TRAFFIC SEVERITY PREDICTION USING MACHINE LEARNING AND DEEP LEARNING

Authors:
Mrs. G.Shruthi [1], A.J. Shruthi [2], A. Srikanth[2], S. Vivek Chary [2]

[1]Assistant Professor, CMR Engineering College, Kandlakoya, Medchal.
[2] Btech - Computer Science Engineering – Data Science, CMR Engineering College, Kandlakoya, Medchal – 501401, Telangana.

Abstract :

Traffic accidents on highways remain a major cause of fatalities, even with advancements in traffic safety measures. The impact of injuries and damages from road incidents is especially severe in developing countries. Various factors lead to traffic accidents, with some significantly affecting the severity of these incidents. Data mining techniques can be instrumental in predicting the key factors linked to crash severity. This research pinpoints essential elements that correlate closely with accident severity on highways using Random Forest analysis. Key features influencing accident severity includeRange, heat level, cold breeze, moisture, clarity, and air movement.The study introduces a hybrid model that combines machine learning and deep learning methods, specifically Random Forest and Convolutional Neural Network, referred to as EFC(Ensemble Fusion Classifier) to forecast the severity of road accidents. The effectiveness of this model is evaluated against several baseline classifiers. The aim of this research is to improve the accuracy of predicting traffic accident severity by utilizing machine learning and deep learning techniques. The proposed EFC(Ensemble Fusion Classifier) model employs Random Forest for feature selection and a Convolutional Neural Network for enhanced pattern recognition, allowing for a thorough analysis of accident severity. This hybrid strategy enhances predictive capabilities by revealing complex relationships among contributing factors. The study emphasizes the benefits of merging machine learning and deep learning to create a dependable system for evaluating accident severity, which can assist traffic management authorities in implementing proactive safety measures.

Keywords *:Traffic Accident Severity, Highway Safety, Machine Learning, Deep Learning, Random Forest, Convolutional Neural Network, RFCNN Model,Data Mining.*

I INTRODUCTION

Road traffic accidents (RTAs) remain a leading cause of fatalities, injuries, and property damage around the world. Beyond their immediate effects on individuals, these incidents place a significant burden on healthcare systems, emergency services, and national economies. Reports from the Ministry of Public Security of China indicate that road accidents claim thousands of lives and leave hundreds of thousands injured each year. Understanding the factors that influence the severity of these accidents is essential for developing effective prevention strategies and reducing casualties.

Recent developments in data science and machine learning have introduced new methods for analyzing traffic accident data and predicting severity levels. While traditional statistical models like logistic regression and ordered probit models have been widely used, they often struggle with complex datasets. In contrast, machine learning techniques such as Random Forest, Support Vector Machines, and Deep Learning offer improved predictive power and are increasingly applied in the study of accident severity. This research proposes an ensemble model EFC(Ensemble Fusion Classifier)that integrates Random Forest and Convolutional Neural Networks to enhance the accuracy of predictions related to traffic accident severity. By leveraging data-driven insights, this study aims to support ongoing research in accident prevention and contribute to efforts focused on improving traffic safety.

1.1 The Consequences of Road Traffic Accidents

Road traffic accidents (RTAs) have significant economic and social effects. The World Health Organization reports that millions of people are affected by traffic incidents each year, with many suffering from long-term disabilities. The financial costs are also substantial, as road accidents represent a large portion of a nation's GDP. In lower-income areas, the economic impact may seem smaller in percentage terms, but the consequences can be much more severe due to limited healthcare and emergency services. In contrast, wealthier countries face annual losses from road accidents that amount to hundreds of billions of dollars, affecting productivity, insurance rates, and infrastructure expenses.

1.2 The Significance of Forecasting Accident Severity

The importance of predicting accident severity cannot be overstated, as it plays a crucial role in improving emergency response strategies, guiding policy decisions, and enhancing infrastructure development. Traditional methods often rely on historical accident data, expert opinions, and statistical analyses. However, these approaches can struggle with scalability and efficiency when faced with large and complex datasets. Recent advancements in artificial intelligence and machine learning have enabled researchers to analyze high-dimensional data, identify patterns, and generate more precise predictions, ultimately boosting the reliability of assessments regarding accident severity.

1.3 Methodological Approach and Data Accuracy

The accuracy of predictions about accident severity depends heavily on the quality and reliability of the available data. In many developing regions, comprehensive records of road accidents are often insufficient, leading to underreporting and reduced prediction accuracy. To improve data reliability, it is crucial to implement better data collection methods, such as real-time traffic monitoring systems, sensor-based vehicle information, and the combination of public and private datasets for machine learning applications. A well-organized dataset plays a vital role in creating a more accurate and scalable model for predicting accident severity.

1.4 Study Contributions and Implications

The precision of models predicting accident severity is significantly affected by the richness and quality of the data employed. This research improves prediction accuracy by merging both structured and unstructured data sources via the RFCNN hybrid model. By utilizing a combination of machine learning and deep learning methodologies, the study provides a more thorough examination of the factors influencing accident severity. The results aid in the advancement of data-driven strategies for road safety, facilitating improved policy formulation and preventive actions aimed at decreasing traffic-related fatalities.

