INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

Data-Centric Approach to Road Crash Severity Estimation Using Recursive Feature Selection and Deep Learning

[1]Syed Nurja, [2]Saidu Soumya, [3]Veldurthi Venkat Siddarth, [4]Vadde Chethana Priya

[1]Assistant Professor[2]Student,[3]Student,[4]Student
[1]Dept. Computer Science and Engineering (Data Science)
CMR Technical Campus, Hyderabad, Telangana, India.

Abstract : Injuries and deaths in the world are among the main causes of road accidents. Timely and correct emergency assistance is required to ensure lives are spared. Emergency response is important in saving lives because road accidents still remain as one of the principal causes of injuries and deaths in the world. Conventional analysis of accident severity makes use of manual reporting which may lead to delays in reporting and lead to the impediment of the real-time decision making. This research paper proposes a hybrid algorithm combining the use of Random Forest Recursive Feature Elimination (RF -RFE) with a deep learning system, as a solution to this issue. RF-RFE selects the most significant characteristics, including vehicles information, the kind of the accident, weather, and demographics. These enhanced features categorize the deep learning model as mild, moderate or severe accidents. The new model, which is tested faster and more accurate than ancient machine-learning methods, indicates that the new avenue has clearly shown promising outcomes. In an emergency system, it may accelerate the speed of severity assessments, enhance the process of sharing resources and accelerate medical services, increasing survivor probability.

Index Terms **- Deep learning, Feature selection, Road accidents, Random Forest, Recursive Feature Elimination**

1.INTRODUCTION

Accidents on roads are a major worldwide concern in the safety. They kill and maim a lot of people every year. According to the World Health Organization 2 approximately 1.19 million deaths happen in traffic accidents annually. Other than loss of life, these accidents are very expensive to the country in terms of medical care, lost jobs and damaged roads and buildings. In most countries, these expenses constitute a conspicuous constituent of the GDP of the country. One of the most important actions in curbing the number of death cases that occur as a result of road traffic accidents is an efficient emergency response system. The response may take long, which increases the severity of the situation with a patient. An understanding of the degree of the badness of an accident would aid in the placement of emergency help where it is most needed. This study seeks to improve the response to an emergency by establishing a model that predicts hazardous accidents beforehand so that ambulance and medical staff could respond faster. There are numerous types of data utilized to predict the severity of the accident: measurements of speed and weather, police report text, and camera photos. This complicated information can be processed with ensemble learning and deep learning methods. Accidents can be categorized as severe or not, or as minor, serious, fatal or no injury, and this would mean that models would be less precise. This is to predict the severity of bad road accidents with the aim of using sophisticated models to predict and understand the factors that influence such predictions. They aim at enhancing emergency strategies and providing evidence-based recommendations to reduce the effects of road accident.

- It presents the RF-RFE technique of selecting the most significant aspects, condensing the data, and enhancing the prognoses of the accidents.
- It combines Tomek Link under sampling and SMOTE oversampling to form the SMOTE TOMEK technique. This fixes the non-uniform data and provides a better and balanced training set.
- The CNN -Bi LSTM -Attention model combines convolutional nets, bidirectional LSTM and attention. It identifies important characteristics, discovers spatial patterns as well as enhances predictive accuracy.
- SHAP describes the entire model, providing clear insights on the decisions as well as demonstrating the key factors that drive the severity of an accident.
- A step-by-step performance appraisal examines the impact of each component of the model on general performance, so that the most suitable method is achieved in predicting the severity of the accident and the emergency response.

2.LITERATURE REVIEW.

2.1 A. Sharma, R. Gupta, and S. Kulkarni, "Explainable AI for Road Accident Severity Prediction Using Ensemble Learning," 2024.

