DOI: https://doi.org/10.14256/JCE.4319.2025

Primljen / Received: 7.5.2025.
Ispravljen / Corrected: 16.8.2025.
Prihvaćen / Accepted: 1.9.2025.
Dostupno online / Available online: 10.1.2026.

Predicting unsafe road sections using machine learning

Authors:

Riste Ristov, MSc. CE
University of St. Cyril and Methodius, N. Macedonia
Faculty of Civil Engineering
ristov@gf.ukim.edu.mk
Corresponding author

Prof. **Slobodan Ognjenović**, PhD. CE
University of St. Cyril and Methodius, N. Macedonia
Faculty of Civil Engineering
ognjenovic@gf.ukim.edu.mk

Prof. **Zlatko Zafirovski**, PhD. CE
University of St. Cyril and Methodius, N. Macedonia
Faculty of Civil Engineering
zafirovski@gf.ukim.edu.mk

Original research paper

Riste Ristov, Slobodan Ognjenović, Zlatko Zafirovski

Predicting unsafe road sections using machine learning

This paper presents an ML methodology to predict hazardous road segments, using the weighted accident index (Wi). The analysis covers 161 road segments in North Macedonia (~1,300 km)—with 23+1 variables categorized into Road, Traffic, Environmental, and Accident data. Feature influence is evaluated using six models with an 80/20 training/testing split. Weighted SHAP is applied to obtain a single variable ranking; XGBoost with 15 inputs is the final predictor. The model achieves a validated performance ($R^2 = 0.65$), while operational prioritization yields AUROC = 0.69 at Wi ≥ 10.13, enabling timely identification of hazardous segments and interventions by relevant authorities.

Key words:
road safety, machine learning, prediction, SHAP, weighted accident index, traffic analysis

Izvorni znanstveni rad

Riste Ristov, Slobodan Ognjenović, Zlatko Zafirovski

Predviđanje nesigurnih cestovnih dionica pomoću strojnog učenja

U ovome je radu opisana metodologija strojnog učenja za predviđanje opasnih cestovnih dionica primjenom ponderiranog indeksa nesreća (Wi). U analizu uključena je 161 cestovna dionica u Sjevernoj Makedoniji, ukupne duljine oko 1300 km, pri čemu su razmotrene 23 varijable svrstane u skupine koje se odnose na cestu, promet, okoliš i nesreće. Utjecaj značajki vrednovan je primjenom šest modela, uz podjelu podataka na skup za učenje i skup za testiranje u omjeru 80 : 20. Primjenom ponderiranog SHAP-a izvedeno je jedinstveno rangiranje varijabli, dok je konačni prediktivni model XGBoost temeljen na 15 ulaznih značajki. Potvrđena učinkovitost modela iznosi $R^2 = 0,65$, a u sklopu operativne prioritizacije postignut je AUROC = 0,69 pri Wi ≥ 10,13, što nadležnim institucijama omogućuje pravodobnu identifikaciju opasnih dionica i odgovarajuće intervencije.

Ključne riječi:
sigurnost u prometu, strojno učenje, predviđanje, SHAP, ponderirani indeks nesreća, analiza prometa

1. Introduction

Road traffic accidents represent a major global threat to public safety, particularly affecting the younger populations. According to the World Health Organization, road crashes are the leading cause of death among individuals aged 5 to 29 years, with over 1.19 million annual fatalities [1]. Systematic measures and infrastructure improvements have reduced mortality rates in developed countries; by contrast, the average number of fatalities in the European Union remains high, at 45.5 per million inhabitants [2].

In the Republic of North Macedonia, the situation is even more concerning, with 69.5 fatalities per million inhabitants [3], which is significantly above the European average. These figures highlight the urgent need for developing new analytical approaches to improving road safety, enabling timely identification of high-risk segments, and planning appropriate interventions.

Modern analytical techniques, such as machine learning and spatial-statistical methods, facilitate a proactive approach to detecting unsafe road sections before critical events occur. These technologies enable the identification of hazardous zones based on the influence of multiple factors as opposed to relying solely on historical accident data.

This study primarily aimed to develop a methodology for predicting unsafe road sections based on an analysis of road geometric characteristics, pavement conditions, traffic intensity, and external factors. Data from 161 sections of the primary road network, covering approximately 1300 km, were used in this study. Advanced machine learning models can be applied in identifying key factors that influence the weighted accident index (Wi) and developing predictive tools to enhance road safety.

Previous research has shown that the risk of road traffic accidents is influenced by various factors, such as longitudinal slope, curvature radius, pavement condition, lane width, and limited visibility [4-6]. Machine learning models, particularly the random forest (RF) and gradient boosting (GB) algorithms (XGBoost and CatBoost), have been increasingly employed to analyse such parameters owing to their ability to handle complex, nonlinear relationships and identify the most influential variables [7, 8].

Similar approaches have also been applied in the domain of infrastructure cost estimation, wherein ensemble learning techniques such as RF and boosting models have demonstrated strong predictive capabilities in capturing complex interactions between multiple input features [9]. Studies have also demonstrated the application of artificial neural networks (ANNs) for modelling crash frequency, particularly in cases with limited or partially available data; nevertheless, their interpretability remains restricted [10]. Several recent studies has reported on the use of SHapley Additive exPlanations (SHAP) analysis as a supplementary tool for interpreting the results from highly accurate models, thereby enabling a better understanding of the relative importance of factors such as traffic volume, road type, and mountainous terrain [11].

Additionally, composite indices and predictive risk maps have been developed wherein data are categorised based on the influence level and ranking instead of a simple binary classification [12, 13]. These approaches have proven useful for resource allocation and the prioritisation of interventions.

However, most existing analyses have been conducted in countries with well-developed data collection systems. In the context of the Republic of North Macedonia and the broader region, there is a lack of models that account for local constraints and the absence of comprehensive datasets [14]. This study aimed to address this gap by applying several machine learning models (XGBoost, RF, GB, CatBoost, LightGBM, and multilayer perceptron (MLP)), gradually reducing the number of input variables based on their influence on Wi. Thus, high-risk road sections were identified using representative national data.

2. Literature review

Research on predicting traffic accidents and their severity increasingly involves the application of advanced machine learning algorithms and explainable models for analysing influential factors. This section presents a review of relevant studies that have employed state-of-the-art techniques to forecast crash occurrences using real-world data and provides an overview of the geographic scope, number of cases, applied models, and performance metrics.

In a study conducted in China, Chen et al. [15] employed a hybrid MSCPO-XGBoost model on a dataset of 13,000 cases, achieving a coefficient of determination (R^2) = 0.918. They analysed factors related to crash severity by combining optimisation and machine learning. Iranmanesh et al. [16] apply XGBoost, decision tree (DT), and RF models on data from 784 crashes on rural roads in a province in Iran, achieving a maximum R^2 of 0.873. They applied these models to identify road segments with a high accident risk.

In a study using data from South Korea, Lee et al. [17] applied an interpretative approach with data augmentation to 11,689 records by using SHAP to identify infrastructure-related influences and achieved $R^2 = 0.842$. Mengistu et al. [18] analysed 1,037 cases involving drivers, roads, and environmental data in Ethiopia by applying XGBoost, achieving $R^2 = 0.863$. In both studies, model transparency in the factor explanation was highlighted as a key advantage.

