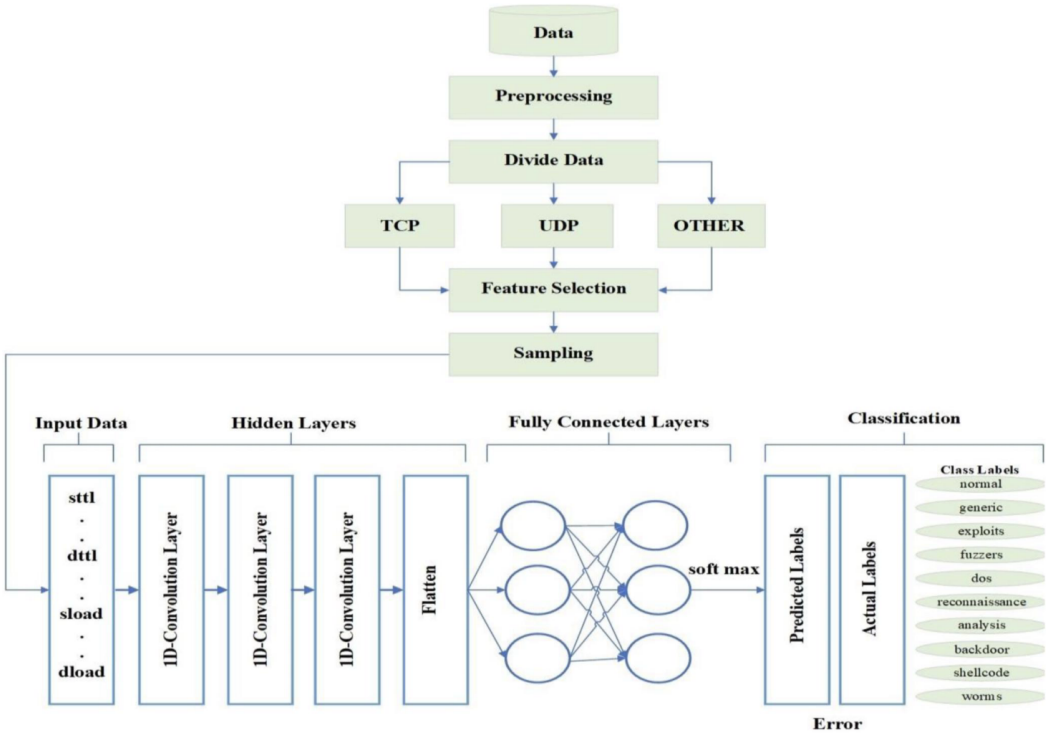

Figure 1: CNN Overview Model

PerformanceAnalysisOfIntrusionDetectionSystemsUsingAFeatureSelectionMethodOnTheUNSW-NB15_Dataset:[5][Accessed 13JAN2024] This paper shows the implementation of intrusion de-

tection system(IDS) using ML techniques(KNN)(SVM)(DT)(ANN)(LR) on UNSW-NB15 dataset .The paper shows the importance of ids's and the crucial role it plays in the computer network security.It focuses on showing the importance of using effective feature selection methods to enhance the performance of the (IDS).The research evaluates different machine learning algorithms and feature selection techniques,showing the results and discussing the possible enhancements of IDS performance which shows that using(ANN) model on multi class classification the accuracy is (77.51%).Furthermore,the paper supply insights into the experimental setup, hardware environment and the performance metrics used for evaluation. Overall, the paper provides a detailed analysis of IDS performance and feature selection methods.

HybridDeepLearningBasedAttackDetectionForImbalancedDataClassification:[6][Accessed 7JAN2024] This paper presents a novel hybrid deep learning model for Intrusion Detection Systems (IDS) in IoT industry, tending to the test of imbalanced information in network intrusion detection. The proposed model joins Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) algorithms, and utilizes a synthetic data generation method to solve the imbalanced data problem issue in the UNSW-NB15 dataset. The review analyzes the exhibition of the proposed model with basic CNN models, benchmark machine learning models (Decision Tree and Random Forest), and related works. The results exhibit that the proposed model accomplished an accuracy of 92.10%, beating different models in terms of Area Under the Curve (AUC), Recall, Precision, and F1 score. Furthermore, the paper examines the feature selection and preprocessing methods used to enhance the model's result, giving significant experiences to future exploration and practical applications in IoT security.

