https://doi.org/10.3311/PPtr.39701

Periodica Polytechnica Transportation Engineering

Road Traffic Accident Prediction by Machine Learning and GIS: Case Study in Thanh Hoa Province, Vietnam

Ha Le Thi[1], Thao Vu Thi Phuong[2*], Thao Do Thi Phuong[3]

[1] Faculty of Civil Engineering, Campus in Ho Chi Minh City, University of Transport and Communications, 450-451 Le Van Viet Street, Tang Nhon Phu Ward, 700000 Ho Chi Minh City, Vietnam
[2] Faculty of Environment, Hanoi University of Mining and Geology, Ministry of Education and Training of Vietnam, 18 Duc Thang, 129000 Bac Tu Liem, Hanoi, Vietnam
[3] Department of Geodesy - Map and land management, Hanoi University of Mining and Geology, Ministry of Education and Training of Vietnam, 18 Duc Thang, 129000 Bac Tu Liem, Hanoi, Vietnam
[*] Corresponding author, e-mail: vuthiphuongthao@humg.edu.vn

Received: 22 December 2024, Accepted: 20 August 2025, Published online: 27 November 2025

Abstract

Traffic accidents pose significant challenges for communities worldwide, particularly in Vietnam, where many individuals live on low to middle incomes and where infrastructure often struggles to keep pace with rapid mechanization. This study uses a historical data set of traffic accidents from 2020 to 2023 in Thanh Hoa province, Vietnam, as input data for Random Forest and Spline Regression machine learning models to predict the number of deaths and injuries from traffic accidents. A traffic accident prediction map for 2024 is established from the predicted results of the death and injury numbers obtained from Random Forest combined with GIS technology. The prediction results show superiority in providing detailed information about accidents to intervene to make traffic safer, especially in areas at high risk and during peak periods of accidents. The Random Forest model demonstrated superior performance to Spline Regression, achieving a mean absolute error of 0.012072 for deaths and 0.036323 for injuries, with the R^2 values of 0.998663 and 0.996552, respectively. Including lagged variables and adjusting for seasonal effects further improved the accuracy of daily predictions. The study offers an approach to solving traffic accidents in low- and middle-income countries, where traffic accident prediction methods based on historical data sources are still not widely used, with the hope of applying machine learning and GIS in road safety management shortly.

Keywords

historical data sources, low- and middle-income countries, random forest algorithm, spline regression, traffic accident prediction

1 Introduction

Traffic accidents are a serious problem affecting global public health, causing tremendous pressure on countries' economies, especially in low- and middle-income countries. Injuries from road traffic accidents are among the leading causes of death worldwide, with the highest number of deaths in low- and middle-income countries where infrastructure development is often slower than the rapid increase in industrialization and urbanization (Ahmed et al., 2023; Pelicic et al., 2024). In Vietnam, the frequency of road traffic accidents and their severe consequences are especially evident due to rapid urbanization, poor road conditions, changes in traffic density, unfavorable environmental factors, etc.

Traffic accidents have a profound impact on the economy if not controlled, not monitored and not found the cause. Therefore, the road traffic accident prediction model is built by analyzing locations with high risk of accidents and the related factors. In previous years, traditional statistical techniques, such as linear regression, were commonly used to analyze road traffic accidents (Azami et al., 2024; Isler et al.,2024; Phaphan et al., 2023). However, traffic accidents often have complex, non-linear interactions that traditional statistical techniques do not provide high accuracy or interpretability. In several recent publications, to evaluate the influence of factors related to traffic accidents, the Random Forest algorithm has been used because of its robustness in processing high-depth dimensional data sets with complex interactions, as well as the ability to rank the importance of features (Bhele et al., 2024; Tejashwini et al., 2024). Meanwhile, Spline regression is used to capture non-linear and seasonal trends in traffic data, making it especially

Cite this article as: Thi, H. L., Phuong, T. V. T., Phuong, T. D. T. (2025) "Road Traffic Accident Prediction by Machine Learning and GIS: Case Study in Thanh Hoa Province, Vietnam", Periodica Polytechnica Transportation Engineering, https://doi.org/10.3311/PPtr.39701

effective for analyzing long-term trends and cyclical fluctuations. period (Ajao et al., 2023; Shelevytsky et al., 2020). These models overcome the limitations of traditional methods, providing improved accuracy and interpretability in traffic accident prediction (Alok et al., 2024; Das et al., 2024). Besides, GIS is a powerful tool for analyzing spatial problems and is used to map accidents and create visual and vivid images (Ulu, 2024).

This study uses a combination of GIS with machine learning models and a traffic accident data set from 2020 to 2023 in Thanh Hoa province to create a map to predict the number of deaths and injuries due to traffic accidents in Thanh Hoa province, one of the areas that illustrate infrastructure challenges and complex terrain conditions. The unique combination of rural and urban roads, combined with seasonal climate variations and diverse traffic patterns, creates an ideal case for analyzing traffic safety dynamics (Fahad et al., 2023; Wang et al., 2024). The study incorporated lagged variables and removed seasonal effects, allowing the models to capture temporal dependence and accommodate recurring patterns in traffic accident trends.

2 Materials and methods

2.1 Study area

The study was conducted in Thanh Hoa province, Vietnam, which is a large province of the North Central region with geographical coordinates ranging from 20°40' to 19°18' North latitude and from 106°04' to 104°22' East longitude (Fig. 1). Thanh Hoa province is one of the five largest provinces in Vietnam, with an area of 11.080,8 km^2. The population of Thanh Hoa province ranks third in Vietnam, with 3.640.128 people in 2019 statistics (Ha et al., 2024). Thanh Hoa province has terrain sloping from Northwest to Southeast. In particular, the mountainous and midland regions occupy most of the province's area. The province's transportation system is very convenient both within and internationally, with the Ho Chi Minh Road and many other national highways.

