Jordan Journal of Civil Engineering, Volume **20**, No. **4**, **2026**
DOI: https://doi.org/10.14525/JJCE.v20i4.10

Jordan Journal of Civil Engineering

Journal homepage: https://jjce.just.edu.jo

Comparison of Weight of Evidence and Random Forest for Traffic Crash Mapping Using GIS and Remote Sensing

Ha Le Thi [1]***, Khanh Giang Le*** [2] ****, Thi Thao Tran*** [2]

[1] *Campus in Ho Chi Minh City, University of Transport and Communications, Ho Chi Minh City, Vietnam.*
[2] *Faculty of Civil Engineering, University of Transport and Communications, Hanoi, Vietnam.*
**Corresponding Author. Email: giangl k@utc.edu.vn.*

ARTICLE INFO

Article History:
Received: 8/5/2026
Accepted: 21/7/2026

ABSTRACT

Road traffic crashes pose a significant challenge to socio-economic development, particularly in developing countries, where rapid urbanization and increasing traffic demand intensify safety risks. This study evaluates and compares the performance of the Weight of Evidence (WoE) and Random Forest (RF) models for traffic crash susceptibility mapping using Geographic Information System (GIS), remote sensing, and multi-source spatial data in Thanh Hoa Province, Vietnam. Sentinel-2 remote sensing imagery was used to derive land use/land cover (LULC) information, while Digital Elevation Model (DEM), road network, and 1,183 accident points (2020-2023) were processed to construct key spatial explanatory variables, including terrain slope, road gradient, intersection density, and crash density. Accident density was incorporated as a spatial explanatory variable to represent historical accident concentration patterns within the study area. The dataset was divided into training (80%) and validation (20%) subsets, and model performance was evaluated using ROC curves, AUC, Accuracy, and F1-score. The results indicate that both models effectively capture spatial crash patterns; however, the RF model outperformed the WoE model, achieving an AUC of 0.93 and an Accuracy of 0.88, compared with an AUC of 0.87 and an Accuracy of 0.84 for the WoE model. High- and very-high-susceptibility road segments account for only 14.3% of the total road network, but contain 83.1% of recorded crashes, indicating a strong spatial concentration of crash occurrences. The resulting maps primarily represent spatial accident susceptibility based on historical accident patterns and environmental conditions. These findings demonstrate that integrating GIS with remotely sensed LULC information, topographic variables, road-network characteristics, historical crash density, and machine learning provides a reliable and effective framework for traffic crash susceptibility assessment, thereby supporting traffic safety management and transportation infrastructure planning.

Keywords: Traffic crashes, Risk mapping, Weight of evidence, Random forest, GIS, Remote sensing.

INTRODUCTION

Road traffic crashes remain one of the leading causes of death and injury worldwide, accounting for approximately 1.3 million fatalities each year and imposing substantial social and economic burdens on society (WHO, 2023). The problem is particularly severe in developing countries, where rapid urbanization, increasing vehicle ownership, and limitations in transportation infrastructure and traffic safety management contribute to elevated crash risks. In Vietnam, the rapid growth of motorized vehicles, especially motorcycles, has significantly increased traffic demand and crash exposure, making road safety a critical challenge for transportation authorities and policymakers.

In recent years, Geographic Information Systems (GIS)s and remote sensing technologies have emerged as powerful tools for traffic safety analysis and spatial risk assessment. These technologies facilitate the integration of diverse spatial datasets, including land use, terrain, transportation networks, and crash records, enabling the identification of accident-prone locations and high-risk zones. GIS-based approaches provide a spatially explicit framework for understanding crash patterns and supporting transportation planning, infrastructure management, and safety improvement strategies (Deressa et al., 2025). Building upon these GIS capabilities, numerous statistical and computational approaches have been developed to model crash occurrence, identify high-risk locations, and improve traffic safety assessment.

Traditionally, traffic crash analysis has relied on statistical approaches, such as logistic regression, Poisson regression, Negative Binomial models, and generalized linear models, to investigate the relationships between crash occurrence and contributing factors (Lord & Mannering, 2010; Savolainen et alk., 2011). These methods have provided valuable insights into the mechanisms underlying traffic crashes and have been widely applied in safety studies. However, conventional statistical models often require predefined functional relationships between explanatory variables and crash outcomes. Consequently, their ability to capture complex nonlinear interactions among environmental, roadway, and traffic-related factors may be limited, particularly when large and heterogeneous spatial datasets are involved (Chen et al., 2016).

To address these challenges, researchers have developed a variety of advanced methodological frameworks, including spatial analysis techniques, econometric models, and heterogeneity-based modeling approaches. Econometric methods, such as Poisson, Negative Binomial, zero-inflated, and random-parameter models, have been extensively used to analyze crash frequency and injury severity while accounting for overdispersion, spatial dependence, and unobserved heterogeneity in crash data (Lord & Mannering, 2010; Mannering et al., 2016). These approaches have substantially improved the understanding of crash-generating mechanisms and the factors influencing crash occurrence. Nevertheless, despite their strong explanatory capabilities, such models may still encounter difficulties in representing highly complex nonlinear relationships and interactions among multiple spatial variables. Alongside these advances in statistical and econometric modeling, machine learning approaches have emerged as powerful alternatives for capturing complex nonlinear crash patterns from large and heterogeneous datasets.

More recently, with the increasing availability of large-scale spatial datasets, machine learning techniques have gained increasing attention in transportation safety research due to their capability to learn complex patterns directly from data without requiring strict assumptions regarding model structure. These approaches have demonstrated considerable potential for handling large datasets, modeling nonlinear relationships, and improving predictive performance (Le, 2025; Le, 2026). Among various machine learning algorithms, Random Forest (RF) has become one of the most widely adopted techniques because of its ensemble-learning structure, robustness against overfitting, ability to process high-dimensional datasets, and capability to estimate the relative importance of explanatory variables. Previous studies have reported that RF-based models can effectively support traffic crash susceptibility assessment, hotspot identification, and traffic safety analysis (Farhangi et al., 2021; Berhanu et al., 2024).

Besides machine learning approaches, probabilistic spatial susceptibility mapping methods have also been increasingly adopted for traffic safety and hazard assessment. Among them, the Weight of Evidence (WoE) method has been widely applied in spatial susceptibility and risk assessment studies. Based on Bayesian probability theory, WoE quantifies the relationship between conditioning factors and the occurrence of a target phenomenon by assigning positive and negative weights to different variable classes. Due to its simplicity, transparency, and compatibility with GIS environments, WoE has been successfully applied in various hazard and susceptibility mapping applications, including landslide, flood, and environmental risk assessments (Bonham-Carter, 1994; Agterberg & Bonham-Carter, 1990). Unlike many machine learning approaches that often operate as black-box models (Wen et al., 2021), the WoE method explicitly assigns statistical weights to each conditioning-factor class, making the influence of each conditioning factor on crash susceptibility explicitly quantifiable and easier to interpret (Gadtaula & Subodh, 2019).

