Real-Time Intrusion Detection System

A PROJECT REPORT

Submitted by

R Keerthana Prasad (1RVU23CSE363)

in partial fulfillment for the award of the degree

of

Bachelor of Technology (H) in Computer Science

School of Computer Science and Engineering
RV University
RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112

June, 2024

DECLARATION

I, **R Keerthana Prasad (1RVU23CSE363),** student third semester B.Tech in **Computer Science ,** at School of Computer Science and Engineering, **RV University,** hereby declare that the project work titled “Real Time Intrusion Detection System” has been carried out by us and submitted in partial fulfilment for the award of degree during the academic year **2023-2024**. Further, the matter presented in the project has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Name: R Keerthana Prasad Signature
USN: 1RVU23CSE363

Place: Bangalore

Date: 18/12/2024

School of Computer Science and Engineering

RV University
RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112

CERTIFICATE

This is to certify that the project work titled "Real Time Intrusion Detection System" is performed by R Keerthana Prasad **(1RVU23CSE363),** a bonafide students of Bachelor of Technology at the School of Computer Science and Engineering, RV university, Bangaluru in partial fulfilment for the award of degree Bachelor of Technology in Computer Science & Engineering, during the Academic year **2020-2021**.

Mr. Shivakumar D	**Dr. Sudhakar KN**	**Dr. G Shobha**
PhD Scholar SOCSE RV University Date: 18/12/2024	Head of the Department SOCSE RV University Date: 18/12/2024	Dean SOCSE RV University Date: 18/12/2024

Name of the Examiner Signature of Examiner

1.

2.

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to the School of Computer Science and Engineering, RV University, for providing us with a great opportunity to pursue our Bachelor's Degree in this institution.

In particular we would like to thank Dr. G. Shobha, Dean, School of Computer Science and Engineering, RV University, for his constant encouragement and expert advice.

It is a matter of immense pleasure to express our sincere thanks to Dr. Sudhakar K. N, Head of the department, Computer Science & Engineering University, for providing right academic guidance that made our task possible.

We would like to thank our guide Shivakumar D PhD, RV University, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.

Date: 18/12/2024

Place: Bangalore

R Keerthana Prasad

1RVU23CSE363

Section-B (3rd Sem)

TABLE OF CONTENTS

ABSTRACT

The field of cybersecurity faces constant challenges due to the evolution of cyber threats. Traditional systems that rely on pre-defined attack signatures or static rules are no longer sufficient to counteract the increasing sophistication of modern attacks. This project, titled "Real-Time Intrusion Detection System," leverages machine learning to enhance the detection and mitigation of potential cyber threats in network environments. By analysing network traffic patterns, the system aims to identify abnormalities and differentiate between normal and malicious activity in real time.

Unlike rule-based systems, this approach adapts dynamically to new forms of attacks by learning from historical data. This paper details the methodology of the project, including the use of datasets like NSL-KDD and supervised learning algorithms such as Support Vector Machines (SVM) and Random Forests. The system incorporates robust data preprocessing steps, feature selection techniques, and a thorough evaluation of multiple machine learning models. This project not only aims to achieve high detection accuracy but also ensures scalability and practical applicability in real-world scenarios.

The study emphasizes a comparative analysis of classifiers, presenting detailed insights into their performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. By providing a systematic approach to building and evaluating an IDS, the project contributes to the broader goal of strengthening cybersecurity defences in today’s digital age.

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol	Explanation
IDS	Intrusion Detection System
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

1. INTRODUCTION

1.1 Background

The rapid growth of digital technologies has significantly transformed industries, governments, and individual lives. However, this dependency on interconnected systems has also amplified the risk of cyber threats, which can lead to data breaches, financial losses, and reputational damage. An **Intrusion Detection System (IDS)** serves as a critical layer of defence in protecting networks by identifying and mitigating malicious activities.

The Need for IDS

Traditional IDS methods, such as signature-based systems, rely on predefined patterns of known attacks. While effective against previously encountered threats, these systems fail to address **zero-day attacks** or sophisticated forms of malware that exhibit unseen patterns. An advanced IDS must have the capability to learn and adapt to these emerging threats.

Objectives of the Project

The primary aim of this project is to design and implement a **real-time IDS** using machine learning techniques. This involves the following key objectives:

1. **Data Analysis:** Understand the characteristics of normal and malicious network traffic using benchmark datasets.
2. **Model Development:** Train and evaluate multiple machine learning models to identify anomalies and classify network traffic.
3. **Performance Optimization:** Apply preprocessing techniques and feature selection to enhance the accuracy and efficiency of the system.
4. **Scalability:** Ensure the system can handle high volumes of data without compromising on detection speed or accuracy.