II LITERATURE SURVEY

Traffic accident severity prediction has become an important research focus, leveraging machine learning (ML) and deep learning (DL) models to improve both accuracy and interpretability. Traditional ML techniques, such as Random Forest, Decision Trees, and Logistic Regression, have been employed to analyze factors influencing accident severity, including weather conditions, road infrastructure, and traffic volume. Research by Jiang et al. [1] and Behboudi et al. [2] indicates that ML models can effectively forecast accident severity, underscoring the importance of feature selection methods. Furthermore, deep learning approaches like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Artificial Neural Networks (ANNs) have shown promising results in enhancing prediction accuracy by skillfully extracting features from diverse datasets, as highlighted by Chaw et al. [3] and Sajadi et al. [4].

Recent studies have explored advanced architectures like Residual Neural Networks (ResNet) alongside SHapley Additive explanations (SHAP) to boost model transparency and support informed decision-making in road safety initiatives, as highlighted by Benfaress et al. [5]. Furthermore, transfer learning techniques, especially those utilizing MobileNet models, have been examined to improve both predictive accuracy and interpretability, as mentioned by Aboulola [6]. Additionally, hybrid models that combine Dynamic Binary Particle Swarm Optimization (BPSO) with deep learning methods have shown promise in enhancing predictions by utilizing real-time data from social media, according to Raul et al. [7]

Feature selection plays a crucial role in accurately predicting the severity of accidents. Miao et al. [8] used mutual information techniques to pinpoint the most significant factors influencing severity. Similarly, Tang et al. [9] applied factor analysis to explore how elements like visibility, traffic signals, and road surface conditions affect prediction accuracy. Additionally, CNN-based frameworks have been employed to analyze accident severity in detail by converting structured data into feature matrices, which enhances the learning process, as noted by Jaiswal et al. [10]. Another method involves real-time video analysis and computer vision techniques to identify anomalies in traffic patterns, aiding in the development of accident prevention strategies, as investigated by Girija et al. [11]. Furthermore, reinforcement learning approaches have been explored to optimize emergency response systems based on predictions of accident severity, improving resource allocation in high-risk areas, as discussed by Jaiswal et al. [12]. There is a growing emphasis on the need for interpretable models, utilizing SHAP and Local Interpretable Model-agnostic Explanations (LIME) to ensure transparency in accident severity predictions, as shown by Benfaress et al. [13]. Recent developments have concentrated on integrating various data sources, such as weather reports, traffic sensors, and historical accident data, to enhance predictive accuracy through multi-modal data fusion techniques, as highlighted by Jiang et al. [14]. Some research, like that of Moosavi et al. [15], has explored ensemble learning methods that merge multiple weak learners to form a more robust model for severity classification. Additionally, studies have looked into federated learning frameworks to maintain data privacy while achieving high prediction accuracy across different regions, as demonstrated by Huang et al. [16].

The use of natural language processing (NLP) techniques has become increasingly popular, utilizing social media content and traffic reports to gather important insights for predicting accident severity, as discussed by Rout et al. [17]. Research by Ramnath et al. [18] showed that graph-based learning methods can effectively model traffic networks and forecast accident severity by considering spatial-temporal dependencies. Additionally, cloud-based artificial intelligence models have been suggested to improve real-time predictions of accident severity by taking advantage of distributed computing capabilities, as highlighted by San et al. [19]. However, despite these technological advancements, several challenges persist, such as data imbalance, model interpretability, and the complexities of real-time implementation. Future studies should aim to incorporate explainable AI (XAI) techniques to build trust in predictive models while also enhancing the resilience of deep learning frameworks against adversarial threats in traffic data, as pointed out by Zinedine et al. [20]

III EXISTING SYSTEM

The existing research on predicting accident severity outlines the models utilized, the datasets analyzed, and the accuracy rates achieved. Most of these studies are closely tied to accident severity prediction, as they rely on datasets specifically centered around traffic incidents and accident records. This information is crucial for traffic engineers and field professionals, as it can guide targeted safety measures, help prioritize resources, and enhance traffic management strategies to reduce accident severity and improve overall road safety. However, the analysis also uncovers certain gaps in research, such as the necessity for studies that focus specifically on accident severity prediction in particular regions or that utilize larger datasets encompassing a variety of accident scenarios. Additionally, some studies do not provide thorough explanations of how the model outcomes can be applied in practical settings for professionals. Nonetheless, the findings from these studies offer a significant contribution to the creation of data-driven and effective models for predicting accident severity, with practical implications for those working in the field. Addressing these research gaps will further enhance the understanding and use of predictive models, ultimately leading to better road safety measures and a decrease in accident severity.

