Machine Learning Models for Proactive Road Safety: Evaluating Regression and Ensemble Techniques Based on Road Geometry

P. Manoj
Department of Civil Engineering, National Institute of Engineering, Mysore, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India | Vidyavardhaka College of Engineering, Mysore, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India
manoj.p@vvce.ac.in (corresponding author)

K. C. Manjunath
Department of Civil Engineering, National Institute of Engineering, Mysore, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India
kcmnie@gmail.com

Punith B. Kotagi
Department of Civil Engineering, National Institute of Engineering, Mysore, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India
punithbkotagi@nie.ac.in

Received: 18 April 2025 | Revised: 21 May 2025, 4 June 2025, and 16 June 2025 | Accepted: 18 June 2025

 DOI: https://doi.org/10.48084/etasr.11569

ABSTRACT

The complex interplay of factors contributing to road accidents necessitates the use of advanced predictive techniques capable of identifying the high-risk zones before incidents occur. The present research addresses this need by using a dataset with 16 road characteristics, such as curve radius, entry/exit speeds, and sight distance, and developing several models for this task, namely traditional linear models, including Simple Linear Regression (SLR), Ridge, Lasso, ElasticNet, ensemble techniques, including Random Forest Regressor (RFR), gradient boosting, Support Vector Regressor (SVR), and Extreme Gradient Boosting (XGBoost), and advanced gradient boosting frameworks, like LightGBM, and CatBoost. Among these, SLR achieved the best performance, with a Root Mean Square Error (RMSE): 6.92, and R^2: 0.94 on the test set, while XGBoost ranked highest among the ensemble methods (RMSE: 14.55, R^2: 0.75). The feature importance analysis revealed that the superelevation e (%), entry speed V(entry) (km/h), and mid-section speed V(mid) (km/h) were the most significant predictors across the models. This analysis offers valuable insights for the transportation authorities to predict accident-prone areas and implement targeted safety measures to reduce the road accidents.

Keywords-accident forecasting; machine learning; road safety; regression models; ensemble methods; feature importance

I. INTRODUCTION

Road traffic incidents are a significant global public challenge [1]. A proactive approach to road safety involves predicting accident-prone areas using road geometric characteristics, which can serve as vital indicators of risk [2]. Although there have been attempts to use traditional statistical methods for predicting accidents, they often fall short in capturing the nonlinear and complex interdependencies inherent in road safety data. However, the advances in machine learning present greater potential for reliable road safety prediction compared to traditional models [2], owing to their capacity to uncover the complex patterns within multidimensional data. This capability enables authorities to take proactive measures in high-risk areas [3]. Nevertheless, the accuracy of such models remains highly dependent on the quality of the input data and the appropriateness of the selected modeling approach [4].

Additionally, the integration of deep learning with modern technologies has enabled real-time risk evaluation [5, 6], while further advancements, such as trajectory analysis, hyperparameter tuning, and the integration of contextual variables, like traffic volume and weather, have significantly

improved the model accuracy and reliability [7–12]. Hybrid modeling approaches, including the fusion of grey systems with neural networks, have also shown promising results [13]. Moreover, ensemble methods, known for their robustness in handling complex, imbalanced data, contribute to enhanced predictive performance [14]. The integration of ML frameworks with the Internet of Things (IoT) infrastructure has opened new frontiers in accident detection and traffic management [15-17], while climate-aware predictive models offer an additional layer to improve reliability [18, 19].

However, existing studies seldom compare multiple modeling techniques or systematically assess the importance of the road geometry features across models [20-24]. This study addresses these gaps by developing and comparing a range of machine learning models for predicting traffic accidents based on the road geometry. Through a systematic analysis of these regression and ensemble models, the research identifies the most accurate and robust approaches, as well as the geometric parameters that most strongly contribute to the accident risk. The findings provide practical insights for i) integrating machine learning-based prediction into transportation infrastructure planning and policy development, and ii) by pinpointing the high-risk segments of road networks with greater precision, this research offers a data-driven foundation for implementing targeted safety interventions, ultimately aiming to reduce the frequency and severity of traffic accidents

II. METHODOLOGY

Figure 1 presents the proposed methodology for predictive accident forecasting to enhance the road safety. The approach follows a structured six-step workflow. The methodology begins with data collection of the road geometric features, followed by feature engineering to preprocess the data for modeling. Multiple predictive models were developed across three major categories: traditional linear models, ensemble learning methods, and advanced boosting algorithms. Feature importance analysis was conducted to identify the most significant predictors of accident risk.

