Feature-Driven Road Accident Risk Prediction Using Gradient Boosting Regression

Yasir Ul Hassan[1], Shamshad Lakho[1], Imran Ali Memon[2], Faryal Arshad[1], Mubashir Ul Hassan[3], Muhammad Bilal Qazi[3]

[1]Computer Science Quaid-e-Awam University of Engineering, Sciences & Technology Nawabshah, Pakistan

[2]Shaheed Benazir Bhutto University, SBA Nawabshah, Pakistan

[3]Data Science Quaid-e-Awam University of Engineering, Sciences & Technology Nawabshah, Pakistan

*Correspondence: yasirulhassan.official@gmail.com, shamshad.lakho@quest.edu.pk, imran.asif.memon@sbbusba.edu.pk, faryaladil03@gmail.com, akhtarmubashir809@gmail.com, bilalmubarik32@gmail.com

Citation | Hassan. Y. U, Lakho. S, Memon. I. A, Arshad. F, Hassan. M. U, Qazi. M. B, "Feature-Driven Road Accident Risk Prediction Using Gradient Boosting Regression", IJIST, Vol. 7 Issue. 10 pp 68-82, November 2025

Received| November 02, 2025 **Revised|** November 23, 2025 **Accepted|** November 26, 2025 **Published|** November 29 2025.

Road traffic accidents remain a major global cause of fatalities and economic loss, posing significant challenges to public safety and urban mobility. Despite advancements in vehicle safety systems and road infrastructure, accurately predicting accident risk remains a complex task that requires advanced analytical techniques. This study develops a predictive framework using Gradient Boosting Regressor (GBR), an ensemble machine learning algorithm, to estimate road accident risk based on environmental and infrastructural features. The analysis incorporates multiple factors, including road type, number of lanes, road curvature, speed limits, lighting conditions, weather patterns, road signage, and historical accident records. The dataset used in this study contains more than 517,000 road condition observations with 13 predictive features and was obtained from the Kaggle Playground Series. During preprocessing, duplicate records were removed to ensure data quality, numerical variables were normalized using Min Max Scaler, and categorical variables were encoded systematically for model compatibility. Experimental results demonstrate that the proposed GBR model achieves strong predictive performance with an accuracy of approximately 88.57% in estimating accident risk levels across diverse road conditions. The findings highlight the significant influence of environmental and infrastructural factors on accident risk and demonstrate the potential of machine learning–based approaches in transportation safety analysis. The proposed framework can assist transportation authorities and policymakers in identifying high-risk road segments and implementing targeted safety interventions to reduce accident occurrences.

Keywords: Accident Risk Prediction, Artificial Intelligence, Gradient Boosting Regressor (GBR), Intelligent Transportation Systems (ITS), Machine Learning, Predictive Modeling.

Introduction:

Accidents on roads remain among the main causes of death and injury across the world and also produce substantial losses in the economy. Drivers continue to face conditions that present danger, even with developments in systems that provide safety in vehicles and in approaches to the design of roads. This occurs from interactions between the form of the road, factors in the environment, and rules for traffic. The form of roads includes curves that present difficulty, capacity of lanes that shows limitation, and limits on speed that appear high. Factors in the environment include lighting that provides inadequate visibility and conditions of weather, such as rain or fog, that decrease the extent to which drivers can see. These factors affect the ability of drivers to perceive conditions, the time that drivers require to respond, and the degree of control that drivers maintain over vehicles, and this increases the likelihood that accidents occur. The challenges appear particularly strong in regions that show low and middle levels of income. In these regions, variation in the structures that support transport and the degree to which environmental conditions present exposure further increases the likelihood that accidents occur.

Given this context, models that predict accident risk are important in the process of identifying segments of roads that contain high risk and that allow approaches to safety using data. This study develops a method using machine learning that allows interpretation and that uses parameters relating to infrastructure, environment, and time to provide estimates of risk for accidents on roads and to support planning for networks of roads that show more safety [1][2].

From the viewpoint of drivers, these factors increase the work that driving requires and increase the degree of uncertainty that occurs. Curves that show sharp features limit the distance for sight, lanes that contain limited space increase the level of interaction between vehicles, and weather conditions such as fog or rain decrease visibility and decrease the degree of grip on roads, and these factors collectively increase delays in response and increase the risk that control is lost.

Traditional statistical methods, such as ARIMA/ARIMAX time-series models and Poisson-type count models, have been widely used in predicting aggregated crash counts and examining the effects of explanatory variables. Such methods are convenient for temporal forecasting and also provide interpretable parameter estimates; however, they may face issues with nonlinearity, multicollinearity, and over-dispersion, which are commonly inherent in accident data. Moreover, case-control and aggregated approaches cannot often detect detailed spatial patterns, which are an important feature for practical applications in urban environments [3].

During the last ten years, machine learning approaches have become prominent for accident prediction because they are able to model nonlinear interactions and handle high-dimensional inputs with more flexibility than traditional methods. Neural networks and deep learning architectures have resulted in high predictive performance across multiple studies, particularly in settings with rich feature sets. Ensemble tree methods, most notably gradient-boosting frameworks such as XGBoost and LightGBM, have also reported high accuracy for traffic-related tasks such as speed-class forecasting and injury-severity classification. Of particular note is that boosting algorithms conduct stage-wise learning to balance bias and variance, thus enhancing robustness to diverse feature types and noisy data [4][5][6][1].

