Open Access and Peer Reviewed Journal ISSN 2395-1303 https://www.ijetjournal.org/

RESEARCH ARTICLE OPEN ACCESS

HIGHWAY ACCIDENT RISK PREDICTION AND BLACKSPOT DETECTION USING MACHINE LEARNING

Dungavath Dushyanth Naik[1], Dr. S. Swarnalatha[2]

[1]PG Student, Department of ECE, Sri Venkateswara University College of Engineering, Tirupathi,

[2]Professor, Department of ECE, Sri Venkateswara University College of Engineering, Tirupathi,

Email – 1 dungavathdushyanth@gmail.com, 2 swarnasvu09@gmail.com

Abstract

Road traffic accidents continue to be one of the greatest challenges to highway safety, and there is a need for proactive approaches that can be used to identify locations where accidents are likely to occur and to predict accident risk. In this study, an integrated framework for highway accident risk prediction and blackspot detection is presented by applying machine learning with spatial analysis and supervised learning. The haversine distance metric is used for DBSCAN to detect spatial concentration of accidents in blackspots. Kernel Density Estimation (KDE) is used to visualise a continuous representation of the accident-density pattern. Random Forest (RF) and Extreme Gradient Boosting (XGBoost) classifiers are used to predict the risk and are evaluated by accuracy, precision, recall, F1-score and AUC-ROC. In the principal model evaluation reported in the study, XGBoost obtained 91.2% accuracy, 90.5% precision, 89.8% recall, 90.1% F1-score and 0.947 AUC-ROC, while Random Forest got 89.4% accuracy. The DBSCAN algorithm found 14 clusters corresponding to highway black spots with an optimised epsilon value of 1.0 km and a minimum of 5 samples, and a silhouette score of 0.58.

Keywords: *Highway safety; Accident risk prediction; Blackspot detection; DBSCAN; Kernel Density Estimation; Random Forest; XGBoost; Machine learning; Spatial analysis; Streamlit*

1. Introduction

Road traffic accidents continue to pose a significant problem for road safety, as roads are responsible for significant loss of life, economic damage and social impacts. As highlighted by previous studies, geographic, temporal, behavioural, environmental and road-related data can be valuable to accident analysis, especially when large data sets are analysed through computational methods (Ahmadi et al., 2025).

The traditional method of accident analysis primarily used statistical methods to determine accident rates. These techniques give valuable information on the pattern of past accidents, but they are limited in their usefulness when multiple accident-related factors are related to each other in a nonlinear way. Machine-learning approaches have therefore become more relevant as they can consider complex relationships based on variables like speed, road features, weather, visibility, traffic conditions, and accident history (Santos et al., 2021).

Highway safety analysis benefits from an important complementary perspective from spatial techniques. Road accidents have been represented using Kernel Density Estimation (KDE), in order to determine the spatial distribution of road accidents and the areas with high concentration of accidents (Anderson, 2009). Geographical accident data is especially well suited to density-based clustering techniques as they are able to detect clusters with arbitrary shapes, not requiring the number of clusters to be

known in advance, and can detect isolated observations as noise.

Random Forest is an ensemble learning method that consists of several decision trees, and has shown its applicability to the prediction of accidents and accident severity due to the robustness of the method and the simplicity with which it models nonlinear relationships (Breiman, 2001). In addition, XGBoost, a gradient boosting framework developed by Chen and Guestrin (2016), has been shown to achieve good prediction results on structured datasets and has been applied to traffic accident prediction and classification problems. Recent studies employing big data from telematics and weather factors further show that XGBoost and other ensemble models are well suited to accident-risk analysis, and that spatial and environmental factors are important (Ahmadi et al., 2025).

Although these advances have been made, the reviewed literature for this study shows that the three issues of accident prediction, spatial hotspot detection and practical decision-support systems are studied separately. The dissertation suggests the necessity for integrating multiple tasks, such as blackspot detection, accident-density analysis, risk prediction and practical visualisation in one framework. The proposed study aims to accomplish this by combining DBSCAN, KDE, Random Forest and XGBoost in a single analytical framework. The system also features a nearby blackspot identification mechanism using GPS, a feature importance analysis, and a safe-speed recommendation mechanism via an interactive dashboard built with Streamlit.