1.2 Objectives

The IDS developed in this project focuses on binary classification—differentiating between normal and attack traffic. It leverages a combination of **supervised learning algorithms** and robust evaluation metrics to achieve high reliability. The project serves as a foundation for future enhancements, such as multi-class classification and the integration of deep learning models. By addressing these objectives, the project aims to contribute to the development of more resilient cybersecurity systems capable of countering modern threats.

2. Related work

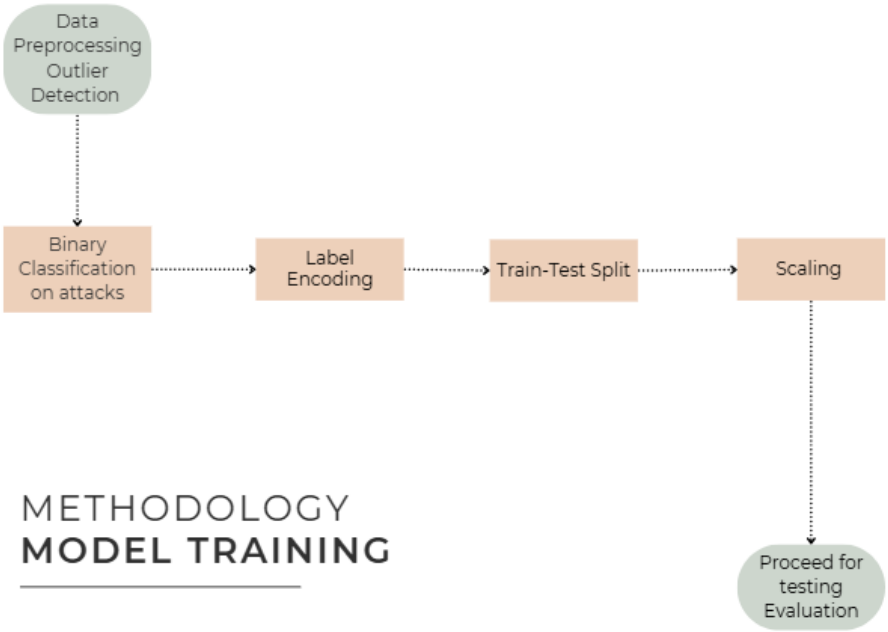
Several machine learning and deep learning-based approaches have been proposed, each improving on the performance and efficiency of IDS models. Passban IDS (2020) addresses the limitations of traditional IDS by using machine learning techniques such as iForest and LOF for anomaly detection in IoT networks. The key advantage is that it operates autonomously at the edge, reducing false alarms and improving real-time detection at IoT gateways, which previous systems struggled to achieve due to high false-positive rates. Deep Learning-Based IDS Review (2021) highlights advancements in deep learning models like CNNs and RNNs, which have become more effective in detecting complex attack patterns. Previous systems faced difficulties in handling large-scale datasets, but deep learning models now offer better scalability, reducing the need for manual updates. Supervised Machine Learning IDS (2021) compares different algorithms, with Random Forest emerging as the most accurate. The system emphasizes ensemble methods, which provide higher accuracy and reduce errors, making them more suitable for modern cybersecurity challenges.

VANET IDS (2021) uses Random Forest to detect DoS/DDoS attacks in vehicular networks, improving upon earlier IDS by offering real-time detection with low false positive rates. This innovation addresses the scalability and dynamic nature of VANETs, which previous systems failed to accommodate. IoT IDS Review (2021) outlines the advantages of hybrid IDS approaches, combining anomaly and signature-based techniques, which better cater to the resource-constrained nature of IoT devices. This approach overcomes earlier limitations where signature-based systems were too rigid, failing to detect new types of attacks. Wireless IDS (2020) presents a deep learning-based framework using CDBN and SCAE, achieving 97.4% accuracy, which surpasses older models that failed to address wireless network vulnerabilities effectively. RDP IDS for Flight Servers (2019) focuses on securing remote desktop connections, improving detection of malicious packets, and addressing vulnerabilities often overlooked in earlier systems. DoS Attack Prevention with Deep Learning (2022) integrates models like LSTM to enhance detection capabilities, preventing sophisticated threats. ML-Based IDPS (2020) uses MLP to detect complex attacks, including DoS, Probe, and R2L, with real-time prevention using iptables. End-to-End IDS Framework (2021) introduces AB-Trap, enhancing IDS performance in LAN environments by focusing on data management and adapting to evolving network threats.