Alshehri et al. [19] used DT and RF models on a dataset of 3,228 crashes in Saudi Arabia; they achieved an AUC (*Area Under Curve*) of up to 0.78 for predicting fatality across age groups and crash types. Additionally, the best-performing model achieved a precision of 0.81 and recall of 0.75, indicating a well-balanced capability to detect positive cases. Ahmed et al. [20] analysed 3,146 incidents in New Zealand by using explainable

Table 1. Comparative summary of representative studies

References	Region/Scope	Task	Models	Size	Metrics
Chen et al. [15]	China (nationwide)	Regression (severity)	MSCPO-XGBoost	13.000	R^2 = 0.918
Iranmanesh et al. [16]	Iran (rural roads)	Regression (risk/segments)	XGBoost, DT, RF	784	R^2(maks.) = 0.873
Lee et al. [17]	South Korea (national)	Regression + XAI	Interpretable ML + SHAP	11.689	R^2 = 0.842
Mengistu et al. [18]	Ethiopia (regional)	Regression (severity)	XGBoost	1037	R^2 = 0.863
Alshehri et al. [19]	Saudi Arabia (multi-city)	Classification (fatality risk)	DT, RF	3228	AUC ≤ 0.78; Precision = 0.81; Recall = 0.75
Ahmed et al. [20]	New Zealand (urban)	Hybrid (regression + classification)	XGBoost, LIME	3146	R^2 = 0.839; AUC = 0.87
Alpalhão et al. [22]	Portugal – Lisabon (urban)	Hybrid (GB)	GB	28.649	RMSE = 0.332

models such as XGBoost and local interpretable model-agnostic explanations (LIME), achieving R^2 = 0.839. As regards the classification component, the best model achieved an AUC of 0.87, highlighting its consistent performance in predicting severity levels. Furthermore, the practical applicability of these models is emphasised through factor impact visualisation.

Megnidio-Tchoukouegno and Adedeji [21] utilised the STATS19 database containing 45,000 records from the United Kingdom. They applied GB and RF models, with the best model achieving an R^2 value of 0.881. Alpalhão et al. [22] analysed 28,649 cases from Lisbon by using a hybrid regression/classification GB model, with a reported root mean square error (RMSE) of 0.332. These studies are particularly important for urban areas where data availability enables complex modelling. Guido et al. [23] focused on the Cosenza region in Italy with 1,349 cases and employed XGBoost, support vector machine, and RF models. The highest R^2 achieved was 0.896, and the models were used to analyse the number of vehicles involved and the road characteristics. Through geospatial analysis and machine learning, they demonstrated the applicability of the models in rural environments.

Xiao and Duan [24] developed a deep learning framework for multitask prediction by using input data from 10,563 cases, achieving R^2 = 0.894 and mean absolute error (MAE) = 0.243. Their study included a detailed SHAP analysis to visualise the contribution of each variable. It combined interpretability and multifunctionality in analysing crash severity.

Table 1 provides a concise overview of representative studies (region/scope, task, models, sample size, and metrics) for easier and more transparent cross-study comparisons, enabling a methodologically consistent assessment of the reported findings.

Building upon the reviewed research, this study applied six machine learning models (CatBoost, GB, XGBoost, RF, LightGBM, and MLP), all trained under the same conditions. The key contribution of this study lies in the systematic comparison of model performances in Wi prediction, along with the development of a methodology for parameter ranking based on combined feature importance scores. Furthermore, the results have practical implications in identifying high-risk segments in target road networks.

3. Data overview

3.1. General Information on the Road Network

The road network in the Republic of North Macedonia has a total length of 14,475 km and is classified into motorways, regional roads, and local roads [25]. The primary road network, which is 897 km long, represents a key segment of the national and trans-European transport infrastructure [26]. It includes motorways, expressways, and two-lane roads that provide the main traffic connections across the country and with neighbouring countries.

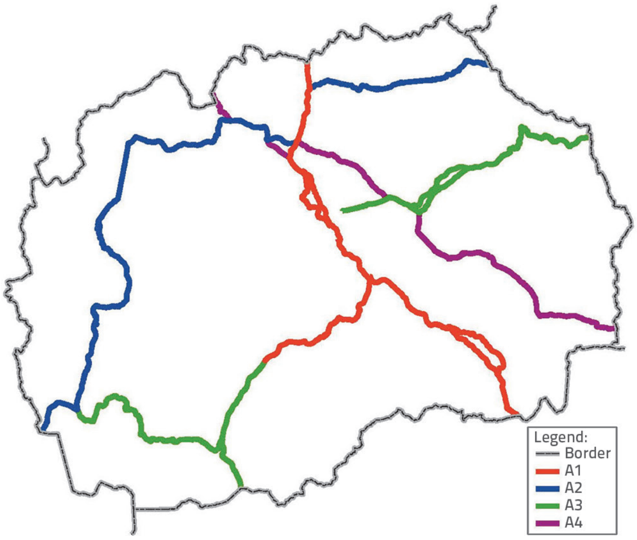


Figure 1. Overview of category A roads

This study focused on the primary roads A1, A2, A3, and A4, which differ in technical characteristics and geometric elements. As illustrated in Figure 1, although the official length of the primary road network is 897 km, the analysis covers approximately 1,300 km because of the separate treatment of the two road directions with divided carriageways (motorways). This approach enables a more detailed and objective assessment of the impacts of various factors on road safety.

3.2. Description and Processing of Data

Combining temporal categorisation and classification by characteristics yields a comprehensive and systematic approach to data processing. Temporal categorisation highlights changes over the years, while classification by characteristics allows for a precise understanding of the role and impact of each individual factor. The data were processed using GIS tools, statistical techniques, and machine learning methods, and the spatial and temporal trends were identified.

The final analytical database comprises 161 road sections (≈1,300 km) with complete records for the period 2014–2023. Before modelling, all layers were re-projected onto WGS 84, spatially joined by a kilometre mark, and cross-checked against duplicate IDs.

Road characteristics

This category included various geometric and functional parameters such as speed limits, alignment curvature, curve radii, longitudinal slopes, and elevation. In addition, the side forces in the curves, stopping sight distance, pavement roughness, rut depth, surface friction coefficient, and pavement condition index (PCI) were analysed. Data regarding the density of intersections, bridges, and viaducts, as well as the conditions of vertical and horizontal signage, were also included [27, 28].

Traffic characteristics

Traffic intensity is expressed through the annual average daily traffic (AADT) based on fixed and mobile automatic traffic counts. This parameter provides insights into the impact of traffic volume on accident risk [29].

Environmental characteristics

Climatic factors are represented by the average and extreme annual values for precipitation and temperature collected over a ten-year period from relevant meteorological stations. The data were processed using geospatial methods to ensure high-resolution coverage at the level of individual road sections [30].

Traffic accident data

The frequency and spatial distribution of traffic accidents were analysed using the index Wi, which considers both the number and severity of crashes. Fatal, injury, and property-damage-only accidents were weighted 85, 10, and 1, respectively, after which the score was normalised by the section length. This index serves as a key indicator for comparing the safety performances of different road sections [31].

Quality control included imputation of less than 3 % missing continuous values with the median of each variable, one-hot encoding of categorical indicators, and z-score standardisation of all continuous inputs. The geometric, traffic, and inventory layers were sourced from the official WebGIS portal of the Public Enterprise for State Roads, ensuring the consistency of measurement and attribution. Appendix B provides a complete list of variable definitions and composite index formulas.

