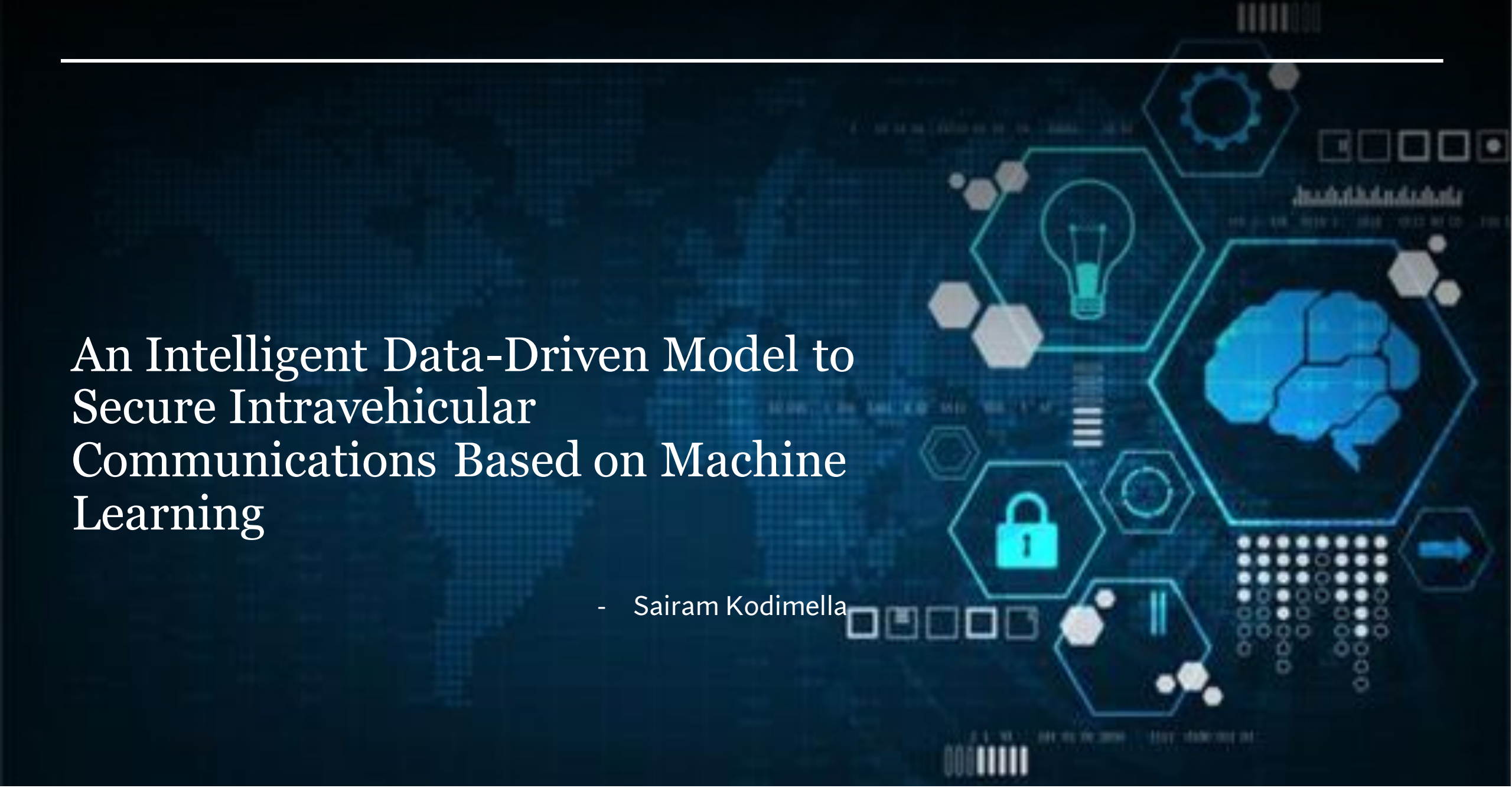
An Intelligent Data-Driven Model to Secure Intravehicular Communications Based on Machine Learning



- Sairam Kodimella

Name :- Sairam Kodimella

Email:- sk31349n@pace.edu

GitHub:- https://github.com/sairam-kodimella/capestone-project.git

Research Question

How effective is the implementation of machine learning algorithms in preventing malicious intruder attacks on the communication between sensors and Electronic Control Units (ECU) in electric vehicles, as compared to conventional methods such as AES encryption and CONTROLLER AREA NETWORK (CAN) protocol?