The goals of this study are then to detect highway accident blackspots with the DBSCAN algorithm, analyse spatial highway accident density with the KDE, construct and compare Random Forest and XGBoost models for predicting highway accidents, assess the predictive capability of these models, and incorporate the obtained analytical functions in an interactive decision support system.

2. Literature Review

Previous research mainly focused on accident frequency, regression-based approaches and geographical analysis to locate hazardous areas. With the amount of data available for accidents, geographical, traffic and environmental data, there has been a growing interest in the use of machine learning techniques for predicting and categorising the severity of traffic accidents (Santos et al., 2021).

In recent years, several studies have explored the analysis of traffic accidents and prediction of traffic hotspots using machine learning. Santos et al. (2021) investigated machine learning methods to analyse accidents and predict hotspots. Khanum et al. (2023) particularly studied the Random Forest for accident-severity prediction on Indian highways. But one of their methods was to predict only, and they did not combine spatial blackspot detection with density estimation based on KDE.

A spatial analysis technique has also been extensively applied to determine accident hotspot locations. Anderson (2009) proved the potential of using KDE and K-means clustering to profile road accident hotspots. KDE offers a continuous density representation, not limited to pre-defined spatial units. But KDE is essentially a density estimation method and is not a supervised accident-risk prediction mechanism.

DBSCAN is especially useful for accident locations as it can detect dense clusters using neighbourhood relationships, without needing to be known beforehand. It is suitable for geographical accident datasets due to its ability to recognise irregularly-shaped clusters and to distinguish noise observations from other data.

Random Forest is an ensemble learning method that uses multiple decision trees and gives feature-importance information, which can help to explain outputs of the forest model (Breiman, 2001). XGBoost is an efficient ensemble-learning method for structured classification problems, which extends the ensemble-learning approach by incorporating sequences of gradient boosting and

regularisation (Chen and Guestrin, 2016). Large-scale research conducted in recent years with the aid of telematics and weather information also revealed the effectiveness of XGBoost along with the significance of spatial, road, behavioural, and environmental factors in the prediction of accidents (Ahmadi et al., 2025).

The reviewed studies for the present study tend to focus on discrete parts – blackspot identification, continuous density analysis, risk classification, and an operational decision-support interface – rather than integrating them.

2.1 Research Gap and Contributions

2.1.1 Research Gap

The literature shows that highway accident analysis is typically approached either separately in the spatial or predictive dimension. The reviewed studies also have a limited level of integration of blackspot detection, continuous density estimation, predictive modelling, and an interactive decision-supporting interface within one framework.

In this regard, the present study aims to combine the following methods: DBSCAN and KDE for discrete detection of dangerous points and continuous estimation of accident density, respectively; Random Forest and XGBoost for risk prediction; and a Streamlit dashboard for visualisation and decision support.

2.1.2 Contributions

1. Spatial blackspot detection: Spatial blackspot detection with the DBSCAN algorithm is implemented on geographical accident coordinates based on the haversine distance without any prior specification of the number of blackspots.
2. Continuous accident-density analysis: KDE can be used in conjunction with DBSCAN to show the concentration of accidents in a continuous spatial density surface.
3. Random Forest and XGBoost – for a task of classifying accidents

Streamlit dashboard that offers blackspot visualisation, risk prediction, feature-importance analysis, GPS-based lookup and safe-speed suggestions.

