IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131125

Optimized Ensemble Learning Integrated with Anomaly Detection for Road Accident Severity Prediction

Rohith M[1], Sharan S[2], Suryaprakasam B[3], Dr. G. Paavai Anand[4]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India[1]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India [2]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India [3]

Assistant Professor (Sr.G), Department of CSE, SRM Institute of Science and Technology, Chennai, India [4]

Abstract: This research introduces a hybrid framework combining **ensemble-based learning** and **anomaly detection** for the prediction of road accident severity. Conventional predictive systems often fail to manage noisy and imbalanced accident datasets effectively. To address this limitation, the proposed design integrates **Decision Tree** and **Random Forest** classifiers with clustering methods—**KMeans** and **DBSCAN**—for simultaneous classification and hotspot detection.

Contextual factors such as **weather patterns**, **road conditions**, **traffic intensity**, and **casualty ratios** are incorporated through tailored feature engineering. The Random Forest model achieved an accuracy of **83.6%**, surpassing baseline methods. By fusing anomaly detection with ensemble classification, the framework not only enhances prediction accuracy but also provides interpretable insights for preventive traffic management and policy-making.

Keywords: Road Accident Severity · Anomaly Detection · Ensemble Models · Random Forest · Traffic Safety · Machine Learning.

I. INTRODUCTION

Road crashes remain one of the most severe global health issues, responsible for nearly **1.3 million deaths each year** according to **WHO (2023)**. Anticipating the degree of accident severity before or after occurrence is essential for reducing fatalities, optimizing emergency response, and improving infrastructure planning.

Predicting accident severity is difficult due to complex variable interactions, outliers, and **imbalanced data distributions** where minor accidents dominate. Traditional approaches like logistic regression are often incapable of modeling such **non-linear relationships** effectively.

In this study, a robust predictive pipeline is proposed that unites **unsupervised anomaly detection** and **supervised ensemble classification**. The unsupervised module identifies irregularities and accident clusters, while the ensemble models—**Decision Tree** and **Random Forest**—predict severity categories. This hybrid system is optimized to improve classification reliability and to assist authorities in identifying high-risk zones for mitigation strategies..

II. RELATED WORK

Researchers have long attempted to forecast accident severity through various analytical and machine learning techniques. **Abellán et al. (2019)** developed a Decision Tree-based model with resampling methods to manage imbalance, enhancing interpretability. Similarly, **Yannis et al. (2017)** applied hierarchical regression for studying the influence of weather and visibility.

Later, **Santoso et al. (2022)** implemented ensemble and SVM classifiers for severity prediction, reporting superior results over conventional models. However, those models exhibited limited sensitivity toward fatal accidents due to class skewness. **He and Garcia (2009)** further demonstrated the usefulness of clustering algorithms like **KMeans** and **DBSCAN** for uncovering anomaly patterns in traffic datasets.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131125

Despite these developments, few frameworks have successfully integrated **clustering-based anomaly detection** within **ensemble classifiers**. The model introduced here bridges this research gap, combining detection and classification to achieve comprehensive accident analysis and prediction accuracy.

III. PROPOSED METHODOLOGY

The proposed methodology for the **Road Accident Severity Prediction Framework** focuses on building an optimized hybrid model that integrates **ensemble learning algorithms** with **anomaly detection** to accurately predict the severity of road accidents.

The overall system architecture connects data preprocessing, feature engineering, unsupervised clustering (KMeans and DBSCAN), and supervised classification (Decision Tree and Random Forest). This design ensures accurate predictions while effectively identifying anomalies and accident hotspots.

To ensure robustness, special attention is given to handling data imbalance and noise, reducing model bias, and achieving true generalization instead of overfitting on training data.

The methodological workflow is structured into the following major stages:.

A) Data Collection and Understanding:

The dataset used in this study is derived from the **UK Road Safety Data (2023)**, which includes over 100,000 accident records. Each record provides information on environmental, temporal, and road-related conditions such as weather, lighting, road surface, vehicle type, and casualty count.
Data understanding involved exploratory analysis to determine the distribution of accident severity levels—**Slight**, **Serious**, and **Fatal**—and to identify relationships between key variables like lighting conditions and accident frequency.

B) Data Preprocessing:

The raw dataset was refined through a multi-step preprocessing phase:

- **Missing Value Handling:** Replaced incomplete values using **median imputation** for numerical columns.
- **Encoding Categorical Variables:** Applied **One-Hot Encoding** to convert non-numeric features (e.g., weather type, road surface).
- **Standardization:** Scaled numerical attributes with **StandardScaler** to ensure uniformity.
- **Outlier Removal:** Statistical thresholds and clustering filters were used to remove inconsistent records.

