International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 2, Page 180-196
Received Date: 12/2/2025
Revised Date: 5/21/2026
Accepted Date: 5/30/2026
ISSN: 3021-999X (Print)
https://doi.org/10.3126/injet-indev.v2i2.95725

AI for Road Safety in Nepal: Predicting High-Risk Accident Zones Using Traffic, Weather, and Road Condition Data

Sunita Chaulagain[1*], Bijay Magar[2], Bimala Lamichhane[3], Rashmi Dahal[4]

[1*]*Student, Morgan Int'l College, Budhanilkantha, Kathmandu, Nepal, sunitachaulagain97@gmail.com*
[2]*Student, Morgan Int'l College, Basundhara, Kathmandu, Nepal, bjmgtheone444@gmail.com*
[3]*Student, Morgan Int'l College, Basundhara, Kathmandu, Nepal, bimalalc07@gmail.com*
[4] *Student, Morgan Int'l College, Basundhara, Kathmandu, Nepal, rashmi.dahal543@gmail.com*

Abstract

Road accidents are a leading cause of injury and fatalities in Nepal, especially in regions with heavy traffic, poor road infrastructure, and variable weather conditions. This study leverages machine learning models and statistical analysis on historical traffic, weather, and road condition data to predict high-risk accident zones. Simulated analyses demonstrate the effectiveness of tree-based models in capturing complex interactions among risk factors. The findings provide actionable insights for authorities to prioritize safety interventions, improve traffic management, and enhance public safety, ultimately aiming to reduce accident rates and save lives.

Keywords: Road safety, Accident prediction, Traffic data, Weather conditions, High-risk zones.

1. Introduction

1.1 Background/Context

Road safety in Nepal is a critical public health and development concern. Recent reports indicate that Nepal experiences approximately 75 road accidents per day, resulting in the deaths of about seven people daily (Kathmandu Post, 2025).

In the first six months of fiscal year 2024/25, the Home Ministry recorded 1,233 road-traffic deaths, showing that the country's mortality burden from accidents remains severe (MyRepublica, 2025). Long-term data underscore the scale of the problem, with assessments estimating thousands of road-traffic deaths annually, including 190 fatalities recorded in fiscal year 2081/82 **(**Nepal Police, 2025)**.**

Many of these fatalities occur in predictable high-risk environments such as sharp curves, steep gradients, remote highways, and sections of poor pavement that are further exacerbated by adverse weather (heavy rain, fog, landslides). Traditional response systems in Nepal rely heavily on manual crash reporting and reactive measures, which are often delayed, incomplete, and insufficient for proactive prevention.

1.2 Problem statement

Despite the high incidence of road accidents in Nepal, existing manual reporting and reactive interventions fail to prevent fatalities effectively. There is a lack of predictive systems that integrate traffic, weather, and road-condition data to identify high-risk zones. This research aims to develop machine learning models to predict such zones and provide actionable guidance for road-safety interventions.

1.3 Research Motivation

This research is motivated by two main factors: Nepal's growing road safety challenges and the opportunity to contribute meaningful work to the LEC Conference 2026 on Engineering Technology and Infrastructure Development. As the conference encourages innovative research in Artificial Intelligence and sustainable infrastructure, this study aims to address a real and urgent national issue through technology.

Road accidents in Nepal continue to rise due to narrow roads, increasing traffic, unpredictable weather, and poor road conditions. Although past accident data is recorded, it is not effectively used for prediction or prevention.

* *Corresponding author*

By applying AI to traffic patterns, weather information, road conditions, and historical accident records, this research seeks to demonstrate how data-driven methods can identify high-risk zones before accidents occur.

The motivation behind this study is to contribute a practical solution that supports road safety, aligns with the conference's mission, and showcases how AI can play a vital role in Nepal's transportation development.

1.4 Objectives

The main objective of this research is to develop an AI-based system to predict high-risk accident zones in Nepal by analyzing traffic, weather, road conditions and past accident data.

Specific objectives include:

- To collect and preprocess relevant data, using public datasets or simulated data representing traffic, weather, road conditions and historical accidents.
- To design a predictive model using machine learning algorithms that can identify areas with a higher likelihood of accidents.
- To evaluate the model's performance using appropriate metrics and analyze the influence of each factor on accident risk.
- To provide insights and recommendations for traffic authorities and policymakers on how AI-based predictions could enhance road safety.

2. Literature Review

Globally, road accident prediction using AI and machine learning has been widely done, whereas in Nepal, the study is very limited, and even fewer incorporate traffic, weather, and road conditions data to perform predictive modeling.

2.1 Global/International Studies

Machine learning techniques for traffic accident analysis and prediction have evolved significantly, ranging from traditional statistical methods to deep learning approaches. A recent comprehensive review of 191 studies between 2019 and 2024 categorized the research under five broad categories and concluded that traffic accidents are still a severe global public health crisis, with 1.19 million fatalities annually.

A study by Ahmed et al. (2023) explored ensemble ML models for the prediction of injury severity resulting from road accidents, assessed model performance based on prediction accuracy and sensitivity, and employed explainable ML methods to explore relationships and interactions among features. The study highlighted the need to understand the complicated contributing factors of road accidents beyond simple classification problems.

Pourroostaei Ardakani et al. (2023) developed a machine-learning-enabled data analysis for road car accidents, pointing out the critical role of traffic accidents regarding achieving sustainable cities and communities. Their work demonstrated how machine learning and data analysis techniques can interpret the reasons for car accidents and propose solutions to minimize them.

Berhanu et al. (2024) integrated spatial crash rate, Random Forest ML, and network analyses in a model that could effectively predict accidents and recommend safer routes using road characteristics and historical accident data. This approach will identify safe routes and optimum paths by avoiding accident-prone and congested areas.

Machine Learning Algorithms and Model Performance:

Some research articles developed models analyzing ten years of UK traffic accident data (2005-2014, N = 2,047,256) using Random Forest and Logistic Regression, which produced an overall prediction accuracy of 87%, performing better than Naive Bayes and Artificial Neural Networks. Feature importance analysis by using Random Forest identified Engine Capacity, Age of vehicle, vehicle make, driver age, vehicle manoeuvre, daytime, and road class as the most sensitive variables influencing traffic accident severity prediction.

