#### INTRUSION DETECTION SYSTEM USING EXPLAINABLE AI (XAI)

**A Mini Project Report**

***Submitted by***

###### HRISHAB KAKOTY[202102022116]

###### SANGEETA SARKAR[202102021046]

###### PRABAL BAISHYA[202102022082]

***Under the supervision of***

**Mr. BIKRAMJIT CHOUDHURY**

(Assistant Dean Dept. Of CSE, CITK)

**B. Tech. 6th Semester**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

केन्द्रीय प्रौद्योगिकी संस्थान कोकराझार

**CENTRAL INSTITUTE OF TECHNOLOGY KOKRAJHAR**

**(Deemed to be University under MoE, Govt. of India)**

**May 2024**

**CERTIFICATE OF APPROVAL**

This is to certify that the mini project entitled “**Intrusion Detection System using Explainable AI**” submitted by **Hrishab Kakoty (202102022116), Prabal Baishya (202102022082), Sangeeta Sarkar (202102021046)** of B.Tech. 6th Semester to the department of Computer Science & Engineering, Central Institute of Technology Kokrajhar is a record of bonafide work carried out by them under our supervision and guidance.

#### (Signature) (Signature)

#### Guide Head of the Department

#### Name: Name:

**(**Designation) (Designation)

Dept. of Computer Science & Engineering Dept. of Computer Science & Engineering

##### Certificate by the Board of Examiners

This is to certify that the project work entitled “Intrusion Detection System using Explainable AI” submitted by Hrishab Kakoty (202102022116), Prabal Baishya (202102022082), Sangeeta Sarkar(202102021046), to the department of Computer Science & Engineering of Central Institute of Technology Kokrajhar has been examined and evaluated.

This project work has been prepared as per the regulations of Central Institute of Technology and has been qualified for acceptance.

#### (Signature) (Signature)

#### Project Coordinator Board of Examiners

#### Name: Name:

**(**Designation) (Designation)

Dept. of Computer Science & Engineering Dept. of Computer Science & Engineering

## **ABSTRACT**

Network intrusion detection systems (NIDS) play an important role in network security by detecting malicious or unauthorized access to computer networks. Traditional signature-based NIDS has limited ability to detect new or unknown attacks.   
  
This project presents a machine learning method for network intrusion detection using the UNSW-NB15 dataset containing network log information. (KNN), decision trees, and random forests are used to classify network connections as normal or malicious. First layer the data and select relevant features for model training. Evaluate the training model using metrics such as accuracy, data distribution, and fuzzy numbers.   
  
LIME creates a local interpretation of a prediction, showing support for classification decisions and their consequences. To decide.   
  
This project contributes to the development of a robust and flexible NIDS that can identify known and unknown attacks while providing insight into the decision-making process.

## 

## 

## i

## **ACKNOWLEDGEMENTS**

We would like to express our sincere gratitude to Mr. Bikramjit Choudhury for his invaluable guidance, support and encouragement throughout this project. Their expertise and insights helped shape the direction and successful completion of this work.

We also thank the Central Institute of Technology for providing the necessary resources and facilities to carry out this research.

We acknowledge the contributions of researchers and authors cited in this report. Finally, we appreciate the unwavering support from colleagues, friends, and family.

ii

# CONTENTS

1 [Introduction](#_bookmark8) 1

* 1. [Problem](#_bookmark9) Statement 2
  2. [Objectives](#_bookmark10) 2

1. [Literature Survey](#_bookmark11) 3
   1. Background details 3
   2. [Related Work](#_bookmark18) 4
   3. [Other](#_bookmark14) required section
      1. [Sub](#_bookmark15) section
2. [Proposed](#_bookmark19) System 5
   1. [System Architecture](#_bookmark21) 5
   2. [Flow Chart](#_bookmark23)/DFD/E-R Diagram 5
   3. [Hardware Components](#_bookmark28) 5
   4. [Software](#_bookmark33) Tools 5
   5. [Other](#_bookmark14) required section
      1. [Sub](#_bookmark15) section
3. [Implementation and Results](#_bookmark43) 6
   1. [Implementation](#_bookmark44) details 6
   2. [Results](#_bookmark47) 6
   3. Analysis………………………………………………………………………...6
   4. [Discussion](#_bookmark63) 6
4. [Conclusion and Scope](#_bookmark64) 7
   1. [Conclusion](#_bookmark65) 7
   2. [Future Scope](#_bookmark66) 7