Two salient methodological challenges persist in the literature. First, accident data are usually highly imbalanced, with true events (crashes) occurring infrequently relative to non-events; this can bias probability estimates and raise false-positive rates when naive models are used. Methods that explicitly treat accidents as rare events or that decompose prediction into classification and conditional regression stages have been shown to reduce this bias and improve operational usefulness. Second, model interpretability is critical to practitioner

adoption: although "black-box" machine learning models may return strong predictive performance, their recommendations are difficult to trust or translate into policy without accompanying explanations. Recent work that integrates boosting models with SHAP (Shapley Additive ex Planations) or comparable interpretability techniques offers a path to achieve high performance with transparent, actionable insights for planners and safety engineers [3][7][5].

Spatial and temporal granularity is a recurring concern: many earlier studies aggregate observations into coarse units, such as city blocks, long road segments, or monthly counts, hence losing the street-level resolution that emergency services and urban planners rely on. Methods delivering fine-grained, segment-level risk estimates that can be recalculated under hypothetical changes to road features or weather are particularly valuable for real-world decision-making [3].

At the early stages of the study, it became clear that the dataset was too imbalanced to use a simple statistical model, and we have done several tests that confirmed that the minority accident cases were being overshadowed, and that's why we moved to Gradient Boosting Regression and built a workflow around it. So, the idea was pretty straightforward: 1) use finer spatial detail, 2) clean the imbalance as much as possible, and 3) then check which features actually affect the predictions. This mixed strategy maintains operating relevance while providing strong predictive performance consistent with the most recent achievements found using the boosting-based approaches, in contrast with the two purely statistical or fully black-box approaches [5][6]. The following sections cover the data and pre-processing methods, the gradient boosting model and its training process, the cross-validation and error analysis evaluation strategy, and the ability to understand results that highlight the main environmental and infrastructural predictors of estimated risk. We offer a clear and easy-to-use tool for identifying high-risk areas within this framework to enable targeted road safety interventions.

Recent studies have increasingly explored machine learning techniques for intelligent transportation and traffic prediction systems. For example, [8] developed a machine learning–based traffic prediction framework that integrates big data analytics, genetic algorithms, and deep learning techniques to analyze traffic conditions influenced by road repairs, traffic signals, and other real-world factors. The proposed system achieved a prediction accuracy of approximately 93.5%, demonstrating the growing effectiveness of data-driven models for improving traffic management and supporting future autonomous transportation technologies [8].

To provide a clearer overview of existing research developments, the relevant studies are summarized in a separate literature review section.

Literature Review:

Traditional Statistical Approaches:

Initial work on predicting accidents occurring on roads used models that follow particular approaches in analysis. The models include those that examine counts using a method that provides measures of occurrences and approaches that examine patterns across time. These methods allow analysis to provide measures of crashes in groups and to examine how factors relate to outcomes over time. The models show clear results that indicate relationships and allow work that examines patterns in time. However, the approaches show limitations when data present particular features. These features include relationships that differ from simple patterns, factors that relate to other factors in the analysis, and cases that show more variation than the model indicates. Also, methods that examine groups and that compare cases show limitations. The limitations appear in the ability to examine patterns across different locations. These patterns provide important information for work that examines safety in areas with high populations.

The rareness of the crash events was dealt with by introducing rare-events logit models to counteract probability underestimation in regular logistic regression. This method

enhances prediction accuracy when accidents are a small proportion of samples, such as motorway conditions. It has also been shown that crash prediction can benefit from including some explainable features in the predictive models; thus, such complex models as ARIMAX, in which environmental/behavioral information makes up for a more sophisticated crash rate estimate compared with simple ARIMA explorative modeling techniques [2].

The studies show that factors relating to the structure of the road, the conditions in the environment, the patterns in traffic, and the time of occurrence appear in the analysis and provide important contributions to the occurrence of accidents and the degree of harm that results. The factors relating to the structure of the road include the degree of bending in the road and the arrangement of lanes. The conditions in the environment include the weather and the level of lighting. The patterns in traffic include the rate of movement and the density of vehicles. These factors show a strong association with outcomes in the data.

Boosting-Based and Ensemble Models:

Recent work indicates that methods combining multiple approaches using a particular form of analysis provide strong results for identifying risk at the level of individual roads in areas with high density. One approach that shows this uses analysis in two separate stages. The first stage identifies locations that show the possibility of accidents occurring. The second stage measures the level of risk at a more specific level relating to individual streets. This strategy addresses the issue that data show imbalance between locations while maintaining the level of detail relating to position. The approach appears suitable for use in planning responses to situations requiring immediate action and for providing treatment focused on specific locations.

The approach that uses Light Gradient Boosting Machine shows performance that differs from other methods combining models in work examining the severity of incidents and predicting the risk of accidents. This occurs as the method provides computation efficiency and allows handling of feature groups that differ in type. Work in the current period combines explanations using a particular analysis approach with these models that use boosting. This combination allows clear identification of factors that affect outcomes. The factors include features of road structure, conditions of lighting, and patterns in weather. Methods combining approaches also show capability in this area. These methods use learning from structures that resemble trees with representations of space or network features. The methods reveal patterns in congestion and differences in speed. These patterns relate to the risk of crashes. The findings from these studies establish that models using boosting provide results with significant accuracy. The models also show relevance for use in settings that involve traffic safety in actual conditions.