These current implementations have made significant progress in addressing key issues like false positives, real-time detection, scalability, and handling sophisticated attack types. The next phase of research will focus on optimizing these models, improving real-time performance, and adapting IDS for emerging network environments like 5G and IoT, where threats continue to evolve rapidly.

3. METHODOLOGY

Real-time intrusion detection system (IDS) involves a systematic approach to data processing, feature transformation, model training, and evaluation. Outliers are identified and treated, where values above a predefined threshold are classified as cyberattacks, while those below are considered legitimate anomalies. The "attack" field is then converted into a binary classification, simplifying the task into detecting abnormal traffic versus normal traffic. Categorical features, such as protocol_type, service, and flag, are encoded into numerical values using Label Encoding, ensuring compatibility with machine learning algorithms. The dataset is split into training and testing subsets, and feature scaling is applied using StandardScaler to standardize the data, improving the efficiency of the machine learning models.
Next, Mutual Information is computed to identify the most relevant features for predicting the binary attack classification. This helps reduce dimensionality and ensures that only the most important features are used, enhancing model performance. Various machine learning models are trained on the processed data and evaluated based on key metrics such as accuracy and recall. The best-performing model is selected and fine-tuned to optimize its ability to detect potential intrusions. Through this methodology, the system is designed to effectively identify abnormal network traffic in real time, thereby providing an efficient and robust intrusion detection solution.



4. IMPLEMENTATION

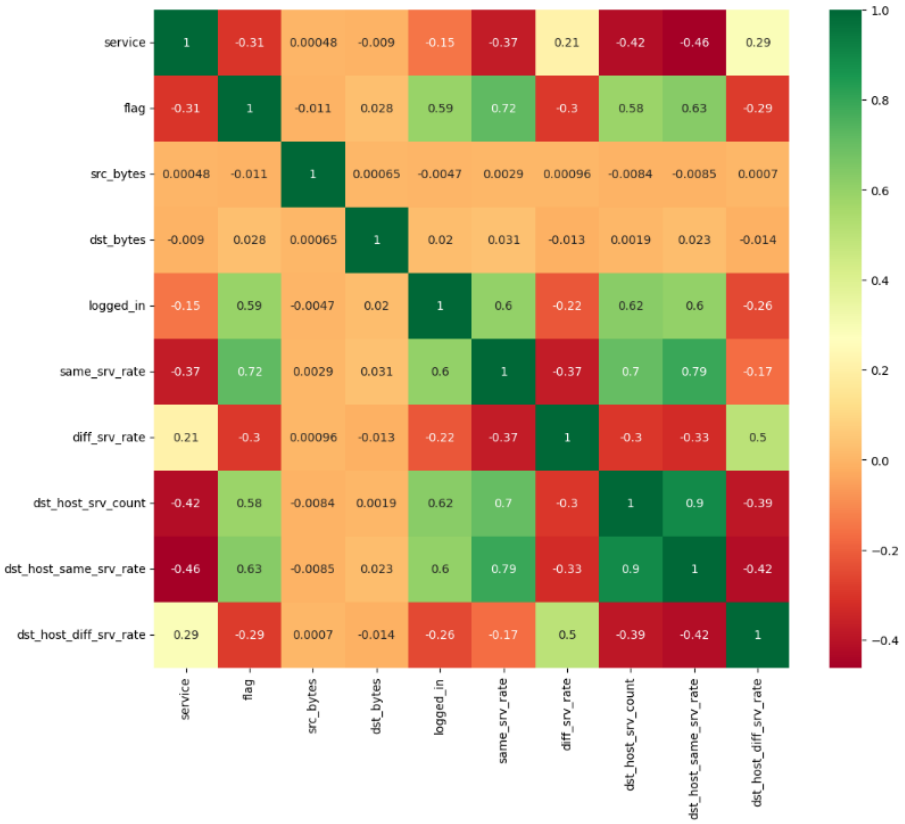
The implementation of the real-time intrusion detection system follows a systematic process of data preprocessing, feature engineering, and model training. Initially, the raw data is cleaned by detecting outliers through a predefined threshold. Outliers above this threshold are considered cyberattacks, while those below are deemed noise or legitimate anomalies. Following this, the "attack" column is transformed into a binary classification, separating data into "normal" and "abnormal" classes. This simplifies the classification task, allowing the model to focus solely on detecting deviations from normal traffic.