References 8

iii

# Chapter 1

# Introduction

In recent years, IoT has been gaining popularity, and with the advancement of technologies, the internet, and big data, security has become essential for IoT networks. Researchers are seeking attention to the intrusion detection system for IoT networks for detecting malicious activities.

However, Signature-based IDS are proven to be inefficient in today’s scenario for two main reasons. First, it needs the predetermined knowledge of signatures or attacks and is thus incapable of detecting new or zero-day attacks.

Anomaly-based IDS, on the other hand, can detects abnormal behavior and is thus capable of detecting new or unknown attacks which are different from normal ones.

With recent development in Machine Learning (ML)/Deep Learning (DL) techniques, these techniques are employed in Anomaly-based IDS to remove the drawbacks. Anomaly-based detection using ML/DL techniques can detect intrusions with higher accuracy and is attracting many researchers for solutions in the direction of network security in IoT networks.

Given the continuous onslaught of attacks in IoT systems, there is a pressing need to identify and mitigate threats to safeguard the network. In IoT architecture, comprising perception, network, and application layers, protecting each layer is critical for overall system security.

Chapter 1. *Introduction* 2

### Problem Statement

Network intrusion detection systems (NIDS) play a crucial role in identifying malicious activities and protecting computer networks from cyber threats. However, traditional machine learning models used for NIDS often lack interpretability, making it challenging to understand the rationale behind their decisions. Explainable AI (XAI) techniques can provide insights into the decision-making process of these models, enabling better understanding and trust in their predictions.

### Objectives

The primary objective of this study is to analyze the performance of different machine learning models, including K-Nearest Neighbors (KNN), Decision Trees, and Random Forests, for network intrusion detection using the UNSW\_NB15 dataset. Additionally, we aim to leverage XAI techniques, specifically the LIME (Local Interpretable Model-Agnostic Explanations) framework, to explain the decisions of these models and gain insights into the important features contributing to the detection of network intrusions.

Chapter 2

# Literature Survey

### Background Details

Network intrusion detection system (NIDS) is a security tool designed to monitor network connections and detect threats or illegal activity. Traditional signature-based NIDS rely on predefined patterns or rules to detect attacks, making them ineffective against new attacks or zero-day attacks.

However, NIDS based on machine learning can learn from network data and adapt to discover new attack patterns in Network system.

They play a very important part in ensuring the security of networks by raising awareness of probable danger and cautioning managers who may then intervene promptly at minimal expense.

Chapter 2. *Literature Survey*

2

### 

### Related Work

Researchers have explored various XAI techniques to interpret and explain the decisions of machine learning models in the cybersecurity domain. For example, Shrikumar et al. (2019) used LIME to explain the predictions of deep learning models for malware classification. Feng et al. (2019) applied SHAP (SHapley Additive exPlanations) to interpret the decisions of random forest models for network anomaly detection.

**2.3 Other Relevant Sections**

Within the annals of literature, myriad perspectives have been explored, ranging from the amalgamation of Random Forests and Support Vector Machines (SVMs) to the pioneering forays into deep learning architectures like CNNs and Long Short-Term Memory (LSTM) networks. Each endeavour contributes a distinctive facet to the tapestry of knowledge, thereby enriching our understanding and paving the way for future innovations.