Figure 2: Comparison between papers and this paper

2.2 IDS flowchart

Figure 3: Deep Instinct

The Deep Instinct Prevention is a Platform that block unknown threats before they can do any action with lowest false positive rate and highest accuracy in the industry.

Figure 4: Darktrace

Darktrace Detect by meticulously examining a vast array of metrics, this system uncovers subtle anomalies that might point to an emerging threat, even if those threats involve novel malware or previously unseen techniques. It expertly distinguishes between malicious and benign activities, effectively identifying attacks that typically go undetected.

3 System Description

3.1 Problem Statement

The field of cybersecurity is dynamic and changes rapidly, Intrusion Detection System became essential. Implementing the suitable deep learning models is very important to enhance the network security. The UNSW-NB15 and KDD are widely the most known datasets which contains several types of attacks. However, There is a lack of comparative analysis of different deep or machine learning techniques. This gap needs to be addressed to find the optimal solution for the real world IDS. A lot of models struggle to detect the attacks effectively which cause a very high false positive rate, Achieving high accuracy with achieving effective monitoring is a problem that most intrusion detection systems face[1][Accessed 13JAN2024].

3.2 System Overview

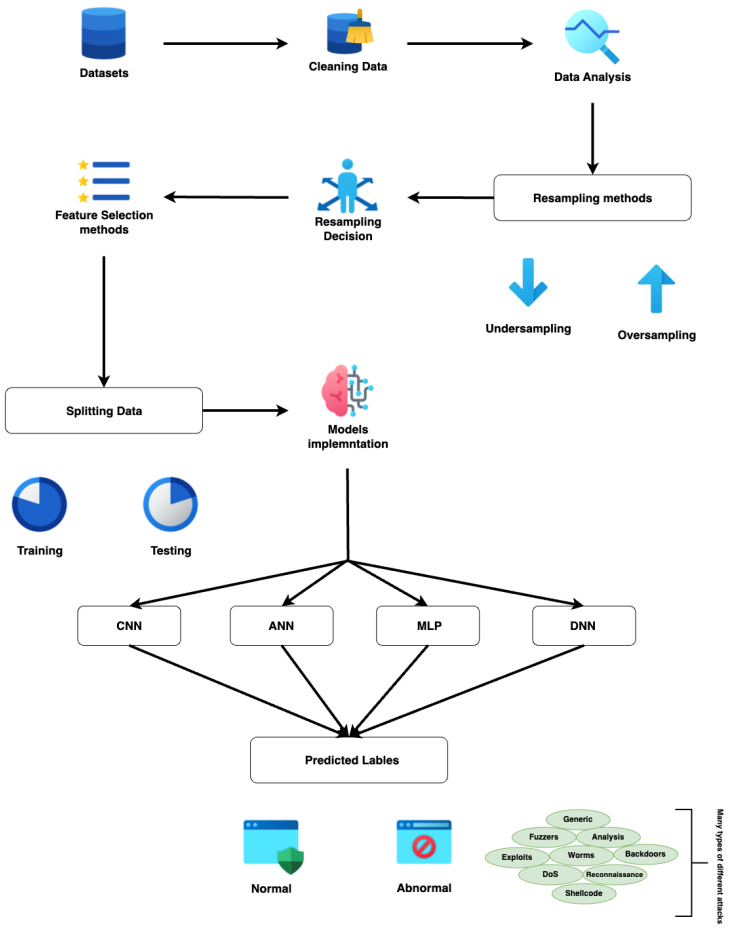

Figure 5: System_Overview

As shown in Figure 5 the description of the overview diagram:

- **Dataset:** Start with importing dataset file
- **Cleaning Data:** Removing noise, duplicates, and irrelevant information.
- **Data Analysis:** Analyzing the data to comprehend its characteristics and discover any imbalances.
- **Resampling Methods:** Choosing or combining between undersampling (reduce the majority class) and oversampling (increase the minority class) to balance the dataset.
- **Resampling Decision:** Comparing each method and choose the most effective and suitable for the models' accuracy.
- **Feature Selection Methods:** Applying methods to choose the most important features for the models.
- **Splitting Data:** Splitting the dataset into 2 parts; 80% for training and 20% for testing.
- **Models Implementation:** Implementing several suitable deep learning models with the effective output from the resampling and feature selection methods. Then comparing each model accuracy.
- **Predicted Labels:** Classifying the network packets into normal and abnormal, which will be classified into several types of attacks.

