*A Progress Report*

*on*

**Network Anomaly Detection Using Effective Machine Learning Techniques**

*carried out as part of the course CSE CS3270 Submitted by*

***Pradyumna Singh***

**209301309**

***VI-CSE***

*in partial fulfilment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

In

**Computer Science & Engineering**



**Department of Computer Science & Engineering,**

**School of Computing and IT,**

**Manipal University Jaipur,**

***March 2022***

**Acknowledgement**

This project would not have completed without the help, support, comments, advice, cooperation and coordination of various people. However, it is impossible to thank everyone individually; I am hereby making a humble effort to thank some of them.

I acknowledge and express my deepest sense of gratitude of my internal supervisor ………………………………….for his/her constant support, guidance, and continuous engagement. I highly appreciate his technical comments, suggestions, and criticism during the progress of this project “………………………………….”.

I owe my profound gratitude to …………………………………., Head, Department of CSE, for his valuable guidance and facilitating me during my work. I am also very grateful to all the faculty members and staff for their precious support and cooperation during the development of this project.

Finally, I extend my heartfelt appreciation to my classmates for their help and encouragement.

|  |  |
| --- | --- |
| **Registration No.** | **Student Name** |
| **209301309** | **Pradyumna Singh** |



**Department of Computer Science and Engineering**

**School of Computing & Information Technology**

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_

**CERTIFICATE**

This is to certify that the project entitled ***Network Anomaly Detection Using Effective Machine Learning Techniques*** is a Bonafede work carried out as ***Minor Project Midterm Assessment (Course Code: CS3270)***  in partial fulfilment for the award of the degree of Bachelor of Technology in Computer Science and Engineering, under my guidance by ***Pradyumna Singh*** bearing registration number **209301309** during the academic semester *VI of year 2022-23.*

Place: Manipal University Jaipur, Jaipur

Name of the project guide:

Signature of the project guide: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ABSTRACT**

The field of network security has become increasingly complex due to the proliferation of sophisticated cyber threats. To address this challenge, researchers have proposed various machine learning-based techniques for network anomaly detection….. to be continued

Table of Contents

[1: Introduction 6](#_Toc130294835)

[1.1: Motivation 6](#_Toc130294836)

[2: Literature Review 6](#_Toc130294837)

[2.1: as required 6](#_Toc130294838)

[2.2: as required 6](#_Toc130294839)

[2.3: Outcome of Literature Review 6](#_Toc130294840)

[2.4: Problem Statement 6](#_Toc130294841)

[2.5: Research Objectives. 6](#_Toc130294842)

[3: Methodology and Framework 6](#_Toc130294843)

[3.1: System Architecture 6](#_Toc130294844)

[3.2: Algorithms, Techniques etc. 6](#_Toc130294845)

[3.3: Detailed Design Methodologies (as applicable) 6](#_Toc130294846)

[4: Work Done 6](#_Toc130294847)

[4.1: Details as required. 6](#_Toc130294848)

[4.2: Results and Discussion 6](#_Toc130294849)

[4.3: Individual Contribution of project members (in case of group project) 6](#_Toc130294850)

[5: Conclusion and Future Plan 6](#_Toc130294851)

[6: References 7](#_Toc130294852)

# 1: Introduction

In today's world, the reliance on technology and network systems has become ubiquitous. With an ever-increasing number of devices connected to these systems, it's essential to have the ability to identify and respond to any unusual events in real-time. That's where our project comes in. We aim to develop a cutting-edge machine learning solution that can effectively and efficiently detect network issues as they arise.

The primary focus of this project is to find and address any network problems before they escalate into bigger problems that might cause significant damage. Our team will conduct thorough research on various machine learning algorithms and techniques to determine the most suitable approach to solving this issue. The outcome of this research will be documented in a comprehensive research paper, which will provide valuable insights into the problem and how our solution can help.

Aside from the research component, we will also be developing a practical application that leverages the findings from our research. This application will allow network administrators to quickly identify any network issues, allowing them to respond to them promptly. The goal of this project is to make a real difference in the world of technology by helping to ensure the smooth and uninterrupted operation of network systems.

## 1.1: Motivation

Network anomalies can seriously harm a system, resulting in security breaches, data loss, and system degradation. Therefore, it is important to detect these anomalies in real time. Current methods for detecting network anomalies are limited and can be time-consuming. This project aims to overcome this limitation by applying effective machine learning techniques to detect anomalous networks. The use of machine learning algorithms in network anomaly detection will provide a faster and more accurate method for detecting network anomalies.

# 2: Literature Review

## 2.1: Introduction and selection of dataset for IDS

Intrusion detection systems (IDS) are critical components of cybersecurity that help detect and prevent attacks on computer networks. One of the key factors that affect the effectiveness of IDS is the dataset used for training and evaluation purposes. As network attacks evolve, it is important to update datasets to capture new attack patterns and scenarios.

Several characteristics have been identified as critical for building a complete and efficient IDS dataset, including diversity of attacks, anonymity, available protocols, complete network traffic capture, complete network interaction capture, complete network configuration definition, feature set, labelled data samples, heterogeneity, and metadata. However, publicly available datasets may not always match current technological demands and can be statistically deficient.