3. Materials and Methodology

3.1 Overall Research Framework

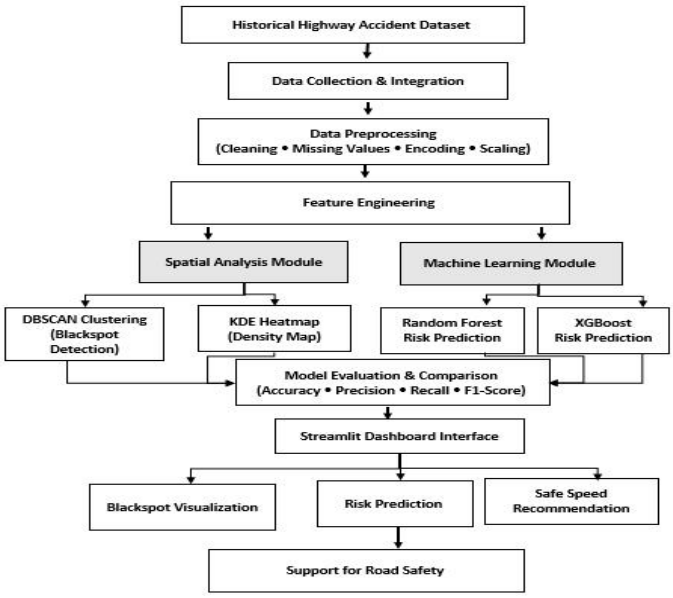

Figure 1 Overall Research Methodology

The historical accident records are first preprocessed and then feature-prepared. Geographical coordinates are then analysed using DBSCAN to find the locations of accidents with a high spatial concentration, and Kernel Density Estimation (KDE) is used to visualise the continuous distribution of accident density. The following models are then trained for an accident risk classification: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Evaluated models and spatial analysis outputs are incorporated into a Streamlit dashboard to be used for visualisation and decision support.

3.2 Dataset Description

The study involved three datasets for accidents: one from Andhra Pradesh with 500 records, Tirupati with 200 and Indian Roads with 20,000.

Table 1 Dataset characteristics

Dataset	Records	Application
Andhra Pradesh	500	Regional accident-risk analysis
Tirupati	200	Local accident-risk analysis
Indian Roads	20,000	Larger-scale accident analysis

3.3 Data Preprocessing

Before model development, the uploaded accident data is examined for validity. Latitude/longitude data is checked against geographic limits, and any records with invalid/undecodable geographic data are deleted. Selected variables, for which values may be missing, are dealt with during the pre-processing phase. The class labels of the categorical target variable are preserved for interpretation and transformed into a set of binary variables for machine-learning analysis. Where possible, constant-value features are not included in the models, and numerical variables are chosen for model building. A training-to-testing ratio of 75% to 25% is used with stratification when available in the class distribution.

In many previous accident prediction studies, data cleaning and handling missing values have already been done before the machine learning modelling process (Ahmadi et al., 2025).

3.4 DBSCAN-Based Blackspot Detection

Geographical clusters of accident locations are identified by the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. Unlike K-means, DBSCAN does not need the number of clusters to be known in advance, and groups the observations based on the density of the local area. In addition, it can classify single observations as noise; thus, it can be applied to accident location analysis under the condition that there are some accidents that do not appear in systematic clusters (Ester et al., 1996). The two most important parameters for DBSCAN are the neighbourhood radius (ε) and the minimum number of observations that make up a dense region (MinPts).

To evaluate, a variety of parameter combinations were tested with the ε values of 0.5, 1.0, 1.5, 2.0 and 3.0 km and different MinPts values. The reported optimum was ε = 1.0 km and MinPts = 5; these resulted in 14 clusters and a reported silhouette score of 0.58.

3.5 Kernel Density Estimation

KDE is used as a complementary spatial-analysis tool to display the accident concentration as a continuous density surface. KDE has been used in the past in the context of road-accident hotspot analysis to define spatial areas where the number of accidents is greater than normal, measured by the density of accidents per unit area (Anderson, 2009).

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{d(x,x_i)}{h}\right)$$

Where n is the number of observations of the accidents, K is the kernel function, d(x, xi) is the distance between the two locations, and h is the bandwidth. The distance function is the haversine distance, and the implementation is based on a Gaussian kernel.

3.6 Random Forest Model

One of the supervised learning models for accident-risk classification is the Random Forest model. The algorithm builds several decision trees from different subsets of features and different bootstrap samples of data and then merges the predictions of these trees to produce a final classification (Breiman, 2001).