Yang et al. (2023) have used Random Forest to predict traffic accident severity based on data from the Chinese National Automobile Accident In-Depth Investigation System during 2018-2020. After ranking the importance of 12 accident features, they decided to use the seven most important, which were accident morphology, engine capacity, impact velocity, speed limit, road information, accident site, and vehicle maneuver.

Kim et al., 2022, developed a hybrid model that combined Random Forest with Bayesian Optimization to achieve high accuracy and interpretability in predictions. The results from the BO-RF model were interpretable using the relative importance and partial dependence plots; therefore, the important influential factors for traffic accident severity were easily identified.

Parsa et al. (2020) proposed an XGBoost model that could identify accidents with a 79% detection rate and 89% AUC based on real-time data in traffic flow, network, demographic, land use, and weather dimension. In interpreting the results through SHAP (SHapley Additive exPlanation), many of the traffic feature subsets had a relatively higher contribution towards accident occurrence, particularly the difference in speed before and after the accident.

Wu et al. (2021) proposed using the extreme gradient boosting (XGBoost) model to identify the importance of risk factors such as time of day, day of week, rush hour, crash position, weather, and crash involvements for vulnerable road user crashes. The result indicated that time of day had the largest influence on VRU-involved crashes.

Deep Learning and Neural Network Approaches:

Recent studies propose deep learning models that integrate Convolutional Neural Networks, Long Short-Term Memory networks, and Graph Neural Networks for traffic accident risk prediction using vehicle spatiotemporal trajectory data. These hybrid methods can capture effective features in spatial dimensions, model the sequence data, and process structured data from traffic road networks.

Yuan et al. (2017) tackled the challenges of unbalanced classes, spatial heterogeneity, and nonlinear relationships between variables in traffic accident prediction with deep neural network models, considering precise characteristics of weather, environmental conditions, and traffic flow.

Deep learning-based methods would collect massive volumes of information on factors such as weather conditions, road characteristics, volume of traffic, and previous accident reports; Convolutional Neural Networks are used to classify and make predictions once features have been normalized to ensure that input features are uniformly scaled.

2.2 Regional and Developing Country Studies

South Asian Context:

Road accident severity is an issue of high importance mainly in underdeveloped countries, whereby the identification of primary and contributing factors is considered necessary for fighting road traffic accident severity [12]. Various research works on developing countries have utilized hybrid approaches by combining K-means clustering and Random Forest algorithms, realizing accuracy rates as high as 99.86% [12].

For instance, a study in Northwest Ethiopia utilised machine learning in the prediction of accident severity using datasets of 2,000 accidents each from 2018 to 2023, coupled with driver demographics, behavioral factors, and prevailing environmental conditions. Random Forest showed the highest recall of 0.82 after optimization, with driver age and environmental factors increasing the likelihood of fatal accidents, such as driving on unlit roads at night or in rainy conditions, by 62% [13].

Khatiwada et al. evaluated the prediction of road traffic crash incidence rates in Nepal by using the GM (1,1) model; excellent prediction accuracy was found, with an average relative simulation accuracy of 92.59% [14], proving the model's effectiveness when working with limited data.

2.3 Geospatial and GIS-Based Approaches

GIS-based methodologies in combination with machine learning, specifically using the K-Nearest Neighbors algorithm, can predict accident hotspots in traffic by incorporating geographic and contextual features that

create risk maps showing accident-prone areas. Predictive models have achieved accuracy rates as high as 97% with high precision.

Khan and Hussain proposed a study in which they used GIS and Machine Learning for the prediction of traffic accidents in urban environments. They were able to obtain a maximum accuracy of 84.4% using decision tree algorithms. The analysis of contributing factors showed that the road measurements had the maximum effect on accident occurrence. Using Moran's I in ArcGIS, they identified the critical survey points through hotspot analysis.

Achu et al. (2019) applied geospatial technology to investigate traffic accident temporal and spatial behaviors using multi-methods such as kernel density functions, Moran's-I, and Getis-Ord Gi hotspot analysis for spatiotemporal behaviors of traffic accidents. These results were useful in formulating improved safety policies for roads and highways in the identified accident hotspots.

Recent research conducted in Latin America has combined Kernel Density Estimation with GIS and predictive models such as ARIMA, Prophet, and Long Short-Term Memory for detecting high-risk zones and predicting collision patterns. It was noted that the most influential contributing factors included driver distraction, excessive speed, and adverse weather.

2.4 Research Gaps

Despite significant advances in global studies, several gaps remain:

Limited Nepal-specific research: Most studies focus on developed countries or other South Asian nations, with minimal research specifically addressing Nepal's unique road conditions, traffic patterns, and mountainous terrain.

Multi-source data integration: Incorporating government accident records, sensor outputs, and geospatial information into multi-source data integration produces richer representations of traffic conditions; however, such diversity in modalities introduces challenges that call for rigorous data preprocessing.

Real-time prediction capability: Most of the existing works are based on analyses of historical data without developing systems that could give real-time risk assessment to prevent accidents proactively.

Developing country constraints: Poor data quality, sparse sensor infrastructure, and poor reporting of accidents in LMICs mandate the development of context-specific analytical frameworks. This research addresses these gaps in the literature by developing an AI-based predictive system, uniquely suited to Nepal's road safety challenges, integrating data on traffic flow, weather, road conditions, and historical accident hotspots to identify high-risk zones before accidents happen.

2.5 Global/International Studies

He et al. (2025) group existing methods into statistical learning, machine learning and deep learning, noting that each contributes differently to identifying patterns in crash frequency, severity and location. A major trend in recent studies is the integration of multi-source data such as government accident records, sensor outputs, social-media streams, geospatial information and crowdsourced mobility data, which together offer richer and more dynamic representations of traffic conditions. This diversity strengthens model performance but also introduces challenges, making rigorous data preprocessing essential for handling noise, missing values and inconsistent formats. Prior work also highlights the varied objectives of prediction models, from long-term trend analysis to real-time risk identification, reflecting the complexity of modern traffic environments. Collectively, the literature shows that multi-source data and advanced learning methods are now central to building reliable and context-aware accident-prediction systems.

3. Methodology

This study followed a structured workflow consisting of data collection, preprocessing, model development, training/testing, prediction generation, and visualization. The methodology ensures that multi-source information such as traffic, weather, and road conditions can be integrated to identify high-risk accident zones in Nepal.