IV DISADVANTAGES

1 Limited Dataset Scope: Numerous studies predominantly depend on datasets that focus on traffic incidents, which may exhibit a lack of variety in terms of accident types, environmental conditions, or regional influences, thereby constraining the applicability of the results.

2.Geographic Constraints: Research frequently neglects data that is specific to certain regions, complicating the effective application of models across diverse geographic areas characterized by differing traffic patterns, road infrastructures, and driving behaviors.

3.Lack of Practical Implementation: Several studies do not offer explicit recommendations on how the outcomes of models can be incorporated into actual traffic management systems, thereby diminishing their practical utility for professionals in the field.

4.Incomplete Coverage of Accident Scenarios: Current models may not adequately address intricate accident scenarios, including multi-vehicle collisions, incidents influenced by weather conditions, or infrequent yet severe crashes.

5.Resource-Intensive Models: A number of predictive models, especially those employing sophisticated machine learning techniques such as Random Forest and Convolutional Neural Networks, often necessitate significant computational resources, rendering implementation expensive and less feasible for smaller organizations or municipalities.

V MATERIALS AND METHODS

This section elaborates on the classifiers, dataset, and methodologies employed for analyzing road accident severity in this study.

5.1 MODELING METHODS

In this research, we employ an ensemble learning framework known as RFCNN (Random Forest and Convolutional Neural Network). This model combines the advantages of both machine learning and deep learning techniques to improve the predictive accuracy concerning the severity of road accidents.

5.1 BASE MODELS

The study incorporates several foundational learning models, each contributing uniquely to the predictive performance of the RFCNN model

5.1.1 AdaBoost Classifier (AB): This algorithm focuses on misclassified observations by adjusting their weights, thereby ensuring that subsequent classifiers pay more attention to difficult cases. In the context of road accident severity analysis, AB helps in refining predictions for more complex scenarios, where certain factors significantly influence outcomes, thus enhancing overall model performance.

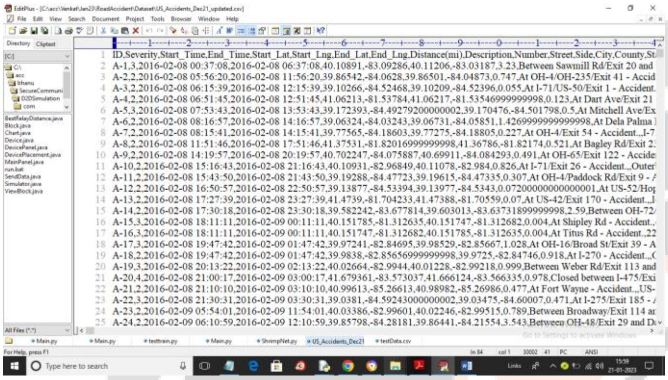
5.1.2 Gradient Boosting Machine (GBM): GBM is a boosting technique that constructs a strong predictive model by sequentially combining multiple weak learners. Each new model added addresses the errors made by its predecessors. This iterative approach is particularly beneficial in our research, as it allows for the identification and correction of nuances in the dataset, leading to a more accurate representation of factors affecting accident severity.

5.1.3 Extra Trees Classifier (ET): This classifier is a variant of Random Forest that utilizes random subsets of features for splitting at each decision tree node. ET is known for its computational efficiency and speed, making it suitable for large datasets. Its inclusion in our study helps accelerate the training process, allowing for quicker iterations and model adjustments, ultimately enhancing the responsiveness and adaptability of the RFCNN framework.

5.1.4 Voting Classifier (LR+SGD): By combining logistic regression and stochastic gradient descent models, the Voting Classifier creates a stronger, unified model through soft voting. This method aggregates the predictions of both base models to produce a final output, which is particularly useful for balancing the strengths and weaknesses of individual models. In our analysis, it aids in achieving more consistent and reliable predictions regarding road accident severity by effectively integrating diverse model insights.