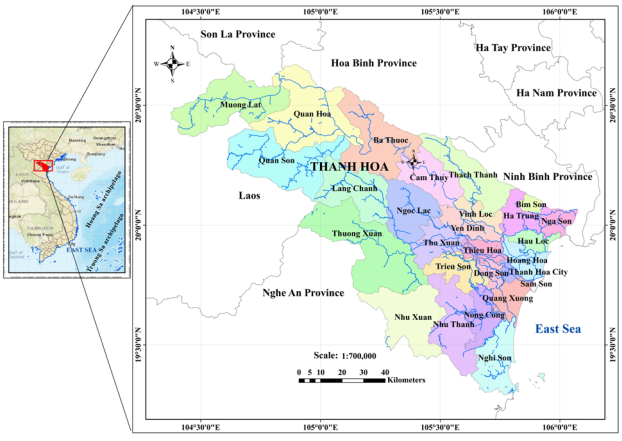


Fig. 1 Study area - Thanh Hoa province of Vietnam

2.2 Methodology

The study was conducted in four steps to predict the accident risk space as shown in Fig. 2. First, the process begins by integrating the geographic database including road layer, boundary layer, and surface cover with the traffic accident database from the geographic information system (GIS) to create a map of accident points in the study area. Second, data preprocessing is an important step to ensure input data quality for the analysis process. Third, the machine learning part includes dividing the data set into 20% test and 80% training, then applying Spline Regression and Random Forest models for training and testing. From there, evaluate the model performance using the MAE, MSE, and R^2 indices. Finally, the road traffic accident risk prediction map is established through the best model.

2.2.1 Map of historical road traffic accidents

The study used two data sets to establish a map of historical road traffic accidents. First, the road network map of Thanh Hoa province and administrative boundaries were digitized from remote sensing image data in shapefile form. Road attribute data includes length, width, road type, etc. collected from reports provided by Thanh Hoa Department of Natural Resources and Environment.

The second data set of 1,594 traffic accidents in 4 years (2020-2023) was provided by Thanh Hoa Provincial Traffic Police Department. This data set contains essential features such as geographical coordinates (latitude and longitude), time information (day, month, year), road conditions, lighting and weather conditions, and results of accidents (deaths and injuries). Other specific data such as the age of

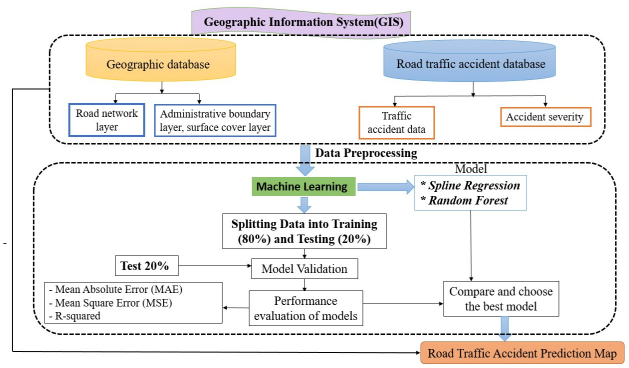


Fig. 2 Study methodology

the victim, means of transportation, and number of vehicles involved. Fig. 3 illustrates the locations of accident incidents in Thanh Hoa from 2020 to 2023.

2.2.2 Data description

The data are depicted in histogram Fig. 4. This is a comprehensive analysis of the frequency distribution of continuous variables (attributes of incidents) in the study, including geographical variables (latitude, longitude, etc.) degree), time factors (day, month, year), environmental conditions (light and weather) and specific factors of the accident (number of vehicles, age of victims, road safety index, number of deaths and injuries). Analyzing these variables is essential in providing the necessary information for the prediction model.

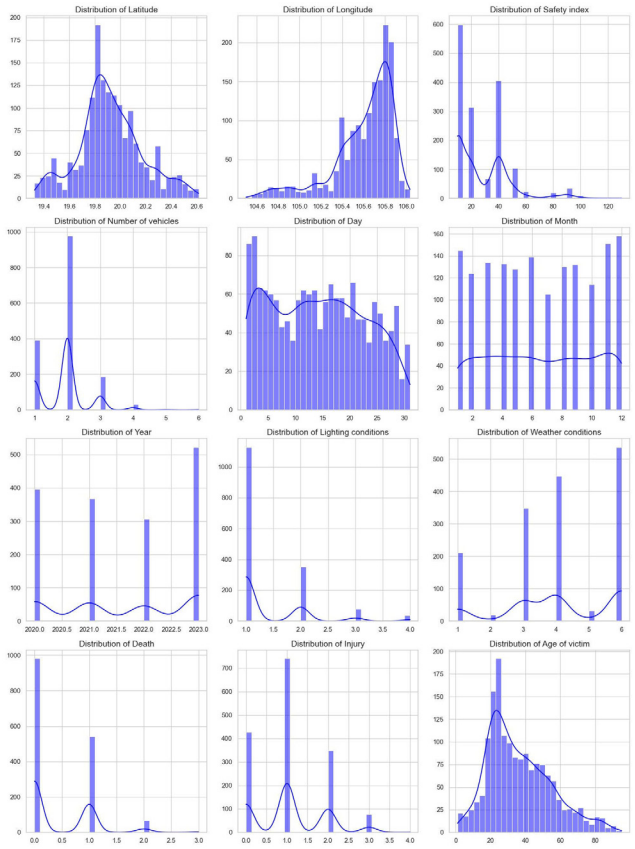


Fig. 4 Frequency distribution of main continuous variables in traffic accident incident data

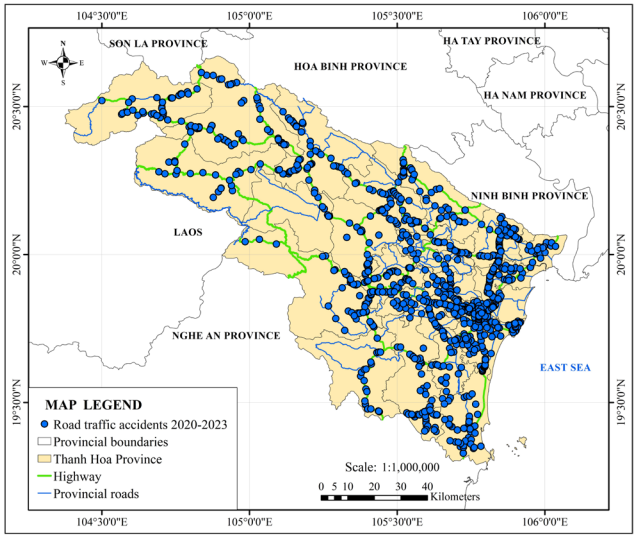


Fig. 3 Map of road traffic incidents in Thanh Hoa province (2020-2023)

2.3 Machine learning model

Two machine learning models are used to predict deaths and injuries: Spline Regression and Random Forest Regression.

