Road Accident Analysis and Classification System

Anuj Jain
Department of Computer Science
KIET Group of Institutions, Delhi-NCR, Ghaziabad
Uttar Pradesh, India
anuj.2024cs1160@kiet.edu

Ayush Pratap Singh
Department of Computer Science
KIET Group of Institutions, Delhi-NCR, Ghaziabad
Uttar Pradesh, India
ayush.2024cs1052@kiet.edu

Arth Srivastava
Department of Computer Science
KIET Group of Institutions, Delhi-NCR, Ghaziabad
Uttar Pradesh, India
arth.2024cs1180@kiet.edu

Harsh Khatter
Department of Computer Science
KIET Group of Institutions, Delhi-NCR, Ghaziabad
Uttar Pradesh, India
harsh.khatter@kiet.edu

***Abstract*—Road accidents persist as a critical global concern, resulting in substantial loss of life and economic burden. This research paper delves into the underlying causes of these accidents, which often stem from a multitude of factors including weather conditions, road surface conditions, and driver behavior. Leveraging machine learning algorithms and data visualization tools, we have developed a predictive model that identifies accident severity. The system integrates real-time weather data, providing a dynamic and comprehensive view of road conditions. By accurately forecasting accident severity, we aim to significantly reduce response times for emergency services and enhance overall traffic safety. This paper provides an overview of our approach, detailing the technologies employed and their potential impact on mitigating the pervasive issue of road accidents.**

***Keywords*— Machine Learning Algorithms, Predictive Model, Road Accidents, Road Surface Conditions, Weather Conditions.**

I. Introduction

Road accidents are a pervasive problem worldwide, contributing significantly to injury and mortality rates. These incidents often occur due to a complex interplay of factors, including adverse weather conditions, road surface conditions, and driver behavior. Prompt and effective response by emergency services is crucial in mitigating the severity of accidentsand reducing their impact on traffic safety.

This research paper leverages machine learning algorithms to develop a predictive model that assesses the severity of road accidents. By analyzing real-time data on weather conditions, road surface conditions, and other relevant parameters, the model aims to provide valuable insights for emergency services. It not only predicts accident severity but also offers recommendations for optimized response times and strategies.In addition to its predictive capabilities, this paper emphasizes data visualization techniques to present the findings inan accessible manner. By making use of data-driven methods, stakeholders can better understand the contributing factors to road accidents, leading to more informed decisions and improved traffic safety measures. This research represents a significant step towards harnessing technology to address the multifaceted challenges posed by road accidents, ultimately striving for safer roads and more efficient emergency responses.

II. Related Work

In the study conducted by Pakgohar et al. [1], it was found that human factors played a significant role in road accidents, contributing to 97.5% of all recorded incidents.Environmental factors, while still significant, accounted for 70.5% of road crashes, followed by vehicle-related factors at 31.5%. Moreover, the research highlighted that in 2006, more than 22% of fatalities resulting from road accidents were non-passengers, including pedestrians and cyclists.

According to Mohad Fedder Musa's research [2], an examination of accident patterns on Malaysian federal roads, comprising 1067 accident cases, revealed noteworthy insights. The most prevalent accident category reported on these federal roads was labeled as 'out of control' incidents. Interestingly, a significant proportion of these accidents transpired during favorable weather conditions, regardless of the time of day.

As highlighted in the study conducted by Mouyid BIN ISLAM [3], human factors remain a pivotal element in crash analyses, a trend observed consistently across various case studies. Nevertheless, it's crucial to acknowledge that both vehicle-related and road-environment factors exert substantial influence on driver behavior, both in the pre-crash and crash phases. These factors merit deeper investigation and more comprehensive analysis, particularly in the contextof developing nations, where vehicle and road-environment standards often differ from those of developed countries.