To address these issues, the CIC-IDS-2017 and CSE-CIC-IDS-2018datasets were developed with the aforementioned characteristics in mind. These datasets provide a comprehensive understanding of attacks conducted and conceptual knowledge of application models, network devices, and protocols. The datasets were generated using the CICFlowMeter tool and include detailed information such as source and destination addresses and ports, timestamps, and attack labels.

Researchers have implemented different classifiers using these datasets, and the files present in the datasets are used for both binary and multi-class classification. However, these datasets also have limitations, including the tedious task of processing large numbers of data instances in each file, merging files to include each attack label, missing and redundant data records, and high-class imbalance. Pre-processing techniques such as feature engineering and relabelling or sampling data samples can address these limitations.

Overall, the availability and quality of datasets play a crucial role in the effectiveness of IDS. While publicly available datasets may have limitations, efforts to develop comprehensive and efficient IDS datasets, such as CIC-IDS-2017 and CSE-CIC-IDS-2018, have been made to improve the accuracy and efficiency of IDS.

### 

## 2.2: Evolution of CSE-CIC-IDS-2018

The CSE-CIC-IDS2018 dataset is an evolution of the ISCXIDS2012 dataset created by the Information Security Centre of Excellence (ISCX) at the University of New Brunswick (UNB) in 2012. The CSE-CIC-IDS2018 dataset was prepared from a much larger network of simulated client-targets and attack machines resulting in a dataset that contains 16,233,002 instances gathered from 10 days of network traffic, with about 17% of instances being attack traffic.

The dataset has a class imbalance and is structurally similar to its predecessor CICIDS2017, which was created by the Canadian Institute of Cybersecurity (CIC) to address the limitations of ISCX2012. CICIDS2017 contains 2,830,743 instances, with attack traffic amounting to about 19.7% of this total number.

The CSE-CIC-IDS2018 dataset is the most recent intrusion detection dataset that is big data, publicly available, and covers a wide range of attack types. The dataset is distributed over ten CSV files that are downloadable from the cloud.

Studies using CSE-CIC-IDS2018 have shown unusually high performance scores, possibly due to overfitting, and a lack of concern for class imbalance and data cleaning, which could hinder reproducibility of experiments.

## 2.3: An Evaluation of Intrusion Detection Models using CSE-CIC-IDS2018

In a recent study, Mohammadi et al. (2021) evaluated the performance of six IDS models using the CSE-CIC-IDS2018 dataset. The models included in the study were Deep Learning-based IDS, Random Forest, Support Vector Machine, k-Nearest Neighbour, Decision Tree, and Naïve Bayes. The authors reported that the Deep Learning-based IDS model achieved the highest detection rate (97.3%) among all the models evaluated.

Similarly, in another study, Almalki et al. (2021) evaluated the performance of four IDS models using the CSE-CIC-IDS2018 dataset. The models included in the study were Random Forest, Decision Tree, k-Nearest Neighbour, and Naïve Bayes. The authors reported that the Random Forest model achieved the highest detection rate (98.32%) among all the models evaluated.

In a more recent study, Akinsanmi et al. (2022) evaluated the performance of three IDS models using the CSE-CIC-IDS2018 dataset. The models included in the study were the Random Forest, k-Nearest Neighbour, and Convolutional Neural Network. The authors reported that the Convolutional Neural Network model achieved the highest detection rate (99.1%) among all the models evaluated. The study also compared the performance of the models using different feature selection techniques and reported that the Recursive Feature Elimination technique performed the best.

## 2.3: Outcome of Literature Review

Overall, these studies demonstrate the effectiveness of the CSE-CIC-IDS2018 dataset in evaluating the performance of IDS models. The results show that Deep Learning-based models and Random Forest models perform well in detecting various types of attacks in network traffic data. Additionally, the studies also highlight the importance of feature selection techniques in improving the performance of IDS models. Future research can continue to explore the use of the CSE-CIC-IDS2018 dataset to develop more accurate and efficient IDS models to improve network security.

## 2.4: Problem Statement

## 

## 2.5: Research Objectives.

# 3: Methodology and Framework

## 3.1: System Architecture

## 3.2: Algorithms, Techniques etc.

## 3.3: Detailed Design Methodologies (as applicable)

# 4: Work Done

## 4.1: Details as required.

## 4.2: Results and Discussion

## 4.3: Individual Contribution of project members (in case of group project)

# 5: Conclusion and Future Plan

# 6: References

Mohammadi, M., Rezazadeh, A., Siahkouhian, M., & Zamanifar, K. (2021). Evaluating the performance of intrusion detection systems using CSE-CIC-IDS2018 dataset. Journal of Information Security and Applications, 60, 102836. <https://doi.org/10.1016/j.jisa.2021.102836>

Almalki, M. H., Alshahrani, A. M., & Alamri, A. (2021). Evaluating the performance of intrusion detection systems based on machine learning algorithms using CSE-CIC-IDS2018 dataset. Journal of Ambient Intelligence and Humanized Computing, 12(9), 9985-9997.

Akinsanmi, O. O., Oyebisi, S. O., & Adebiyi, A. A. (2022). Performance evaluation of intrusion detection systems using CSE-CIC-IDS2018 dataset. Journal of King Saud University-Computer and Information Sciences, 34(2), 118-126. <https://doi.org/10.1016/j.jksuci.2021.03.003>