5G-NIDD - A comprehensive network Intrusion detection dataset generated over 5G Wireless network.

Under the Guidance of Dr. Abdallah Moubayed
Team: F- 33
Adit Sandeep Virkar
Arnav Raviraj
Shivanjay Vilas Wagh
Vinay Kantilal Chavhan

Sprint Plan

Sprint 1: Conduct comprehensive literature review, categorize studies, highlight critical theories, discuss future scope.

Sprint 2: Find a suitable model and feed the encoded dataset to it for evaluation and optimization using the binary classification label.

Sprint 3: Improve boosting model performance, explore feature selection, implement ANN and GAN models.

Sprint 4: Implement input sampling, boosting models, GAN-based models, and ANN models for multiclass classification label.

Progress Report

Overview

- As of sprint 4, team is working on advancing the research and apply the multi classification models.
- As of now, We have applied three types of models. Boosting Algorithms, GAN Based, and ANN.
- Used different data processing and sampling techniques to increase the model performance.
- Made comparable progress on existing research.
- Github: https://github.com/Araviraj8/SER_517_F_33

Progress Report

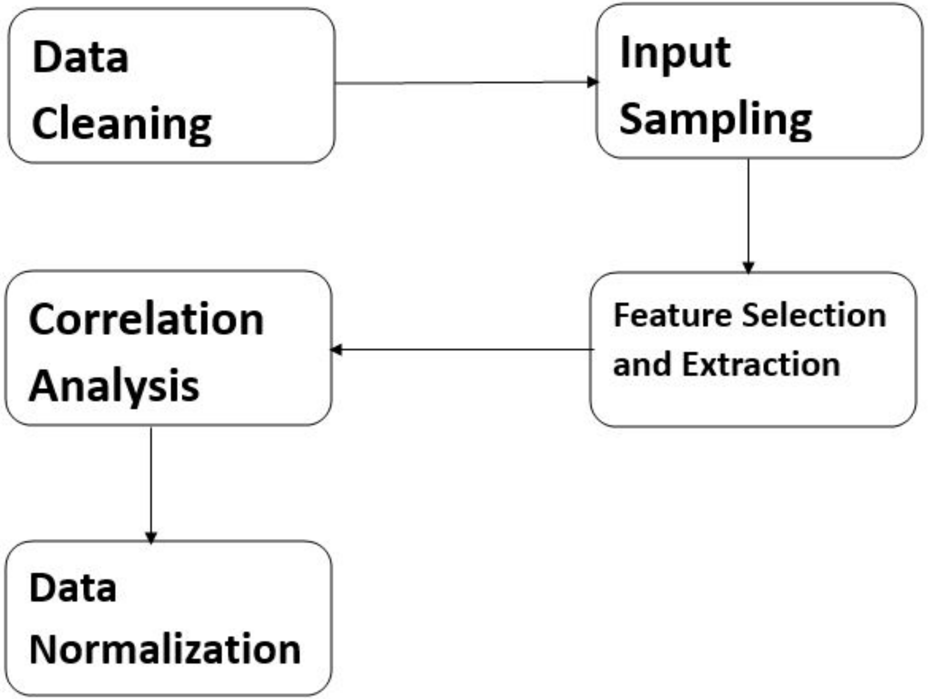
Problems Encountered:

- Challenges such exploring the new models which are suitable for the project.
- Managing large dataset and execution of the complex models

Solutions:

- Categorized the model search and explored it
- We split the dataset using the sampling techniques

Data Processing



Data Cleaning

Input Sampling

Feature Selection and Extraction

Correlation Analysis

Data Normalization

1. Data Cleaning: Removal of Zero columns.
2. Input Sampling: Balance class distribution.
3. Feature Selection & Extraction: Random Forest.
4. Correlation Analysis: Pearson & Spearman.
5. Data Normalization: Min-Max Scaling.

Models

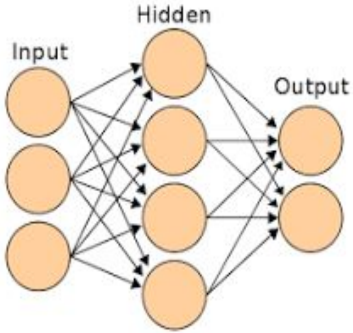
1. Boosting models
   XGBoost: A fast and accurate boosting algorithm for large datasets, known for its reliability in predictive modeling.

   AdaBoost: Trains using the AdaBoostClassifier,ability to sequentially train weak learners which focuses more on misclassified instances.

   CatBoost: Excels in handling categorical variables without preprocessing, making it ideal for classification tasks, especially with tabular data.

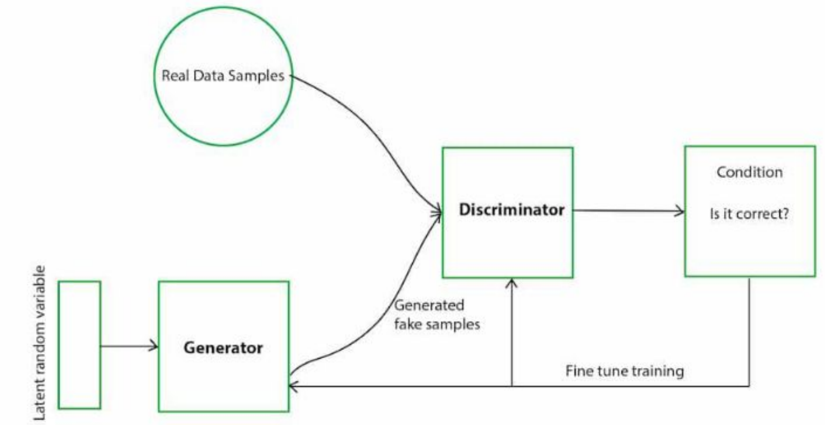
2. ANN

An ANN using TensorFlow and Keras is set up for binary classification, processing data by fixing missing values, normalizing, and encoding labels. The model has four layers, uses ReLU for hidden layers and sigmoid for output, trained with SGD optimizer, 10 epochs, and 10% validation split.