For feature transformation, categorical attributes such as protocol_type, service, and flag are encoded into numerical values using **LabelEncoder** from **scikit-learn**. This ensures that the machine learning models can process the data correctly. The data is then split into training and testing sets, preserving the integrity of the evaluation process. The dataset is carefully encoded and split in a way that prevents inconsistencies between the training and testing phases.

The model is trained using several algorithms, including Logistic Regression, SVM, Decision Trees, and Random Forest, among others. For effective model training, mutual information is computed for each feature to assess its importance in predicting the target variable. This step helps select the most relevant features, improving model efficiency and accuracy. Additionally, data scaling using **StandardScaler** ensures that the features are on a comparable scale, avoiding any bias toward features with larger numerical ranges.

Once the models are trained, they are evaluated on the test dataset to assess their performance. The SVM model with the Radial Basis Function (RBF) kernel provides the best results, achieving the highest accuracy, precision, and recall values. The final model's effectiveness is confirmed by its superior Receiver Operating Characteristic (ROC) curve, indicating its ability to balance sensitivity and specificity. The model is then fine-tuned to ensure optimal performance before being deployed for real-time threat detection.

5. RESULT AND DISCUSSION



The correlation analysis highlights the importance of selecting relevant features, with features such as flag and service having a strong positive correlation, thus improving the model's performance after proper scaling.

	Accuracy	Precision	Recall
Logistic Regression	0.718062	0.809490	0.718062
Support Vector Machines linear	0.724805	0.820728	0.724805
Support Vector Machines polynomial	0.767078	0.835162	0.767078
Support Vector Machines RBF	0.809617	0.851766	0.809617
Decision Trees	0.777990	0.836621	0.777990
Random Forest	0.777502	0.842784	0.777502
Naive Bayes	0.726845	0.831605	0.726845
K-Nearest Neighbor	0.793825	0.844492	0.793825

Based on the results, the Support Vector Machines RBF classifier achieved the highest performance with an accuracy of 0.8096, precision of 0.8518, and recall of 0.8096. This outperforms other models, such as Random Forest, which had an accuracy of 0.7775, precision of 0.8428, and recall of 0.7775, and K-Nearest Neighbor with an accuracy of 0.7938, precision of 0.8445, and recall of 0.7938. The Receiver Operating Characteristic (ROC) curve further supports this finding, as the SVM RBF model has the highest AUC value, demonstrating the best trade-off between sensitivity (true positive rate) and specificity (false positive rate).

6. CONCLUSION

The Real-Time Intrusion Detection System (IDS) developed in this project effectively integrates machine learning techniques to identify and respond to cyber threats in real-time. The model training phase involved preprocessing the dataset by checking for outliers, converting the target variable (attack) to binary classification (normal or abnormal), and performing label encoding. These preprocessing steps ensured that the data was ready for analysis and model training.
The models evaluated in this project included Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, Naive Bayes, and K-Nearest Neighbors. The system was trained and tested on a dataset that was split into training and testing sets, and the performance was measured using metrics such as accuracy, precision, recall, and ROC curves.

Key Results:

- **Support Vector Machines (RBF kernel)** emerged as the best model, achieving the highest **accuracy** of **80.96%**, **precision** of **85.18%**, and **recall** of **80.96%**.
- Other models such as **K-Nearest Neighbor** and **Random Forest** also performed well with accuracies of **79.38%** and **77.75%**, respectively.
- **Logistic Regression** and **Naive Bayes** had lower performance compared to other models, with accuracies of **71.81%** and **72.68%**, respectively.

ROC Curve analysis further confirmed the performance of the models, with **Support Vector Machines (RBF)** achieving the highest **Area Under the Curve (AUC)**.

These results demonstrate that the selected machine learning models are capable of detecting network intrusions effectively. The system's success in classifying normal and abnormal activity provides a robust foundation for real-time intrusion detection

7. FUTURE SCOPE

The future scope of this Real-Time Intrusion Detection System (IDS) involves several key areas of enhancement and development.