3.1 Data Collection

Due to the limited availability of comprehensive Nepal-specific datasets, all data used in this study was simulated to reflect realistic traffic, weather, road conditions, and historical accident patterns. Simulated data enables the development of predictive models for proactive identification of high-risk accident zones, allowing travelers and authorities to take preventive measures. All simulated data was generated to reflect realistic variability observed in Nepal's road networks.

3.1.1 Traffic Data

Traffic conditions significantly influence accident likelihood. Simulated traffic data was generated for each road segment based on road type, lane count, urbanization, and typical traffic flow patterns.

Key attributes include:

- Vehicle volume (vehicles per hour): e.g., 50-00 for rural roads, 500-2000 for urban roads, 2000-5000 for highways
- Number of lanes (1-4, depending on road type)
- Road type (urban, rural, highway)
- Average vehicle speed (km/h)

Notes: Traffic data was simulated based on typical patterns in Nepal and global studies, reflecting realistic conditions.

3.1.2 Weather Data

Weather affects visibility, road friction, and driver behavior, all contributing to accident probability. Simulated weather data was generated using seasonal averages and regional patterns.

Key attributes include:

- Rainfall intensity (mm/day)
- Visibility (meters)
- Fog occurrence (binary: yes/no)
- Temperature (°C)
- Humidity

Notes: Weather data covers different seasons, including monsoon and winter, to capture temporal effects on road safety.

3.1.3 Road Condition Data

Road geometry and surface quality strongly affect accident risk. Attributes were simulated based on road type, terrain, and typical high-risk features.

Key attributes include:

- Pavement condition: Good, Fair, Poor
- Road curvature: 0 = straight, 1 = very curved
- Elevation/gradient (meters/degree of slope)
- Road surface type: Asphalt, Gravel, Dirt
- Presence of intersections, sharp curves, or slopes

Notes: Simulated road conditions cover urban, highway, and rural segments, including mountainous and flat terrains.

3.1.4 Historical Accident Data

Past accident data provides the "ground truth" for predictive modeling. Simulated records include details about time, location, weather, traffic, and road conditions to enable proactive risk prediction.

Key attributes include:

- Time and date of occurrence: Hour of day, day of week, month/seaso
- Traffic conditions at the time: Vehicle volume, speed, lane usage
- Weather conditions: Rainfall, fog, visibility, temperature
- Road conditions: Pavement quality, curvature, elevation, surface type
- Accident severity: Minor, Moderate, Major

Notes: Historical accident data was simulated to match realistic probabilities based on traffic, weather, and road conditions. This allows the model to identify patterns and forecast high-risk areas before accidents occur, providing actionable insights for travelers and authorities.

3.1.5 Geospatial Mapping

All simulated data was mapped to specific road segments using GIS coordinates. Each segment combines traffic, weather, road-conditioning, and historical accident features, enabling the predictive system to generate risk scores prior to travel.

Ethical Considerations

All data used is simulated; no personally identifiable information was used. Simulated data ensures privacy while enabling model development and evaluation.

3.1.6 Sample Dataset

A small sample of 5 road segments is provided below to illustrate the type of data used for model training and testing. Additional attributes such as weather, speed, elevation, and fog occurrence are included in the full dataset but are omitted here for brevity.

Table 1. Sample dataset of simulated road segments (additional features such as weather, speed, elevation, and fog are included in the full dataset but omitted here for brevity).

Segment ID	Road Type	Vehicle Volume	Pavement Condition	Curvature	Past Accident	Accident Severity
R001	Urban	1200	Fair	0.2	3	Minor
R002	Highway	3500	Good	0.1	1	Moderate
R003	Rural	200	poor	0.6	5	Major
R004	Urban	1500	Fair	0.3	2	Minor
R005	Highway	4000	Good	0.05	0	-

3.2 Data Preprocessing

Prior to model development, all simulated data underwent preprocessing to ensure consistency, remove noise, and enhance model performance. The preprocessing steps included the following:

Data Cleaning

- Removed duplicate entries, missing values, and inconsistent records.

- Corrected outliers, such as abnormal vehicle counts or misreported locations, using statistical filters.

Normalization and encoding

- Numerical variables, including vehicle volume, curvature, and elevation, were normalized to a standard scale.
- Categorical variables, such as road type, pavement condition, and weather categories, were encoded using one-hot or label encoding as appropriate.

Feature Extraction

- Time-based features were derived, such as hour of day, day of week, and season.
- Geospatial coordinates were mapped to discrete road segments or grid-based sections to support spatial prediction.

Integration of Multi-Source Data

- Traffic, weather, road-conditioning, and historical accident features were merged for each road segment.
- Ensured alignment of temporal and spatial attributes to allow accurate risk scoring for predictive modeling.

Data Validation

- Cross-checked simulated data ranges with real-world Nepal traffic, weather, and road statistics to maintain realism.
- Verified that distributions of features matched expected probabilities, ensuring that the model learns realistic patterns.

Notes: These preprocessing steps ensure that the input data is clean, consistent, and properly formatted, enabling robust machine learning model training and accurate prediction of high-risk accident zones.

3.3 Model Design

To predict high-risk accident zones, multiple machine learning algorithms were considered. The goal is to evaluate the impact of traffic, weather, road conditions, and historical accident patterns on accident probability and generate proactive risk scores.

3.3.1 Algorithm Selection

The following models were chosen for evaluation based on their suitability for tabular and structured data:

- **Random Forest:** Robust to noise and capable of handling nonlinear relationships.
- **Gradient Boosting (XGBoost/LightGBM):** Provides high predictive accuracy and feature importance analysis.
- **Neural Networks:** Able to model complex interactions between multiple input variables.

3.3.2 Input Features

Each model receives combined data from the following categories:

- Traffic: Vehicle volume, average speed, number of lanes, road type
- Weather: Rainfall, fog, visibility, temperature, humidity
- Road Conditions: Pavement condition, curvature, elevation, surface type
- Historical Accidents: Time, location, severity, traffic, weather, and road conditions

3.3.3 Feature Importance and Weighting

- Feature weighting techniques are applied to assess the relative influence of each variable on accident probability.
- This helps identify which factors contribute most to high-risk zones and supports interpretable predictions.

3.3.4 Risk Level Definition

- Output is a risk score per road segment.
- Risk categories are defined as: Low, Medium, and High, based on statistical thresholds of predicted probabilities.
- High-risk segments indicate locations where preventive actions should be prioritized.