The reported implementation is based on 200 decision trees, class weighting and a fixed random state of 42. During model evaluation and model-selection activities, a five-fold cross-validation procedure is employed.

3.7 XGBoost Model

The second supervised learning model is the Extreme Gradient Boosting (XGBoost). XGBoost is based on efficient optimisation methods and regularisation, and was originally introduced as a scalable gradient boosting framework (Chen & Guestrin, 2016).

learning rate (0.1) and 100 estimators were the configuration with the highest reported cross-validation accuracy of the different configurations tested, and was therefore the one that was selected.

3.8 Model Evaluation

The performance of the two models is compared by accuracy, precision, recall, F1-score and AUC-ROC. Model selection is also made using five-fold cross-validation.

Random Forest had an accuracy of 89.4%, and XGBoost had an accuracy of 91.2% in the main comparison of the models presented in the study. XGBoost also gave better results in terms of precision, recall, F1-score, and AUC-ROC compared to Random Forest.

Table 2 Performance comparison of Random Forest and XGBoost

Metric	Random Forest	XGBoost
Accuracy	89.4%	91.2%
Precision	88.1%	90.5%
Recall	87.6%	89.8%
F1-score	87.8%	90.1%
AUC-ROC	0.921	0.947

3.9 Risk Prediction and Safe-Speed Advisory

The system that is implemented classifies the prediction of the risk as Low, Medium and High. The thresholds reported are considered Low <30%, Medium 30-60%, and High >60% risk. The safe-speed module kicks in when the forecasted danger-class probability is above the high-risk threshold.

The module recognises the speed feature and then iteratively reduces the input speed by 5 km/h, and then recalculates the predicted risk. This process is repeated until the predicted risk is less than a defined safety level of 30%.

3.10 Streamlit Dashboard Implementation

The entire analytical process is realised by an interactive Streamlit dashboard. Modules for uploading accident data, machine learning models, visualisation of feature importance, live risk prediction and road-speed recommendation.

The implementation is implemented with the help of the following libraries: Pandas, NumPy, scikit-learn and XGBoost (for data processing and machine learning), and Plotly (for interactive visualisation) and is implemented in the Python programming language.

4. Results

4.1 DBSCAN Blackspot Detection

The silhouette score corresponding to that was 0.58, which represented significant discrimination between the identified spatial clusters. In addition, 62 noise observations were obtained for individual accident locations that did not meet the density requirement.