3.3. Statistical summary of the dataset

The descriptive analysis begins with the training subset, wherein 80 % of the data were rescaled to the 0–1 interval to highlight the relative ranges of all input variables and the output

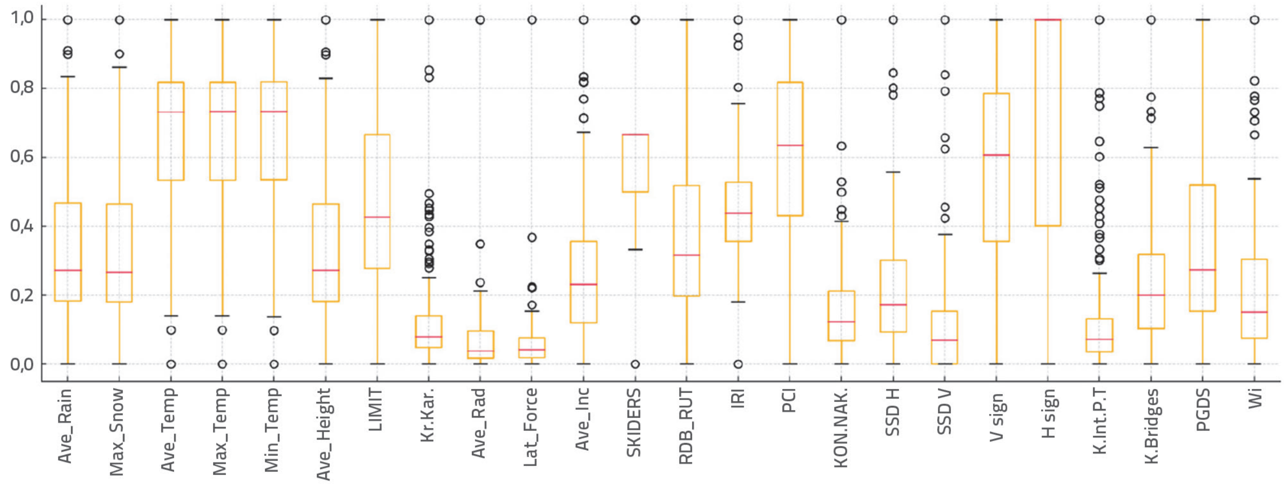


Figure 2. Normalized boxplot of input variables and Wi: training set

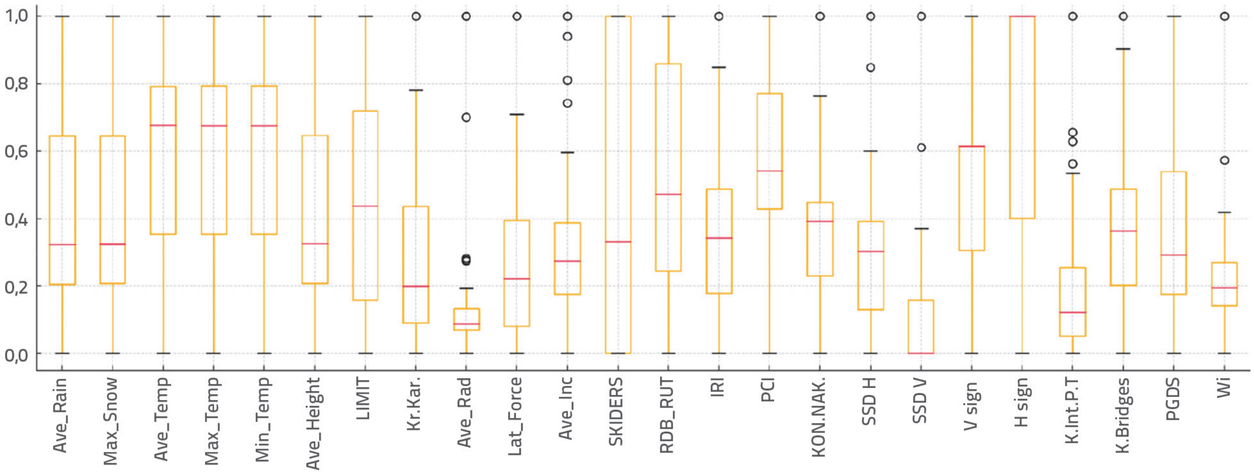


Figure 3. Normalized boxplot of input variables and Wi: test set

parameter (i.e. Wi). The diagram in Figure 2 reveals pronounced interquartile ranges for most parameters, while the isolated dots mark road sections that deviate substantially from the typical pattern - particularly for Wi, PGDS, and PCI.

The remaining 20 % of the data formed the test subset, processed using the same normalisation procedure. The distributions shown in the second diagram retain the shape and width of the interquartile ranges observed in the training set, indicating that the split is statistically representative and that the models are not exposed to a systematically different distribution of values during validation.

Figure 3 presents the distribution of variables in the test subset, enabling a direct comparison with the training data. The consistency between the two subsets provides a sound basis for assessing the general applicability of the developed models. Meanwhile, the discernible differences in box lengths and the number of outliers for individual variables underscore the contribution of each parameter to the variability in road safety conditions across the analysed sections.

4. Methodological Approach for the Development of the Weighted Accident Index (Wi) Prediction Model

A structured seven-step process that entails model selection, tuning, feature optimisation, and evaluation was applied to develop a reliable and interpretable model for predicting Wi. This methodology ensures the gradual refinement of both the model structure and input variables, with careful separation between exploratory analysis and formal validation. The workflow is illustrated in Figure 4.

Initial model screening

An initial exploratory comparison of nine different machine learning models was performed. These include linear models, DT-based models, boosting techniques, kernel-based models, and neural networks. This comparison served to identify algorithms with promising predictive potential on the basis of general trends in R^2 and error metrics [32, 33].

Model selection

Based on preliminary results, models that achieved an $R^2 > 0.50$ were considered sufficiently reliable to be included in the formal validation process.

Hyperparameter tuning

For each selected model, a 20-iteration random search was conducted on the 80 % training subset to identify suitable hyperparameters. The final tuned hyperparameters used in the subsequent 80/20 evaluation are presented in Table 2; all the metrics presented herein were computed using the held-out test set. A fixed random_state = 42 was used across all procedures to ensure reproducibility.

Feature selection and robustness analysis

The full set of 23 input parameters was gradually reduced in size by using the SHAP and permutation-based methods, and the performance was monitored after each removal. The

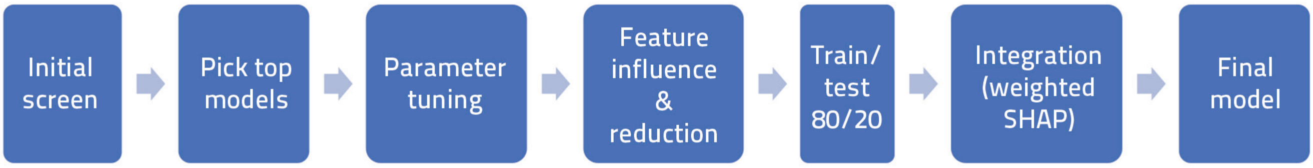


Figure 4. Linear workflow for the development of the Wi-prediction model

Table 2. Final hyperparameters for each algorithm (80:20 split)

Model	Final hyper-parameters
XGBoost (final predictor)	n_estimators = 100; max_depth = 4; learning_rate = 0.3 (default); random_state = 42
Gradient Boosting (GB)	n_estimators = 100; max_depth = 4; learning_rate = 0.1 (default); random_state = 42
Random Forest (RF)	n_estimators = 100; max_depth = 4; min_samples_split = 2; random_state = 42
CatBoost (CB)	iterations = 1000; depth = 6; learning_rate = 0.03; random_state = 42
LightGBM (LGBM)	n_estimators = 300; learning_rate = 0.03; max_depth = 4; min_child_samples = 10; random_state = 42
Multilayer Perceptron (MLP)	hidden_layer_sizes = (100, 50); activation = "relu"; alpha = 0.0005; max_iter = 1000; random_state = 42

final selection only included the most influential features for each model. Selection stability was checked using correlation-based analyses across random seeds/model variants [34, 35]. Permutation-based importance was quantified by the decrease in R^2 when each variable was shuffled.

Training and testing (80/20 split)

With optimised hyperparameters and reduced feature sets, each model was trained on 80 % of the data and tested on the remaining 20 %. This facilitated unbiased performance evaluations based exclusively on the test set [35, 36]. The performance is reported as R^2 (%), MAE, and RMSE for the held-out test set.

Integration for interpretability (no predictive ensemble)

No separate predictive ensemble was used, and integration served only to derive a robust cross-model ranking through weighted SHAP, with the contribution of each model being proportional to its R^2 value [39].

Final model formulation

The final prediction module is the single best-performing model selected under the 80/20 evaluation protocol (XGBoost with 15 inputs), with all metrics reported based on out-of-sample evaluation. If the model is subsequently refitted on the entire dataset, this is solely for deployment on new sections, while the reported results remain based on the 80/20 evaluation [46].