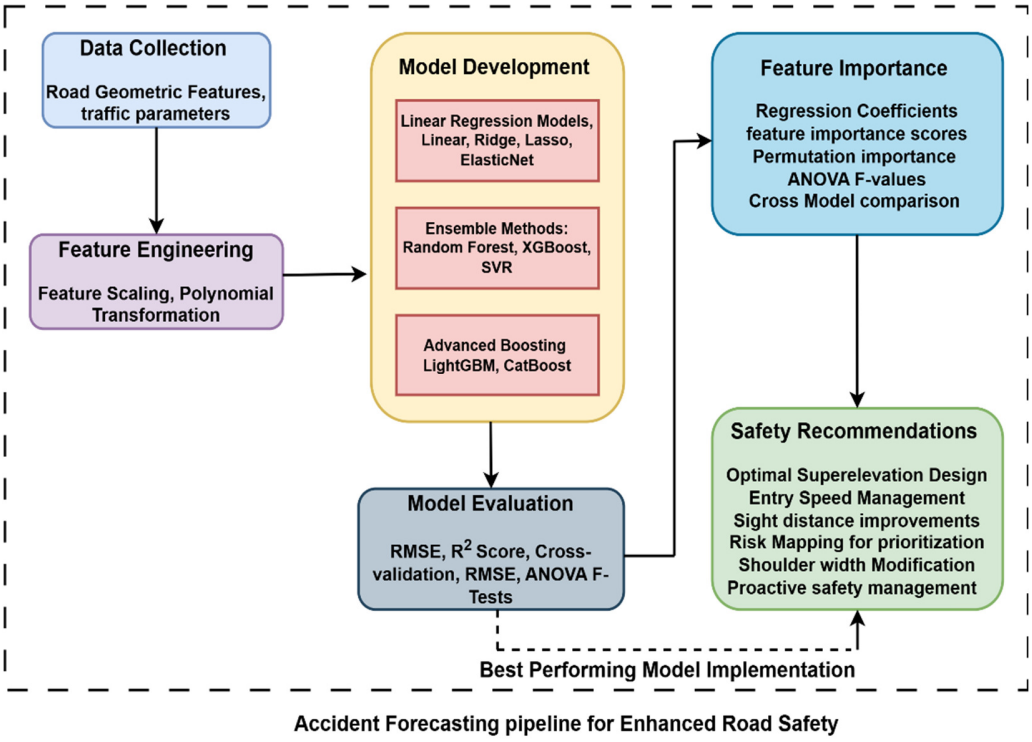

Fig. 1. Proposed methodology.

A. Data Description

The study utilized a comprehensive dataset of road segments with geometric characteristics and corresponding accident records, including 16 features related to road geometry and traffic conditions, as shown in Table I. The dependent variable was the Equivalent Accident Number (EAN). A sample of the dataset is presented in Table II. A total of 67 combinations were compiled from field surveys and were made publicly available via an open-access repository on GitHub [25].

B. Data Preprocessing

Data preprocessing involved the following steps:

- The missing values were addressed using listwise deletion.
- The numerical features were normalized via StandardScaler.
- Feature engineering included polynomial transformations to capture the interaction effects.

- An 80:20 train–test split was performed to evaluate the model generalization.

Additionally, multicollinearity was assessed using variance inflation factors, and appropriate transformations were applied where necessary.

TABLE I. LIST OF INDEPENDENT VARIABLES

No.	Variable	No.	Variable
1	Radius [R] (m)	9	Total Width [TW] (m)
2	Speed @ Entry [V(entry)] (km/h)	10	Carriage Way Width [CW] (m)
3	Speed @ Exit [V(exit)] (km/h)	11	Shoulder Width (Left) [SW(L)] (m)
4	Length of Transition Curve [Ls] (m)	12	Shoulder Width (Right) [SW(R)] (m)
5	Tangent Length [TL] (m)	13	Long Chord (LC) (m)
6	Superelevation [e] (%)	14	Appex Distance (Es) (m)
7	Sight Distance [SD] (m)	15	Mid Speed [V(mid)] (km/h)
8	Deflection Angle [D Angle] (m)	16	Passenger Car Unit [PCU]

TABLE II. SAMPLE DATA OF INDEPENDENT AND DEPENDENT VARIABLES

R (m)	**V(entry) (km/h)**	**V(exit) (km/h)**	**Ls (m)**	**TL (m)**	**e (%)**
92.3	68	70	27.09	38.12	5.67
83.09	70	70	30.09	26.94	5.95
SD (m)	**D Angle (m)**	**TW (m)**	**CW (m)**	**SW(L) (m)**	**SW(R) (m)**
73.87	50.21	19.83	12.72	3.55	3.55
42.9	38.52	13.35	7.5	4	1.85
LC (m)	**Es (m)**	**V(mid) (km/h)**	**PCU**	**EAN**	
80.88	9.63	50	4000	78	
55.85	4.92	49	4179.75	90	

C. Model Development

1) Linear Regression Models

a) Simple Linear Regression

SLR is one of the earliest statistical modeling techniques, used to predict a dependent variable based on a single independent variable. In the context of SLR, road safety can be modeled by estimating the accident frequency using traffic volume as the predictor. SLR requires a strict linear relationship between the variables and assumes uncorrelated error terms, making it less suitable for complex, interdependent datasets. The SLR model is defined as:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

where y is the dependent variable, x is the independent variable, β_0 is the intercept, β_1 is the slope, and ε is the error term, assumed to follow a normal distribution with mean 0 and variance σ^2. The objective is to minimize the sum of squared residuals:

$$Min. \sum_{i=1}^{n} = (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (2)$$

The coefficients are estimated using:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \tilde{x})(y_i - \tilde{y})}{\sum_{i=1}^{n}(x_i - \tilde{x})^2}$$

$$\tilde{\beta}_0 = \tilde{y} - \hat{\beta}_1 \tilde{x} \quad (3)$$

where $\tilde{x}$ and $\tilde{y}$ are the means of x and y.