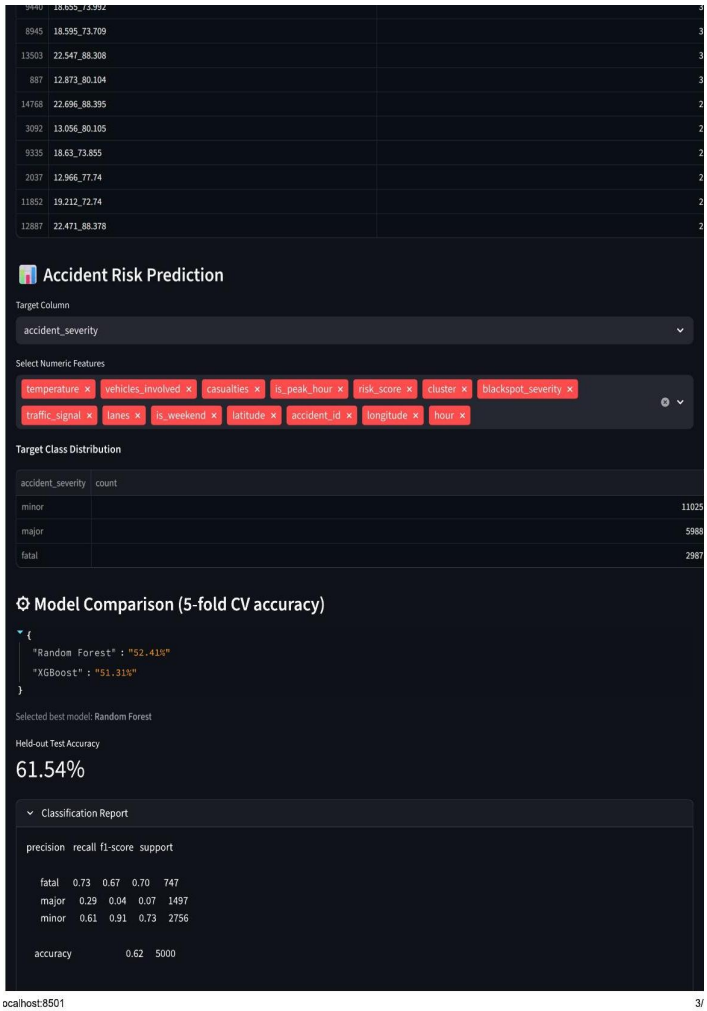

Figure 2 DBSCAN blackspot detection map showing the 14 identified clusters

The largest cluster had 14 accidents identified, and six clusters had >8 accidents and were considered priority locations in the developed analysis.

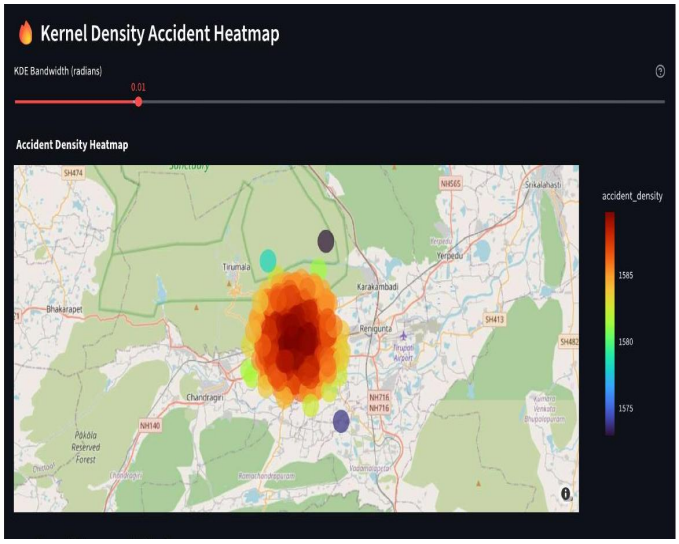

Figure 3 KDE accident-density heatmap showing continuous spatial risk distribution

This resulted in a heatmap indicating areas of high and low concentration of accidents and helped interpret the clusters found in DBSCAN. In particular, KDE enabled a visualisation of areas with high accident density, even if the number of observations did not meet the DBSCAN clustering criterion.

4.3 Random Forest Performance

Figure 4 Random Forest Output

In the main reported model evaluation, the overall accuracy of the Random Forest model was 89.4%, while its precision, recall, F1-score and AUC-ROC were 88.1%, 87.6%, 87.8% and 0.921, respectively.

4.4 XGBoost Performance

XGBoost model gave the best result among the two models used for evaluation. The model recorded 91.2% accuracy, 90.5% precision, 89.8% recall, 90.1% F1-score and an AUC-ROC of 0.947. For XGBoost, the accuracy for the five-fold cross-validation was reported as 91.0%, whereas the accuracy of Random Forest was 89.1%.

Table 3 Performance comparison of the two machine-learning models

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random Forest	89.4%	88.1%	87.6%	87.8%	0.921
XGBoost	**91.2%**	**90.5%**	**89.8%**	**90.1%**	**0.947**

4.5 Hyperparameter Tuning

We used grid search to find appropriate settings for both classifiers. In the case of Random Forest, the reported comparison took place with varying numbers of estimators and maximum depth. The number of estimators was set to 200, and the maximum depth was not limited, giving the highest accuracy of 89.1% reported for the cross-validation.

The combinations of maximum depth, learning rate, and number of estimators were tested for XGBoost. The best reported cross-validation accuracy of 91.0% was obtained with the maximum depth = 6, learning rate = 0.1, and 100 estimators.