2.3.1 Spline regression

Spline regression was chosen for this study because of its ability to model nonlinear relationships, which are common in traffic accident data. Spline functions allow for smooth curve fitting across different parts of the data, effectively capturing complex trends that may not be linear. The spline regression model used in this study was implemented with cubic splines, which provides flexibility in fitting both short-term fluctuations and long-term trends in accident rates (Ajao et al., 2023; Meganfi et al., 2023; Schuster et al., 2022; Yu et al., 2024).The cubic spline function is expressed as follows:

$$y(x) = \beta_0 + \sum_{i=1}^{n} \beta_i B_i(x). \quad (1)$$

$B_i(x)$ are the coefficients estimated from the data. Nodes for the spline were determined using cross-validation to minimize the risk of overfitting.

2.3.2 Random forest

Random Forest was chosen because of its robustness in handling large data sets with many dimensions and its ability to measure the importance of features. The model consists of multiple decision trees, where each tree is trained on a random subset of data and features. The final prediction is made by averaging the predictions from all trees, thus reducing the risk of overfitting (Prasetiyowati et al., 2020; Venugopal et al., 2024).

The Random Forest model is determined by the following parameters:

- n_estimators: number of decision trees in the forest (adjustable from 100 to 300).
- max_depth: maximum depth of each tree (adjustable between 10 and 20 or unlimited).
- min_samples_split: minimum number of samples required to split a node (adjustable from 2 to 10).
- min_samples_leaf: Minimum number of samples required at a leaf node (adjustable between 1 and 4).

- max_features: Number of features to consider when looking for the best split ('auto', 'sqrt' or 'log2').

To improve model accuracy, a grid search with cross-validation was applied to determine the optimal hyperparameters. A TimeSeriesSplit method was used for cross-validation to ensure that the chronological sequence of the data was preserved. TimeSeriesSplit splits the data into training (80%) and validation (20%) sets, ensuring that each validation set only contains data from the future relative to the training set. This method is important for time series prediction, where the goal is to predict future events based on past information.

2.3.3 Incorporation of lagged time variables and seasonal adjustment

To capture short-term temporal dependencies in traffic accident patterns, we generated lagged versions of the target variables, including fatalities and injuries from the preceding one and two days. Specifically, the constructed variables include:

- Fatality_{t-1}, Injury_{t-1}: number of deaths/injuries from the previous day
- Fatality_{t-2}, Injury_{t-2}: values from two days prior

These lagged time variables were included as input features in both the Random Forest and Spline Regression models. Their inclusion improved predictive performance by allowing the models to account for autocorrelation and recent trends in accident severity.

Additionally, the study employed seasonal-trend decomposition using LOESS (STL) on the daily time series of fatalities and injuries. This process separated the data into trend, seasonal, and residual components. By removing seasonal effects (e.g., monthly peaks associated with holidays or weather patterns), the models were trained on de-seasonalized data, enhancing robustness and generalization.

These techniques jointly improved the models' ability to predict daily accident outcomes and reduced sensitivity to short-term fluctuations or periodic spikes.

2.4 Calculate the models

Models are evaluated based on several performance metrics:

- Mean absolute error (MAE): A measure of prediction accuracy, calculating the average absolute difference between the actual and predicted values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_l\right|, \tag{2}$$

where n is the number of observations; y_i is the actual value; $\hat{y}_l$ is the predicted value.

- Mean Squared Error (MSE): This index calculates the square of the difference between the actual value and the predicted value to more severely penalize large errors.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_l\right)^2. \tag{3}$$

Where n is the number of observations; y_i is the actual value; $\hat{y}_l$ is the predicted value.

- R^2: Measures how well the model captures variation in the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \tag{4}$$

Where n is the number of observations; y_i is the actual value; $\hat{y}_l$ is the predicted value; $\bar{y}$ is the mean value of the dependent variable.

Additionally, a 95% confidence interval was calculated for each predicted value using the standard deviation of each tree prediction in the Random Forest model.

2.5 Predict future accidents

Once the models are trained and validated, they are used to predict daily deaths and injuries in 2024. A separate data set of future features is simulated (e.g., weather conditions, traffic flow) generated by randomly sampling values from the historical data set, ensuring that the distribution of these features remains consistent with observed patterns. The predictions are accompanied by confidence intervals, which measure the degree of uncertainty in the prediction. The prediction results, including predicted deaths and injuries and their corresponding confidence intervals, were saved to an Excel file for further analysis and interpretation.

2.6 Visualization and mapping

Geospatial analysis was conducted using predicted values of deaths and injuries, mapped against historical data. These maps, constructed using GIS software, identify high-risk areas and facilitate spatial comparisons with past accident hotspots. This visual representation provides a valuable tool for policymakers and stakeholders to plan targeted interventions.

3 Results

3.1 Performance metrics of the Spline regression model and the Random Forest regression model in predicting fatalities and injuries from traffic accidents

Table 1 show performance metric of the models in predicting deaths and injuries

Table 1 Performance metrics of the models in predicting deaths and injuries

Model	Metric	Deaths	Injuries
Spline regression	MAE	0.0288	0.2748
	MSE	0.0013	0.1305
	R^2	0.996	0.805
Random Forest	MAE	0.012072	0.036323
	MSE	0.000446	0.002306
	R^2	0.998663	0.996552

Evaluation of the Spline regression model showed superior performance in predicting mortality, with a mean absolute error (MAE) of 0.0288 and an R^2 value of 0.996, indicating that the model explains 99.6% of the variance in mortality. This result reflects the model's reliable accuracy in capturing mortality trends. The model predicted lower performance injuries with an MAE of 0.2748 and R^2 of 0.805, explaining 80.5% of the variance. The mean squared error (MSE) values confirm the model's reliability, especially in minimizing significant prediction errors.

