***https://doi.org/10.48047/AFJBS.6.13.2024. 3634-3646***

African Journal of Biological Sciences

# Predictive Data Modeling Using Machine Learning Techniques for Road Crashes

**Ashwini Bagga[1], Dr. Sumit Srivastava[2], Dr. Rajveer Singh Shekhawat[3]**

[1]Research Scholar, Department of Mathematics and Statistics, Manipal University Jaipur, tel.: +91-96360-30303, email: ashwinibagga@gmail.com

[2]Professor, Department of Information Technology, Manipal University Jaipur, email: sumit.310879@gmail.com

[3]Professor(retd.), School of Computing and Information Technology, Manipal University Jaipur, email: rajveersingh.shekhawat@jaipur.manipal.edu

***Corresponding Author:***

Dr. Sumit Srivastava
Professor, Department of Information Technology
Manipal University Jaipur
Jaipur, Rajasthan

Email: sumit.srivastava@jaipur.manipal.edu

Article History
Volume 6, Issue 13, 2024
Received: 18June 2024
Accepted: 02July 2024
doi:***10.48047/AFJBS.6.13.2024.*** 3634-3646

## ABSTRACT

Road crashes have become a common cause of deaths and injuries worldwide. The phenomenon is severe especially in the developing nations. The resulting mortality and morbidity due to road traffic crashes occurs mostly to the vulnerable road users such as motorized two-wheelers and non-motorized transport users. To address the burning issue, use of scientific methods for data collection, analysis and prediction modeling is highly recommended. The previous methods make use of the inference and the statistical models for the crashes in certain categories. However, the Ensemble and Machine Learning algorithm involves the correlations among the independent features and will not be used for casual reasoning. This research focuses on the analysis and prediction of road crashes using machine learning algorithms, particularly in the context of Jaipur city. It emphasizes the importance of scientific methods for data collection, analysis, and prediction modeling to address the severity of road traffic crashes. The research aims to build prediction models based on classification techniques, comparing the performance of various algorithms, and providing details of the contributing features.

***Keywords:***

Road Safety
Data Analytics
Traffic Crash Prediction Machine Learning
Classification Techniques

## 1. INTRODUCTION

Road crashes cause loss of 13,50,000 lives every year worldwide and it has become the primary cause of death for youth and children belonging to the age group five to twenty-nine years and it is the eighth leading cause of death for all age groups. The fatality rate is three times higher in countries belonging to low- and middle-income group [34]. The top ten significant causes of unnatural death in the world are depicted in Table 1 wherein deaths due to road traffic injuries lie at the eighth position.

The vulnerable road users including pedestrians, bicyclists, and motorized two-wheeler drivers are mostly affected by road traffic injuries and bear the loss in terms of deaths or permanent disability. However, ranks on top in terms of road crashes and fatalities in India followed by China and USA [35]. India contributes around 11 percent of the crash related deaths in the World [34], more than 1,53,000 lives were lost and 3,48,000 people got injured in around 4,12,000 accidents reported in year 2021 [36]. The worldwide statistics for the year 2016 reveal that the reported road crash fatalities in India were 1,50,785 whereas WHO estimated 2,99,091 fatalities and the Institute for Health Metrics and Evaluation (IHME) estimated around 2,19,670 deaths in 2016 due to road traffic crashes in India. This indicates a prevalent gap determining under reporting of data related to road crashes in India. The number of road crashes, persons killed and injured in India from 2012 to 2021 is depicted in Table 2.

Table 1. Leading Causes of Death (All Ages) in 2016. Source: WHO

| Rank | Cause | %age of Total Deaths |
|---|---|---|
| 1 | Ischemic Heart Disease | 16.6 |
| 2 | Stroke | 10.2 |
| 3 | Chronic Obstructive Pulmonary Disease | 5.4 |
| 4 | Lower Respiratory Infections | 5.2 |
| 5 | Alzheimer's Disease and Other Dementias | 3.5 |
| 6 | Trachea, Bronchus, Lung Cancers | 3.0 |
| 7 | Diabetes Mellitus | 2.8 |
| 8 | Road Traffic Injuries | 2.5 |
| 9 | Diarrhoeal Diseases | 2.4 |
| 10 | Tuberculosis | 2.3 |

Table 2. Road Accidents, Deaths, and Injuries in India (2012 - 2021) | Source: MoRTH

| Year | Accidents | Killed | Injured |
|---|---|---|---|
| 2012 | 490383 | 138258 | 509667 |
| 2013 | 486476 | 137572 | 494893 |
| 2014 | 489400 | 139671 | 493474 |
| 2015 | 501423 | 146133 | 500279 |
| 2016 | 480652 | 150785 | 494624 |
| 2017 | 464910 | 147913 | 470975 |
| 2018 | 467044 | 151417 | 469418 |
| 2019 | 449002 | 151113 | 451361 |
| 2020 | 366138 | 131714 | 348279 |
| 2021 | 412432 | 153972 | 348448 |

The percentage change in number of persons killed in India from 2012 to 2021 shows a significant increase in number of fatalities by 17 percent in year 2021 whereas a decline of 13 percent was observed in 2020 that was due to nationwide lockdown due to Covid 19.

In addition to loss of lives, road crashes also affect the national economy in several ways. It has been revealed by a media report that road traffic crashes costs around three percent loss of India's GDP which translates to more than $ 58,000 million in terms of value [37]. Another estimation according to Save Life Foundation, New Delhi describes it a loss of Rupees 4.34 lakh crores to the Indian economy [38]. The report by World Bank claims that road crashes cost 7.5 percent of country's GDP costing $172.02 billion [39] in year 2016. Researchers have devised various methods to classify and predict the severity of road crashes. These classifications typically range from property damage only (PDO) to fatality, with intermediate categories like probable injuries and incapacitating injuries [1]. Ma et al. (2018) and Mesa-Arango et al. (2018) offer a simpler three-level system: PDO, injuries, and fatalities[2][4].