Recent research further confirms the effectiveness of boosting-based models for traffic safety analysis. Hamdan and Sipos (2025) evaluated Random Forest, Gradient Boosting, and K-Nearest Neighbors models to predict segment-level crash severity on Hungarian road networks. Their Gradient Boosting model achieved high predictive performance with an accuracy of approximately 95% and an R^2 value exceeding 0.87 for fatal crash prediction, highlighting the capability of boosting algorithms to capture complex relationships in transportation safety data [9].

Deep Learning and ANN Models:

Models using approaches that examine data in multiple stages and structures similar to networks have shown more use for predicting incidents on roads. This occurs because these approaches allow analysis of relationships that differ in form and that involve factors relating to road design, conditions in the environment, and patterns of vehicle movement. A study using a structure with multiple stages of analysis found results that indicate higher performance in prediction compared to other approaches, including those using genetic patterns in programming and models that examine negative outcomes with consideration of differences

between groups. The study examined data from highways, and results show that structures with multiple stages provide advantages when analysis involves large numbers of factors.

However, the study also suggests that approaches showing improved performance in prediction often show reduced capacity for interpretation. This limitation affects use in practice when an explanation of results is required.

In different research, models using network structures that examine data for road systems in Bosnia and Serbia indicate that structures with less complexity can provide strong performance in prediction. These models use features that the research selected with care. The features include the width of roads, the characteristics of terrain, the volume of vehicle movement, and limits on speed. Results from this analysis show that performance remains strong and that patterns the models identify appear consistent when application occurs across different regions.

Recent studies also apply ensemble and boosting algorithms for accident severity prediction using real-world crash datasets. Muktar and Fono (2024) analyzed traffic accident data from Montreal using several machine learning classifiers, including XGBoost, CatBoost, Random Forest, and Gradient Boosting. Their results showed that the XGBoost model achieved the highest prediction accuracy of approximately 96%, while Gradient Boosting reached around 89% accuracy. These findings demonstrate the strong predictive capabilities of boosting-based algorithms for identifying factors associated with severe traffic accidents [10].

Similarly, Roudnitski (2024) evaluated several ensemble learning algorithms—including Random Forest, XGBoost, AdaBoost, LightGBM, and CatBoost—using crash data from New South Wales, Australia. The ensemble framework achieved a moderate predictive performance with a ROC–AUC value of approximately 0.68 and identified vehicle type and collision type as the most influential predictors of crash severity. The study demonstrates how ensemble-based machine learning approaches can provide valuable insights into accident patterns and support data-driven road safety planning [11].

The findings suggest differences in the relationship between model complexity and performance. Models with multiple stages show advantages when the data available for analysis is large. However, structures with less complexity may provide better outcomes when consideration includes both performance in prediction and capacity for interpretation, particularly in contexts where data available for analysis shows limitations.

Bayesian Models:

Approaches using data to update predictions have been used in work on road accidents to consider uncertainty, limited data, and structures with multiple levels. Models that use these updating methods and include multiple levels have been used to examine counts of accidents with different types while considering variation and relationships across levels of injury. In particular, one study provided a method that combines updating through a specific procedure, a form of analysis relating factors, and networks that show probability to produce estimates of accident risk that change with data and reflect uncertainty across large road networks [12].

A main feature of these updating methods is that they show uncertainty in a clear way and include previous knowledge, which provides value for planning over time and examining policy. However, these models often require strong assumptions about the form of data and can require considerable resources for analysis when used with large datasets that include many factors. Also, models using these methods typically focus on total counts of accidents rather than estimates of risk for specific segments. These limitations indicate that different approaches are required that maintain clear interpretation while providing analysis that can be used at a large scale and give predictions of accident risk at high resolution. This motivates the approach using methods for learning from data that are used in this study.

Time-Series Models:

Analysis following over time remains important for examining patterns in road safety, particularly for trends in crashes across time. A study in Anambra State, Nigeria, compared methods for predicting crashes and showed that including factors beyond the data improves prediction. The first method uses only previous crash data. The second method includes other factors such as the behavior of individuals, conditions in the environment, and features of the road. This approach provides more reliable predictions and more useful results.

However, methods using data over time show limitations in handling relationships between factors that differ in complex patterns. These approaches use data that combine across observations, and these limits are used for specific locations or for a detailed assessment of risk. The limitations indicate that different approaches provide advantages. Methods using learning from data allow analysis that includes time, environment, and infrastructure together. These methods also allow analysis at higher levels of detail for location.

Research Gaps:

Although current research has made significant advances in road accident prediction, several key drawbacks remain. There are some key drawbacks that have not been addressed. Conventional statistical models have a hard time modeling nonlinear interactions between the road geometry (and environmental conditions) and time, whereas machine learning and deep learning models tend to have very low levels of interpretability, thus challenging use in a practical form. Moreover, accident data tend to be very imbalanced, so that most of the models tend to make more forecasts about non-accident situations and limit their usefulness in forecasting risky situations.

Cross-context generalization is the other key limitation. A lot of models designed are specific to particular areas or road conditions and may not work properly with dissimilar traffic, weather, or infrastructure conditions. In addition, coarse spatial structures of past research restrict their ability to be used in the determination of risks in the street and smarter interventions.