The Random Forest model showed outstanding performance in predicting deaths and injuries with MAE values of 0.012072 for deaths and 0.036323 for injuries, indicating prediction errors. The MSE values are also very low, 0.000446 for deaths and 0.002306 for injuries. The results show that the model rarely makes large errors in predicting. Additionally, the R^2 values of 0.998663 for deaths and 0.996552 for injuries reflect that the model explains almost all the variation in the data. Of these, 99.87% of the variation was explained by deaths and 99.66% by injuries. These results prove the high accuracy and reliability of the model.

Table 2 provides detailed information on the models' performance in predicting deaths and injuries for training and testing datasets.

The spline regression model demonstrates excellent performance for mortality predictions. On the training dataset, the results are as follows: Mean Absolute Error (MAE) = 0.012447, Mean Squared Error (MSE) = 0.000483, and R^2 = 0.9986. Similarly, the model performs well on the test dataset, with MAE = 0.028747, MSE = 0.001324, and R^2 = 0.996035. These metrics indicate a tight fit and minimal prediction error, showing that the model captures nearly all the variance in the data, even for out of sample predictions. In contrast, the model performed well for injuries on the training set, with MAE = 0.035629, MSE = 0.002413, and R^2 = 0.996454. However, there was a significant decrease in performance on the test set, with MAE = 0.274753, MSE = 0.130498, and R^2 = 0.804929. While the model performs well on the training data, it exhibits signs of overfitting when tested on larger injury samples, indicating difficulty in generalizing. Overall, the Spline regression model shows good generalization for predicting deaths. However, it needs improvement in accurately predicting injuries, particularly for data that it did not encounter during training.

Table 2 Performance metrics of models for predicting deaths and injuries in training and testing datasets

Model	Target	Data set	MAE	MSE	R^2
Spile regression	Deaths	Train	0.012447	0.000483	0.9986
		Test	0.028747	0.001324	0.996035
	Injuries	Train	0.035629	0.002413	0.996454
		Test	0.274753	0.130498	0.804929
Random Forest	Death	Train	0.013154	0.000527	0.998481
		Test	0.012072	0.000446	0.998663
	Injuries	Train	0.036866	0.002447	0.996404
		Test	0.036323	0.002306	0.996552

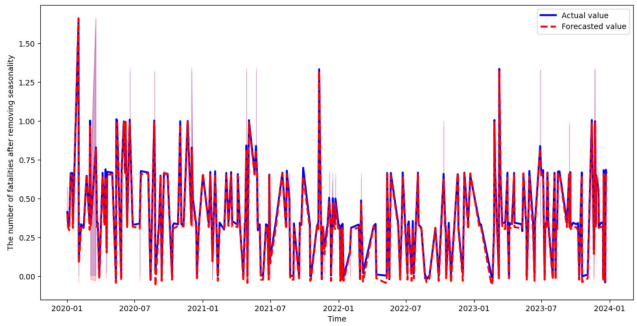
The Random Forest model showed superior performance in predicting deaths and injuries on both training and testing datasets. For fatalities, the MAE is 0.013154 on the training set and 0.012072 on the testing data set, with MSE values of 0.000527 and 0.000446, respectively. *R*-squared values of 0.998481 (training set) and 0.998663 (testing set) indicate an almost perfect model fit, explaining most of the variation in mortality outcomes. This consistency between the training set and testing set highlights the generalizability and stability of the model. For the injury case, the MAE was 0.036866 on the training dataset and 0.036323 on the testing dataset, with MSE values of 0.002447 and 0.002306, respectively. *R*-squared values of 0.996404 (training set) and 0.996552 (testing set) confirm that the model explains more than 99.6% of the variation in injury data, demonstrating predictive accuracy similarly robust on both datasets.

3.2 Predictions for fatalities and injuries from traffic accidents using Spline regression and Random Forest regression models.

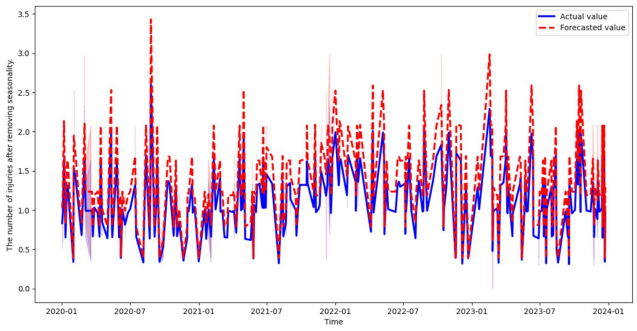
3.2.1 Predictions using Spline Regression model

Fig. 5 displays the forecast results from the Spline regression model. In Fig. 5 (a), we see a comparison of actual versus predicted death counts after removing seasonality. Similarly, Fig. 5 (b) presents the comparison for the number of injuries. These metrics enable us to evaluate the performance of the Spline regression model and highlight key areas for improvement.

Fig. 5 (a) demonstrates excellent accuracy in predicting deaths, as indicated by the close resemblance between actual



(a)



(b)

Fig.5 Comparison of actual versus predicted values for deaths (a) and injuries (b) after removing seasonality using a spline regression model

values (blue) and forecast values (red). The model exhibits strong predictive capability during periods of moderate mortality, with a mean absolute error (MAE) of 0.028747 and a mean squared error (MSE) of 0.001324. Fig. 5 (a), shown in Table 2, indicate minimal forecast error. Moreover, the consistently high R-squared (R^2) value of 0.996035 confirms that the model accounts for 99.6% of the variation in death rates. During certain periods, such as the sharp increase around mid-2023, the model exhibited minor deviations that led to a slight underestimation of the number of deaths. Although these deviations were small, they highlight a significant challenge in accurately assessing accidents involving multiple vehicles or dangerous situations, which affects the model's overall applicability. Despite these challenges, the model demonstrates high forecasting accuracy during both low and high mortality periods, making it a reliable tool for long-term predictions.