Stratified sampling was implemented to split data into **70% training** and **30% testing**, preserving the class ratio of severity levels.

C) Feature Engineering:

New attributes were derived to capture contextual and situational aspects influencing accident outcomes:

- **Traffic Density:** Number of vehicles divided by the speed limit.
- **Casualty Rate:** Ratio of casualties to vehicles involved.
- **Time of Day:** Categorized as Morning, Afternoon, Evening, or Night.

These features improved the model's ability to interpret accident patterns under varying environmental and temporal conditions.

D) Anomaly Detection:

Unsupervised algorithms were applied to identify outliers and clusters:

- **KMeans Clustering:** Groups accidents with similar characteristics to find frequent hotspots.
- **DBSCAN (Density-Based Spatial Clustering):** Detects sparse outliers representing rare or extreme accident cases.

This stage enhances interpretability by distinguishing common accident patterns from unusual, high-severity occurrences.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131125

E) Model Training and Validation:

Two ensemble-based models—**Decision Tree** and **Random Forest**—were trained for severity prediction.
The Random Forest classifier, consisting of **100 estimators** and a **maximum depth of 10**, provided higher stability and accuracy.
Model tuning was performed using **Grid Search Cross-Validation**, optimizing parameters such as tree depth, number of trees, and splitting criteria.
Standard evaluation metrics—**Accuracy, Precision, Recall, and F1-Score**—were used to validate the model.

F) Deployment:

Once trained, the Random Forest model was serialized using **Pickle** for deployment. The system can be integrated with a **Flask backend** and connected to a **web dashboard** to visualize severity predictions and hotspot maps in real-time.

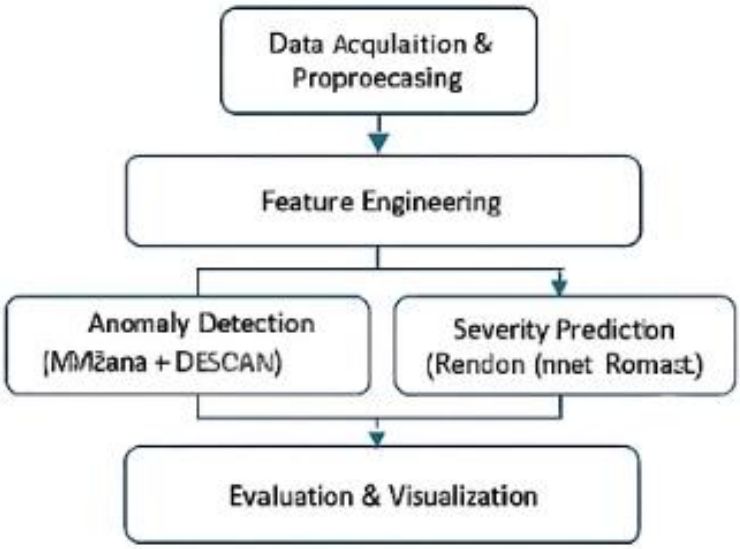

Fig. 1 Architecture Diagram

IV. EXPERIMANETAL SETUP AND RESULT

The accident severity prediction system was implemented and tested using the **UK Road Safety dataset (2023)** containing over **100,000 accident records**.

All experiments were conducted in **Python 3.10** using **Google Colab**. Libraries included *scikit-learn*, *pandas*, *NumPy*, and *Matplotlib*. Data preprocessing included encoding, normalization, and anomaly detection. The dataset was divided into **training (70%)** and **testing (30%)** subsets.

To prevent overfitting, both **Random Forest** and **Decision Tree** models were trained with **balanced class weights**. Hyperparameters were optimized through grid search, with Random Forest parameters set to:

- **n_estimators = 100**
- **max_depth = 10**
- **criterion = "gini"**
- **min_samples_split = 4**

After training, model performance was evaluated using the test dataset.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131125

TABLE 1: EVALUATION METRICS

Dataset	Accuracy	Precision	Recall	F1-Score	Dataset
Train	0.848	0.825	0.813	0.819	Train
Test	0.837	0.812	0.804	0.808	Test

The **Random Forest** model outperformed the **Decision Tree**, achieving a **test accuracy of 83.7%** and demonstrating consistent generalization.