A significant body of research explores statistical and machine learning techniques for crash severity assessment [3] [5-10]. Traditional methods like probit, logit, and their mixed variants have been employed extensively [11-14]. Wang et al. (2021) introduced a more nuanced approach using correlated mixed logit models that account for heterogeneity and temporal fluctuations, capturing both injury severity and vehicle damage[15].

Recent studies have delved into comparisons of statistical and machine learning techniques for crash prediction [17 – 23], but still the crash severity analysis required lot to be explored as the proposed solution. The proposed study is one of such study discussed on around 10,337 cases taken from the various police stations on the road crash with severity measured classes defined. The various machine learning techniques is applied, and results are analysed for various performance metrics.

## 2. LITERATURE REVIEW

There is a vital need to conduct scientific research in the field of road safety that could assist the decision makers to formulate data driven strategies for respective regions. Machine Learning and Deep Learning offers advanced algorithms and techniques that can be used to analyze road crash data to build useful prediction models and formulation of association rules to discover hidden patterns.

Over the past few decades, research on crash injury severity has advanced significantly through the utilization of diverse supervised learning algorithms and sophisticated statistical analysis techniques. Studies by Zinno et al. (2022) and Choo et al. (2022) have delved into methodologies like multivariate regression, autocorrelation, trigonometry, and linear regression, among others, to unravel the complexities of crash injury severity[25][26]. In a groundbreaking work, Song et al. (2021) developed a comprehensive crash severity model that integrated risk indicators associated with both drivers and road conditions [27]. The Bayesian network is deployed to systematically explored interconnections between accident seriousness and various factors, revealing nuanced combinations that exerted a substantial impact on severity outcomes. Similarly, Topuz and Delen (2021) adopted a multi-stage probabilistic inference approach, utilizing a Bayesian belief network (BBN) to discern factors significantly influencing injury outcomes in car crashes [28]. Their method not only yielded interpretable results but also maintained prediction accuracy.

Exploring the impact of road and environmental factors on crash severity, Yang et al. (2022b) introduced the eXtreme Gradient Boosting (XGBoost) model, complemented by the SHapley Additive exPlanation (SHAP) value for model interpretability [21]. Their findings emphasized the efficacy of a holistic approach that considers the synergistic effects of road and environmental conditions on crash severity prediction. Moreover, recent research has seen a surge in interest in AI techniques such as decision trees and Artificial Neural Networks (ANN) as potential solutions for traffic engineering challenges and road safety issues. Shiran et al. (2021) evaluated highway crash severity using Ensemble and Machine Learning approaches, highlighting the effectiveness of the C5.0 method for estimating traffic crash severity levels [29].

In another domain, Hosseinzadeh et al. (2021) investigated variables affecting crash severity involving large trucks, employing Support Vector Machine (SVM) and random parameter LOGIT models to develop a robust prediction model. Furthermore, Mohanty et al. (2022) leveraged binary logistic regression and ANN to assess crash prediction, comparing the strengths and weaknesses of these methods [30]. Additionally, Danesh et al. (2022) explored crash severity in imbalanced datasets, employing data leveling techniques and machine learning methods to achieve optimal results [31]. Their insights highlighted factors contributing to fatal crashes, including head-on collisions, road curvature, and vehicle type.

In a comprehensive crash prediction model developed by Koramati et al. (2023), ANN algorithms played a crucial role in identifying significant factors like the cause of the crash and road geometry, emphasizing their influence on the likelihood of fatal incidents [33].Such methods can help researchers to identify complex factors related to road crashes to devise sustainable solutions backed by data driven strategies. To unleash the potential of Machine Learning in the area of road safety, the literature pertaining to the work done in the past has been systematically reviewed which is presented in this section.

Camilo Gutierrez Osorio, Cesar Pedraza [40] used various algorithms to analyse, characterize, and forecast road crashes and presented a detailed review. The researchers presented a collection of various data sources that were being used by various researchers to carry out their research based on road crashes. These data sources include Open Data Sources, Government Database, Data from Onboard Equipment, social media, and Measurement Technologies data [40]. The techniques such as Bayesian Networks, Support Vector Machines, Artificial Neural Networks and Deep Learning were reviewed and concluded stating when two or more algorithms or techniques are combined offer best results, and also proposed the scope of forecasting pertaining to road traffic models and predictions provide authentic results when used with heterogeneous data sources.

Md. Farhan Labib et. al. [42] used various algorithms such as Ada Boost, K Nearest Neighbour, Decision Tree, and Naive Bayes to classify the gravity of road traffic crashes. The classification of these algorithm categorized the crashes such as Fatal, Grievous, Simple Injury and Motor Collision. They used eleven major factors as features that majorly affected road crashes occurring in Bangladesh. The performance of every algorithm was determined for the four different severity classes defined to categorize road crashes. The results obtained by them show that Ada Boost and Naive Bayes achieved higher accuracy level. In total, the Ada Boost algorithm gave best results with around 80% accuracy score. That could be achieved using the fundamental having iterative classification using the decision tree.

Sakham Nagendra Babu and Jebamalar Tamilselvi [43] suggested a system for prediction for causes of road crashes for different categories of accidents. They implemented various methods of big data analytic using Machine Learning Algorithms to predict accurate information. The Enhanced Expectation Maximization Algorithms and Improved Association Rule Mining - IARM algorithm for different classes of vehicles was implemented. They have also used Traffic Congestion Analyzer and Machine Framework for training the machine and apply various association rules on the data set. The outcome of the algorithms show that their proposed approach is better in prediction of road crashes than the existing approaches.

G Pavan Karthik, Sneha B and Sudalaimuthu T [44], developed Machine Learning based intelligent models that segregate injury severity and checks relationship between various features such as driver behaviour, light conditions, road condition and weather condition etc. They designed a hybrid approach using K Means Clustering, Random Forest, Linear Regression and plotted the results. According to the results of different algorithms the accuracy for fatal injury yielded 96.5 percent and non-fatal category yielded 97.45 percent accuracy.

