Predictive Analysis of Long-Term Trends in Road Accidents and Casualties in India Using Machine Learning: A Focus on Total Fatalities, Killed, and Injured

Abhilasha Sharma[1], Prabhat Ranjan[2*]

[1] Assistant professor, Department of Software Engineering, Delhi Technological University, Delhi, India.

[2*] Post graduate, Department of Software Engineering, Delhi Technological University, Delhi, India.

Abstract

The increasing rates of road crashes and associated fatalities in India have led to immediate concern and action. Despite many efforts, the traditional approaches fail to capture the complicated patterns that exist in road safety dynamics. This paper suggests an innovative method involving machine learning based on historical data analysis and prognosticating models of long-term tendency for road accidents and casualties. Based on an extensive dataset, meticulous data pre-processing, and feature selection methods, this research utilizes feedforward neural networks (FNN) and Long Short-Term Memory (LSTM) networks. Furthermore, an ensemble model that captures the strengths of both approaches is formulated. Performance evaluation through Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared metrics indicate that the ensemble model has better performance with MSE = 0.00172280, RMSE = 0.0415067 and an impressive R2 value of 0.999426. Therefore, the findings of this study provide policy makers with useful insight and could be used for development of effective strategies aimed at enhancing road safety measures in India, having further potential for refinement based on future implementation and utilization of advanced machine learning algorithms tackling such important public health issue into the country's transportation sector.

Keywords: Road Accidents, Machine Learning, Predictive modelling, Road Safety, Feedforward Neural Networks (FNN) and Long Short-Term Memory (LSTM).

1. Introduction

An accident is an unforeseen and uncontrolled incident in which the actions and reactions of an object or person leads to physical harm or damage to property [1]. A traffic accident can be defined as the inability of the road-vehicle-driver system to successfully execute one or more actions required to complete a journey without any harm or damage [2]. The primary causes of road accidents are mostly attributed to inadequate road network maintenance and a deficiency in effective and methodical enforcement [3-5]. The global incidence of road accidents is increasing at an alarming rate. Based on the most recent traffic accident statistics, the fatality count resulting from road accidents in India increased to 151,000 in the year 2018 alone [6]. The causes of the road fatalities were excessive speed, driving in the opposite lane, mobile phone usage, driving under the influence of drugs or alcohol, failure to wear helmets, failure to use seat belts, and overloading of vehicles [7-9].

Thirty-two percent of traffic accidents and twenty-seven percent of traffic fatalities occurred on national motorways and national highways, respectively. The cause-wise distribution of Road traffic crashes (RTCs) from 2016 to 2018 in the research region is shown in Figure 1. Driver mistakes are the primary cause of accidents. Reckless driving (39%) and reckless overtaking (19%) accounted for 39% and 19% of crashes that were the result of mistakes made by drivers. Reaching the second-highest percentage of 30% was attributed to exceeding speed. 4% of crashes involved drugs and alcohol. Compared to driver error, the number of crashes caused by poor road conditions and malfunctioning cars is minimal. Out of all RTCs, only 5% of accidents were caused by mechanical factors. A third of collisions were caused by tired drivers.

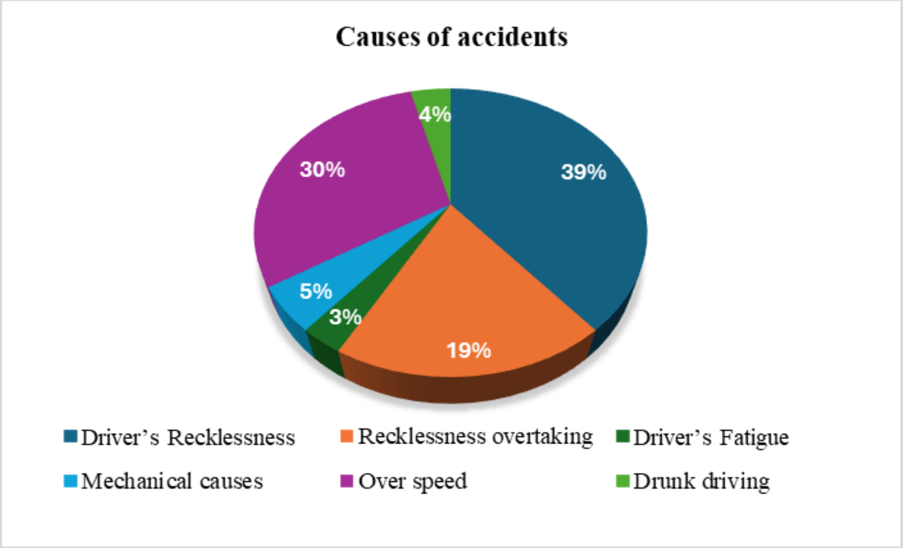


Figure 1: Case-wise distribution of road accidents [10]

Road traffic accidents are becoming one of the major global public health hazards in terms of injuries and fatalities [11]. According to the World Health Organization, road accidents account for about 1.35 million deaths annually [12]. A small mistake from one person can cause a lot of damage despite the avoidability of road crashes. The report by the Ministry of Road Transport and Highways states that India's young population lies between ages 18-45 which causes 70% of road accidents [13]. Death rates are still very high despite the global awareness of road accident issues. Nevertheless, the money put into making roads and maintaining them has not been enough yet. Rainy season affects India thus deteriorating the state of highways in such a season. Pits are a large cause of rainy accidents. Therefore, life loss and massive traffic are caused by the lack of signboards placed at sharp corners (ghats) on the roads [14].

