Sigma J Eng Nat Sci, Vol. 44, No. 5, pp. 3567–3582, October, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00104

Research Article

Advanced predictive approaches for accident forecasting to enhance road safety

Prasad MANOJ[1,2,*], Kerekoppa Chandrashekar MANJUNATH[2], Kotagi B. PUNITH[2], S. K. VEENA[3]

[1]*Department of Civil Engineering, Vidyavardhaka College of Engineering, Mysore, (Affiliated to Visvesvaraya Technological University, Karnataka), 570002, India*
[2]*Department of Civil Engineering, The National Institute of Engineering, Mysore, (Affiliated to Visvesvaraya Technological University, Karnataka), 570008, India*
[3]*Department of Computer Engineering, Maharaja Institute of Technology, Mysore, (Affiliated to Visvesvaraya Technological University, Karnataka), 571438, India*

ARTICLE INFO

Article history
Received: 06 February 2025
Revised: 05 May 2025
Accepted: 10 July 2025

Keywords:
Blackspot Analysis; Machine Learning; Predictive Modeling; Road; Road Traffic Accidents; Safety

ABSTRACT

As a public health problem, it is important to understand the factors associated with an increased risk of injury severity in road traffic accidents. To improve road safety, make better decisions about road safety, and lessen the severity of crashes in the future, it is crucial to identify these elements. This study was designed to gather traffic data, analyze it for selection of suitable variables to be used in the prediction of accidents, and to create an accident predictive model that would be suitable for regional conditions. Random Forest, Support Vector Regression, and Multi Linear Regression models were developed to predict traffic accidents. The dataset comprised of 67 blackspots, each containing 17 variables collected using different survey from the identified blackspot of Karnataka state. Mean Absolute Error, Root Mean Squared Error, and Coefficient of Determination metrics were used to assess the models' performance after the data was divided into training and validation sets, backward stepwise Multi Linear Regression fared best with R^2 of 0. 87 for validation. The study emphasises the potential of machine learning to improve traffic safety.

Cite this article as: Manoj P, Manjunath KC, Punith KB, Veena SK. Advanced predictive approaches for accident forecasting to enhance road safety. Sigma J Eng Nat Sci 2026;44(5):3567–3582.

INTRODUCTION

The World Health Organization's Global Status Report on Road Safety (GSR) 2023 estimates that global deaths from road traffic crashes reached 1.19 million, or 15 deaths per 100,000 people, in 2021. Road traffic injuries (RWTS) have become the 12th leading cause of death among all age groups and remain the major cause of death of the 5-29 age group. In addition, traffic collisions cost a considerable

***Corresponding author.**
*E-mail address: manoj.p@vvce.ac.in

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey
© Author. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

amount of money worldwide; it has been estimated that the global economic impact of traffic accidents is as high as US$1.8 trillion, which is 10-12% of the world's gross domestic product [1]. This is a huge social and economic problem. Minimizing the severity of collisions is a great goal throughout the world. Vulnerable road users, like cyclists, motorcyclists and pedestrians, account for about 54% of all traffic deaths worldwide, and 43% in Europe [2]. While all traffic accidents are concerning, vulnerable road users, like cyclists, motorcyclists and pedestrians, represent a significant proportion of all traffic deaths around the world. According to Chang and Wang, the most significant factor influencing the severity of injuries is the category of vehicle, with susceptible drivers having the highest risk of suffering a serious injury or death [3]. Compared to car occupants, motorbike riders have a 30 times higher risk of dying in a traffic accident [4].

To increase road safety, a thorough examination of traffic accidents and the identification of the variables influencing the severity of injuries sustained in the crash interventions are essential. For statistical and machine learning techniques to effectively forecast the severity of injuries, the accessible data must be dependable and inclusive of the target population. After analysing data on traffic accidents in a few countries in southeast Europe, Laiou et al. [5] concluded that to increase road safety in such nations, it is critical to improve the quality of the data sets. One of the most crucial elements in enhancing traffic safety is data, as without it, it would be impossible to gauge the effectiveness of the current road safety initiatives [6].

The comprehensive study involved the identification and analysis of black spots in Karnataka State, focusing on comparing various analytical methods viz., Multi Linear Regression, SVR, and Random Forest algorithms for predicting accident occurrences. The overall methodology is demonstrated in the Figure 1. The uniqueness of this study is in its comprehensive approach to data collection and processing techniques, using region-specific data gathered from various accident black spots across the Karnataka state.

Literature Review

Artificial neural networks are the most widely used machine learning technique for risk assessment in engineering, with Support Vector Machines (SVM) coming in second. Various techniques, including decision trees, Bayesian networks, and Random Forest, have been developed over the years to create models for predicting the severity of injuries in traffic accidents.

SVM and Ordered Probit (OP) models were used by Li et al. [7] to analyse crash injury severity in China. This research was conducted on 1800 crash casualties. Several variables were taken into consideration, including the exit ramp's length and the shoulder width. 48.8% of the SVM model's predictions were accurate, compared to 44.0% for the OP model. Furthermore, it was found that the SVM model outperformed the OP model in terms of sensitivity and specificity [7]. Ibrahim and Far [8] created a Real-Time Transportation Data Mining approach. This technique could be used to forecast future Real Time Accidents (RTA) related data and assess a set of real-time traffic data. The study focused on the application of two decision tree types viz., Java 48 (J48) and Active Directory trees to 1385 accident records collected by the UK Department of Transport. The Active Directory tree approach registered 85.9%, while J48 registered 87.2%, based on the forecast accuracy statistics [8].

Mohamed [9] forecasted the causes of traffic accidents based on 1000 real crashes in Dubai City using SVM with Gaussian Radial Basis function (RBF). The accuracy of

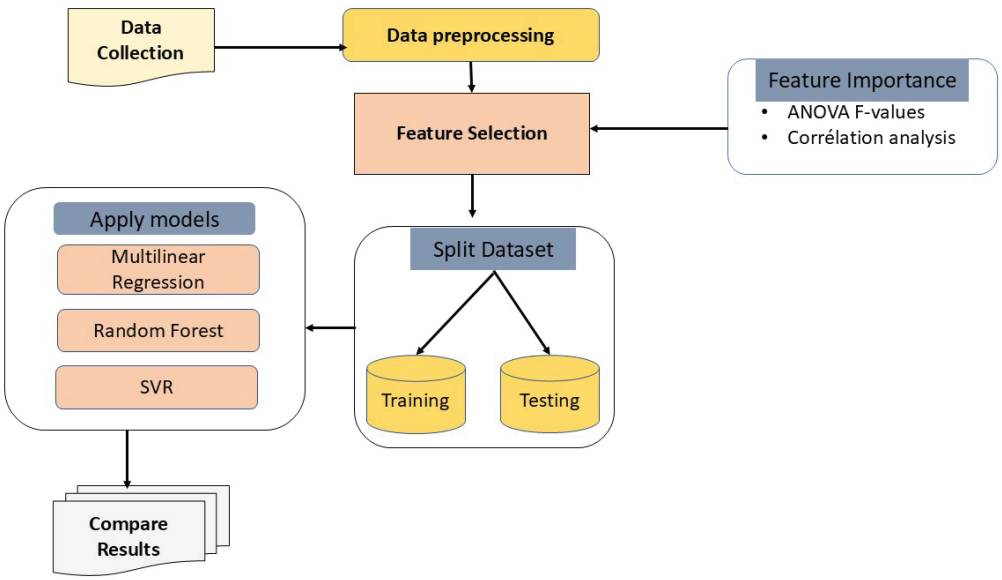


Figure 1. Proposed model.

this multi-class SVM model was over 75% [9]. To find the key factors affecting the severity prediction of crash datasets in Iran, Effati et al. [10] used an integrated method of Machine Learning (ML) that included SVM, coactive Neuro-Fuzzy, and ANN inference systems. This innovative integration technique registered good forecasts with an accuracy of 85.49% [10]. Perone [11] used SVM, Logistic Regression, Random Forest, KNN, and Naïve Bayes to create a new prediction model to evaluate the severity of injuries in Brazil using 20798 accident records from the city of Porto Alegrers. According to African Union Commission (AUC) records, the highest satisfaction rates were achieved by SVM and Logistic Regression (94%), Random Forest (93%), KNN (90%), and Naïve Bayes (83%) [11].

