Integrating AI Algorithms and GIS Technology for Traffic Accident Prediction: Towards Safer Roads Through Advanced Modeling

Khanh Giang LE [a,*], Quang Hoc TRAN [a]

[a] *Faculty of Civil Engineering, University of Transport and Communications, Hanoi, Vietnam*
[*] *E-mail: gianglk@utc.edu.vn (Corresponding author)*

Abstract:
This study explores the integration of Artificial Intelligence (AI) algorithms and Geographic Information System (GIS) technology for traffic accident prediction and prevention. Conventional statistical methods often fail to capture the spatial and temporal complexity of accident occurrences, while AI and GIS offer a more comprehensive and data-driven approach. By combining AI's predictive capabilities with GIS's spatial analysis, the authors develop a robust framework for identifying accident hotspots and assessing risk levels. This study employs optimized Support Vector Machines (SVMs) enhanced with the Sparrow Search Algorithm (SSA) for superior predictive performance. Kernel Density Estimation (KDE) is applied to refine spatial clustering of high-risk locations. A case study in the United Kingdom demonstrates the effectiveness of this approach, showing significant improvements in predictive accuracy and actionable insights for traffic safety planning. The results emphasize the need for real-time data integration and ongoing refinement of AI-GIS models to further enhance road safety strategies.

Keywords: Traffic Accident Prediction, Artificial Intelligence (AI), Machine Learning (ML), Geographic Information System (GIS), Road Safety.

1. Introduction

Road traffic accidents continue to be a major global public health issue. According to the World Health Organization (2023), approximately 1.19 million fatalities and 50 million injuries occur annually, with the burden most acutely felt in low- and middle-income countries [1]. In addition to the tragic loss of life, traffic accidents impose significant economic costs and disrupt the efficiency of urban transportation networks.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have transformed predictive analytics in various fields, including traffic safety. By harnessing deep learning (DL) and other advanced algorithms, researchers and transportation agencies can analyze vast and complex datasets—ranging from historical crash records to real-time environmental data—to identify risk factors and forecast high-risk locations. However, traditional statistical methods struggle to capture the spatial complexity of traffic accident data [2].

Geographic Information System (GIS) technology plays a crucial role by providing a robust platform for spatial data integration and visualization. When combined with AI, GIS not only supports the cleaning and enrichment of large datasets but also allows for detailed mapping and analysis of accident-prone regions. This paper proposes a comprehensive framework that integrates AI and GIS, thereby offering a data-driven approach for effective traffic accident prediction and prevention.

The remainder of this paper is structured as follows: Section 2 reviews previous studies on AI, ML, and GIS applications in traffic safety. Section 3 describes the methodology and data sources used in this study. Section 4 presents the results and discusses their implications. Finally, Section 5 concludes the paper, identifies limitations, and suggests directions for future research.

2. Literature Review

Integrating AI, ML, and GIS technologies has emerged as a promising solution to overcome the limitations of traditional accident analysis methods. Prior research has demonstrated that AI-driven models, such as Random Forests and Support Vector Machines (SVMs), can effectively predict accident severity and frequency. However, these studies often treat spatial information as an ancillary component rather than an integrated part of the analytical framework [3-5].

On the other hand, GIS excels in spatial data management, allowing for the consolidation of diverse datasets—including traffic volumes, road geometries, and meteorological conditions—into a unified system. This integration supports root-cause analysis and the precise mapping of high-risk areas [6-10]. Each technology has its own strengths; when GIS and AI are combined, their power is maximized, especially in the field of transportation in general and traffic accident analysis in particular, as big data is fully utilized. Key benefits of AI and GIS integration include:

- Enhanced data management: GIS ensures accurate geocoding and seamless integration of spatial data, which is critical for high-quality predictions.
- Advanced analytical capabilities: AI algorithms uncover hidden patterns in large datasets, while GIS provides spatial context, improving model interpretability.
- Real-time processing: The combined system can process data in real time, enabling timely interventions.
- Improved mapping accuracy: GIS-driven methods yield highly detailed risk maps that inform targeted safety measures.

