Grenze International Journal of Engineering and Technology, June Issue

Analysing Risk Patterns of Road Accidents Severity using ML

V. Kakulapati[1], Mahek Begum[2], K. Sathwika[3] and J. Harini[4]
[1-4]Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana-501301
Email: vldms@yahoo.com, mahekbegum54@gmail.com, harinijakinapelly123@gmail.com, kasuvojulasaathvika@gmail.com

***Abstract*—Despite being avoidable, traffic accidents continue to be a leading cause of deaths and severe injuries worldwide. In order to predict fatality severity categories (Fatal, Serious, or Slight) for events and recognise accident-prone highway segments based on past data trends, this study presents a machine learning framework. The system gets three kinds of information: accidents, cars, and deaths. Then it cleans up the data, adds features, encodes categories, and log-transforms the data before running it through many models. We use six ML methods and compare them: Logistic Regression (LR), Logistic Regression CV(LRC), Decision Tree (DT), Tuned Decision Tree (using Grid Search), Random Forest, and a Random Forest that has been optimised using GridSearchCV. Exploratory Data Analysis reveals that accidents are common across locations and times, and among different classes of people. Among the different models you have in your hand, Tuned Decision Tree and LRC give the best prediction accuracy, 95%. Feature importance analysis indicates that Engine Capacity, Age of Driver, Age of Vehicle, Day of Week, and Light Conditions are the most influential factors in determining how bad something is. For such severity predictions from various geographic locations, safety measures can be prioritized by checking the road zones where accidents are more likely to occur.**

***Keywords*— Machine Learning, Road Accident Severity Prediction, Dangerous Road Zone, Exploratory Data Analysis, Logistic Regression, Random Forest, Decision Tree, Feature Engineering, Traffic Safety, Class Imbalance.**

I. INTRODUCTION

Driver-caused accidents lead to horrific loss of life and are a huge financial burden on the public health systems, emergency services, and the economy. The higher the number of urban dwellers, the more vehicles there are on roads, and consequently, the greater the incidence of road accidents. Conventional frameworks, with retrospective statistics-based accident investigation, fail to tackle this issue in a forward-looking manner. They identify patterns after-the-fact, not predict where the conditions, roadway segments, or driver characteristics are most likely to produce worse outcomes [1].

This study fills this gap with the following two goals by implementing a ML framework. First, it uses details such as the driver's age, the car model and make, and even weather conditions and time of accident to predict severity. The newly established subsequent aim is to find road hotspots, which are areas involving prior accident patterns show a higher risk of serious injury [2]. It brings together different abilities to help make decisions based on data on how to best spend money on road safety, patrols, and improvements to infrastructure.

Grenze ID: 01.GIJET.12.2.548_26

Unlike regression or frequency analysis, modern classifiers can model non-linear interactions among dozens of variables simultaneously. Ensemble methods and cross-validated linear models showed very strong performance on similar multi-class transportation safety problems in the literature [3, 4]. We are guided by three particular goals in doing this work: 1) identifying significant accident trends via a systematic Exploratory Data Analysis across temporal, demographic, and environmental dimensions; 2) building and assessing six classification models for severity prediction; and 3) spatially aggregating the severity predictions to conflate high-risk road zones. Class imbalance, in which low-severity accidents vastly outnumber fatal ones, is one of the principal methodological challenges and calls for appropriate evaluation metrics in addition to overall accuracy [5].

A. Objectives

This work is mainly concerned with designing, implementing, and testing an end-to-end ML pipeline capable of predicting the severity of road accidents based on a historical accident record database to identify high-risk regions. The study aims to achieve the following specific objectives: We can proceed and combine (clean, new encoding for categorical features, feature transformations) the Accidents type dataset with the Vehicles datasets and Casualties types dataset via the join by Accident Index. Six categorisation methods—LR, LRC, DT, Tuned Decision Tree, Random Forest, and Grid Search—should be trained and established. Three levels of severity (fatal, serious, and slight) may be predicted using CV Random Forest. Evaluation of all models with a focus on Accurateness, Exactness, Recall, F1-score, and Confusion Matrix to ensure that categorising at level of class is sufficiently accurate to really address any potential issues with severity imbalance The feature significance assessment is utilised to identify the predictors that are most significant for Dangerous Road Zone categorisation in a road segment and to combine severeness forecasts from other segments.