To determine the key factors driving most RTAs, Sharma et al. [12] collected 300 real accident cases from India using Multi Layers Perceptron (MLP) and SVM (Gaussian kernel). Three categories were applied to their data set: 70% for training, 20% for cross-validation, and 10% for testing. LIBSVM, a support vector machine library, was integrated with Octave. The results of the investigation demonstrated that SVM with a Gaussian kernel function achieved 94% prediction accuracy, whereas conventional MLP achieved only 60% [12]. Gu et al. [13] forecasted fatal traffic incidents in China using SVM. The goal of this study was to compare SVMs, KNNs, and Bayesian networks. The results showed that the particle swarm with mutation optimization-SVM-based traffic fatalities prediction model had a higher prediction precision (97%) and fewer mistakes (9%) in training and testing data. [13].

To determine the primary environmental features of RTA in the UK, Al-Radaideh and Daoud [14] used ANN backpropagation, SVM (polynomial Kernel), Decision Trees and Random Forests. The experimental results of the study showed that Random Forests, generated the highest accurate forecast of accident severity in the United Kingdom (80.6%) [14].

Farhat et al. [15] employed several ML techniques, including decision trees and artificial neural networks (ANNs), to predict traffic accidents in Lebanon. The findings show that an artificial neural network (ANN) using a Multi Layers Perceptron (MLP) with two hidden layers and 42 neurones in each layer was the best strategy with an accuracy rate of prediction (94.6%) and an AUC (95.71%) [15].

Karthik et al. [16] used various ML namely J48, Random Forest and Naïve Bayesian in the Thanjavur region of India for prediction of cause of major accidents in the area. Ten years of accident data provided by the RTA were obtained, in addition to other RTA parameters, such as surface, road boundary, accident time, and accident location. J48 outperformed the other classifiers with an accuracy rate of 56.96%, followed by Random Forest with an accuracy rate of 49% and Naïve Bayesian with an accuracy rate of 54% [16].

Apart from the existing methods of accident black spot identification and summarization, the researchers Cui et al. [17] proposed a new method called accident spacing distribution method to avoid the issue of road segmentation and provide more accurate accident black spot identification [17].

For enhancing the safety of the road, a study by Karamanlis et al. [18] was conducted in Northern Greece to identify road accident black spots using traditional and modern approaches (binary logistic regression and various machine learning techniques) [18].

Mbarek et al. [19] proposed a methodology to identify, classify and assess accident black spots on rural road networks in Morocco using the Weighted Severity Index (WSI), Bagging Extreme Learning Machine (B-ELM) and Ordinal Regression (OR) by studying the impact of environmental factors and road conditions on accident severity [19].

The authors employed artificial neural network to classify the severity of the road accidents. The model was promising, but was not successful at recognizing less frequent, more serious crashes because of data imbalance [20]. Spatial regression analysis was performed for the analysis of road traffic accidents in Türkiye (2013–2018) to elucidate space–time patterns. Results showed that accidents were not random, but instead, the pattern revealed a clustering of accidents by province. Vehicle numbers and asphalt road length were positively associated with accident rates and non-asphalt roads were not significant. Spatial factors were important but neighbouring territories' characteristics had minimum influence on local accident rates [21]. Manoj et al. [22] predicted accidents using machine learning in Karnataka with blackspot data, achieving an R2 of 0.67, demonstrating the potential of machine learning in improving road safety.

In the studies, machine learning techniques are highlighted as most important in predicting the severity of road traffic accidents, and in most cases, the SVM is able to achieve the highest accuracy; this is summarised in Table 1. Many people are interested in making better predictions, but it is difficult to overcome problems like imbalance and differences in data sets from different regions. One of the major challenges is the lack of consistency in comparisons between datasets.

We have also done a detailed review analysis, where we focused on the precise identification and management of road accident black spots to enable timely and effective remedial actions. We have reviewed existing methodologies and highlighted the open research gaps that have not been thoroughly addressed in current literature. The study proposes a multi-stage approach to improve black spot analysis: first, by managing highly complex and heterogeneous accident data; second, by resolving ambiguity in black spot detection through real-time data modelling; and third, by developing a cost-effective and intelligent classification system. The core aim is to create a reliable model capable of continuous decision-making and prediction using real-time data streams, ultimately enhancing the accuracy of accident analysis and informing geographically and geometrically targeted interventions for accident mitigation [23].

Table 1. Literature summary

Author	Methods Used	Dataset / Region	Key Findings / Accuracy
Li et al. [7]	SVM, Ordered Probit (OP)	1800 crash cases / China	SVM (48.8%) outperformed OP (44.0%) in accuracy, sensitivity, and specificity
Ibrahim & Far [8]	J48, Active Directory Tree	1385 records / UK	J48: 87.2%, Active Directory Tree: 85.9%
Mohamed [9]	SVM (RBF kernel)	1000 crashes / Dubai	Multi-class SVM accuracy: >75%
Effati et al. [10]	SVM, Coactive Neuro-Fuzzy, ANN	Crash data / Iran	Integrated model accuracy: 85.49%
Perone [11]	SVM, Logistic Regression, Random Forest, KNN, Naïve Bayes	20,798 records / Brazil (Porto Alegre)	SVM & Logistic Regression: 94%, RF: 93%, KNN: 90%, NB: 83%
Sharma et al. [12]	SVM (Gaussian kernel), MLP	300 accident cases / India	SVM: 94%, MLP: 60%
Gu et al. [13]	SVM (PSO-optimized), KNN, Bayesian Network	Traffic data / China	Optimized SVM accuracy: 97%, error: 9%
Al-Radaideh & Daoud [14]	ANN (Backprop), SVM (poly kernel), Decision Trees (RF, C4.5, CART, J45)	RTA data / UK	Random Forest accuracy: 80.6%
Farhat et al. [15]	ANN (MLP, 2 hidden layers, 42 neurons)	Accident data / Lebanon	MLP accuracy: 94.6%, AUC: 95.71%
Karthik et al. [16]	J48, Random Forest, Naïve Bayes	10 years of data / Thanjavur, India	J48: 56.96%, NB: 54%, RF: 49%
Karamanlis et al. [18]	Binary Logistic Regression, ML techniques	Northern Greece	Identified blackspots using both conventional and ML methods

MATERIALS AND METHODS

Objective

The current study demonstrates how various datasets; ML algorithms and methodologies are used to address the complex and vital issue of road safety. Overall, it demonstrates the multiple nature of research in forecasting and analysing the severity of traffic accidents. Through data-driven insights, the research contributes to the advancement of prediction models that may help avoid and mitigate the severity of traffic accidents.

Study Area and Data Collection

A road accident black spot is defined as a stretch of highway approximately 500 meters in length where, over the last three calendar years, either five road accidents involving fatalities or grievous injuries have occurred, or ten fatalities have taken place as per the guidelines of Ministry of Road Transport and Highways (MoRTH), 2019.

The Planning and Road Asset Management Centre (PRAMC), a state government undertaking situated in Bangalore, Karnataka has identified several black spots across Karnataka State, India. However, 67 blackspots were considered for the study, and their locations are shown in Figure 2.

Primary Data: Primary data collection was done using multiple surveys such as Road Inventory survey and Spot speed study data were collected and recorded in an excel sheet for further analysis. The road inventory and detailed survey consisted of using a Total Station to take road geometric measurements. Black spot areas were also documented with further details in each area. A radar speed gun was used to measure the speed of traffic by detecting the Doppler effect on the frequency emitted by the radar gun. A handheld radar gun has been used exclusively on

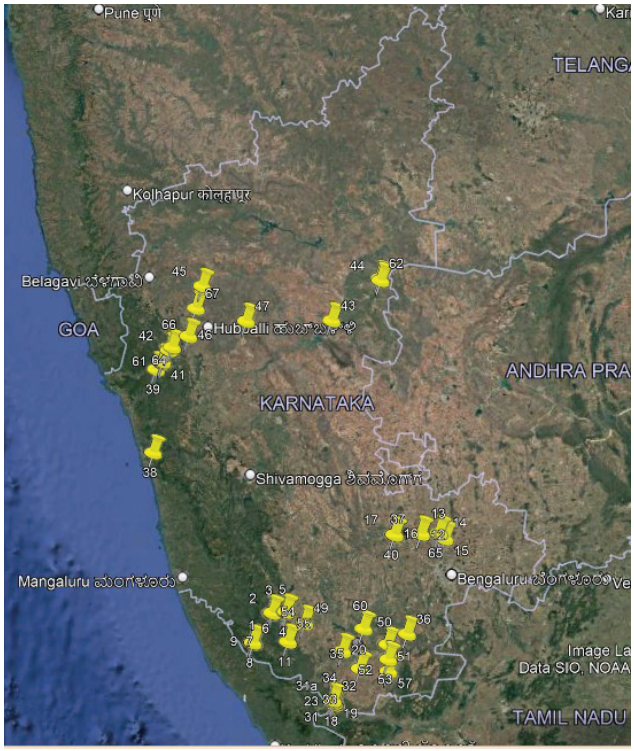


Figure 2. Locations of selected blackspots in State of Karnataka, India.

two-wheelers and four-wheelers to determine speeds at all black spots. The average speed at the spots was then calculated and used for modelling.