Sharma et al. (2024) proposed an ensemble-based framework integrating Random For est, XG Boost, and Light GBM combined with Explainable AI (XAI) techniques such as SHAP for interpreting crash severity predictions. The study used Indian road accident 3 datasets (2016–2022) and achieved an accuracy of 88.7%, outperforming standalone models. Key influencing factors included road conditions, weather, driver behavior, and time of accident. The major contribution was interpretability, enabling authorities to understand model decisions. However, the model lacked real-time deployment capability and did not integrate live traffic or IoT sensor data. The authors suggested incorporating streaming data and edge computing for real-time emergency response systems

2.2 L. Wang, Y. Chen, and H. Zhao, "Spatiotemporal Deep Learning for Traffic Accident Severity Prediction," 2024.

Wang et al. (2024) developed a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to capture both spatial and temporal dependencies in accident data. Using large-scale Chinese ur ban traffic datasets (2018–2023), the model achieved 86.3% accuracy and demonstrated superior performance over traditional ML models. Important features included traffic density, weather conditions, and time-of-day patterns. Despite its effectiveness, the model required high computational resources and lacked interpretability. The study recommended integrating attention mechanisms and lightweight architectures for scalable real-time deployment.

2.3 M. Kashifi, Deep hybrid attention network: Deep hybrid attention network to predict the severity of road-crashes in real time: 2023.

Kashifi used a very limited number of inputs such as the approximate time and location of the incident to form a Deep Hybrid Attention Network (DHAN) and forecast the crash severity in real time. A good performance was observed in the DHAN model, which was approximately 76 per cent accurate and 80 per cent recalls in the sample data (2011-2017) of crash data of France. Its weaknesses were that it was limited to the association of spatial and temporal variables and only resulted into a binary outcome, although more detailed categories can be helpful. The authors also stated that emergency medical service (EMS) operators should prevent unnecessary dispatching and also optimize the recall and false-alarm rates. To have a more accurate response, they proposed to check multi-class severity prediction and a vast variety of contextual data.

2.4. S. Patel and M. Verma, "Accident Severity Prediction Using Federated Learning in Smart Cities," 2025.

Patel and Verma (2025) introduced a federated learning-based framework for accident severity prediction, enabling multiple smart city nodes to collaboratively train models without sharing sensitive data. Using datasets from multiple Indian metropolitan cities, the model achieved 84.9% accuracy while preserving data privacy. The approach effectively captured regional variations in accident patterns. However, communication overhead and model convergence issues were observed. The authors suggested optimizing communication protocols and integrating edge AI for faster inference in emergency response scenarios.

2.5. Graph SAGE H. Sattar et al., Graph SAGE crash severity prediction, 2022.

To determine which accident data, need to be linked, Sattar and colleagues constructed a K -nearest neighbor graph using the UK accident data. They used Graph SAGE and its models to make predictions based on the severity of accidents. The graph-based model achieved approximately 85.5 per cent accuracy compared with the traditional machine learning methods including XG Boost, Random Forest and artificial neural networks which had lower accuracy rates. This demonstrates that accident data can be complex and non-linear using graph methodology. Nonetheless, they managed class imbalance by random under sampling thereby decreasing the variety of data and this may compromise robustness. The authors suggested the use of cost-sensitive learning and improved sampling methods to enhance usefulness in the real world by increasing inference and proper trustworthiness of the model in emergency response mechanisms.

3.EXISTING SYSTEM

The most popular machine-learning algorithms used in traditional models to predict the extent of an accident are the Random Forest, the naive bayes, the SVM and the Logistic regression. They deal with organized table like data, and severity as a yes/no problem or a few category problems. These models are able to perform satisfactorily though there are some issues that prevent these from being exemplary. Several methods fail to select the most desirable features, and hence the data is noisy and highly vast. Since fatal crashes are infrequent, the models have a tendency to be biased toward common and minor classes. They also find it difficult to describe the interaction of the complex road environments, driver behavior, vehicle information, and weather. Lastly, due to their difficulty in elucidation, they may prove counterproductive in cases where emergency teams require fast straight forward answers.