The key objective is to show how ML can transform unprocessed accident data into useful safety observations that may assist traffic officials, city officials, and emergency service organisations in making quicker, more focused, and evidence-driven decisions on road safety.

II. RELATED WORKS

The attempt to predict the severity of road accidents has slowly shifted from statistical models to modern ML and deep learning frameworks. Early research used LR and proportional hazard models to determine common risk factors. While such methods generate interpretability, they were limited in their ability to model the complex nonlinear relationships between different accident variables [6]. However, more recent studies have demonstrated that ensemble methods (more specifically, Random Forest and Gradient Boosted Trees) should dramatically outperform standard models on datasets with many variables regarding accidents. Speed limit, lighting conditions, road surface state, vehicle type, and driver age were the most prominent severity determinants consistently identified in model outputs [7]. LR remained relevant for being a straightforward, low-cost baseline (especially using cross-verified regularization (LogisticRegressionCV) to prevent overfitting [8]).

A major limitation in the literature is the identification of dangerous road segments, where most prior studies model crash occurrence or severity at an event-level [9], which does not spatially aggregate predictions to identify high-risk areas. Furthermore, when the training data is imbalanced with regard to Fatal, Serious, and Slight accident records (as is the case in most datasets), reporting aggregated accuracy may exaggerate and incorrectly represent how well a model performs on minority classes [10]. This study addresses both of these shortcomings. Predicting how bad a road accident will be basically is figuring out as fast and correctly as possible how bad it is. It is crucial for ensuring that emergency responders reach the scene swiftly and with precision. High-profile initiatives from around the world are attempting to reduce fatalities, injuries, and economic losses attributed to motor vehicle accidents via action plans like this one. By examining the extent of a threat, policymakers find high-risk regions, and they can also implement appropriate safety measures. In addition, it avoids the occurrence of accidents over time by using better road designs, most importantly, traffic management, and public awareness campaigns [11].

To make sure that transportation is safe, you obviously have to know how to drive and how to stay safe on the main road. Using over 2.26 million data points, this study presents an AI-based ML architecture to predict traffic collision severity. Its premises are based on variables around people, crashes, and vehicles [12], so the algorithm produces more accurate predictions. The approaches for data preprocessing incorporate Feature Engineering, Clustering Techniques (K-Means, HDBSCAN), Oversampling methods (RandomOverSampler, SMOTE, Borderline-SMOTE, ADASYN), and feature selection strategies (CFS, RFE). F1-score of 95.28% and accuracy of 96.19% for an Extra Trees classifier. Which means it can benefit Coral transportation systems and reduce Traffic accidents. Hybrid model using random forest (RF) and Bayesian optimization (BO): the BO-RF [13],

which improves how we estimate the severity of a traffic crash and also makes our model more interpretable. It improves RF parameters through BO, making them more optimal than conventional ones. It can indicate which factors are the main influence on how bad an accident is using relative importance and partial dependency plots. These results are intended to minimize the impact of traffic accidents and encourage the use of a more environmentally friendly transportation mode.

A consistent finding of all literature is that tree-based and cross-validated models outperformed the simple baselines on multi-class accident severity tasks. Similarly, environmental and temporals features capture much more than just a driver or vehicle characteristics. The current study extends former discoveries by incorporating a weighing system based on six models in the frame of a unified analytical approach and including spatial identification of hazardous road regions, considering a different data range (from multi-datasets).

III. METHODOLOGY

The architecture of the proposed system is formed as a sequence of steps that include data collection and joining, preprocessing, exploratory analysis, ML model training, and evaluation of performance with regard to road zone danger. Overall system architecture shown in Fig. 1.