Despite these methodological advances, several research gaps remain. First, many previous studies have focused on the application of a single statistical, econometric, or machine learning technique, making it difficult to objectively evaluate the relative strengths and limitations of different approaches under comparable experimental conditions. Second, relatively few published studies have systematically compared the WoE model and the RF model within a unified GIS-based framework for traffic crash susceptibility mapping. Such comparisons are important, because WoE provides explicit and interpretable statistical relationships between conditioning factors and crash occurrence, whereas RF is capable of capturing complex nonlinear interactions among multiple variables. Third, although GIS, remote sensing, and machine learning techniques have each been widely applied in traffic safety studies, relatively few studies have integrated remotely sensed land-use information, road-network characteristics, topographic variables, and historical crash data within a unified GIS-based traffic crash susceptibility assessment framework. Furthermore, evidence regarding the trade-off between model interpretability and predictive capability under identical datasets, predictor variables, spatial resolution, and validation procedures remains limited.

Addressing these gaps is essential for improving the understanding of the relative strengths and limitations of different spatial modeling approaches and for selecting appropriate methods that balance predictive performance, model interpretability, and practical applicability in GIS-based traffic crash susceptibility mapping. Therefore, this study aims to:

(i) develop traffic crash susceptibility maps by integrating GIS, remote sensing, and road-network-derived variables;
(ii) compare the predictive performance and spatial prediction characteristics of WoE and RF models under the same experimental framework; and
(iii) identify high-risk road segments to support traffic safety management and infrastructure planning.

The primary methodological contribution of this study lies in establishing a unified GIS-based experimental framework for objectively comparing an interpretable probabilistic model (WoE) and a nonlinear machine learning model (RF). Unlike many previous studies that employed only a single modeling approach, both models were developed, trained, and evaluated using identical predictor variables, spatial resolution, training and validation datasets, and validation procedures. By adopting the same predictor variables, spatial resolution, training and validation datasets, and evaluation metrics for both models, the proposed framework ensures a fair and objective comparison of predictive accuracy, spatial prediction characteristics, and model interpretability. This experimental design minimizes methodological bias arising from inconsistent model settings and provides a more reliable assessment of the trade-off between interpretability and predictive capability in traffic crash susceptibility mapping.

In addition, the proposed framework demonstrates the feasibility of integrating multi-source spatial information, including remotely sensed land-use data, terrain characteristics, road-network variables, and historical traffic crash records, into a reproducible GIS-based crash susceptibility assessment workflow. This integration provides a practical methodology that can be readily adapted to other regions where comprehensive traffic exposure data is limited.

The findings provide empirical evidence regarding the comparative performance of statistical and machine learning approaches for GIS-based traffic crash susceptibility mapping in Thanh Hoa Province, Vietnam. The resulting susceptibility maps provide spatially explicit information that may support traffic safety management, transportation planning, and infrastructure investment prioritization. More broadly, this study contributes to the growing body of GIS-based traffic safety research by providing a reproducible comparative framework that integrates remote sensing, GIS, and machine learning techniques for traffic crash susceptibility assessment in developing-country contexts.

The remainder of this paper is organized as follows. The next section describes the study area, datasets, and methodology, followed by presenting the results. Then, the findings are discussed, and the last section concludes the study.

DATA AND METHODS

Study Area

Thanh Hoa Province is located in North Central Vietnam and plays an important role in regional transportation (Fig. 1). The province extends from

19°18′ to 20°40′ North latitude and from 104°22′ to 106°04′ East longitude (Ha et al., 2024). The province has diverse topography, including coastal plains and mountainous areas, together with a rapidly developing road network. Increasing traffic demand and variations in infrastructure conditions have contributed to elevated traffic accident risk in the region.

Figure 1. Study area in Thanh Hoa Province, Vietnam

Data Sources

To develop the traffic accident risk model, multiple spatial datasets were used, including traffic accident records, Sentinel-2 satellite imagery, DEM data, and road network data. All spatial datasets were transformed into the WGS 84 / UTM Zone 48N coordinate system prior to analysis. To ensure spatial consistency among datasets from different sources, all raster layers were standardized to a common spatial resolution of 30 m. Sentinel-2 imagery originally acquired at 10 m resolution was resampled to 30 m, while DEM-derived variables retained their original 30 m resolution. Vector-based road network variables were subsequently rasterized to the same grid size and spatial extent before model implementation.

The data sources used in this study are summarized in Table 1.

Table 1. Data sources used in this study

Data	Source	Description
Traffic crashes	Thanh Hoa Traffic Police	1183 points (2020-2023)
Sentinel-2	ESA Copernicus / Google Earth Engine	Multispectral imagery, 10 m resolution
DEM (SRTM)	NASA / USGS	Elevation data, 30 m resolution
Road network	National geospatial database/ OpenStreetMap	Vector data

A total of 1183 accident points and 1183 non-accident points were used to construct the binary dataset. Each accident record included geographic coordinates, date and time of occurrence, road class, lighting condition, weather condition, vehicle type, victim age, and casualty information, such as fatalities and injuries. Prior to analysis, duplicate records, incomplete entries, and records with inconsistent spatial information were removed. Non-accident points were randomly generated along the road network using the Create Random Points tool in ArcGIS Pro. To reduce potential spatial overlap and local spatial autocorrelation, the Near tool was used to calculate the distance between accident and non-accident points, and non-accident points located too close to accident locations were excluded. The final dataset was randomly divided into 80% for training and 20% for validation.

Sentinel-2 imagery was used to derive the land use/land cover (LULC) map, while DEM data was used to extract terrain-related variables, such as slope. Road network data was utilized to derive road gradient and intersection density. The LULC classification achieved an Overall Accuracy of 90.3% and a Kappa coefficient of 0.71, indicating satisfactory classification quality for subsequent spatial analysis. All processed spatial layers were subsequently used as input variables for the WoE and RF models.

Following data standardization, geometry validation, duplicate removal, and visual inspection of spatial alignment were conducted to ensure positional consistency among all datasets. Accident locations were also visually cross-checked against the road network to verify that each accident point was correctly aligned with the corresponding road segment before extracting explanatory variables. Although the datasets originated from different sources, the environmental and infrastructure variables (e.g., DEM, road network, and LULC) represent relatively stable characteristics over the accident observation period (2020-2023), thereby minimizing potential temporal inconsistencies during spatial analysis. These preprocessing and quality-control procedures were implemented to minimize uncertainties associated with multi-source data integration and to improve the reliability, robustness, and reproducibility of the subsequent traffic accident susceptibility analysis.

METHODOLOGY

Research Workflow

The overall workflow for traffic crash risk mapping consists of four main steps: data preparation, variable extraction, model development, and model evaluation (Fig. 2).