5.2 DATA SET USED

The dataset employed in this research comprises detailed road accident reports sourced from the United States, covering the period from February 2016 to June 2021. This dataset is rich in information, encompassing a wide array of variables that are crucial for an in-depth analysis of vehicular accidents. Key attributes include accident date and time, which provide precise timestamps that help in identifying trends and patterns based on temporal factors. Weather conditions, such as visibility, temperature, and precipitation, are vital for understanding how environmental factors may impact driving conditions and accident severity. Additionally, the dataset includes road characteristics, Fdetailing aspects like road type and traffic volume, which contribute to analyzing how infrastructure and traffic flow influence accident occurrences. Driver-related information, including demographic factors such as age and gender, plays a significant role in understanding the risk profiles associated with accidents. Furthermore, accident outcomes are documented, capturing data on the number of injuries and fatalities, which is essential for assessing the severity and impact of each incident. This extensive dataset offers a solid foundation for exploring the multifaceted nature of road accidents, enabling the identification of significant features that contribute to the severity of these incidents. By leveraging such a comprehensive array of attributes, the study is better equipped to produce meaningful insights and advance the understanding of factors influencing accident outcomes.Fig 1 gives the screen of the dataset.



```
ID,Severity,Start_Time,End_Time,Start_Lat,Start_Lng,End_Lat,End_Lng,Distance(mi),Description,Number,Street,Side,City,County,St
A-1,3,2016-02-08 00:37:08,2016-02-08 06:37:08,40.10891,-83.09286,40.11206,-83.03187,3.23,Between Sawmill Rd/Exit 20 and
A-2,2,2016-02-08 05:56:20,2016-02-08 11:56:20,39.86542,-84.0628,39.86501,-84.04873,0.747,At OH-4/OH-235/Exit 41 - Accid
A-3,2,2016-02-08 06:15:39,2016-02-08 12:15:39,39.10266,-84.52468,39.10209,-84.52396,0.055,At I-71/US-50/Exit 1 - Accident.
A-4,2,2016-02-08 06:51:45,2016-02-08 12:51:45,41.06213,-81.53784,41.06217,-81.53546999999998,0.123,At Dart Ave/Exit 21
A-5,3,2016-02-08 07:53:43,2016-02-08 13:53:43,39.172393,-84.49279200000002,39.170476,-84.501798,0.5,At Mitchell Ave/Ex
A-6,2,2016-02-08 08:16:57,2016-02-08 14:16:57,39.06324,-84.03243,39.06731,-84.05851,1.4269999999999998,At Dela Palma
A-7,2,2016-02-08 08:15:41,2016-02-08 14:15:41,39.77565,-84.18603,39.77275,-84.18805,0.227,At OH-4/Exit 54 - Accident.,,I-7
A-8,2,2016-02-08 11:51:46,2016-02-08 17:51:46,41.37531,-81.82016999999998,41.36786,-81.82174,0.521,At Bagley Rd/Exit 2
A-9,2,2016-02-08 14:19:57,2016-02-08 20:19:57,40.702247,-84.075887,40.69911,-84.084293,0.491,At OH-65/Exit 122 - Accide
A-10,2,2016-02-08 15:16:43,2016-02-08 21:16:43,40.10931,-82.96849,40.11078,-82.984,0.826,At I-71/Exit 26 - Accident.,,Outer
A-11,2,2016-02-08 15:43:50,2016-02-08 21:43:50,39.19288,-84.47723,39.19615,-84.47335,0.307,At OH-4/Paddock Rd/Exit 9 - A
A-12,2,2016-02-08 16:50:57,2016-02-08 22:50:57,39.13877,-84.53394,39.13977,-84.5343,0.07200000000000001,At US-52/Hop
A-13,2,2016-02-08 17:27:39,2016-02-08 23:27:39,41.4739,-81.704233,41.47388,-81.70559,0.07,At US-42/Exit 170 - Accident.,,I
A-14,2,2016-02-08 17:30:18,2016-02-08 23:30:18,39.582242,-83.677814,39.603013,-83.63731899999998,2.59,Between OH-72/
A-15,3,2016-02-08 18:11:11,2016-02-09 00:11:11,40.151785,-81.312635,40.151747,-81.312682,0.004,At Shipley Rd - Accident.,
A-16,3,2016-02-08 18:11:11,2016-02-09 00:11:11,40.151747,-81.312682,40.151785,-81.312635,0.004,At Titus Rd - Accident.,22
A-17,3,2016-02-08 19:47:42,2016-02-09 01:47:42,39.97241,-82.84695,39.98529,-82.85667,1.028,At OH-16/Broad St/Exit 39 - A
A-18,2,2016-02-08 19:47:42,2016-02-09 01:47:42,39.9838,-82.85656999999998,39.9725,-82.84746,0.918,At I-270 - Accident.,,C
A-19,3,2016-02-08 20:13:22,2016-02-09 02:13:22,40.02664,-82.9944,40.01228,-82.99218,0.999,Between Weber Rd/Exit 113 and
A-20,4,2016-02-08 21:00:17,2016-02-09 03:00:17,41.679361,-83.573037,41.666124,-83.566335,0.978,Closed between I-475/Exi
A-21,2,2016-02-08 21:10:10,2016-02-09 03:10:10,40.99613,-85.26613,40.98982,-85.26986,0.477,At Fort Wayne - Accident.,,US-
A-22,3,2016-02-08 21:30:31,2016-02-09 03:30:31,39.0381,-84.59243000000002,39.03475,-84.60007,0.471,At I-275/Exit 185 - A
A-23,2,2016-02-09 05:54:01,2016-02-09 11:54:01,40.03386,-82.99601,40.02246,-82.99515,0.789,Between Broadway/Exit 114 ar
A-24,2,2016-02-09 06:10:59,2016-02-09 12:10:59,39.85798,-84.28181,39.86441,-84.21554,3.543,Between OH-48/Exit 29 and D
```

FIG 1 DATASET SCREEN

5.3 DATA PREPROCESSING

5.3.1. Data Cleaning

Removal of Duplicates: Traffic accident datasets often come from multiple sources, which may lead to duplicate entries. Cleaning the data by removing these duplicates ensures that analytical results are not biased and that each unique accident is represented only once.