b) Ridge Regression

Ridge regression extends SLR to address multicollinearity, common in accident forecasting datasets, where predictors like traffic volume, road size, and speed are highly correlated. It also mitigates the coefficient inflation by introducing an L2 regularization term, which penalizes large coefficients while retaining all predictors. This regularization reduces the model variance and improves generalization. The ridge regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \quad (4)$$

The objective is to minimize:

$$Min. \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (5)$$

where y_i is the observed accident outcome for the i[th] observation, x_{ij} is the value of the j[th] predictor and i[th] observation, β_j is the coefficient for the j[th] predictor, λ is the regularization parameter ($\lambda \geq 0$) controlling the penalty strength, and $\sum_{j=1}^{p} \beta_j^2$ is the L2 penalty term. In matrix form, the solution is:

$$\tilde{\beta} = (X^T X + \lambda I)^{-1} X^T y \quad (6)$$

where X is the design matrix, Y is the response vector, and I is the identity matrix.

c) Lasso Regression

Lasso regression applies L1 regularization to reduce overfitting and select key features by shrinking the irrelevant coefficients to zero. In accident prediction, it highlights the impactful variables (e.g., speed, traffic) while discarding the relevant ones, aiding interpretability and strategic intervention. However, with highly correlated predictors, Lasso's feature selection may become unstable, requiring careful tuning. The model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \quad (7)$$

The objective is to minimize:

$$Min. \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (8)$$

where $|\beta_j|$ is the absolute value of the j[th] coefficient, forming the L1 penalty, and λ is the regularization parameter controlling sparsity.

d) ElasticNet

When dealing with datasets characterized by multicollinearity and a large number of predictors, ElasticNet offers a robust solution for accident forecasting. By combining the strengths of both Ridge and Lasso regression, ElasticNet applies both L1 and L2 regularization, striking a balance between feature selection (L1) and coefficient shrinkage (L2). The ElasticNet objective function is:

$$Min. \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} |\beta_j| +$$
$$+ (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \quad (9)$$

where $a \in [0,1]$ is the mixing parameter, with $a = 1$ corresponding to Lasso and $a = 0$ to Ridge regression.

2) Ensemple Methods

a) Random Forest Regressor

RFR is an ensemble learning technique that employs multiple decision trees to predict continuous outcomes by modeling complex nonlinear relationships. Also, it is particularly effective with noisy datasets and offers valuable feature importance rankings that help identify the most influential variables. The RFR prediction is computed as the average of T decision trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (10)$$

where $\hat{y}$ is the predicted outcome, and $h_t(x)$ is the prediction from the t[th] decision tree for input x. Each tree minimizes the Mean Squared Error (MSE) over its training subset:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_t(x_i))^2 \quad (11)$$

where y_i is the observed outcome, and n is the number of samples in the subset.

b) Gradient Boosting Regressor

Gradient boosting regressor builds an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessors. Its flexibility in optimizing different loss functions and handling heterogeneous data makes it powerful, though it requires careful regularization to prevent overfitting. The prediction is given by:

$$\hat{y} = F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \quad (12)$$

where $F_M(x)$ is the final prediction after M iterations, $h_m(x)$ is the weak learner (decision tree) at iteration m, and γ_m is the learning rate. The loss function minimized is:

$$L = \sum_{i=1}^{n}(y_i - F(x_i))^2 \quad (13)$$

At each iteration m, the model is updated as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (14)$$

where $h_m(x)$ is fit to the negative gradient of the loss:

$$h_m(x) \approx \frac{\partial L}{\partial F_{m-1}(x_i)} \quad (15)$$

c) Support Vector Regressor

SVR enables accurate accident forecasting, even with high-dimensional or non-linear data, by learning a function that approximates the target values within a specified margin of error. In road safety applications, SVR can leverage features, such as speed and traffic volume to model the non-linear frequency of accidents, particularly using Radial Basis Function (RBF) kernels. SVR seeks to minimize:

$$\begin{aligned} &min\,\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &y_i - (w^T\phi(x_i) + b) \leq \in + \xi_i \\ &(w^T\phi(x_i) + b) \leq \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0 \end{aligned} \quad (16)$$

where w is the weight vector, $\phi(x_i)$ is the feature mapping (e.g., via RBF kernel $K(x_i, x_j)$), b is the bias term, ε is the margin of tolerance, C is the penalty parameter for errors, ξ_i, ξ_i^* are slack variables that allow deviations outside the ε-tube. The SVR prediction is given by:

$$\hat{y} = \sum_{i \in SV}(\alpha_i - \alpha_i^*)K(x_i, x) + b \quad (17)$$

d) Extreme Gradient Boosting

XGBoost is highly effective for forecasting due to its speed, scalability, and ability to model complex, non-linear relationships. The algorithm is particularly advantageous for feature selection, as it demonstrates resilience to missing data, limited samples, and heterogeneous inputs. It also supports parallel computation, enabling efficient training on large and multi-source datasets. The prediction function is defined as:

$$\hat{y} = \sum_{m=1}^{M} f_m(x) \quad (18)$$

where $f_m(x)$ is the prediction of the m[th] tree. The objective function combines a loss term and a regularization term:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \quad (19)$$

where $l(y_i, \hat{y}_i)$ is the loss function, $\Omega(f_m)$ is the regularization term, which typically depends on the number of leaves T and the associated leaf weights w.