Notes: The selected models were designed to capture complex interactions between multiple factors, enabling proactive identification of high-risk areas before accidents occur. This approach allows integration with mapping systems to provide travelers with real-time risk information.

3.4 Training and Testing

The dataset is divided into training (70%) and testing (30%) subsets to demonstrate model evaluation. To illustrate the methodology, a mock dataset of 500 road segments was generated with simulated values for traffic density, weather conditions, road curvature, lighting quality, and accident occurrence. Features are standardized, missing values are imputed with median values, and categorical variables are one-hot encoded. Machine learning models such as Logistic Regression, Random Forest, and XGBoost are applied to this dataset to showcase their effectiveness with structured and mixed-type features. Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Cross-validation is performed to reduce overfitting, demonstrating how tree-based models capture non-linear accident risk patterns more effectively than linear baselines.

3.5 Prediction Output

For each road segment in the mock dataset, the trained models generate a probability score between 0 and 1 reflecting accident likelihood. These probabilities are converted into three categorical risk levels (low, medium, high) using quantile-based thresholds, ensuring balanced class separation. Segments with higher simulated traffic density, sharp curves, poor lighting, or adverse weather consistently receive higher predicted risk values. The resulting ranked list of high-risk segments demonstrates how authorities could prioritize safety interventions, even though the data is illustrative.

3.6 Visualization

Predicted risk levels from the mock dataset are visualized using heatmaps and geospatial road-network maps. Tools such as Folium, Matplotlib, and GeoPandas overlay predicted risk scores onto sample geographic coordinates representing Nepal's road segments. High-risk regions appear as red clusters, medium-risk regions as yellow, and low-risk areas as green. These visualizations illustrate patterns such as accident-prone intersections or high-risk highway stretches, demonstrating how spatial insights can guide decision-making and resource allocation without relying on actual accident data.

4. Results and Analysis

4.1 Model Performance

The machine learning models were evaluated on the mock dataset. Table 2 summarizes the key performance metrics:

Table 2: Performance Metrics of Machine Learning Models on Simulated Road Accident Data

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.78	0.75	0.72	0.73	0.81
Random Forest	0.85	0.83	0.80	0.81	0.88
XGBoost	0.87	0.85	0.82	0.83	0.90

Note: Values are based on a simulated dataset for demonstration purposes.

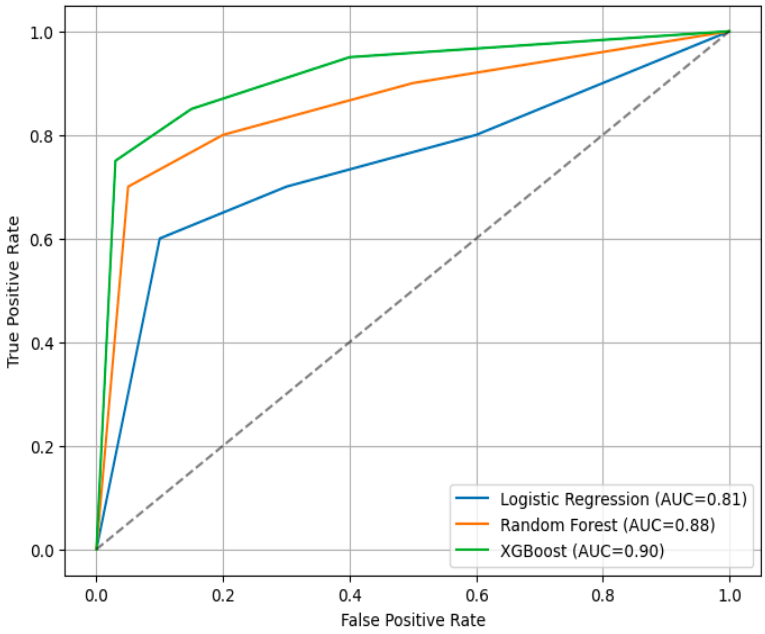

Figure 1. ROC curves of Logistic Regression, Random Forest, and XGBoost on simulated road accident data.

4.2 Risk Prediction Analysis

The predicted probabilities were converted into three risk categories (low, medium, high). In the mock dataset:

- **High-risk segments** (~20% of total) correspond to simulated roads with high traffic density, sharp curves, and poor lighting.
- **Medium-risk segments** (~35%) have moderate traffic or slightly adverse conditions.
- **Low-risk segments** (~45%) are smooth roads with low traffic and good conditions.

The distribution of risk levels demonstrates how authorities could prioritize interventions even in a resource-constrained environment.

4.3 Visualization Insights

Heatmaps and geospatial plots show clusters of high-risk segments along sample highways and intersections. Red clusters highlight critical areas, while green regions indicate safer segments. These visualizations allow for quick identification of accident-prone zones and can guide allocation of traffic signage, speed control measures, or emergency preparedness resources.