Although several studies have explored the integration of GIS and machine learning (ML) in traffic accident analysis, most have been limited to clustering techniques and association rule mining to identify key accident-related factors [11]. One such study [12] employed ML models in R Studio and ArcGIS to predict accident severity. While challenges such as technology costs, data quality, and infrastructure deployment persist, the integration of AI and GIS presents a highly promising approach to improving traffic safety. By leveraging the strengths of both technologies, this combination has the potential to drive significant advancements in traffic safety management, ultimately reducing accidents and saving lives.

Despite this potential, few studies have fully harnessed the synergistic benefits of AI-GIS integration. To address this gap, this research develops a comprehensive framework that combines advanced AI algorithms with GIS-based spatial analysis, enhancing the accuracy and reliability of accident prediction.

3. Methodology and Study Data

3.1. Methodology

The integration of AI and GIS has revolutionized the approach to predicting and preventing traffic accidents. An effective AI-GIS accident prediction system comprises several key components that work in tandem to provide accurate and actionable insights. These components include a data collection module, a geospatial analysis engine, a machine learning core, and a visualization interface.

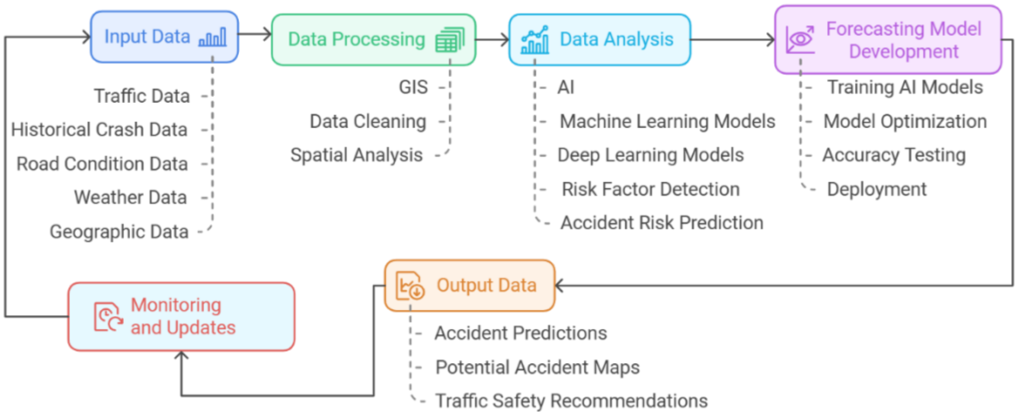

Figure 1. Diagram of the traffic accident prediction analysis model based on AI and GIS

Figure 1 illustrates the methodological framework for traffic accident prediction and safety enhancement. The process starts with input data acquisition, which includes various sources such as traffic, historical crash, road condition, weather, and geographic data, etc.

Next, data preparation and enrichment are performed. The collected data undergoes rigorous cleaning and validation to remove errors, duplicates, and inconsistencies. GIS tools are then used for feature extraction—deriving metrics such as road curvature, proximity to intersections, and population density—and for temporal-spatial aggregation. This stage ensures that the data is both comprehensive and structured for analysis.

The next steps are data analysis and forecasting model development. This leads to data analysis, where AI, ML, and deep learning models detect risk factors and predict accident risks. Following analysis, the forecasting model development stage involves training AI models, optimizing, testing accuracy, and deploying them. The framework employs advanced ML algorithms, including optimized SVM models. Various heuristic optimization techniques (e.g., Sparrow Search Algorithm) are applied to enhance the SVM's performance. Concurrently, spatial analysis models (such as the Kernel Density Estimation) are utilized to capture the influence of neighboring regions on accident occurrences.

The output data generated includes accident predictions, accident risk maps, and traffic safety recommendations. Finally, a monitoring and updates process ensures continuous refinement, improving model performance and adapting to new data.

This methodological flow supports a comprehensive approach to analyzing and forecasting traffic risks and enhancing safety recommendations based on data-driven insights.

3.1.1. Kernel Density Estimation Model

Kernel Density Estimation is a non-parametric method used to estimate the probability density function of a random variable based on a given dataset. The general formula for the kernel density function is [13]:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \qquad (1)$$

where:

$f(x)$ is the estimated density at point x.

x_i represents observed data points.

n is the number of observed data points.

h is the bandwidth, a parameter that controls the smoothness of the density estimate.