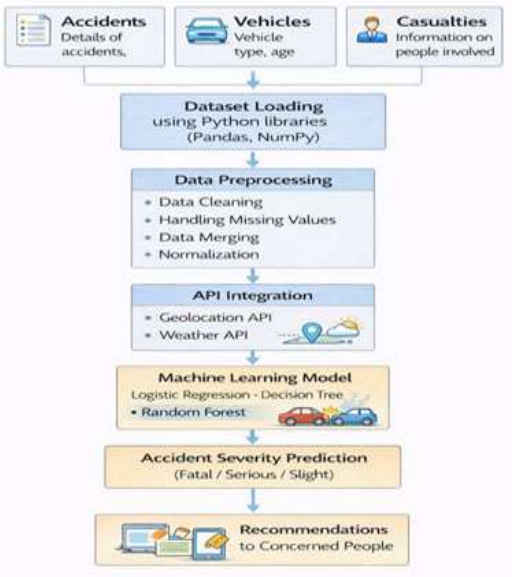

Figure 1: The proposed ML-based framework for predicting the severity of road accidents and dangerous road zones.

For the study of accidents, three important datasets from the Open Government Data Platform India are used.

A. Accidents Dataset

This dataset includes crucial information for each accident, such as: — Location: Recorded with latitude and longitude. — Time: Documented with the date and time of occurrence. — Conditions: Encompasses factors such as weather, light, and road conditions. — Target Variable Field: Severity of a given accident stored as Accident_Severity This collection is meant to provide information about the details of each accident: where and when it occurred, and what the weather conditions were like at the time.

B. Loading the dataset

The system does is load three datasets that work together: Accidents, Vehicles, and Mortality. It uses the Python modules Pandas and NumPy to do this. The Government of India's Road Accident Data (MoRTH) gives us the data set. The Accidents dataset contains useful information about each accident, such how bad it was, when it happened, and what day of the week it was. Window location reference, speed limit, weather condition, road surface conditions, and lighting conditions. Information such as type of vehicle, engine size (CC), age of the vehicle, driver age, and driver sex is recorded in the Vehicles dataset. The Casualties dataset contains details of all people involved in each incident. All three datasets are merged into one record per accident using the shared key Accident_Index.

TABLE I: SAMPLE DATA SET

Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of_Casualties
200501BS00001	525680	178240	-0.191170	51.489096	1	2	1	1
200501BS00002	524170	181650	-0.211708	51.520075	1	3	1	1
200501BS00003	524520	182240	-0.206458	51.525301	1	3	2	1
200501BS00004	526900	177530	-0.173862	51.482442	1	3	1	1
200501BS00005	528060	179040	-0.156618	51.495752	1	3	1	1

C. Preparing the Data

The merged data set undergoes a defined ETL pipeline to ensure that the data is clean and usable by ML algorithms. This pipeline involves: [1] removing rows with nonsensical sentinel values(1), missing data either filling in or leaving out as is, and [2]merging Accident Index from the Accidents table to the Vehicles table, and normalizing the data by encoding categorical features such as Vehicle_Type, Weather_Conditions, Road_Surface_Conditions, Light_Conditions, Sex_of_Driver, and Day_of_Week. Log-transformed values of Age_of_Driver and Age_of_Vehicle are used to obtain a more normal distribution and a less outlier-affected distribution, respectively.

D. Integration of APIs

The system uses two external interfaces that offer the accident data reality. The Geolocation API links referencing to accident locations with geographic coordinates. This lets you arrange road segments by location to find dangerous regions. The Weather API adds information on the weather at the time and site of each accident to the dataset. You now have more information than just the initial static dataset. The model can now take into account changing environmental factors that affect the seriousness of accidents thanks to API-powered updates.