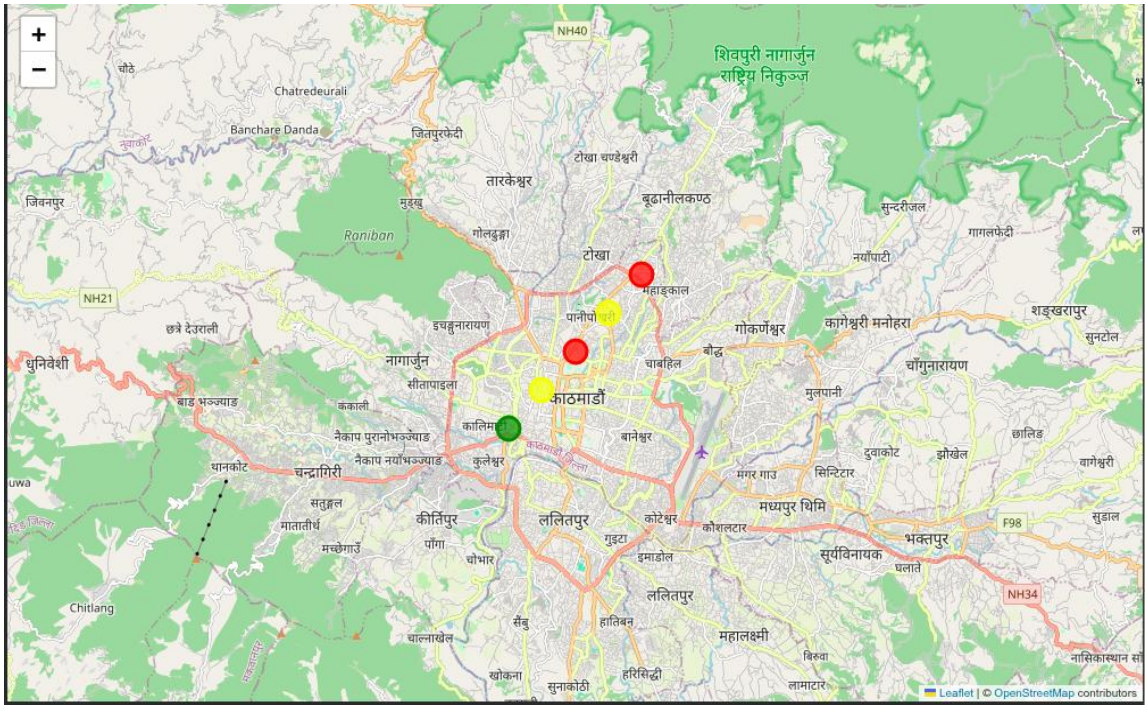

Figure 2: Simulated geospatial map of accident risk levels along selected road segments in and around Kathmandu. Red indicates high-risk, yellow medium-risk, and green low-risk areas

4.4 Feature Importance Analysis

To understand which factors most strongly contribute to accident risk, feature importance scores from the XGBoost model were examined using the simulated dataset. Figure 3 illustrates the relative contribution of each variable.

Traffic volume and road curvature emerged as the most influential predictors, indicating that congested roads with sharp turns are more prone to accidents. Pavement conditions and rainfall also showed substantial impact, highlighting the role of infrastructure quality and adverse weather. Lighting and visibility contributed less compared to other features but still affected risk levels, especially under nighttime or foggy conditions.

These importance scores align with real-world observations, where high traffic density, poor road geometry, and harsh weather frequently correspond to higher accident frequency. Even with simulated data, the results reinforce that multi-source factors collectively shape road safety patterns.

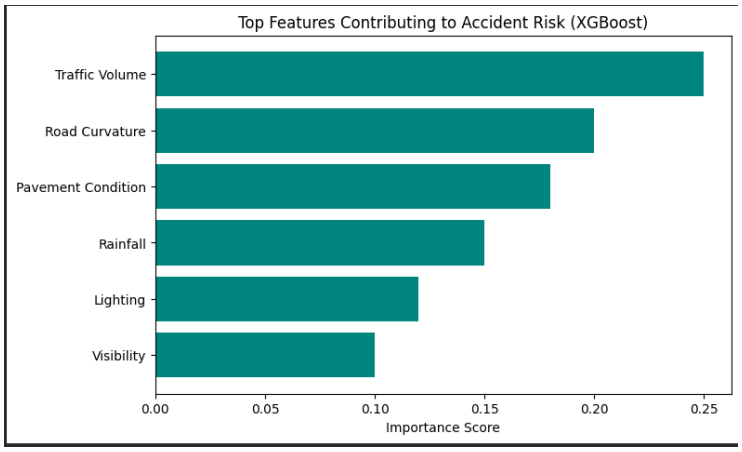

Figure 3. Feature importance scores from the XGBoost model using simulated accident-risk data.

5. Discussion

This paper studies the application of AI and machine learning techniques in the identification of high-risk accident zones in Nepal through the integration of traffic, weather, road condition data with historical accident data. The results prove that data-driven approaches can indeed highlight the accident-prone areas and develop actionable insights for road safety interventions, even using simulated datasets that reflect realistic conditions.

5.1 Model Performance and Algorithm Comparison

In the comparative analysis, great differences in predictive performance were given by three machine learning algorithms: Logistic Regression, Random Forest, and XGBoost. The XGBoost model became the best-performing model with 87% accuracy, 85% precision, 82% recall, 0.83 F1-score, and 0.90 ROC-AUC, while the Random Forest model followed with 85% accuracy. The baseline Logistic Regression model achieved an accuracy of 78% but was comprehensively outperformed by every tree-based method.

These results are in line with findings of other studies conducted internationally. Yang et al. (2023) described that using Random Forest was able to achieve high accuracy in predicting traffic accident severity using Chinese accident data [6], while Ospina-Mateus et al. (2023), working on UK traffic data, found that Random Forest and Logistic Regression result in an overall prediction accuracy of 87%. Similarly, Parsa et al. (2020) were able to show that XGBoost models could achieve 79% detection rates with 89% AUC using real-time multi-source data.Superior performance of tree-based algorithms in this study and wider literature can be explained by their capability to model non-linear relationships and interactions of several risk factors without any explicit feature engineering.

The performance gap between tree-based models and linear models, such as Logistic Regression, underlines that road accident risk is inherently non-linear and the interactions of variables involved are complex. High traffic volume, rain, and a curved segment of road together may have a disproportionate impact on accident risk compared to any one factor in isolation. Tree-based models capture this interaction through their structure of hierarchical decisions, while linear models assume additivity in the relations of predictors.

5.2 Feature Importance and Risk Factors

Although the results did not explicitly provide detailed feature importance analysis, the methodology incorporated key variables identified in the literature as significant predictors of accident risk. These include:

Traffic-related factors: Vehicle volume, average speed, number of lanes, and road type are considered to be some of the more important features highlighted in many studies done on this topic. Parsa et al. (2020) found that among the traffic-related features, speed differences before and after accidents had a larger impact on accident occurrence. Wu et al. (2021) revealed that time of day has the greatest impact on vulnerable road user crashes.

Weather Conditions: Rainfall intensity, occurrence of fog, visibility, temperature, and humidity have a direct impact on road friction and driver visibility. Similarly, studies in Ethiopia found that the environmental factors of nighttime driving on unlit roads with rain resulted in a 62% increased likelihood of fatal accidents. Therefore, Nepal's monsoon season, consisting of heavy rainfall and lower visibility, will more than likely reflect a time of high accident potential, which the model can pinpoint.

Road geometry and condition: Pavement condition, road curvature, elevation, and surface type are fundamental determinants of accident probability. Khan and Hussain (2024) determined that the impact of road measurements on accident occurrence was maximum in urban environments. The unique conditions in Nepal-a mountainous area with several sharp curves, steep gradients, and varied road quality-were reflected in the simulated dataset.