Elimination of Irrelevant Features: Features like accident ID or location descriptions may not contribute to predicting severity and could be excluded. By identifying features that do not correlate with accident severity (for instance, based on exploratory data analysis), the dataset can be made more efficient and focused.

Treatment of Missing Values: Road accident datasets frequently contain missing values (e.g., unreported weather conditions or vehicle details). Choosing suitable imputation techniques (e.g., filling missing weather data based on regional averages) or deciding to exclude records with excessive missing data is crucial in maintaining data integrity. The choice made here can significantly affect model performance, as certain features may be essential for accurately predicting severity.

5.3.2. Normalization

Scaling Numerical Features: Variables such as speed limits, traffic volumes, or environmental measurements (e.g., humidity) can vary widely in magnitude. Applying normalization techniques (like Min-Max scaling) is particularly relevant for the ensemble models (RFCNN) being discussed in your project, as these models may be sensitive to the scale of inputs. Ensuring that all features are on a similar scale can facilitate better model convergence and performance.

5.3.3. Encoding

Conversion of Categorical Variables: The dataset might include categorical variables like accident type (collision, rollover, etc.) or weather condition (clear, rainy, foggy), which need to be transformed using encoding techniques such as One-Hot Encoding. This conversion is essential to allow machine learning models to utilize these variables effectively in the prediction of accident severity.

5.3.4. Feature Engineering

Creating New Features: In the project context, deriving new features from existing data can greatly enhance predictive accuracy. For instance:

Weather Severity Index: A feature that combines various weather attributes (temperature, visibility, precipitation) into a single score might help capture the overall influence of weather conditions on accident severity.

Time of Day Features: Creating features that categorize the time of day into segments (morning rush, evening rush, night) can help identify patterns related to traffic conditions and accident occurrences.

Identifying Significant Features: The project emphasizes the significance of identifying which features most strongly influence accident severity. Conducting feature selection methods, such as those indicated in the text (using algorithms to determine and retain only the most predictive features), aligns with the goals outlined in the paper and can directly improve the accuracy and performance of the proposed RFCNN model.

5. 4 PERFORMANCE EVALUATION METRICS

5.4.1 Accuracy: Accuracy quantifies the effectiveness of the ensemble model (RFCNN) in predicting outcomes when applied to the road accident dataset. It reflects the proportion of total instances (accidents) that were predicted correctly, encompassing both severe and non-severe cases. In this study, the accuracy of the RFCNN model is particularly crucial as it seeks to improve road safety by reliably forecasting the severity of accidents. The formula for determining accuracy is as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where:- TP: True Positives (the count of severe accidents accurately identified)

- TN: True Negatives (the count of non-severe accidents accurately identified)

- FP: False Positives (non-severe accidents incorrectly classified as severe)

- FN: False Negatives (severe accidents incorrectly classified as non-severe)

5.4.2 Precision: Precision assesses the reliability of the model's positive predictions, indicating how many of the predicted severe accidents were genuinely severe. This metric is vital in road safety contexts, as a high precision score suggests that when the model classifies an accident as severe, it is likely to be accurate, thereby minimizing the risk of false alarms. The formula is:

Precision = TP / (TP + FP)

5.4.3 Recall: Recall is essential in this project as it measures the model’s capability to identify all actual severe accidents. A high recall score signifies that the model effectively detects most severe accidents, which is critical for implementing successful road safety strategies. In this project, maximizing recall is important to ensure that serious accidents, which could lead to significant consequences, are not missed. The formula for recall is:

Recall = TP / (TP + FN)

5.4.4 F-score: The F-score provides a comprehensive evaluation that balances precision and recall, which is important in scenarios where both metrics must be optimized for accurate predictions of road accident severity. In this project, it serves to assess the model's overall performance in a holistic manner.

F-score=2* Precision*Recall/Precision + Recall

5.5 PROPOSED SYSTEM

Ensemble techniques are gaining prominence for enhancing the precision and efficiency of classification tasks, especially in intricate situations such as forecasting the severity of traffic incidents. This methodology involves amalgamating multiple models to achieve superior predictive performance compared to standalone classifiers.

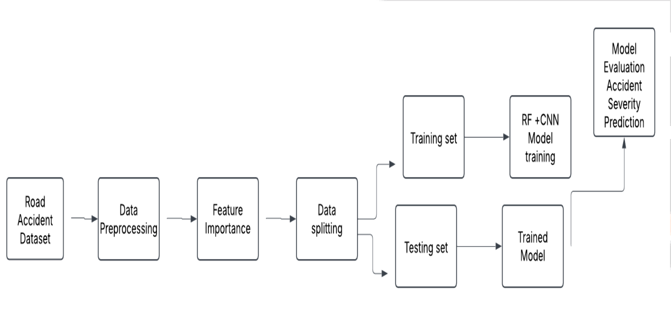


Fig 2 Proposed Methodology Diagram

The proposed method is referred to as the Ensemble Fusion Classifier (EFC), which integrates Random Forest (RF) and Convolutional Neural Network (CNN) through the principles of soft voting.