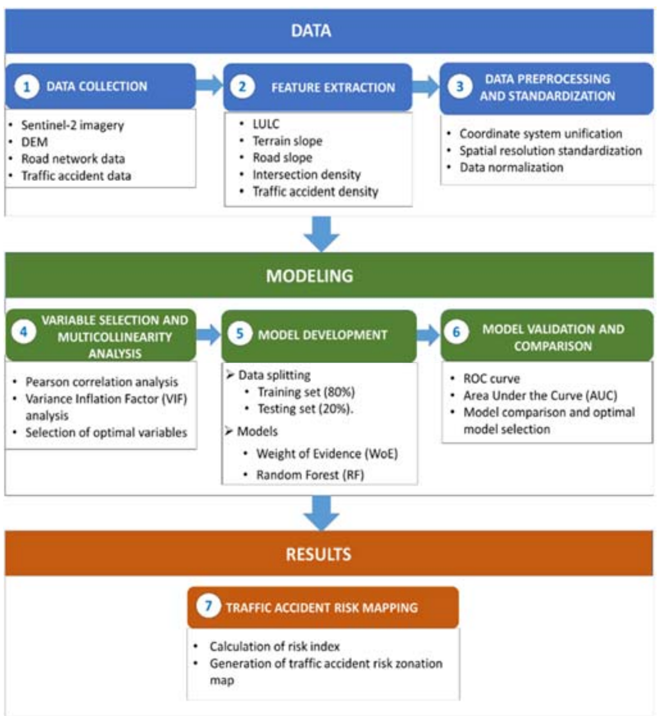

Figure 2. Research workflow

First, multi-source datasets, including Sentinel-2 imagery, DEM, road network data, and traffic crash records, were collected and preprocessed. Input variables were then derived within the GIS environment, including LULC, terrain slope, road gradient, intersection density, and crash density. All datasets were standardized to ensure consistency in coordinate systems, spatial resolution, and data format.

Next, explanatory variable selection and validation were conducted to ensure the scientific relevance and reliability of the input factors. Pearson correlation analysis and Variance Inflation Factor (VIF) analysis were applied to reduce redundancy and assess multicollinearity among variables prior to model implementation.

After preparing the dataset, two models, WoE and RF, were implemented. The dataset was divided into 80% for training and 20% for validation. Model performance was evaluated using the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). Finally, the best-performing model was used to generate the traffic accident risk map. Spatial distribution of accident risk was further analyzed across the road network.

Data processing and analysis were conducted using Google Earth Engine and ArcGIS Pro. The WoE model was implemented in ArcGIS Pro, while the RF model was developed using programming-based approaches.

Input Variable Construction

The explanatory variables were selected according to four criteria: (i) documented relationships with traffic accident occurrence reported in previous studies; (ii) physical relevance to the mechanisms influencing traffic accident susceptibility; (iii) availability, spatial completeness, and consistency of reliable datasets covering Thanh Hoa Province; and (iv) suitability for GIS-based spatial analysis at a common spatial resolution. Based on these criteria, five explanatory variables were selected to represent complementary dimensions influencing traffic accident occurrence, including land use conditions (LULC), terrain characteristics (terrain slope and road gradient), road network characteristics (intersection density), and historical traffic exposure patterns (crash density) (Fig. 3). The statistical suitability of the selected variables was subsequently evaluated through Pearson correlation analysis and Variance Inflation Factor (VIF) analysis, as described later.

LULC was derived from Sentinel-2 imagery using a Random Forest classification algorithm supported by the spectral indices NDVI, NDBI, and NDWI to improve classification accuracy. The resulting LULC map was used to represent spatial urbanization and land-use characteristics associated with traffic accident susceptibility. Reference samples were collected from the ESA WorldCover 2020 dataset using stratified random sampling for five LULC classes, including forest, agricultural land, built-up land, water, and bare land. The dataset was randomly divided into 70% for training and 30% for validation. Classification accuracy was evaluated using a confusion matrix, Overall Accuracy (OA), and the Kappa coefficient. The resulting classification achieved an Overall Accuracy of 90.3% and a Kappa coefficient of 0.71, indicating satisfactory classification quality for subsequent spatial analysis. Terrain slope was derived from the 30 m SRTM DEM. Road gradient was calculated by integrating the DEM with the road network and computing elevation differences along individual road segments within the GIS environment. Intersection density was calculated from the spatial distribution of road intersections extracted from the road network.

To minimize potential information leakage during model development, the accident dataset was first randomly divided into training (80%) and validation (20%) subsets. Kernel Density Estimation (KDE) was then performed exclusively using the training accident samples to generate the accident density layer representing the spatial concentration of historical traffic accidents. Validation accident samples were not involved in the KDE generation process. The resulting accident density layer was subsequently integrated with the remaining explanatory variables and used for model development and validation.

The KDE analysis was performed in ArcGIS Pro using a spatial resolution of 30 m and the default quartic kernel function. The bandwidth parameter was set to 10,250 m based on a previous Incremental Spatial Autocorrelation analysis conducted in Thanh Hoa Province, where this distance corresponded to the strongest spatial clustering tendency of traffic accidents (Ha et al., 2024). Accident density was incorporated as an explanatory variable to characterize historical spatial clustering patterns of traffic accidents. This procedure ensured that the accident density variable represented historical spatial information without introducing information from the independent validation dataset.