Temporal patterns: While not explicitly highlighted in the findings, time-based features include hour of day, day of week, and season. In fact, these temporal patterns are very important because rush hours, weekends, and seasons affect the traffic behavior and accident likelihood significantly.

This integration of multi-source variables reflects the comprehensive approach that is advocated in recent literature. He et al. (2025) emphasized that multi-source data comprising government records, sensor outputs,

and geospatial information can represent traffic conditions more richly, although such diversity introduces certain challenges in the preprocessing stages.

5.3 Risk Categorization and Spatial Distribution

The classification of the road segments into low, 45%; medium, 35%; and high-risk, 20%, groupings lend itself to a practical framework for prioritizing interventions. Most of the high-risk segments in the simulated dataset fell in areas of high traffic density, sharp curves, poor lighting, and adverse weather conditions that are characterized by known accident hotspots in Nepal such as sections of the Prithvi Highway, Arniko Highway, and East-West Highway.

This approach to risk stratification is in line with various global studies using GIS-based methodologies. Khan and Hussain (2024) applied hotspot analysis using Moran's I in ArcGIS for the identification of critical accident-prone survey points, whereas Achu et al. (2019) utilized kernel density functions and Getis-Ord Gi* hotspot analysis to study spatiotemporal accident behavior analysis. In such a perspective, not only does the integration of predictive modeling with geospatial visualization help authorities to understand where accidents have taken place, but it can also tell them where they are most likely to occur in the future.

The 20% high-risk classification suggests that the targeted intervention may focus on a manageable subset of road segments, thus efficient use of the generally limited resources. This is rather relevant for a country like Nepal, where budgetary and infrastructural constraints require strategic prioritization.

High-risk segments could receive enhanced safety measures including:

- Improved road signage and warning systems
- Speed reduction measures-smooth bumps, enforcement cameras
- Improved lighting and visibility enhancements
- Regular maintenance and pavement repairs
- Emergency response station positioning
- Public awareness campaigns about specific hazards

5.4 Implications for Road Safety in Nepal

The road safety crisis in Nepal is severe, with about 75 accidents daily and seven fatalities on average per day. During the first six months of fiscal year 2024/25, 1,233 people lost their lives due to road-traffic injuries, thus requiring immediate interventions that are effective. Conventional methods that depend on the manual crash reporting process and use remedial measures have proved to be ineffective in containing the trend.

This proposed AI-based predictive system has a number of advantages over conventional methods:

Proactive Prevention: Instead of reacting to an accident after it has happened, the system pre-locates high-risk zones to allow preventive measures to be applied well in advance to avoid tragedies.

Data-Driven Decision Making: Predictions involve objective analysis of a number of factors, as opposed to subjective assessments or anecdotal evidence, which will support evidence-based policy formulation.

Resource Optimization: The identification of high-risk 20% of the road segments helps focus limited resources in ways that yield the best results.

Dynamic Risk Assessment: The model can be retrained as new data becomes available, reflecting changes in conditions, seasonality, and intervention effectiveness.

Scalability: After validation with real-world data, the system can be scaled up to cover more road networks across Nepal-from urban areas to remote mountain highways.

5.5 Comparison with Regional Studies

The methodological approach and performance metrics of this study compare well with research conducted in similar contexts. Khatiwada et al. (2025) achieved a prediction accuracy of 92.59% for road traffic crash

incidence rates in Nepal using the GM (1,1) model [14], thus showing predictive modeling is feasible even with limited data. Other studies in Ethiopia that utilize Random Forest obtained 82% recall [13], while studies in developing countries using hybrid K-means and Random Forest approaches attained an accuracy rate of 99.86% [12].

However, direct comparisons are difficult due to variations in data quality, sample sizes, and prediction objectives across studies.

5.6 Limitations and Challenges

Several limitations must be acknowledged in the interpretation of findings of this study:

Simulated Data: The most significant limitation is the use of simulated rather than actual historical data. The simulated dataset, while designed to reflect realistic conditions based on known patterns and statistical distributions, cannot fully capture complexity, variability, and unpredictability of real-world traffic and accident dynamics. Real data may include unforeseen patterns, outliers, and interactions not represented in the simulation.

Data Availability: Nepal lacks integrated, centralized databases that incorporate traffic flow, weather conditions, road characteristics, and accident records. Data, if available, is fragmented across multiple agencies like the Nepal Police, Department of Roads, and Department of Hydrology and Meteorology, which are inconsistently formatted and incompletely recorded. Underreporting of accidents, especially minor ones and those occurring in remote areas, further compromises the quality of such data.

Temporal Dynamics: The present model does not consider real-time conditions of traffic flow and dynamic weather conditions. The actual deployment should hence be integrated with the real-time data streams from the traffic sensors, weather stations, and possibly crowdsourced information from traffic and navigation apps, which are presently limited in Nepal.

Model Interpretability: Even though tree-based algorithms offer some sense of interpretability via feature importance metrics, the complex decision paths are often hard to explain to a non-technical audience. This may make acceptance and implementation difficult for policymakers and traffic authorities needing clear, actionable explanations.

Validation Constraints: Without actual accident data for validation purposes, the study cannot conclusively show that the predicted high-risk zones reflect real-world accident hotspots. Future work must include validation of model predictions against historical accident locations and prospective monitoring to see whether predicted high-risk areas indeed experience higher rates of accidents.

Socioeconomic Factors: The model does not consider several socioeconomic variables, such as driver education levels, vehicle maintenance standards, enforcement effectiveness, or even cultural factors related to driving behavior, which may influence the accident risk in Nepal.

5.7 Comparison with Global Best Practices

AI-based traffic safety with real-time monitoring has already been implemented in several developed nations. For instance, ITS in Europe and North America includes sensor networks, automated enforcement, and predictive analytics integrated in one system to reduce accidents. Real-time XGBoost models have achieved results as high as 79% for detection rates using SHAP interpretability [8]. Deep learning methods based on trajectory prediction remain promising using combined architectures such as CNN, LSTM, and GNN.

Given infrastructure constraints, opportunities include mobile data, satellite imagery, and international partnerships:

Mobile data: Navigation apps based on smartphones, such as Google Maps or local versions, create data on traffic flow that might complement traditional sensors.

Satellite images: Remote sensing keeps track of road conditions and recognizes the need for maintenance in remote areas.