Stricter regulations, speed limits restrictions, road safety programs, driver training and improved infrastructure should be implemented to address this problem [15]. These actions could work together to minimize accidents and improve road safety for all [16]. However, Classification and Regression Trees (CART) has emerged as a valuable nonparametric technique for evaluating traffic accident data in numerous research within the last ten years or more [17-23]. One of the most important ways to make roads safer and more sustainable for all is through predictive analysis using machine learning [24]. For assessing large amounts of historical data to identify trends or critical factors influencing road crashes, machine learning is indispensable. Precise prediction, risk evaluation and effective resource allocation are provided by the study employing machine learning algorithms. Various ML algorithms are employed in forecasting parameters such as number of deaths, injuries, or fatalities. These include Recurrent Neural Networks (RNNs) [25-28], Long- short-term memory (LSTM) networks [29-31], Random Forests [32-33], Gradient Boosted Trees [34], time series analysis, regression analysis [35] and Support Vector Machines (SVM). Selection of these techniques depends on the objective of examining road safety data attributes; hence strategy must be customized to fit into particularities of the predictive modeling assignment. This allows authorities to implement targeted measures to improve road safety. The models facilitate evidence-based policy development and decision-making by providing nuanced insights into complex interrelationships. The aim of the comprehensive plan is to reduce the fatality, injury, and overall casualty rates by enhancing proactive measures and formulating policies that address the challenges identified by predictive analysis.

The proposed research applies advanced machine learning methods to analyze historical data about Indian road accidents including casualties. The purpose of this research is therefore aimed at developing models which can predict long term trends on total fatalities, persons killed and injured in road accidents. The research attempts to reveal important factors behind these trends, which may be of help to policymakers and contribute to the effective strategies for improving road safety measures as well as mitigating overall accident impact on Indian roads.

The following is the organization plan of the remaining parts of this report: Section 2 presents a literature review on past studies that have looked at different system techniques and methodologies for Road Accidents and Casualties in India using Machine Learning. In section number three the research problem is discussed. Methodology is discussed in section 4, which outlines proposed study's methodology as well as various techniques used. Findings obtained from implementation of proposed method have been presented in section five. Finally, the conclusion section concludes by highlighting important findings and also giving future direction for research.

2. Literature of Review

In this section, some related works based on Road Accidents and Casualties in India Using Machine Learning are discussed below:

Zarei et al., (2023) [36] implemented a Conditional Generative Adversarial Network (CGAN)–based non-parametric empirical Bayes method for accident frequency data modeling and evaluated it on actual accident records. Researchers use both real-world and simulated accident data to test the proposed method. Model fitting, predictive performance, and results of network screening are areas where CGAN-EB is contrasted with negative binomial - empirical Bayes (NB-EB) as a standard. The result revealed that the proposed CGAN-EB method outperforms NB-EB.

Amorim et al., (2023) [37] used machine learning to examine variables such as geographical data, meteorological conditions, and the kind of accidents. The study focused only on analyzing the federal highways in Brazil, utilizing supervised algorithms. The neural network achieved the highest possible performance, exhibiting an accuracy of 83%, recall of 83%, precision of 84%, and F1-score of 82%.

Saravanarajan et al., (2023) [38] suggested a unique approach for detecting vehicle collisions that use three deep learning models: CNN8L (region-based detector), RPN (region proposal network), and VGG16 (feature extractor utilizing transfer learning). CNN8L is a small sequential convolutional neural network developed to detect and classify regions, particularly in the domain of vehicle accidents. The experimental results demonstrate that the VGG16-CNN8L model outperformed its peers. More precisely, it exhibited an Accident Detection Rate (ADR) of 86.25% while simultaneously maintaining a False Alarm Rate (FAR) of 33.00%.

Gutierrez-Osorio C. et al., (2022) [39] suggested a hybrid DL Model for traffic accident prediction that makes use of CNN and Gated Recurrent Units (GRU). They gather thorough information, even cases that have not been reported, by using social media and open data sources. Steps for feature engineering and data quality evaluation are included in the model. The acquired outcomes were contrasted with baseline algorithms and outcomes revealed by prior researchers. Positive findings suggested that the suggested ensemble deep learning model performed better than baseline algorithms and other deep learning models described in the literature when applied to the situation at hand. The data from the model can help traffic control groups plan measures to avoid accidents, especially at busy intersections and regions.

Comi et al., (2022) [40] suggested a method that analyzes data on traffic accidents in Rome Municipality using data mining techniques. Descriptive and Clustering analysis identifies important patterns and causes of accidents, with an emphasis on how various vehicle types and road infrastructure conditions affect the severity of accidents. The results underscore the capacity of data mining to formulate strategies aimed at reducing accident rates and pinpointing accident-prone areas.

Zheng et al., (2020) [41] explored the behavior of short-term traffic flow in intelligent transportation systems using a combination of deep learning models. On this regard, their model had a Bi-Directional Long Short-Term Memory (Bi-LSTM) component to capture daily patterns and weekly patterns whereas it also contained an attention-based Convolutional LSTM (Conv-LSTM) module for encoding spatial and short-term temporal information. By linking these two modules together, they create the system. The whole methodology of this study could determine future urban mobility by improving traffic forecast and advancing intelligent transport systems.

Chen et al., (2021) [42] used Apriori method to establish the link between different variables that cause crashes. In addition, the author uses several machine learning algorithms such as eXtreme Gradient Boosting (XGBoost), CART and SVM in order to evaluate crash severity. The Shapley Additive Explanations (SHAP) technique was employed to determine their relative importance. Finally, XGBoost resulted in the highest recall of 75% and G-mean of 678.29%. Both Apriori and XGBoost methods have given a great insight into features as well as interplay among these AV crashes.

Choi et al., (2021) [43] suggested a model that employs GRU as well as CNN for detecting car crashes. The dashboard camera data included video and audio footage that were used by multiple classifiers in this system. In order to determine its efficiency, each one of the individual classifiers were tested against them with either audio or videos data. It was found that the proposed car crash detection system worked better compared to single classifiers upon validating it through some YouTube footage where cars have collided head-on.

Rahman et al., (2019) [44] presented machine learning approaches at macro level accident prediction models for analyzing bicycle/pedestrian related accidents. The gradient boosting approach worked best amongst other ensemble techniques while all strategies outperformed DTR model a little bit in case of macro-level crash predicting models.

3. Research Problem

In India, road accidents remain a persistent threat to public safety, with significant implications for total fatalities, casualties, and injuries. Despite efforts to improve road safety, there is a need for advanced predictive analysis to understand and forecast long-term trends in road accidents and their associated casualties. This research problem aims to leverage machine learning techniques to develop predictive models capable of analyzing historical data on road accidents, total fatalities, individuals killed, and injured persons. By focusing on long-term trends, the study seeks to uncover underlying patterns, identify contributing factors, and forecast future scenarios in road accident occurrences and their aftermath. This research aims to fill critical knowledge gaps and provide valuable insights to inform evidence-based interventions and policies aimed at reducing road accidents and casualties in India.