4.PROPOSED METHODOLOGY

The paper presents a new system of prediction of serious accidents. It employs a number of methods: RF-RFE feature selection, SMOTE-Tomek class balancing, and a neural network, which is a combination of CNN, BiLSTM, Attention. RF -RFE selects the most significant features, thus the model will be trained on useful data. SMOTE-Tomek corrects the imbalance of the classes by eliminating those that may belong to multiple classes and also increasing the sample of the smaller classes. This makes the model effective in all the levels of severity. CNN component learns feature to feature connections whereas BiLSTM learns temporal patterns. The focus is laid on the most important attributes in terms of injury severity. Lastly, SHAP justifies the predictions as a result, the emergency responders are able to view what the reason why an accident is expected to be serious.

5.IMPLEMENTATION

5.1. Data Collection & Pre-processing:

This module gathers the data on the accidents based on official sources. It contains information on the type of the vehicle, the road conditions, weather, the characteristics of the driver, time, place of the accident, and the severity of the accident. The data normally contains gaps and irregularities, and noise. To correct this, we preprocess: that is, we fill in missing values, bring numeric values to a common scale, convert categorical data to numbers, convert time and date data to usable form, and delete non-useful records. The data was cleaned and then di vided into training, validation and testing sets. This ensures that there is clean, consistent and reliable data that will be used in future.

5.2. Feature Selection using RF-RFE:

The In this module, the Random Forest -Recursive Feature Elimination (RF -RFE) technique is applied to identify which features con tribute the greatest to the severity of accidents. RF-RFE prioritizes the inputs into the system based on their usefulness in making predictions, and finds the least useful ones one after another. This reduces the variables and preserves the crucial information, and this enhances accuracy and less time is required to train deep-learning models.

5.3. Class Balancing Module (SMOTE-Tomek):

The difference in the number of each type of accident often has a significant change in accident data. Accidents that result in death are a lot fewer in contrast to minor and no-injury accidents. This is assisted by the use of the SMOTE (Synthetic Minority Over-sampling Technique)-Tomek method. SMOTE synthesizes samples of the rare types. Tomek links collapse too similar or noisy examples across types, which enhances the balance of the data set and improves it.

5.4. Deep Learning Severity Prediction (CNN– BiLSTM–Attention):

This system component makes predictions. It employs three classes of methods: convolutional steps that recognize patterns within space, BiLSTM steps that identify time and sequence and attention that directs focus to the most significant aspects. These are determined jointly on the severity of an accident minor, no injury, serious, and fatal. Such forecasts assist emergency workers to have a sense of priorities when assisting those in need.

5.5. Explainability Model (SHAP):

The explainability module takes advantage of SHAP and elucidate the manner in which the model arrives at predictions. It determines the degree to which each input assists in determining the severity of the accident and therefore emergency responders and the analysts can understand why an accident is considered severe or fatal. This facilitates the system and leads to trust, and can be used to identify potential biases or strange trends in the conduct of the model.

5.6. Admin Module:

Admin Module can enable an administrator to manage data, initiate the cleaning of data, retrain the models, and monitor prediction outcomes. Admins can also view the information about the accuracy of the model, confusion tables, balances in the classes, and the features that are of most importance. They are also able to manage user accounts, give them permission to access and view in-depth reports on severity of accidents based on predictions. This module ensures that the system is functioning as it should and the admins manage the back-end.

5.7. User Model:

The User Module provides a simplified interface to the emergency managers, any analysts, or any system users. Journalists are able to log in, input the data about the accidents, and receive the prompt forecast concerning the extent of the accident severity. The module gives elaborate outcomes, the intensity and straightforward explanation of each prediction. Report exporting can also be done by the user to assist in emergency planning, decision making about ambulance dispatch, and policy formation.