Nejdet Dogru [4] presents an innovative real-time traffic accident detection method, combining vehicle-to-vehicle communication with machine learning. Tested in the SUMO simulation, the study highlights the effectiveness of Random Forest, Artificial Neural Networks, and Support Vector Machine algorithms in accident recognition. It offers estimated accident locations, aiding quick responses and preventing secondary incidents. The study also demonstrates using machine learning to detect accidents as data outliers. Random Forest stands out for its simplicity and effectiveness.

In the study by Liling Li [5], data mining techniquesare applied to traffic accident data to enhance road safety. Using the Fatal Accidents Dataset for 2007, the research explores association rule mining and classification methods, specifically the Naive Bayes technique. The findings reveal that environmental factors like weather and light conditions have a limited impact on fatality rates compared to human factors like alcohol consumption and collision type. Clustering analysis identifies regions with higher fatality rates, suggesting heightened caution when driving in these areas [11]. The study emphasizes the potential for more robust insights with additional data, such as non-fatal accidents, weather, and mileage data.

In Mohamed K Nour's research [6], a data analytics framework analyzes UK traffic accident data spanning 2005to 2019, aiming to predict injury severity. By integrating information from various sources, including 63 attributes, the study tackles data quality and imbalance issues. It demonstrates that XGBoost excels in handling imbalanced data, outperforming logistic regression, support vector machines, and neural networks. Future work involves exploring parallel processing and comparing rule-based techniques. Notably, decision tree methods like XGBoost and Random Forest perform better due to the categorical attribute nature, despite increased processing time with larger datasets.

In the investigation led by Jayesh Patil [7], a comprehensive analysis delves into factors such as age, gender, weather conditions, vehicle and road conditions, and the driver's mental state. Employing k-means clustering across multiple datasets, the study prioritizes accuracy. The resulting model delivers precise accident predictions across diverse Indian locations, enabling swift accident cause identification and prevention. The analysis notably reveals a higher incidenceof accidents due to speeding during nighttime compared to daytime.

III. Supervised Learning & Classification Algorithms

Machine learning is a data analysis process in which data undergoes preprocessing before being input into a live system to obtain anticipated outcomes. These techniques can be broadly classified into three types, i.e., supervised, unsupervised, and reinforcement learning. We have used various classification techniques, which are a part of supervised learning.

These techniques are pivotal in delineating patterns and relationships within the data, thus enabling accurate categorization of elements into predefined classes or groups [8]. The employed classification algorithms, Decision Tree, Gaussian Naive Bayes, Random Forest, and Logistic Regression, offer distinctive methodologies and attributes.

A. Decision Tree

- **Splitting Data**: Decision Tree begins with the entire dataset and selects the attribute that best splits the data into subsets with the least impurity [12].
- **Recursive Partitioning**: The dataset is recursively divided into smaller subsets based on the selected attributes, forming a tree-like structure.
- **Classification**: When a new instance enters the tree, it traverses through the nodes until it reaches a leaf node, which determines the class label. The code is shown in figure 1.

```
if stopping criteria met:
        return leaf node with the majority class
    else:
        select the best attribute to split on
        create a decision node for the selected attribute
        for each value of the attribute:
            split data into subsets
            attach the child node with the subset
            buildDecisionTree(subset)
```

Fig. 1. Decision Tree Pseudo Code

B. Gaussian Naive Bayes

- **Probabilistic Classification**: Gaussian Naive Bayes estimates the probability of an instance belonging to a class based on the distribution of its features.
- **Independence Assumption**: It assumes that features are conditionally independent, simplifying the calculations.
- **Likelihood Estimation**: It calculates the likelihood of features for each class and uses Bayes' theorem to determine the most likely class [13]. Figure 2 shows the pseudo code of GNB.

```
calculate class prior probabilities
    for each feature:
        calculate mean and variance for each class
```