# Chapter 3

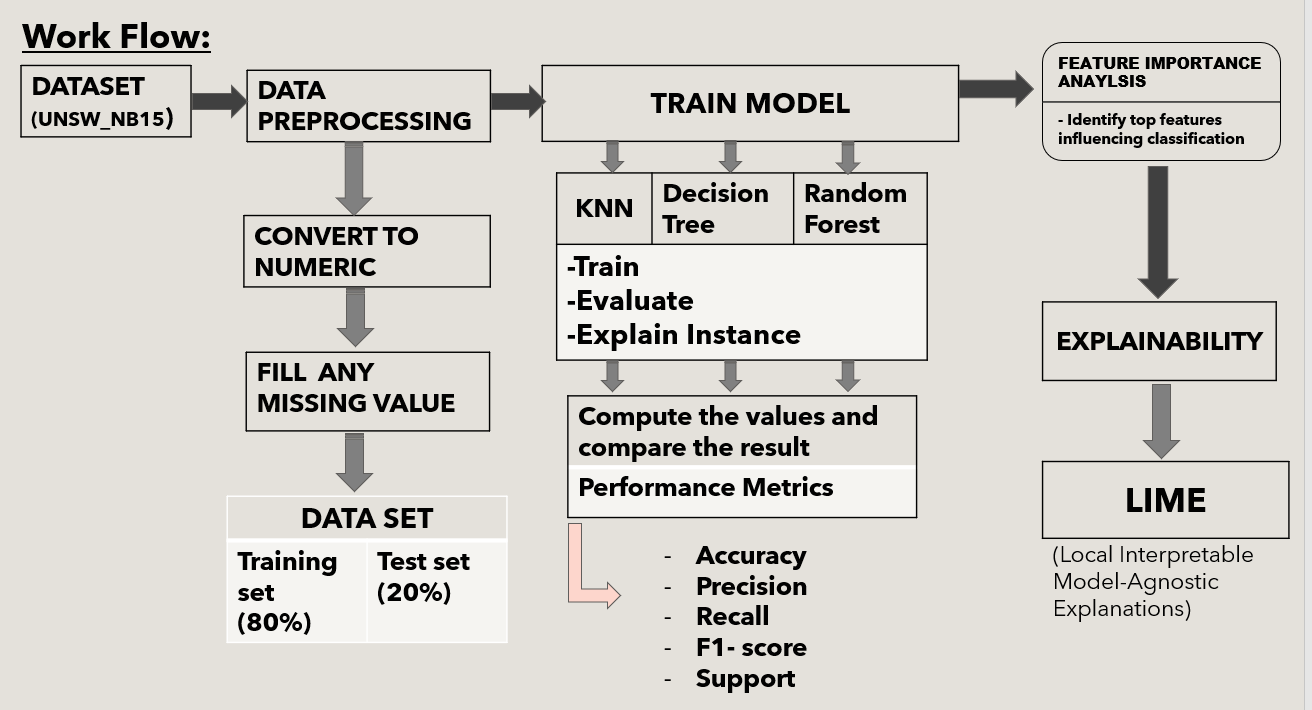
# Proposed System

### System Architecture

The planning process has the following important elements:

* + 1. Data Preprocessing: This component is responsible for loading the UNSW-NB15 dataset, converting the data into numbers and filling in missing values ​​with average values.
    2. Feature Selection: Select relevant features from the dataset based on domain knowledge or key criteria.
    3. Training Model: Machine learning models such as KNN, decision trees, and random forests are trained on previous data.
    4. Model Evaluation: Evaluate the training model using metrics such as accuracy, distribution distribution, and confusion matrices.
    5. LIME Explainations: LIME is used to create local annotations for the prediction of the training model and to help understand the factors that influence classification.

### Flow Chart/ Data Flow Diagram/E-R Diagram:

****

### Hardware Components

The system can be implemented on a standard computer or server with sufficient RAM and processing power to handle the computational requirements of the machine learning algorithms and data processing tasks. The specific hardware requirements will depend on the size of the dataset, the complexity of the models, and the desired performance.