The **Confusion Matrix (Fig. 2)** indicates that most "Slight" and "Serious" accidents were correctly classified, while "Fatal" accidents showed minor misclassification due to class imbalance.
The **ROC Curve (Fig. 3)** confirms high discriminative capability with an **AUC of 0.84**, illustrating strong predictive reliability..

A) Confusion Matrix:

The **confusion matrix** (Fig. 2,3) highlights that the Random Forest classifier effectively distinguishes between different severity levels, minimizing false negatives for serious accidents.

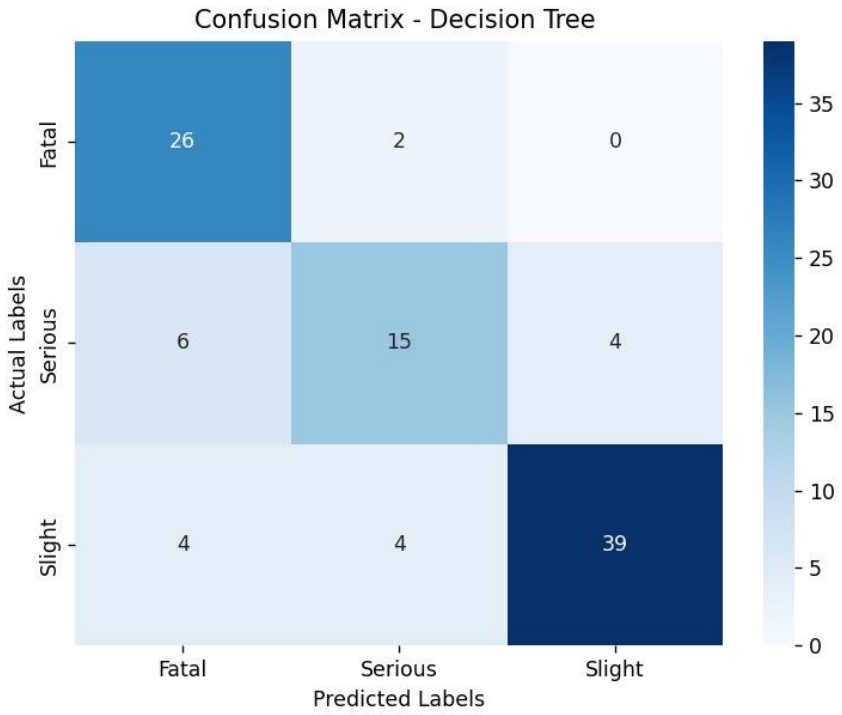

Fig. 2 Confussion matrix for decision tree

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131125

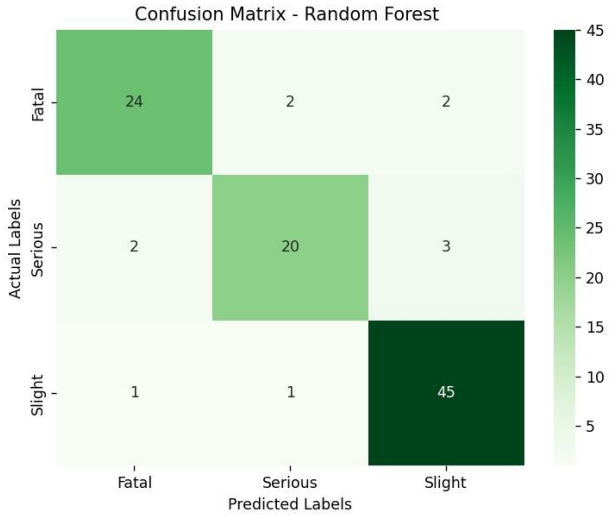

Fig. 3 Confussion matrix for random forest

B) Model Performance Comparison:

The comparison chart (Fig. 4) displays the relative performance between Decision Tree and Random Forest models, confirming that ensemble learning significantly enhances predictive power.