3.3 System Scope

Our project is mainly designed to detect the anomaly network activities and to identify them to normal and abnormal with classification[7][Accessed 9JAN2024].
Re-sampling:[8][Accessed 9JAN2024]
UNSW-NB15 and KDD datasets are unbalanced, by using resampling methods (SMOTE, ADASYN) to balance out our dataset, then compare between the methods to identify the best one with the best results.
Feature selection[9]: [10Accessed 7JAN2024]
By using these methods (correlation, random forest, information gain, selectkbest, igrf, rfe, chi2), this process not only improves model accuracy by eliminating redundant or irrelevant data that can lead to overfitting but also reduces computational complexity, leading to faster and more efficient model training.
Classify into Several different types of attacks:

- Fuzzers
- Analysis
- Backdoors

- DoS
- Exploits
- Generic
- Reconnaissance
- Shellcode
- Worms
- R2l
- U2R
- Probe

Implementation:
Implement a software which will Visualize the model and it produces graphs and diagram for the use to make the output understandable.

3.4 System Context

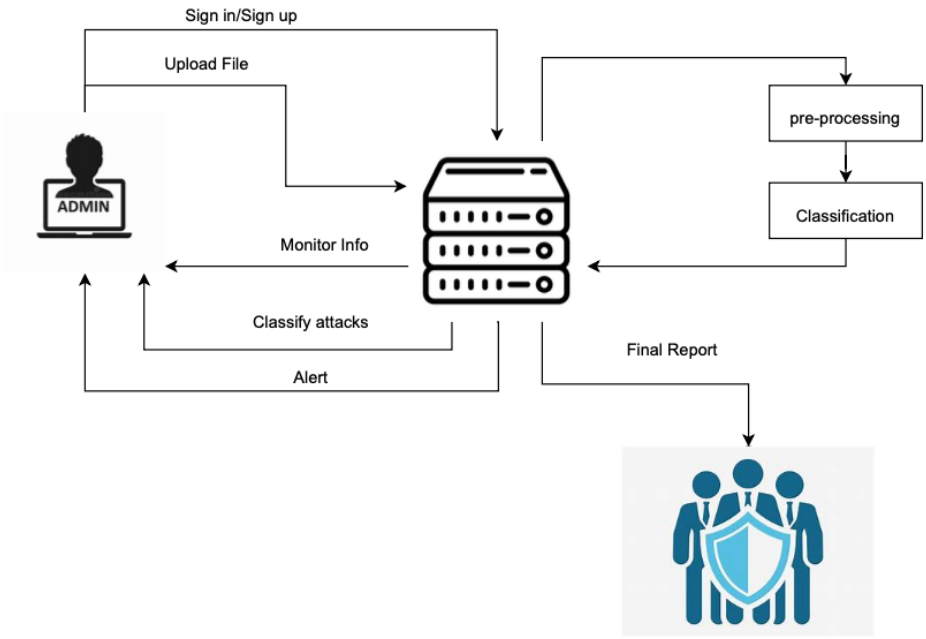

Figure 6: Context_Diagram

As shown in Figure 6, the diagram characterize an organization network security checking work process. An administrator signs in, import network traffic information, which is then preprocessed and classified to detect anomalies. The system effectively monitors the organization network and recognizes potential threats, and issues alerts. At the end a final report with all the activities and classifying the network traffic[10][Accessed 10JAN2024].

3.5 Objectives

- The main aim for the project is to developing an System that can effectively identify and mitigate network anomalies.
- The System will start with emphasis on detecting the 9 types of attacks as specified in the UNSW-NB15 dataset, with the potential for further scalability.
- The system will utilize suitable feature selection and re-sampling algorithms to enhance its detection capabilities.
- The key objective of this project is to achieve high accuracy in detecting network intrusions through the effective implementation of our models.
- A further goal is to minimize the rates of false positives and false negatives, thereby enhancing the reliability and efficiency of the IDS.

3.6 User Characteristics

For normal user:

- Needs to know what is deep learning and how it works
- Some General information about anomaly detection types of attacks
- Needs to know the difference between deep learning models
- Needs to know General information about feature selection
- Needs to know some general information about re-sampling and it's methods

For organisations and corroborates:

- Need to know how to implement different deep learning models
- Need to choose the suitable model from the information gained from the comparison

4 Functional Requirements

The Use Case diagram for Admin and Server that will be using the system is shown in figure 7.