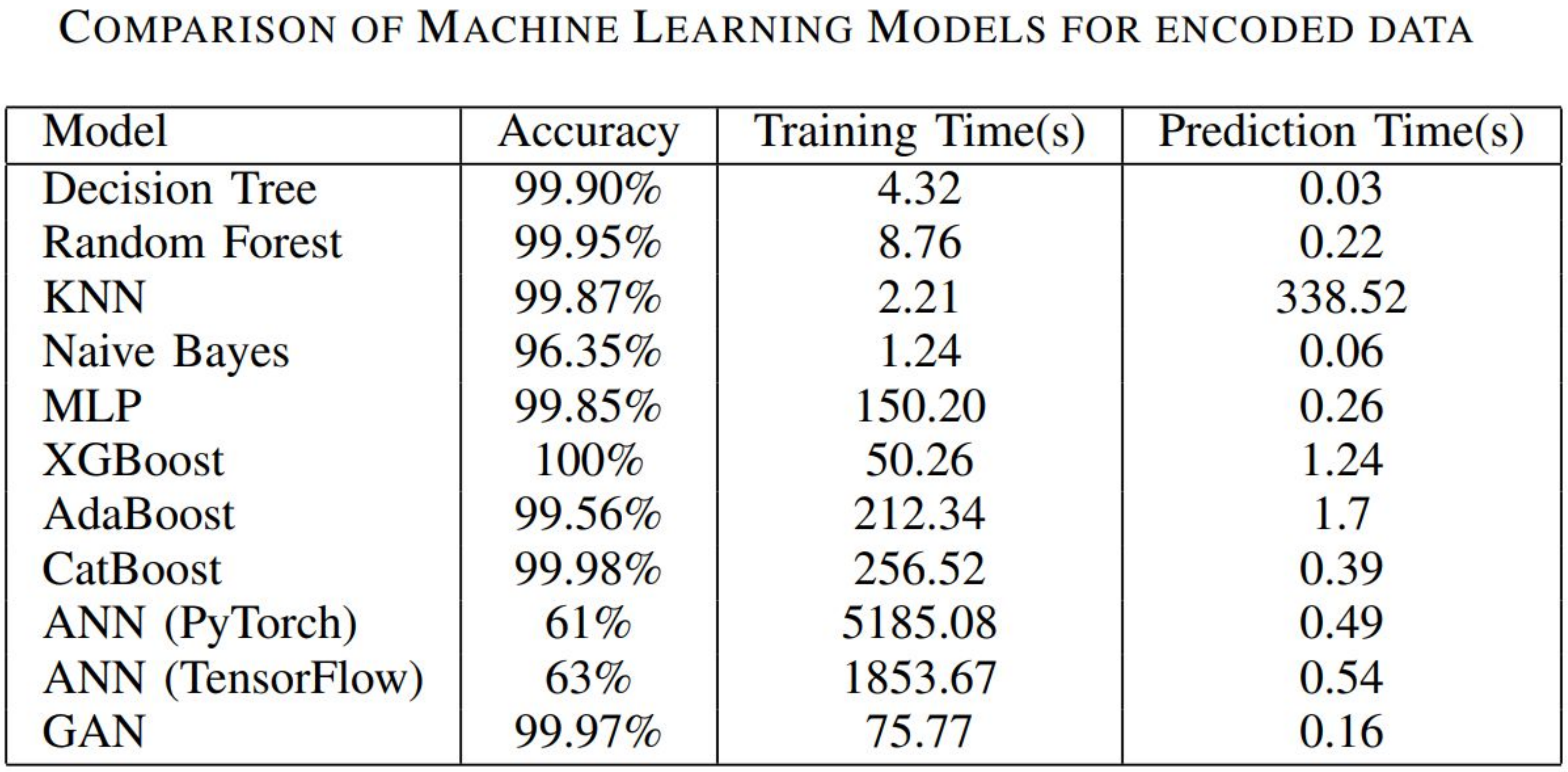


3. GAN Based Network

- GANs are made up of two neural networks, a generator and a discriminator.
- The Generator attempts to fool the Discriminator, which is tasked with accurately distinguishing between produced and genuine data, by producing random noise samples.
- Types of GANs
  - Vanilla GAN
  - Conditional GAN (CGAN)
  - Deep Convolutional GAN (DCGAN)
  - Laplacian Pyramid GAN (LAPGAN)
  - Super Resolution GAN (SRGAN)



Demo



Comparison of Machine Learning Models for encoded data

Model	Accuracy	Training Time(s)	Prediction Time(s)
Decision Tree	99.90%	4.32	0.03
Random Forest	99.95%	8.76	0.22
KNN	99.87%	2.21	338.52
Naive Bayes	96.35%	1.24	0.06
MLP	99.85%	150.20	0.26
XGBoost	100%	50.26	1.24
AdaBoost	99.56%	212.34	1.7
CatBoost	99.98%	256.52	0.39
ANN (PyTorch)	61%	5185.08	0.49
ANN (TensorFlow)	63%	1853.67	0.54
GAN	99.97%	75.77	0.16

FUTURE PLANS

1. Exploring new architectures like transformers.

2. Working on making the algorithm robust from adversarial attacks.

3. Increasing inference time by utilizing methods like quantization.

4. Reducing model size and complexity.

5. Exploring ensembling techniques with multiple models.

THANK YOU