K is the kernel function, which determines the weighting of data points based on their distance from x.

In this study, Epanechnikov kernel is selected to determine the weighting of data points based on their distance from x. The general formula for the Epanechnikov kernel is:

$$K(u) = \frac{3}{4}(1 - u^2), if\ |u| \le 1, otherwise\ K(u) = 0 \qquad (2)$$

where:

u is a normalized variable, defined as follows:

$$u = \frac{x - x_i}{h} \qquad (3)$$

3.1.2. Support Vector Machines (SVMs) Algorithms in Risk Assessment

Support Vector Machines (SVMs) are powerful supervised learning algorithms used in classification and regression tasks. In risk assessment, SVMs are particularly useful for identifying high-risk areas, predicting potential hazards, and classifying risk levels based on various factors. The fundamental idea behind SVMs is to find an optimal hyperplane that separates different classes in a given dataset with the maximum margin [14].

The goal of SVM is to find a decision boundary (hyperplane) defined as:

$$w^T x + b = 0 \qquad (4)$$

where w is the weight vector, x is the feature vector, b is the bias term.

SVM objective function: The goal is to optimize the margin between classes and the hyperplane. The objective function is:

$$min\frac{1}{2} \parallel \mathrm{w} \parallel^2 \qquad (5)$$

subject to:

$$y_i(w^T x + b) \quad \ge 1, i = 1, \dots, n.$$

This ensures that all training samples are correctly classified with a margin of at least 1.

3.1.3. Sparrow Search Algorithm (SSA) for SVM

The Sparrow Search Algorithm (SSA) is a metaheuristic optimization technique inspired by the foraging and anti-predation behavior of sparrows. It has been widely used to optimize machine learning models, including SVMs, by fine-tuning hyperparameters such as the regularization parameter (C) and kernel function parameters (e.g., γ in RBF kernel) to improve classification accuracy. The formula for updating a sparrow's position is [15]:

$$x_i^{t+1} = x_i^t * e^{\left(-\frac{i}{max\_iter}\right)} \qquad (6)$$

where:

x_i is position of sparrow i;

t: iteration number;

max_iter: maximum number of iterations.

SSA-SVM optimization process: The SSA is applied to optimize SVM hyperparameters by:

- Encoding Hyperparameters: Each sparrow represents a candidate solution (C, γ, etc.) for SVM training.
- Evaluating Fitness: The classification accuracy of an SVM model trained with the candidate hyperparameters is used as the fitness function:

$$f(X) = \text{SVM Accuracy}(\mathrm{C}, \gamma) \qquad (7)$$

- Updating Positions: The SSA updates sparrow positions iteratively using producer, scrounger, and predator phases.

- Convergence Check: The algorithm terminates when a stopping criterion is met (e.g., maximum iterations or fitness improvement threshold).
- Best Solution Selection: The best sparrow (hyperparameter set) is used to train the final SVM model.

The SSA efficiently optimizes SVM hyperparameters by balancing exploration and exploitation. By fine-tuning SVM parameters, SSA enhances classification accuracy, making it a valuable approach for risk assessment, anomaly detection, and predictive modeling.

3.2. Study Area and Data

The study area is the United Kingdom (UK). The UK covers a total area of about 244,376 km², with 242,741 km² being land. Its population stands at 67,837,434. The country has a radial road network comprising 46,904 km of main roads, 3,497 km of motorways, and 344,000 km of paved roads. As of 2022, there were 40.8 million licensed vehicles in the UK.

In 2022, the UK recorded 1,711 road fatalities, a 2% drop from 2019, and 29,742 cases of death or serious injury, down 3%. Total casualties of all severities were 135,480, a 12% decrease. Vehicle travel returned to pre-COVID levels, with 328 billion miles traveled, but fatalities per billion miles rose by 2% to 5%.