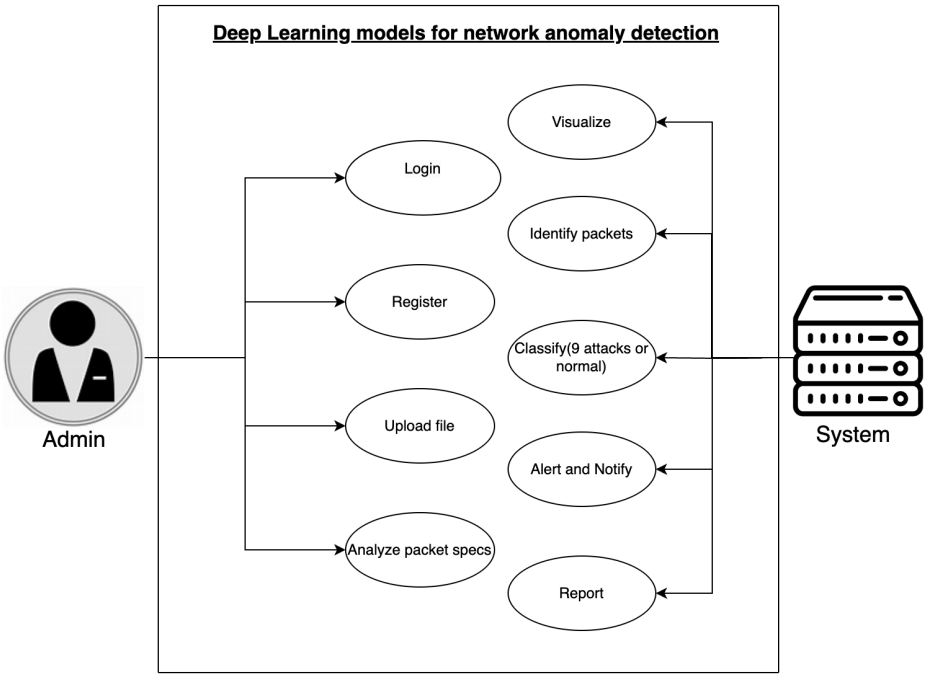

Figure 7: Use_Case_Diagram

4.1 System Functions

- **ID:1 Login:** Admin can login using username and password to access the system.
- **ID:2 Register:** register new user with username and password to allow him/her to login.
- **ID:3 Upload file:** allow user to upload packet files to detect and analyze it.
- **ID:4 Analyze packet specs:** admin can view, and extract needed or required information from each packet.
- **ID:5 Visualize** visualize the output in formative way to be understandable for normal user
- **ID:6 Identify and Classify:** categorize each packet into normal or abnormal which is divided into 9 different attacks.
- **ID:7 Alert and Notify:** Alert the admin if any packet had been classified abnormal.
- **ID:8 Report:** create a report that includes a summarized information about abnormal detected packets.

4.2 Detailed Functional Specification

Table 2: Login

Name	**Admin login**
Code	Fn1
Priority	Very High
Critical	Essential for the admin to access all services on system
Description	It searches in a database for the username and password if valid or not
Input	Username and Password
Output	Boolean(found or not found)
Pre-Condition	Admin must already have a created account
Post-Condition	If found direct admin to dashboard if not say that username or password incorrect
Dependency	none
Risk	A previous session did not end.

Table 3: Visualize

Name	**Visualize**
Code	Fn5
Priority	Low
Critical	none
Description	Visualize the output in formative way to be understandable for normal user
Input	Dataset
Output	Graph and Diagram
Pre-Condition	finishing classification
Post-Condition	the output will be shown in the system
Dependency	Fn8
Risk	none

Table 4: Classification

Name	**Identify and Classify**
Code	Fn6
Priority	High
Critical	None
Description	Classify packets into normal and abnormal is classified into 9 attacks
Input	Packets
Output	Attack types
Pre-Condition	Should receive monitored packets from system
Post-Condition	none
Dependency	Fn3
Risk	Cannot detect day zero attacks

Table 5: Alert

Name	**Alert**
Code	Fn7
Priority	High
Critical	none
Description	It alerts admin that a malicious packet is found
Input	Malicious packets
Output	Alert message
Pre-Condition	none
Post-Condition	Admin can take immediate actions or further investigations.
Dependency	Fn6
Risk	none

5 Design Constraints

5.1 Standards Compliance

Minimum of 16 GB RAM PC

Data-sets: As needed

5.2 Hardware Limitations

Computer: minimum 2000$

server: undetermined

6 Non-functional Requirements

- **Usability:** The user interface should be friendly to make the process of the administrators easier.
- **Security:** The system should only be accessed by the user and disallow any unauthorized access.
- **Reliability:** The system should be available whenever needed and should be able to continue functioning in the presence of hardware and software failures.
- **Performance:** The time of response should be acceptable and define a certain maximum capacity.
- **Maintainability:** the requirements for upgrading the system and adding other types of attacks the can be identified by the system should be addressed.