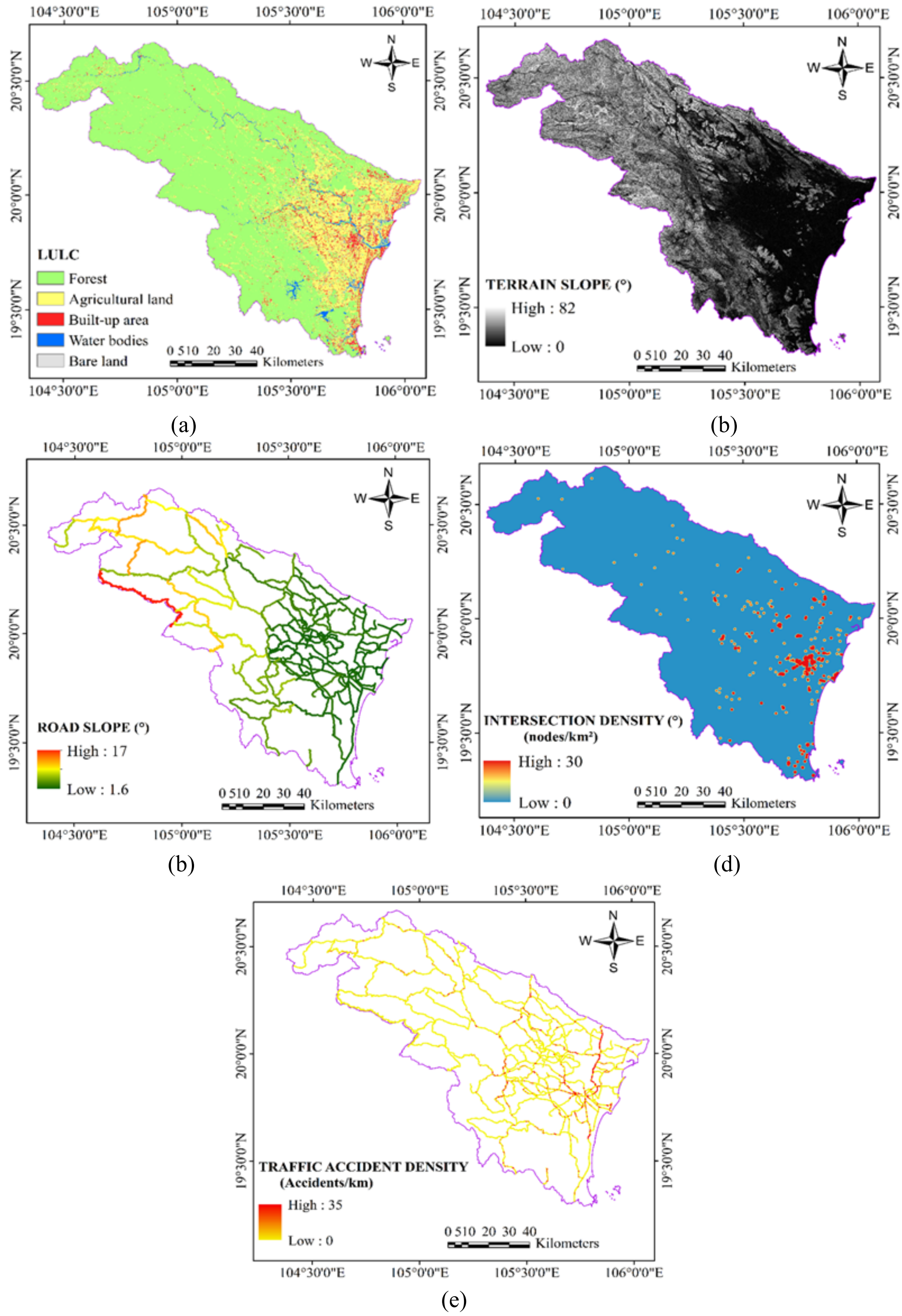

Figure 3. Environmental factors; (a) LULC, (b) Terrain slope, (c) Road gradient, (d) Intersection density, (e) Traffic crash density

Variable Validation

The explanatory variables were further evaluated using Pearson correlation analysis and VIF analysis to assess redundancy and multicollinearity prior to model implementation.

Pearson Correlation Analysis

The linear relationships among input variables were evaluated using the Pearson correlation coefficient (r), which ranges from −1 to 1. Values close to ±1 indicate strong linear relationships, whereas values near 0 indicate weak or no correlation.

Variables with an absolute correlation coefficient ($|r| \geq 0.8$) were considered highly correlated, and one of the correlated variables was removed to reduce redundancy (Dou et al., 2023). This step helps mitigate multicollinearity, reduce overfitting, and improve model generalization.

Multicollinearity Test

Multicollinearity among input variables was further assessed using the VIF. A VIF value of 1 indicates no multicollinearity, values between 1 and 5 indicate low to moderate multicollinearity, and values greater than 10 indicate severe multicollinearity (Abdulhafedh, 2022; Moyo & Dzinavatonga, 2025).

This analysis ensures the independence of input variables and enhances model stability and reliability.

Weight of Evidence

The WoE model is a Bayesian-based statistical approach used to quantify the relationship between crash occurrence and environmental factors.

For each class i, positive and negative weights are calculated as:

$$W_i^+ = \ln((P(B \mid D))/(P(B \mid \bar{D}))) \quad (1)$$

$$W_i^- = \ln((P(\bar{B} \mid D))/(P(\bar{B} \mid \bar{D}))) \quad (2)$$

where P is probability; B and $\bar{B}$ represent crash and non-crash points; D and $\bar{D}$ denote total crash and non-crash points in the study area.

The contrast value is computed as:

$$C_i = W_i^+ - W_i^- \quad (3)$$

A higher C_i indicates a stronger influence of the factor on crash occurrence.

Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using random subsets of training samples and predictor variables. Each tree generates an individual prediction, and the final output is obtained by aggregating the predictions from all trees through a voting process. This ensemble learning strategy helps reduce overfitting and improves model robustness by combining the outputs of multiple decision trees.

In this study, the RF model was implemented in Google Earth Engine using the smileRandomForest classifier with 300 decision trees and five predictor variables. The minimum leaf population and bag fraction were set to 1 and 0.5, respectively. The number of variables considered at each split followed the default setting of the smileRandomForest classifier, which automatically selects the square root of the total number of predictor variables. No restriction was imposed on the maximum number of tree nodes. The RF model was trained and validated using the training and validation dataset previously described. Since the dataset was balanced between accident and non-accident samples, class weighting was not applied. Extensive hyperparameter tuning was not conducted, because the primary objective of this study was to provide a consistent comparative evaluation of the WoE and RF models under the same experimental framework rather than to optimize predictive performance.

The model output can be expressed as:

$$Z(x) = 1/N \sum Y_i\,(x) \quad (4)$$

where $Z(x)$ is the predicted output, $Y_i\,(x)$ is the prediction of the i-th tree, and N is the number of trees.

Model Evaluation

Model performance was evaluated using the ROC curve and the AUC. A higher AUC indicates better classification performance.

The classification results were categorized into four groups: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The ROC curve represents the relationship between the true positive rate (TPR) and false positive rate (FPR):

$$TPR = TP/(TP + FN) \quad (5)$$

$$FPR = FP/(FP + TN) \quad (6)$$

In addition, Accuracy and F1-score were used to evaluate classification performance, particularly for imbalanced datasets.

RESULTS

Variable Selection and Validation

Pearson Correlation Analysis

Pearson correlation analysis was conducted to evaluate the linear relationships among continuous input variables. The LULC variable was excluded from this analysis, because it is categorical.

The correlation matrix is presented in Table 2. The correlation coefficients range from −0.312 to 0.781, indicating low to moderate relationships among most variable pairs. The highest correlation was observed between road gradient and terrain slope (r = 0.781), suggesting a relatively strong relationship between these two variables.

According to previous studies, variables with an absolute correlation coefficient ($|r| \geq 0.8$) are considered highly correlated and should be carefully examined to avoid redundancy (Farhangi et al., 2021). In this study, no variable pair exceeded this threshold, indicating that all variables can be retained for subsequent multicollinearity analysis using VIF.