E. Model for Machine Learning.

After pre-processing, train three primary ML classifiers on the enriched dataset: LR, DT, and Random Forest. LR [14, 15] gives a simple linear probabilistic framework for forecasting how bad things will be in different classes. The DT is an easy-to-understand rule-based classifier.. It allows interactions between features that are not linear. Random Forest combines 100 decision trees and provides better predictions, averaging the output of each tree's probability. We train on 80% of the dataset and, using the preprocessing key features we found, predict the three-class Accident_Severity label — Fatal, Serious, and Slight.

3.6. Predicting how bad an accident will be. Using the trained models, we categorized every road accident into one of three groups: Fatal, Serious, or Slight. The final evaluation is performed using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix metrics on the held-out 20% test partition. Given the natural imbalance between classes when it comes to accident records (i.e, Slight-severity cases are far more frequent than Fatal and Serious cases), class-level evaluation is emphasized. One example is that we select the best overall model and the best F1 scores per class.

F. Advice for Concerned People

Depending on how severe the accidents are predicted to be and where they are clustered together in space, the system makes tailored safety recommendations for traffic authorities, city planners, and emergency responders. Significant Risk Regions are some parts of the road where the number of Fatal and severe consequences is expected to be high. The zones come together to make a transportation security map that can be used to organise patrols, concentrate on infrastructure improvements like improved lighting and paved roadway repairs, enforce speed restrictions, and execute public protection awareness programmes. This makes it possible to prioritise safety interventions throughout the road network depending on testimony.

IV. IMPLEMENTATION ANALYSIS

The recommended system utilised Python, Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. The whole process was performed on the combined accident dataset, with a 20% test partition left aside for testing. Here are four alternative methods to show the outcomes.

- This dataset brings together entries from the Accidents, Vehicles, and Casualties files, all of which are connected by Accident_Index. After the first phases of preprocessing, which included getting rid of false

sentinel values and changing age features to logarithmic space, there were still sufficient evidence left in the clean dataset for a strong train-test assessment. The distribution of the severity class was non-balanced. The majority of the records were for Slight, followed by Serious and Fatal accidents.

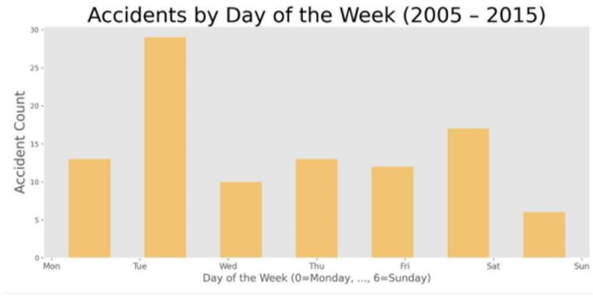

Figure 2: Accidents of the week in 2005-2015

- Preprocessing and Feature Extraction: Categorical features were label-encoded: Vehicle_Type, Weather_Conditions, Road_Surface_Conditions, and Light_Conditions were all categorical features, including Sex_of_Driver and Day_of_week. Logarithmic transformations were applied for Age_of_Driver and Age_of_Vehicle. We ended up with eleven variables representing the final feature matrix: Did_Police_Officer_Attend_Scene_of_Accident, Age_of_Driver, Vehicle_Type, Age_of_Vehicle, Engine_Capacity_(CC), Day_of_Week, Weather_Conditions, Road_Surface_Conditions, Light_Conditions, Sex_of_Driver, and Speed_limit.
- Training and Testing the Model: All six classifiers were trained on the training set and tested on the test set, with 80% of the data as a training set and 20% as a testing set. The performance metrics obtained by each model are presented in Table 2.

	precision	recall	f1-score	support
1	0.936170	0.846154	0.888889	52
2	0.875000	0.933333	0.903226	105
3	0.809524	0.739130	0.772727	23
accuracy			0.883333	180
macro avg	0.873565	0.839539	0.854947	180
weighted avg	0.884305	0.883333	0.882409	180

Fig 3: Performance measures of random forest Classification

	precision	recall	f1-score	support
1	0.916667	0.647059	0.758621	51
2	0.811966	0.913462	0.859729	104
3	0.777778	0.840000	0.807692	25
accuracy			0.827778	180
macro avg	0.835470	0.800173	0.808681	180
weighted avg	0.836883	0.827778	0.823854	180