3) Advanced Boosting Algorithms

a) LightGBM Regressor

The LightGBM regressor is particularly useful in road safety applications, as it offers optimized speed and scalability with its high-performance gradient-boosting framework. Using features, like traffic density, road type, and speed conditions, it can predict the frequency of accidents with high accuracy due to its histogram-based learning and leaf-wise tree growth methods. LightGBM is ideal as it requires no extensive data preprocessing and offers high accuracy and parameter tuning.

b) CatBoost Regressor

CatBoost regressor is another gradient boosting algorithm, specifically designed to handle categorical features natively. It performs well in predicting the severity or frequency of accidents using several features because it automatically encodes categorical variables and constructs patterns without requiring significant data preparation. It mitigates overfitting by employing ordered boosting, symmetric tree structures, and robust regularization techniques that enhance generalization.

D. Model Evaluation

RMSE was used to assess the prediction accuracy by measuring the average magnitude of errors between the predicted and observed accident outcomes (e.g., frequency or severity). RMSE is expressed in the same units as the target variable and is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (20)$$

where y_i is the observed outcome, $\hat{y}_i$ is the predicted accident outcome, and n is the number of observations. In road safety forecasting, a lower RMSE indicates more accurate predictions, facilitating a better identification of the high-risk conditions.

The coefficient of determination (R^2) was used to measure the proportion of variance in the accident outcomes explained by the model. R^2 is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (21)$$

where $\sum_{i=1}^{n} = (y_i - \hat{y}_i)^2$ is the sum of squared residuals and $\sum_{i=1}^{n} = (y_i - \bar{y})^2$ is the total sum of squares.

Cross-validation RMSE (CV-RMSE) was employed to assess the model generalization and robustness across different subsets of the dataset. In k-fold cross-validation, the dataset is partitioned into k subsets (k=5); the model is trained on k-1 folds and validated on the remaining fold, which is repeated k times. The mean RMSE across folds is computed as:

$$\text{CV} - \text{RMSE} = \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{1}{n_j}\sum_{i \in fold_j}(y_i - \hat{y}_i)^2} \quad (22)$$

where n_j is the number of observations in fold j. CV-RMSE provides a more reliable estimate of the performance on unseen data, essential for accident prediction under varying traffic or environmental conditions.

Additionally, Analysis of Variance (ANOVA) F-tests were conducted to evaluate the statistical significance of the model features. The test compares the variance explained by each predictor to the residual variance, determining whether a feature (e.g., traffic density, speed, or road type) contributes significantly to the model. The F-statistic is calculated as:

$$F = \frac{MSR}{MSE}$$

$$MSR = \frac{SSR}{df_R}$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i^{\,full} - \hat{y}_i^{\,reduced})^2 \quad (23)$$

where MSR is the mean square regression, SSR is the sum of squares attributed to the feature, and df_R corresponds to the degrees of freedom. The error variance MSE is:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p} \quad (24)$$

where n is the number of observations, and p is the number of parameters in the full model.

III. RESULTS AND DISCUSSION

Table III presents a comprehensive comparison of the performance metrics across all evaluated models in predicting the number of traffic accidents. Figures 2-4 illustrate the performance of each model in terms of RMSE, R^2, and CV-RMSE, respectively. The SLR model outperformed all other models, achieving the lowest RMSE of 6.92 and a high R^2 of 0.94 on the test dataset, surpassing even more complex algorithms. This unexpected result suggests that the relationship between the road geometric features and accident risk in the dataset may be predominantly linear. Lasso regression followed closely, with an RMSE of 7.42 and R^2 of 0.93. Notably, Lasso outperformed both Ridge regression (RMSE: 14.58, R^2: 0.75) and ElasticNet (RMSE: 13.48, R^2: 0.78), indicating that in this particular prediction task, sparsity contributes more to performance than coefficient shrinkage.

Despite the ensemble methods' theoretical advantage in capturing non-linear relationships, they did not achieve superior test accuracy. XGBoost performed best among them (RMSE: 14.55, R^2: 0.75), while RFR (RMSE: 19.07, R^2: 0.57) and LightGBM (RMSE: 21.83, R^2: 0.44) performed significantly worse. These results suggest a potential overfitting in the more flexible models, likely due to their higher parameter complexity, relative to the dataset size. SVR demonstrated the weakest performance (RMSE: 27.76, R^2: 0.09), indicating that the chosen kernel or hyperparameter settings may not be well-suited for this application, and that tuning SVR may be impractical for this forecasting task. Interestingly, although SLR achieved the highest test set metrics, it recorded the worst CV-RMSE score of 16.59 among the linear models, suggesting a lack of stability across the data partitions. In contrast, Lasso regression showed the best balance between test accuracy and generalization, with a CV-RMSE of 14.69, highlighting its robustness and consistency across folds.

In summary, Lasso regression emerges as the most reliable model, offering a strong combination of accuracy, interpretability, generalization, and stability. These characteristics make it particularly suitable for accident forecasting, where dependable and actionable predictions are essential.