Table 2. Pearson correlation matrix of input variables

Variables	Crash density	Intersection density	Road gradient	Terrain slope
Crash density	1			
Intersection density	0.182	1		
Road gradient	-0.312	-0.221	1	
Terrain slope	-0.285	-0.183	0.781	1

Multicollinearity Test

Multicollinearity among input variables was evaluated using the VIF computed with the Generalized Linear Regression (GLR) tool in ArcGIS Pro. A VIF value of less than 5 indicates no significant multicollinearity.

The results (Table 3) show that VIF values range from 1.08 to 2.68. All variables have VIF < 5, indicating that multicollinearity is not significant and all variables are suitable for model development.

Table 3. VIF results of input variables

No.	Variables	VIF
1	LULC	1.13
2	Crash density	1.14
3	Intersection density	1.08
4	Road gradient	2.68
5	Terrain slope	2.60

WoE Model Results

In the WoE model, input variables were classified into intervals to calculate the probability of crash occurrence within each class. The positive weight (W^{+}), negative weight (W^{-}), and contrast (C) values were computed based on Bayesian probability (Table 4).

Table 4. WoE weights of influencing factors

Factor	Class interval	W^{+}	W^{-}	C
	0 - 1	-2.506	1.695	-4.201
	1 - 3	1.188	-0.361	1.549
Accident density (accidents/km)	3 - 6	2.149	-0.320	2.469
	6 - 12	2.479	-0.194	2.673
	12 - 35	2.927	-0.058	2.985
	0 - 1	-0.205	0.850	-1.055
Intersection density (nodes/km²)	1 - 4	0.637	-0.088	0.725
	4 - 9	1.276	-0.070	1.346

	9 - 16	1.749	-0.020	1.770
	16 - 30	0	0	0
Road gradient (°)	1.6 - 2.7	0.561	-0.438	0.999
	2.7 - 4.5	-0.305	0.082	-0.387
	4.5 - 7.9	-0.110	0.028	-0.138
	7.9 - 12.0	-1.442	0.150	-1.592
	12.0 - 17.0	-0.606	0.021	-0.628
Terrain slope (°)	0 - 8	0.281	-0.910	1.191
	8 - 18	-0.641	0.098	-0.739
	18 - 27	-1.145	0.086	-1.231
	27 - 37	-1.653	0.041	-1.694
	37 - 82	-1.946	0.006	-1.952
LULC	Forest	-0.829	0.829	-1.658
	Agricultural land	-0.338	0.338	-0.677
	Built-up land	0.494	-0.494	0.988
	Water	-1.099	1.099	-2.197
	Bare land	0.326	-0.326	0.652

For accident density, the lowest density class (0-1 accidents/km) exhibits a strongly negative contrast value (C = −4.201), whereas all higher-density classes (1-35 accidents/km) show positive contrast values, with the contrast increasing progressively from 1.549 to 2.985. This trend indicates that road segments with a greater historical concentration of traffic accidents are associated with higher accident susceptibility. The increasing contrast values further suggest that historical accident density is an important indicator of future accident occurrence, reflecting the spatial clustering characteristics of traffic crashes.

Similarly, intersection density classes from 1-16 nodes/km² exhibit positive contrast values, with the highest contrast observed in the 9-16 nodes/km² class (C=1.770). This finding suggests that accident susceptibility increases in areas with dense road intersections, where multiple traffic streams converge and vehicle conflict points become more frequent. In contrast, the lowest intersection-density class (0-1 nodes/km²) exhibits a negative contrast value (C = −1.055), indicating relatively lower accident susceptibility.

For road gradient, only the 1.6-2.7° class shows a positive contrast value (C = 0.999), whereas steeper gradient classes exhibit negative contrast values. This result suggests that traffic accidents in the study area occur more frequently along road segments with relatively gentle gradients, which generally experience higher traffic volumes than mountainous roads with steep gradients.

Terrain slope analysis shows a similar pattern. The relatively flat terrain class (0-8°) exhibits a positive contrast value (C = 1.191), while progressively steeper terrain classes show increasingly negative contrast values. This finding suggests that accidents are more likely to occur in low-relief areas characterized by denser transportation networks, greater urban development, and higher traffic exposure, whereas mountainous regions generally experience fewer accidents because of lower traffic demand.

Regarding LULC, built-up land (C = 0.988) and bare land (C = 0.652) exhibit positive contrast values, indicating relatively higher accident susceptibility. In contrast, forest, agricultural land, and water bodies exhibit negative contrast values. The positive association between built-up areas and accident occurrence is likely related to intensive transportation activities, higher traffic demand, and more frequent interactions among road users within urban environments.

Overall, the WoE analysis demonstrates that historical accident density, intersection density, terrain characteristics, and land use all contribute to the spatial distribution of traffic accident susceptibility in Thanh Hoa Province. Positive contrast values associated with accident density, dense intersections, and built-up land indicate that traffic accidents tend to concentrate in highly urbanized transportation corridors with intensive traffic activities. Conversely, mountainous areas and steep terrain generally exhibit lower accident susceptibility, reflecting reduced traffic exposure and fewer transportation interactions. More importantly, the WoE model explicitly quantifies the contribution of each conditioning-factor class through statistical weights, providing an interpretable framework for understanding how individual spatial variables influence traffic accident susceptibility and supporting evidence-based traffic safety planning.

Traffic Crash Risk Map Based on the WoE Model

The traffic crash risk index was calculated by integrating WoE weights of all input factors within the GIS environment. The results are presented as a traffic crash risk map (Fig. 4).

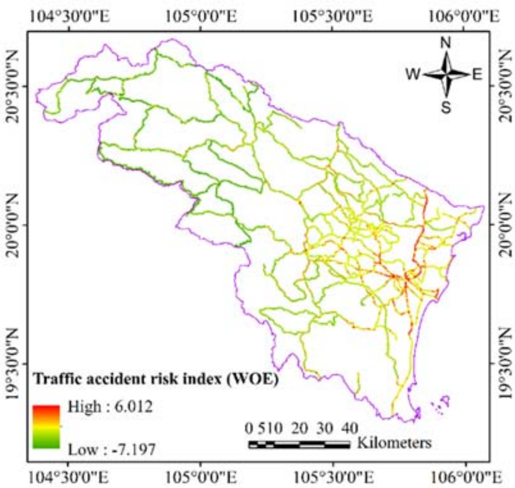

Figure 4. Traffic crash risk map (WoE)

The WoE-based risk map reveals clear spatial variability in crash risk across the study area. High-risk zones are predominantly concentrated in the central and coastal plain regions of Thanh Hoa Province, where road networks are dense and major transportation corridors are located. In contrast, the western mountainous areas exhibit low-risk levels, reflecting lower traffic density and more complex terrain conditions.

Overall, crash risk tends to cluster in areas with intensive traffic activities, particularly along major roads and urbanized regions. These findings indicate that the WoE model effectively captures the spatial heterogeneity of traffic crash risk.