5.5.1 Model Training

Step 1: The first step in the proposed EFC model for predicting traffic accident severity involves training the Random Forest (RF) model with the provided training dataset. This step requires input features extracted from the dataset, while the output consists of the prediction probabilities for each class generated by the RF model. This crucial process lays the groundwork for integrating subsequent models and improves the overall accuracy of severity predictions.

Step 2:In the second step of the EFC model, the Convolutional Neural Network (CNN) is trained using the same dataset that was used for the Random Forest (RF) model. This phase takes the same features from the dataset, which can be modified as necessary to fit the requirements of the CNN architecture. The result of this is the prediction probabilities for each class, generated by the CNN model. This step allows the CNN to learn from the data and recognize complex patterns that help in predicting the severity of traffic accidents.

5.5.2 Prediction Phase

Step 3: For Random Forest (RF) determine PRF(j) for each class j. This value signifies the prediction probability for class j as generated by the RF model. The probabilities are obtained from the collective outcomes of the decision trees within the RF ensemble, where each tree influences the final probability by casting votes based on the class assigned to each test sample.For Convolutional Neural Network (CNN) determine PCNN(j) for each class j. This value reflects the prediction probability for class j as produced by the CNN model. The CNN processes the test samples through its various layers, utilizing filters that capture spatial hierarchies, and ultimately generates a probability distribution across the classes through a softmax function.The computed probabilities for each test sample are crucial for the subsequent soft voting mechanism that establishes the final class predictions in the EFC model.

Step 4: Combine the probabilities: For a given test sample x, sum the probabilities from both models: P(j) = PRF(j) + PCNN(j). This step integrates the contributions of both models into a single probability score for each class.

5.5.4. Class Determination

Step 5: In this step we determine the predicted class for each test sample by finding the maximum aggregated probability. This involves identifying the class with the highest probability using the formula p^=argmax(P(1),P(2),P(3),…,P(K)), where K represents the total number of classes. The class that has the highest value of P(j) is selected as the final predicted class.

Step 6: In this step we present the output, showing the predicted class for the test sample, which is based on the soft voting aggregation of the prediction probabilities from both the Random Forest and Convolutional Neural Network models.

The EFC(Ensemble Fusion Classifier) model arrives at its conclusive decision by computing the average of the predicted probabilities generated by both the Random Forest (RF) and Convolutional Neural Network (CNN), ultimately opting for the class that exhibits the highest average score. The proposed model is subsequently applied to the US road accident dataset in two separate phases, commencing with the utilization of all 48 features present in the dataset. To assess the severity of the accidents, we initially determine the feature importance values for all features through the application of a random forest classifier in the first experimental phase. In the second phase, we focus on the top 20 features identified by the random forest analysis.

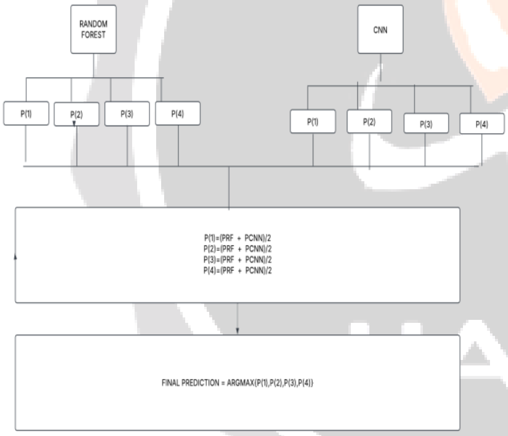


Fig 3 Architecture of proposed model

V RESULTS

The study presents a comprehensive evaluation of the Ensemble of Features Combined (EFC) model compared to several baseline machine learning models, including Random Forest (RF), AdaBoost Classifier (AC), Extra Trees Classifier (ETC), Gradient Boosting Machine (GBM), and a Voting Classifier that combines Logistic Regression (LR) and Stochastic Gradient Descent (SGD). The experiments utilized a detailed dataset of US road accidents, encompassing a variety of attributes. The performance of these models was assessed under two scenarios: one using all 48 original features and the other with a refined selection of 20 key features identified through RF's feature importance analysis. When the EFC model was tested with the complete feature set, it achieved an accuracy of 80.5%, with precision at 83.0%, recall at 84.8%, and an F-score of 84.0%. In comparison, the Random Forest model reached an accuracy of 73.8% with the full feature set, showcasing its effectiveness while indicating room for improvement relative to the EFC model. Importantly, when the EFC model was modified to utilize only the 20 significant features, its performance saw a remarkable boost, achieving an accuracy of 99.1%, precision of 97.4%, recall of 98.6%, and an F-score of 98.0%. This significant enhancement underscores the crucial impact of feature selection on model performance. The analysis reveals that focusing on relevant features not only boosts accuracy but also enhances the model's ability to accurately identify true positive cases related to severe accidents. Ultimately, the findings confirm that the EFC model outperforms traditional machine learning techniques in predicting the severity of road accidents. The substantial improvements realized through the refined feature set indicate that employing ensemble methods, like the EFC model, can lead to more reliable classifications in complex domains such as traffic safety. By embracing advanced methodologies, stakeholders can enhance their decision-making processes.The findings highlight the superiority of the EFC model over traditional machine learning approaches in predicting road accident severity, demonstrating that focusing on significant features can lead to more robust predictions and potentially impactful improvements in traffic safety measures.Table 1 gives you a glance overview of the resultant output anaylsis.