Secondary Data: First Information Reports (FIRs), vehicle insurance claim reports and traffic volume data was used as secondary data for the study. The accident data was collected from FIRs, which were obtained with the permission of the Superintendent of Police and collected from police stations in Mysuru, Tumkuru, Ramanagaram, Chamarajanagar, and Kodagu. The reports from 2017-2021 have reported accidents based on various Indian Penal Code (IPC) sections for different vehicles such as tempos, trucks and buses. Copies of FIRs were also downloaded from the Karnataka State Police portal from the black spot areas and these detailed information about the accidents included the type of accident, time, accident involved individuals, causes of accident, road condition, and vehicle type. IPCs sections classified accidents as minor (Section 337), major (Section 338) and fatal (Section 304(A)).

Further, traffic volume data such as PCU values were collected from the Office of the Chief Engineer (Communication and Buildings, South), Bangalore, which gave the traffic volume information for some black spot stretches in districts.

This research has analyzed a small-scale data set of 67 subjects with significant context. The data were manually collected during real time field surveys on curved road sections, a road area well known for its high accident numbers. The small sample size is a limitation in generalizability of findings, but the data provide important location-specific insights. The aim of this exploratory approach is to gain a better understanding about curved segments that are prone to accidents. Furthermore, the data collection process and methodological framework have been transparently described to acknowledge and address the inherent limitations of the study.

Data Preparation

On two-lane undivided rural highways, black spots were examined with an emphasis on curved road segments only. Software viz., MS Excel, AutoCAD and E-Survey CAD were used to process survey data. Accidents were categorised as Minor, Major and Fatal using FIR data and Equivalent Accident Number (EAN) was calculated based on the severity of accidents as per Indian Road Congress (IRC) 131-2022 under section 4.5.1.

A weighted total of accident severity, with severity weights obtained from IRC 131-2022 (minor = 2, major = 5, fatal = 10), was used to calculate EAN. More accurate accident analysis was made possible by this weighted system. The equation1 for calculating EAN is represented below.

$$EAN = (2X minor\ accidents) + (5X major\ accident) + (10X Fatal\ accident) \quad (1)$$

The computed EAN for each Black Spot was then fixed as a Dependent Variable (DV) and processed data consisting of 16 variables were considered Independent Variables (IDV) for model development.

Selection of variables

List of variables: After processing the data, 17 variables were selected for model development and the statistical details of which 16 Independent Variables (IDV) which are used for the prediction of EAN are presented in Table 2. For every variable, important statistics are given, including

Table 2. The statistical details of selected variables

Sl. No.	Variables	Mean	Median	Mode	Standard deviation	Minimum	Maximum
1	**Radius [R] in m**	126	97.5	247	96.9	13.4	414
2	**Speed @ Entry [V(entry)] in Kmph**	65.3	66	66	3.07	58	72
3	**Speed @ Exit [V(exit)] in Kmph**	66.3	67	67	3.4	56	73
4	**Length of Transition Curve [Ls] in m**	46.9	26.2	10	60.5	6.05	382
5	**Tangent Length [L(tangent)] in m**	42.9	36.2	47.8	27.8	3.55	118
6	**Superelevation [e] in mm**	4.95	4.92	4.64	0.853	2.5	6.8
7	**Sight distance [SD] in m**	67.1	50	20	46.5	10	200
8	**Deflection Angle [D(angle] in m**	53.5	48	22.3	26	17.9	177
9	**Total width [TW] in m**	11.4	11	11	3.8	5.5	21
10	**Carriage Way Width [CW] in m**	8.22	7.5	5.5	2.69	3.75	15
11	**Shoulder width (Left) [SW(L) in m**	1.81	1.5	0	1.97	0	8.94
12	**Shoulder width (Right) [SW(R)] in m**	1.44	1.2	0	1.61	0	7
13	**Long Chord (LC) in m**	91.7	78.5	96.7	59.2	7.9	256
14	**Appex Distance (Es) in m**	12.5	8.16	4.77	12.6	1.07	77.3
15	**Mid Speed [V(mid)] in m**	50.5	52	56	10.3	20	67
16	**Passenger Car Unit [PCU]**	7109	5879	5462	3751	1769	25000
17	**Equivalent Accident Number [EAN]**	58.9	62	43	25.4	0	99

the mean, median, mode, standard deviation, minimum, and maximum. For instance, the speed at entry and exit varies very little, with a mean of about 65–66 m, and the mean radius is 126 m with a standard deviation of 96.9. To illustrate the distribution of these variables with their corresponding extremes, the dataset additionally contains measurements such as the Length of Transition curve, Tangent Length, Carriage Way width and Shoulder width. Some variables, such as Passenger Car Unit (PCU) have a wide range of values (7109 to 25000), which indicate a wide range of observed conditions.

Correlation analysis: The correlation analysis of the variables used to estimate the Equivalent Accident Number (EAN) is shown in Figure 3. The correlation matrix heatmap above displays the relationships between various variables related to road geometry, vehicle speed, sight distance, and accident data in terms of EAN. The colour scale indicates the strength and direction of correlations, with dark red representing strong positive correlations and dark blue indicating strong negative correlations. For instance, "Radius" shows a strong positive correlation with "Sight distance" (0.88) and "Tangent Length" (0.83), suggesting that a larger road radius may be associated with increased sight distance and tangent length. On the other hand, there is a strong negative correlation between "Sight distance" and "EAN" (-0.79), which means that increased visibility may help to decrease the frequency or severity of accidents. Likewise, "Speed at Entry" and "Speed at Exit" show moderate positive correlation with "EAN" (0.45 and 0.47 respectively) meaning that higher speeds at these points could be associated with more severe accidents. Designing safer road infrastructure may benefit from this analysis and it facilitates in identifying the traffic and road features most frequently linked to accidents.

Data analysis

Statistical analysis and ANOVA tests are carried out for the complete data set to check the suitability of data for model development and the details are shown in Table 3. With a multiple R of 0.94, an R-squared value of 0.88, and an adjusted R-squared value of 0.85, the regression analysis shows high predictive performance. This means that the model explains around 88.4% of the variations in the dependent variable. The ANOVA results indicate that there is a significant overall model fit ($F = 23.83$, $p < 0.001$). Radius (negative effect, p<0.001p < 0.001p<0.001), Sight Distance (negative effect, p=0.014), D Angle (negative effect, p<0.001p < 0.001p<0.001), Shoulder Width (L) (negative effect, p=0.007), Appex Distance (Es) (positive effect, p=0.031), Mid Speed (positive effect, p<0.001), Speed @ Entry (positive effect, p<0.001), Ls (positive effect, p<0.001), e (positive effect, p<0.001p), and PCU (negative effect, p=0.01). Tan Length, Total Width, and Speed @ exit are non-significant predictors (p>0.05). The statistical robustness of the model is further supported by the intercept (−308.99, p<0.001). These findings highlight how crucially geometrical and speed-related elements affect the outcome variable.

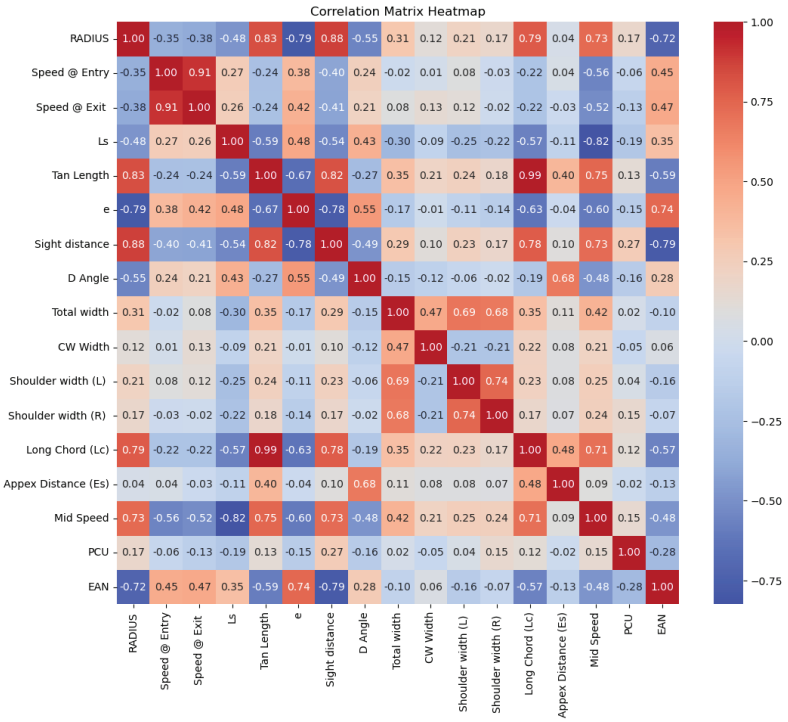


Figure 3. Correlation between the variables.