This methodology provides a coherent and reproducible framework for Wi prediction. Each step, from initial model screening to final deployment, was carefully structured to ensure transparency, robustness, and scientific rigor. The selected models, along with their tuning and validation procedures, made the developed tool accurate and practically applicable in road safety assessment contexts.

5. Analytical assessment of the performance of predictive models

The evaluation of various machine learning approaches to predict Wi is presented in stages. The evaluation covers the R^2 values of the models during training with the full dataset, testing using an 80/20 split, selection of optimal parameters, and development of the final predictive model.

5.1. Indicative training evaluation and model screening

In the initial phase of the analysis, all the selected models were trained using the entire dataset. This approach enabled a rapid assessment of the capabilities of the different methods for the predictive modelling of Wi [43, 45]. The analysis included various mathematical approaches, such as linear models, DT-based models, boosting models, kernel methods, and ANNs.

These nine models were selected on the basis of their compatibility with the nature of the available data and their proven applicability in previous road safety studies [6, 8].

The training results demonstrate the general potential of the models and were used to identify those suitable for further analysis. The selection criterion was achieving an R^2 value greater than 50 %, which was considered the minimum threshold for capturing the variability in Wi.

Figure 5 shows the initial performance of all nine models in terms of their R^2 and RMSE values. Models with R^2 values above 50 % were considered adequate for modelling this type of data and retained for further validation.

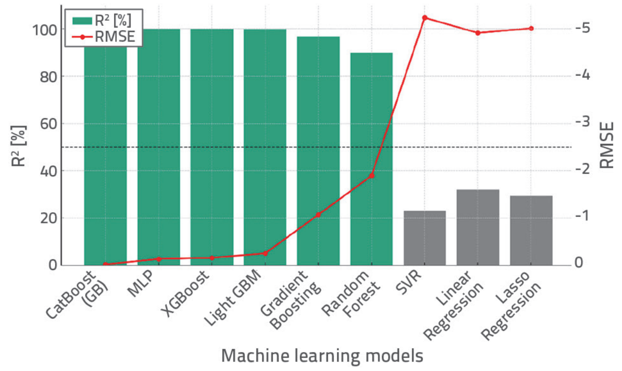


Figure 5. Initial R^2 and RMSE performance of the nine predictive models

As a result of this initial screening, six models, CatBoost [37], MLP, XGBoost [47], LightGBM [38], GB, and RF [40], were selected for further evaluation using the structured procedure outlined in the methodology. The excluded models failed to meet the threshold for explanatory power and were therefore not considered in the subsequent optimisation steps.

5.2. Defining the optimal number of influential parameters

A combined analysis was conducted using SHAP values and permutation importance to elucidate the influence of individual parameters on Wi. Both techniques provide a transparent interpretation of the role of each input variable in the final prediction, which is essential for drawing practical conclusions and defining the forecasting formulas.

The SHAP values originate from game theory and express the contribution of each parameter to a particular prediction. These values correspond to the principle of fair distribution of influence on the output of the model. SHAP provides nuanced insights into which parameters exert the greatest influence and whether that influence is positive or negative based on the direction and magnitude of the values [42-44]. The value for a parameter *i* can be calculated as follows:

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\left(|F|-|S|-1\right)!}{|F|!}\left[f\left(S \cup \{i\}\right)-f\left(S\right)\right] \qquad (1)$$

where Φ_i represents the SHAP value for parameter *i*, *S* is a subset of the remaining parameters, and *f*(*S*) is the model output for the input set *S*.

Additionally, the SHAP methodology facilitates indirect observation of interactions between parameters through their cumulative effect on the output of the model.

A permutation importance analysis was conducted to verify these results. This approach assesses the importance of each parameter by measuring the change in the model performance when the values of a particular parameter are replaced with permuted values. If this replacement causes a significant decrease in accuracy, the parameter is considered highly influentia [40]. The influence of the parameter is defined by the difference in the R² values, as follows:

$$\text{Importance}_i = R^2_{\text{original}} - R^2_{\text{permuted},i} \qquad (2)$$

where R^2_{original} is the explained variance of the original model and $R^2_{\text{permuted},i}$ is the value obtained after permuting the values of parameter *i*. The greater the difference, the more influential is the parameter on the model predictions.

For five of the six models (GB, RF, CatBoost, LightGBM, and XGB), the same SHAP explainer based on TreeExplainer was used, whereas for the MLP model, KernelExplainer was used because of its neural network structure.

Figure 6 shows the normalised importance of each of the 23 parameters compared across models. The numbers in the squares represent absolute SHAP values rounded to two decimal places, while the colour visually indicates the relative importance on a scale from 0 to 1. This visualisation enables a direct comparison of the parameter influences across all models; PGDS, LIMIT, and K.Int. P.T are clearly the most consistently influential parameters. Simultaneously, parameters such as KON.NAK., H sign, and Max_Temp show lower or selective importance only in individual models.

To finalise the parameter ranking, the SHAP values obtained from the six models were weighted according to their respective R² values from the 80/20 set analysis. These values are listed in Table 3.

Table 3. R² performance values used as weights in the final SHAP computation

Model	R² score	Weight
Gradient Boosting (GB)	0.5381	0.225
Random Forest (RF)	0.5126	0.214
CatBoost (CB)	0.4664	0.195
LightGBM (LGBM)	0.3447	0.144
Multilayer Perceptron (MLP)	0.1607	0.067
XGBoost (XGB)	0.1043	0.044