The above gaps specify that the modeling framework should be precise, interpretable, and strong to data imbalance at the fine spatial resolution. To deal with these issues, the current work suggests a feature-guided Gradient Boosting Regression model that balances predictive capability with explanation by allowing accurate predictions of accident-risk action and proactive and informed road safety planning based on these predictions.

Objectives:

This research is about building a robust machine learning framework for predicting the risk of a road accident through Gradient Boosting Regression. The methodological objectives undertaken in this study to achieve the above-mentioned aim include:

Collect and analyze a large-scale dataset of road conditions with infrastructural, environmental, and temporal features concerning accidents.

Preprocess the dataset: removing duplicates, encoding categorical variables, scaling numerical features, and partitioning data into training and testing subsets to be ready for modeling.

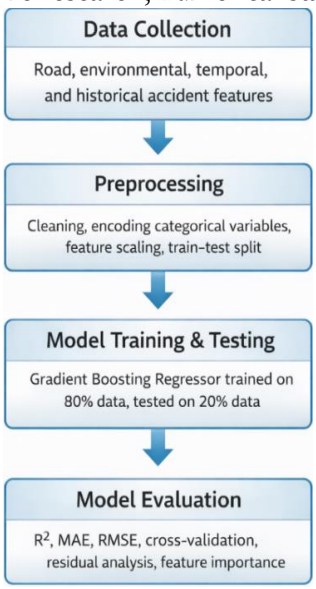
To capture nonlinear relationships between road features and accident risk scores, set up and train a Gradient Boosting Regressor.

In addition to generally using 10-fold cross-validation, residual analysis, and feature importance interpretation, model performance will be measured using the metrics R2, MAE, MSE, and RMSE

Research Methodology:

The section outlines a logical methodological framework used to create a reliable and understandable accident risk prediction model. Data collection, preprocessing, model training

and testing, and model evaluation are the four stages of this workflow of this workflow. Every step was taken with proper scientific research, numerical stability, and reproducibility.



Flow of Study

System Work Flow Architecture:

This methodology consists of four main stages: data collection of roads, environmental, temporal, and historical features; data preprocessing, including cleaning, encoding, scaling, and data splitting; model training and testing using Gradient Boosting Regression; and model evaluation using performance metrics, cross-validation, residual analysis, and feature importance.

Data Collection:

The data for this study were obtained from the Kaggle Playground Series, Season 5, Episode 10. This section describes the simulated data used in the experimental procedure. The data was prepared for machine learning, and the dataset contains 517,754 observations.

The feature set includes thirteen items. These variables belong to four different categories.

Road Infrastructure:

road_type, number of lanes, curvature, speed_limit

Environmental Conditions:

lighting, weather

Regulatory Indicators:

road_signs_present, public_road

Temporal & Seasonal Factors:

time_of_day, holiday, school_season

Another variable, num_reported_crashes, indicates the number of crashes recorded in the past. The variable accident_risk represents the target variable of the prediction model. It is a continuous value ranging between 0 and 1, where higher values indicate a greater probability of accident occurrence. This assumption reflects conditions similar to those observed in controlled settings.

Table 1. Description of dataset variables used in the accident risk prediction model

Feature	Category	Description
road_type	Infrastructure	Type of road (urban, rural, highway)
num_lanes	Infrastructure	Number of lanes on the road segment
curvature	Infrastructure	Degree of road curvature
speed_limit	Infrastructure	Degree of road curvature

lighting	Infrastructure	Lighting condition (daylight, dim, night)
weather	Environmental	Weather condition (clear, rainy, foggy)
road_signs_present	Regulatory	Presence of traffic signs
time_of_day	Regulatory	Indicate whether the road is public
holiday	Temporal	Time period of day (morning, afternoon, evening)
school_season	Temporal	Whether the observation occurred during a holiday
num_reported_accidents	Temporal	Indicator of school season
accident_risk	Historical	Number of previously reported accidents
school_season	Target Variable	Continuous risk score between 0 and 1

Data Preprocessing:

Data preparation is crucial for ensuring data cleanliness and consistency before model training. The important steps that the study follows are given below:

Data Cleaning:

The data were analyzed to confirm their suitability for use. This process is based on various Procedures:

Checking for missing or null values

Verifying data types

Detecting and removing duplicate records

This process ensured a clean dataset without redundancy or corrupted observations.

Categorical Encoding:

All categorical variables were converted into numerical values using a structured approach:

road_type: {urban = 1, rural = 2, highway = 3}

lighting: {daylight = 1, dim = 2, night = 3}

weather: {rainy = 1, clear = 2, foggy = 3}

time_of_day: {afternoon = 1, evening = 2, morning = 3}

Variables such as road_signs_present, public_road, holiday, and school_season indicate true or false conditions and were converted to binary values (0 or 1). This approach provides a compact representation and ensures compatibility with tree-based methods.

Feature Scaling:

Although gradient boosting is generally robust to feature scaling, key numerical variables were scaled using Min Max Scaler to improve stability in computations:

number of lanes

curvature

speed_limit

num_reported_accidents

All selected variables were obtained and scaled between zero and one.

Dividing Data for Training and Testing:

To test whether the model shows better performance than the training set shared 80 percent of the data, 100th, and 20th:

80% for training

20% for testing

The fixed random_state = 42 was used for the output variables.

Model Training and Testing:

The model was developed using the Gradient Boosting feature, Gradual Bridge Replacement (GBR). GBR is well-suited for modeling nonlinear relationships and interactions among mixed-type features.