Table 3. Regression results

Regression Statistics					
	Multiple R	**R Square**	**Adjusted R Square**	**Standard Error**	**Observations**
	0.94	0.88	0.85	9.92	67.00
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	16.00	37541.85	2346.37	23.83	5.95E-18
Residual	50.00	4923.83	98.48		
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**
Intercept	-308.99	53.64	-5.76	5.15262E-07	-416.73
RADIUS	-0.19	0.04	-4.49	4.24471E-05	-0.28
Speed @ Entry	4.03	0.99	4.09	0.000155253	2.05
Speed @ Exit	0.13	0.93	0.14	0.89	-1.74
Ls	0.20	0.05	4.19	0.000111663	0.11
Tan Length	-0.63	0.95	-0.67	0.51	-2.54
e	12.71	2.71	4.68	2.20093E-05	7.26
Sight distance	-0.16	0.06	-2.55	0.013778276	-0.28
D Angle	-0.70	0.15	-4.84	1.30506E-05	-1.00
Total width	1.15	1.71	0.68	0.50	-2.28
CW Width	-2.23	1.61	-1.38	0.17	-5.46
Shoulder width (L)	-5.49	1.95	-2.81	0.01	-9.41
Shoulder width (R)	3.83	2.12	1.81	0.08	-0.42
Long Chord (Lc)	0.34	0.44	0.77	0.44	-0.54
Appex Distance (Es)	0.68	0.30	2.23	0.03	0.07
Mid Speed	1.99	0.38	5.23	3.3664E-06	1.22
PCU	-0.0009	0.0004	-2.6780	0.01	-0.002

Model Development

The dataset was divided into 80% testing and 20% validation for the application of regression models and the following regression models were applied.`

Multi Linear Regression: Stepwise Backward Multi Linear Regression is a feature selection technique that finds the most influential variables by methodically eliminating the least significant predictors from a regression model. It iteratively eliminates variables with the highest p-values from the initial set of predictors, then refits the model until only statistically significant predictors are left. This approach focuses on the most important predictors, which helps to avoid overfitting and improves the model's interpretability and simplicity. It is useful for modelling simplification, but may be sensitive to multicollinearity and may not be able to capture non-linear interaction. It is helpful in accident analysis for pinpointing important causes of accidents and improving the precision and focus of models.

The algorithm starts with the formation of an initial multiple linear regression model with all variables. The equation of the model is written as follows: equation 2

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon \quad (2)$$

where: Y is the dependent variable,

$X_1, X_2,\ldots, X_p$ are the predictors,

$\beta 0, \beta_1,\ldots,\beta_p$ are the coefficients of the predictors, and ϵ is the error term.

The predictors are checked for statistical significance (p-value) and least significant predictor (p-value>0.05) is eliminated step by step. The model is refitted after each removal, and this process continues until all predictors in the model have statistically significant p-values. The final model is expressed as in equation 3:

$$Y = \beta 0 + i = 1 \sum k\beta iXi + \epsilon \quad (3)$$

where k ≤ pk and each X_i represents a statistically significant predictor.

Support Vector Machine (SVM) Regression: SVR is a version of the SVM algorithm that is used for linear and non-linear data, it uses Kernel functions to map the data in a higher dimension. In this investigation, SVR looks for a function that deviates from the actual data points by the least amount, given a given margin of tolerance. This feature allows for capturing of the complex relationships between features and accident severity, avoiding overfitting. SVM is very robust to outliers because it works on the basis

of accident data that can include some anomalous values in the presence of extremely rare events [24].

The goal of the SVR is to obtain a function in equation 4.

$$f(x) = \omega^T \phi(\mathrm{x}) + \mathrm{b} \quad (4)$$

that has at most ϵ epsilon ϵ-deviation from the actual target values y, while minimizing the norm of *w*

Random Forest Regression: Random Forest is an ensemble method, which uses multiple decision trees while training and combines their predictions by averaging them to reduce overfitting and get better prediction accuracy. It is valuable for collision prediction problems where multiple factors (traffic speed and road geometry) are all affecting the severity of the collision at the same time, because it works well with complex, non-linear relationships among the factors. The model is also able to calculate the feature importance, which helps in guiding decision making for road safety improvements, understanding the factors with the greatest influence on accident frequency and severity [25].

The prediction of Random Forest Regression is the average of predictions from all individual decision trees as in equation 5:

$$y = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \quad (5)$$

Where:

T Number of trees in the forest.

$h_t(x)$: Prediction from the tth decision tree.

Model Validation: Two metrics to measure the regression models are Error rate and Coefficient of determination.. If the discrepancy between actual and predicted values for the training and validation data sets is negligible and unbiased, the regression model is good. Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R2 value are used to check the performances of the regression models.

The model evaluation metrics MAE, RMSE, and R^2 show good predictive performance, with training R^2 at 0.86 and validation R^2 at 0.87, confirming its robustness on unseen data. And is represented in Table 4.

The models Random Forest and Support Vector Regression (SVR) gave lower R^2 values than the Multi Linear Regression model, also shown in Table 3 and Figure 4, suggesting they may be less appropriate for this dataset compared to the Multi Linear Regression model.

The Random Forest model exhibits clear signs of overfitting, as indicated by a high training R^2 of 0.95 and a significantly lower validation R^2 of 0.58. The model is able to fit very well on the training set, with a low MAE (3.8) and RMSE (5.79); however, the performance of the model is not good when tested on the validation set with MAE (16.4) and RMSE (18.81) which showed significant increase. This indicates that the model has learned patterns and noise from the training data, and is not able to generalize well to new unseen data. Random Forest can suffer from overfitting due to its complexity, for example, trees that are too deep, not enough regularization, and it can be mitigated by pruning, making trees shallower, and cross-validation to boost generalization performance.

Table 4. Performance of regression models

Model Name	Model Stage	MAE	RMSE	R^2
Multi Linear Regression	Training	7.02	9.54	0.86
	Validation	6.81	8.50	0.87
Random Forest	Training	3.8	5.79	0.95
	Validation	16.4	18.81	0.58
Support Vector Regressor	Training	20.74	24.97	0.05
	Validation	19.72	24.31	0.30

RESULTS AND DISCUSSION

Visualising the actual versus expected values, Figure 5 shows how three machine learning models viz., Multiple

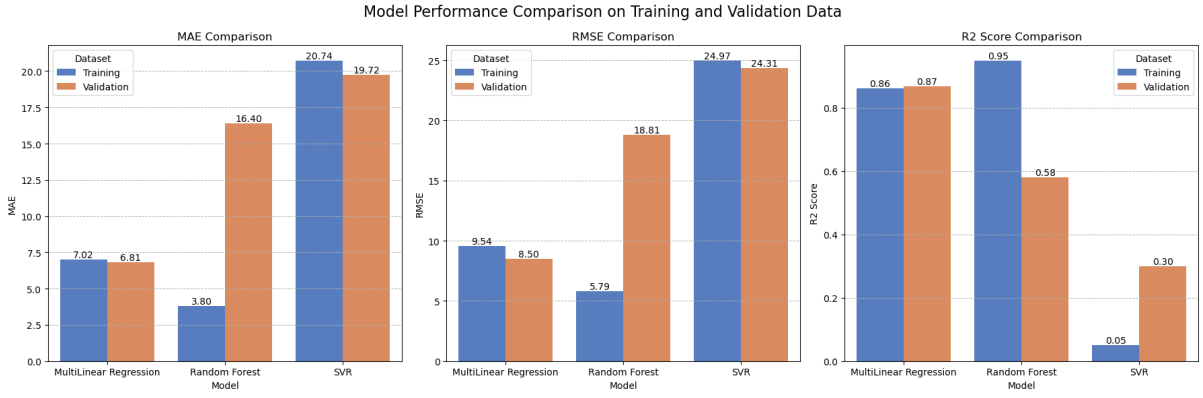


Figure 4. Comparison of different models' performance.

Linear Regression, Random Forest, and Support Vector Regression (SVR) perform on training and validation data. Model predictions for the training data are displayed in the first row. The points in the vicinity of the regression line in the plots suggest that Random Forest (b) and Multiple Linear Regression (a) are well fit to the actual values. Support Vector Regression (c) makes less accurate predictions as it has more variance and wide confidence interval. Although the variance has somewhat risen in comparison to the training data.