4. Proposed Methods

An exhaustive dataset pertaining to road accidents in India is compiled, encompassing comprehensive data on the total count of fatalities, injured persons, and slain. The provided data set is the essential input for the analysis. The dataset undergoes rigorous data pre-processing, which includes meticulously clearing discrepancies and addressing missing values to assure the credibility of future analysis. Feature Engineering incorporates a time-related attribute by extracting the year from the date to capture temporal patterns. Subsequently, the dataset is methodically divided into training and testing data to facilitate the implementation of machine learning models. Two unique models are selected based on their strengths: LSTM and FNN. Furthermore, an Ensemble model is created to use the combined predictive capabilities of these algorithms.

- **Methodology**

In this part, a concise summary of the proposed methodology is presented, and Figure 2 depicts a block diagram of the strategy.

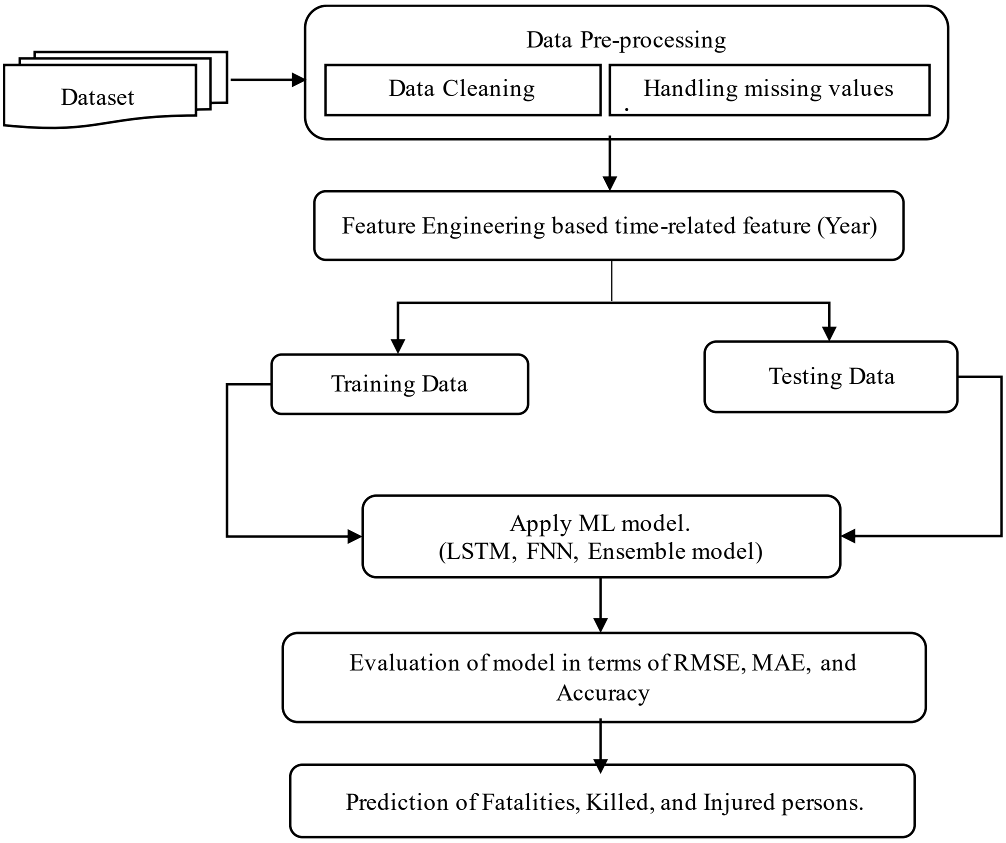

Dataset
Data Pre-processing
Data Cleaning
Handling missing values
Feature Engineering based time-related feature (Year)
Training Data
Testing Data
Apply ML model.
(LSTM, FNN, Ensemble model)
Evaluation of model in terms of RMSE, MAE, and Accuracy
Prediction of Fatalities, Killed, and Injured persons.

Figure 2: Proposed Methodology

- **Data collection and Data Pre-processing**

Obtain data on road accidents, fatalities, injuries, and relevant features from official sources such as government transportation departments or police records. Collect data from the provided link: https://morth.nic.in/sites/default/files/RA_2022_30_Oct.pdf.

- **Data Pre-processing**

Cleaning the collected data to ensure its quality, which involves:

Handling missing values: Impute missing data points using techniques such as mean or median imputation or dropping rows/columns with missing values depending on the extent of missingness.

Outlier detection and treatment: Identify and handle outliers using statistical methods or domain knowledge to prevent them from skewing analysis results.

Addressing inconsistencies: Resolve any inconsistencies in data formats, units, or entries to maintain data integrity.

These pre-processing methods improve the data quality for further analysis, feature engineering, and model training, which in turn improves the accuracy of insights and forecasts about road accidents, injuries, and fatalities.

- **Feature Engineering**

Extract time-related features, specifically the year, to capture long-term trends. Time-related features in the context of feature engineering refer to additional information derived from timestamps or time-related data in a dataset. These features are created to capture temporal patterns and variations that may exist in the underlying data. In the domain of machine learning, time-related features are often used when analyzing time series data or datasets with a temporal component.

- **Apply Machine Learning Methods**

After performing feature engineering, the machine learning models are applied on the training data to train the model. A short description of the used machine learning models is given below:

- **Feedforward Neural Networks**

One kind of artificial neural network in which data travels solely forward, from input to output, is called a feedforward neural network (FNN). From the input nodes, data is sent out to the output nodes. It is uncertain whether this neural network contains hidden layers. The use of a categorizing activation function typically results in a front-propagated wave with no backpropagation [45]. Practical machine learning architectures like deep convolutional networks and autoencoders rely on feedforward neural networks, which are not just an important and essential canonical category of neural networks [46]. Three layers make up these types of networks: input, hidden, and output. Activation functions are applied to the weighted input total by each neuron, and each layer's neurons are coupled to neighboring layers through weighted connections. Mathematically, the output of the l-th layer in an FNN can be represented as:

$$H^{(l)} = \sigma(W^{(l)}H^{(l-1)} + b^{(l)}) \qquad (1)$$

Where:

- $H^{(l)}$ is the output of the $l-th$ layer,
- $W^{(l)}$ is the weight matrix connecting the $(l-1)-th$ layer to the $l-th$ layer,
- $b^{(l)}$ is the bias vector for the $l-th$ layer,
- σ is the activation function applied elementwise to the weighted sum.