Random Forest Model Results

The RF model was developed using five explanatory variables, including accident density, terrain slope, road gradient, LULC, and intersection density. The relative importance of each variable, calculated based on the RF model, is presented in Table 5.

Table 5. Variable importance in the random forest model

No.	Variables	Importance (%)
1	Crash density	42.7
2	Terrain slope	21.1
3	Road gradient	20.1
4	LULC	8.2
5	Intersection density	7.9

Among the five explanatory variables, accident density exhibits the highest importance (42.7%), indicating that the spatial distribution of historical traffic accidents is the strongest predictor of future accident susceptibility within the study area. This result suggests that locations with a history of frequent accidents tend to remain accident-prone, because they often reflect persistent roadway characteristics, traffic demand, geometric design, and operational conditions.

Terrain slope (21.1%) and road gradient (20.1%) are the second and third most influential variables, respectively. Their relatively high importance indicates that topographic conditions substantially influence traffic accident susceptibility in Thanh Hoa Province. Steeper terrain and greater road gradients may affect vehicle operating speeds, braking performance, driver visibility, and maneuverability, thereby increasing the complexity of traffic conditions. These findings highlight the importance of considering terrain-related factors when assessing spatial traffic accident susceptibility, particularly in regions with diverse topographic characteristics.

In comparison, LULC (8.2%) and intersection density (7.9%) contribute less to the overall prediction. Nevertheless, these variables remain relevant, because they characterize the surrounding built environment and road-network configuration. Different land-use types influence traffic activity and travel demand, whereas intersection density reflects potential traffic conflict points where vehicle interactions are more frequent.

Based on the trained RF model, a continuous traffic accident susceptibility map was generated (Fig. 5).

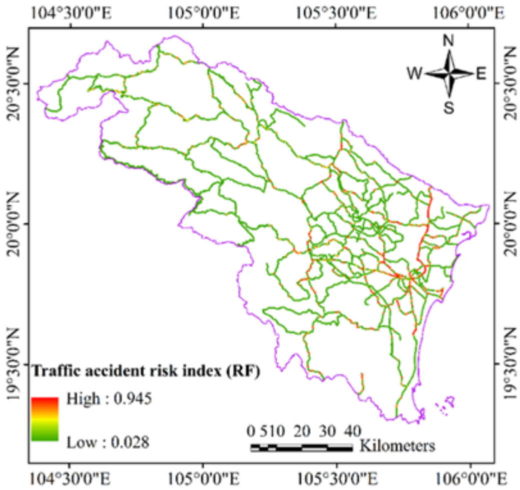

Figure 5. Traffic accident susceptibility map generated using the RF model

The RF-derived susceptibility map indicates that high-risk areas are primarily concentrated along major transportation corridors and in the central part of Thanh Hoa Province, where road networks are denser and traffic activities are more intensive. Conversely, low-susceptibility areas are mainly distributed in the western mountainous region, which is characterized by lower road density and relatively limited traffic activity. These spatial patterns are consistent with the distribution of historical accident records and indicate that the RF model effectively captures the complex spatial variability of traffic accident susceptibility.

Overall, the RF model demonstrates a strong capability to model nonlinear relationships and interactions among multiple spatial variables. The dominance of accident density, terrain slope, and road gradient indicates that both historical traffic accident concentration and topographic conditions play critical roles in explaining accident susceptibility. Compared with conventional statistical approaches, the RF model can simultaneously account for complex interactions among predictor variables without requiring predefined functional relationships, making it particularly suitable for GIS-based traffic accident susceptibility assessment.

Model Performance Comparison

The performance of the WoE and RF models was evaluated using confusion matrix-based metrics and ROC analysis. Table 6 summarizes the comparative performance of the two models. The RF model achieved higher Accuracy, Recall, F1-score, Kappa, and AUC values than the WoE model, indicating better predictive capability and stronger discrimination between accident and non-accident areas.

Table 6. Performance comparison of the WoE and RF models

Metric	WoE	RF
Accuracy	0.84	0.88
Precision	0.88	0.86
Recall	0.79	0.91
Specificity	0.90	0.85
F1-score	0.84	0.89
Kappa	0.69	0.76
AUC	0.87	0.93

The ROC curves (Fig. 6) further confirm the superior discriminative performance of the RF model compared to the WoE model.

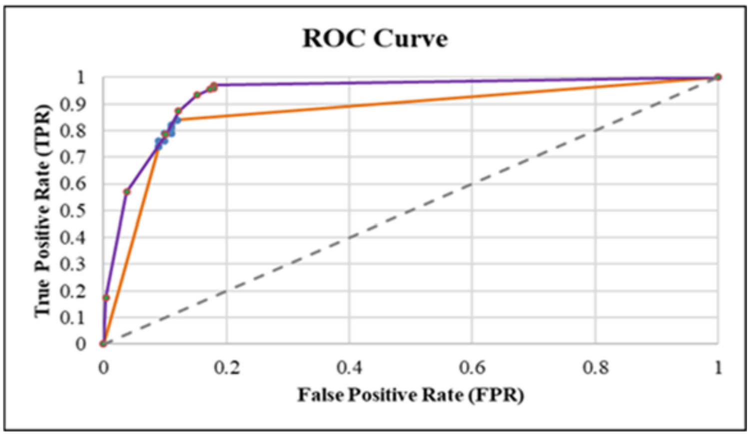

Figure 6. ROC comparison of the WoE and RF models

Overall, the RF model demonstrated better capability in identifying traffic accident risk and was therefore selected for traffic accident risk mapping in the study area. crash and non- crash areas.

Traffic Crash Risk Zoning Based on the Random Forest Model

Based on the RF model results, the traffic crash risk map was classified into five levels: very low, low, moderate, high, and very high (Fig. 7).