7 Data Design

There is 2 dataset the were included in this paper which are UNSW-NB15 and KDD-CUP99, Both datasets are benchmarks in the field of security as both of them contain many different types of attacks. Both datasets are available on the internet and there links is added to the references We will talk about each one separately:
UNSW-NB15[11][Accessed 8JAN2024]: Developed at the University of New South Wales (UNSW), is designed to serve as a contemporary benchmark for intrusion detection research in network security. The dataset file format is CSV. It contains a wide range of different network traffic scenarios to mimic real world conditions. The dataset contains 9 different types of attacks which are Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms, beside covering Normal packets also to provide a balanced representation. The features was extracted different algorithms like Argus and Bro-IDS which leaded us to have a dataset containing 49 different feature like the source and destination IP addresses, timestamps, protocol types, and attack labels. The dataset comes with 2 different versions which are the completely unbalanced version which contains a total of 2,540,044 records. It is considered imbalanced due to the number of packets of attacks compared to the normal packets like worms which has a representation with less than 200 row in the dataset. It also comes with 4 CSV file which is indicated for the training and testing of the models. The Training set: 175,341 records while the Testing set: 82,332 records. They are also considered imbalanced but they are more realistic to work on and provides greater efficiency.

No	Class	TCP	UDP	OTHERS	ALL
1	normal	1,436,890	766,685	15,189	2,218,764
2	generic	3118	210,600	1763	215,481
3	exploits	27,443	874	16,208	44,525
4	fuzzers	15,474	6043	2729	24,246
5	dos	3336	527	12,490	16,353
6	reconnaissance	6965	4890	2132	13,987
7	analysis	622	-	2055	2677
8	backdoor	323	34	1972	2329
9	shellcode	750	761	-	1511
10	worms	153	21	-	174
	Total	**1,495,074**	**990,435**	**54,538**	**2,540,047**

Figure 8: UNSW-NB15 Description

The KDD-CUP99: originally introduced for the KDD Cup '99 competition, is a benchmark dataset widely used in the field of intrusion detection and network security research. The dataset file format is CSV. One of the earliest benchmark datasets for evaluating intrusion detection systems. Encompasses a wide range of network attacks, including denial-of-service (DoS), probing, user-to-root (U2R), and remote-to-local (R2L) attacks. Each of those types of attacks has presence in the dataset with many different formats (each one category contains several types of attacks). It includes many features like protocol types, source and destination IP addresses, service types, and attack labels. The dataset is separated into 2 different files, One for training and the other for testing. The dataset consist of 5 million row of network traffic data and also considered imbalanced as more than 8 attacks has less than 10 rows presence in the dataset.