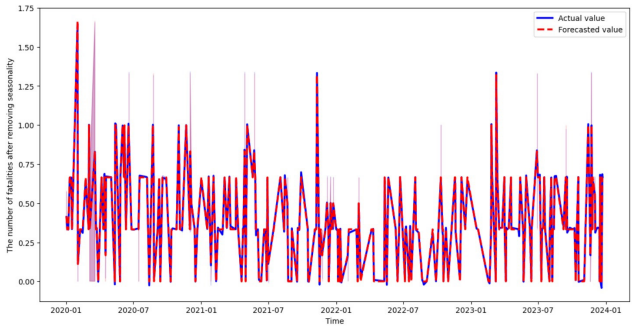
The forecast of injuries, as shown in Fig. 5 (b), reveals significant discrepancies between actual and predicted values. The Mean Absolute Error (MAE) for injuries is 0.274753, and the Mean Squared Error (MSE) is 0.130498. Both error metrics for Fig. 5 (b) are considerably higher than the corresponding mortality metrics for Fig. 5 (a) in Table 2. This suggests that injury data exhibit more complex variability, potentially influenced by factors such as road conditions, driver behavior, or the severity of accidents, which the model does not fully account for. The model effectively captures broad trends and seasonal patterns in injury data, as indicated by a relatively strong R^2 value of 0.804929, meaning it explains 80.5% of the variation in the dataset. However, the model struggles during periods of high injury rates, specifically in mid-2020 and late 2023, highlighting its challenge in generalizing to more unpredictable conditions. This could be attributed to the erratic nature of injury severity, where factors like speed, vehicle type, and weather conditions can significantly impact the outcomes of traffic accidents.

Thus, the Spline Regression model provides highly accurate predictions for deaths, with low forecast error and almost perfect fit, as shown in Fig. 5 (a). For injuries, although the model captures general trends, the higher MAE (0.274753) and higher MSE (0.130498) in Fig. 5 (b) suggest that injury prediction is challenging due to the complex and variable nature of injury data. To enhance the model's performance, incorporating additional explanatory variables and exploring hybrid models may improve results, particularly when predicting injuries in extreme or fluctuating conditions.

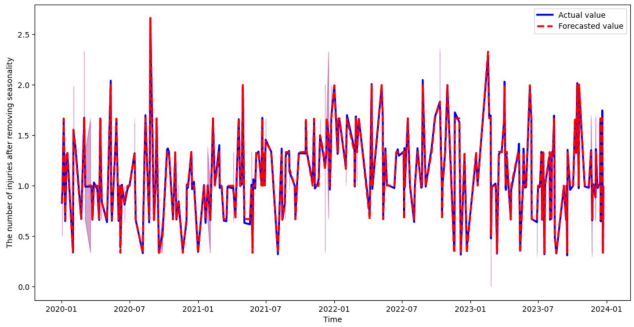
3.2.2 Predictions using Random Forest regression model

The Random Forest model demonstrates outstanding accuracy in predicting deaths Fig. 6 (a). Actual values (blue) and forecast values (red) nearly coincide throughout the time series, with slight deviations observed. The forecast effectively captured both high-frequency fluctuations and long-term trends in mortality data. The MAE (mean absolute error) value is 0.012072, and the MSE (mean squared error) value is 0.000446 (Table 2), showing the model's ability to minimize errors and make predictions with high precision.

Additionally, the high R^2 value of 0.998663 shows that the model explains almost completely the variation in mortality data. Even during mortality peak periods (e.g., mid-2023), the agreement between forecast and actual values highlights the stability and ability to handle mortality events at moderate and extreme levels of the model. The model's performance in predicting injuries, as shown in Fig. 6 (b), was also robust, with a high degree of agreement between actual and predicted values. The MAE value for injuries was 0.036323, and the MSE was 0.002306, a slightly higher level of bias than for deaths, but the overall results were still very accurate and consistent with the data. The R^2 value of 0.996552 demonstrates that the model explains 99.66% of the variation in injury data, demonstrating the model's superior predictive ability. While the model shows a very close fit during low-injury periods (Fig. 6 (b)), the model tends to underestimate peak periods slightly (e.g., late 2023). However, these



(a)



(b)

Fig.6 Comparison of actual versus predicted values for deaths (a) and injuries (b) after removing seasonality using a Random Forest model

deviations are still within a small range. The model effectively captured the overall trend and seasonality of injury cases and became a reliable tool for predicting injury rates over time.

It is found that the Random Forest model demonstrated superior performance in predicting both fatalities and injuries, with low error levels and high R^2 values, indicating its strong ability to capture the factors influencing traffic crash outcomes. Due to these outstanding results, the Random Forest model is used to predict daily traffic deaths and injuries in 2024.

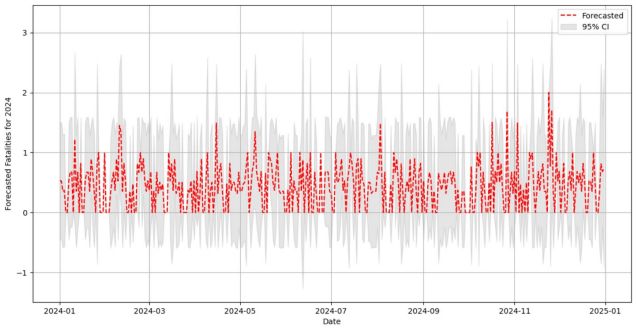
4 Daily prediction of traffic deaths and injuries in 2024 using the Random Forest model

The 2024 death prediction, illustrated in Fig. 7 (a), offers precise and detailed predictions of daily variations over time. The Random Forest model, which was optimized through hyperparameter tuning and temporal cross-validation, effectively captured the daily fluctuations in death counts. Most forecast values range from 0 to 2 deaths per day. However, significant peaks exceeding 2 deaths on certain days suggest periods of higher risk, potentially linked to known seasonal or temporal factors.

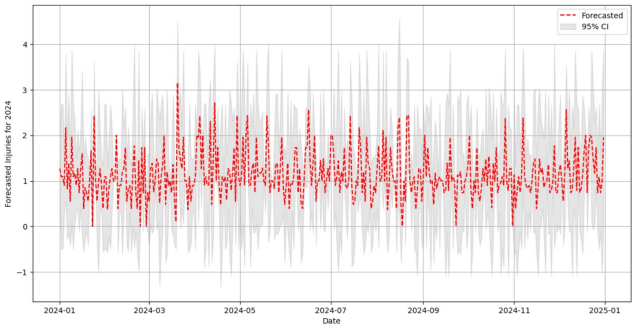
95% confidence intervals (CI) provide information about prediction uncertainty. Confidence intervals widen over some periods, indicating higher levels of unpredictability, while they narrow in other periods, showing higher levels of confidence. This volatility reflects the model's ability to adapt to changing conditions throughout the year while maintaining stability in its predictions. These results provide a strong foundation for understanding risk patterns over time and highlight the potential for preventative measures, focusing on identified periods of high risk.