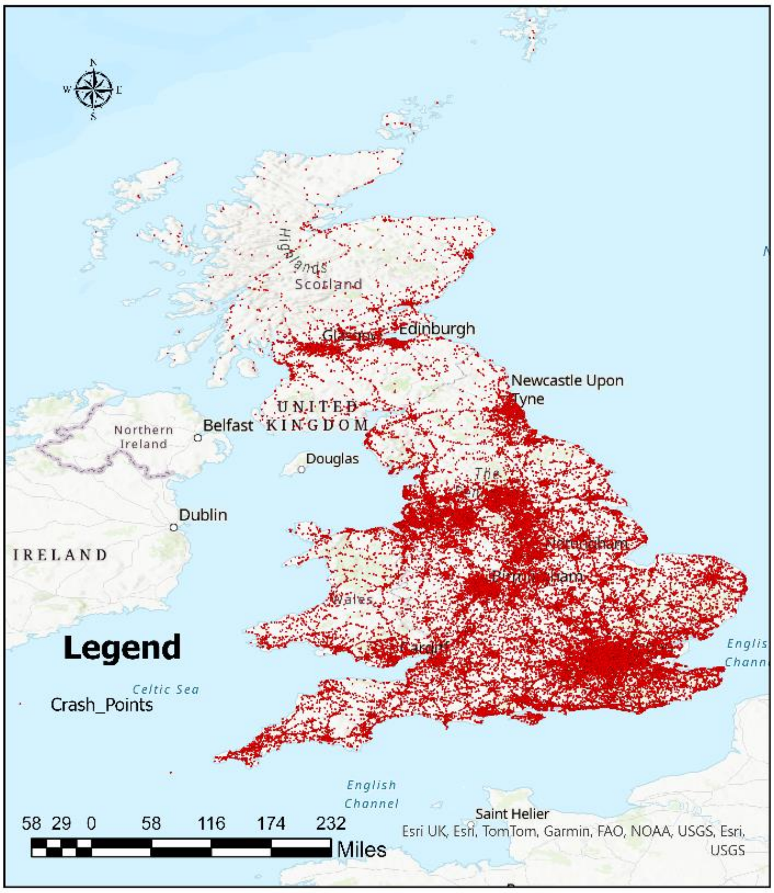

Figure 2. The locations of traffic crashes in the UK in 2022 displayed on the GIS map.

The traffic accident data used for analysis was collected from January to December 2022, with a total of 106,005 accidents, including full spatial information (such as accident location, as shown in Figure 2) and attributes (such as date, time, cause, vehicle type, accident type, etc., as shown in Table 1), all provided via the website. Additionally, the road network data was sourced from the OpenStreetMap website, including both spatial and attribute data of the road network.

Table 1. The sample of study data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	accident_index	accident_year	accident_reference	location_easting_osgr	location_northing_osgr	longitude	latitude	police_force	accident_severity	number_of_vehicles	number_of_casualties	date	day_of_week	time
2	2.02E+12	2022	10352073	525199	177928	-0.198224	51.486454	1	3	2	1	05-01-22	4	16:40
3	2.02E+12	2022	10352573	546214	179866	0.105042	51.49883	1	3	2	1	01-01-22	7	1:17
4	2.02E+12	2022	10352575	551119	174789	0.173482	51.451924	1	3	2	1	01-01-22	7	1:15
5	2.02E+12	2022	10352578	528889	192230	-0.139873	51.614153	1	3	2	2	01-01-22	7	2:24
6	2.02E+12	2022	10352580	539773	190404	0.016495	51.595151	1	3	4	3	01-01-22	7	2:30
7	2.02E+12	2022	10352588	543159	181261	0.061626	51.512146	1	2	1	5	01-01-22	7	2:55
8	2.02E+12	2022	10352591	535435	168744	-0.0544	51.401567	1	3	2	1	01-01-22	7	4:20
9	2.02E+12	2022	10352594	546736	189194	0.116442	51.58251	1	3	4	4	01-01-22	7	2:19
10	2.02E+12	2022	10352596	539501	178262	0.007763	51.486112	1	3	1	1	01-01-22	7	5:05
11	2.02E+12	2022	10352600	534597	190727	-0.05806	51.599313	1	2	1	1	01-01-22	7	2:15
12	2.02E+12	2022	10352601	532694	183413	-0.088278	51.534038	1	3	1	1	01-01-22	7	2:10
13	2.02E+12	2022	10352605	534094	182549	-0.068434	51.525943	1	2	1	1	01-01-22	7	1:13
14	2.02E+12	2022	10352606	546301	179840	0.106283	51.498574	1	2	2	3	01-01-22	7	3:51
15	2.02E+12	2022	10352610	508766	174030	-0.435995	51.454815	1	2	1	2	01-01-22	7	5:45
16	2.02E+12	2022	10352611	509531	182582	-0.422338	51.531533	1	3	3	4	01-01-22	7	7:40
17	2.02E+12	2022	10352612	528471	189140	-0.147037	51.58648	1	3	2	1	01-01-22	7	10:45
18	2.02E+12	2022	10352613	533304	178605	-0.081305	51.490688	1	3	2	1	01-01-22	7	6:40
19	2.02E+12	2022	10352615	529403	177887	-0.137725	51.48514	1	3	2	1	01-01-22	7	4:49
20	2.02E+12	2022	10352616	538111	177157	-0.012675	51.476523	1	3	4	1	01-01-22	7	9:05