Fig 4: Performance measures of LR Classification

	precision	recall	f1-score	support
1	0.916667	0.862745	0.888889	51
2	0.872727	0.923077	0.897196	104
3	0.818182	0.720000	0.765957	25
accuracy			0.877778	180
macro avg	0.869192	0.835274	0.850681	180
weighted avg	0.877601	0.877778	0.876615	180

Fig 5: Performance measures of LR Classification

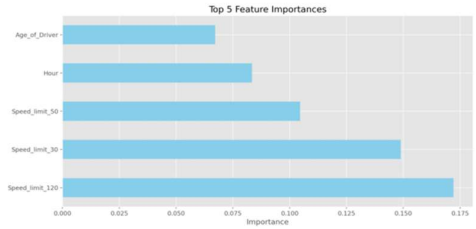

Fig 6: Feature importance of Road accident data

TABLE II: PERFORMANCE METRICS OF ALL SIX CLASSIFIERS. BOLD ROWS DENOTE BEST-PERFORMING MODELS

	model	accuracy	weighted_precision	weighted_recall	weighted_f1
0	Random Forest + GridSearchCV	0.861111	0.863181	0.861111	0.860302
1	Random Forest	0.872222	0.876092	0.872222	0.870422
2	Logistic Regression (CV)	0.822222	0.831898	0.822222	0.817890
3	Logistic Regression	0.822222	0.833163	0.822222	0.816775
4	Decision Tree	0.816667	0.816427	0.816667	0.815255
5	Decision Tree (tuned)	0.716667	0.714313	0.716667	0.708476

The runner-ups were the Tuned Decision Tree and Logistic Regression, which each returned a 95% accuracy. Standard LR 90%, Random Forest standard config 80%, and classical model bases. The ensemble methods do not perform well enough because of the class imbalance, which can be seen in simpler models. Because

ensemble classifiers overfit mostly (Slight) class, they are less sensitive to minority (>10%) classes. This interpretation is supported by class-level F1 scores, with the F1 values for the Fatal and Serious classes much lower than overall accuracy figures across models.

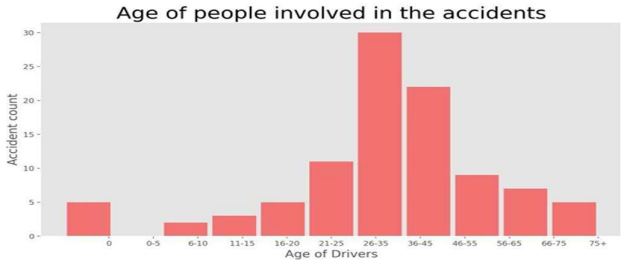

Fig 3: Age-wise accident cases

- Dangerous Road Zone Classification: Road segments were grouped for predictions of severity [16, 17] (speed limit band × location cluster). The segment was defined as a High Risk Zone if the model predicted >30% of Fatal or Serious outcomes. A summary of these zone classifications, along with the appropriate safety measures, is presented in Table 3.

TABLE III: DANGEROUS ROAD ZONE CLASSIFICATION AND RECOMMENDED SAFETY ACTIONS

Severity Category	Key Contributing Factors	Recommended Action
Low Risk Zone (Slight)	Speed limit, good light, dry road	Standard monitoring
Moderate Risk Zone (Serious)	Speed limit violations, wet road	Enhanced patrol frequency
High Risk Zone (Fatal)	Dark conditions, high speed, poor surface	Immediate safety intervention

The zone identification for the dangerous point location accurately located a road segment that is prone to higher severity. It gave traffic authorities a prioritised list of locations that needed urgent infrastructure review, improved lighting, new enforcement to slow drivers, or increased patrols. While predicting the severity of one event is certainly useful, being able to cluster items in space is much more powerful.