The 2024 injury forecast, shown in Fig. 7 (b), shows a relatively stable trend, with most daily prediction values fluctuating between 1 and 2 injuries per day. However, the model also predicts higher peaks, especially in March and late November, which may be related to external factors such as seasonality or other influences that need further investigation. The model's 95% confidence intervals (CIs) reveal significant variability, with wider intervals indicating greater uncertainty in the predictions. This suggests that the model may have a limited ability to fully account for the complex interactions that affect injury rates over time. Nevertheless, the predictions demonstrate noticeable daily patterns, offering valuable insights into how injury numbers fluctuate throughout the year.

The wide confidence intervals indicate that further enhancements are necessary to improve forecast accuracy and reduce uncertainty. This can potentially be achieved by utilizing more detailed data or refining feature selection. The analysis underscores the model's effectiveness as a



(a)



(b)

Fig. 7 Prediction of daily traffic deaths (a) and injuries (b) for 2024

forecasting tool while also pinpointing areas that need optimization to better support critical decisions related to traffic injury prevention. With future improvements, the model could offer a more reliable and robust foundation for developing interventions aimed at preventing road traffic incidents.

The map in Fig. 8 (a) illustrates the number of traffic deaths forecast for 2024 in Thanh Hoa province using the Random Forest model. The model's predictions (indicated by red points) match actual traffic incidents from 2020 to 2023 (blue points). Notable overlap can be observed, especially in high-risk areas along major roads and national highways, demonstrating the model's ability to accurately identify at-risk regions where future accidents occur.

Predictions focus mainly on urban centres and major transportation routes with higher traffic density and accident risk. However, the model also points to emerging risk areas in remote and less populated areas. This result emphasizes the stability and robustness of the Random Forest model in predicting traffic fatalities and its potential to support traffic management and policy interventions. The strong correlation between historical data and forecast results increases the reliability of the Random Forest model in predicting traffic fatalities. Additional analysis is essential for areas where death projections are available but no corresponding historical information.

The map in Fig. 8 (b) shows the prediction number of traffic injuries in 2024 in Thanh Hoa province by using the Random Forest model. The forecast locations (yellow points) match traffic incidents recorded between 2020 and 2023 (blue points), showing significant overlap, especially on the main roads and national highways. High-risk areas are identified in urban centers and essential traffic routes, where congestion and traffic accidents frequently occur.

The model also points to emerging injury risk areas, especially in less populated areas. The close match between forecast data and historical data demonstrates the reliability of the Random Forest model in predicting the number of traffic injuries. These results can provide valuable information to support traffic management, helping local authorities focus on high-risk areas to implement preventive measures and improve traffic safety. However, further research is needed to clarify discrepancies between predicted injuries and historical records in certain regions.

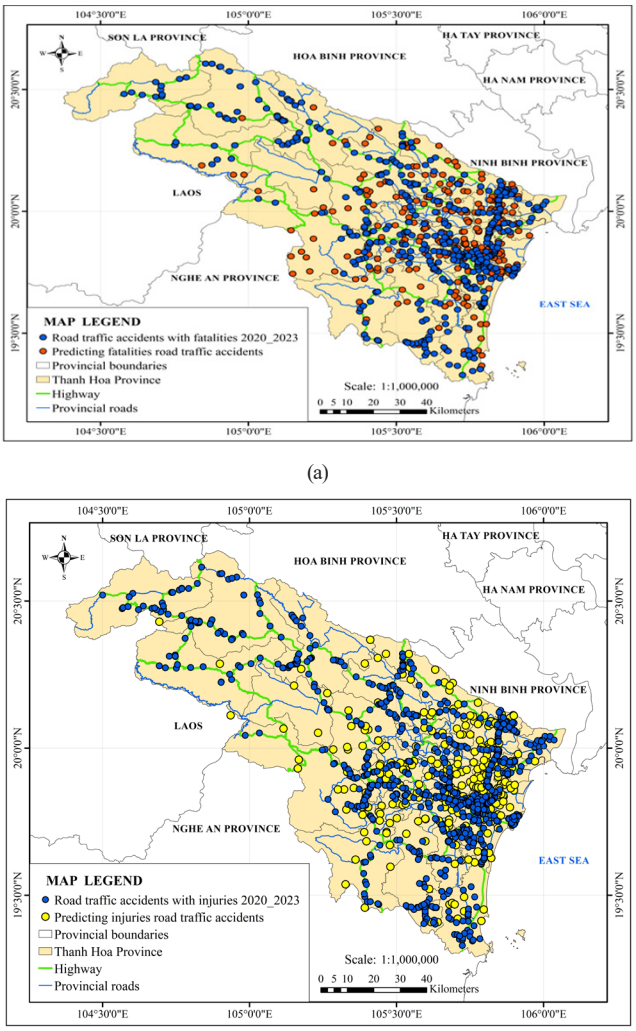


Fig. 8 Map predicting the number of deaths due to traffic accidents (a) and the number of injuries (b) in Thanh Hoa province using the Random Forest model

5 Discussion

This study evaluates the predictive ability of Spline Regression and Random Forest models to predict the number of deaths and injuries due to road traffic accidents in Thanh Hoa province. Also, it provides details of performance analysis and comparisons with existing documents. The results show that the Random Forest model has outstanding predictive power, especially in predicting the number of injuries, which is consistent with previous studies that have used machine learning to predict events. traffic problem.

While effective in capturing long-term trends in traffic fatalities, the Spline Regression model is ineffective when applied to highly variable outcomes such as injury cases. This difference is consistent with the results of Kazi Redwan Shabab et al. (2024) (Shabab et al., 2024), who showed that Spline-based models often have difficulty in highly variable contexts highly dynamic, where non-linear factors such as sudden weather changes, traffic surges, or human behaviors can impact forecast results in unpredictable ways. The relatively poor performance of the Spline Regression model in predicting the number of injuries (with mean absolute error (MAE)

significantly higher than expected) highlights the limitations of this model in handling highly variable data, especially with non-linear outcomes such as injury severity.