The validation data results in the Figure 6 show that Multiple Linear Regression (a) and Random Forest (b) are still doing well, with predictions nearly matching actual values. On the other hand, SVR (c) has a broader confidence interval and a flatter trend, indicating that it may not be a good fit for the validation data.

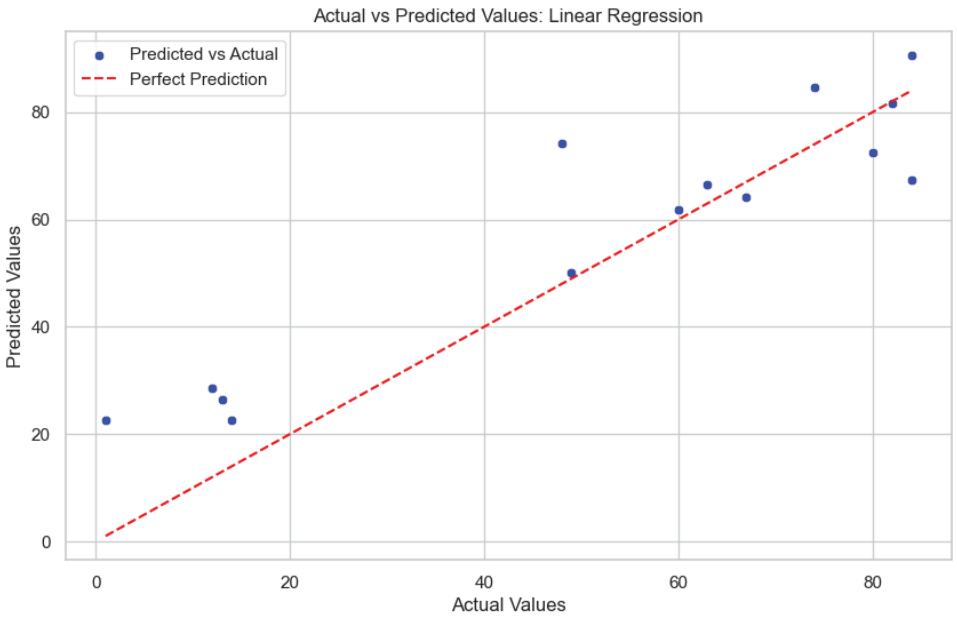


Figure 5. a) The actual and predicted values for training data: Multi Linear Regression.

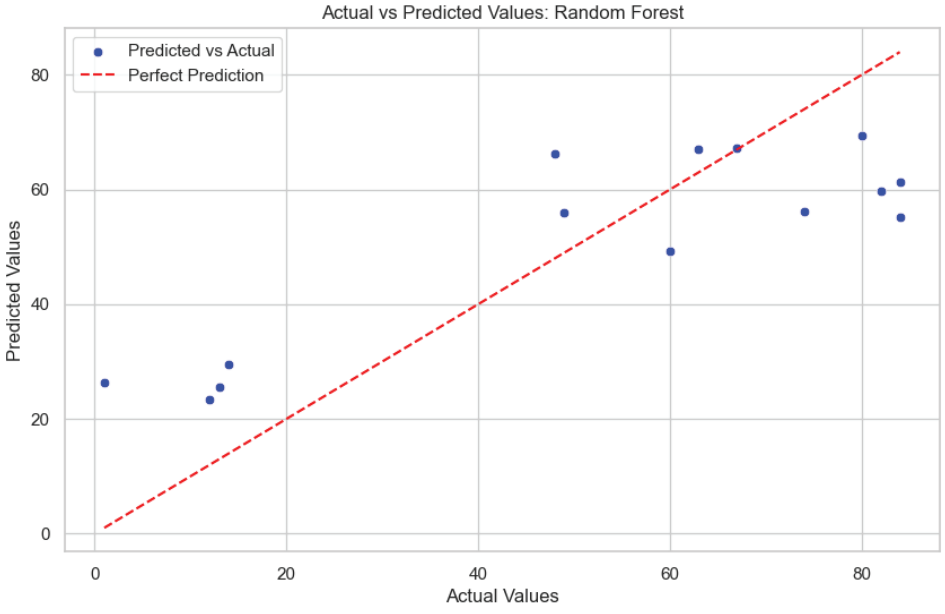


Figure 5. b) The actual and predicted values for training data : Random Forest.

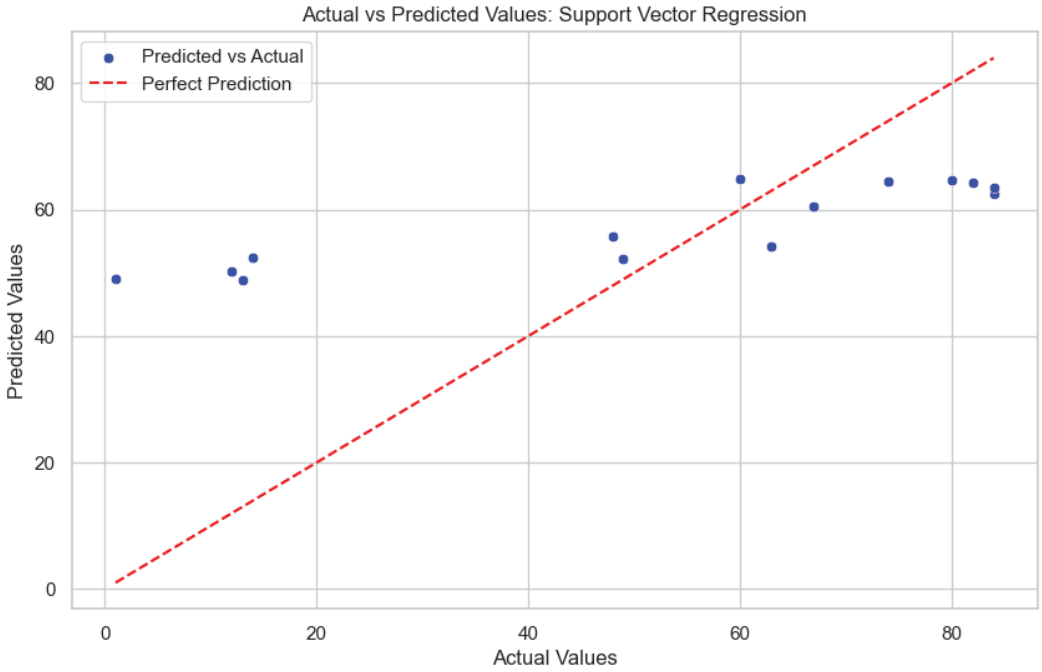


Figure 5. c) The actual and predicted values for training data : Support Vector Regression.

Overall, the generalisability and predictive capability of these two models (Random Forest and Multiple Linear Regression) are superior to that of the SVR model.

Multiple linear regression is a regression model where each of the predictors is present in the model simultaneously to estimate what role each one plays in the response. Multiple linear regression is a type of regression in which the model is refined by the backward approach, also known as backward Multi Linear Regression or backward elimination. It starts by adding all possible predictors and then discards the least significant ones, based on criteria such as P-values, until only those that have a significant effect



Figure 6. a) The actual and predicted values for validation data: Multi Linear Regression.

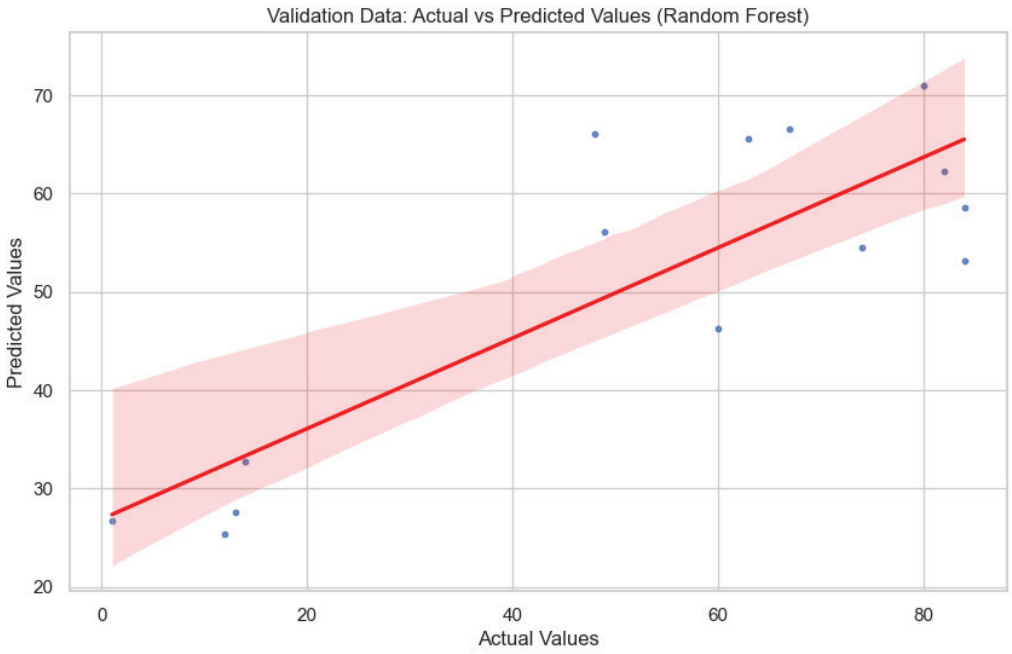


Figure 6. b) The actual and predicted values for validation data: Random Forest.