Motivation

As the electric vehicle is growing on popularity and usage and hacking electric vehicle for hackers benefit might cost lives. We need to build reliable technology which makes it difficult for hackers to interrupt the intra-vehicular communication and save lives.

Dataset

Label - The column 'Label' indicates whether the data row is to be considered as normal (Label=0) or as intrusion (Label=1).

ID - The column 'ID' contains the identifiers for the IDs that are 'id1', ..., 'id10'.

Time - In the column 'Time' the time stamp of the current message is represented in milliseconds.

Signal1_of_ID
Signal2_of_ID
Signal3_of_ID
Signal4_of_ID

} The columns contain the actual signal values

Out[11]:

	Label	Time	ID	Signal1_of_ID	Signal2_of_ID	Signal3_of_ID	Signal4_of_ID
0	0	8.100845e+07	id3	0.200000	1.000000	NaN	NaN
1	0	8.100846e+07	id9	0.370003	NaN	NaN	NaN
2	0	8.100846e+07	id7	0.044139	0.000000	NaN	NaN
3	0	8.100846e+07	id8	0.170534	NaN	NaN	NaN
4	0	8.100846e+07	id5	0.173044	0.874886	NaN	NaN
...	...	...	...	...	...	...	...
2575798	0	8.550892e+07	id10	0.336289	0.555556	0.913794	0.352531
2575799	0	8.550892e+07	id8	0.215258	NaN	NaN	NaN
2575800	0	8.550892e+07	id4	0.233920	NaN	NaN	NaN
2575801	0	8.550892e+07	id3	0.800000	1.000000	NaN	NaN
2575802	0	8.550892e+07	id2	0.000000	0.626573	0.210451	NaN

2575803 rows × 7 columns

Literature Review Summary

The reviewed literature presents various approaches to detecting intrusions in vehicular networks.

The proposed project can benefit from the techniques and approaches presented in the reviewed literature, such as machine learning-based intrusion detection systems and anomaly-based intrusion detection systems. The project can also explore the use of deep learning techniques and identity-based encryption schemes to improve the detection accuracy and security of the proposed model. The project can evaluate the proposed model's performance and compare it with existing intrusion detection systems in terms of accuracy, speed, and reliability, using metrics such as accuracy, AUC, IOU, specificity, and others.

It aims to develop an intelligent model using machine learning algorithms to analyze and monitor intravehicle communications in real-time, identify potential security vulnerabilities and threats, and develop appropriate countermeasures to prevent attacks. The success of the project will depend on the accuracy, efficiency, and robustness of the model compared to conventional methods.

The project's success criteria include the ability of the model to more accurately, efficiently, and robustly avoid hostile attacks on communication between sensors and Electronic Control Units in electric vehicles than conventional methods can.

Preliminary preprocessing and EDA

Out[11]:

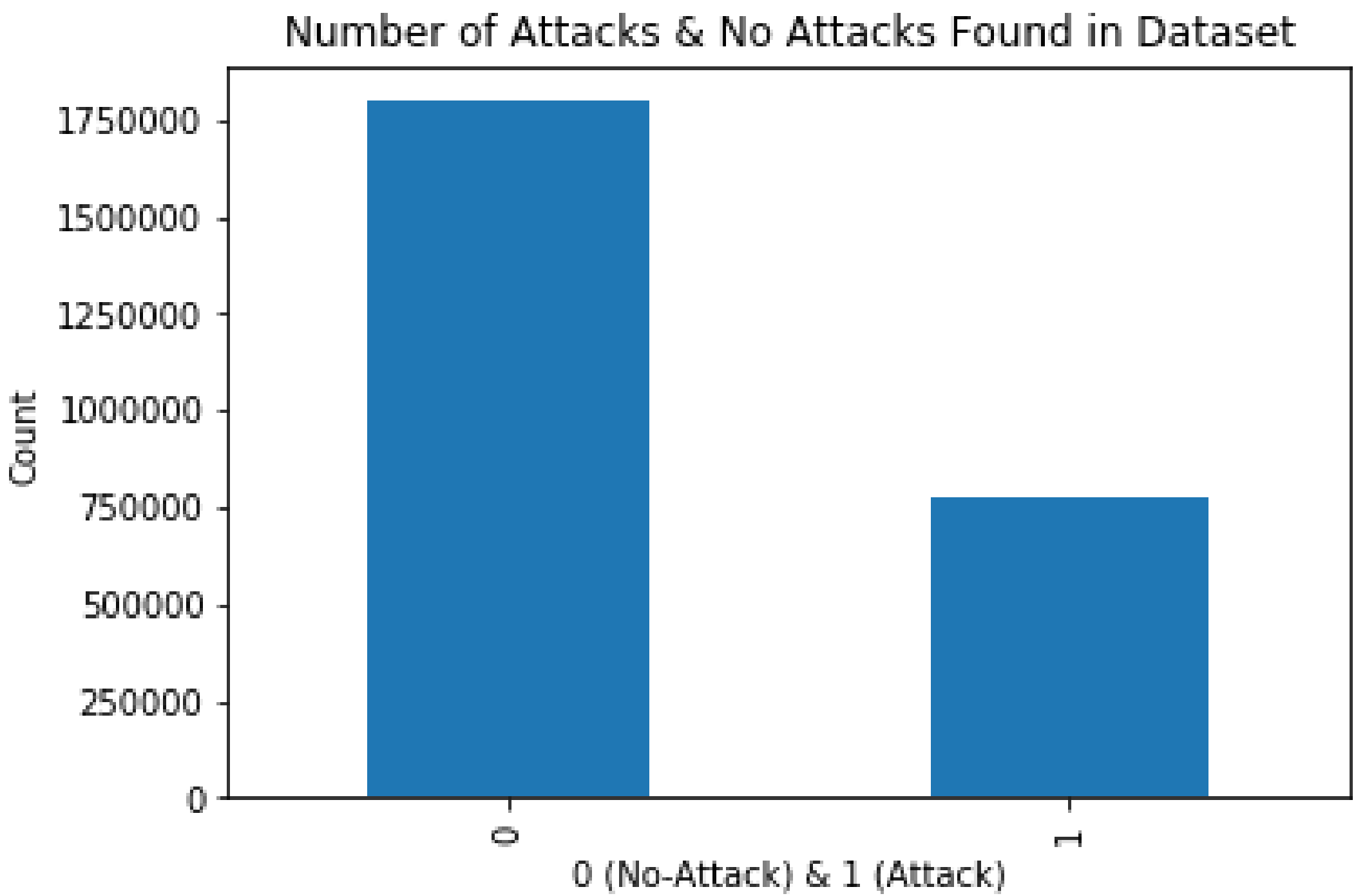
	Label	Time	ID	Signal1_of_ID	Signal2_of_ID	Signal3_of_ID	Signal4_of_ID
0	0	8.100845e+07	id3	0.200000	1.000000	NaN	NaN
1	0	8.100846e+07	id9	0.370003	NaN	NaN	NaN
2	0	8.100846e+07	id7	0.044139	0.000000	NaN	NaN
3	0	8.100846e+07	id8	0.170534	NaN	NaN	NaN
4	0	8.100846e+07	id5	0.173044	0.874886	NaN	NaN
...	...	...	...	...	...	...	...
2575798	0	8.550892e+07	id10	0.336289	0.555556	0.913794	0.352531
2575799	0	8.550892e+07	id8	0.215258	NaN	NaN	NaN
2575800	0	8.550892e+07	id4	0.233920	NaN	NaN	NaN
2575801	0	8.550892e+07	id3	0.800000	1.000000	NaN	NaN
2575802	0	8.550892e+07	id2	0.000000	0.626573	0.210451	NaN