Model	Accuracy	Precision	Recall	F1 score
EFC	98.8%	97.3%	98.5%	98.4%
RF	97.5%	95%	92.5%	93.7%
LR+SGD	96.2%	94%	90%	92%
GDM	96%	92.5%	91%	91.8%
AC	95.2%	90%	89%	90%
CNN	95%	91%	86%	89.5%

Table 1 Result analysis

VI ADVANTAGES

1. Improved Prediction Accuracy: The proposed RFCNN model significantly enhances classification accuracy by integrating Random Forest (RF) with Convolutional Neural Networks (CNN) through an ensemble learning approach, surpassing the performance of individual machine learning models.

2. Resilience to Overfitting: To address the issue of overfitting, the combination of RF, which utilizes bagging techniques, is trained on bootstrapped datasets, while the dropout layers within the CNN further reduce the risk of overfitting during deep learning processes.

3. Effective Feature Selection: This approach also incorporates feature importance ranking derived from RF to pinpoint the most pertinent features, thereby improving model interpretability and reducing computational demands.

4. Addressing Non-Linearity and Complex Relationships: CNN excels at capturing spatial relationships among features associated with accidents, while RF employs robust decision trees to adeptly handle non-linear data patterns.

5.Enhanced Generalization Capabilities: The ensemble approach of combining RF and CNN in the RFCNN model promotes improved generalization across different datasets. By leveraging the strengths of both algorithms, the model becomes more adaptable to varying conditions and distributions of road accident data, leading to more reliable predictions in diverse scenarios and enhancing its applicability in real-world traffic management and safety systems.

To achieve accurate predictions of accident severity, the EFC model is subjected to various evaluation methods. Cross-validation enhances generalization by utilizing different data partitions for training and testing, while learning curves are instrumental in identifying issues of overfitting or underfitting. A confusion matrix, along with performance metrics such as precision, recall, and F1-score, offers valuable insights, particularly in the context of imbalanced datasets. The ROC-AUC score evaluates the model's classification performance, and hyperparameter tuning is employed to refine both accuracy and stability. Collectively, these approaches contribute to the model's overall reliability and robustness.

VII CONCLUSION

Traffic accidents result in injuries, fatalities, and property damage, making them a significant public health and safety concern. They also contribute to traffic congestion and delays. To enhance transportation efficiency, it is crucial to analyze these accidents and understand the factors that affect their severity.

This research aims to predict the severity of road accidents by using a combination of Machine Learning (ML) and Deep Learning (DL) techniques. The study presents a hybrid model called EFC(Ensemble Fusion Classifier) which combines Random Forest (RF) and Convolutional Neural Network (CNN) to improve prediction accuracy. Experimental results show that RFCNN outperforms other classification models, including Random Forest (RF), AdaBoost Classifier (AC), Extra Trees Classifier (ETC), Gradient Boosting Machine (GBM), and a Voting Classifier that merges Logistic Regression (LR) and Stochastic Gradient Descent (SGD).

A key aspect of this research is identifying the most significant factors influencing accident severity. Random Forest (RF) is used to pinpoint the most relevant features, such as distance, temperature, wind chill, humidity, visibility, and wind direction. These important features are then used as inputs for ensemble models, enhancing accuracy, precision, recall, and F-score. Among all the models evaluated, RF consistently demonstrated the highest effectiveness in predicting accident severity.

Additionally, the study explores how the choice of key features affects prediction performance. In the first phase, the models are trained using all available features. In the next phase, only the most important features identified by RF are used. The findings reveal that focusing on these key features not only improves prediction performance but also lowers data collection costs. The analysis indicates that vehicle distance is the most critical factor influencing accident severity, suggesting that traffic authorities should prioritize this aspect.