1. **Integration with Emerging Technologies (5G, IoT, Edge Computing):** As IoT devices and 5G networks proliferate, the IDS system can be further optimized to handle the unique challenges posed by these technologies. Future research could focus on creating models that can handle the massive data influx and speed of 5G networks, while also addressing the security concerns in IoT environments.
2. **Adaptation to Zero-Day Attacks:** Current IDS solutions often struggle to detect unknown, zero-day attacks. By incorporating techniques such as anomaly detection, behavioral analysis, and advanced machine learning models, your project could evolve to better identify novel attack vectors that haven't been encountered before.
3. **Deep Learning and Reinforcement Learning:** Incorporating advanced techniques such as deep reinforcement learning could allow your IDS to adapt dynamically to network traffic, learning and evolving from new attack patterns. This would enhance its ability to predict and prevent complex, sophisticated attacks that traditional IDS may miss.
4. **Distributed IDS Systems for Large-Scale Networks:** As networks scale, IDS models could evolve into distributed systems capable of monitoring multiple nodes across different network segments. This would provide a more holistic and integrated approach to network security, where different IDS units collaborate in real time to detect and respond to threats.

Firstly, the integration of more advanced machine learning models, such as deep learning algorithms, could improve the system's accuracy and its ability to detect novel types of attacks that were not previously recognized. Furthermore, expanding the dataset to include a wider variety of attack scenarios would make the system more robust and adaptable to new and emerging threats.

REFERENCES

1. Zhang, Y., Wang, S., and Ji, G. (2019) 'A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications', IEEE Access, 7, pp. 101-120.

2. Ahanger, A.S., Masoodi, F., and Khan, S.M. (2021) 'An Effective Intrusion Detection System using Supervised Machine Learning Techniques', *Proc. 5th Int. Conf. on Computing Methodologies and Communication (ICCMC)*.

3. Zang, M., and Yan, Y. (2021) 'Machine Learning-Based Intrusion Detection System for Big Data Analytics in VANET', *Proc. IEEE 93rd Vehicular Technology Conf. (VTC2021-Spring)*.

4. Buono, A., Nugroho, E.P., Sitanggang, I.S., et al. (2021) 'A Review of Intrusion Detection System in IoT with Machine Learning Approach: Current and Future Research', *Proc. 6th Int. Conf. on Science in Information Technology (ICSITech)*, pp. 260-265.

5. Li, J., Yin, L., Yang, L., et al. (2020) 'Real-Time Intrusion Detection in Wireless Network: A Deep Learning-Based Intelligent Mechanism', *IEEE Access*, 8.

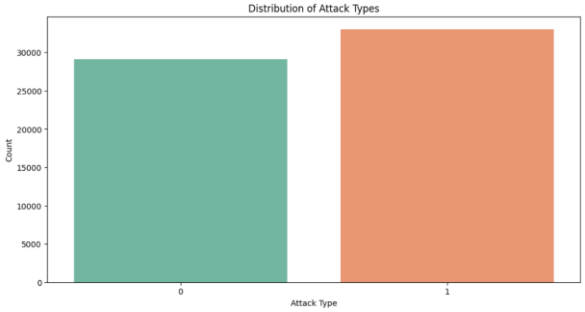
6. Shabtai, A., and Bitton, R. (2019) 'A Machine Learning-Based Intrusion Detection System for Securing Remote Desktop Connections to Electronic Flight Bag Servers', *IEEE Trans. on Dependable and Secure Computing*, 18.

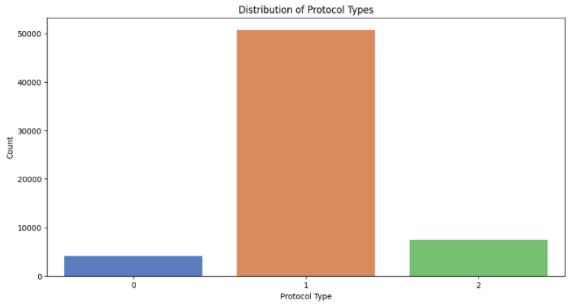
7. Taborda Blandon, G.E., and Cañola Garcia, J.F. (2022) 'A Deep Learning-Based Intrusion Detection and Prevention System for Detecting and Preventing Denial-of-Service Attacks', *IEEE Access*, 10.