Table 4 Selected hyperparameters and cross-validation performance

Model	Selected parameters	5-fold CV accuracy
Random Forest	n_estimators = 200; max_depth = None	89.1%
XGBoost	max_depth = 6; learning_rate = 0.1; n_estimators = 100	91.0%

4.6 XGBoost Confusion Matrix and Classification Results

For the reported test set, the model correctly classified 87 of 100 Low-risk observations, 82 of 100 Medium-risk observations and 88 of 100 High-risk observations. The class-level F1-scores were 0.91, 0.895 and 0.915, respectively.

The XGBoost confusion matrix and classification report are shown in Figure 11.

The overall accuracy was 91.2 % for the held-out test set of 300 samples reported.

4.7 Feature Importance

The feature-importance analysis results revealed that the XGBoost model highlighted that speed is the most significant feature, with 34.2% of the reported feature importance. The contribution of the accident density derived from KDE was 22.8%, followed by the road type (14.6%) and blackspot severity (12.3%). The importance attributed to weather and the type of vehicle accounted for smaller percentages.

The findings suggest that the individual characteristics of the accident and spatial information all played a role in predicting the accident. In particular, the role of accident density and blackspot severity illustrates the utility of combining the outputs from the spatial analysis with supervised machine learning.

4.8 Risk Prediction and Safe-Speed Advisory

The created dashboard gave the risk classifications (Low, Medium and High) and risk probability. The safe-speed recommendation module was triggered when the predicted danger-class probability was greater than 60%. The actual implementation involved decreasing the input speed by 5 km increments and retesting the model again and again until the output risk was below 30%.

The project reports that the advisory did a good job of bringing the risk prediction from the High-risk range down to the Low-risk range in the evaluated scenarios.

Figure 5 Streamlit Dashboard Interface

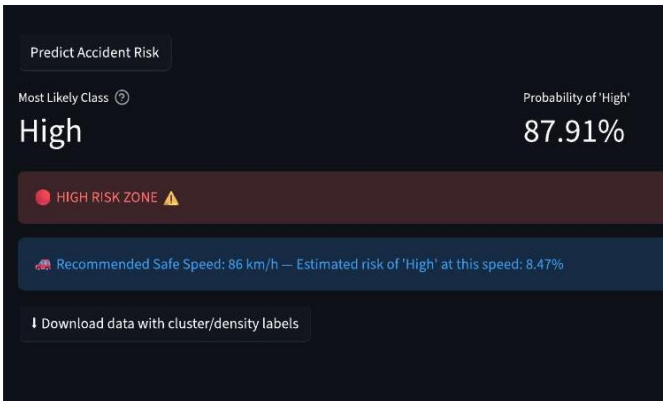

Figure 6 Dashboard Output Results

4.9 Integrated Dashboard Results

The Streamlit implementation was able to successfully integrate all the separate parts in a single interface. The dashboard offers data upload, mapping with DBSCAN, visualisation using kernel density estimation, road-segment risk analysis, model training and evaluation, feature importance, live model prediction and safe-speed recommendation.

5. Discussion

The results show that the proposed framework can integrate spatial analysis and machine-learning prediction to get a more comprehensive impression of highway accident risks. The DBSCAN method detected 14 areas of high spatial concentration of blackspots (clusters), and the KDE method yielded a continuous density surface that extended beyond the area of the clusters.

Using machine learning comparison, XGBoost outperformed Random Forest on all evaluation metrics. The accuracy of XGBoost is 91.2%, which is higher than 89.4% for Random Forest, while the precision, recall, F1-score and AUC-ROC of XGBoost are also higher. This is consistent with the ability of gradient-boosting methods to gradually improve predictions by directing subsequent learners to the errors of previous learners as outlined in Chen & Guestrin (2016).

The feature importance analysis also suggests the importance of the combination of spatial and non-spatial variables. The reported features by XGBoost in descending order of importance were speed (34.2%), accident density from KDE (22.8%), road type (14.6%), and blackspot severity (12.3%). This is in line with earlier studies showing that spatial factors are crucial for modelling accident risk (Ahmadi et al., 2025).