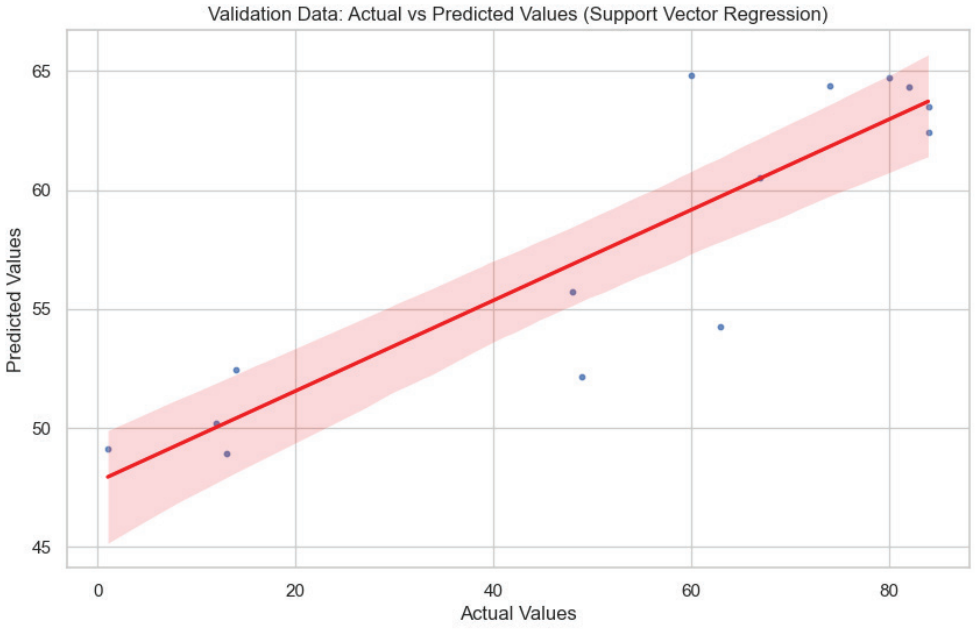


Figure 6. c) The actual and predicted values for validation data: Support Vector Regression.

are left. It will eliminate variables unrelated to the dependent variable and will prevent overfitting the model and make it easier to interpret. In conclusion, backward Multi Linear Regression is a linear regression with additional stages of selection of the "best" predictors by their elimination. In summary, while linear regression fits a model to all provided predictors, backward Multi Linear regression involves an additional step of selecting the most relevant predictors through elimination.

Figure 5[a-c] show the scatter plots comparing actual and predicted values on the training data set for three regression models. The red dashed line represents the

ideal (perfect) prediction line. Closer clustering of points along this line indicates better predictive performance. Multi Linear Regression (a) seems to be more accurate than Random Forest (b) and SVR (c), as the data points are closer to the ideal line.

Figure 6[a-c] is the scatter plot that compares the actual values (x-axis) with the predicted values (y-axis) for the training dataset. The red line is the regression line and the shaded area is the confidence interval. The nearer points are to the red line, the more accurate the predictions. Multi Linear Regression for the training data set shows that the R squared (R2) value is high, meaning that the model is a good fit for the training data and that there is a good correlation between the actual values and the values predicted by the model, as evidenced by the points being close to the regression line. (b) Random Forest Regression on the validation set shows that the overall trend is captured, but has larger scatter, particularly at the extremes of actual values, indicating moderate generalization ability. SVR on the validation set is shown with a small prediction range and the majority of predictions are grouped around the regression line. The overall linear trend is obvious, but the limited variation may indicate that the model is underfitting the data or it might lack flexibility.

The combined plots of the experimental, predicted and absolute errors for each model are presented in Figure 7[a-c]

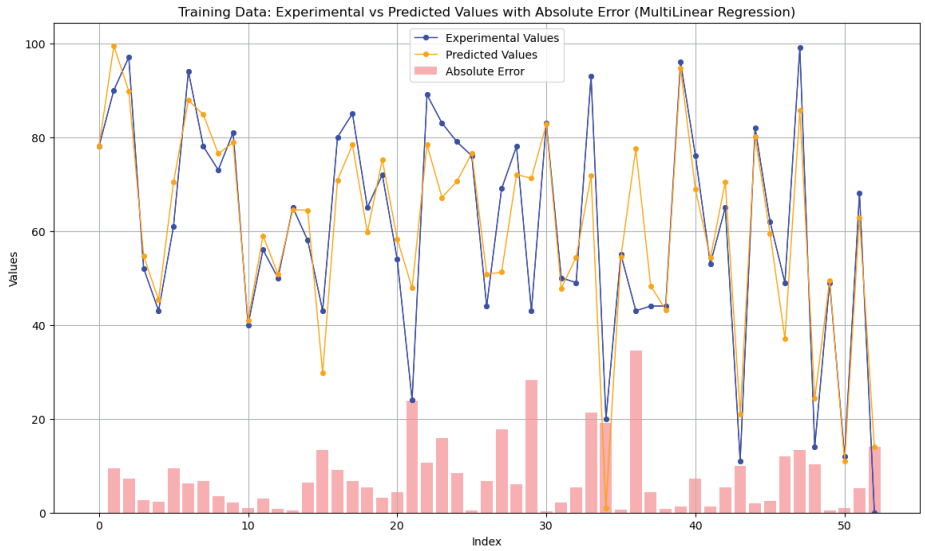


Figure 7. a) The absolute error values of the models for training data: Multi Linear Regression.

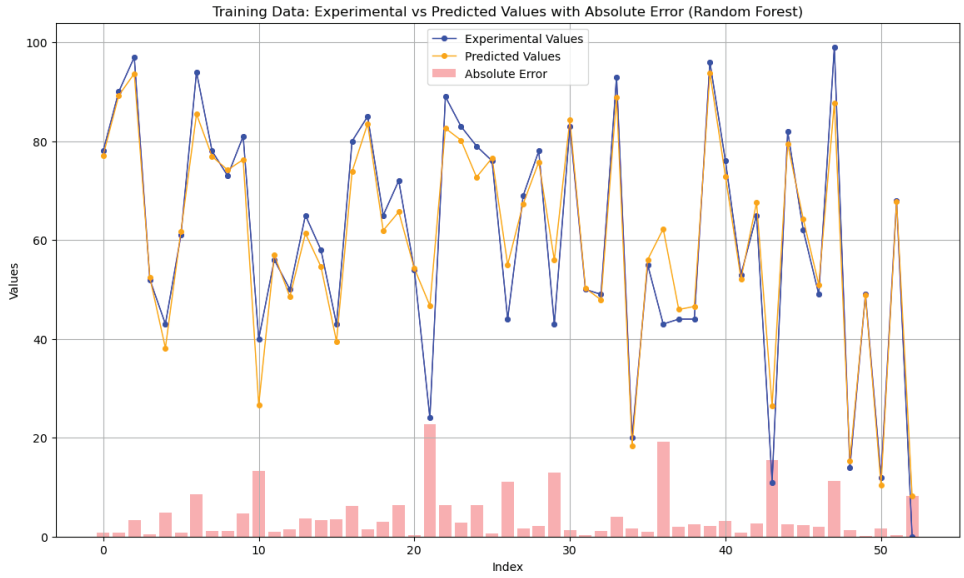


Figure 7. b) The absolute error values of the models for training data: Random Forest.