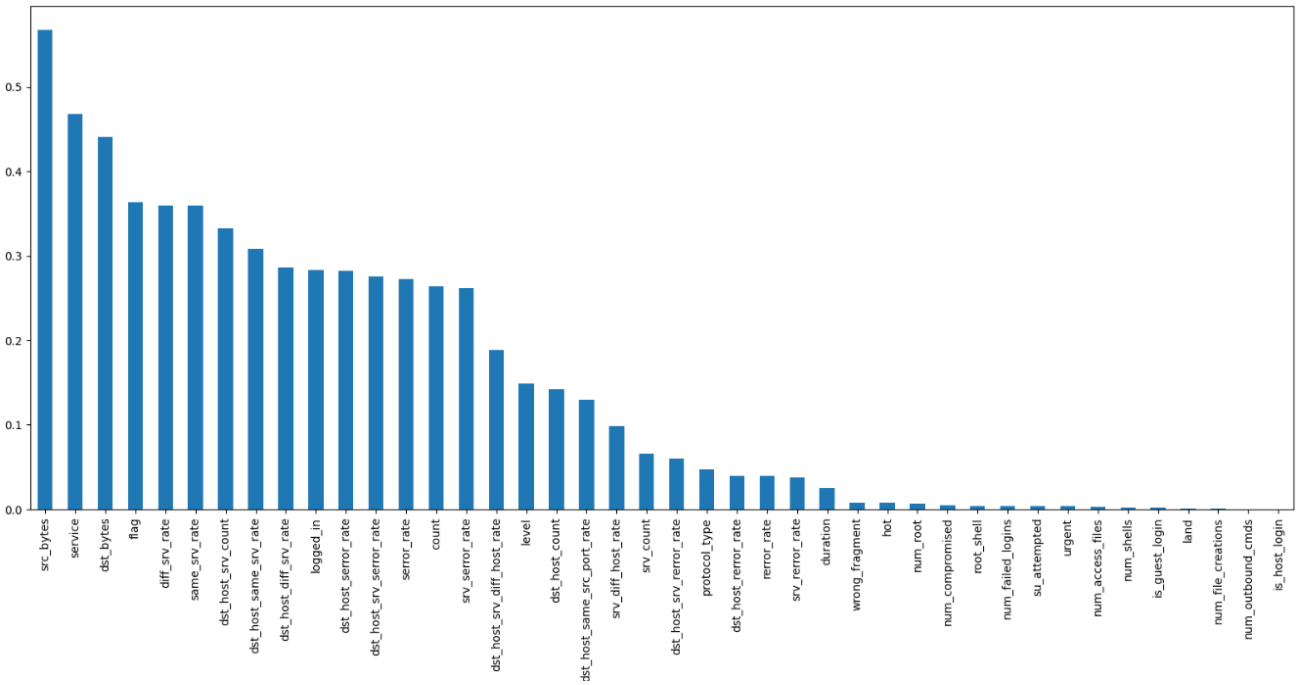
8. Krishna, A., Lal, A.M.A., Mathewkutty, A.J., et al. (2020) 'Intrusion Detection and Prevention System Using Deep Learning', *Proc. Int. Conf. on Electronics and Sustainable Communication Systems (ICESC)*.

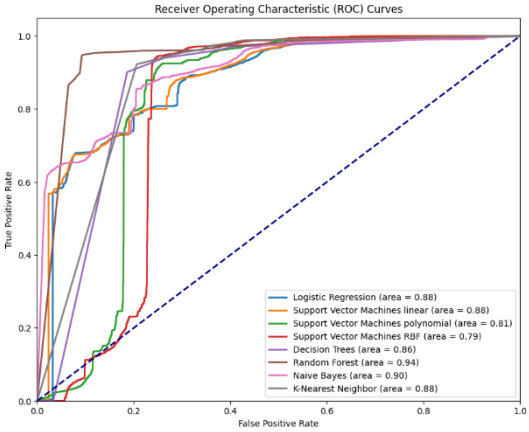
9. Lansky, J., Ali, S., Karim, S.H.T., et al. (2021) 'Deep-Learning Based Intrusion Detection Systems: A Systematic Review', *IEEE Access*, 9.

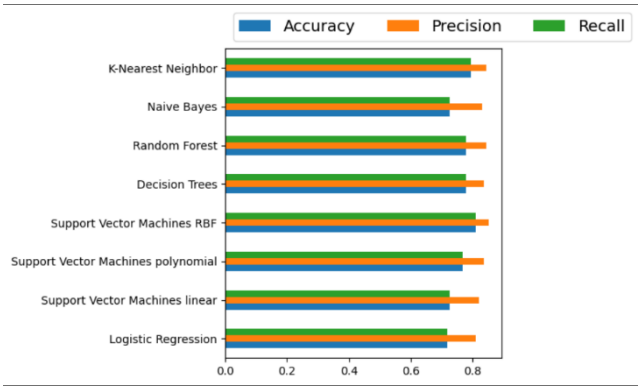
APPENDIX











Source code GitHub Link: https://github.com/Keerthana-prasad/Dhee-AIML-SEE-Cybersecurity-.git