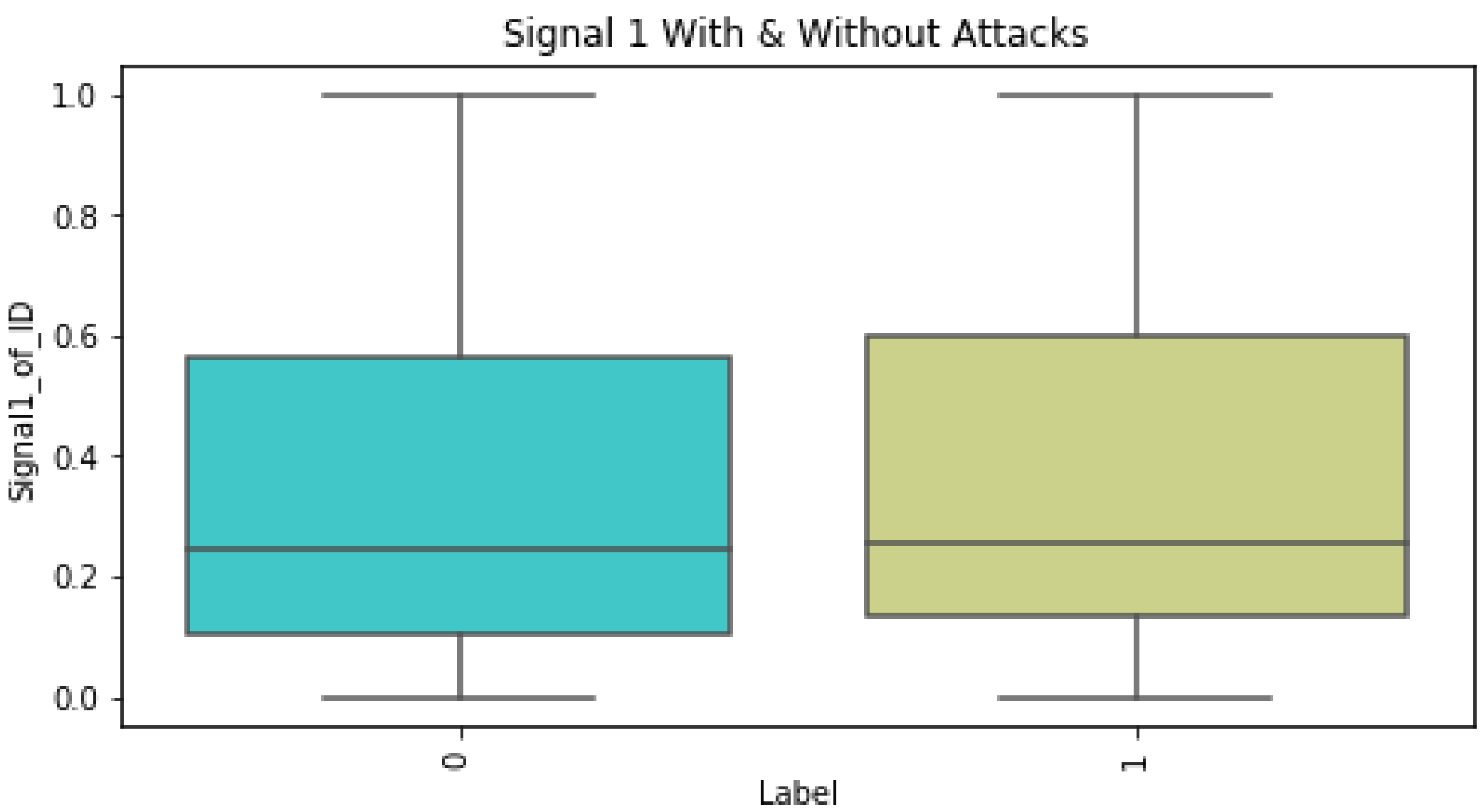
2575803 rows × 7 columns

Out[12]:

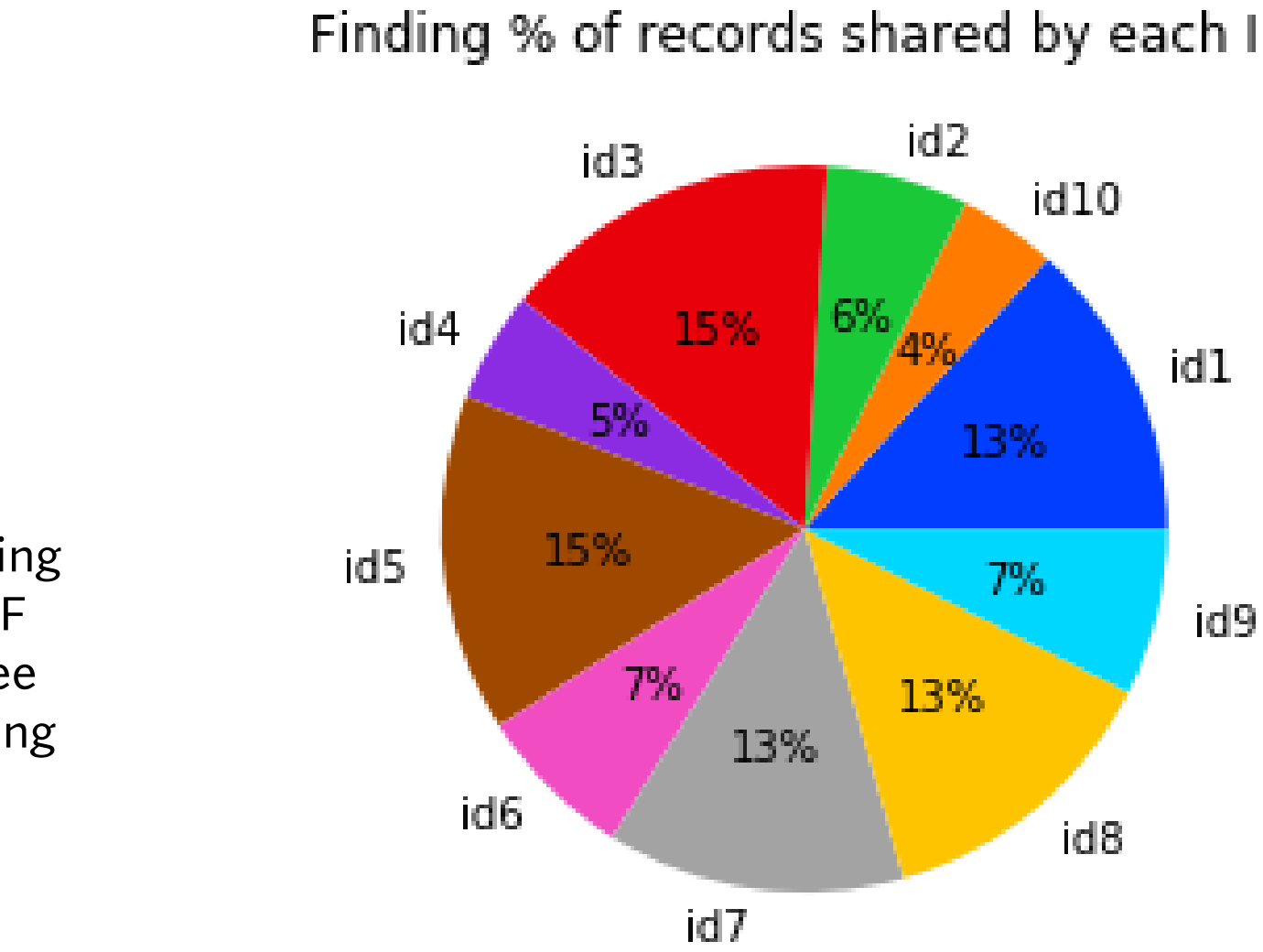
	Label	Time	Signal1_of_ID	Signal2_of_ID	Signal3_of_ID	Signal4_of_ID
count	2.575803e+06	2.575803e+06	2.575803e+06	1.918455e+06	275736.000000	112921.000000
mean	3.025169e-01	8.325403e+07	3.393861e-01	6.495193e-01	0.386199	0.387427
std	4.593479e-01	1.299185e+06	2.932694e-01	3.339861e-01	0.302042	0.066162
min	0.000000e+00	8.100845e+07	0.000000e+00	0.000000e+00	0.000000	0.222089
25%	0.000000e+00	8.212913e+07	1.172740e-01	4.025682e-01	0.194169	0.346848
50%	0.000000e+00	8.324363e+07	2.529841e-01	7.500000e-01	0.240539	0.384280
75%	1.000000e+00	8.438190e+07	5.789474e-01	9.993689e-01	0.555807	0.423877
max	1.000000e+00	8.550892e+07	1.000000e+00	1.000000e+00	1.000000	0.747278