Crowdsourcing: Public reporting of hazards and near-misses could enrich the dataset.

International collaboration may include partnerships with development agencies and research institutions for technical expertise and funding.

5.8 Policy Recommendations

Based on the findings and discussion, several policy recommendations emerge:

Establish a Centralized Data Infrastructure: Implement a unified national database that integrates accident records, traffic flow, weather data, and road conditions using uniform reporting standards across agencies.

Pilot Implementation: The predictive model shall be validated and the approach further refined through actual data on pilot studies of high-traffic corridors, such as the Kathmandu-Pokhara highway, before its implementation countrywide.

Capacity Building: Training of traffic police, road engineers, and policy makers in data-driven decision-making and interpretation of AI-based predictions.

Public Engagement: Developing mobile applications or online platforms that offer real-time risk assessment for planned routes, thereby empowering users to take safer travel decisions.

Evidence-Based Interventions: Use model predictions to drive infrastructure investments and direct the improvement of roads, signage, and enforcement efforts toward predicted high-risk zones.

Continuous Monitoring: Establish feedback loops to monitor the effectiveness of interventions; that is, whether safety measures in areas predicted as high-risk reduce accident rates.

Interagency Collaboration: Promote coordination between Nepal Police, the Department of Roads, the Ministry of Home Affairs, and meteorological services to ensure full data sharing and unified action planning.

5.9 Future Research Directions

The following avenues are now open for future research:

Real-World Validation: Apply methodology to real-world accident data from Nepal to validate predictions and further develop models

Real-Time Systems: Develop operational platforms with integrated live data to support dynamic risk assessment. Deep Learning: Explore advanced architecture, including LSTM for temporal patterns, GNN for road network analysis, and attention mechanisms for feature weighting. Causal Analysis: Moving beyond correlation to causal inference, this identifies which interventions will be most effective in reducing accidents. Cost-Benefit Analysis: Quantify the economic impact of predicted interventions versus accident costs (medical expenses, lost productivity, property damage) Behavioral Modeling: Driver patterns and human factors should be integrated into predictive models. Climate Change Impact: Assess how changing weather patterns and increasing extreme events will impact future accident risk. 5.10 Contribution to the Field Notwithstanding its limitations, the present study develops some key contributions. This is the first holistic framework designed only for the country's road safety challenges using AI and machine learning. This study demonstrated that multi-source data-integration of traffic, weather, and road conditions-is feasible in accident prediction for resource-constrained settings. Comparison of algorithms provides practical guidance toward algorithm selection for similar applications in developing countries. Lastly, the proposed visualization approach using heat maps and geospatial mapping will make it easier to present these complex predictions to policymakers and the public. Integrating artificial intelligence into the management of traffic flow in Nepal is about more than an opportunity; it is a dire necessity as the country is continuously developing its road infrastructure and dealing with growing private vehicle ownership and surging volumes of traffic. The way forward lies in continued data collection, inter-agency coordination, capacity building, and evidence-based policymaking. Where these are present, there is little doubt that AI-based predictive systems can save lives and build safer roads for all Nepalis.

5.10 Contribution to the Field

Despite its limitations, this study makes several important contributions. It provides the first comprehensive framework specifically designed for Nepal's road safety challenges using AI and machine learning. It demonstrates the feasibility of integrating multi-source data (traffic, weather, road conditions) for accident

prediction in a resource-constrained context. The comparative evaluation of algorithms offers practical guidance on model selection for similar applications in developing countries. Finally, the visualization approach using heatmaps and geospatial mapping provides an accessible format for communicating complex predictions to policymakers and the public.

As Nepal continues to develop its road infrastructure and cope with increasing vehicle ownership and traffic volumes, the integration of artificial intelligence into traffic management represents not just an opportunity but a necessity. The path forward requires commitment to data collection, inter-agency collaboration, technical capacity building, and evidence-based policymaking. With these elements in place, AI-based predictive systems can play a transformative role in saving lives and building safer roads for all Nepalis.

6. Conclusion

This research work presents the capability of AI and machine learning in predicting high-risk accident zones in Nepal by fusing traffic flow, weather and road conditions, and historical accident data. Through the development and evaluation of different machine learning models-Logistic Regression, Random Forest, and XGBoost-on simulated data representing realistic road conditions in Nepal, this paper has shown that the tree-based algorithms are particularly good at capturing the complex, non-linear relationships between various risk factors contributing to road accidents.

The best performance was achieved by XGBoost with an accuracy of 87%, precision of 85%, recall of 82%, and 0.90 ROC-AUC, closely followed by Random Forest with an accuracy of 85%. These models significantly outperformed the baseline Logistic Regression model, demonstrating their capability in identifying high-risk segments based on multiple interacting variables, including vehicle volume, road curvature, pavement condition, weather patterns, and visibility.

The predictive system was used successfully to categorize the road segments into three levels of risk: low, medium, and high. It allowed for prioritization of safety interventions in resource-constrained environments. Visualization using heatmaps and geospatial mapping showed that authorities can locate accident-prone zones, such as sharp curves, poorly maintained roads, and high-traffic intersections, especially under adverse weather conditions. These insights provide actionable guidance for targeted measures such as improved signage, speed control, road maintenance, and preparedness for emergency responses.

Although this study is based on simulated data because of difficulties in accessing comprehensive real data from Nepal, it lays a sound methodological framework that can be directly applied to actual traffic, weather, and accident records once such data becomes available. The current road safety infrastructure in Nepal relies heavily on manual reporting and reactive measures; these are mostly delayed and cannot accommodate proactive accident prevention.

7. Key Contributions:

Methodological Framework: The integrated AI-based system will be developed for road safety challenges in Nepal using combined data from different sources for comprehensive risk assessment.

Model Comparison: Systematic evaluation of multiple machines learning algorithms, establishing that ensemble methods like Random Forest and XGBoost are most suitable for accident prediction tasks.

Practical Application: The establishment of a risk categorization system, by which authorities can target interventions and invest resources only in priority situations.

Visualization Tools: Integration of geospatial mapping and heatmap visualization to intuitively identify high-risk zones by policymakers and traffic management authorities.

8. Limitations and Future Work:

Although this research gives some insights, it has a few limitations that should be noticed. The first limitation is that the data used are simulated, rather than real historical records, which cannot reflect all complexities of real-world traffic patterns and accident dynamics in Nepal. Besides, real-world applications require real-time data streams and dynamic traffic conditions, yet these are not included in this study.