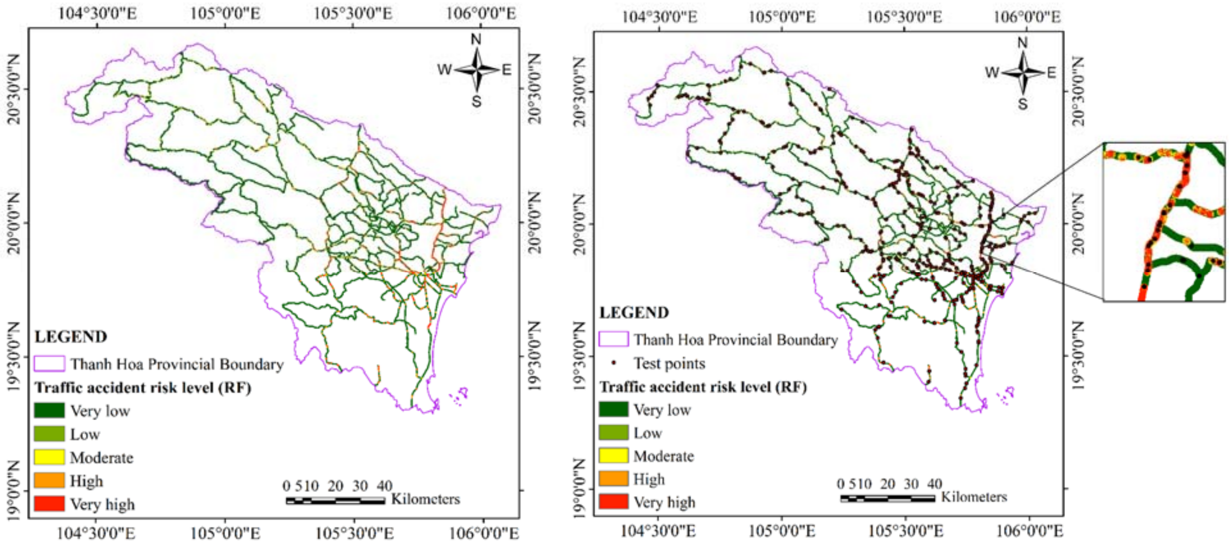

Figure 7. Traffic crash risk zoning based on the random forest model; (a) Spatial distribution of traffic crash risk; (b) Detailed view of high- and very high-risk segments with validation points

High- and very high-risk road segments are mainly concentrated along major highways in Thanh Hoa Province, particularly National Highways 1A, 10, 45, and 47, where traffic density and intersection frequency are high. In contrast, roads in the western mountainous areas, such as National Highways 15, 15C, and 217, are predominantly classified as low and very low risk.

The relationship between road length and crash proportion by risk level is shown in Table 7.

Table 7. Distribution of road length and crash proportion by risk level

Risk level	Road length (%)	Crash (%)
1 (Very low)	77.2	3.4
2 (Low)	3.2	5.5
3 (Moderate)	5.3	8.0
4 (High)	5.4	24.5
5 (Very high)	8.9	58.6

Very low-risk roads account for 77.2% of the total road length, but only for 3.4% of crashes. In contrast, high- and very high-risk roads represent only 14.3% of the network, but account for 83.1% of total crashes.

These results indicate a strong concentration of crashes in high-risk segments, demonstrating that the RF-based risk zoning effectively reflects the spatial distribution of traffic crashes and accurately identifies high-risk areas.

DISCUSSION

The results clearly indicate that the RF model outperforms the WoE model in predicting traffic crash risk, as reflected by higher Accuracy, F1-score, and AUC values. This finding is consistent with previous studies showing that tree-based machine learning algorithms generally provide superior predictive performance compared to conventional statistical approaches, because they can effectively capture complex nonlinear relationships and interactions among spatial variables (Farhang et al., 2021; Khan & Hussain, 2024; Zhang et al., 2022).

The variable importance analysis reveals that crash density is the most influential factor (42.7%) in the RF model. This reflects the inherent spatial clustering characteristics of traffic crashes, where events tend to concentrate in specific locations, commonly referred to as hotspots in GIS-based studies (Lan & Hoc, 2024; Le et al., 2022; Le et al., 2019). Although accident density was generated exclusively from the training dataset to minimize potential information leakage during model development, some degree of spatial dependence may still exist, because traffic crashes are inherently spatially clustered. Therefore, the contribution of accident density should be interpreted as representing historical spatial concentration patterns rather than an independent predictor of future crash occurrence.

The results also indicate that flat terrain and moderate road gradients are associated with higher accident risk. Similar findings have been reported in previous studies, where lowland and urbanized areas often exhibit higher accident frequencies due to greater traffic exposure and denser transportation networks (Berhanu et al., 2024; Le et al., 2024). In Thanh Hoa Province, these conditions are mainly distributed in densely populated plains with intensive transportation activities and higher traffic volumes compared to mountainous regions. Therefore, the observed relationship reflects not only terrain conditions, but also the combined effects of urbanization, road density, and traffic intensity.

Although intersection density is commonly associated with accident occurrence, its relative importance in the RF model was lower than expected. Previous studies have shown that intersections increase traffic conflicts and accident probability, particularly in urban areas (Lan & Hoc, 2024; Le et al., 2024). However, in this study, the influence of intersection density may be partially represented by other spatial variables, especially accident density and urban road concentration. In addition, the RF model evaluates variables simultaneously within a nonlinear multivariate framework, which may reduce the standalone contribution of variables sharing overlapping spatial information.

The risk zoning map further shows that high-risk road segments are primarily distributed along major highways, such as National Highways 1A, 45, and 47. Similar spatial concentration patterns along major transportation corridors have also been reported in previous traffic accident susceptibility studies conducted in rapidly urbanizing regions (Berhanu et al., 2024; Le et al., 2024). These roads typically have high traffic volumes and dense intersections with local road networks, increasing traffic conflicts and accident likelihood. Traffic exposure variables, such as traffic volume, AADT, and vehicle speed, may also influence the spatial distribution of traffic accidents, particularly along major highways and urbanized corridors with intensive transportation activities. Due to the limited availability of spatially continuous exposure data for the entire study area, the present framework primarily relied on spatial variables derived from GIS, remote sensing, road network characteristics, and historical accident patterns. This result is consistent with previous findings in Thanh Hoa Province, where accident hotspots are concentrated in high-traffic areas (Le et al., 2024).

In addition, statistical analysis indicates that high and very high-risk road segments account for only 14.3% of the total road length, but represent up to 83.1% of traffic accidents. This highlights the strong spatial concentration of accidents and demonstrates the effectiveness of the RF model in identifying critical high-risk segments.

Overall, the results demonstrate the capability of the proposed framework to identify spatial patterns and high-risk areas in Thanh Hoa Province. Since accident density was derived from historical accident records and incorporated as an input factor, the resulting maps primarily reflect the spatial distribution tendencies and relative susceptibility levels under existing environmental and transportation conditions. Therefore, the proposed approach should be interpreted as a GIS-based traffic accident susceptibility assessment framework rather than an independent future accident prediction model.

In addition, the validation strategy in this study was based on random data partitioning to maintain consistency in the comparative evaluation of the WoE and RF models. Although this approach is commonly adopted in susceptibility mapping studies, nearby samples may share similar spatial characteristics, potentially resulting in optimistic performance estimates. Therefore, spatially explicit validation strategies, such as spatial cross-validation or block cross-validation, may provide a more rigorous assessment of model robustness and generalization performance in future studies. Future studies may also incorporate additional traffic exposure variables, such as AADT, traffic volume, and vehicle speed, to further improve the interpretation and predictive capability of traffic crash susceptibility models.