In above screen we are applying data exploration technique to find MIN, MAX, count and other statistics from dataset


Number of Attacks & No Attacks Found in Dataset
1750000
1500000
1250000
1000000
750000
500000
250000
0
Count
0
1
0 (No-Attack) & 1 (Attack)


Signal 1 With & Without Attacks
Signal1_of_ID
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Label

In above graph we are processing dataset to finding NUMBER OF MISSING values and we can see signals contains so many missing values



Methodology

The methodology can be divided into the following steps:

- **Loading Dataset:** The code loads the dataset from the given file path and displays the values.
- **Dataset Exploration**: The code explores the dataset by describing each column in terms of count, mean, standard deviation, etc., and finding the count of different signals on different IDs.
- **Dataset Visualization**: The code performs data visualization by finding and plotting graphs of attacks from the dataset and signal graphs with and without attacks.
- **Dataset Preprocessing**: The code preprocesses the dataset by converting non-numeric data into numeric values using Label Encoder and replacing missing values with mean.
- **Model Building**: The code then builds three different models for classification, including SVM, KNN, and Decision Tree, using genetic selection for feature selection. It splits the dataset into training and testing sets and trains the models on the training set. The accuracy, precision, and confusion matrix are then calculated on the testing set.
- **Model Evaluation**: Finally, the code evaluates the models based on their accuracy, precision, and confusion matrix and selects the best-performing model.

Dataset Loading

Dataset Exploration

Dataset Visualization

Dataset Preprocessing

Model Building

EVALUATION

Model Evaluation


Experiments

The methodology described above can be justified as it follows standard data preprocessing and model building steps for solving a cybersecurity problem. The EDA and data visualization steps help in understanding the dataset and identifying any patterns or anomalies present in the data. The data preprocessing steps ensure that the dataset is ready for the model building process and that the models can work efficiently. The model building process involves selecting the best features for classification and training different models on the training set. Finally, the evaluation step helps in selecting the best-performing model for the given problem.

In conclusion, the above methodology provides a systematic approach to solve the cybersecurity problem, and the experiments conducted in the code provide a good starting point for further analysis and improvements.

Conclusion, Future Work and Limitations

Conclusion:-

This proposed project a novel intelligent and secured anomaly detection model for cyberattack detection and avoidance in the electric vehicles. From the cyber security point of view, the proposed model could successfully detect malicious behaviors while letting the trusted message frames broadcast in the CAN protocol. The high HR% and FR% indices prove the true positive and true negative decisions made by the proposed model. Regarding the MR% and CR% indices, the very low values which most of them are around the upper and lower bounds of the message frame frequency, show the highly trustable performance of this model. The project will assess the effect of other cyberattacks on the performance of different anomaly detection models in the future works.

Future Work:-

1. **Developing a real-time intrusion detection system**: The current project focuses on offline detection of intrusions using a pre-collected dataset. However, a real-time intrusion detection system that can detect intrusions as they happen could be more useful in preventing unauthorized access.
2. **Improving the accuracy of the model**: The accuracy of the intrusion detection model could be improved by using more advanced machine learning techniques or incorporating other data sources.
3. **Testing the model on different types of vehicles**: The current project only uses data from a specific type of vehicle. Testing the model on different types of vehicles could determine if the model's accuracy varies depending on the type of vehicle.

Limitations:-

1. **Privacy concerns**: The project involves using data from the CAN bus, which contains sensitive information about the vehicle and its passengers. Ensuring that the data is **anonymized** and the privacy of the passengers is protected is important.

2. **Limited data availability**: The availability of data from the CAN bus could be limited depending on the vehicle and its manufacturer. Accessing the data could require specialized equipment or permission from the manufacturer.

3. **Adversarial attacks**: An intruder with knowledge of the intrusion detection model could attempt to evade detection by modifying the data sent over the CAN bus. Developing techniques to detect and prevent such attacks could be challenging.

Thank you