Fig. 2. Gaussian NB Pseudo Code

C. Random Forest

- **Ensemble of Decision Trees**: Random Forest is an ensemble learning method that combines multiple Decision Trees.
- **Bootstrapped Data**: It creates random subsets of the data through bootstrapping, allowing each tree to learn from a different subset [14].
- **Voting or Averaging**: The algorithm combines the predictions from individual trees through voting (classification) or averaging (regression).
- Figure 3 shows the code.

```
for i from 1 to num_trees:
        create a bootstrapped dataset
        build a decision tree on the bootstrapped data
        save the decision tree
```

Fig. 3. Random Forest Pseudo Code

D. Logistic Regression

- **Sigmoid Function**: Logistic Regression models the probability of an instance belonging to a class using the sigmoid function, which produces values between 0 and 1.
- **Maximum Likelihood Estimation**: It estimates the model parameters to maximize the likelihood of the observed data.
- **Thresholding**: A threshold is applied to the predicted probabilities to classify instances into classes, typically using 0.5 as a cutoff. Figure 4 shows the sample code.

```
for i from 1 to num_trees:
    create a bootstrapped dataset
    build a decision tree on the bootstrapped data
    save the decision tree
```

Fig. 4. Logistic Regression Pseudo Code

IV. Proposed System

In this section, we detail the methodology employed in our research, which encompasses the choice of classification models, feature extraction techniques, and the overall workflow to predict road accident severity accurately. Our study considers a range of factors, such as age, gender, atmospheric conditions, vehicle conditions, road conditions, and the mental state of drivers, which play pivotal roles in understanding the causes of road accidents. Figure 5 shows the proposed model of the system.