S. Krishnaveni and Dr. M. Hemalatha [45] evaluate performance of classification using J48, Naıve Bayes, PART, Ada Boost, and Random Forest classifiers for accident dataset for the year 2008 for three different scenarios i.e., based on accident, casualty, and vehicle information. In the first scenario accident dataset was analyzed using the given algorithms for attributes District Council, Weather, Junction Control, Vehicle Movement, No of Casualties and Types of Collision. Among these Random Forest classification algorithm took highest percentage when compared with other classification algorithms. For the second scenario i.e., applying Genetic Algorithm for feature selection in Casualty Dataset for attributes like Age, Sex, Role of Casualty and Location of Accident Random Forest classification algorithm took highest percentage when compared with other classification algorithms. For the third scenario Genetic Algorithms were applied for feature selection in Vehicle Dataset involving attributes involving Driver Age, Vehicle Class, and Year of Manufacture. Again, in this case Random Forest classification algorithm took highest percentage when compared with other classification algorithms.

Jongtae Lee, Taekwan Yoon, Jonghak Lee, and Sangil Kwon [46] used accident data of nine years from 2007 to 2015 for the Neebu Expressway. The research revealed that according to Korean Transport Safety Authority (KTSA), the road traffic accidents dropped but fatalities increased in accidents caused due to rain in the period 2013 to 2016, specifically in Seoul. The methodologies involved Decision Tree, Artificial Neural Networks (ANN) and Random Forest Algorithms on the described datasets. According to the results the Random Forest method yielded more accurate predictions. The output revealed that attributes Rainfall Intensity, Curve Length and Driver Gender could be considered as important factors affecting the number of road crashes depending on surface condition of Neebu Expressway.

The main objective behind the current research to explore the machine learning models and its variants to investigate the severity of the road crashes. Also, the methodology to handle the im-balancing of the data may also be targeted in the research. The literature survey has shown some light on the Artificial Neural Network (ANN) and related algorithms which also compared during the current study of road crashes in Jaipur.

## 3. ROAD CRASHES IN JAIPUR

Jaipur is the capital of Rajasthan which also accounts for highest number of road crash fatalities in the state. Jaipur district comprises a large area which comes under different jurisdictions. According to the Commissionerate policing system, Jaipur is divided into two zones Jaipur Rural and Urban. Jaipur Urban is further segregated into East, West, North and South zones. Jaipur Rural comprises a larger area including distant towns including Kotputli, Shahpura, Chomu and Dudu. Majority of the highways such as NH 48 and NH 52 pass through these towns. According to the road crash data of 2022, Jaipur Rural reported highest share of road crashes followed by Jaipur East, West, South and North. Number of crashes, persons killed and injured in Jaipur district in year 2022 is depicted in Table 3 and percentage share of crashes, deaths and injuries in each zone is depicted in Figure 2.

Table 3. Accidents, Deaths, and Injuries in Different Zones of Jaipur (2022) | Source: Transport Department, Rajasthan

| S. N. | Zone | Crashes | Killed | Injured |
|---|---|---|---|---|
| 1 | East | 908 | 256 | 708 |
| 2 | West | 899 | 261 | 741 |
| 3 | North | 240 | 49 | 212 |
| 4 | South | 640 | 199 | 506 |
| 5 | Rural | 1248 | 562 | 1198 |
| **TOTAL** | | 3935 | 1327 | 3365 |

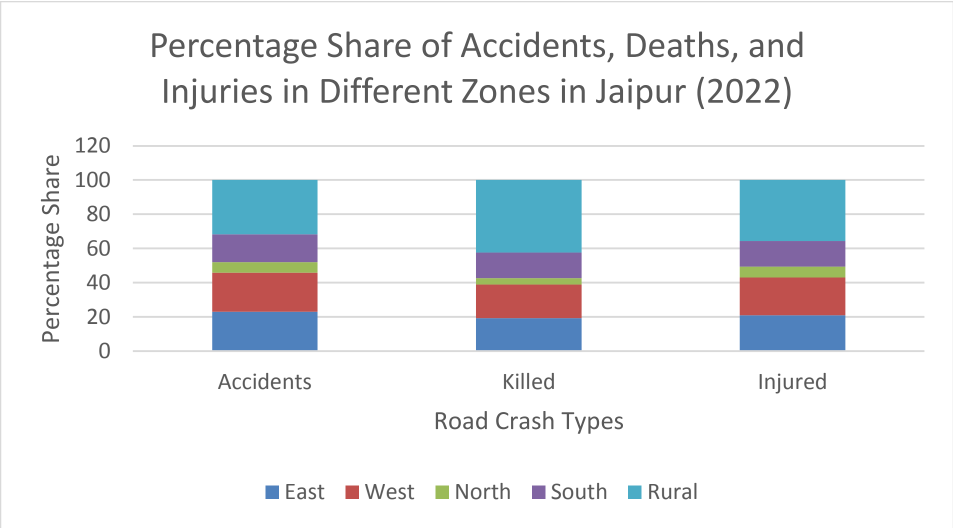


Figure 1. Percentage Share of Crashes, Deaths, and Injuries in Different Zones in Jaipur (2022)

## 4. DATA DESCRIPTION

### 4.1 The Problem Statement

The prediction models were build based on road crashes reported in selected areas of Jaipur city during the year 2019, the FIR reports were collected from various Police Stations namely Adarsh Nagar, Amer, Bhatta Basti, Brahampuri, Galta Gate, Gandhi Nagar, Jalupura, Jawahar Circle and Shyam Nagar.

The FIR had been studied in detail to retrieve relevant information about accidents such as date, time and day of accident, vehicles or road users involved, number of persons killed or injured, type of collision, cause of accident, location of accident, Zone and Circle of Police Station, distance, and direction of accident spot from police station etc. The accident location was not recorded by the police officials in FIR. Hence, the GPS coordinates were recorded manually for each accident. After detailed analysis of FIRs, a dataset was generated such that the Machine Learning algorithms could be applied on them. The structure of data source created using the FIRs is depicted in Table 4.