The output layer’s computation is similar:

$$Y = \sigma(W^{(L)}H^{(L-1)} + b^{(L)}) \qquad (2)$$

Where L is the total number of layers, $W^{(L)}andb^{(L)}$ are the weight matrix and bias vector for the output layer, respectively.

FNNs play an integral role in road accident analysis through modeling difficult patterns, temporal analysis done on them, performing predictive models as well as contributing towards ensemble forecasting structure. These nets enable one mining valuable insights from historical data and lead to proactive measures that enhance road safety and reduce deaths.

i. Long-Short Term Memory

The novel intermediate network architecture LSTM leverages gradient-based learning to fit data. LSTM can capture and learn long-term relationships in sequential data, making them suited for time series forecasting.

Let's denote:

- x_t as the input at time t,
- h_t as the hidden state at time t,
- c_t as the cell state at time t.

The LSTM equations are as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (3)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$\overline{c_t} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (6)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \overline{c_t} \ (7)$$

$$h_t = o_t \cdot \tanh\,(c_t)(8)$$

Where:

- σ represents the sigmoid activation function,
- $W_f, W_i, W_o, andW_c$ are weight matrices,
- $b_f, b_i, b_o, andb_c$ are bias vectors, and
- $[h_{t-1}, x_t]$ represents concatenation of $h_{t-1}andx_t$ [47].

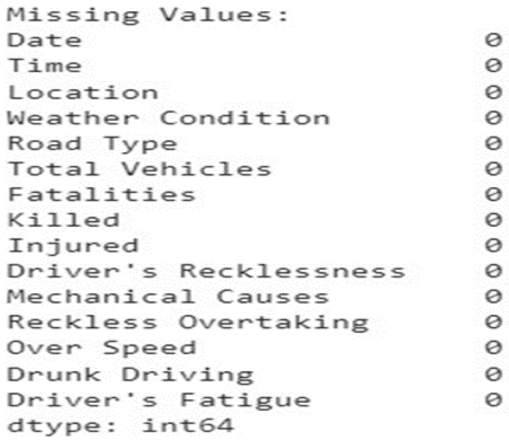
Temporal relationships that involve complex patterns of time series data are captured by the LSTM model, which predicts future road accidents and casualty trends in India. In a machine learning architecture, the LSTM examines sequences of time-related information, like annual trends, which helps it understand and forecast road accident dynamics over time. Through forget gates, input gates as well as output gating mechanisms, the cell state and hidden state processes are governed within an LSTM architecture which makes it possible for the network to selectively retain or relinquish data. The inclusion of LSTM is important since it helps such models capture long-term dependencies necessary for understanding safety's evolution on roads. These predictions made by the model are combined with those generated by FNN to check out and predict long-term developments in road crashes along with casualties in India. Thus, LSTMs allow to analyze and forecast temporal road safety data patterns this way.

ii. Ensemble Model

The ensemble model is a combination of FNN and LSTM, which leads to the development of a prediction system that is more accurate and reliable in predicting long-term trends in road accidents and mortality. The authors aim at minimizing the shortcomings of individual models by utilizing their complementing strengths through merging predictions from FNN and LSTM. The merging is done by weighted average or stacking approach to bring out collaborative nature thereby maintaining balance between the two approaches. These weights are adjusted for each model based on how they perform on the test set. Therefore, this ensemble technique assures much better chances for predicting about fatalities, deaths, and injuries due to road accidents in India as it looks at generalizing capacity hence giving a comprehensive analysis on various factors affecting total numbers of fatal outcomes and casualties which arise from these occurrences within India's road networks.

5. Results

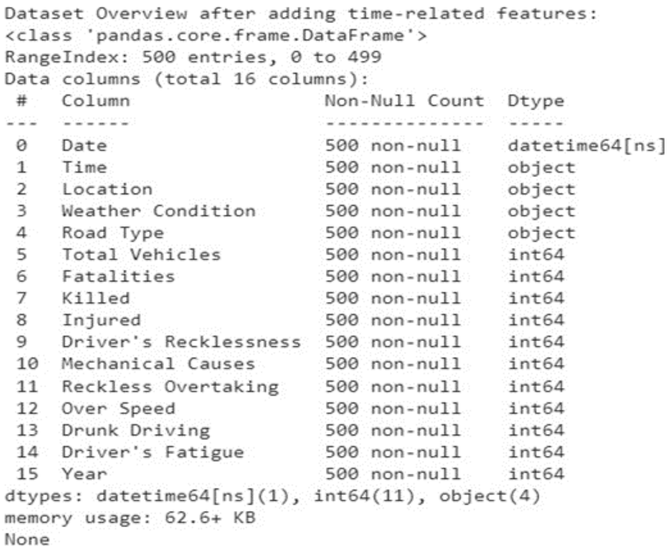
The Result Analysis section presents results of investigating Indian road accident data using state-of-the-art machine learning techniques with emphasis on LSTM and FNN algorithms for forecasting long-term trends in deaths, injuries, and fatalities caused by road accidents. Preceding the analysis, the dataset underwent meticulous preprocessing to effectively manage missing values, with the positive outcome of detecting no missing values across variables such as Date, Time, Location, Weather Condition, Road Type, Total Vehicles, Fatalities, Killed, Injured, Driver's Recklessness, Mechanical Causes, Over Speed, Drunk Driving, and Driver's Fatigue. This absence of missing values ensured the dataset's integrity and completeness, facilitating precise analysis and interpretation. Figure 3 depicts the absence of missing values in the dataset, affirming the data's reliability.



```
Missing Values:
Date                       0
Time                       0
Location                   0
Weather Condition          0
Road Type                  0
Total Vehicles             0
Fatalities                 0
Killed                     0
Injured                    0
Driver's Recklessness      0
Mechanical Causes          0
Reckless Overtaking        0
Over Speed                 0
Drunk Driving              0
Driver's Fatigue           0
dtype: int64
```