Algorithm Rationale:

We chose Gradient Boosting Regression because it:

Sequentially minimize error through stage-wise boosting.

Makes good use of different types of products.

It works well on large structured datasets.

Training Configuration:

The model is trained in the following optimized multiple parameters:

n_estimators: 200

learning_rate: 0.1

max_depth: 4

random_state: 42

During training, the model learned relationships between road attributes and accident-risk scores.

Testing was conducted on the withheld 20% dataset to evaluate out-of-sample performance.

Model Evaluation (Methods Only):

The model was evaluated using standard regression metrics to quantify predictive performance:

Coefficient of Determination (R^2):

Measures the proportion of variance in accident risk explained by the model.

Mean Absolute Error (MAE):

MAE=1n∑|ytrue−ypred|MAE = \frac{1}{n} \sum |y_{true} - y_{pred}|MAE=n1∑|ytrue−ypred|

Represents average absolute prediction error.

Mean Squared Error (MSE):

MSE=1n∑(ytrue−ypred)2MSE = \frac{1}{n} \sum (y_{true} - y_{pred}) ^2MSE=n1∑(ytrue−ypred)2

Penalizes larger errors more strongly.

Root Mean Squared Error (RMSE):

RMSE=MSERMSE = \sqrt {MSE}RMSE=MSE

Evaluates prediction accuracy in the original scale.

Cross-Validation:

A **10-fold cross-validation** procedure was used to confirm model stability across multiple data splits.

Residual Diagnostics:

Residual analysis was conducted to evaluate:

error distribution

presence of bias

potential heteroscedasticity

model fit quality

These diagnostic methods ensure the reliability and validity of the trained model.

Features with near-zero importance were retained to preserve model completeness and allow future datasets or regions to reveal different contribution patterns.

Results & Discussion:

The model results are discussed in this section. Summary statistics, the lessons we learned from the model itself, and the tests we performed to evaluate its performance are all included in the discussion. Figures are presented in a logical order to support the reasoning behind each point.

Exploratory Data Analysis (EDA):

Distribution of Accident Risk:

Figure 1 illustrates the distribution of the target variable accident_risk.

X-axis: Accident risk score (continuous values between 0 and 1).
Y-axis: Frequency of observations.
Blue curve: Kernel Density Estimate (KDE) representing the smoothed probability density.
Red dashed line: Mean accident risk value.

The distribution is slightly right-skewed, indicating that low-risk scenarios dominate the dataset, while high-risk situations occur less frequently. This confirms the suitability of a regression-based approach without transforming the target variable.

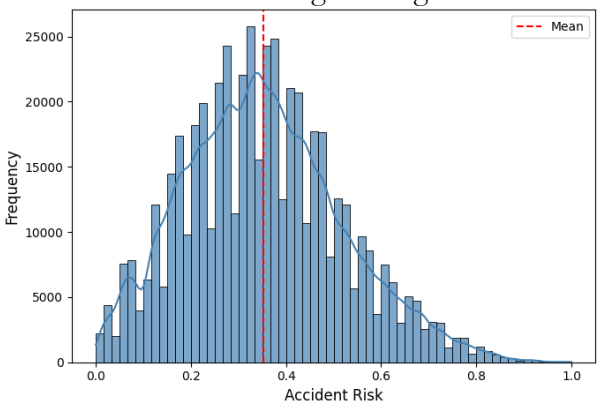


Figure 1. Distribution of the accident_risk target variable

Numerical Feature Distributions:

To understand the behavior of the numerical features, four key numerical values were examined:

Figure 2a illustrates the variability of road curvature across the dataset, indicating the presence of diverse road geometries that may contribute differently to accident risk.

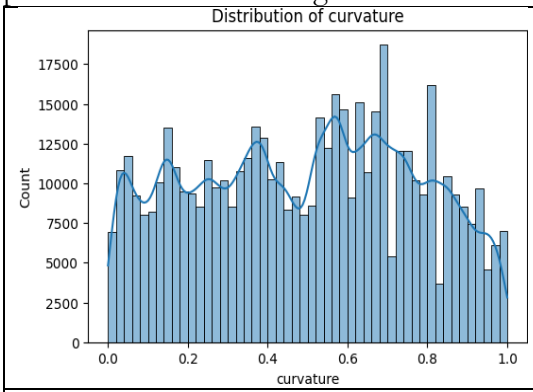


Figure 2a. Distribution of road curvature values across the dataset.

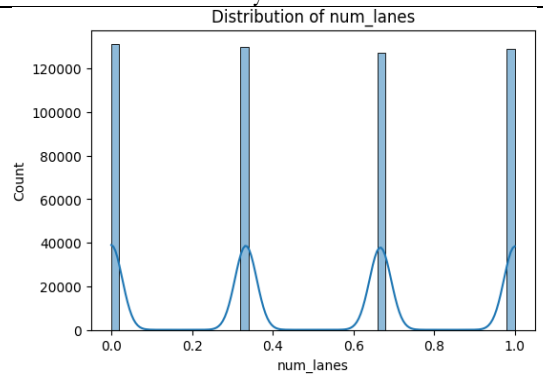


Figure 2b. Distribution of the number of lanes for observed road segments

Figure 2b shows that num_lanes follows a distribution, and the maximum number of paths is observed in most observations.