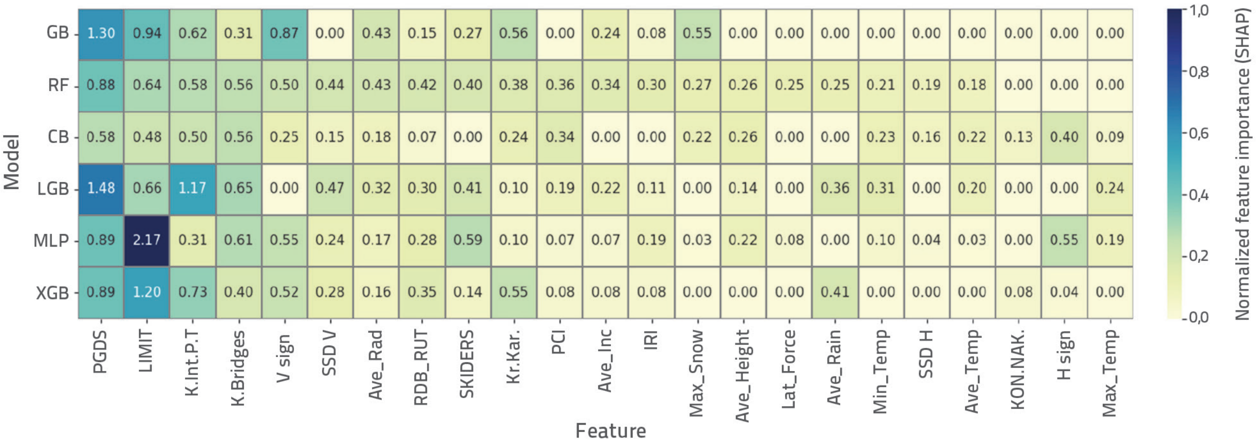


Figure 6. Normalized SHAP feature importance values across all models

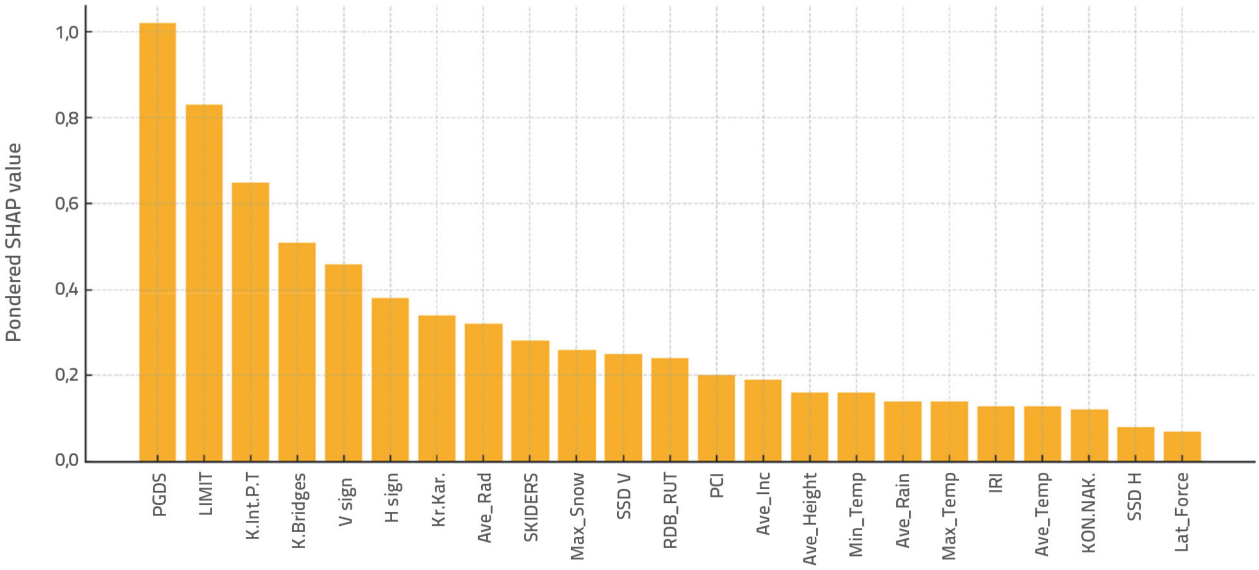


Figure 7. Feature Importance (weighted SHAP values)

Furthermore, to integrate the results from all the models into a synthetic assessment, a weighted SHAP value was calculated for each parameter, as follows:

$$S_j = w_1 \cdot S_{j1} + w_2 \cdot S_{j2} + \ldots + w_6 \cdot S_{j6} \tag{3}$$

Equation (3) represents the sum of the products of SHAP values for a given parameter *j* obtained from each of the six models (S_{j1}, S_{j2}, ..., S_{j6}) and their corresponding weights (w_1, w_2, ..., w_6), which were determined in proportion to the R² accuracy of each model. This yields a resultant SHAP value that integrates all the models into a unified importance metric.

Figure 7 shows the resulting influence values for each parameter, expressed using the weighted SHAP value. This bar chart enables the ranking of parameters according to their aggregated importance across all models. The highest values were observed for PGDS, LIMIT, and K.Int. P.T, indicating their consistent and dominant influence on the predictions across all the models. By contrast, parameters such as KON.NAK., Max_Temp, and H sign showed the lowest weighted SHAP values, suggesting that their effect on the output variable was minimal or only selectively significant in a limited number of models. This diagram allows for a visual assessment of the key factors for future analyses and potential reduction in the number of variables.

For further analysis, the parameters were ranked based on their weighted SHAP values, with the goal of gradually reducing the number of variables. This approach enables the identification of the most influential parameters without a classic stepwise method but with an integrated evaluation across all models.

5.3. Optimization of the number of parameters and selection of the most accurate model

In this phase, the six models were evaluated for their capability to predict Wi, with a focus on how prediction accuracy (i.e. R²) changes as the number of input parameters is reduced. The analysis was based on an 80/20 data split, with parameters sequentially removed according to their predefined importance rankings [41].

For each iteration, a subset of the most relevant features is selected, followed by model training and testing for the same split. The MLP model included input standardisation through a pipeline, whereas the LightGBM model employed specialised hyperparameters to control model complexity. A fixed random_state of 42 was used to ensure reproducibility of the results.

Figure 8 illustrates the variation of the R² values across different numbers of parameters for each model. The graph enables a visual comparison of the sensitivity and robustness of the model to dimensionality reduction. Notably, certain models, such as GB, exhibit stable performance over a wider parameter range, whereas others, such as MLP, show sharp fluctuations, particularly with fewer inputs.

Table 4 presents the maximum R² value achieved by each model, along with the corresponding number of parameters sorted in descending order of accuracy.

Table 4. Maximum R² (%) and corresponding number of parameters for each model

Model	R² [%]	Number of parameters
XGBoost	65.05	15
GB	57.13	7
MLP	56.39	9
LightGBM	55.21	9
CatBoost	52.06	7
RF	48.40	6

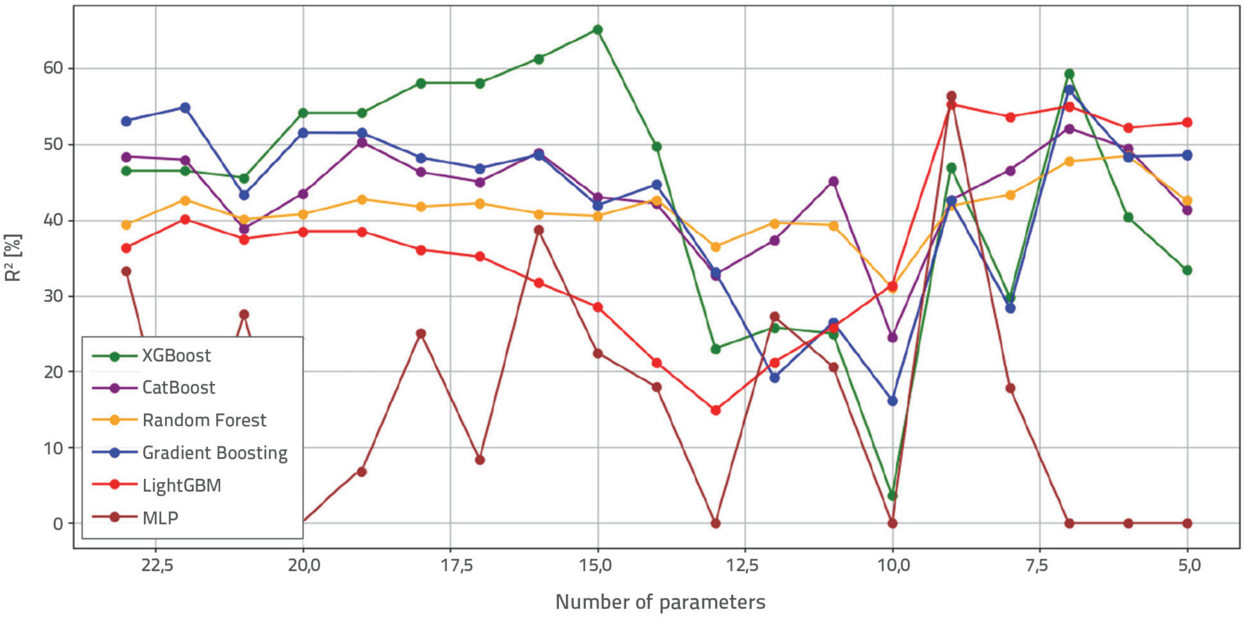


Figure 8. Variation of R² (%) with number of input parameters for all models

Based on these results, XGBoost was identified as the most suitable model for further applications, having achieved the highest R² value. Although GB and CatBoost operate with fewer parameters, they afford competitive accuracy and exhibit high stability, making them remarkably effective when data availability is limited. MLP and LightGBM achieved similar R² values but only under specific parameter conditions and with less consistency across the range.

These findings indicate that selecting the appropriate model and number of parameters can significantly improve prediction accuracy, even without relying on the full set of input variables. This analysis supports the formulation of a balanced trade-off between dimensionality and model stability, which is crucial for practical implementation.

5.4. Testing and validation

The process of testing and validation is essential to determine whether the developed predictive model for Wi is stable, accurate, and applicable in real-world conditions. As detailed in this section, regression-based testing was used to quantify the accuracy of the predictions, while classification-based validation was employed to examine the ability of the model to identify and rank higher-risk sections so as to support practical decisions on road safety interventions.