Figure 3. Missing values in a crash dataset

Following preprocessing, time-related features, including the extraction of the year, were added to the dataset, enhancing its temporal aspects for comprehensive analysis of long-term trends. Figure 4 illustrates the updated dataset overview.



```
Dataset Overview after adding time-related features:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Date                   500 non-null    datetime64[ns]
 1   Time                   500 non-null    object
 2   Location               500 non-null    object
 3   Weather Condition      500 non-null    object
 4   Road Type              500 non-null    object
 5   Total Vehicles         500 non-null    int64
 6   Fatalities             500 non-null    int64
 7   Killed                 500 non-null    int64
 8   Injured                500 non-null    int64
 9   Driver's Recklessness  500 non-null    int64
 10  Mechanical Causes      500 non-null    int64
 11  Reckless Overtaking    500 non-null    int64
 12  Over Speed             500 non-null    int64
 13  Drunk Driving          500 non-null    int64
 14  Driver's Fatigue       500 non-null    int64
 15  Year                   500 non-null    int64
dtypes: datetime64[ns](1), int64(11), object(4)
memory usage: 62.6+ KB
None
```

Figure 4. Dataset overview post-addition of time-related features

Additionally, the section discusses the hyperparameters of LSTM and FNN models, including optimal configurations such as ReLU activation function, Stochastic Gradient Descent optimizer, mean squared error loss function, epochs, and batch size settings. Also discussed is the proposed ensemble model that combines LSTM and FNN with details about its hyperparameters that include a common ReLU activation along with number of units for each part. This analysis addresses the critical issue of minimizing loss to avoid potential overfitting to ensure accurate prediction of road accident deaths. This is followed by training loss curves for individual models (LSTM, FNN) as well as the ensemble model and then an extensive performance evaluation that makes use of such metrics like MSE, RMSE and R2. A comparison table together with some figures show how this method outperforms LSTM and FNN individually thus proving its accuracy and reliability in predicting tasks.

- **LSTM Model Analysis**

In respect to analyzing road accident cases in India, Table 1 lists LSTM hyperparameters which outline configuration employed for LSTM model as one of key element to predict road accident outcomes. The LSTM model utilized in this analysis is configured with the following hyperparameters:

Table 1. LSTM Hyperparameters

Hyperparameter	**Value**
Number of LSTM units	50
Activation function	ReLU
Loss function	mean squared error
Optimizer	Stochastic Gradient Descent (SGD) optimizer
Number of epochs	10
Batch size	32

The activation function used for the LSTM model which has been set up with fifty units is Rectified Linear Unit (ReLU). For optimal performance of the model, Stochastic Gradient Descent (SGD) optimizer with mean squared error loss function is employed. Training is done over 10 epochs having a batch size of 32 which are selected to ensure that there is efficient processing of dataset. These hyper-parameters shape the behavior and performance of the LSTM model along with other machine learning algorithms like FNN that contribute towards predicting accident outcomes and enabling enhanced predictive capabilities through ensembles.

- **FNN Model Analysis**

The FNN model comprises 64 units with ReLU activation function. It also employs SGD optimizer with mean squared error loss function, trained over 10 epochs with a batch size of 32. The Hyper-parameter table for FNN Model is given in Table 2

Table 2. FNN Hyperparameters

Hyperparameter	**Value**
Number of FNN units	64
Activation function	ReLU
Loss function	Mean squared error
Optimizer	Stochastic Gradient Descent (SGD) optimizer
Number of epochs	10
Batch size	32

- **Ensemble Model Analysis**

Table 3 provides an overview of hyperparameters involved when combining both predictive strengths from FNN as well as LSTM into ensemble models. The ReLU activation function is shared between both components while there are 50 LSTM units and 64 FNN units inside the architecture.

Table 3. Ensemble Model Hyperparameters

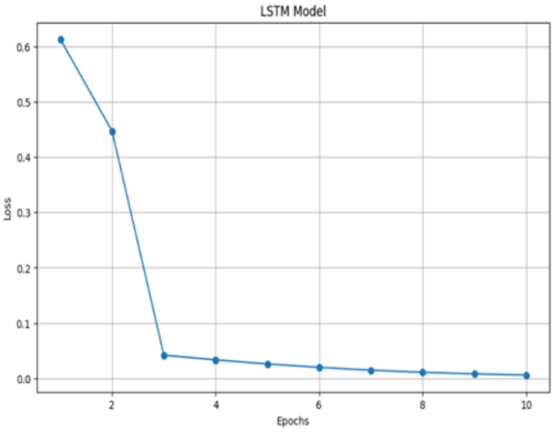
Hyperparameter	**Value**
Number of LSTM units	50
Number of FNN units	64
Activation function	ReLU
Loss function	Mean Squared Error
Optimizer	Stochastic Gradient Descent (SGD) optimizer
Number of epochs	10
Batch size	32

The effectiveness and resilience of the ensemble model are enhanced by these hyperparameters, which allow for precise forecasting of the results of road accidents.

5. 1.Road Accident Fatality Prediction Loss

The proposed model shows promise in learning from the dataset, it's crucial to address potential overfitting to ensure its effectiveness in predicting road accident fatalities accurately. Loss functions are fundamental components in machine learning models that quantify the model's performance by comparing its predictions to the ground truth. Minimizing loss during model training ensures that the machine learning model effectively learns patterns from the data to make accurate predictions of road accident fatalities.

- Figure 5(a) illustrates the loss incurred during the training of an LSTM model for predicting road accident fatalities across multiple epochs.
- Figure 5(b) displays the training loss of an FNN model over 10 epochs for the task of predicting road accident fatalities. The x-axis denotes the number of epochs, ranging from 1 to 10, while the y-axis represents the corresponding loss values.
- Figure 5(c) demonstrates the training loss of an ensemble model, which integrates a FNN and an LSTM model, for the prediction of road accident fatalities. The x-axis denotes the number of training epochs (iterations), while the y-axis illustrates the loss values.