### 4.2 Challenges Faced Regarding Data Collection

The analysis using Machine Learning algorithms on a dataset has various pre-requisites in terms of format, missing data, and type of data etc. According to the category of data, certain algorithms could be applied on numerical and certain on categorical data. Since Ensemble and Machine Learning algorithms are based on mathematical models including probability and statistical analysis, it was necessary to convert categorical data into its equivalent numerical form. Similarly, an algorithm cannot be applied if dataset contains missing values. Therefore, necessary pre-processing steps were applied on the raw data for carrying out prediction modeling using Ensemble and Machine Learning algorithms [47-49].

With respect to the accident data collected using FIRs received from various police stations, lot of issues were faced to generate a functional dataset. The major challenges faced in terms of data collection for writing this paper involves the following:

i. Time consumed for collecting data from various police stations,
ii. Data extraction from physical FIRs by going through each FIR in detail,
iii. Conversion of categorical data into equivalent numeric value,
iv. Handling missing values,
v. No scientific method used for writing FIR by police,
vi. Duplicate records for a single incident
vii. Non-standard format of FIRs makes difficult to extract accident related data,
viii. No proper analysis done for reporting the cause of the accident,
ix. Mapping of police station and police circle for each FIR,
x. Evaluation of GPS coordinates for each accident reported in FIR

Table 4. Structure of Data Source Created using FIR. Source: Rajasthan Police

| S. No. | Column Name | Data Type | Description |
|---|---|---|---|
| 1 | FIR No | Integer | The number assigned to FIR by Police Station |
| 2 | Date | Date Time | Date of the accident |
| 3 | Time | Date Time | Time of the accident |
| 4 | Day | Numeric | Day of the accident |
| 5 | User 1 | String | First road user category |
| 6 | User 2 | String | Second road user category |
| 7 | Collision Category | String | Such as vehicle to vehicle or vehicle to pedestrian etc. |
| 8 | Killed | Integer | Number of persons killed in accident |
| 9 | Injured | Integer | Number of persons injured in accident |
| 10 | Age | Integer | Age of the person killed (it can be multiple) |
| 11 | Violation Type | String | Such as overspeed or use of mobile etc. |
| 12 | Collision Type | String | Such as Hit from back or Head on etc. |
| 13 | Cause of Fatality | String | Such as head injury, grievous injury etc. |
| 14 | Category of Injury | String | Such as grievous or minor injury etc. |
| 15 | Zone | String | Such as East, West, North or South |
| 16 | Circle | String | Police circle within which a police station lies |
| 17 | Police Station | String | Name of the police station |
| 18 | Distance from PS | Float | Distance of accident from police station |
| 19 | Direction From PS | String | Direction of location from police station |
| 20 | MV Act | String | Sections of MV Act under which case is booked |
| 21 | IPC | String | Section of IPC under which case is booked |
| 22 | Latitude | Float | Latitude of the accident location |
| 23 | Longitude | Float | Longitude of the accident location |
| 24 | Address | String | Address of the accident location recorded in FIR |

To overcome these issues, lot of manual work was done to extract the data. The GPS coordinates for accident locations were evaluated for each record. After creation of the dataset, the values were mapped to generate heatmap of accidents that occurred in Jaipur [Fig 3].

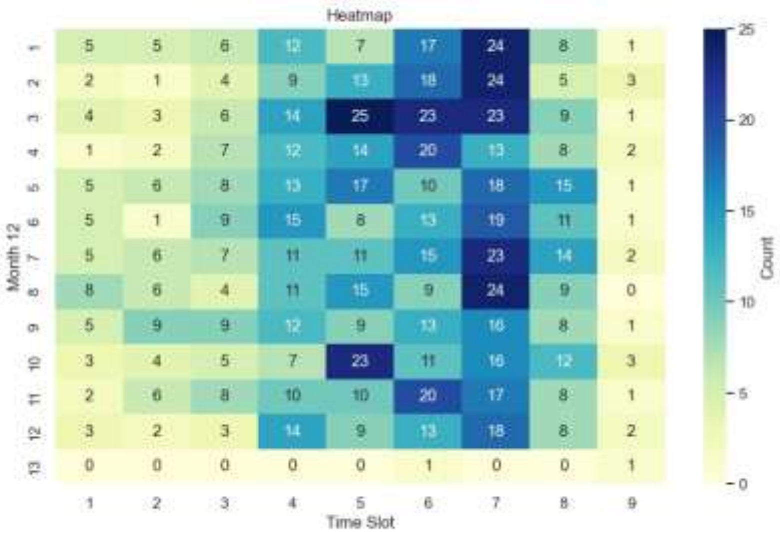


Figure 2. Heat Map of the Accident counts for each month at various time slot.

In addition, many categorical values were encoded using the Ordinal Encoder class in Python. For example, to apply Machine classification algorithms on the Police Station feature, the name of the police station has been encoded to its equivalent numeric code [50]. For example, encoded value for police station is depicted in Table 5.

Table 5. Encoded Value for Police Station Feature

| S. No. | Police Station | Encoded Value |
|---|---|---|
| 1 | Adarsh Nagar | 1 |
| 2 | Amer | 2 |
| 3 | Bhatta Basti | 3 |
| 4 | Brahampuri | 4 |
| 5 | Galta Gate | 5 |
| 6 | Gandhi Nagar | 6 |
| 7 | Jalupura | 7 |
| 8 | Jawahar Circle | 8 |
| 9 | Shyam Nagar | 9 |

## 5. METHOD

The research aims at investigating the road accident prediction based on the classification of the given injuries as "Fatal," "Grievous Injury," "Minor Injury," and "No Injury.", a multiclass problem in which ensemble algorithms were used to analyze the dataset. The ensemble algorithm combines several base models with feature and their influence prompt analysis is performed on the model, it reduces the dispersion of the data and result acceptability is increased. The performance for various ensemble algorithms is measured and analyzed in terms of predicting accurately the road accident severity, its precision and calculation of the F1-score and recall. The algorithms utilized can be briefly summarized as follows:

### 5.1 Logistic Regression (LS)

The Logistic regression is a binary classifier, which uses loss function(cross-entropy) and summarizes it for all classes in terms of Probability Distribution for multinominal regression [54]. However, with limitation to the one-vs-rest for the multi-class capabilities it majorly used as meta-learner. In the given problem it checks for “Fatal” Injury against all classes.