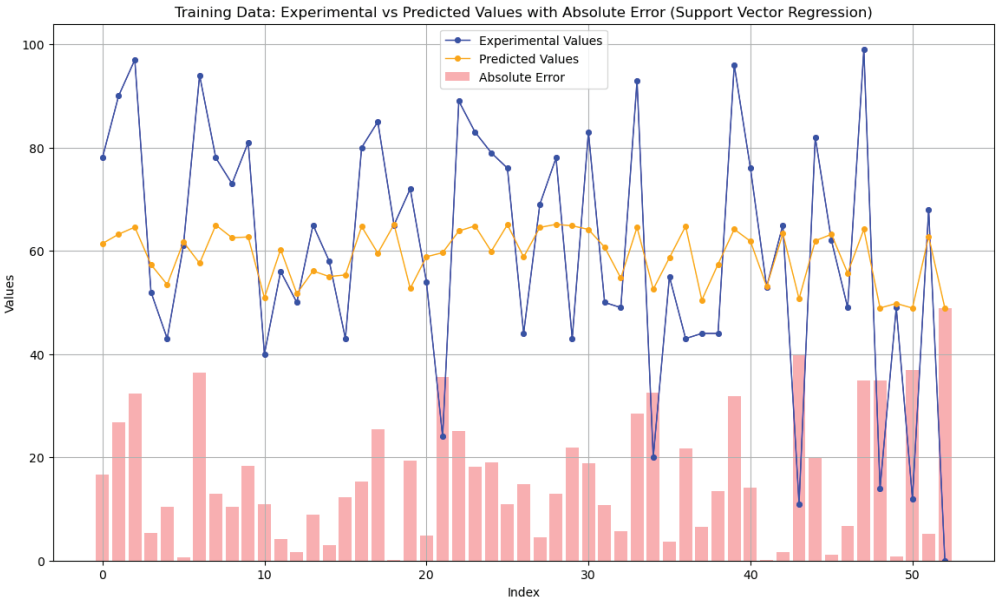


Figure 7. c) The absolute error values of the models for training data: Support Vector Regression.

to help visualize the separation between these three. The data points that were actually observed during the experiment are represented by a blue line with circular markers. The values predicted are displayed with an orange line with dot markers, and it is easy to compare with the experimental values. Pink vertical bars indicate the absolute error which displays the difference between the experimental value and the prediction. This colour scheme is useful for immediately spotting areas where the model is successful (small bars and closely aligned lines) and areas where there is significant prediction error (large bars and spread out lines). The use of these colors in all three models—Multi Linear Regression, Random Forest, and Support Vector Regression—promotes the interpretability and facilitates a good visual comparison of the models' performances.

Random Forest(b) has the lowest error bars and the best agreement between Random Forest and experimental values, among the 3 models, and thus has the highest accuracy. For Multi Linear Regression(a), the accuracy is moderate, and there are deviations and fluctuating error bars. SVM(c), on the

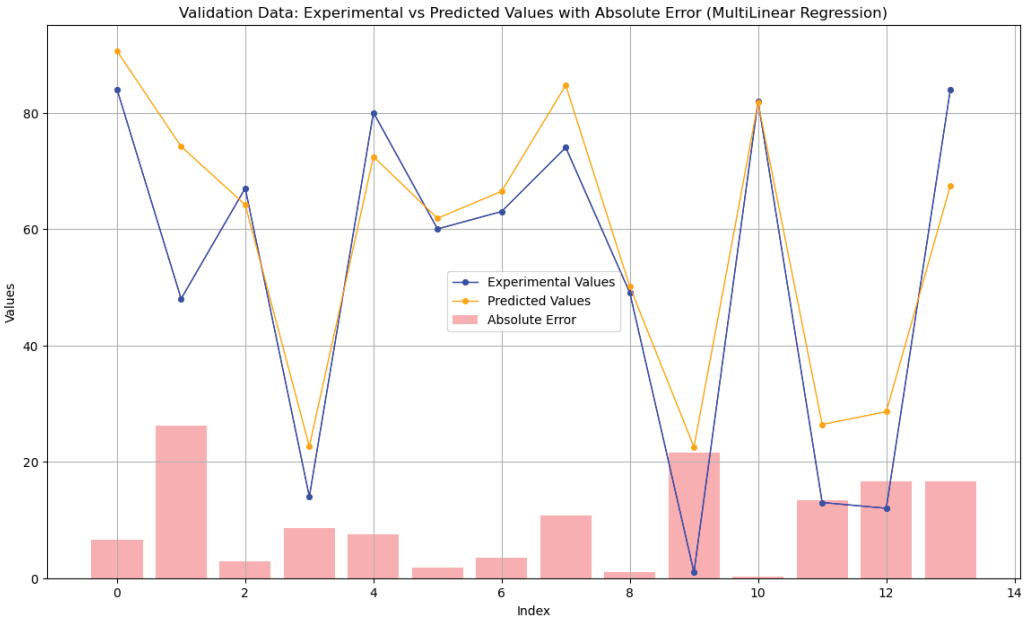


Figure 8. a) The absolute error values of the models for validation data: Multi Linear Regression.

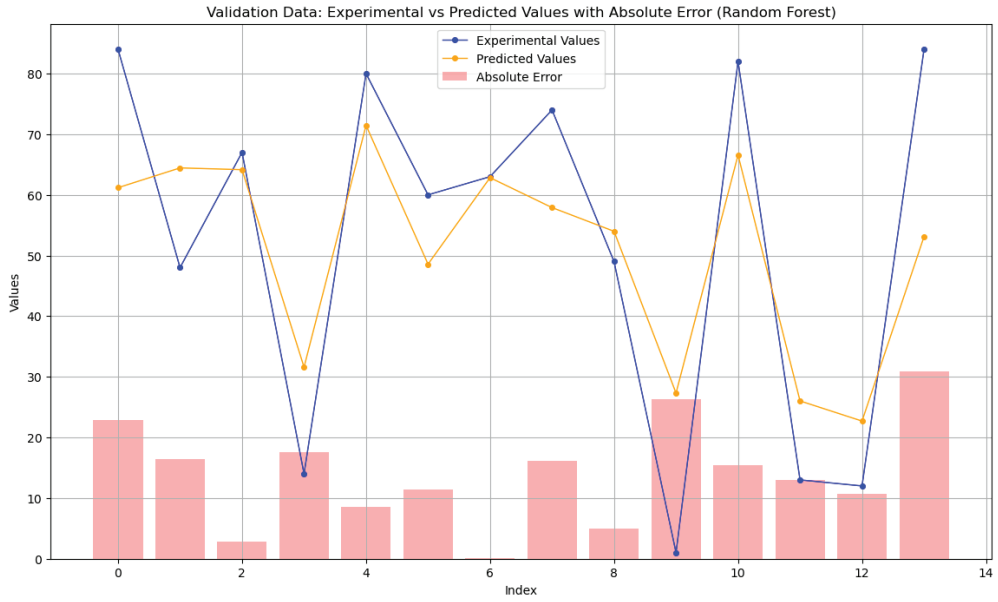


Figure 8. b) The absolute error values of the models for validation data: Random Forest.

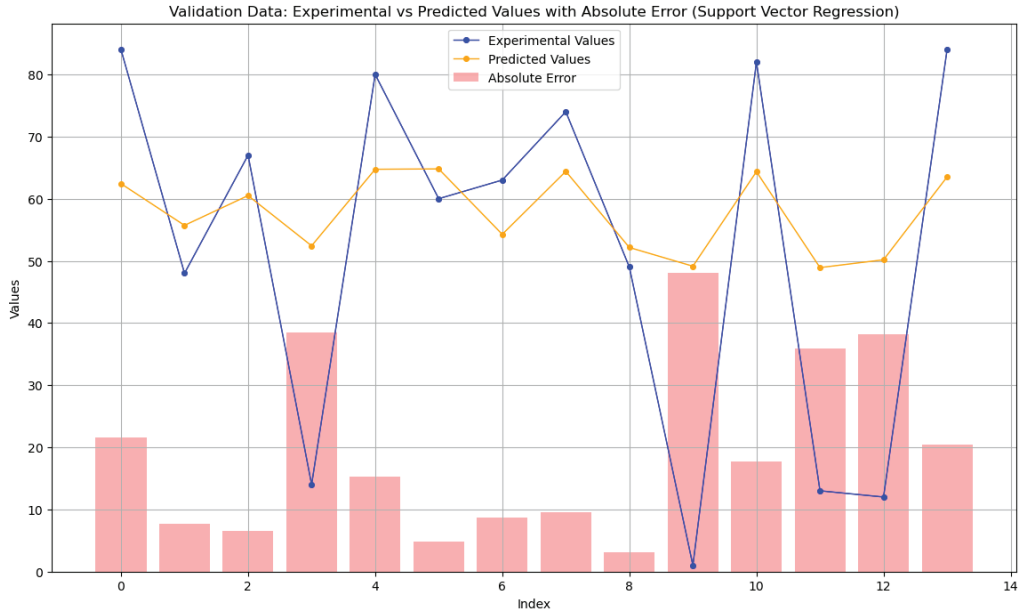


Figure 8. c) The absolute error values of the models for validation data: Support Vector Regression.

contrary, has very low accuracy, high and frequent error bars with flat predictions that don't reflect the variability of the data.