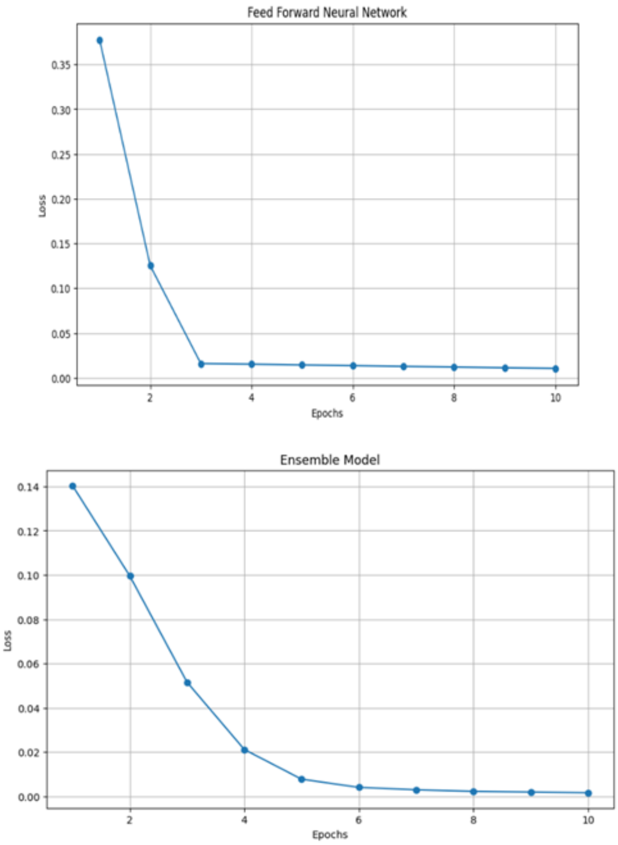


Figure 5.The training loss curves for three models predicting road accident fatalities: LSTM 3 (a), FNN 3 (b), and an Ensemble 3 (c).

5. 2.Performance Analysis

In assessing the effectiveness of a model, performance analysis in predictive modeling entails the use of different metrics such as MSE, RMSE, and R2. MSE measures the mean square distance between the predicted and actual values. On the other hand, RMSE is used to measure it more precisely by taking its root. Moreover, R2 assesses how well the model fits with observed data. Comparative studies using these metrics across various prediction methods like LSTM, FNN, and Ensemble models give important results. The Ensemble model outperforms both LSTM and FNN with significantly lower MSEs and RMSEs while R2 values are higher indicating better predictability and dependability.

i. Mean of Square Error (MSE)

An estimate is used to measure the mean of square error, which is the average of the squares distance among expected and reliable products. The expected value of the squared error loss is inversely proportional to this threat function [48].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (9)$$

Where, the number of samples is represented by n. The target's actual value for the ith sample is denoted by y_i. For the ith sample, the anticipated target value is denoted as $\hat{y}_i$.

ii. Root Mean Squared Error (RMSE)

One common way to measure how inaccurate a model is when making predictions about quantitative data is by looking at its root mean square error. Here is the formal definition:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

iii. R-squared

R squared is a statistical measure that assesses how well a model fits the data that has been observed. In terms of mathematics, the calculation of R-squared is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (11)$$

Where:

- $\bar{y}$is the mean of the observed data.

The fourth table compares the road accidents mortalities prediction methods based on three performance metrics: MSE, RMSE and R2. However, the LSTM model has an MSE value of 0.00489319, RMSE of 0.6995137 and an astonishing R2 of 0.9983709 showing that it predicts quite well. Similarly, the FNN model also displays competitive outcomes with MSE=0.00976033, RMSE=0.9879442 and R2=0.9967505. However, combining these two models LSTM and FNN in ensemble model performs better than any single one in terms of MSE (0.00172280), RMSE (000415067) and R2(0999426).

Table 4. Performance Comparison of Prediction Techniques

Technique	MSE	RMSE	R^2
LSTM	0.00489319	0.6995137	0.9983709
FNN	0.00976033	0.9879442	0.9967505
Ensemble	0.00172280	0.0415067	0.999426

Figure 6 below shows how LSTM, FNN and Ensemble compare with each other in terms of their performances.

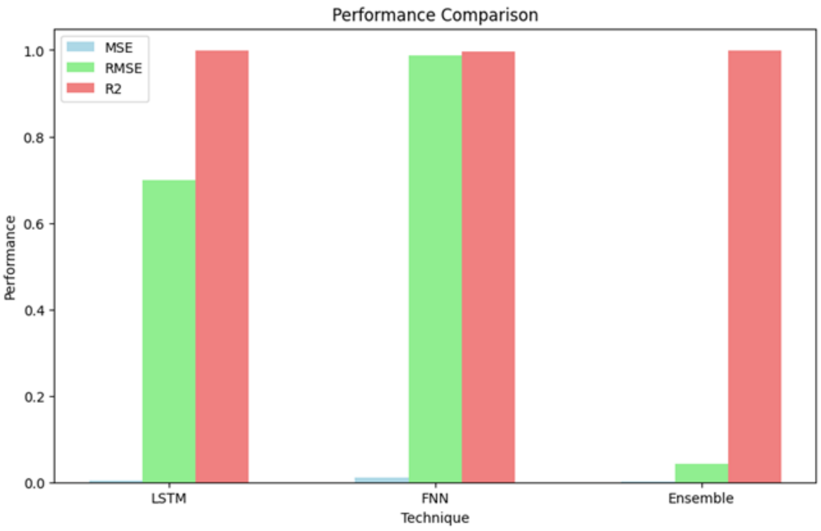


Figure 6. Performance Comparison of Techniques

The Ensemble method provides better results for R2, RMSE, and MSE compared to LSTM and FNN respectively suggesting that it is more accurate and dependable when it comes to making predictions.