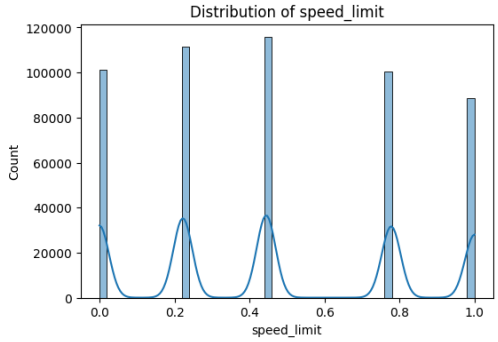


Figure 2c. Indicates that speed limits vary according to road type, with higher limits typically associated with highway segments and lower limits appearing in urban environments.

Categorical Feature Representation:
Figures 3a–3b summarize categorical variables using encoded values.

Encoding note:
Categories are encoded numerically for modeling purposes; numeric labels represent classes, not ordinal magnitude.

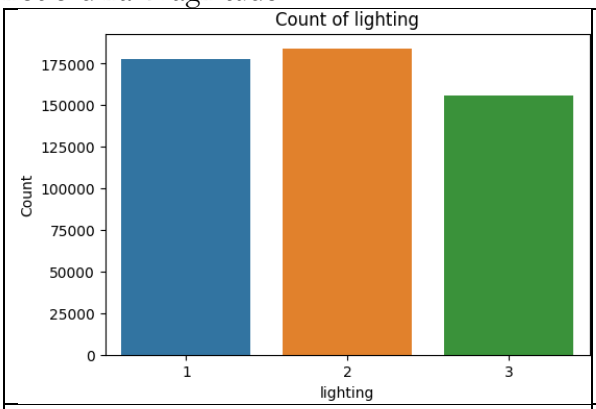


Figure 3a. (Lighting):
Encoded as 1 = Daylight, 2 = Dim, 3 = Night. Daylight conditions dominate, while low-light scenarios remain substantial and safety-critical.

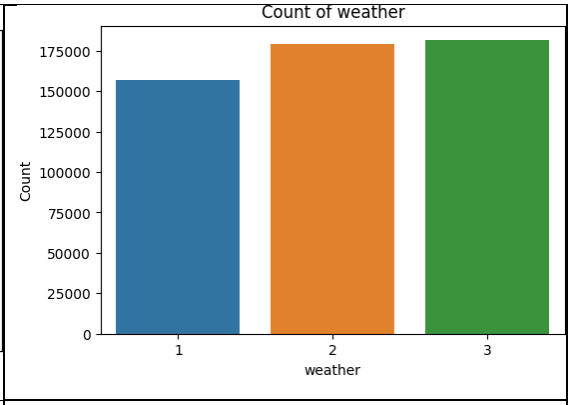


Figure 3b. (Weather):
Encoded as 1 = Clear, 2 = Rainy, 3 = Foggy. Rain and fog appear frequently, which justifies their inclusion as key environmental risk factors.

Correlation Analysis:
The correlation matrix in Figure 4 illustrates the relationships between continuous predictors and the target variable.

Key findings include:
Curvature and speed_limit show strong positive correlations with accident_risk.
num_reported_accidents has a moderate correlation, indicating that historical crashes contribute to the risk estimation.
Low multicollinearity among features supports including all numerical variables in model training.

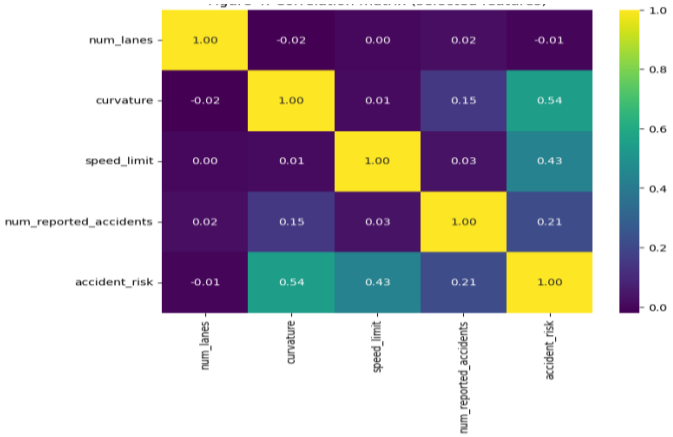


Figure 4. Calculated correlation heatmap of accident risk

Model Insights:

Feature Importance:
The feature importance results of the Gradient Boosting model (Figure 5) identify the most impactful predictors:
Curvature contributes the highest importance score. Speed limit, weather, lighting conditions, and previous accident count appear as strong predictors.

Temporal variables such as *time_of_day* and *holiday* contribute less but still offer additional context.

These concepts are important in guiding road safety interventions.

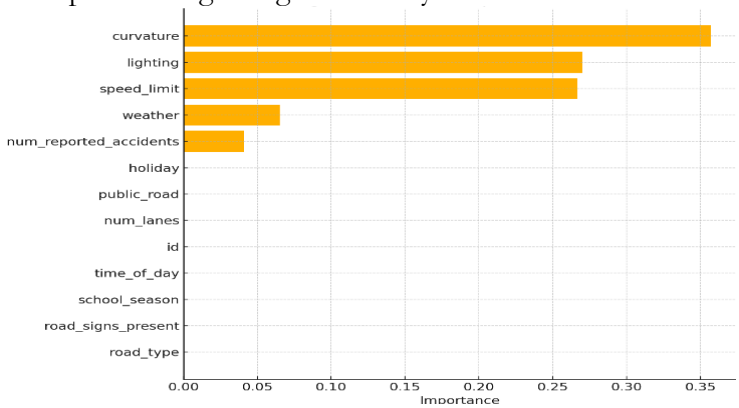


Figure 5. Feature importance scores obtained from the Gradient Boosting Regression model.