3.2.1. Data Preparation and Enrichment

An efficient AI-GIS accident prediction system is built on the foundation of data preparation and enrichment. To guarantee the accuracy, consistency, and applicability of the data utilized in the model, this critical stage entails a number of tasks.

3.2.1.1. Data Cleaning and Validation

The initial phase of data preparation involves cleaning and validating the data. This entails the removal or modification of data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. Data cleaning can be a time-intensive process, often consuming up to 45% of a data scientist's time. It includes correcting spelling and syntax errors, standardizing datasets, rectifying issues like empty fields, and identifying duplicate entries.

To maintain data quality, a systematic approach is required, which involves:

1) Eliminating duplicate observations and irrelevant data
2) Filtering out unwanted outliers
3) Correcting structural errors, such as inconsistent naming conventions or improper capitalization
4) Addressing any missing data
5) Validating the cleaned dataset

For machine learning applications, a dataset should ideally contain at least 1,000 rows and 5 columns, with the first column designated as an identifier. The data should be consolidated into a single file or table, with minimal missing values and no personally identifiable information.

3.2.1.2. Feature Extraction from GIS

GIS is essential for extracting significant features for accident prediction models. This involves generating new metrics such as distance, proximity, and density, which can be fed into machine learning algorithms. Techniques like spatial clustering and spatial interpolation are commonly used to create these features.

To effectively integrate geospatial data, a strong framework is necessary to unify information from various sources, including GPS, satellite imagery, and social media. GIS tools excel in this area, allowing for the consolidation and alignment of disparate datasets into a comprehensive geospatial database.

Key static features obtained from GIS data include:

1) Road geometry (e.g., curvature)
2) Speed limits
3) Population density

4) Road orientation
5) Sinuosity
6) Traffic signs

3.2.1.3. Temporal and Spatial Aggregation

Temporal and spatial aggregation are critical for preparing data for accident prediction models. Temporal aggregation involves grouping data into specific time frames, such as hourly, daily, weekly, or monthly intervals, enabling the model to identify time-dependent patterns in accident occurrences. In contrast, spatial aggregation groups data based on geographic areas, such as countries, cities, or states, which helps identify spatial patterns and accident hotspots.

Incorporating both spatial and temporal dimensions is vital for models reliant on dynamic location data. Temporal-spatial models that consider changes over time and space offer a more comprehensive context for analysis. These models can incorporate dynamic features such as:

1) Weather conditions
2) Traffic volumes
3) Time variables (hour, month, day)

To manage the extensive geospatial datasets needed for accident prediction, efficient data management strategies are essential. This may involve using optimized storage solutions and cloud-based platforms. Additionally, employing spatial indexing techniques, like R-trees or Quadtrees, can enhance the speed of spatial queries and minimize computational overhead.

By implementing these data preparation and enrichment steps, organizations can establish a solid foundation for their AI-GIS accident prediction models, leading to more accurate and reliable outcomes.

4. Results and Discussion

Table 2 compares the performance of the standard SVM model and its optimized version (SSA-SVM). Each model demonstrates significant differences in effectiveness, evaluated through metrics such as Accuracy, F1 Score, Precision, Recall, and Area Under the Curve (AUC).