TABLE III. MODEL PERFORMANCE COMPARISON

Model	RMSE	R^2	CV-RMSE
SLR	6.92	0.94	16.59
Ridge regression	14.58	0.75	15.02
Lasso regression	7.42	0.93	14.69
ElasticNet	13.48	0.78	15.23
RFR	19.07	0.57	18.20
Gradient boosting	16.10	0.69	20.08
SVR	27.76	0.09	22.19
XGBoost	14.55	0.75	21.37
LightGBM	21.83	0.44	21.55
CatBoost	16.46	0.68	17.5

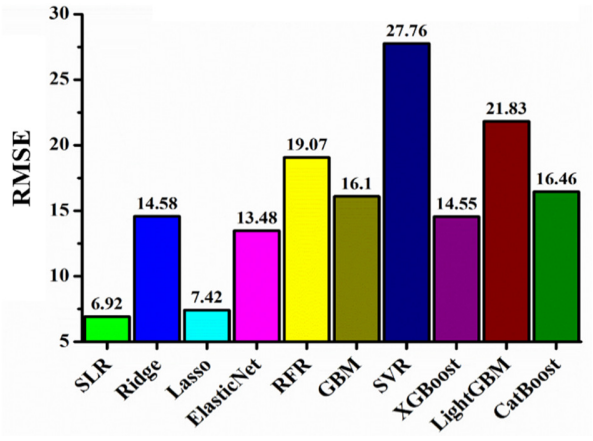

Fig. 2. RMSE obtained from different models.

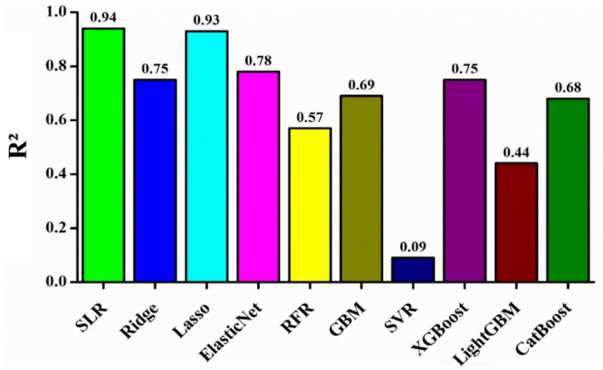

Fig. 3. R^2 obtained from different models.

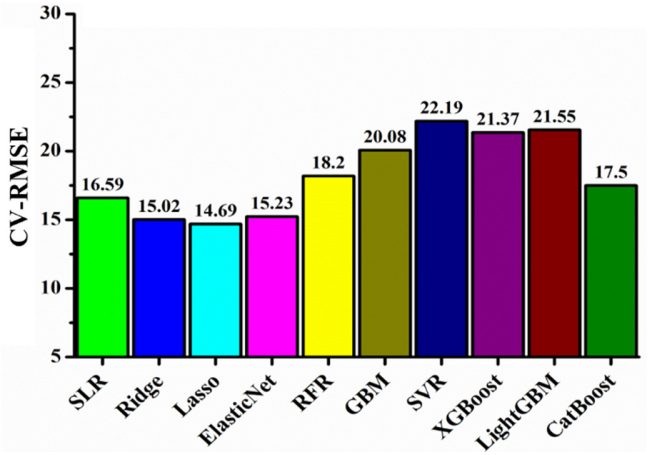

Fig. 4. CV-RMSE obtained from different models.

A. *Feature Importance*

The feature importance analysis revealed consistent patterns across the evaluated modeling approaches. Figure 5 presents the feature importance rankings derived from the polynomial linear regression model. In this model, e emerged as the most influential feature, with a coefficient of 10.85, followed by SW(L) (–4.34), V(entry) (3.44), and SW(R) (3.23).

Figure 6 depicts the ANOVA F-test results for each of the features used, and Table IV lists the corresponding p-values for each feature. Notably, V(entry) ($F = 68.24$, $p < 0.001$), V(exit) ($F = 63.92$, $p < 0.001$), and e ($F = 59.67$, $p < 0.001$) demonstrated the highest F-values, reinforcing their relevance in accident prediction.

Figures 7 and 8 illustrate the corresponding feature importances from the ElasticNet model and the ensemble models (XGBoost, LightGBM, and CatBoost). Feature rankings from ElasticNet and ensemble models generally aligned with those of the linear models, with minor variations. In ElasticNet, e had the highest importance score (7.8), followed by V(entry) (4.3). The XGBoost model also ranked e highest (0.42), followed by V(entry) (0.13) and SW(R) (0.115). The CatBoost model similarly identified e (14.8) and V(entry) (13.9) as the most influential predictors.

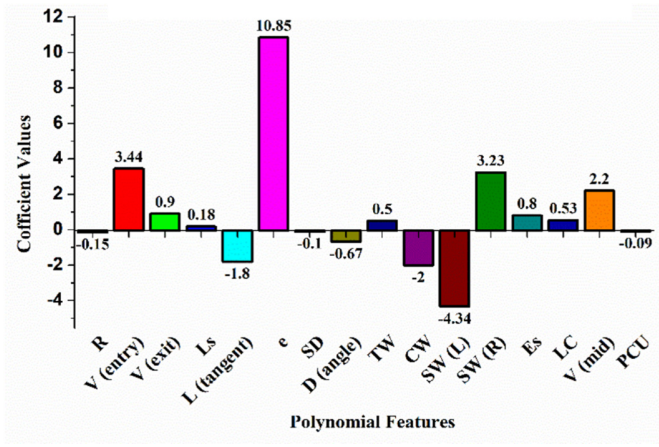

Fig. 5. Polynomial regression coefficients.