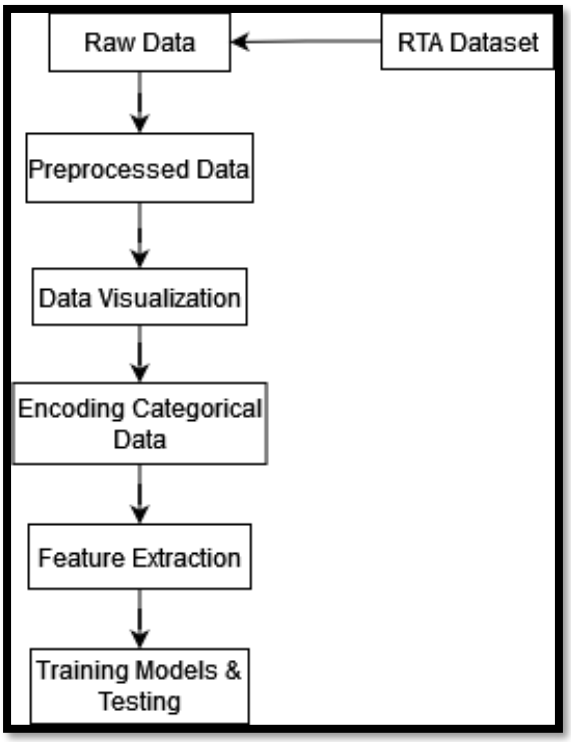


Fig. 5. Block Diagram of Model

A. Data Gathering

Our Dataset was collected from the RTA Dataset of UK [9].

```
Data columns (total 33 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Accident_Index                              20000 non-null  object
 1   Location_Easting_OSGR                       20000 non-null  int64
 2   Location_Northing_OSGR                      20000 non-null  int64
 3   Longitude                                   20000 non-null  float64
 4   Latitude                                    20000 non-null  float64
 5   Police_Force                                20000 non-null  int64
 6   Accident_Severity                           20000 non-null  int64
 7   Number_of_Vehicles                          20000 non-null  int64
 8   Number_of_Casualties                        20000 non-null  int64
 9   Date                                        20000 non-null  object
 10  Day_of_Week                                 20000 non-null  int64
 11  Time                                        20000 non-null  object
 12  Local_Authority_(District)                  20000 non-null  int64
 13  Local_Authority_(Highway)                   20000 non-null  object
 14  1st_Road_Class                              20000 non-null  int64
 15  1st_Road_Number                             20000 non-null  int64
 16  Road_Type                                   20000 non-null  object
 17  Speed_limit                                 20000 non-null  int64
 18  Junction_Detail                             0 non-null      float64
 19  Junction_Control                            15597 non-null  object
 20  2nd_Road_Class                              20000 non-null  int64
 21  2nd_Road_Number                             20000 non-null  int64
 22  Pedestrian_Crossing-Human_Control           20000 non-null  object
 23  Pedestrian_Crossing-Physical_Facilities     20000 non-null  object
 24  Light_Conditions                            20000 non-null  object
 25  Weather_Conditions                          20000 non-null  object
 26  Road_Surface_Conditions                     20000 non-null  object
 27  Special_Conditions_at_Site                  20000 non-null  object
 28  Carriageway_Hazards                         20000 non-null  object
 29  Urban_or_Rural_Area                         20000 non-null  int64
 30  Did_Police_Officer_Attend_Scene_of_Accident  20000 non-null  object
 31  LSOA_of_Accident_Location                   20000 non-null  object
 32  Year                                        20000 non-null  int64
dtypes: float64(3), int64(15), object(15)
```

This Dataset comprised of various attributes which were crucial for classification. Figure 6 shows the dataset attributes.

Fig. 6. Attributes in the Dataset

B. Data Preprocessing

Data preprocessing is a crucial step in ensuring the dataset's quality and uniformity [10]. We carried out data cleaning to handlemissing values, data normalization to standardize the scaleof attributes, and data transformation to create a consistent

```
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Junction_Detail   0 non-null      float64
 1   Junction_Control  15597 non-null  object
dtypes: float64(1), object(1)
```

format. Figure 7 shows the columns with null values.

Fig. 7. Columns with Null Values

C. Data Visualization

Data visualization played a pivotal role in our research. Visualizations were employed to gain insights into the data and its patterns, revealing trends and relationships that might not be apparent through raw data. This step was vital in our data exploration process. Figure 8 shows the significance of speed limit.

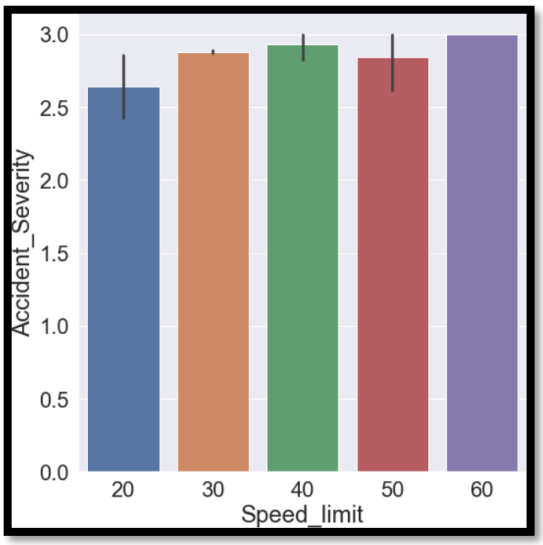


Fig. 8. Significance of Speed Limit

D. Encoding Categorical Data

Many attributes in our dataset were categorical, and machine learning models require numerical inputs. Thus, we encoded categorical data into numerical format, making it suitable for our models using One-Hot Encoding technique.

E. Feature Extraction

Feature extraction is a critical step to determine the most significant attributes for modeling. We utilized the following attributes for our analysis, which is shown in figure 9.

```
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Location_Easting_OSGR   20000 non-null  int64
 1   Location_Northing_OSGR  20000 non-null  int64
 2   Longitude               20000 non-null  float64
 3   Latitude                20000 non-null  float64
 4   Day_of_Week             20000 non-null  int64
 5   Speed_limit             20000 non-null  int64
 6   2nd_Road_Class          20000 non-null  int64
 7   Number_of_Vehicles      20000 non-null  int64
 8   Light_Conditions        20000 non-null  int32
 9   Weather_Conditions      20000 non-null  int32
 10  Road_Surface_Conditions 20000 non-null  int32
 11  Year                    20000 non-null  int64
dtypes: float64(2), int32(3), int64(7)
```

Fig. 9. Selected Attributes

F. Data Training & Analysis

With a cleaned and preprocessed dataset and our chosen features, we proceeded to train our classification models. We divided our data into training and testing sets, where 75% of data was used for training purposes and 25% for testing purpose.