6. Conclusion

India is experiencing a pressing problem due to the increase of road accidents resulting in fatalities. The issue of traffic safety in India is multidimensional, thus demanding unusual and data-directed approaches than traditional ones. Thus, development of models for prediction using sophisticated technologies such as machine learning can reveal concealed foresights. This study thus intends to fill this important knowledge gap towards better policy making and interventions that address the persisting challenges of road safety in India. Using machine learning algorithms such as LSTM and FNN, the study intends to analyze historical data on road accidents and casualties in India for purposes of predicting fatal trends over time, no of persons killed and injured. This research uses its data driven insights to determine key factors influencing dynamics of road accidents giving important information from which policy makers can develop focused interventions and policies. Ensemble modelling enhances predictability as well as reliability making it better than individualistic models. The ensemble model has better predictive performance than all other individual models: MSE=0.00172280, RMSE=0.0415067 and R2=0. 999426. The implication here is that using advanced machine learning techniques can help reduce losses emanating from road traffic crashes hence making policies based on real evidence. Additionally, future research should explore if more variables can generally improve the prediction power or if other ensemble methods have similar behavior thus assisting in targeting specific measures meant for checking the safety levels on roads across India among others going beyond it.

Refrences

[1] Bhardwaj, Utkarsh, A. P. Teixeira, and C. Guedes Soares. "Casualty analysis methodology and taxonomy for FPSO accident analysis." Reliability Engineering & System Safety 218 (2022): 108169.

[2] Ahmed, Sirwan K., Mona G. Mohammed, Salar O. Abdulqadir, Rabab G. Abd El-Kader, Nahed A. El-Shall, Deepak Chandran, Mohammad E. Ur Rehman, and Kuldeep Dhama. "Road traffic accidental injuries and deaths: A neglected global health issue." Health science reports 6, no. 5 (2023): e1240.

[3] Maqbool, Younus, Ankit Sethi, and Jagdeep Singh. "Road safety and road accidents: an insight." International Journal of Information And Computing Science, Volume 6 (2019): 93-105.

[4] Celine, Thalappillil Mathew, and Jimmy Antony. "A study on injuries sustained in road traffic accidents at a tertiary care level." International Journal of Environmental Health Engineering 3, no. 1 (2014): 23.

[5] Singh, Dalbir, Satinder P. Singh, M. Kumaran, and Sonu Goel. "Epidemiology of road traffic accident deaths in children in Chandigarh zone of North West India." Egyptian journal of forensic sciences 6, no. 3 (2016): 255-260.

[6] World Health Organization. Global status report on road safety 2015. World Health Organization, 2015.

[7] Cai Q.: Cause analysis of traffic accidents on urban roads based on an improved association rule mining algorithm. IEEE Access 8, 2020, 75607–75615

[8] Sobhana M., et al.: A Hybrid Machine Learning Approach for Performing Predictive Analytics on Road Accidents. 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2022

[9] Road Accidents in Malaysia: Top 10 Causes & Prevention. Kurnia, 21 Sept. 2022 [http://www.kurnia.com/blog/road-accidents-causes].

[10] Giri, Om Prakash, Padma Bahadur Shahi, and Sandeep Poddar. "Analysis of road traffic crashes in the Narayanghat to Mugling road segment in Nepal." In AIP conference proceedings, vol. 2854, no. 1. AIP Publishing, 2023.

[11] Koramati, Siddardha, Bandhan Bandhu Majumdar, Agnivesh Pani, and Prasanta K. Sahu. "A registry-based investigation of road traffic fatality risk factors using police data: A case study of Hyderabad, India." Safety science 153 (2022): 105805.

[12] Jumaniyozov, K. Y., N. N. Inoyatova, M. M. Olimova, and S. A. Rakhmonova. "INJURY AND MORTALITY IN CAR ACCIDENTS." Ta'lim innovatsiyasi va integratsiyasi 9, no. 1 (2023): 27-34.

[13] Kumar, Batchu Pavan, Vuyyuru Vidya Devi, Tapas Kumar Bandyopadhyay, and Syed Mehmood Hussaini. "Factors influencing road traffic accidents causing maxillofacial injuries in Nalgonda District: prospective survey of 366 cases." Journal of injury and violence research 15, no. 1 (2023): 27.

[14] Patil, Jayesh, Mandar Prabhu, Dhaval Walavalkar, and Vivian Brian Lobo. "Road accident analysis using machine learning." In 2020 IEEE Pune Section International Conference (PuneCon), pp. 108-112. IEEE, 2020.

[15] Östh J., et al.: Driver kinematic and muscle responses in braking events with standard and reversible pre-tensioned restraints: validation data for human models. SAE Technical Paper, 2013, 2013-22-0001.

[16] Chen M.-M., Chen M.-Ch.: Modeling road accident severity with comparisons of logistic regression, decision tree, and random forest. Information 11(5), 2020, 270

[17] Abdel-Aty, Mohamed, Joanne Keller, and Patrick A. Brady. "Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression." Transportation Research Record 1908, no. 1 (2005): 37-45.

[18] Chang, Li-Yen, and Hsiu-Wen Wang. "Analysis of traffic injury severity: An application of non-parametric classification tree techniques." Accident Analysis & Prevention 38, no. 5 (2006): 1019-1027.

[19] Yan, Xuedong, and Essam Radwan. "Analyses of rear-end crashes based on classification tree models." Traffic injury prevention 7, no. 3 (2006): 276-282.

[20] Qin, Xiao, and Junhee Han. "Variable selection issues in tree-based regression models." Transportation Research Record 2061, no. 1 (2008): 30-38.

[21] Elmitiny, Noor, Xuedong Yan, Essam Radwan, Chris Russo, and Dina Nashar. "Classification analysis of driver's stop/go decision and red-light running violation." Accident Analysis & Prevention 42, no. 1 (2010): 101-111.

[22] Pande, Anurag, Mohamed Abdel-Aty, and Abhishek Das. "A classification tree based modeling approach for segment related crashes on multilane highways." Journal of safety research 41, no. 5 (2010): 391-397.

[23] Akhoondzadeh, Mehdi. "Decision Tree, Bagging and Random Forest methods detect TEC seismo-ionospheric anomalies around the time of the Chile,(Mw= 8.8) earthquake of 27 February 2010." Advances in Space Research 57, no. 12 (2016): 2464-2469.