Figure 5 shows the feature importance scores from the model used for prediction. Curvature appears as the most influential factor in accident risk, and conditions for lighting follow this, along with the limit for speed and conditions for weather. These factors influence driver visibility, vehicle control, and stopping distance, explaining their key role in risk assessment, and this explains the main role that these factors play in the process of assessing risk.

Various factors, such as the number of lanes, the type of road, the presence of signs on the road, and indicators for holidays, show importance that is near zero. This does not suggest that these elements lack real-world relevance; rather, it indicates that, within the data that the study examines, these factors provided minimal contribution to reducing error in predictions. The model assigns importance only to features used in its decision splits, and features that show limited power for discrimination receive scores that are negligible in a natural way. The results in general suggest that risk for accidents in the data that the study examines is driven in a primary way by conditions that are geometric and environmental conditions that are environmental rather than by categorical or features that relate to regulation.

Predictive Performance Evaluation:

Predicted vs Actual Values:

Figure 6 illustrates the relationship between predicted and actual accident-risk values for the testing dataset. The close clustering of points around the diagonal reference line indicates strong agreement between predicted and observed values, demonstrating high predictive accuracy.

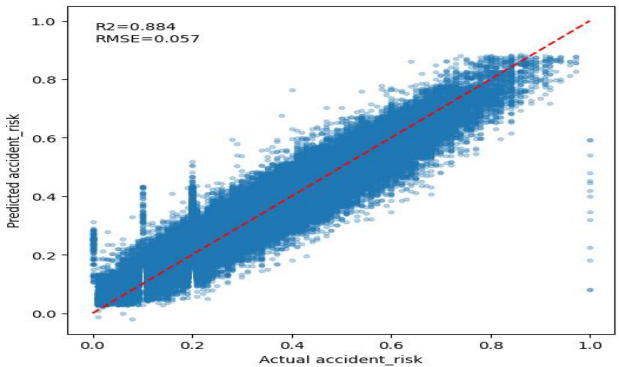


Figure 6. Comparison between predicted and actual accident-risk values in the testing dataset.

Table 2. Performance metrics of the Gradient Boosting Regression model.

Metric	Description	Value
R^2	Coefficient of determination	0.89
MSE	Mean Squared Error	0.00325
RMSE	Root Mean Squared Error	0.057
Cross-validation	Model validation strategy	10-fold

Residual Error Distribution:

The residual diagnostics shown in Figures 7 and Figure 8 are used to verify these statistical assumptions:

Figure 7 shows that the residuals approximately follow the normal trend, the distribution is centered at zero, indicating that predictions are impartial.

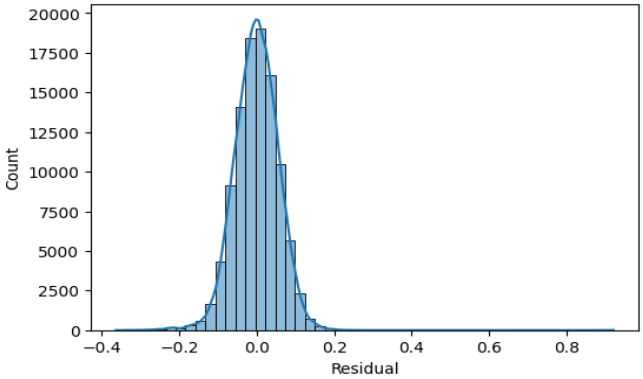


Figure 7. Histogram of residual errors from the Gradient Boosting Regression model.

Figure 8 plots residuals against predicted values and exhibits no funneling or directional patterns, confirming homoscedasticity and stable variance.

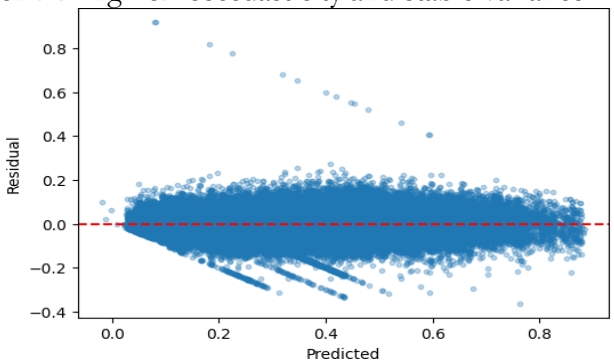


Figure 8. Scatter plot of residual errors versus predicted accident-risk values.

Cross-Validation Performance:

The model's robustness was evaluated using 10-fold cross-validation. Figure **9** reports R^2 scores across all folds, showing minimal variability and confirming that the model generalizes well beyond the training dataset.