In contrast, the Random Forest model better predicted deaths and injuries. The model achieved an MAE of 0.012072, an R^2 of 0.998663 for deaths, and an MAE of 0.036323 with an R^2 of 0.996552 for injuries. These results surpass the Spline Re regression model and outperform other studies. For example, Tejashwini AG et al. (2024) (Tejashwini et al., 2024) reported an R^2 of 0.868 using a Random Forest model with a similar dataset, which is lower than the results I achieved. The higher accuracy is because we integrate lagged variables and seasonal adjustments, which helps the model capture temporal dependencies and reduces the risk of overfitting. This approach is consistent with Tejashwini et al. (2024), the authors found that the use of lagged variables significantly improved the ability to predict safety-related outcomes road by capturing the delayed effects of traffic density and external factors such as weather.

Incorporating lagged predictor variables is essential in improving the prediction accuracy of Random Forest models, especially for injuries. Random forests' ability to model complex, time-dependent interactions between factors such as vehicle density, road surface conditions, and weather events is a significant advantage over other models, such as linear models like Spline Regression. The research results coincide with the findings of the authors Khanum et al. (2023), Scarano et al. (2023), Sufian et al. (2024), Taamneh and Taamneh, (2019), found that using time-lagged variables significantly improved injury predictions by capturing the delayed effects of traffic congestion and the behavior of drivers. By modelling these temporal dependencies, the model produced more accurate predictions during high-risk periods, such as March and November, when accidents often increase sharply due to seasonal fluctuations in traffic density and weather conditions. Similar patterns of seasonal fluctuations were also noted by Briand et al. (2022) and Sufian et al. (2024), the authors suggested that these increases were often related to holiday travel traffic and damaged road surface conditions.

Spatial analysis provided through maps forecasting the number of deaths and injuries in 2024 highlights high-risk areas along national roads and provincial roads, especially in urban centers and near important transportation routes.

Fig. 8 (a) illustrates the concentration of predicted traffic fatalities. With red points overlapping green points representing historical incidents showing a high degree of concordance between historical data and future predictions. This is consistent with the study of Chaudhuri et al., (2023), the authors emphasized that deaths due to traffic accidents are often concentrated in corridors with dense traffic flow. Similarly, Fig. 8 (b) depicts the predicted distribution of the number of injuries with a wider dispersion, reflecting the inherent unpredictability of injury outcomes. This wide distribution is consistent with the study of Zhou et al. (2020), the authors observed that injuries are likely to occur in urban and rural areas due to many factors such as traffic volume, vehicle speed and human behavior. The high overlap between historical and forecast data in both maps emphasizes the robustness of the Random Forest model in capturing spatial and temporal patterns of traffic accident incidents. The results of this study have important implications for traffic safety management in Thanh Hoa province and similar areas. By identifying high-risk periods and locations, predictive models can support specific interventions, such as increased traffic enforcement or road maintenance during peak accidents, especially on holidays or when weather conditions are adverse. Furthermore, information from this study can help guide traffic management strategies, such as adjusting traffic signal timing or applying speed limits in high-risk areas. However, to exploit the full potential of machine learning in traffic safety, future models need to integrate more detailed sociological data, such as driver age or type of vehicle, along with real-time environmental factors to refine predictions and deliver actionable information more effectively.

6 Conclusion

This study has shown the effectiveness of employing machine learning in conjunction with GIS to capture the non-linear and time-dependent interactions in traffic accident data, which helps predict deaths and injuries resulting from road traffic accidents. The Random Forest model demonstrated strong performance, with high mean absolute error (MAE) and R^2 values, demonstrating its reliability in predicting deaths and injuries. Model enhancement through the integration of lagged variables and adjustment for seasonal effects further improved daily forecast accuracy. Key findings from the analysis highlight several factors significantly influencing traffic accident outcomes in Vietnam's Thanh Hoa province, which has diverse road conditions, including complex topography and quality, heavy traffic density, lighting conditions, and highly variable weather. This information provides valuable guidance to policymakers, opening opportunities to implement proactive traffic safety measures. Predicting accident outcomes daily allows authorities to deploy strategic resources during high-risk periods, such as holidays or bad weather, contributing to improved management and overall traffic safety.

Acknowledgement

Thanks to the staff of the Thanh Hoa Department of Natural Resources and Environment and the Thanh Hoa Provincial Traffic Police Department for their generous time, materials, suggestions, and assistance throughout.

References

Ahmed, S. K., Mohammed, M. G., Abdulqadir, S. O., El-Kader, R. G. A., El-Shall, N. A., Chandran, D., Ur Rehman, M.E., Dhama, K. (2023) "Road traffic accidental injuries and deaths: A neglected global health issue", Health Science Reports, 6(5), e1240. https://doi.org/10.1002/hsr2.1240

Ajao, I. O., Adaraniwon, A. O., Ilugbusi, A. O. (2023) "Beyond Linearity: Harnessing Spline Regression Models to Capture Non-linear Relationships", Ilorin Journal of Science, 10(1), pp. 28–43. https://doi.org/10.54908/iljs.2023.10.01.003

Alok, K., Kamalraj.R, (2024) "Machine Learning in Transportation", International Journal of Innovative Research in Computer &Communication Engineering, 12(5), pp. 6036–6040. https://doi.org/10.15680/IJIRCCE.2024.1205151

Azami, M. F. A. M., Misro, M. Y., Hamidun, R. (2024) "Road crash dynamics in Malaysia: Analysis of trends and patterns", Heliyon, 10(18), e37457. https://doi.org/10.1016/j.heliyon.2024.e37457

Bhele, R., Subedi, A., Thapa, B., Rajchal, P. (2024) "Exploring the Efficiency of Random Forest, Support Vector Machine, and Artificial Neural Network in Traffic Crash Data", In: 2024 International Conference on Inventive Computation Technologies (ICICT). IEEE, pp. 12–17. ISBN 979-8-3503-5929-9 https://doi.org/10.1109/ICICT60155.2024.10544787

Briand, J., Deguire, S., Gaudet, S., Bieuzen, F. (2022) "Monitoring Variables Influence on Random Forest Models to Forecast Injuries in Short-Track Speed Skating", Frontiers in sports and active living, 4. https://doi.org/10.3389/fspor.2022.896828