6.SYSTEM ARCHITECTURE

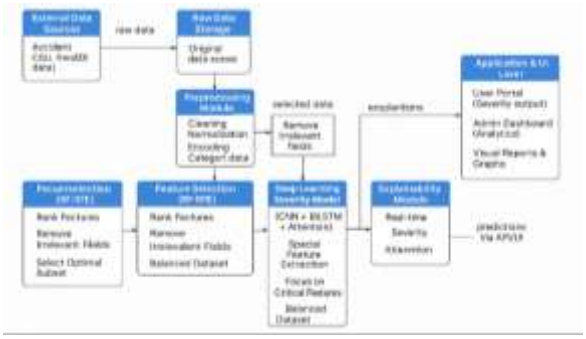

Fig. 1. System architecture of Data-Centric Approach to Road Crash Severity Estimation Using Recursive Feature Selection and Deep Learning

7.METHODOLOGY

The framework of the accident severity prediction also takes a simple stepwise method that is easy to understand to ensure that the data is good, learning is fair and prediction is accurate. It starts by cleaning and normalizing accident records provided by official sources, handling missing values, numeric features that need scaling, handling categorical features that need to be converted to numbers and time data that needs to be converted to analyzable format. Then a Random Forest-Recursive Feature Elimination (RF-RFE) technique is used to select the most relevant features in order to have the model to concentrate on what actually influences the 7 severities of accidents. Due to the low number of fatal cases, the SMOTE-Tomek method gives a balance to the data through increased sample of minorities and elimination of rough overlaps. The clean data subsequently gives training to a hybrid CNN-BiLSTM-Attention model. In this case, the CNN layers will be used to record spatial relationships, Bi-LSTM layers will be used to learn temporal relationships, and the key factors will be emphasized by the attention mechanism. SHAP based explainability is used to make predictions understandable and enhance transparency. Lastly, the results are provided via a user-friendly interface, which provides data management tools, retraining models, and performance monitoring, being useful in making fast and effective decisions during an emergency response

8.RESULTS

Fig. 2. Home page for the user

Fig. 3. Admin login page

Fig. 4. Uploading dataset for batch prediction

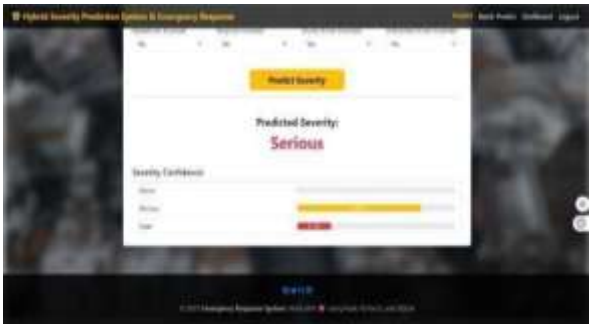

Fig. 5. Severity prediction of a geographical area

Fig. 6. Feature reduction for batch prediction

9.CONCLUSION

Proposed accident severity prediction architecture works with the assistance of the sophisticated machine learning and deep learning and assists in emergency response decisions. Using the RF-RFE and class balancing feature selection method on the input data, the system transforms the data into high-quality and organized. The CNN BiLSTM-Attention model also learns spatial relationship, time patterns and significant factors contributing to accidents, thus enhancing prediction of various degree of levels of accidents. With the inclusion of SHAP explanations the results become clear and reliable thus the responders can understand why such severity was expected. In general, the system is strong and can be extended to predict the severity of accidents particularly in busy cities where immediate decisions are needed. It can assist traffic administrators, hospitals and EMS in multitasking resources as well as prioritizing serious cases resulting in reduced response time and higher life-saving mechanisms.

REFERENCES

[1] M. Kashifi and K. Ahmad, “Road accident severity prediction using histogram-based gradient boosting,” 2022.
[2] Alotaibi, J., et al., “Enhancing traffic accident severity prediction using machine learning techniques,” Vehicles, 2025.
[3] M.Kashifi, “Deep hybrid attention network for real-time road crash severity prediction,” 2023
[4] H. Sattar et al., “Crash severity prediction using Graph SAGE on KNN-based accident graphs,” 2022.
[5] A.Assi, “Hybrid PCA–MLP–SVM model for traffic accident severity prediction,” 2021.
[6] M. Ahmed et al., “An explainable machine learning model for road accident injury severity in New Zealand,” 2022.
[7] N. Rezashoa et al., “Light GBM–Optuna-based traffic accident severity prediction on U.S. roads,” 2023.