Table 2. Comparison of prediction results between AI algorithms.

Model	SVM	SSA-SVM
Accuracy	0.76	0.87
F1 Score	0.66	0.87
Precision	0.66	0.94
Recall	0.76	0.86
AUC	0.74	0.89

The standard SVM model exhibits the lowest performance among all models. Specifically, its accuracy is only 0.76, with an F1 Score and Precision of 0.66, indicating weak classification capability, especially in balancing Precision and Recall. The AUC for standard SVM is 0.74, reflecting average performance in distinguishing between data classes.

In contrast, the SSA-SVM model, which employs the Sparrow Search Algorithm, is the best-performing model in the table. With accuracy and F1 Score both at 0.87 and an outstanding Precision of 0.94, SSA-SVM showcases superior classification capability, particularly in minimizing false positives. Its AUC of 0.89, the highest among all models, indicates excellent ability to differentiate between classes. These findings highlight the potential of optimized SVM models to improve the accuracy and reliability of traffic accident risk assessments.

Crucially, these AI-generated predictions were validated through GIS-based spatial analysis. By overlaying the predicted accident hotspots onto historical accident density maps, we observed a high degree of spatial congruence. This validation confirms that the integration of GIS with AI not only improves predictive performance but also enhances the interpretability and practical utility of the results. Figure 3 illustrates the classification of accident risk levels, ranging from low to very high, providing a clear visual tool for traffic managers and urban planners.

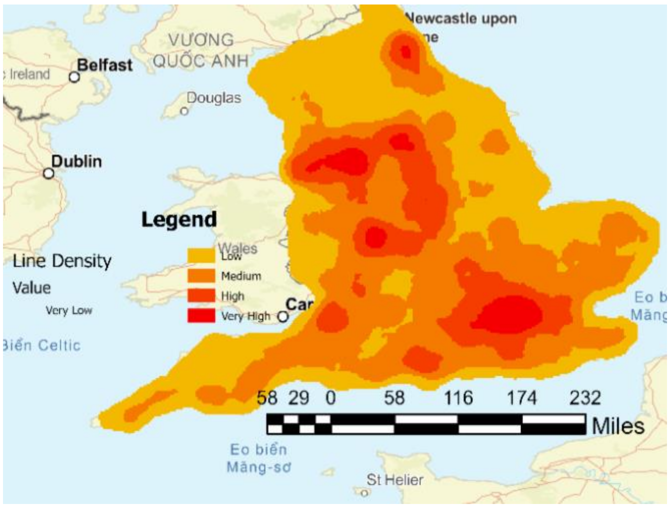

Figure 3. The dangerous accident locations.

Improved prediction accuracy has real-world safety implications, including:

1) Accident hotspot identification: By identifying areas with high accident densities, transportation agencies can prioritize these locations for safety improvements.
2) Proactive safety interventions: Accurately predicting high-risk zones enables targeted interventions such as:
   - Adjusting speed limits;
   - Improving road design;
   - Enhancing visibility;
   - Implementing traffic calming measures.
3) Enhanced emergency response: Predicting accident severity can optimize resource allocation for emergency response teams, reducing fatalities and injury severity. This could involve:
   - Dispatching appropriate resources;
   - Planning efficient routes;
   - Preparing hospitals for trauma cases.
4) Informed urban planning: Reliable models guide policymakers in designing safer transport systems, considering traffic volume, road geometry, and population density.
5) Optimized traffic management: These insights can be used to improve traffic flow control and inform better transportation planning. This could lead to:
   - Reduced congestion;
   - Smoother traffic flow;
   - Minimized travel delays.
6) Data-driven decision making: AI-GIS models offer insights into the factors contributing to accidents, allowing transportation planners to prioritize interventions with the greatest impact on safety.

These advancements translate into measurable safety enhancements, such as a reduction in accident rates and associated costs, ultimately saving lives and fostering safer communities.

5. Conclusions, Limitations, Mitigation Strategies, and Future Directions

5.1. Conclusions

This study demonstrates that integrating AI algorithms with GIS technology offers a powerful framework for traffic accident prediction. The SSA-optimized SVM model, combined with GIS-based clustering, enhances predictive accuracy and facilitates targeted safety interventions. By merging advanced machine learning techniques with spatial analysis, the proposed approach achieves high predictive accuracy while providing actionable insights for road safety improvements. The spatial validation of AI predictions underscores the importance of GIS in contextualizing and refining data-driven models.