Future research should focus on several areas of importance:

Real-World Data Integration: In cooperation with the Nepal Police, the Department of Roads, and meteorological departments to access real traffic accident records, road condition surveys, and weather data for model validation and refinement.

Real-Time Prediction System: Development of a live prediction platform that continuously refreshes risk assessments based on current traffic flow, weather conditions, and time-of-day traffic patterns to allow proactive warnings to drivers/authorities.

Mobile Application Development: Develop user-friendly mobile apps that shall be able to deliver real-time, route-specific risk assessments to commuters and identify safer alternatives.

Expanded Feature Set: Addition of more variables like driver behavior pattern, vehicle characteristics, road lighting conditions, and enforcement activity data to enhance the accuracy of the prediction.

Deep Learning Exploration: The investigation of advanced deep learning architectures, such as LSTM networks for temporal pattern recognition, and Graph Neural Networks for road network analysis.

Policy Impact Assessment: Longitudinal studies measuring the effectiveness of AI-guided interventions in reducing accident rates and improving road safety outcomes. Broader Implications: This research contributes to AI applications for road safety in developing countries and supports UN SDGs 3 and 11. In Nepal, where more than 1,200 deaths occur every six months due to road accidents, AI-based predictive systems can significantly reduce this burden through proactive rather than reactive approaches. In conclusion, this study successfully demonstrates that machine learning techniques can effectively predict high-risk accident zones by analyzing multiple contributing factors. The proposed framework provides a foundation for developing operational systems that can transform Nepal's approach to road safety from reactive incident response to proactive risk prevention. Indeed, as Nepal continues to expand its road infrastructure and faces increasing traffic volumes, integrating artificial intelligence with traffic management will be sure to be one of the keys to building safer, more sustainable forms of transportation. Successful implementation of such predictive systems based on adequate data collection infrastructure and inter-agency collaboration holds tremendous potential to save thousands of lives and meaningfully contribute to Nepal's development objectives.

Acknowledgement

We gratefully acknowledge Lalitpur Engineering College (LEC), Safe and Sustainable Travel Nepal (SSTN), and the Society of Transport Engineers Nepal (SOTEN) for providing the opportunity to carry out this research on AI-based road safety in Nepal. Their initiative to promote data-driven approaches to transportation safety created the foundation for this study.

We are grateful to the faculty members and mentors who offered guidance during the development of this work, especially in shaping the methodology and refining the scope of the research. Their insights helped the research team navigate the technical and practical challenges involved in predicting high-risk accident zones.

We also acknowledge the various public data sources in Nepal that make traffic, weather, and road-condition information accessible. Even though the dataset used in this study is limited, these platforms enabled the creation of a realistic framework for demonstrating an AI-based solution.

Finally, we are grateful to our friends and peers who supported us throughout the process, encouraged them to keep going, and assisted with writing and revising this paper.

References:

Abebe, M.T., Sharew, N.T. and Mihretie, G.N., 2025. Machine learning-based prediction of road traffic accident severity in Northwest Ethiopia: model development and comparison. s.l., s.n.

*Achu, A.L., Aju, C.D. and Reghunath, R., 2022. Spatial modeling of road accident hotspots: AHP and Getis-Ord Gi comparison in Thrissur District, India. s.l., s.n.**

Ahmed, S.K., Mohammed, M.G., Abdulqadir, S.O., El-Kader, R.G.A., El-Shall, N.A., Chandran, D., Rehman, M.E.U. and Dhama, K., 2023. Road accident prediction and contributing factors using explainable machine learning models. s.l., s.n.

Berhanu, Y., Alemayehu, E. and Girma, K., 2024. Machine learning approach for predicting road accident severity and suggesting safe routes. s.l., s.n.

Beshah, T. and Hill, S., 2020. Road accident prediction using hybrid K-means and random forest model. s.l., s.n.

He, M., Meng, G., Wu, X., Han, X. and Fan, J., 2025. Road traffic accident prediction based on multi-source data. s.l., s.n.

Khan, M.A. and Hussain, E., 2024. GIS and machine learning for traffic accident prediction in urban environments. s.l., s.n.

Khatiwada, S., Chalise, S., Dhakal, A. and Bhandari, D., 2025. Prediction of road traffic crash incidence rates in Nepal using GM (1,1). s.l., s.n.

Kathmandu Post, 2025. Nepal sees at least 75 road accidents on average daily. s.l., s.n.

Kim, J.K., Ulfarsson, G.F., Kim, S. and Shankar, V.N., 2022. Traffic accident severity prediction based on random forest. s.l., s.n.

Kumar, S. and Gupta, D., 2024. Deep learning-based traffic accident prediction system. s.l., s.n.

Loor-Zambrano, H.Y., García-Pineda, J.L. and Santos-Roldán, L., 2025. Analysis of traffic accident prediction models in Latin America. s.l., s.n.

MyRepublica, 2025. Road accidents claim 1233 lives in six months. s.l., s.n.

Nepal Police, 2025. Annual accidental description. s.l., s.n.

Ospina-Mateus, H., Quintana Jiménez, L.A., Lopez-Valdes, F.J. and Salas-Navarro, K., 2023. Effectiveness of machine learning in forecasting traffic accident severity. s.l., s.n.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A., 2020. Application of XGBoost and SHAP for real-time accident detection. s.l., s.n.

Pourroostaei Ardakani, S., Cheshmehzangi, A. and Chen, H., 2023. Machine-learning-enabled data analysis for road car accidents. s.l., s.n.

Shatnawi, N., 2024. GIS-based traffic accident hotspot prediction using machine learning. s.l., s.n.

Wu, P., Meng, X., Song, L. and Zhao, W., 2021. Application of XGBoost in PM2.5 prediction: case study of Shanghai. s.l., s.n.

Yang, C., Chen, M., Yuan, Q. and Fu, R., 2023. Prediction of traffic accident severity using random forest. s.l., s.n.

Zhang, X., Yao, H., Hu, G., Zhao, M. and Yang, R., 2024. Traffic accident forecasting in intelligent transportation systems. s.l., s.n.

Zhang, Y., Li, Q., Wang, Z. and Liu, B., 2024. Deep learning model for traffic accident risk prediction based on trajectory data. s.l., s.n.