Chaudhuri, S., Juan, P., Mateu, J. (2023) "Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation", Journal of Applied Statistics, 50(16), pp. 3229–3250. https://doi.org/10.1080/02664763.2022.2104822

Fahad, K., Joarder, M. F., Nahid, M., Tasnim, T. (2023) "Road Accidents Severity Prediction Using a Voting-Based Ensemble ML Model", Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning, Springer, pp. 793–808. ISBN 978-981-99-8936-2 https://doi.org/10.1007/978-981-99-8937-9_53

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), pp. 16–26. https://doi.org/10.52939/ijg.v20i9.3537

Isler, C. A., Huang, Y., de Melo, L. E. A. (2024) "Developing accident frequency prediction models for urban roads: A case study in São Paulo, Brazil", IATSS Research, 48(3), pp. 378–392. https://doi.org/10.1016/j.iatssr.2024.07.002

Khanum, H., Garg, A., Faheem, M. I. (2023) "Accident severity prediction modeling for road safety using random forest algorithm: an analysis of Indian highways", F1000 Research, 12:494. https://doi.org/10.12688/f1000research.133594.2

Meganfi, A., Chamidah, N., Sediono, S., Anna, E. (2023) "Modelling the number of traffic accident using negative binomial regression spline", AIP Conference Proceedings, 2554(1),030017. https://doi.org/10.1063/5.0104855

Pelicic, D., Ristic B., Radevic S., (2024) "Epidemiology of traffic traumatism", Sanamed, 19(2), pp. 233–236. http://dx.doi.org/10.5937/sanamed0-51914

Phaphan, W., Sangnuch, N., Piladaeng, J. (2023) "Comparison of the Effectiveness of Regression Models for the Number of Road Accident Injuries", Science & Technology Asia, 28(4), pp. 54–66. https://doi.org/10.14456/scitechasia.2023.71

Prasetiyowati, M. I., Maulidevi, N. U., Surendro, K. (2020) "The Speed and Accuracy Evaluation of Random Forest Performance by Selecting Features in the Transformation Data", In: Proceedings of the 2020 The 9th International Conference on Informatics, Environment, Energy and Applications, pp. 125–130. ISBN 9781450376891 https://doi.org/10.1145/3386762.3386768

Scarano, A., Riccardi, M. R., Mauriello, F., D'Agostino, C., Pasquino, N., Montella, A. (2023) "Injury severity prediction of cyclist crashes using random forests and random parameters logit models", Accident Analysis & Prevention, 192, 107275. https://doi.org/10.1016/j.aap.2023.107275

Schuster, N. A., Rijnhart, J. J. M., Twisk, J. W. R., Heymans, M. W. (2022) "Modeling non-linear relationships in epidemiological data: The application and interpretation of spline models", Frontiers in Epidemiology, 2, 975380. https:/doi.org/10.3389/fepid.2022.975380

Shabab, K. R., Bhowmik, T., Zaki, M. H., Eluru, N. (2024) "A systematic unified approach for addressing temporal instability in road safety analysis", Analytic Methods in Accident Research, 43, 100335. http://dx.doi.org/10.1016/j.amar.2024.100335

Shelevytsky, I., Shelevytska, V., Semenova, K., Bykov, I. (2020) "Regression Spline-Model in Machine Learning for Signal Prediction and Parameterization", In: Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2019, pp. 158–174. ISBN 978-3-030-26473-4 http://dx.doi.org/10.1007/978-3-030-26474-1_12

Das, S., Bandyopadhyay, T., Mukherjee, S., Acharyya, S. (2024) "Revolutionizing Traffic Management: Advanced Machine Learning Techniques for Accurate Traffic Flow Prediction", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 4(6), pp. 313–316. http://dx.doi.org/10.48175/IJARSCT-17649

Sufian, M. A., Varadarajan, J., Niu, M. (2024) " Enhancing prediction and analysis of UK road traffic accident severity using AI: Integration of machine learning, econometric techniques, and time series forecasting in public health research", Heliyon, 10(7), e28547. http://dx.doi.org/10.1016/j.heliyon.2024.e28547

Taamneh, S., Taamneh, M. (2019) "Evaluation of the Performance of Random Forests Technique in Predicting the Severity of Road Traffic Accidents", In: Proceedings of the AHFE 2018 International Conference on Human Factors in Transportation, Springer, pp. 840–847. ISBN 978-3-319-93884-4
http://dx.doi.org/10.1007/978-3-319-93885-1_78

Tejashwini A G, Singh, C. K., Sinha, C., Karnani, D., Pandey, A. (2024) "Road Accident Severity Prediction Using Random Forest Algorithm", International Journal for Research in Applied Science & Engineering Technology (IJRASET), 12(4), pp. 1205–1210.
http://dx.doi.org/10.22214/ijraset.2024.59991

Ulu M., Kilic, E., Türkan, Y. S. (2024) "Prediction of Traffic Incident Locations with a Geohash-Based Model Using Machine Learning Algorithms", Applied Sciences, 14(2), 725.
https://doi.org/10.3390/app14020725

Venugopal, Y. R, Dr. Srikanth, V. (2024) "Ensemble Learning Approaches for Improved Predictive Analytics in Healthcare", International Journal of Research Publication and Reviews, 5(3), pp. 757–760.
http://dx.doi.org/10.55248/gengpi.5.0324.0629

Wang, Y., Zhai, H., Cao, X., Geng, X. (2024) "A Novel Accident Duration Prediction Method Based on a Conditional Table Generative Adversarial Network and Transformer", Sustainability, 16(16), 6821.
https://doi.org/10.3390/su16166821

Yu, Z., Yang, J., Huang, H. H. (2024) "Smoothing regression and impact measures for accidents of traffic flows", Journal of Applied Statistics, 51(6), pp. 1041–1056.
https://doi.org/10.1080/02664763.2023.2175799

Zhou, Z., Wang, Y., Xie, X., Chen, L., Zhu, C. (2020) "Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective", IEEE Transactions on Knowledge and Data Engineering, 34(8), pp. 3786–3799.
http://dx.doi.org/10.1109/TKDE.2020.3034312