CONCLUSION

This study evaluated and compared the performance of WoE and RF models for traffic crash risk mapping using remote sensing and GIS data in Thanh Hoa Province. The results demonstrate that the RF model outperforms the WoE model, achieving better model performance in terms of Accuracy and F1-score.

Variable importance analysis indicates that crash density is the most influential factor, followed by terrain slope and road gradient. The risk zoning results show that high-risk road segments are mainly concentrated along major highways, such as National Highways 1A, 45, and 47, where traffic density and intersection frequency are high. Notably, high- and very high-risk segments account for only 14.3% of the road network, but represent up to 83.1% of total crashes.

These findings confirm that the integration of remote sensing, GIS, and machine learning provides an effective approach for identifying high-risk areas and supporting traffic safety management.

However, this study is based on traffic crash data collected during the period 2020-2023, which experienced certain fluctuations due to the impact of COVID-19. Nevertheless, the dataset still reflects the actual spatial distribution characteristics of traffic crashes, thereby providing a reliable basis for model development and evaluation. Future studies may consider incorporating additional factors, such as traffic volume, weather conditions, and driver behavior, to further improve model performance and enhance traffic accident risk assessment.

Acknowledgements

This research is funded by the University of Transport and Communications (UTC) under grant number T2026-PHII_CT-009.

Data Availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request. The authors declare that there are no restrictions on data access relevant to the findings of this research.

Conflict of Interests

The authors report that there is no conflict of interests to declare.

REFERENCES

Abdulhafedh, A. (2022). Modeling vehicle crash frequency when multicollinearity exists in vehicle crash data: Ridge regression versus ordinary least squares linear regression. *Open Access Library Journal, 9*(6), 1-17. https://doi.org/10.4236/oalib.1108873

Agterberg, F.P., & Bonham-Carter, G.F. (1990). Deriving weights of evidence from geoscience contour maps for the prediction of discrete events. In *Proceedings of the 22nd APCOM Symposium* (pp. 381-395). Berlin, Germany.

Berhanu, Y., Schröder, D., Wodajo, B.T., & Alemayehu, E. (2024). Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks. *Results in Engineering, 23*, 102737. https://doi.org/10.1016/j.rineng.2024.102737

Bonham-Carter, G. (1994). *Geographic information systems for geoscientists: Modeling with GIS*. Elsevier. https://doi.org/10.1016/C2013-0-03864-9

Chen, C., Zhang, G., Qian, Z., Tarefder, R.A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention, 90*, 128-139. https://doi.org/10.1016/j.aap.2016.02.011

Deressa, B.F., Habtegiogis, K.A., Endashaw, D.K., Al-Ramadan, B.M., & Al-Ahmadi, H.M. (2025). A systematic review of GIS-driven road traffic accident evaluation. *Vehicles, 7*(4), 161. https://doi.org/10.3390/vehicles7040161

Dou, H.Q., Huang, S.Y., Jian, W.B., & Wang, H. (2023). Landslide susceptibility mapping of mountain roads based on machine learning combined model. *Journal of Mountain Science, 20*(5), 1232-1248. https://doi.org/10.1007/s11629-022-7657-2

Farhangi, F., Sadeghi-Niaraki, A., Razavi-Termeh, S. V., & Choi, S. M. (2021). Evaluation of tree-based machine learning algorithms for accident risk mapping caused by driver lack of alertness at a national scale. *Sustainability, 13*(18), 10239. https://doi.org/10.3390/su131810239

Gadtaula, A., & Dhakal, S. (2019). Landslide susceptibility mapping using weight of evidence method in Haku, Rasuwa district, Nepal. *Journal of Nepal Geological Society, 58*, 163-171.

Ha, L.T., Thao, V.T.P., & Thao, D.T.P. (2024). Analyzing road traffic incident hotspots using cluster analysis in Thanh Hoa Province of Vietnam. *International Journal of Geoinformatics, 20*(9), 16-26.

Khan, A.A., & Hussain, J. (2024). Utilizing GIS and machine learning for traffic accident prediction in urban environment. *Civil Engineering Journal, 10*(6), 1922-1935.

Lan, H.H.T., & Hoc, T.Q. (2024). Applying a two-step cluster algorithm in traffic accident data analysis. *Transportation and Communications Science Journal, 75*(4), 1673-1687. https://doi.org/10.47869/tcsj.75.4.16

Le, H.T., Vu, T.P.T., & Do, T.P.T. (2024). A study on identifying and analyzing road traffic incident hotspots on National Highway 1A, Thanh Hoa Province, Vietnam, employing statistical and GIS techniques. *Journal of Mining and Earth Sciences, 65*(6), 22-33. https://doi.org/10.46326/JMES.2024.65(6).03

Le, K.G. (2025). Improving road safety: Supervised machine learning analysis of factors influencing crash severity. *Science Journal of Silesian University of Technology. Series Transport, 127*, 129-153. https://doi.org/10.20858/sjsutst.2025.127.8

Le, K.G. (2026). Safety-oriented and explainable machine learning for KSI crash risk prediction: Evidence from the United Kingdom. *PLOS ONE, 21*(4), e0347873.

Le, K.G., Liu, P., & Lin, L.T. (2022). Traffic accident hotspot identification by integrating kernel density estimation and spatial autocorrelation analysis: A case study. *International Journal of Crashworthiness, 27*(2), 543-553. https://doi.org/10.1080/13588265.2020.1824747

Le, K.G., Liu, P., & Lin, L.T. (2019). Road traffic accident black spot determination by using kernel density estimation algorithm and cluster statistical significant evaluation. In *FIG Working Week*.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*(5), 291-305. https://doi.org/10.1016/j.tra.2010.02.001

Mannering, F.L., Shankar, V., & Bhat, C.R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research, 11*, 1-16.

Moyo, E., & Dzinavatonga, K. (2025). Logistic regression modeling of road traffic accident severity: A study on driver characteristics in Zimbabwe. *International Journal of Research and Innovation in Social Sciences, 9*, 3189-3206. https://doi.org/10.47772/IJRISS.2025.91100254

Savolainen, P.T., Mannering, F.L., Lord, D., & Quddus, M.A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention, 43*(5), 1666-1676. https://doi.org/10.1016/j.aap.2011.03.025

Wen, X., Xie, Y., Jiang, L., Pu, Z., & Ge, T. (2021). Applications of machine learning methods in traffic crash severity modeling: Current status and future directions. *Transport Reviews, 41*(6), 855-879.

World Health Organization. (2023). *Global status report on road safety 2023: Country and territory profiles*. World Health Organization. https://iris.who.int/server/api/core/bitstreams/46275f9f-ef66-4892-8ddd-a496ef8c1b74/content

Zhang, Z., Qian, Z., Wei, Y., Zhu, X., & Wang, L. (2022). Evaluation of geological disaster sensitivity based on the WOE-RF model. *Sustainability, 14*(23), 16247. https://doi.org/10.3390/su142316247