5.4.1. Testing (regression, 80/20)

The XGBoost model, trained on the 15 most influential parameters (defined via SHAP analysis), was evaluated using an 80 % for training and 20 % for test. The “Predicted vs. Actual” comparison with the ideal *y* = *x* line facilitates a direct visual examination of the agreement between model outputs and observed Wi values (Figure 9).

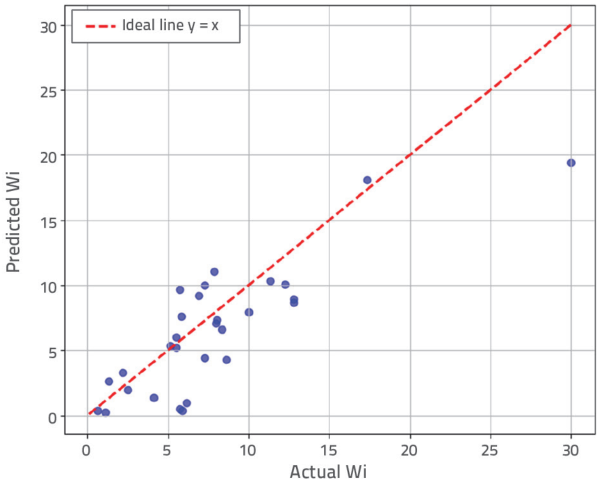


Figure 9. Comparison of predicted and actual Wi values from XGBoost with 80/20 training/test split

Most points were clustered near the ideal line, with the expected but limited spread at the extremes. Quantitatively, R² = 0.6505 indicates that a substantial share of the variance in Wi can be attributed to the model, while MAE = 2.72 and RMSE = 3.63 confirm moderate absolute and quadratic errors. Collectively, these results support the use of XGBoost as a solid basis for operational, section-level risk assessment.

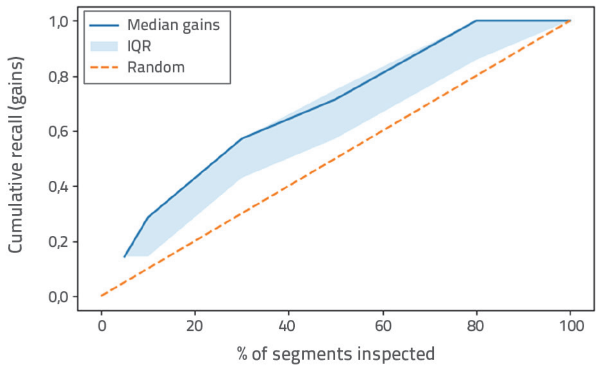
5.4.2. Validation (classification for risk prioritization)

Beyond accurate regression predictions, practical deployment requires the model to prioritise the highest-risk sections at the top of the ranked list. The regression model was adapted to a classification setting with the threshold Wi ≥ 10.130 (prevalence ≈ 20.5 %) in order to assess this aspect, and the performance was summarised over 100 bootstrap iterations. The choice of class-imbalance-aware metrics (precision/recall at fixed inspection rates, PR-AUC) and rank-oriented diagnostics (gains/lift) is methodologically appropriate for prioritisation tasks in road safety analytics and aligns with recent applications of machine learning in crash risk modelling [42, 43]. Aggregated results (medians with 95 % CI) are summarised as follows:

- AUROC: 0.692 [0.519, 0.834]
- PR-AUC: 0.420 [0.229, 0.696]
- Precision at 10 %: 0.500 [0.250, 0.750]
- Recall at 10 %: 0.286 [0.143, 0.429]
- Lift at 10 %: 2.857 [1.429, 4.286]

Figure 10 presents the gains curve. Inspecting only the top 10 % of road sections using the model's risk scores identified approximately half of all truly high-risk sections. This marks a significant improvement over random selection and a clear indication of operational usefulness for ranking.

Figure 10. Gains curve for identifying sections with Wi ≥ 10.130



(bootstrap median, 100 iterations)

Figure 11 shows the lift curve, which quantifies the advantage relative to random choice. In the top 10 % of the ranking, lift is approximately 2.9, implying that the model concentrates a markedly greater share of 'positive' cases near the top precisely the behaviour desired for effective prioritisation.

Collectively, the classification-based validation demonstrates that the mode delivers accurate Wi predictions and effectively ranks sections by risk. Combined with the regression test results, this provides a consistent and sufficiently strong confirmation of the practical applicability of the model for systematic road safety planning.

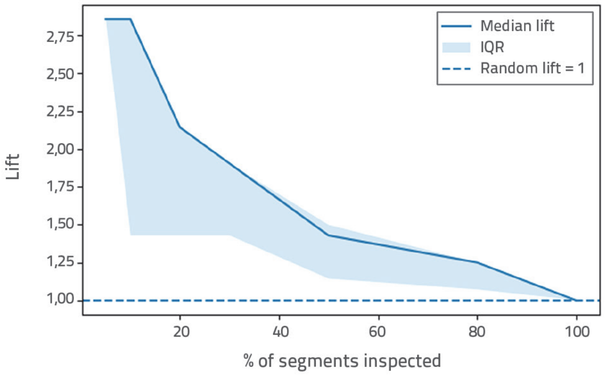


Figure 11. Lift curve for evaluating relative gain in ranked selection (bootstrap median, 100 iterations)

6. Discussion of the results

6.1. Analysis of the results

The analysis covered six models (XGBoost, CatBoost, GB, RF, LightGBM, and MLP) and followed two tracks: R^2-based evaluation (R^2, MAE, and RMSE) under an 80/20 train–test split and explainability via SHAP and permutation importance. The variable influence was computed for all six models and averaged with weights proportional to the R^2 value of each model from the 80/20 evaluation protocol (GB: R^2 = 0.5381, weight = 0.225; RF: R^2 = 0.5126, weight = 0.214; CatBoost: R^2 = 0.4664, weight = 0.195; LightGBM: R^2 = 0.3447, weight = 0.144; MLP: R^2 = 0.1607, weight = 0.067; and XGB: R^2 = 0.1043, weight = 0.044). This weighted SHAP aggregation produces a unified cross-model ranking that is stable and less sensitive to the idiosyncrasies of a single algorithm.

The resulting rankings show that the PGDS (AADT), LIMIT, and K.Int. P.T. were the most consistently influential factors, with PCI and Ave_Inc among the leading infrastructure/geometry variables. Conversely, KON.NAK., H sign, and Max_Temp exhibit selective/low influence. Mechanistically, exposure and operating conditions increase the baseline risk, while pavement conditions and geometry modulate it through friction, stability, and sight conditions.

Ablation analysis (R^2 as a function of the number of inputs) indicated a clear trade-off between compactness and accuracy. Peak R^2 (%) and optimal input counts per model are as follows: 65.05 (15 inputs) for XGBoost; 57.13 (7) for GB; 56.39 (9) for MLP; 55.21 (9) for LightGBM; 52.06 (7) for CatBoost; and 48.40 (6) for RF. Thus, the best generalisation is achieved with a reduced yet informative subset, as opposed to a full input space.

In the formal 80/20 validation, XGBoost [47] with 15 SHAP–permutation–selected inputs achieved R^2 = 0.6505, MAE = 2.72, and RMSE = 3.63, with test points concentrated around the ideal y = x line, confirming a strong agreement between the predicted and observed Wi values.

For operational validation (prioritisation), the regression output was adapted to a classification scenario using the threshold Wi ≥ 10.130 (prevalence ≈ 20.5 %) and 100× bootstrap summarisation. The results were as follows: AUROC = 0.692, PR-AUC = 0.420, Precision@10 % = 0.500, Recall@10 % = 0.286, and Lift@10 % = 2.857. These results imply that inspecting only the top 10 % of segments captured a substantially larger share of truly high-risk segments than random selection (as also evident from the gains/lift curves).