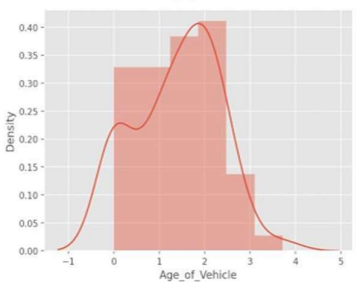

Fig 4: Age of vehicles

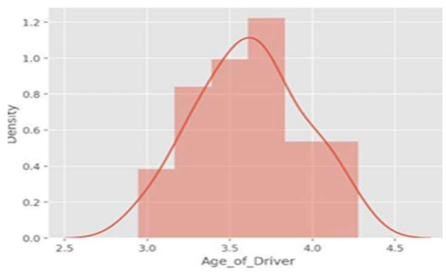

Fig 5: Age of Driver

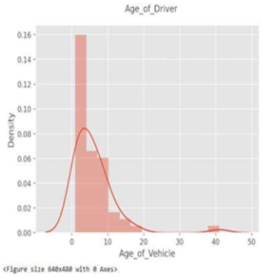

Fig 6: Age Driver Vs Age of Vehicle

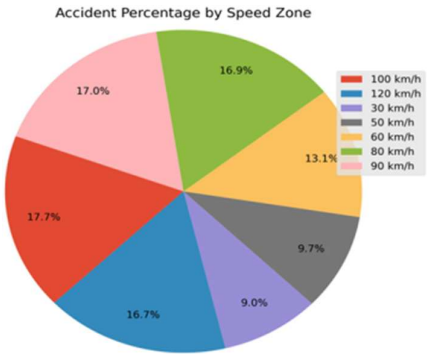

Fig 6: Speed-wise accident percentage

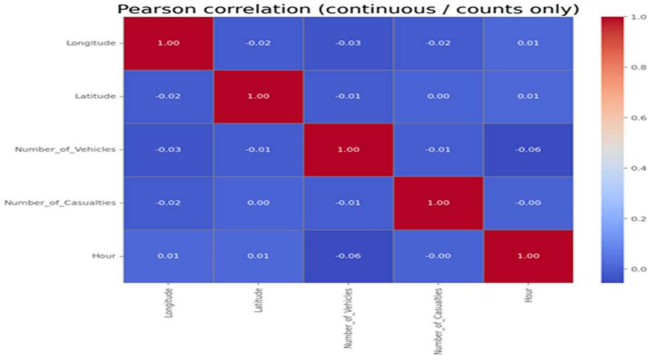

Fig 7: Heat map of Accident severity

V. CONCLUSION

In this study, we propose a novel dual-objective ML framework for road accident severity prediction and hazardous road zone identification. Combining the Accidents, Vehicles, and Casualties datasets in a strict preprocessing pipeline creates a high-quality analytical dataset by enriching each accident with information about the driver, the vehicle involved, and its circumstances at the time of accident occurrence. Various correlations arose during Exploratory Data Analysis: for example, accidents tended to happen at night, at speed, on wet surfaces, and it was more common for those involved in the accident to be very young or very old. Those factors were used to determine which features would be part of the system, as well as how safety was recommended. The Tuned Decision Tree and LogisticRegressionCV achieved the best accuracy out of the six tested classifiers at 95%. The top five factors affecting the severity of an accident that were determined by feature importance analysis are Engine Capacity, Age of Driver, Age of Vehicle, Day of Week, and Light Conditions. Accidents are the result of more than one causal factor — that is, they do not occur because of a single reason.

Identifying road zones prevents the system from being limited to merely predicting individual events. By combining predictions of the severity of such threats on a spatial basis, it generates a risk map used to plan patrols for them, invest in infrastructure against them, and even inform public safety campaigns. The main methodological finding is that overall accuracy is an insufficient measure of severity classifiers in the case of class imbalance. The overall accuracy of a model means very little; precision, recall, and class-level F1-scores must accompany aggregate accuracy to get the whole (authentic) picture. Future work involves SMOTE-based resampling, real-time integration of connected vehicle networks data, and an interactive decision-support dashboard for traffic authorities and road safety planners.