### 5.2 Support Vector Machine (SVM)

The SVM process is like LR but approaches the problem in the different way. The algorithms create a binary classifier for every unique pair of classes. For the given ‘N’ class, then we N*(N-1)/2 classifiers. It checks for probable two classes ‘A’ and ‘B’ with most votes is assigned while predicting. Another approach is to have separate classifier for each class and each class is distinguished its assigned class from all others. During Prediction, the model with the highest score is selected for the chosen point.

### 5.3 Random Forest (RF)

A Random Forest is one of the finest classifiers for multi-class problems, with multitude of decision trees created as weak leaner through the technique called as bagging [52]. Each tree is trained on the random sample raised diversity and prevent overfitting of the data. At each point a random subset of features is considered. During the next stage of voting, data points is classified on all trees trained, the class with majority votes is determined among all trees.

### 5.4 Gradient Boosting

Unlike Random Forest, which trains individual trees independently, Gradient Boosting adopts a sequential methodology. It constructs a model iteratively, each stage dedicated to enhancing overall performance by learning from preceding errors. The fundamental units are typically decision trees (weak learners), chosen for their simplicity and flexibility. At each iteration, a tree is trained to forecast pseudo-residuals for data points. These pseudo-residuals signify the discrepancies (i.e., differences between actual labels and predictions) of the previous model within the ensemble. The newly trained tree joins the ensemble, with its predictions shaping the collective model's output. This cycle persists for a predetermined number of iterations, progressively honing the model's capacity to categorize data points.

### 5.5 Neural Network

The Neural Network is the ideal choice for multi-class classification with Multi-Layer perceptron used as the standard feed-forward neural network architecture commonly used for networks. It takes various features related to the road accident (e.g., vehicle types, speed, weather conditions) as input and processes them through hidden layers with activation functions [53]. The final output layer uses a softmax activation function to predict the probability distribution across all injury severity classes (minor, moderate, severe, fatal). The network learns by adjusting the weights between neurons based on the difference between predicted and actual injury severity. Backpropagation, a powerful algorithm, helps calculate these adjustments efficiently. Unlike logistic regression, neural networks can capture complex non-linear relationships between features and injury severity, leading to potentially more accurate predictions. Deep learning architectures (a specific type of neural network with many layers) can even learn feature representations from raw data like images or sensor data from the accident scene, reducing the need for manual feature engineering. Neural networks can handle large datasets effectively, which is often the case with road accident data.

## 6. RESULTS AND DISCUSSION

The classification report for different Ensemble and Machine Learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting and Neural Network model. These reports are providing performance metrics for each class predicted by the models as well as some overall averages. The data seems to be related to some kind of injury classification with classes being "Fatal," "Grievous Injury," "Minor Injury," and "No Injury."

The Logistic Regression performs best for the "Fatal" class with a precision of 0.64 and an F1-score of 0.74. The "Minor Injury" class has a high precision (0.78) but a low recall (0.19), which results in a moderate F1-score (0.31). "No Injury" has zero precision and recall, indicating that this class was not correctly predicted at all. The overall accuracy is 0.61, and the macro and weighted average F1-scores are 0.37 and 0.52 respectively.

The Support Vector Machine has a lower overall performance compared to Logistic Regression. The "Fatal" class still has the best F1-score (0.78), but with a lower precision (0.58) compared to Logistic Regression. The other classes have significantly lower scores across all metrics, with "Minor Injury" and "No Injury" again having zero precision and recall. The overall accuracy is 0.55, and the macro and weighted average F1-scores are 0.23 and 0.45 respectively.

The Random Forest precision, recall, and F1-score for the "Fatal" class are 0.59, 0.71, and 0.64, respectively, indicating a reasonable performance, with a decent balance between precision and recall. The "Minor Injury" class has very low recall (0.05) despite a moderate precision (0.85), resulting in a low F1-score (0.09). The "No Injury" class has low precision and recall, both at 0.33, with the F1-score also at 0.33, suggesting poor performance in predicting this class. The overall accuracy of the Random Forest model is 0.51. The macro average F1-score is 0.33, and the weighted average F1-score is 0.47, indicating moderate performance. [Fig.3]

The Gradient Boosting "Fatal" class performance is like the Random Forest model, with a slightly higher recall at 0.74 and an F1-score of 0.65. The "Minor Injury" class shows a very low precision (0.25) and a low recall (0.08), resulting in a very low F1-score (0.12). The "No Injury" class again shows zero precision and recall, indicating that this class was not correctly predicted, like the Logistic Regression and SVM models. The overall accuracy is 0.52, with the macro average F1-score at 0.29 and the weighted average F1-score at 0.48, which is slightly lower than the Random Forest model. [Fig.4]

Logistic Regression

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fatal | 0.64 | 0.88 | 0.74 | 117 |
| Grievous Injury | 0.48 | 0.39 | 0.43 | 41 |
| Minor Injury | 0.78 | 0.19 | 0.30 | 37 |
| No Injury | 0.00 | 0.00 | 0.00 | 7 |
| accuracy | | | 0.62 | 202 |
| macro avg | 0.48 | 0.36 | 0.37 | 202 |
| weighted avg | 0.61 | 0.62 | 0.57 | 202 |

Support Vector Machine

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fatal | 0.58 | 0.90 | 0.70 | 117 |
| Grievous Injury | 0.33 | 0.15 | 0.20 | 41 |
| Minor Injury | 0.00 | 0.00 | 0.00 | 37 |
| No Injury | 0.00 | 0.00 | 0.00 | 7 |
| accuracy | | | 0.55 | 202 |
| macro avg | 0.23 | 0.26 | 0.23 | 202 |
| weighted avg | 0.40 | 0.55 | 0.46 | 202 |

Figure 3. Results of Logistic Regression & Support Vector Machine

Random Forest

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fatal | 0.59 | 0.71 | 0.64 | 117 |
| Grievous Injury | 0.38 | 0.41 | 0.40 | 41 |
| Minor Injury | 0.15 | 0.05 | 0.08 | 37 |
| No Injury | 0.33 | 0.14 | 0.20 | 7 |
| accuracy | | | 0.51 | 202 |
| macro avg | 0.36 | 0.33 | 0.33 | 202 |
| weighted avg | 0.46 | 0.51 | 0.47 | 202 |