The safe-speed module takes the framework from prediction to an operational recommendation. In the scenarios included in this study, this process brought the predicted risk back down into the Low-risk range. However, the recommendation is not meant to replace statutory speed limits, engineering assessments or real-time traffic and weather data.

6. Limitations and Future Scope

The existing regime predictions depend on the quality, completeness and geographical precision of historical accident records. The present validation is also conducted on the datasets used in the study, and thus should be validated in other geographical regions for wider generalisation. Moreover, the existing framework doesn't include real-time traffic data, weather data, or connected-vehicle data. Class imbalance (the fact that some classes may have a small number of observations) could also influence the risk classification reported.

Future development can thus be extended with live traffic and weather data, mobile driver alerts, urban roads and temporal and deep learning models like LSTM and GNN approaches.

7. Conclusion

1. For the selected parameters (ε = 1.0 km and MinPts = 5), DBSCAN was able to detect 14 concentrated highway accident clusters.
2. Accident-density assessment: In addition to the output of DBSCAN, KDE also produced a continuous surface of accident density

performance was XGBoost, with a score of 91.2% accuracy, 90.5% precision, 89.8% recall, 90.1% F1-score and 0.947 AUC-ROC.

4. The top four variables in the XGBoost feature-importance analysis were speed (34.2%), accident density (22.8%), road type (14.6%), and blackspot severity (12.3%).
5. Operational risk advisory: The developed safe-speed module was able to identify high predicted risk and then incorporate actionable recommendations for a speed reduction by an iteration of 5 km/h.
6. Finally, the last Streamlit implementation showed the integration of the data processing, the blackspot detection, the KDE mapping, the evaluation of the machine-learning models, the feature-importance analysis, the prediction of the risk in real time, and the calculation of the safe speed in a single platform.

References

[1] Anderson, T.K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3), 359–364.

[2] Ahmadi, N., Ataei, S.M.-N., Khosravi, S., Jafari, A., Shahraz, S. & Farzadfar, F. (2025). Interpretable machine learning models to predict errors in road accident hotspots using telematics data. *PLOS ONE*, 20(7), e0326483.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[4] Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[5] Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data*

traffic accident analysis and hotspot prediction. *Computers*, 10(12), 157.

[7] Khanum, H., Garg, A. & Faheem, M.I. (2023). Accident severity prediction on Indian highways using Random Forest. *F1000Research*, 12, 494.

[8] Bokaba, T., Doorsamy, W. & Paul, B.S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. *Applied Sciences*, 12(2), 828.

[9] AlMamlook, R.E., Kwayu, K.M., Alkasisbeh, M.R. & Frefer, A.A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*.

[10] Zangeneh, A., Najafi, F., Karimi, S., Saeidi, S. & Izadi, N. (2018). Spatial-temporal cluster analysis of mortality from road traffic injuries using geographic information systems in the west of Iran during 2009–2014. *Journal of Forensic and Legal Medicine*, 55, 15–22.

[11] Kaygisiz, Ö., Düzgün, Ş., Yildiz, A. & Senbil, M. (2015). Spatio-temporal accident analysis for accident prevention in relation to behavioural factors in driving: The case of South Anatolian Motorway. *Transportation Research Part F: Traffic Psychology and Behaviour*, 33, 128–140.

[12] Hamdar, S.H., Qin, L. & Talebpour, A. (2016). Weather and road geometry impact on longitudinal driving behaviour: Exploratory analysis using an empirically supported acceleration modelling framework. *Transportation Research Part C: Emerging Technologies*, 67, 193–213.

[13] Khan, M.Q. & Lee, S. (2019). A comprehensive survey of driving monitoring and assistance systems. *Sensors*, 19(11), 2574.

[14] Siami, M., Naderpour, M. & Lu, J. (2021). A mobile telematics pattern recognition framework for driving behaviour extraction. *IEEE Transactions on Intelligent Transportation Systems*, 22(3), 1459–1472.

Python. *Journal of Machine Learning Research*, 12, 2825–2830.