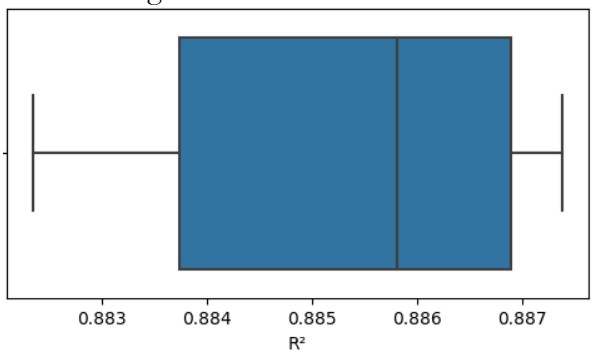


Figure 9. 10-fold cross-validation R^2 scores

Future Work:

The Gradient Boosting Regression model provides insightful information about accident-risk trends. However, this study could be further improved and operationalized in a number of ways. For example, real-time data integration with traffic, weather, and sensor-based inputs could be used in place of static inputs. In this manner, the accident-risk monitoring model would continue to be dynamic, improving its suitability for use in an Intelligent Transportation System. We could try models like XGBoost or CatBoost in the future, and explore a small deep learning setup to assess potential performance improvements. To improve predictions, it may also be beneficial to incorporate behavioral or spatial features and eventually turn the model into a simple tool with a dashboard for agencies.

Conclusion:

This study developed a feature-driven Gradient Boosting Regressor (GBR) model to predict road accident risk using environmental, infrastructural, and temporal variables. Using a dataset of more than 517,000 road condition observations with 13 features, the proposed model achieved strong predictive performance with an R^2 value of approximately 0.89. Model validation through cross-validation and residual diagnostics confirmed the stability and reliability of the predictions. Feature importance analysis indicated that road curvature, lighting conditions, speed limits, weather conditions, and historical accident frequency are the most influential predictors of accident risk. These findings highlight the importance of environmental and road design factors in traffic safety analysis. The proposed framework provides a practical tool for identifying high-risk road segments and supports data-driven decision-making for transportation authorities and policymakers. Future research may integrate real-time traffic and environmental data to further enhance accident risk prediction in intelligent transportation systems.

References:

[1] Dragan Gatarić, Nenad Ruškić, "Predicting Road Traffic Accidents—Artificial Neural Network Approach," *Alogrithms*, vol. 16, no. 5, p. 257, 2023, [Online]. Available: https://www.mdpi.com/1999-4893/16/5/257

[2] Chukwutoo C. Ihueze, Uchendu O. Onwurah, "Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria," *Accid. Anal. Prev.*, vol. 112, pp. 21–29, 2018, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0001457517304542

[3] N. Alpalhão, P. Sarmento, and Bruno Jardim, "Assessing the risk of traffic accidents in lisbon using a gradient boosting algorithm with a hybrid classification/regression approach," *Transp. Res. Interdiscip. Perspect.*, vol. 32, p. 101495, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590198225001745

[4] G. Singh, M. Pal, Y. Yadav, and T. Singla, "Deep neural network-based predictive modeling of road accidents," *Neural Comput. Appl. 2020 3216*, vol. 32, no. 16, pp. 12417–12426, Jan. 2020, doi: 10.1007/s00521-019-04695-8.

[5] A. K. Sheng Dong, "Predicting and Analyzing Road Traffic Injury Severity Using Boosting-Based Ensemble Learning Models with SHAPley Additive exPlanations," *Int. J. Environ. Res. Public Heal.*, vol. 19, no. 5, p. 2925, 2022, [Online]. Available: https://www.mdpi.com/1660-4601/19/5/2925

[6] N. A. Kenan Menguc, "A Data Driven Approach to Forecasting Traffic Speed Classes Using Extreme Gradient Boosting Algorithm and Graph Theory," *Phys. A Stat. Mech. its Appl.*, vol. 620, 2023, [Online]. Available: https://ideas.repec.org/a/eee/phsmap/v620y2023ics0378437123002935.html

[7] A. Theofilatos, G. Yannis, P. Kopelias, and F. Papadimitriou, "Predicting Road Accidents: A Rare-events Modeling Approach," *Transp. Res. Procedia*, vol. 14, pp. 3399–3405, 2016, doi: 10.1016/J.TRPRO.2016.05.293.

[8] S Govindaraju, M Indirani, Siti Sarah Maidin, Jingchuan Wei, "Intelligent Transportation System's Machine Learning-Based Traffic Prediction," *J. Appl. Data Sci.*, vol. 5, no. 4, 2024, [Online]. Available: https://bright-journal.org/Journal/index.php/JADS/article/view/364

[9] Noura Hamdan, Tibor Sipos, "Predicting Segment-Level Road Traffic Injury Counts Using Machine Learning Models: A Data-Driven Analysis of Geometric Design and Traffic Flow Factors," *Futur. Transp*, vol. 5, no. 4, p. 197, 2025, doi: https://doi.org/10.3390/futuretransp5040197.

[10] Bappa Muktar, Vincent Fono, "Toward Safer Roads: Predicting the Severity of Traffic Accidents in Montreal Using Machine Learning," *Electronics*, vol. 13, no. 15, p. 3036, 2024, doi: https://doi.org/10.3390/electronics13153036.

[11] Alexei Roudnitski, "Evaluating Road Crash Severity Prediction with Balanced Ensemble Models," *Findings*, 2024, doi: 10.32866/001c.116820.

[12] M. S. Markus Deublein, "Prediction of road accidents: A Bayesian hierarchical approach," *Accid. Anal. Prev.*, vol. 51, pp. 274–291, 2013, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0001457512004101