Gradient Boosting

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fatal | 0.58 | 0.74 | 0.65 | 117 |
| Grievous Injury | 0.38 | 0.37 | 0.37 | 41 |
| Minor Injury | 0.25 | 0.08 | 0.12 | 37 |
| No Injury | 0.00 | 0.00 | 0.00 | 7 |
| accuracy | | | 0.52 | 202 |
| macro avg | 0.30 | 0.30 | 0.29 | 202 |
| weighted avg | 0.46 | 0.52 | 0.48 | 202 |

Figure 4. Results of Random Forest & Gradient Boost

The Neural Network "Fatal" class has relatively balanced precision and recall scores (0.61 and 0.69 respectively), with an F1-score of 0.65. This indicates a reasonable performance for this class. The "Grievous Injury" and "Minor Injury" classes have identical precision and F1-scores of 0.40 and 0.36 respectively. However, "Grievous Injury" has a slightly better recall of 0.41 compared to 0.30 for "Minor Injury". The "No Injury" class has the lowest performance with zero precision and recall, which results in an F1-score of 0.00. This suggests the Neural Network model is unable to correctly predict this class at all. The overall accuracy of the model is 0.54. The macro average F1-score is 0.36, and the weighted average F1-score is 0.53. [Fig.5]

Neural Network

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fatal | 0.61 | 0.69 | 0.65 | 117 |
| Grievous Injury | 0.40 | 0.41 | 0.41 | 41 |
| Minor Injury | 0.46 | 0.30 | 0.36 | 37 |
| No Injury | 0.00 | 0.00 | 0.00 | 7 |
| accuracy | | | 0.54 | 202 |
| macro avg | 0.37 | 0.35 | 0.36 | 202 |
| weighted avg | 0.52 | 0.54 | 0.53 | 202 |

Figure 5. Results of Neural Network

Both Random Forest and Gradient Boosting models perform slightly better than the SVM but are comparable or slightly worse than the Logistic Regression model in terms of overall accuracy and weighted average F1-score. When comparing the Neural Network model to the previous models, The Neural Network's performance is like that of the Logistic Regression and Random Forest models, with slightly better accuracy than the SVM and Gradient Boosting models. Like the other models, the Neural Network struggles with the "No Injury" class due to the low number of instances (support is 7), which may be causing issues related to class imbalance.

The overall performance of the Neural Network is moderate, with room for improvement. Techniques such as adjusting network architecture, hyperparameter tuning, or incorporating techniques to handle class imbalance might improve the performance. Given the consistent performance across multiple models, it's likely that the inherent difficulty in predicting certain classes is due to the nature of the dataset itself, potentially requiring a review of the data and feature engineering to achieve significant improvements.

## 7. CONCLUSIONS

The project on road crash prediction modeling using machine learning and ensemble classification techniques has provided valuable insights into the factors contributing to road accidents and the severity of the resulting injuries. Through the analysis of road crash data from selected areas of Jaipur city, the study has demonstrated the effectiveness of ensemble and machine learning algorithms in building prediction models. The research has highlighted the importance of scientific methods for data collection, analysis, and prediction modeling in addressing the critical issue of road safety. The findings of the study can be utilized to formulate data-driven interventions for reducing road crashes and resulting fatalities not only in Jaipur but also in other cities and locations. The project has laid the groundwork for further research and the development of sustainable solutions backed by data-driven strategies to improve road safety and prevent irreversible loss to human life and public assets caused by road accidents.

Further approaches can be worked out in next stage of learning which include neural networks, approximate clustering, and deep learning techniques. Information about unknown vehicles projected using the best performance model could be of extreme importance in averting collisions, as well as developing improved road safety plans.

## ACKNOWLEDGEMENTS

The data used for analysis and classification using Ensemble and Machine Learning algorithms for writing this paper has been officially sourced from various Police Stations of Jaipur City. The Department of Transport & Road Safety, Government of Rajasthan is the nodal department for road safety initiatives being taken in the state and all the line departments function in collaboration with the Lead Agency setup within the Transport Department and directly reports to the Supreme Court Committee on Road Safety. An official written consent has also been obtained by the author from the Department of Transport & Road Safety, Government of Rajasthan for collecting, processing, and analyzing road crash data for the purpose of research and development in the field of road safety.

## REFERENCES

[1] A. Shaban and S. Sattar, "Mobility and transport infrastructure in Mumbai Metropolitan Region: growth, exclusion and modal choices," Urban, Planning and Transport Research, vol. 11, no. 1, 2023. [Online]. Available: https://doi.org/10.1080/21650020.2023.2212745

[2] C. Ma, W. Hao, W. Xiang, and W. Yan, "The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents," Journal of Advanced Transportation, 2018.

[3] X. Pei, S. C. Wong, and N. N. Sze, "A joint-probability approach to crash prediction models," Accident Analysis & Prevention, vol. 43, no. 3, pp. 1160–1166, 2011.

[4] R. Mesa-Arango, V. G. Valencia-Alaix, R. A. Pineda-Mendez, and T. Eissa, "Influence of socioeconomic conditions on crash injury severity for an urban area in a developing country," Transportation Research Record, vol. 2672, no. 31, pp. 41–53, 2018.

[5] Q. Zeng and H. Huang, "A stable and optimized neural network model for crash injury severity prediction," Accident Analysis & Prevention, vol. 73, pp. 351–358, 2014.

[6] N. Fiorentini and M. Losa, "Handling imbalanced data in road crash severity prediction by machine learning algorithms," Infrastructures, vol. 5, no. 7, p. 61, 2020.

[7] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," Accident Analysis & Prevention, vol. 108, pp. 27–36, 2017.

[8] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing prediction performance for crash injury severity among various machine learning and statistical methods," IEEE Access, vol. 6, pp. 60079–60087, 2018.