In summary, 15 inputs are sufficient for stable generalisation in XGBoost; GB and CatBoost deliver competitive performance with even fewer inputs (useful under data constraints). Beyond predicting Wi with strong out-of-sample agreement (R^2), the approach effectively concentrates the highest-risk segments at the top of the list for inspection and intervention. Interpretation relies on validated (out-of-sample) results; fits obtained on the full development set were used only exploratorily and not for formal evaluation.

6.2. Comparison with previous research

The validated results for Wi under an 80/20 split (XGBoost, 15 inputs: R^2 = 0.6505, MAE = 2.72, RMSE = 3.63) were aligned with current practices that favour tree-based ensembles and explainability (SHAP). In the regression studies summarised in Table 1, reported R^2 values typically fall in the vicinity of 0.84 to 0.92 for national/urban settings with larger and richer datasets [15,17,18,21,23]. In hybrid/interpretability-focused settings with an urban emphasis, results are commonly in the vicinity of R^2 ≈ 0.83 to 0.87 [20]. These differences are expected, arising from target mismatch (while many studies predict severity, this study modelled continuous Wi), spatial scale, and attribute richness. Hence, comparisons are intended to be methodological (emphasising the combined use of boosting and explainable artificial intelligence (XAI)) rather than direct numerical equivalence.

In classification and hybrid setups, studies have typically reported AUC/PR-AUC/precision/recall, with AUC approximately within 0.78 to 0.87 depending on task and data [19–20]. For operational comparability, a prioritisation check was also conducted in this study: at a threshold of Wi ≥ 10.130 (prevalence ≈ 20.5 %; 100× bootstrap), the results are AUROC = 0.692, PR-AUC = 0.420, Precision@10 % = 0.500, Recall@10 % = 0.286, and Lift@10 % = 2.857, indicating effective concentration of the highest-risk segments at the top of the list. The higher R^2/AUC values reported in the literature are partly attributable to larger and richer datasets (spatially and temporally), which provide broader variability and more efficient learning; the setup with 161 segments naturally imposes stricter conditions for generalization.

Regarding determinants, the findings indicate that exposure and operating environment (PGDS/AADT, speed limit, and intersection density) dominate, while pavement condition and geometry (PCI, longitudinal grade) modulate risk. These are consistent with studies that combine boosting with SHAP for explainability [17, 20, 24]. This supports the use of weighted SHAP aggregation 'across models' and justifies input reduction without materially compromising generalisation.

In summary, the out-of-sample results are in line with contemporary approaches (ensemble-based approaches coupled with SHAP) and are operationally useful for prioritisation. Table 1 serves as a reference frame for a methodologically consistent comparison of the tasks, metrics, and scales.

7. Limitations and future directions

Although the application of advanced machine learning models demonstrated validated out-of-sample performance (R^2 = 0.6505; MAE = 2.72; RMSE = 3.63) in predicting Wi, certain limitations need to be considered.

Careful management of the risk of overfitting is necessary, particularly for models involving a large number of parameters [48]. In the classification-based validation used for prioritisation, performance was bounded by the class prevalence (~20.5 %); this should be considered when interpreting AUROC/PR-AUC.

The main limitation arises from the availability and detail of the input data. Updated information regarding the condition of vertical and horizontal road signage, current state of road pavements, and complete climatic parameters is lacking.

Furthermore, the traffic accident data were limited in terms of detailed descriptions of causes, conditions at the time of accidents, and exact geographic locations of incidents, thereby influencing the precision of the models.

The models were trained using data from the main road network. Thus, applying the same approach to other road categories would require additional adaptation.

For future research, it is recommended to expand the database to include information on the current condition of the infrastructure, more specific climatic conditions, and factors related to road user behaviour.

Moreover, the use of combined algorithms and explainable machine learning techniques can further improve the predictive accuracy and interpretability of the model results.

8. Conclusion

This study demonstrated that advanced machine learning methods can reliably predict Wi at the road segment level and support operational decision-making. Under formal 80/20 validation, XGBoost with 15 SHAP-permutation-selected inputs achieved R^2 = 0.6505, MAE = 2.72, and RMSE = 3.63, indicating a strong out-of-sample agreement between the predicted and observed values. A cross-model, weighted SHAP aggregation, provided a stable ranking of determinants with AADT (PGDS), speed limit

(LIMIT), and intersection density (K.Int. P.T) as the most influential, followed by the pavement condition (PCI) and longitudinal grade (Ave_ Inc.). An operational check tailored for prioritisation further showed that, at Wi ≥ 10.130 (prevalence ≈ 20.5 %), the model effectively concentrates risk (AUROC = 0.692; PR-AUC = 0.420; Precision@10 % = 0.500; Recall@10 % = 0.286; Lift@10 % = 2.857), making it suitable for targeting inspections and interventions when resources are limited.

The main contributions of this study are as follows. a standardised out-of-sample comparison of ensemble and baseline models trained under equal conditions for Wi prediction; an integrated variable-ranking procedure (weighted SHAP across six models) that guides dimensionality reduction without materially sacrificing generalisation; and an operational validation framework that links regression outputs to the actionable rank-based screening of high-risk segments.

Although the results are promising, they remain conditioned by data granularity and coverage (e.g. detailed signage state, current pavement condition, precise crash geolocation, and richer climatic descriptors). Expanding and updating these inputs, incorporating spatial/temporal structures, and conducting external validation on additional networks are expected to improve explanatory power and transferability. In practice, the findings indicate the need for a combined strategy of managing exposure and the operating speed environment (e.g. speed-limit policies and junction management), while maintaining and upgrading infrastructure (surface condition, drainage, and sight distance). The proposed approach offers a transparent, explainable, and operationally meaningful pathway for prioritising road safety measures at scale.

REFERENCES

[1] World Health Organization: Road traffic injuries, Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed March 2025)., 2024

[2] European Commission: Road Safety Statistics 2023, Available at: https://road-safety.transport.ec.europa.eu/european-road-safety-observatory_en (accessed March 2025).

[3] State Statistical Office of the Republic of North Macedonia: MakStat - Statistical Database. Available at: https://makstat.stat.gov.mk/PXWeb/pxweb/en/ (accessed March 2025), 2025

[4] Ziakopoulos, A., Yannis, G.: A review of spatial approaches in road safety, Accident Analysis & Prevention, 135 (2020), Paper 105323, https://doi.org/10.1016/j.aap.2019.105323

[5] Aguero-Valverde, J., Jovanis, P.P.: Spatial analysis of fatal and injury crashes in Pennsylvania, Accident Analysis & Prevention, 38 (2006) 3, pp. 618–625, https://doi.org/10.1016/j.aap.2005.12.006

[6] Silva, P.B., Andrade, M., Ferreira, S.: Machine learning applied to road safety modeling: A systematic literature review, Journal of Traffic and Transportation Engineering (English Edition), 7 (2020) 6, pp. 775–790, https://doi.org/10.1016/j.jtte.2020.07.004

[7] Almahdi, A., Al Mamlook, R.E., Bandara, N., Almuflih, A.S., Nasayreh, A., Gharaibeh, H., Alasim, F., Aljohani, A., Jamal, A.: Boosting ensemble learning for freeway crash classification under varying traffic conditions: A hyperparameter optimization approach, Sustainability, 15 (2023) 22, Paper 15896, https://doi.org/10.3390/su152215896

[8] Dong, S., Khattak, A., Ullah, I., Zhou, J., Hussain, A.: Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley Additive exPlanations, International Journal of Environmental Research and Public Health, 19 (2022) 5, Paper 2925, https://doi.org/10.3390/ijerph19052925

[9] Kovačević, M., Ivanišević, N., Petronijević, P., Despotović, V.: Construction cost estimation of reinforced and prestressed concrete bridges using machine learning. Građevinar, 73 (2021) 1, pp. 1–13, https://doi.org/10.14256/JCE.2738.2019

[10] Gatarić, D., Ruškić, N., Aleksić, B., Đurić, T., Pezo, L., Lončar, B., Pezo, M.: Predicting road traffic accidents-Artificial neural network approach. Algorithms, 16 (2023) 5, Paper 257, https://doi.org/10.3390/a16050257