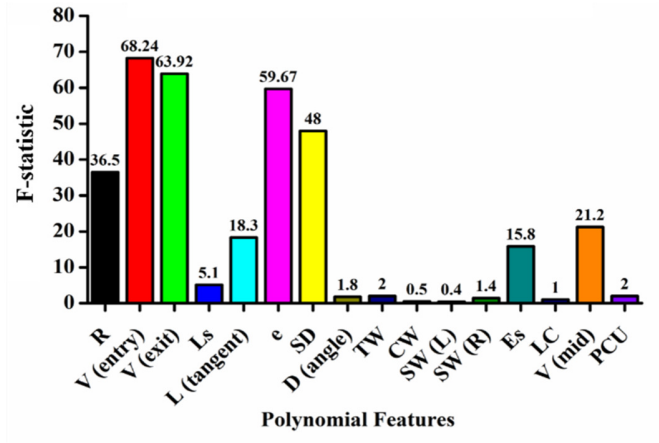

Fig. 6. ANOVA F-test for feature significance.

TABLE IV. P-VALUES FOR FEATURE SIGNIFICANCE

Variable	p-value	Variable	p-value
R (m)	1.89×10^{-7}	TW (m)	0.1635
V(entry) (km/h)	6.59×10^{-11}	CW (m)	0.4828
V(exit) (km/h)	1.69×10^{-10}	SW(L) (m)	0.53
Ls (m)	0.0283	SW(R) (m)	0.2423
TL (m)	8.50×10^{-5}	LC (m)	0.00023
e (%)	4.44×10^{-10}	Es (m)	0.3221
SD (m)	7.77×10^{-9}	V(mid) (km/h)	2.87×10^{-5}
D Angle (m)	0.1858	PCU	0.1635

B. *Discussion*

Across all models, the following features consistently ranked among the most impactful: e, V(entry), V(mid), R, and SD. These findings indicate that both the geometric roadway characteristics and vehicle dynamics are central to understanding accident risk. The identification of e, V(entry), and R as key predictors aligns with established engineering principles and prior research [20].

These findings suggest that roadway engineers should prioritize optimal e design, particularly on curves with high V(entry), to enhance the vehicle stability and reduce the accident risk. Additionally, improving SD is critical to ensuring adequate driver response time. Finally, shoulder width

modifications should be considered, although the observed discrepancy in the effects of left and right shoulders warrants further investigation to better understand their role in road safety.

Additionally, the residual analysis of the SLR model indicated no significant violations of the underlying regression assumptions. The residuals were approximately normally distributed and exhibited no discernible patterns when plotted against predicted values, suggesting homoscedasticity and model adequacy. The multi-model validation further revealed that although linear models achieved the highest accuracy on the current dataset, the ensemble methods demonstrated greater stability across different data partitions, as reflected in their more consistent cross-validation scores. This finding suggests that while linear models are well-suited for the current data, the ensemble approaches may offer superior generalization to unseen scenarios. Accordingly, the transportation authorities should consider integrating both linear and ensemble modeling strategies when developing predictive systems for accident forecasting, balancing immediate accuracy with long-term reliability across diverse traffic and environmental conditions.

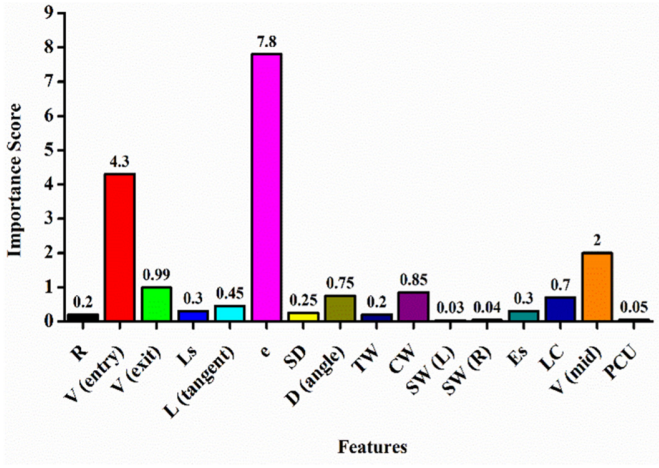

Fig. 7. ElasticNet feature importance.