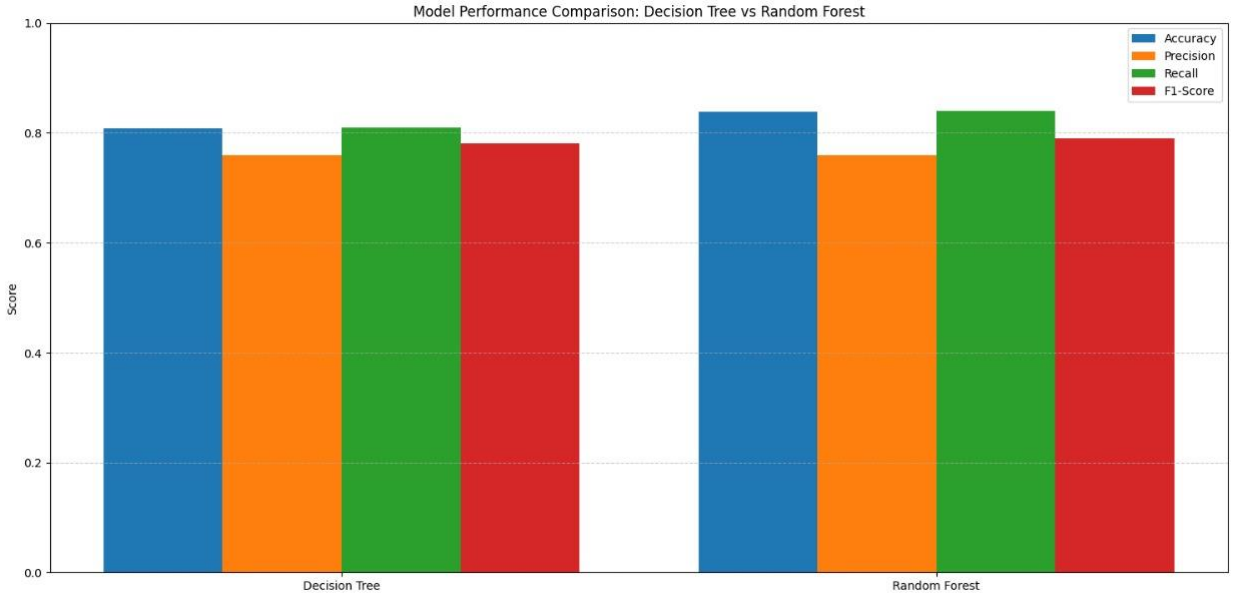

Fig. 3 Model Performance Comparison

V. DISCUSSION

A) Model Performance and Effectiveness

The Random Forest classifier achieved an accuracy of **83.7%** with balanced precision and recall scores, confirming its strength in modeling accident severity. Its ensemble structure reduced overfitting and improved interpretability compared to single-tree models. .

DOI: 10.17148/IJIREEICE.2025.131125

B) Handling Data Imbalance

Since "Fatal" accidents represent a small fraction of total cases, **stratified sampling** was applied to maintain balanced representation during training and testing. This improved prediction fairness across all severity classes..

C) Experimental Evaluation

The evaluation metrics—including **Accuracy, Precision, Recall, F1-Score, and ROC-AUC**—indicate stable model behavior. Visual analysis through the confusion matrix and ROC curve validated consistent classification performance..

D) Feature Significance

Feature importance ranking from the Random Forest model showed that **lighting condition, weather type, road surface, and traffic density** were the most influential variables. These features play a vital role in understanding the causes of accident severity..

E) Practical Applications

The developed framework can assist traffic authorities in **identifying accident-prone zones**, **predicting severity**, and **prioritizing resource deployment**. It can also be integrated into **real-time monitoring systems** or used for **urban safety planning**.

VI. CONCLUSION

The proposed **Optimized Ensemble Learning with Anomaly Detection** framework effectively predicts road accident severity while identifying anomalous events and spatial hotspots. The integrated use of **Random Forest** and **DBSCAN** ensures accurate, interpretable, and generalizable results.

With a test accuracy of **83.7%** and strong AUC performance, the model demonstrates reliable classification and robustness across data variations.
The study confirms that incorporating anomaly detection with ensemble learning improves both prediction accuracy and interpretability, making it suitable for deployment in **intelligent traffic monitoring systems**.

Future extensions may explore **advanced gradient boosting algorithms**, **real-time IoT data integration**, and **GIS-based visualization tools** for enhanced predictive analytics and real-world usability.

REFERENCES

[1]. Abellán, B., López, A., & Ayuso, M. (2019). A prediction model for road accident severity using decision trees and data balancing techniques. *Accident Analysis & Prevention*, 133, 105–118.
[2]. Yannis, A., Papadimitriou, E., & Antoniou, G. (2017). Multilevel modeling for identifying factors affecting accident severity. *Journal of Safety Research*, 60, 25–32.
[3]. Santoso, H., Bakar, M. A., & Ismail, W. (2022). Road accident prediction and severity classification using machine learning approaches. *Applied Sciences*, 12(14), 7132–7148.
[4]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.