After comparing the results of all the models, we used a voting classifier to maximize the overall accuracy of our system.

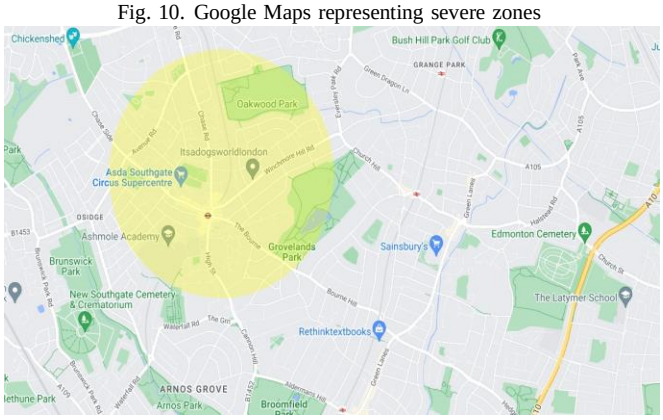
V. Results & Integration

The utilization of the significant attributes identified during our analysis, as mentioned in Section IV, has enabled us to pinpoint high-risk areas and the factors contributing to the severity of accidents. Our results play a crucial role in improving road safety and accident prevention.

Figure 10 provides a visual representation of how our proposed system effectively analyzes specific areas, identifying the most severe zones. These zones are then incorporated into a mapping system, using distinct color codes to denote their severity:

- **Severity Index 1 (Represented in Green)**: This indicates minor accidents, typically involving small injuries or property damage.
- **Severity Index 2 (Represented in Yellow)**: Signifying serious accidents, this category encompasses accidents with moderate injuries and significant property damage.
- **Severity Index 3 (Represented in Red)**: Denoting fatal accidents, this category represents the most severe incidents, often resulting in the tragic loss of life.

Fig. 10. Google Maps representing severe zones



A. Integration with Web Application

To ensure the practical applicability of our findings,we have seamlessly integrated our final predictive model into a user-friendly web application. Our web application incorporates an intuitive graphical user interface (GUI) designed to be user-centric.

This integration provides a real-time and accessible tool for road users, empowering them with knowledge about high-risk zones and accident severity factors. The integration process is aimed at enhancing road safety and reducing accidents.

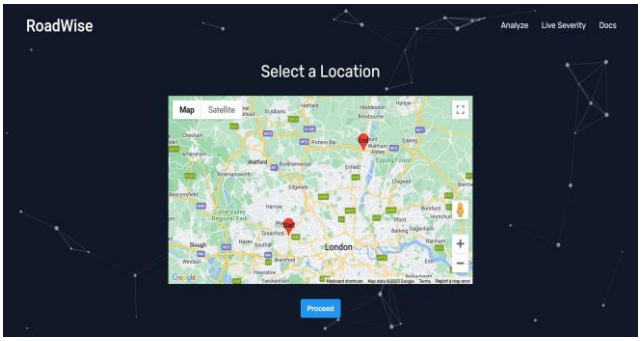


Fig. 11. Select your location from Google Maps

Figure 11 illustrates the user interface functionality, allowing users to interact with the map and select a specific location. This selected location is then seamlessly sent to our integrated model for a comprehensive analysis, enabling real- time predictions of accident-prone areas.