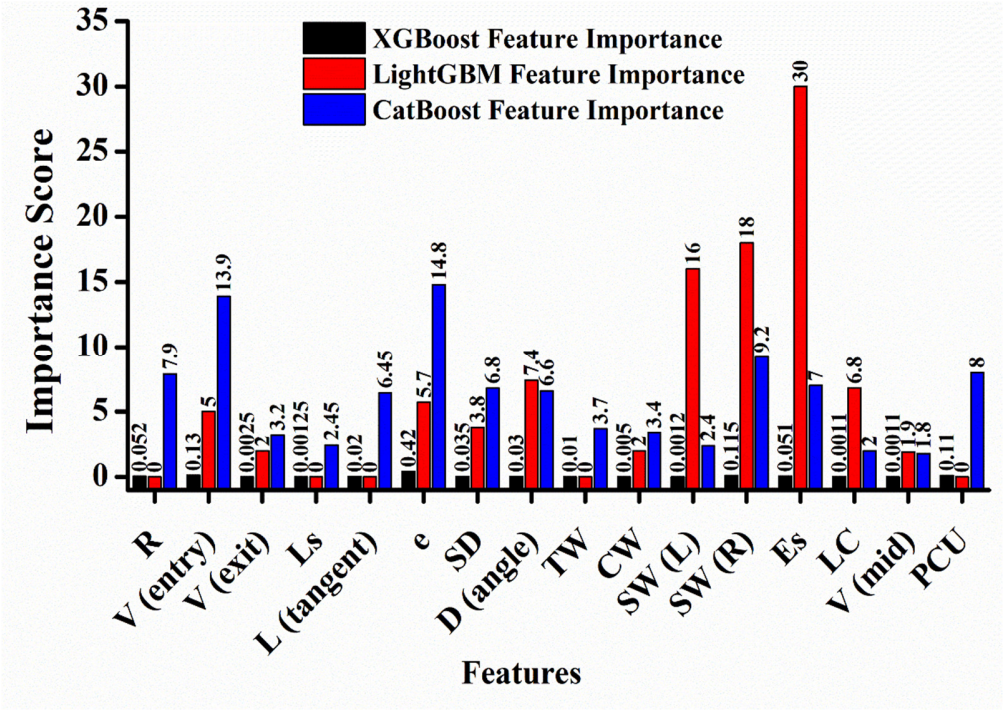

Fig. 8. Feature importance obtained from XGBoost, LightGBM, and CatBoost.

IV. CONCLUSION

This study evaluated road geometry-based traffic accident prediction models using both ensemble and traditional regression techniques. Despite the concerns about overfitting, Simple Linear Regression (SLR) yielded the best performance with a Root Mean Square Error (RMSE) of 6.92 and R^2 of 0.94. Lasso regression outperformed other regularized methods, achieving RMSE = 7.42, R^2 = 0.93, and a Cross-validation RMSE (CV-RMSE) of 14.69, indicating strong generalization. Among the ensemble models, Extreme Gradient Boosting (XGBoost) was the most effective, with an RMSE of 14.55 and R^2 of 0.75. The feature importance analysis consistently identified five key predictors across all models: superelevation e, entry speed V(entry), mid-curve speed V(mid), Radius (R), and Sight Distance (SD). These findings underscore the need for the transportation agencies and roadway engineers to take into consideration these parameters when designing new roads and to ensure safer conditions on the existing roadways.

Future research should build on this work by incorporating larger, multi-regional datasets that include detailed driver behavior and speed profiles. The use of deep learning and hybrid modeling approaches holds promise for uncovering complex patterns and enhancing interpretability. Integrating Geographic Information System (GIS)-based analytics could reveal seasonal trends and high-risk locations, supporting real-time monitoring and intervention. Finally, focusing on model explainability, field validation, transferability, and cost-effectiveness can help shift the road safety strategies from reactive to preventive.

REFERENCES

[1] World Health Organization, *Global status report on road safety 2018*. Geneva: World Health Organization, 2018.

[2] M. L. Siregar, T. Tjahjono, and N. Yusuf, "Predicting the Segment-Based Effects of Heterogeneous Traffic and Road Geometric Features on Fatal Accidents," *International Journal of Technology*, vol. 13, no. 1, Jan. 2022, Art. no. 92, https://doi.org/10.14716/ijtech.v13i1.4450.

[3] S. Alshiya, E. Anoop, K. D. Arpith, K. Sanal, and A. Happy, "Raps-Road Accident Prediction System Using Random Forest," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 4, Apr. 2023, https://doi.org/10.56726/IRJMETS35823.

[4] U. Iliyasu, M. M. Yakudu, and A. A. Abdulwasiu, "Predictive Analysis of Road Traffic Accidents in Katsina State, Nigeria Using Machine Learning Algorithms: A Study on Factors and Mitigation Strategies," *International Journal of Science for Global Sustainability*, vol. 9, no. 2, pp. 188–198, Jul. 2023, https://doi.org/10.57233/ijsgs.v9i2.475.

[5] J. Siswanto, A. S. N. Syaban, and H. Hariani, "Artificial Intelligence in Road Traffic Accident Prediction," *Jambura Journal of Informatics*, vol. 5, no. 2, pp. 77–90, Nov. 2023, https://doi.org/10.37905/jji.v5i2.22037.

[6] S. Olugbade, S. Ojo, A. L. Imoize, J. Isabona, and M. O. Alaba, "A Review of Artificial Intelligence and Machine Learning for Incident Detectors in Road Transport Systems," *Mathematical and Computational Applications*, vol. 27, no. 5, Sep. 2022, Art. no. 77, https://doi.org/10.3390/mca27050077.

[7] Y. Zhang and Y. Sung, "Traffic Accident Detection Method Using Trajectory Tracking and Influence Maps," *Mathematics*, vol. 11, no. 7, Apr. 2023, Art. no. 1743, https://doi.org/10.3390/math11071743.

[8] H. Khanum, A. Garg, and M. I. Faheem, "Accident severity prediction modeling for road safety using random forest algorithm: an analysis of Indian highways," *F1000Research*, vol. 12, Oct. 2023, Art. no. 494, https://doi.org/10.12688/f1000research.133594.2.