[9] Q. Hou, X. Huo, J. Leng, and F. Mannering, "A note on out-of-sample prediction, marginal effects computations, and temporal testing with random parameters crash-injury severity models," Analytic Methods in Accident Research, vol. 33, Article 100191, 2022.

[10] M. Islam and F. Mannering, "The role of gender and temporal instability in driver-injury severities in crashes caused by speeds too fast for conditions," Accident Analysis & Prevention, vol. 153, Article 106039, 2021.

[11] M. Abdel-Aty, "Analysis of driver injury severity levels at multiple locations using ordered probit models," Journal of Safety Research, vol. 34, no. 5, pp. 597–603, 2003.

[12] S. R. Hu, C. S. Li, and C. K. Lee, "Investigation of key factors for accident severity at railroad grade crossings by using a logit model," Safety Science, vol. 48, no. 2, pp. 186–194, 2010.

[13] S. Yasmin, N. Eluru, and S. V. Ukkusuri, "Alternative ordered response frameworks for examining pedestrian injury severity in New York City," Journal of Transportation Safety & Security, vol. 6, no. 4, pp. 275–300, 2014.

[14] G. S. Tulu, S. Washington, M. M. Haque, and M. J. King, "Injury severity of pedestrians involved in road traffic crashes in Addis Ababa, Ethiopia," Journal of Transportation Safety & Security, vol. 9, sup1, pp. 47–66, 2017.

[15] K. Wang, N. Shirani-Bidabadi, M. R. R. Shaon, S. Zhao, and E. Jackson, "Correlated mixed logit modeling with heterogeneity in means for crash severity and surrogate measure with temporal instability," Accident Analysis & Prevention, vol. 160, Article 106332, 2021.

[16] S. Chen, S. Zhang, Y. Xing, and J. Lu, "Identifying the factors contributing to the severity of truck-involved crashes in Shanghai River-Crossing Tunnel," International Journal of Environmental Research and Public Health, vol. 17, no. 9, p. 3155, 2020.

[17] L. Wahab and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," PLoS One, vol. 14, no. 4, Article e0214966, 2019.

[18] M. Ghasedi, M. Sarfjoo, and I. Bargegol, "Prediction and analysis of the severity and number of suburban accidents using logit model, factor analysis and machine learning: A case study in a developing country," SN Applied Sciences, vol. 3, pp. 1–16, 2021.

[19] P. Infante et al., "Comparison of statistical and machine-learning models on road traffic accident severity classification," Computers, vol. 11, no. 5, p. 80, 2022.

[20] Y. Yang, K. He, Y. P. Wang, Z. Z. Yuan, Y. H. Yin, and M. Z. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," Physica A: Statistical Mechanics and Its Applications, vol. 595, Article 127083, 2022a.

[21] Y. Yang, K. Wang, Z. Yuan, and D. Liu, "Predicting freeway traffic crash severity using XGBoost-Bayesian network model with consideration of features interaction," Journal of Advanced Transportation, 2022b, Article 4257865.

[22] S. Jafarzadeh Ghoushchi, S. Shaffiee Haghshenas, A. Memarpour Ghiaci, G. Guido, and A. Vitale, "Road safety assessment and risks prioritization using an integrated SWARA and MARCOS approach under spherical fuzzy environment," Neural Computing and Applications, vol. 35, no. 6, pp. 4549–4567, 2023.

[23] E. A. Atumo, T. Fang, and X. Jiang, "Spatial statistics and random forest approaches for traffic crash hot spot identification and prediction," International Journal of Injury Control and Safety Promotion, vol. 29, no. 2, pp. 207–216, 2022.

[24] M. Ghasedi, M. Sarfjoo, and I. Bargegol, "Prediction and analysis of the severity and number of suburban accidents using logit model, factor analysis and machine learning: A case study in a developing country," SN Applied Sciences, vol. 3, pp. 1–16, 2021.

[25] R. Zinno, S. S. Haghshenas, G. Guido, and A. Vitale, "Artificial intelligence and structural health monitoring of bridges: A review of the state-of-the-art," IEEE Access, vol. 10, pp. 88058–88078, 2022.

[26] B. C. Choo, M. Abdul Razak, A. B. Dayang Radiah, M. Z. Mohd Tohir, and S. Syafiie, "A review on supervised machine learning for accident risk analysis: Challenges in Malaysia," Process Safety Progress, vol. 41, pp. S147–S158, 2022.

[27] Y. Song, S. Kou, and C. Wang, "Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach," Journal of Safety Research, vol. 76, pp. 64–72, 2021.

[28] K. Topuz and D. Delen, "A probabilistic Bayesian inference model to investigate injury severity in automobile crashes," Decision Support Systems, vol. 150, Article 113557, 2021.

[29] G. Shiran, R. Imaninasab, and R. Khayamim, "Crash severity analysis of highways based on multinomial logistic regression model, decision tree techniques, and artificial neural network: A Modeling comparison," Sustainability, vol. 13, no. 10, p. 5670, 2021.

[30] M. Mohanty et al., "Development of crash prediction models by assessing the role of perpetrators and victims: A comparison of ANN & logistic model using historical crash data," International Journal of Injury Control and Safety Promotion, pp. 1–17, 2022.

[31] A. Danesh, M. Ehsani, F. Moghadas Nejad, and H. Zakeri, "Prediction model of crash severity in imbalanced dataset using data leveling methods and metaheuristic optimization algorithms," International Journal of Crashworthiness, vol. 27, no. 6, pp. 1869–1882, 2022.

[32] S. Koramati, A. Mukherjee, B. B. Majumdar, and A. Kar,"Development of crash prediction model using Artificial Neural Network (ANN): A case study of Hyderabad, India," Journal of The Institution of Engineers (India): Series A, vol. 104, no. 1, pp. 63–80, 2023.

[33] Global Status Report on Road Safety, World Health Organization, 2018.

[34] IRF World Road Statistics, Geneva: International Road Federation, 2020.

[35] Road Accidents in India, Ministry of Road Transport and Highways, Government of India, 2021.