[11] Xiao, Y., Duan, Z.: An explainable multi-task deep learning framework for crash severity prediction using multisource data,Scientific Reports, 15 (2025), Paper 39226, https://doi.org/10.1038/s41598-025-09226-1

[12] Behboudi, N., Moosavi, S., Ramnath, R.: Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques, arXiv preprint, arXiv:2406.13968. https://doi.org/10.48550/arXiv.2406.13968, 2024

[13] Li, W., Luo, Z.: Research on traffic accident risk prediction method based on spatial and visual semantics. ISPRS International Journal of Geo-Information, 12 (2023) 12, Paper 496, https://doi.org/10.3390/ijgi12120496

[14] Li, H., Chen, X.: Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories, PLOS ONE, 20 (2025) 7, Paper e0320656, https://doi.org/10.1371/journal.pone.0320656

[15] Chen, F., Liu, X.Q., Yang, J.J., Liu, X.K., Ma, J.H., Chen, J., Xiao, H.Y.: Traffic accident severity prediction based on an enhanced MSCPO-XGBoost hybrid model. Scientific Reports, 15 (2025), Paper 25729, https://doi.org/10.1038/s41598-025-00797-7

[16] Iranmanesh, M., Seyedabrishami, S., Moridpour, S.: Identifying high crash risk segments in rural roads using ensemble decision tree-based models. Scientific Reports, 12 (2022), Article 20024, https://doi.org/10.1038/s41598-022-24476-z

[17] Lee, J., Kim, S., Heo, T.Y., Lee, D.: Identifying the roadway infrastructure factors affecting road accidents using interpretable machine learning and data augmentation, Applied Sciences, 15 (2025, 5), Paper 501, https://doi.org/10.3390/app15020501

[18] Mengistu, A.K., Gedefaw, A.E., Baykemagn, N.D., Walle, A.D., Yehuala, T.Z., Alemayehu, M.A., Messelu, M.A., Assaye, B. T.: Predicting car accident severity in Northwest Ethiopia: A machine learning approach leveraging driver, environmental, and road conditions, Scientific Reports, 15 (2025), Paper 21913. https://doi.org/10.1038/s41598-025-08005-2

[19] Alshehri, A.H., Alanazi, F., Yosri, A.M., Yasir, M.: Comparing fatal crash risk factors by age and crash type using machine learning techniques, PLOS ONE, 19 (2024) 5, e0302171, https://doi.org/10.1371/journal.pone.0302171

[20] Ahmed, S., Hossain, M.A., Ray, S.K., Bhuiyan, M.M.I., Sabuj, S.R.: A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance, Transportation Research Interdisciplinary Perspectives, 19 (2023), Paper 100814, https://doi.org/10.1016/j.trip.2023.100814

[21] Megnidio-Tchoukouegno, M., Adedeji, J.A.: Machine learning for road traffic accident improvement and environmental resource management in the transportation sector (using UK STATS19 data), Sustainability, 15 (2023) 3, Paper 2014, https://doi.org/10.3390/su15032014

[22] Alpalhão, N., Sarmento, P., Jardim, B., de Castro Neto, M.: Assessing the risk of traffic accidents in Lisbon using a gradient boosting algorithm with a hybrid classification/regression approach, Transportation Research Interdisciplinary Perspectives, 21 (2025), Paper 101495, https://doi.org/10.1016/j.trip.2025.101495

[23] Guido, G., Shaffiee Haghshenas, S., Vitale, A., Astarita, V., Park, Y., Geem, Z.W.: Evaluation of contributing factors affecting number of vehicles involved in crashes using machine learning techniques in rural roads of Cosenza, Italy. Safety, 8 (2023) 2, Paper 28, https://doi.org/10.3390/safety8020028

[24] Xiao, Y., Duan, Z.: An explainable multi-task deep learning framework for crash severity prediction using multisource data, Scientific Reports, 15 (2025), Paper 9226. https://doi.org/10.1038/s41598-025-09226-1

[25] Public Enterprise for State Roads, Web-GIS Platform for Spatial Analysis and Visualization, Available at: http://62.77.137.99/pesr/webgis/#/map (accessed March 2025).

[26] Ministry of Local Self-Government, Development Program for Planning Regions for the Period 2021–2026, Government of the Republic of North Macedonia, 2021.

[27] Doncheva, R., Ognjenović, S.: Proektiranje patishta, University "Ss. Cyril and Methodius" – Faculty of Civil Engineering, Skopje, ISBN: 978-608-4510-60-4, 2024.

[28] Tobias, P., de León Izeppi, E., Flintsch, G., Katicha, S., McCarthy, R.: Pavement Friction for Road Safety: Primer on Friction Measurement and Management Methods, Federal Highway Administration (FHWA), Report No. FHWA-SA-23-007, 2023.

[29] Public Enterprise for State Roads, Web-GIS Platform for Spatial Analysis and Visualization, Available at: http://tdps.roads.org.mk/ (accessed March 2025).

[30] Gjeshovska, V., Taseski, G., Ilioski, B.: Intensive Precipitation in the Republic of North Macedonia, University "Ss. Cyril and Methodius" – Faculty of Civil Engineering, Skopje, ISBN: 978-608-4510-56-7, 2024.

[31] Government of the Republic of North Macedonia, Ministry of Transport, Project Implementation Unit, Handbook on Black Spot Management (BSM), July 2024.

[32] Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: an overview, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2 (2012) 1, pp. 86–97, https://doi.org/10.1002/widm.53.

[33] Trajkovski, V.: How to Select Appropriate Statistical Test in Scientific Articles, Journal of Special Education and Rehabilitation, 17 (2016) 3–4, pp. 5–28, https://doi.org/10.19057/jser.2016.7.

[34] Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, https://doi.org/10.1007/978-0-387-21606-5, 2001.

[35] Kuhn, M., Johnson, K.: Applied Predictive Modeling, Springer, https://doi.org/10.1007/978-1-4614-6849-3, 2013.

[36] Raschka, S.: Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, arXiv (2018), https://doi.org/10.48550/arXiv.1811.12808.

[37] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features, arXiv (2017), https://doi.org/10.48550/arXiv.1706.09516.

[38] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree, Advances in Neural Information Processing Systems 30 (NeurIPS 2017), pp. 3149–3157, URL: https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree

[39] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (1998) 3, pp. 226–239, https://doi.org/10.1109/34.667881

[40] Breiman, L.: Random Forests, Machine Learning, 45 (2001) 1, pp. 5–32,https://doi.org/10.1023/A:1010933404324.

[4’] Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29 (2001) 5, pp. 1189–1232, https://doi.org/10.1214/aos/1013203451.

[42] Ahmed, S., Hossain, M.A., Ray, S.K., Bhuiyan, M.M.I., Sabuj, S.R.: A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance, Transportation Research Interdisciplinary Perspectives, 19 (2023), Paper 100814, https://doi.org/10.1016/j.trip.2023.100814.

[43] Alshehri, A.H., Alanazi, F., Yosri, A.M., Yasir, M.: Comparing fatal crash risk factors by age and crash type using machine learning techniques, PLOS ONE, 19 (2024) 5, e0302171, https://doi.org/10.1371/journal.pone.0302171.

[44] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 3 (2003), pp. 1157–1182, https://jmlr.org/papers/v3/guyon03a.html.

[45] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 30 (2017), pp. 4765–4774, https://doi.org/10.48550/arXiv.1705.07874

[46] Molnar, C.: Interpretable Machine Learning, Second edition, self-published, 2022, https://doi.org/10.1177/09726225241252009

[47] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016), https://doi.org/10.1145/2939672.2939785.

[48] Patil, P., Du, J.H., Kuchibhotla, A.K.: Bagging in overparameterized learning: Risk characterization and risk monotonization, arXiv preprint arXiv:2210.11445, (2022), https://doi.org/10.48550/arXiv.2210.11445.