[9] P. Gorzelańczyk, M. Jurkovič, T. Kalina, and M. Mohanty, "Forecasting the road accident rate and the impact of the covid 19 on its frequency in the polish provinces," *Communications - Scientific letters of the University of Zilina*, vol. 24, no. 4, pp. A216–A231, Oct. 2022, https://doi.org/10.26552/com.C.2022.4.A216-A231.

[10] C. D. Wirz, J. L. Demuth, M. G. Cains, M. White, J. Radford, and A. Bostrom, "National Weather Service (NWS) Forecasters' Perceptions of AI/ML and Its Use in Operational Forecasting," *Bulletin of the American Meteorological Society*, vol. 105, no. 11, pp. E2194–E2215, Nov. 2024, https://doi.org/10.1175/BAMS-D-24-0044.1.

[11] H. Xu, Y. Zhao, D. Zhao, Y. Duan, and X. Xu, "Improvement of disastrous extreme precipitation forecasting in North China by Pangu-weather AI-driven regional WRF model," *Environmental Research Letters*, vol. 19, no. 5, May 2024, Art. no. 054051, https://doi.org/10.1088/1748-9326/ad41f0.

[12] M. Girija and V. Divya, "Deep Learning-Based Traffic Accident Prediction: An Investigative Study for Enhanced Road Safety," *EAI Endorsed Transactions on Internet of Things*, vol. 10, Feb. 2024, https://doi.org/10.4108/eetiot.5166.

[13] Q. Guo, B. Guo, Y. Wang, S. Tian, and Y. Chen, "A Combined Prediction Model Composed of the GM (1,1) Model and the BP Neural Network for Major Road Traffic Accidents in China," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–11, Apr. 2022, https://doi.org/10.1155/2022/8392759.

[14] A. P. Kumar and D. Teja Santosh, "Road Accident Severity Prediction Using Machine Learning Algorithms," *International Journal of Computer Engineering in Research Trends*, vol. 9, no. 9, pp. 175–183, Oct. 2022, https://doi.org/10.22362/ijcert/2022/v9/i9/v9i902.

[15] V. Prajwal, R. Abhay, R. D. Gowda, and G. Savitha, "AI-Driven Urban Traffic Optimization to Assess Complex Traffic Patterns for Public Traffic Control and Mobility," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 11, pp. 794–798, Nov. 2023, https://doi.org/10.22214/ijraset.2023.56630.

[16] K. Yao, "AI-driven innovations in automation and urban management," *Applied and Computational Engineering*, vol. 57, no. 1, pp. 160–165, Apr. 2024, https://doi.org/10.54254/2755-2721/57/20241327.

[17] I. De Zarzà, J. De Curtò, G. Roig, and C. T. Calafate, "LLM Multimodal Traffic Accident Forecasting," *Sensors*, vol. 23, no. 22, Nov. 2023, Art. no. 9225, https://doi.org/10.3390/s23229225.

[18] P. Gorzelanczyk, "Impact of information on the number of traffic accidents on the outcome of the forecast," *Technical Sciences*, vol. 26, Nov. 2023, https://doi.org/10.31648/ts.8945.

[19] S. A. Arhin and A. Gatiba, "Predicting Injury Severity of Angle Crashes Involving Two Vehicles at Unsignalized Intersections Using Artificial Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 9, no. 2, pp. 3871–3880, Apr. 2019, https://doi.org/10.48084/etasr.2551.

[20] M. Touahmia, "Identification of Risk Factors Influencing Road Traffic Accidents," *Engineering, Technology & Applied Science Research*, vol. 8, no. 1, pp. 2417–2421, Feb. 2018, https://doi.org/10.48084/etasr.1615.

[21] A. N. Al-Nuaimi, A. K. Jameel, and S. M. Alsadik, "Impact of Variable Speed Limits on Crash Frequency and Crash Rate: Stimulation by Flow Rate and Percentage of Heavy Vehicles," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20335–20341, Feb. 2025, https://doi.org/10.48084/etasr.9660.

[22] H. Khanum, R. Kulkarni, A. Garg, and M. Iqbal Faheem, "Enhancing Road Safety in India: A Predictive Analysis Using Machine Learning Algorithm for Accident Severity Modeling," in *Recent Topics in Highway Engineering - Up-to-Date Overview of Practical Knowledge*, vol. 10, S. Antonio Biancardo, Ed. IntechOpen, 2024.

[23] P. Manoj, K. B. Punith, K. C. Manjunath, "Enhancing road safety through blackspots mitigative measures - a review," *Sigma Journal of Engineering and Natural Sciences – Sigma Mühendislik ve Fen Bilimleri Dergisi*, pp. 1670–1682, 2024, https://doi.org/10.14744/sigma.2023.00077.

[24] P. Manoj, K. C. Manjunath, and P. B. Kotagi, "ANFIS-Based Traffic Accident Prediction Model for Karnataka State," *International Journal of Computational and Experimental Science and Engineering*, vol. 11, no. 2, May 2025, https://doi.org/10.22399/ijcesen.2137.

[25] *Accident-Data-Set-1*. (2025), P. Manoj. [Online]. Available: https://github.com/manojp2690/Accident-Data-Set-1.git.