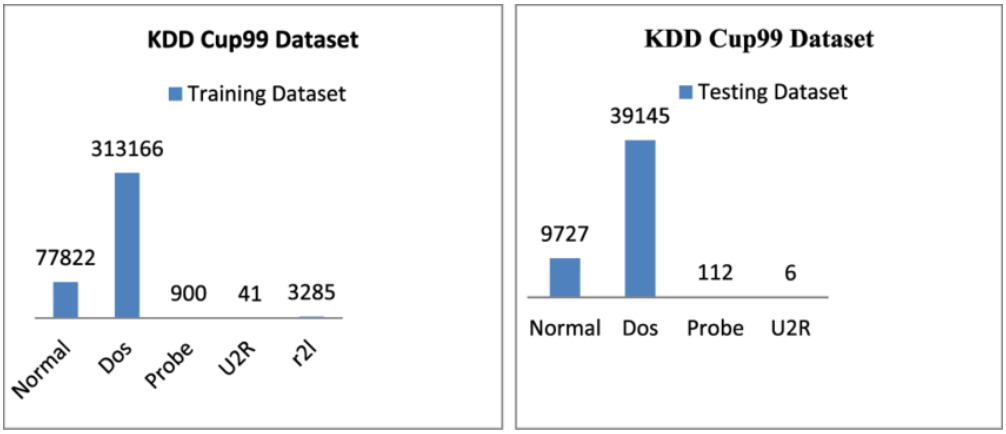

Figure 9: KDD Cup99 Description

8 Preliminary Object-Oriented Domain Analysis

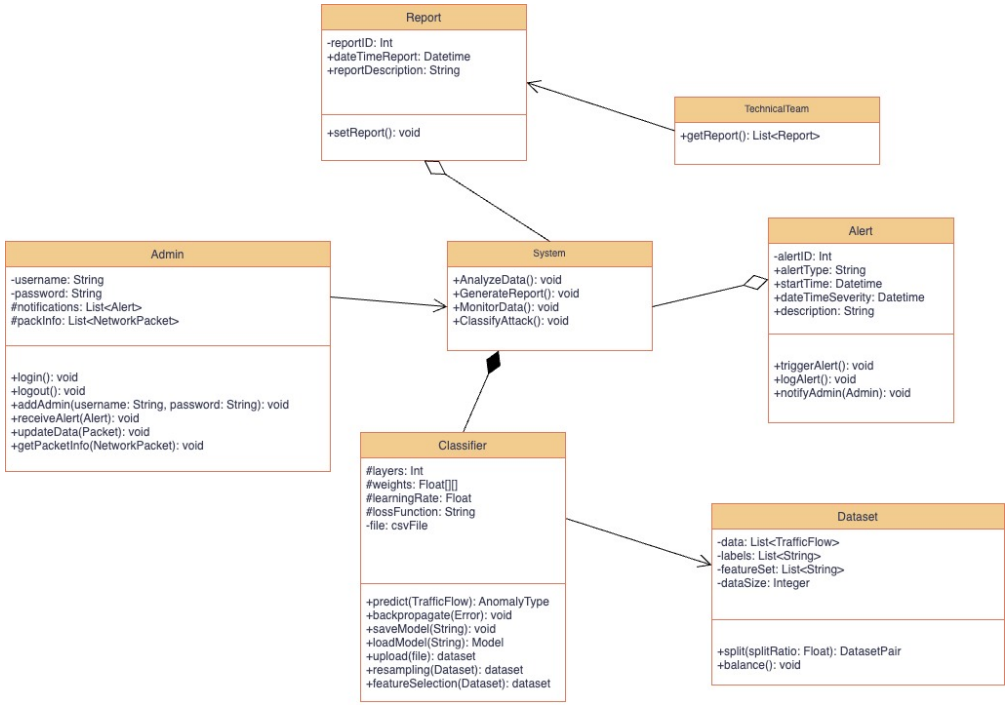

Figure 10: Class_Diagram

9 Operational Scenarios

- **Scenario (1): User login:** The user has the option to login or register a new user.
- **Scenario (2): Network Traffic Monitoring:** Monitoring and analyzing network for anomalies to identify potential security breaches.
- **Scenario (3): Identifying and classifying packets:** The system constantly captures packets, and identify the packets into normal or if the captured packet happens to be abnormal then it is classified to one of the 9 attacks.
- **Scenario (4): Incident response and alerting** Generating alerts or notifications when potential security incidents or anomalies are detected, allowing user to take immediate actions or investigate further.

10 Project Plan

Trello:`https://trello.com/invite/b/5rl7yt4L/ATTI73c75f58975e65e959cdd7ca521f8076A78B89EE/time-plan`

TASK LIST		
MY TASKS	DUE DATE	Role
Idea and Supervisor	22/09/2023	All team members
Information Collection and Researches	19/10/2023	All team members
Survey and Proposal preparation	15/11/2023	All team members
Preprocessing stage	15/11/2023	All team members
Proposal presenation 10%	16/11/20203	All team members
Classify Dataset	04/12/2023	All team members
SRS Preparation	14/01/2023	All team members
SRS Presentation 35%	15/01/2024	All team members
SDD Preparation	06/03/2024	All team members
SDD Presentation 65%	07/03/2024	All team members

Figure 11: Time Plan Table

11 Appendices

11.1 Definitions, Acronyms, Abbreviations

KNN	k-nearest neighbors
SMOTE	Synthetic Minority Oversampling Technique
ADASYN	Adaptive Synthetic Sampling
RFE	Recursive Feature Elimination
IGRF	Hybrid Feature Selection
CHI2	CHI-Square
DOS	Denial Of Service
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory Networks
IDS	Intrusion Detection System
DNN	Deep Neural Network
MLP	Multi Layer Perceptron
SVM	support vector machine
ANN	artificial neural networks
AUC	Area Under Curve
DT	Decision Tree
LR	Logistic Regression