[24] Sobhana, Mummaneni, Nihitha Vemulapalli, Gnana Siva Sai Venkatesh Mendu, Naga Deepika Ginjupalli, Pragathi Dodda, and Rayanoothala Bala Venkata Subramanyam. "URBAN TRAFFIC CRASH ANALYSIS USING DEEP LEARNING TECHNIQUES." Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska 13, no. 3 (2023): 56-63.

[25] Kanakala, Raviteja, and Krishna Reddy. "Modelling a deep network using CNN and RNN for accident classification." Measurement: Sensors (2023): 100794.

[26] Sameen, Maher Ibrahim, and Biswajeet Pradhan. "Severity prediction of traffic accidents with recurrent neural networks." Applied Sciences 7, no. 6 (2017): 476.

[27] Yuan, Jinghui, Mohamed Abdel-Aty, Yaobang Gong, and Qing Cai. "Real-time crash risk prediction using long short-term memory recurrent neural network." Transportation research record 2673, no. 4 (2019): 314-326.

[28] Shaik, Md Ebrahim, Md Milon Islam, and Quazi Sazzad Hossain. "A review on neural network techniques for the prediction of road traffic accident severity." Asian Transport Studies 7 (2021): 100040.

[29] Jiang, Feifeng, Kwok Kit Richard Yuen, and Eric Wai Ming Lee. "A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions." Accident Analysis & Prevention 141 (2020): 105520.

[30] Cura, Aslıhan, Haluk Küçük, Erdem Ergen, and İsmail Burak Öksüzoğlu. "Driver profiling using long short term memory (LSTM) and convolutional neural network (CNN) methods." IEEE Transactions on Intelligent Transportation Systems 22, no. 10 (2020): 6572-6582.

[31] Jia, Shuo, Fei Hui, Shining Li, Xiangmo Zhao, and Asad J. Khattak. "Long short-term memory and convolutional neural network for abnormal driving behaviour recognition." IET Intelligent Transport Systems 14, no. 5 (2020): 306-312.

[32] Chen, Mu-Ming, and Mu-Chen Chen. "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest." Information 11, no. 5 (2020): 270.

[33] Gatera, Antoine, Martin Kuradusenge, Gaurav Bajpai, Chomora Mikeka, and Sarika Shrivastava. "Comparison of random forest and support vector machine regression models for forecasting road accidents." Scientific African 21 (2023): e01739.

[34] Elyassami, Sanaa, Yasir Hamid, and Tetiana Habuza. "Road crashes analysis and prediction using gradient boosted and random forest trees." In 2020 6th IEEE Congress on Information Science and Technology (CiSt), pp. 520-525. IEEE, 2021.

[35] Jamal, Arshad, Muhammad Zahid, Muhammad Tauhidur Rahman, Hassan M. Al-Ahmadi, Meshal Almoshaogeh, Danish Farooq, and Mahmood Ahmad. "Injury severity prediction of traffic crashes with

ensemble machine learning techniques: A comparative study." International journal of injury control and safety promotion 28, no. 4 (2021): 408-427.

[36] Zarei, Mohammad, Bruce Hellinga, and Pedram Izadpanah. "CGAN-EB: A non-parametric empirical Bayes method for crash frequency modeling using conditional generative adversarial networks as safety performance functions." International Journal of Transportation Science and Technology 12, no. 3 (2023): 753-764.

[37] Amorim, Brunna de Sousa Pereira, Anderson Almeida Firmino, Cláudio de Souza Baptista, Geraldo Braz Júnior, Anselmo Cardoso de Paiva, and Francisco Edeverton de Almeida Júnior. "A Machine Learning Approach for Classifying Road Accident Hotspots." ISPRS International Journal of Geo-Information 12, no. 6 (2023): 227.

[38] Saravanarajan, Vani Suthamathi, Rung-Ching Chen, Christine Dewi, Long-Sheng Chen, and Lata Ganesan. "Car crash detection using ensemble deep learning." Multimedia Tools and Applications (2023): 1-19.

[39] Gutierrez-Osorio C., González F. A., Pedraza C. A.: Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data. Computers 11(9), 2022, 126.

[40] Comi, Antonio, Antonio Polimeni, and Chiara Balsamo. "Road accident analysis with data mining approach: evidence from Rome." Transportation research procedia 62 (2022): 798-805.

[41] Zheng, Haifeng, Feng Lin, Xinxin Feng, and Youjia Chen. "A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction." IEEE Transactions on Intelligent Transportation Systems 22, no. 11 (2020): 6910-6920.

[42] Chen, Hengrui, Hong Chen, Ruiyu Zhou, Zhizhen Liu, and Xiaoke Sun. "Exploring the mechanism of crashes with autonomous vehicles using machine learning." Mathematical problems in engineering 2021 (2021): 1-10.

[43] Choi, Jae Gyeong, Chan Woo Kong, Gyeongho Kim, and Sunghoon Lim. "Car crash detection using ensemble deep learning and multimodal data from dashboard cameras." Expert Systems with Applications 183 (2021): 115400.

[44] Rahman, Md Sharikur, Mohamed Abdel-Aty, Samiul Hasan, and Qing Cai. "Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones." Journal of safety research 70 (2019): 275-288.

[45] Islam, Mohaiminul, Guorong Chen, and Shangzhu Jin. "An overview of neural network." American Journal of Neural Networks and Applications 5, no. 1 (2019): 7-11.

[46] Lalapura, Varsha S., J. Amudha, and Hariramn Selvamuruga Satheesh. "Recurrent neural networks for edge intelligence: a survey." ACM Computing Surveys (CSUR) 54, no. 4 (2021): 1-38.

[47] Allcock, Jonathan, Chang-Yu Hsieh, Iordanis Kerenidis, and Shengyu Zhang. "Quantum algorithms for feedforward neural networks." ACM Transactions on Quantum Computing 1, no. 1 (2020): 1-24.

[48] Nagaraj, Nandini, Harinahalli Lokesh Gururaj, Beekanahalli Harish Swathi, and Yu-Chen Hu. "Passenger flow prediction in bus transportation system using deep learning." Multimedia tools and applications 81, no. 9 (2022): 12519-12542.