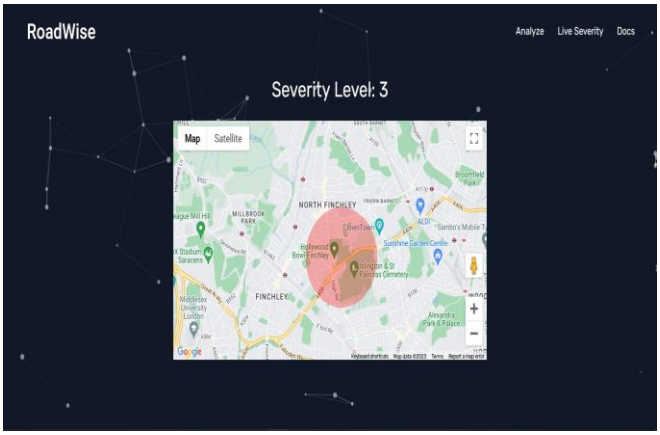


Fig. 11. Get accident severity on your route

Upon selecting your location, our system helps identify accident-prone areas along your chosen route. This feature serves a dual purpose: it aids in making informed decisions during your journey and provides valuable data for in-depth analysis and research.

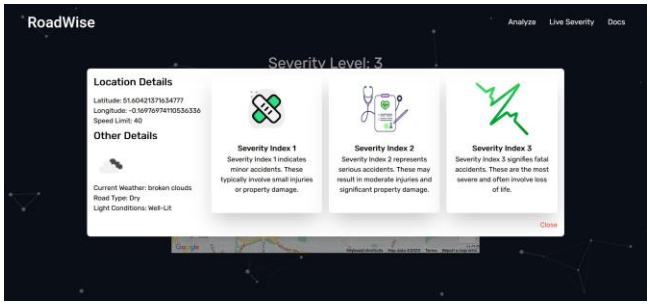


Fig. 12. Additional details about those areas

In addition to the visualization of accident severity zones, users can access precise latitude and longitude coordinates, real-time weather conditions and, essential data concerning road and light conditions.

Also, the defined speed limits withinthese zones are clearly outlined to make safer and more informed decisions while navigating these areas.

B. Architectural Integration

To offer a comprehensive overview, we have attached the architectural layout of our system, highlighting the integrationof the predictive model within the web application. This detailed architectural representation provides insight into how the model functions within the web application, the flow ofdata, and the interaction between different components.

The architecture system comprises two primary components, the frontend, and the backend. Each component plays a critical role in creating a comprehensive and effective solution.

1) Frontend: The frontend of our system is responsiblefor providing an accessible and user-centric interface for road users. We have employed the following technologies to develop the frontend:

- **React**: A popular JavaScript library for building user interfaces, serves as the foundation of our frontend. It facilitated the development of a user-friendly interface through which road users can access real-time information about high-risk areas and receive warnings to enhance safety.
- **Material Tailwind**: An extension of the popular Material-UI framework has been integrated into our frontend to ensure a consistent and visually appealing design.

2) Backend: The backend of our system manages the core functionality of accident prediction and integrates various data sources. It relies on a combination of technologies to ensure the accuracy and reliability of accident data and predictions:

- **Serialization with Pickle**: Pickle, a Python module, has been used for serialization in our system. Serialization allows us to save and load the machine learning models efficiently. This enables the rapid deployment of predictive models within the backend.
- **Flask Server**: The Flask server serves as the heart of the backend operations. It hosts our predictive models and handles real-time data requests from the frontend.

Data Integration: To enhance the accuracy of accident predictions, our system integrates data from external sources, including the OpenWeatherMap API and theGoogle Maps API. The OpenWeatherMap API provides real-time weather information, which is a crucial factorin understanding road conditions and accident likelihood.Google Maps API contributes to location-based data, assisting in pinpointing high-risk areas.

VI. CONCLUSION

Machine learning, a rapidly advancing field in research and analytics, has proven to be a pivotal tool for understanding complex data patterns and making accurate predictions. Inthis study, we used the potential of machine learning to tackle the pervasive issue of road accidents, considering a multitude of factors, and aiming for the utmost accuracy.