5.2. Limitations and Mitigation Strategies

While our findings are promising, several limitations must be acknowledged:

1) Data quality and availability: Inaccuracies or gaps in the input data can adversely affect model performance. Future work should focus on automating data cleaning processes and improving data collection methodologies.
2) Computational complexity: The integrated framework demands substantial computational resources, potentially hindering scalability. Optimizing algorithms and leveraging cloud-based distributed computing could mitigate these issues.
3) Generalizability: The study is based on data from the UK, which may limit its applicability to other regions. Expanding the dataset to include diverse geographic areas and conducting localized model calibration will be important steps for future research.

5.3. Future Directions

Looking ahead, the integration of real-time data streams (e.g., live traffic, weather, and road condition updates) into the AI-GIS framework could further enhance predictive capabilities and responsiveness. Expansion of study regions to improve model generalizability. Development of cloud-based AI-GIS platforms for large-scale deployment.

References

1. Global Status Report on Road Safety 2023, https://www.who.int/publications/i/item/9789240086517, last accessed 2024/08/02.
2. Ahmed S, Hossain MA, Ray SK, Bhuiyan MM, Sabuj SR. A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. Transportation research interdisciplinary perspectives. 2023 May 1;19:100814.
3. Chai, A. B. Z., Lau, B. T., Tee, M. K. T., & McCarthy, C. (2024). Enhancing road safety with machine learning: Current advances and future directions in accident prediction using non-visual data. *Engineering Applications of Artificial Intelligence*, *137*, 109086.
4. Gatera, A., Kuradusenge, M., Bajpai, G., Mikeka, C., & Shrivastava, S. (2023). Comparison of random forest and support vector machine regression models for forecasting road accidents. *Scientific African*, *21*, e01739.

5. Bhele, R., Subedi, A., Thapa, B., & Rajchal, P. (2024, April). 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)* (pp. 12-17). IEEE.
6. Mohammed, S., Alkhereibi, A. H., Abulibdeh, A., Jawarneh, R. N., & Balakrishnan, P. (2023). GIS-based spatiotemporal analysis for road traffic crashes; in support of sustainable transportation Planning. *Transportation research interdisciplinary perspectives*, *20*, 100836.
7. Le KG, Liu P, Lin LT. Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi, Vietnam. Geo-spatial Information Science. 2020 Apr 2;23(2):153-64.
8. Le KG, Liu P, Lin LT. Traffic accident hotspot identification by integrating kernel density estimation and spatial autocorrelation analysis: a case study. International journal of crashworthiness. 2022 Mar 4;27(2):543-53.
9. Le, K. G., & Tran, Q. H. (2021). The sustainable development of railway system in Vietnam by GIS-based Technologies. In *E3S Web of Conferences* (Vol. 310, p. 03003). EDP Sciences.
10. Le, K. G., Liu, P., Lin, L. T. (2019). Road traffic accident black spot determination by using kernel density estimation algorithm and cluster statistical significant evaluation. In *FIG Working Week*.
11. Le KG, Ho TLH, Do VM, Hoc TQ. Applying a two-step cluster algorithm in traffic accident data analysis. Transport and Communications Science Journal. 2024;75(4):1673-87.
12. Al-Mistarehi BW, Alomari AH, Imam R, Mashaqba M. Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS. Frontiers in built environment. 2022 Apr 20;8:860805.
13. Hashimoto, S., Yoshiki, S., Saeki, R., Mimura, Y., Ando, R., & Nanba, S. (2016). Development and application of traffic accident density estimation models using kernel density estimation. *Journal of traffic and transportation engineering (English edition)*, *3*(3), 262-270.
14. Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, *40*(4), 1611-1618.
15. Gharehchopogh, F.S.; Namazi, M.; Ebrahimi, L.; Abdollahzadeh, B. Advances in Sparrow Search Algorithm: A Comprehensive Survey. Arch. Comput. Method Eng. 2023, 30, 427–455.