### Software tools

The code is written in Python and utilizes the following libraries and frameworks:

* NumPy: For numerical computations
* Pandas: For data manipulation and analysis
* Scikit-learn: For machine learning algorithms and evaluation metrics
* LIME: For generating local interpretable model-agnostic explanations

**3.5 Additional Components**

In addition to the core components delineated above, our system embodies a spirit of innovation and adaptability, thereby accommodating a plethora of auxiliary elements aimed at enhancing its efficacy and resilience. These may include but are not limited to real-time data streaming capabilities, integration with existing security infrastructure, and mechanisms for continuous monitoring and adaptation in response to evolving threat landscapes.

# Chapter 4

**Implementation and Results**

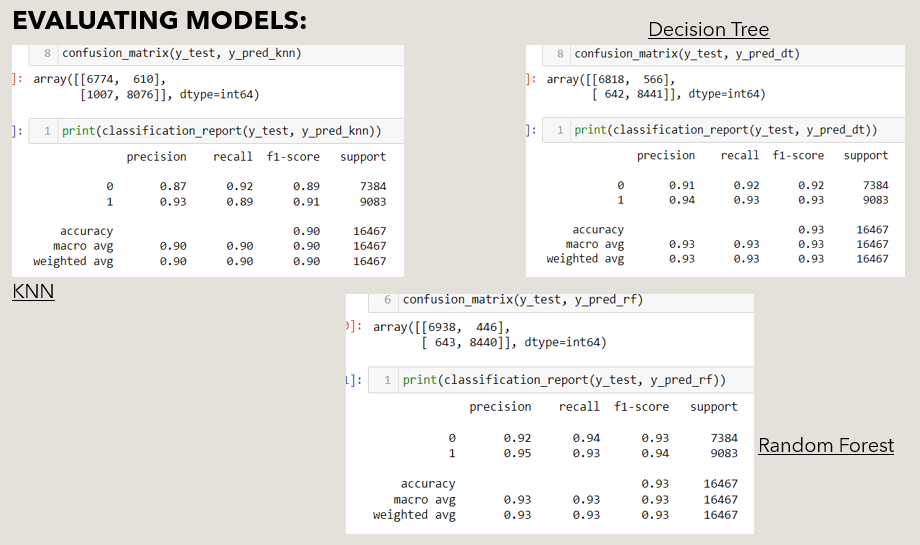
### Implementation Details

The implementation follows these main steps:

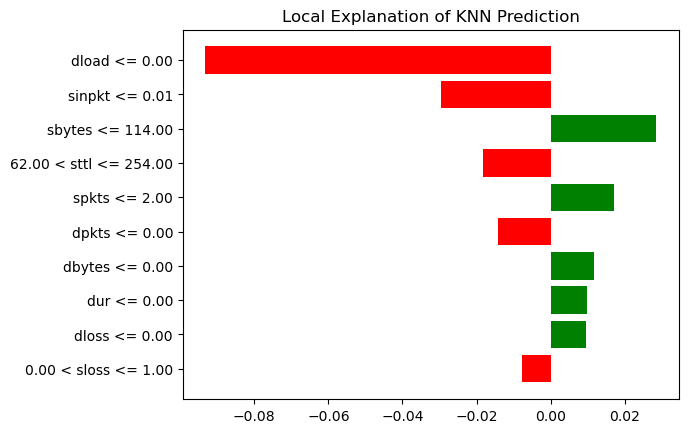
1. Load and preprocess the UNSW-NB15 dataset.
2. Split the data into training and testing sets.
3. Train and evaluate KNN, Decision Tree, and Random Forest models on the selected features.
4. Important feature selection.
5. Generate LIME explanations for the predictions made by each model on a test instance.
6. Visualize and analyze the LIME explanations.