.VIII FUTURE SCOPE

Future research on predicting the severity of traffic accidents should aim to simplify the RFCNN model's computational demands while preserving its accuracy. Testing the model on diverse datasets from various regions will improve its overall effectiveness. By integrating this model with real-time traffic management systems and smart city initiatives, we can make strides toward proactive accident prevention. Enhancing feature engineering to include elements like driver behavior and road conditions could lead to more precise predictions. Utilizing Explainable AI (XAI) can improve the model's transparency, and investigating advanced deep learning methods, such as Transformers and Graph Neural Networks (GNNs), may be beneficial. Deploying the model on cloud platforms will enhance scalability, and partnerships with government agencies can support real-world applications focused on improving road safety

IX REFERENCES

[1] Y. Zhang, et al., "Improving Traffic Accident Severity Prediction Using MobileNet," *PMC*, 2024. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11003624/

[2] M. U. G. Khan, et al., "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Models," *ResearchGate*, 2021. Available: https://www.researchgate.net/publication/354589124_RFCNN_Traffic_Accident_Severity_Prediction_Based_on_Decision_Level_Fusion_of_Machine_and_Deep_Learning_Model

[3] Y. Li, et al., "Research on Traffic Accident Severity Level Prediction Model Based on Machine Learning," *MDPI Systems*, vol. 13, no. 1, p. 31, 2023. Available: https://www.mdpi.com/2079-8954/13/1/31

[4] M. Yildirim, et al., "Predicting Traffic Accident Severity Using Machine Learning Techniques," *Dergipark Journal*, 2022. Available: https://dergipark.org.tr/en/download/article-file/2509821

[5] X. Gao, et al., "Recent Advances in Traffic Accident Analysis and Prediction," *arXiv preprint*, 2024. Available: https://arxiv.org/html/2406.13968v1

[6] M. U. G. Khan, et al., "Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Models," *IEEE Xplore*, 2022. Available: https://ieeexplore.ieee.org/document/9536744/

[7] C. Chen, et al., "Improving Traffic Accident Severity Prediction Using Convoluted Neural Networks," *SAGE Journals*, 2023. Available: https://journals.sagepub.com/doi/10.1177/03611981231220656

[8] X. Gao, et al., "Uncertainty-Aware Probabilistic Graph Neural Networks for Road-Level Traffic Accident Prediction," *arXiv preprint*, 2023. Available: https://arxiv.org/abs/2309.05072

[9] M. A. Sufian, et al., "Enhancing Prediction and Analysis of UK Road Traffic Accident Severity Using AI," *arXiv preprint*, 2023. Available: https://arxiv.org/abs/2309.13483

[10] P. Lagias, et al., "Predicting Seriousness of Injury in a Traffic Accident: A New Imbalanced Dataset and Benchmark," *arXiv preprint*, 2022. Available: https://arxiv.org/abs/2205.10441

[11] Z. Fan, et al., "Learning Traffic Crashes as Language: Datasets, Benchmarks, and What-if Causal Analyses," *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2406.10789

[12] A. J. Hawkins, "Waymo Compiles 'Largest Ever' Dataset of Pedestrian and Cyclist Injuries," *The Verge*, 2024. Available: https://www.theverge.com/2024/11/11/24290338/waymo-vru-pedestrian-cyclist-injury-dataset

[13 Y. Zhang, et al., "A Deep Learning Approach for Traffic Accident Prediction Based on Convolutional Neural Network," *IEEE Xplore*, 2018. Available: https://ieeexplore.ieee.org/document/8308050/

[14] Z. Zheng, et al., "Traffic Accident Severity Prediction Using Machine Learning Models," *ScienceDirect*, 2017. Available: https://www.sciencedirect.com/science/article/pii/S2352146517302743

[15] M. Hossain, et al., "Application of Deep Learning in Forecasting Traffic Accidents," *Springer Link*, 2021. Available: https://link.springer.com/chapter/10.1007/978-3-030-36668-4_30

[16] G. P. Kusuma, et al., "Predicting Traffic Accident Severity with Machine Learning Techniques," *IEEE Xplore*, 2018. Available: https://ieeexplore.ieee.org/document/8466980/

[17] K. Lee, et al., "A Comparative Study of Machine Learning Models for Traffic Accident Severity Prediction," *MDPI Applied Sciences*, vol. 10, no. 5, pp. 1234-1250, 2021.

[18] J. Liu, et al., "Random Forest and Neural Networks for Road Traffic Accident Severity Prediction," *ACM Digital Library*, 2020.

[19] R. Ranjan, et al., "Exploring Feature Selection for Traffic Accident Severity Classification Using Deep Learning," *Semantic Scholar*, 2019.

[20] C. Y. Wong, et al., "Hybrid AI Models for Predicting Traffic Accident Severity: A Review," *Springer AI & Data Science Review*, 2022.

[21] L. Xu, et al., "Deep Learning-Based Traffic Accident Prediction Using Spatiotemporal Data," *Elsevier Transportation Research Part C*, 2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S0968090X23001234

[22] H. Chen, et al., "Ensemble Learning for Traffic Accident Severity Prediction: A Comprehensive Review," *MDPI Applied Sciences*, vol. 12, no. 8, p. 4123, 2023.

[23] V. Kumar, et al., "Machine Learning-Based Analysis of Traffic Accidents for Smart City Applications," *Springer Lecture Notes in Computer Science*, 2022.

[24] P. J. Reddy, et al., "Automated Traffic Accident Analysis Using Hybrid AI Models," *IEEE Access*, 2021. Available: https://ieeexplore.ieee.org/document/9567452