[36] "India losses 3% of its GDP to road accidents: UN study," Economic Times Auto. [Online]. Available: https://auto.economictimes.indiatimes.com/news/industry/india-losses-3-of-its-gdp-to-road-accidents-un-study/55700816

[37] "The problem," New Delhi: Save Life Foundation. [Online]. Available: https://savelifefoundation.org/the-problem/

[38] Guide for Road Safety Opportunities and Challenges: Low- and Middle-Income Countries Country Profiles, Washington, DC., USA: World Bank, 2019.

[39] D. Mohan, G. Tiwari, and K. Bhalla, Road Safety in India: Status Report, Transportation Research & Injury Prevention Programme, Indian Institute of Technology, New Delhi, 2020.

[40] C. Gutierrez – Osorio and C. Pedraza, "Modern Data Sources and Techniques for Analysis and Forecast of Road Accidents: A Review," Journal of Traffic and Transportation Engineering (English Edition), 2020.

[41] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," 7th International Conference on Smart Computing & Communications (ICSCC), 2019.

[42] S. N. Babu and J. Tamilselvi, "Generating Road Accident Prediction Set with Road Accident Data Analysis Using Enhanced Expectation Maximization Clustering Algorithm and Improved Association Rule Mining," Journal Europeen des Systemes Automatises, vol. 52, no. 1, 2019.

[43] G. P. Karthik, S. B. and S. T. Sudalaimuthu, "Analysis of Road Accidents using Machine Learning," International Journal of Advanced Science & Technology, vol. 29, no. 6, 2020.

[44] S. Krishnaveni and M. Hemalatha, "A Perspective Analysis of Traffic Accident using Data Mining Techniques," International Journal of Computer Applications, vol. 23, no. 7, 2011.

[45] J. Lee, T. Yoon, S. Kwon, and J. Lee, "Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study," MDPI Applied Sciences. [Online]. Available: https://www.mdpi.com/2076-3417/10/1/129

[46] S. Ferreira, M. Andrade, and P. B. Silva, "Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review," Journal of Traffic and Transportation Engineering (English Edition), 2020.

[47] P. A. Nandurge and N. V. Dharwadkar, "Analyzing Road Accident Data using Machine Learning Paradigms," International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud), 2017.

[48] B. Bhatnagar and S. Srivastava, "A Robust Model for Churn Prediction using Supervised Machine Learning," 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, pp. 45–49, 2019. doi: 10.1109/IACC48062.2019.8971494

[49] H. Jain, A. Khunteta, and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost," International Conference on Computational Intelligence and Data Science, vol. 167, pp. 101-112, 2020.

[50] A. Bagga, S. Srivastava, and R. Shekhawat, "Review of the Machine Learning Techniques in Road Crashes," International Conference on Computation, Automation and Knowledge Management, pp. 373–376, 2020. doi: 10.1109/ICCAKM46823.2020.9051506

[51] A. Bagga, S. Srivastava, and R. S. Shekhawat, "Road Crash Data Analysis and Prediction Using Machine Learning Algorithms," 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), 2022. doi: 10.1109/ICCCMLA56841.2022.9989171

[52] "Random Forest Algorithm," Javatpoint. [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm

[53] "Decision Tree Classifier Explained in Real Life: Picking a Vacation Destination," Towards Data Science. [Online]. Available: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575

[54] "Regression Trees - Decision Tree for Regression," Medium. [Online]. Available: https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047

## BIOGRAPHIES OF AUTHORS

| | |
|---|---|
|  | **Ashwini Bagga** is Research Scholar at Manipal University Jaipur, Rajasthan, India. He Holds a master's degree in computer science. His research areas include application of Machine Learning Algorithms in Road Safety. At present he is serving as Consultant in State Road Safety Cell, Department of Tranport and Road Safety, Government of Rajasthan. During his continued journey he has been actively involved in varied activities including Policy Formulation, Planning, Participating in Road Safety Audits, Report Writing, Data Analysis, Research & Development, Training & Capacity Building and Writing Books on Road Safety as a Subject Matter Expert. Details of his professional journey in the field of research in road safety can be browsed by visiting https://ashwinibagga.com. He can be contacted at email: ashwinibagga@gmail.com. |
| | |
|  | **Dr. Sumit Srivastava** is currently Professor in department of Information Technology, Manipal University Jaipur(MUJ. He has done his Ph.D. in Data Mining from Univeristy of Rajasthan and has 21 years of Teaching and nearly 12 years of Research Experience. His area of research involves algorithms, data science, knowledge discovery, computational, agent-based modeling, hybrid dynamic systems, decentralized decision making, feature extraction, process mining, and engineering education. He has also been an invited speaker in Algorithms, Machine Learning, and Information & Knowledge Discovery at various short-term courses. He has published around 70+ research papers in SCI, Scopus & peer reviewed journals and conferences of international repute. He can be contacted at email: sumit.srivastava@jaipur.manipal.edu . |
| | |
|  | **Dr. Rajveer Singh Shekhawat** currently Technical Advisor to Persius Ou, a startup offering a creative Skills Platform for B2B. He has been Professor of Computer Science at Manipal University Jaipur during 2015-2022. My research areas include IoT & embedded systems, AI and Soft Computing, Computer Vision, GIS and Spatial Analysis, computer networks & distributed computing. I had directed strategic programs on Smart Energy Systems at Secure Meters Ltd with a special focus on innovation for product design and development. I was at Central Electronics Engg Research Institute (CSIR-CEERI) as a Scientist while driving R&D programs in IT Systems and Industrial Controls systems beginning in 1983. I also worked at University of Bremen and German National Research Centre of IT , Bonn, Germany as Visiting Scientist and Principal Investigator for international collaborative research projects. I have guided 100+ PG students and published 90+ research papers in reputed national and international conferences and as many research reports, apart from a number of patents and copy-rights to his credit. Since July 2015, my focus shifted to transfer his learnings to the budding engineers which I acquired while at R&D institutions and industry. I have obtained PhD from Birla Institute of Technology and Science (BITS) Pilani as well as MS, B. Tech (EEE), and M.Sc. (Hons) Physics. He can be contacted at email: rajveersingh.shekhawat@jaipur.manipal.edu |