In our research, we worked on a predictive model to address the pressing problem of road accidents. By accounting for a wide range of variables, we've sought to minimize deviations and enable the efficient identification of high-risk areas. In a world increasingly reliant on data-driven solutions, our research underscores the pivotal role that machine learning can play in enhancing road safety. In conclusion, our comprehensive analysis and predictive model pave the way for a safer road environment. By leveraging machine learning, we aim to prevent accidents and overcome road safety challenges, working toward a safer world.

REFERENCES

[1] A. Pakgohar, Reza Tabrizi, Mohadeseh Khalili, and Alireza Esmaeili, "The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach," Procedia Computer Science, vol. 3, pp. 764-769, December 2011.

[2] Mohad Fedder Musa, Sitti Asmah Hassan, Nordiana Mashros, "The impact of roadway conditions towards accident severity on federal roads in Malaysia," PLoS One, 2020 Jul 6. DOI: doi.org/10.1371/journal.pone.0235564.

[3] Mouyid Bin Islam, Kunnawee Kanitpong, "Identification of factors in road accidents through in-depth accident analysis," IATSS research, vol. 32, no. 2, pp. 58-67, 2008. Publisher: Elsevier.

[4] Nejdet Dogru, A. Subasi, "Traffic accident detection using random forest classifier," 15th Learning and Technology Conference (L&T), 1 February 2018.

[5] Liling Li, Sharad Shrestha, Gongzhu Hu, "Analysis of road traffic fatal accidents using data mining techniques," International Conference on Software Engineering Research and Applications, 7 June 2017.

[6] Mohamed Nour, Atif Naseer, Basem Alkazemi, Muhammad Abid Jamil, "Road Traffic Accidents Injury Data Analytics," International Journal of Advanced Computer Science and Applications, vol. 11, no. 12, January 2020. DOI: 10.14569/IJACSA.2020.0111287.

[7] Jayesh Patil, Vaibhav Patil, Dhaval Walavalkar, Vivian Brian Lobo, "Road Accident Analysis and Hotspot Prediction using Clustering," 2021 6th International Conference, Publisher: IEEE, Date of Conference: 08- 10 July 2021.

[8] "Classification Algorithms in Machine Learning," online. https://medium.datadriveninvestor.com/classification-algorithms-in-machine-learning-85c0ab65ff4 Accessed on October 26, 2023.

[9] "UK Road Safety: Traffic Accidents and Vehicles," online. https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles Accessed on October 26, 2023

[10] J. Patil, M. Prabhu, D. Walavalkar, and V.B. Lobo, "Road accident analysis using machine learning," In 2020 IEEE Pune Section International Conference (PuneCon), IEEE, pp. 108–112, 2020

[11] Y. Wang and W. Zhang, "Analysis of roadway and environmental factors affecting traffic crash severities," Transportation Research Procedia, vol. 25, pp. 2119–2125, 2017

[12] Harsh Khatter, Amrita Jyoti, Rashmi Sharma, Pooja Malik, Rashmi Mishra "Enhancing Network Efficiency and Extending Lifetime through Delay Optimization and Energy Balancing Techniques", Wireless Personal Communications (2023). https://doi.org/10.1007/s11277-023-10812-7

[13] Xue Liu, Xiaowei Wang, Feng Ren, Ming Zhang, Harsh Khatter, Afroj Alam, "Multi-objective optimization scheduling of sequential charging software for Networked Electric Vehicles", Journal of Sensors (2022), Hindawi, Vol 2022(6968470), pp. 1-8, 29 July 2022, https://doi.org/10.1155/2022/6968470

[14] Amit Kumar Gupta, Ruchi Rani Garg, Arti Sharma, Saurabh and Harsh Khatter, "Posture Identification Using Artificial Neural Network", Presented in 3rd International Conference on Smart Computing and Cyber Security SMARTCYBER, 28-29 June 2023, Kyungdong University, South Korea.