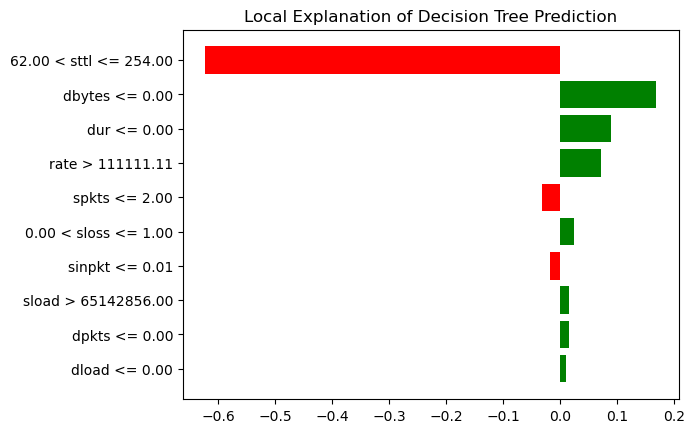
### Results

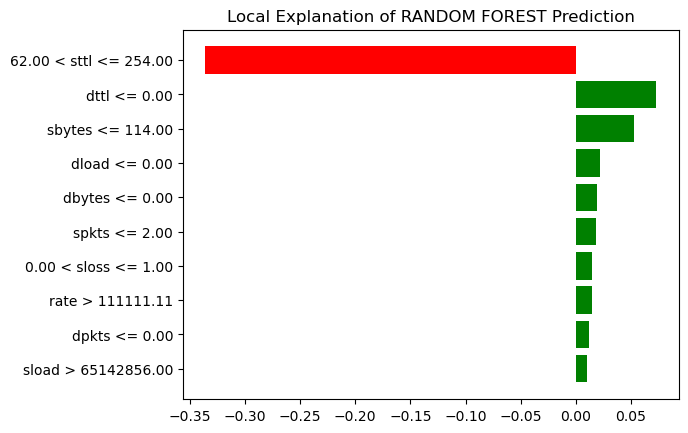
**Evaluation results:**



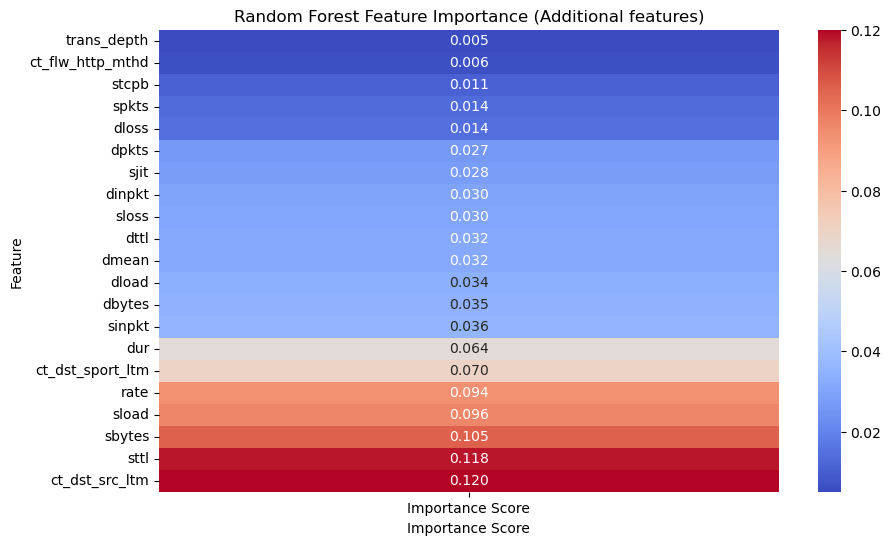
LIME EXPLAINATION:



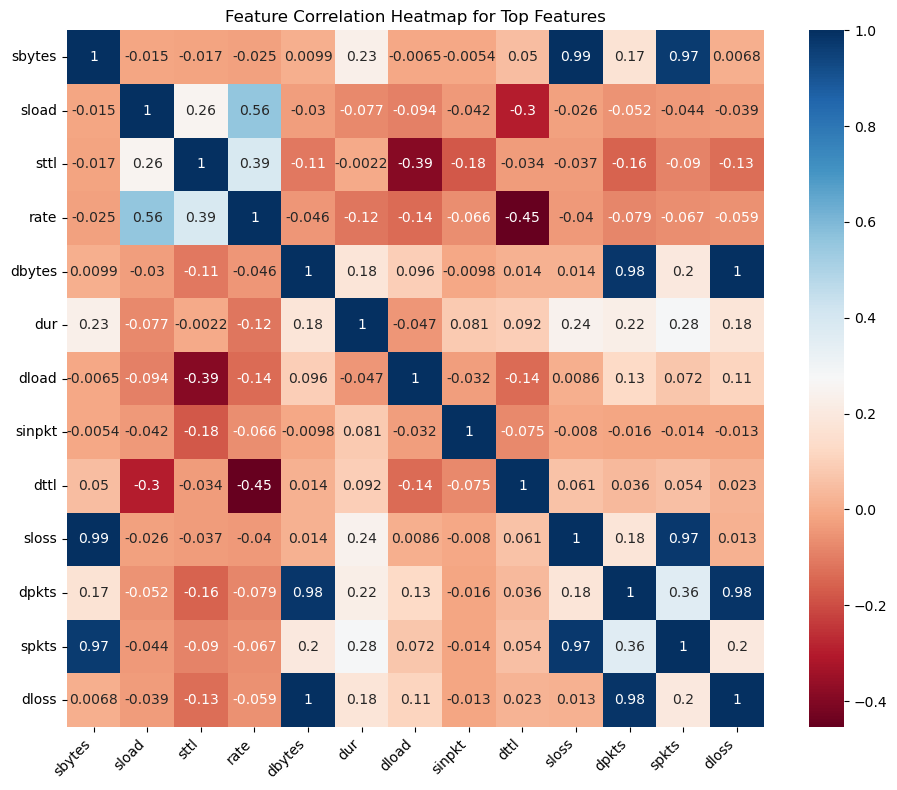




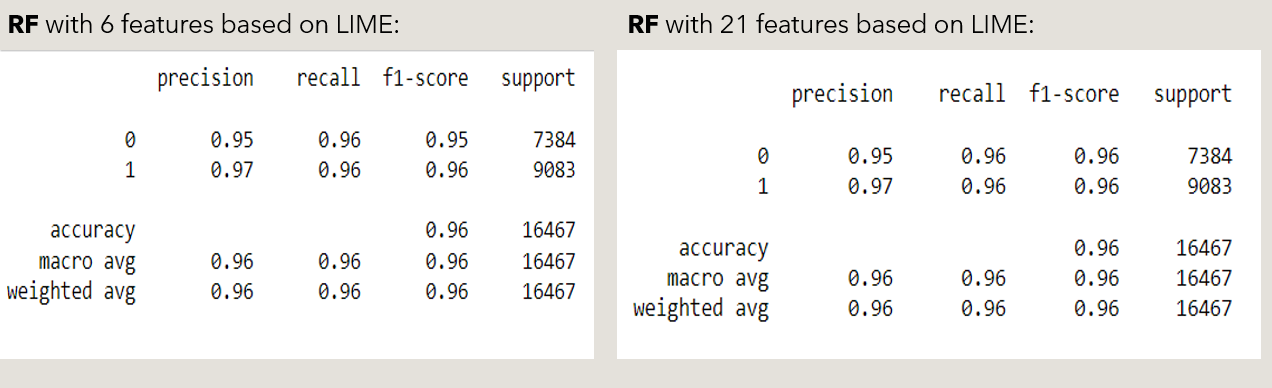
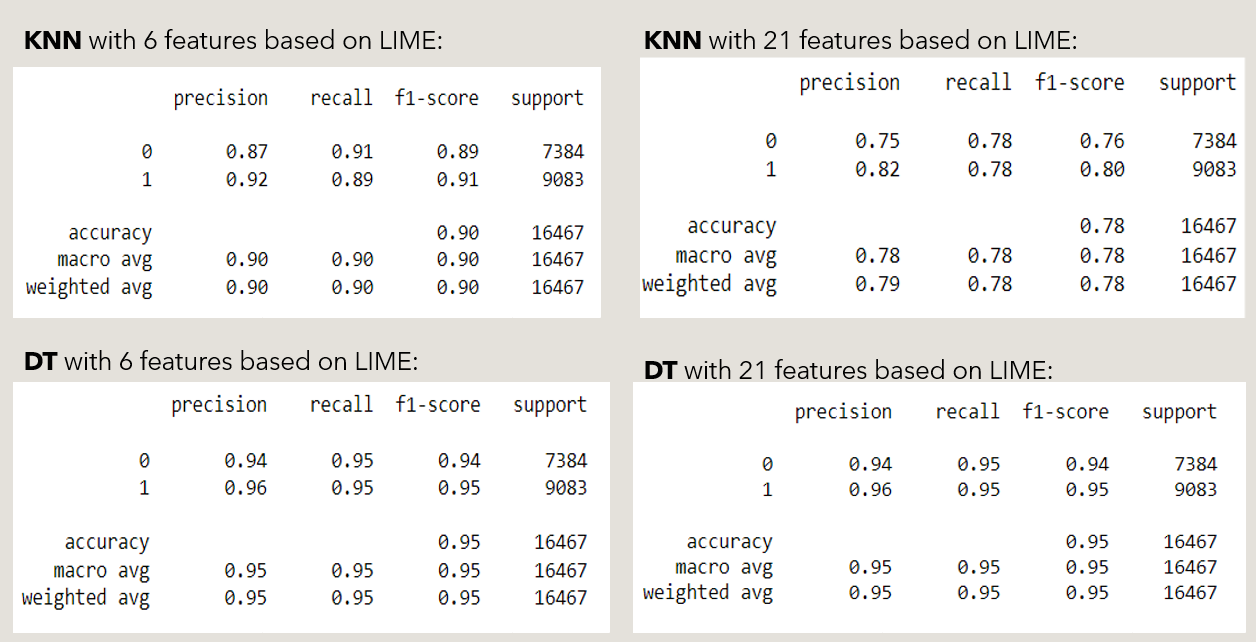
Heatmap of Random Forest:



Feature correlations:



By obtaining the main features:



### Analysis

### A critical analysis of our system's performance metrics unveils a nuanced tapestry of strengths and weaknesses, shedding light on areas ripe for refinement and avenues primed for further exploration. Through the prism of performance analysis, we glean invaluable insights into the efficacy of our approach and chart a course towards continuous improvement and innovation.

### Discussion

The dialogue sparked by our results transcends the realm of mere performance metrics, delving deep into the murky depths of interpretability and explanation. Through the lens of interpretability, we unravel the enigmatic decisions of our machine learning models, illuminating the underlying factors driving classification outcomes and empowering stakeholders with actionable insights to fortify their defenses against the ever-present specter of cyber threats.

# 

# Chapter 5

**Conclusion and Scope**

### Conclusion

In summation, our journey through the labyrinthine landscape of intrusion detection has yielded a bounty of insights and revelations, culminating in the realization of a robust and adaptable NIDS imbued with the power of machine learning and the clarity of interpretability.

### Future Scope

Outline potential areas for future work, such as:

* Exploring other machine learning algorithms or ensemble methods for intrusion detection.
* Incorporating additional features or feature engineering techniques to improve model performance.
* Investigating the application of deep learning models for network intrusion detection.
* Deploying the system in real-world scenarios and conducting further evaluations.

# References

[1] Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach *www.elsevier.com/locate/eswa*

[2] A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks *www.elsevier.com/locate/eswa*