Figure 12: Abbreviations

11.2 Supportive Documents

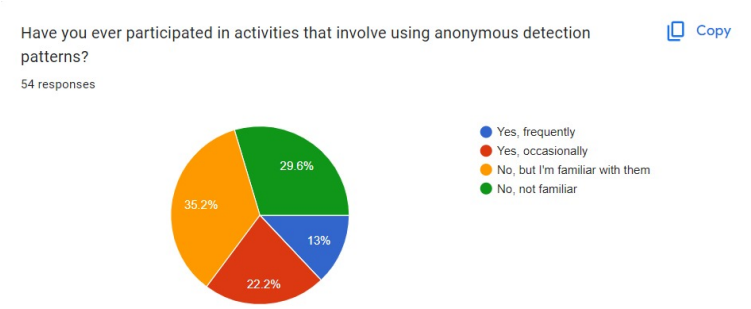

Figure 13: Question 1

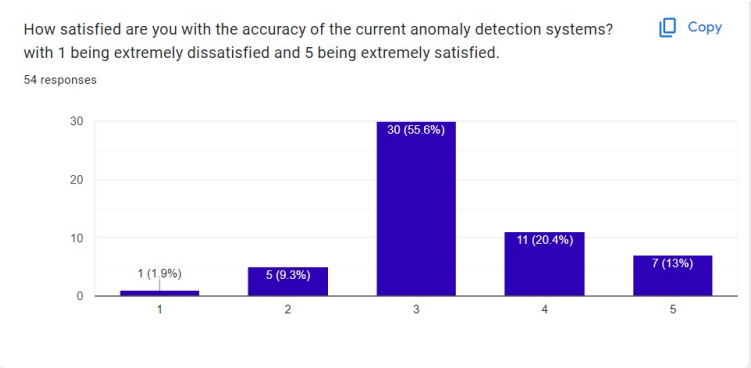

Figure 14: Question 2

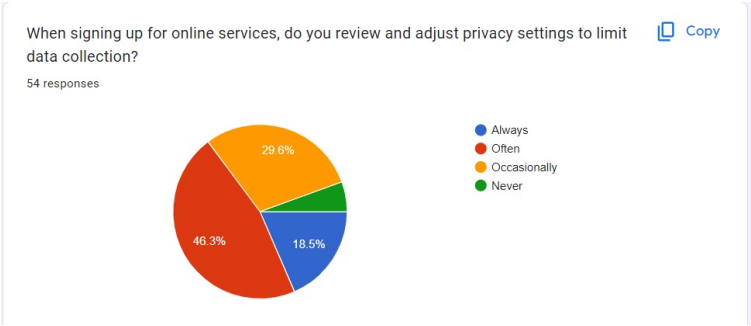

Figure 15: Question 3

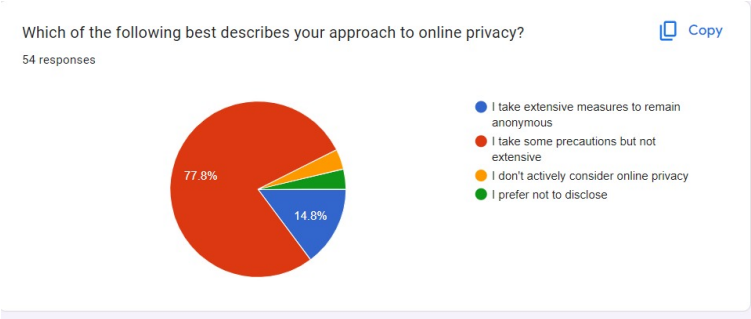

Figure 16: Question 4

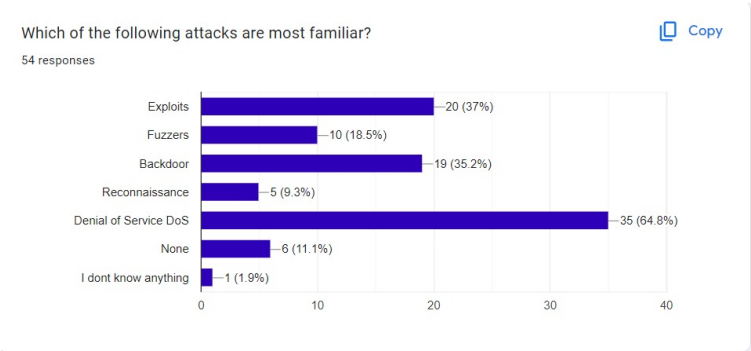

Figure 17: Question 5

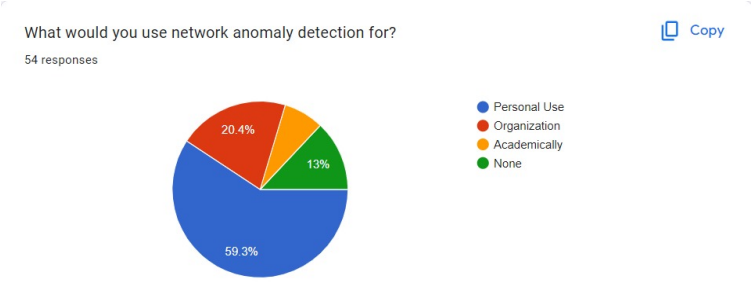

Figure 18: Question 6

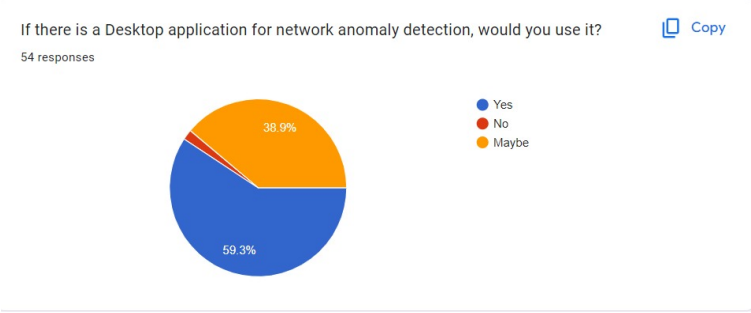

Figure 19: Question 7

References

[1] Utkarsh Dixit, Suman Bhatia, and Pramod Bhatia. "Comparison of Different Machine Learning Algorithms Based on Intrusion Detection System". In: *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*. Vol. 1. IEEE. 2022, pp. 667–672.

[2] Sarika Choudhary and Nishtha Kesswani. "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT". In: *Procedia Computer Science* 167 (2020), pp. 1561–1573.

[3] Ricardo Jorge Santos, Jorge Bernardino, and Marco Vieira. "Approaches and challenges in database intrusion detection". In: *ACM Sigmod Record* 43.3 (2014), pp. 36–47.

[4] Mohammad Kazim Hooshmand and Doreswamy Hosahalli. "Network anomaly detection using deep learning techniques". In: *CAAI Transactions on Intelligence Technology* 7.2 (2022), pp. 228–243.