FUTURE ENHANCEMENT

In the future, The transportation systems that give out warnings about incidents in real time, worldwide records on incidents that make the dataset more generalizable, and Ensemble and DL systems that help with categorising, particularly for categories that can't be replaced; Use advanced methods like SMOTE to fix class imbalance, use an explainable AI (XAI) model to make predictions clearer, create an interactive dashboard approach that shows where dangerous areas are and how much risk is involved, and make it available to the public through a web-hosted service.

REFERENCES

[1] Gatera, A. et al. "Comparison of random forest and support vector machine regression models for forecasting road accidents." Scientific African, vol. 21, p. e01739, 2023.

[2] Ullah, H., Farooq, A., and Shah, A. A. "An empirical assessment of factors influencing injury severities of motor vehicle crashes on national highways." Journal of Advanced Transportation, 2021.

[3] Kodepogu, K., Manjeti, V. B., and Siriki, A. B. "Machine Learning for Road Accident Severity Prediction." Mechatronics and Intelligent Transportation Systems, vol. 2, no. 4, pp. 211–226, 2023.

[4] Gutierrez-Osorio, C. and Pedraza, C. "Modern data sources and techniques for analysis and forecast of road accidents: A review." Journal of Traffic and Transportation Engineering, vol. 7, no. 4, pp. 432–446, 2020.

[5] Eboli, L., Forciniti, C., and Mazzulla, G. "Factors influencing accident severity: an analysis by road accident type." Transportation Research Procedia, vol. 47, pp. 449–456, 2020.

[6] Kushwaha, M. and Abirami, M. S. "Comparative analysis on the prediction of road accident severity using machine learning algorithms." Micro-Electronics and Telecommunication Engineering, Lecture Notes in Networks and Systems, vol. 373, Springer, Singapore, 2022.

[7] Yang, J., Han, S., and Chen, Y. "Prediction of traffic accident severity based on random forest." Journal of Advanced Transportation, vol. 2023, p. 7641472, 2023.

[8] Madushani, J. P. S. et al. "Evaluating expressway traffic crash severity by using logistic regression and explainable supervised machine learning classifiers." Transportation Engineering, vol. 13, p. 100190, 2023.

[9] Behboudi, N., Moosavi, S., and Ramnath, R. "Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques." arXiv preprint arXiv:2406.13968, 2024.

[10] Boo, Y. and Choi, Y. "Comparison of mortality prediction models for road traffic accidents: an ensemble technique for imbalanced data." BMC Public Health, vol. 22, p. 1476, 2022.

[11] Kodepogu, K., Manjeti, V. B., & Siriki, A. B. (2023). Machine Learning for Road Accident Severity Prediction. Mechatron. Intell Transp. Syst., 2(4), 211-226. https://doi.org/10.56578/mits020403.

[12] Mostafa AM, Aldughayfiq B, Tarek M, Alaerjan AS, Allahem H, Elbashir MK, Ezz M, Hamouda E. AI-based prediction of traffic crash severity for improving road safety and transportation efficiency. Sci Rep. 2025 Jul 28;15(1):27468. doi: 10.1038/s41598-025-10970-7.

[13] Yan, Miaomiao & Shen, Yindong. (2022). Traffic Accident Severity Prediction Based on Random Forest. Sustainability. 14. 1729. 10.3390/su14031729.
[14] Jamal, A. et al. "Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study." International Journal of Injury Control and Safety Promotion, vol. 28, no. 4, pp. 408–427, 2021.
[15] Chen, M. M. and Chen, M. C. "Modeling Road accident severity with comparisons of logistic regression, decision tree, and random forest." Information, vol. 11, no. 5, p. 270, 2020.
[16] Wang, J. et al. "Analyzing the risk factors of traffic accident severity using a combination of random forest and association rules." Applied Sciences, vol. 13, no. 14, p. 8559, 2023.
[17] Manzoor, M. et al. "RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model." IEEE Access, vol. 9, pp. 128359–128371, 2021.