Figure 8[a-c] show the performance of the three regression models (Multi Linear Regression, Random Forest Regression, Support Vector Regression (SVR)) when compared to the experimental values of the validation data set. The plot of each experimental and theoretical graph in each case is displayed as a line graph, and in each case the absolute error for each data point is displayed using a bar graph. In terms of prediction accuracy, the Multi Linear Regression model (a) has a moderate level but with unevenness in the magnitude of the errors. Random Forest model (b) shows a fairly good agreement between the predicted and experimental values, but there are some discrepancies. The SVR model (c), on the other hand, shows higher absolute errors in some of the data points, suggesting that it is less accurate than the other models.

The comparisons of experimental values, projected values and absolute errors for three machine learning models: Multiple Linear Regression (MLR), Random Forest (RF) and

Support Vector Regression (SVR) are presented in Figure 7 and 8 by training and validation datasets. The experimental and predicted values are plotted in each plot, while the pink bars indicate absolute errors in each instance. The errors and spread of SVR (c) are larger, indicating a less accurate fit compared to Multiple Linear Regression (a) and Random Forest (b) which almost match the results of the experimental data in the training data results (a-c). Absolute errors are slightly higher for Random Forest as well. Both methods of Random Forest and Multiple Linear Regression fit the experimental values fairly well in the validation data plots, although the absolute errors have increased when compared with the training data, which means that the model is not as accurate as before. The poor efficacy of SVR is also confirmed with the inconsistencies in prediction and more errors on this dataset. In terms of predictive accuracy and consistency, Random Forest seems to do the best overall, followed by Multiple Linear Regression, with SVR trailing behind.

CONCLUSION

1. This study highlights the critical impact of road traffic accidents emphasizing the urgent need for predictive models to analyse the accident data and improve road safety.
2. The Dataset used is collected road inventory and spot speed surveys with road details measured using a Total Station and vehicle speeds recorded via radar guns at black spots and using accident data, which includes FIRs from various districts, later data were processed, and Accidents were categorized into Minor, Major, and Fatal, with the Equivalent Accident Number calculated.
3. Based on the statistical analyse the attributes are selected for prediction. Then the data set is divided into 80% training and 20% validation the regression models are implemented.
4. By comparing regression models viz., Backward stepwise Multi Linear Regression, Random Forest, and Support Vector Regression, the study revealed Multi Linear Regression as the most suitable model for predicting the accident number for Karnataka State.
5. Random Forest and SVR were tested and Backward stepwise linear regression showed better results with an R^2 score of 0.867 while the other two models were R^2 = 0.807 and R^2 = -0.027 respectively. The Multiple Linear Regression model showed improvement in accuracy and stability with lower values of mean absolute error (MAE) and root mean squared error (RMSE).
6. These findings support the model's ability to provide focused enhancements to road safety thereby reducing road accidents.
7. More data can be added to the model for further refinement to improve accuracy and efficiency of developed model.
8. Overall, the Machine Learning model is seen as being preferred over conventional models in accident analysis because it can learn from data automatically, deal with large amounts of data and over time improve the accuracy of those predictions.

Limitations and Future Work: Although the results of the study indicate that the regression models are effective models for predicting the accidents, the study is bound by the size and diversity of the data used. The data collected was area specific (Karnataka state) and these results may not be applicable to other areas with varying traffic characteristics and road surfaces. Additionally, the Support Vector Regression model was poorer, perhaps due to the fact that the hyperparameters were not optimized and this could be further investigated. Future work will use a wider and larger range of datasets, investigate more complex models such as deep learning, and hyperparameter optimization.Ethical considerations: Great care was taken in addressing ethical considerations – all accident data used were obtained from official sources such as FIRs, road inventories, and surveys. Personally identifiable information (PII) was neither collected nor used, and the data were anonymized prior to analysis to maintain privacy and comply with ethical research standards.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The authors declared that they do not have any conflict of interest with other entities or researchers.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCE

[1] World Health Organization. Global status report on road safety 2023: summary. Geneva: World Health Organization; 2023.

[2] World Health Organization. Global status report on road safety 2018. Geneva: World Health Organization; 2019.

[3] Wang X, Kim SH. Prediction and factor identification for crash severity: Comparison of discrete choice and tree-based models. Transp Res Rec 2019;2673:640–653. [CrossRef]
[4] Lin MR, Kraus JF. Methodological issues in motorcycle injury epidemiology. Accid Anal Prev 2008;40:1653–1660. [CrossRef]
[5] Laiou A, Papadimitriou E, Yannis G, Milotti A. Road safety data and information availability and priorities in South-East European regions. Transp Res Procedia 2017;25:3703–3714. [CrossRef]
[6] Sukor NA, Sadullah AM. Addressing the road safety results impasse through an outcome-based approach in the state of Penang, Malaysia. Eng Herit J 2017;1:21–24. [CrossRef]
[7] Li Z, Liu P, Wang W, Xu C. Using support vector machine models for crash injury severity analysis. Accid Anal Prev 2012;45:478–486. [CrossRef]
[8] Ibrahim H, Far BH. Data-oriented intelligent transportation systems. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014); 2014. p. 322–329. [CrossRef]
[9] Mohamed EA. Predicting causes of traffic road accidents using multi-class support vector machines. J Commun Comput 2014;11:441–447.
[10] Effati M, Thill JC, Shabani S. Geospatial and machine learning techniques for wicked social science problems: Analysis of crash severity on a regional highway corridor. J Geogr Syst 2015;17:107–135. [CrossRef]
[11] Perone CS. Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil. arXiv [Preprint]. 2015. arXiv:1502.00245.
[12] Sharma B, Katiyar VK, Kumar K. Traffic accident prediction model using support vector machines with Gaussian kernel. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving: SocProS 2015. Vol. 2; 2016. p. 1–10. [CrossRef]
[13] Gu X, Li T, Wang Y, Zhang L, Wang Y, Yao J. Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization. J Algorithms Comput Technol 2018;12:20–29. [CrossRef]
[14] Al-Radaideh QA, Daoud EJ. Data mining methods for traffic accident severity prediction. Int J Neural Netw Adv Appl 2018;5:1–12.
[15] Farhat Z, Karouni A, Daya B, Chauvet P. Comparative study between decision trees and neural networks to predict fatal road accidents in Lebanon. In: 5th International Conference on Computer Science, Information Technology; 2019. p. 1–14. [CrossRef]
[16] Karthik D, Vijayarekha K. Identifying efficient road safety prediction model using data mining classifiers: Recent advancements. 2020.
[17] Cui H, Dong J, Zhu M, Li X, Wang Q. Identifying accident black spots based on the accident spacing distribution. J Traffic Transp Eng (Engl Ed) 2022;9:1017–1026. [CrossRef]
[18] Karamanlis I, Kokkalis A, Profillidis V, Botzoris G, Galanis A. Identifying road accident black spots using classical and modern approaches. WSEAS Trans Syst 2023;22:556–565. [CrossRef]
[19] Mbarek A, Jiber M, Yahyaouy A, Sabri A. Black spots identification on rural roads based on extreme learning machine. Int J Electr Comput Eng 2023;13:3149–3160. [CrossRef]
[20] Chuerubım ML, Valejo A, Bezerra BS, Da Sılva I. Artificial neural networks restriction for road accidents severity classification in unbalanced database. Sigma J Eng Nat Sci 2019;37:927–940.
[21] Yıldırım V, Kantar YM. Spatial analysis of the road traffic accident statistics in the provinces of Turkey. Sigma J Eng Nat Sci 2020;38:1667–1680.
[22] Manoj P, Manjunath KC, Kotagi PB. ANFIS-based traffic accident prediction model for Karnataka State. Int J Comput Exp Sci Eng 2025;11. [CrossRef]
[23] Manoj P, Punith KB, Manjunath KC. Enhancing road safety through blackspots mitigative measures: A review. Sigma J Eng Nat Sci 2024;42:1670–1682. [CrossRef]
[24] Özer ÖÖ, Subasi N. Unraveling data complexity in the metaverse for anomaly detection with Python on NYC taxi. In: Practical applications of machine learning and AI: Medicine, environmental science, transportation, and education. IGI Global Scientific Publishing; 2025. p. 261–276. [CrossRef]
[25] Manoj P, Manjunath KC, Kotagi PB. Machine learning models for proactive road safety: Evaluating regression and ensemble techniques based on road geometry. Eng Technol Appl Sci Res 2025;15:25951–25958. [CrossRef]