[5] Sydney M Kasongo and Yanxia Sun. "Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset". In: *Journal of Big Data* 7 (2020), pp. 1–20.

[6] Rasha Almarshdi, Laila Nassef, Etimad Fadel, et al. "Hybrid Deep Learning Based Attack Detection for Imbalanced Data Classification." In: *Intelligent Automation & Soft Computing* 35.1 (2023).

[7] Waad Falah Kamil and Imad Jasim Mohammed. "Deep learning model for intrusion detection system utilizing convolution neural network". In: *Open Engineering* 13.1 (2023), p. 20220403.

[8] Abhishek Divekar, Meet Parekh, Vaibhav Savla, et al. "Benchmarking datasets for anomaly-based network intrusion detection: KDD CUP 99 alternatives". In: *2018 IEEE 3rd international conference on computing, communication and security (ICCCS)*. IEEE. 2018, pp. 1–8.

[9] Souhail Meftah, Tajjeeddine Rachidi, and Nasser Assem. "Network based intrusion detection using the UNSW-NB15 dataset". In: *International Journal of Computing and Digital Systems* 8.5 (2019), pp. 478–487.

[10] Zeinab Zoghi and Gursel Serpen. "Unsw-nb15 computer security dataset: Analysis through visualization". In: *arXiv preprint arXiv:2101.05067* (2021).

[11] Ahmed Aleesa, MOHAMMED Younis, Ahmed A Mohammed, et al. "Deep-intrusion detection system with enhanced UNSW-NB15 dataset based on deep learning techniques". In: *Journal of Engineering Science and Technology* 16.1 (2021), pp. 711–727.