Spatial Analysis of Road Infrastructure and Road Accident Risk in Bangkok

Student ID: 5758736

***Abstract*—Road traffic accidents remain a major public safety concern, especially in today's world of rapidly developing and urbanising cities. This study focuses on Bangkok, Thailand, and analyses accident data to investigate how road-environment and contextual factors influence accident frequency and severity. Exploratory data analysis revealed substantial variation in accident rates across, with motorways and major arterial roads exhibiting the highest accident frequencies per kilometer. A Poisson regression model confirmed that road class is a statistically significant predictor of accident occurrence.**

To build on these findings, a Random Forest Classifier was developed to predict accident severity. A baseline model was trained using only the road-type features, in which it identified motorways, primary and secondary roads, and intersection presence to be the most influential indicators of accident severity. Feature importance rankings closely aligned with descriptive and statistical analyses, indicating that the model captures meaningful real-world patterns. This first model achieved an accuracy of only 0.595 suggesting that road-environment features are strong predictors but they do not fully explain accident severity. A secondary Random Forest classifier was trained and evaluated, incorporating attributes such as vehicle type, weather condition, road geometry, and slope descriptors. While vehicle type and motorway class emerged as the most influential predictors, the extended model did not improve predictive performances with an accuracy of 0.548 over the simpler road-only model, suggesting that increased feature dimensionality and sparse categorical representation limit generalisation.

Overall, the study demonstrates that road infrastructure characteristics are dominant determinants of accident severity in Bangkok. These findings highlight major roads and intersections as key targets for traffic safety interventions and provide a foundation for future work incorporating additional temporal and spatial factors to further enhance accident risk modelling.

I. Introduction

Road traffic accidents constitute a major global public safety challenge. According to the World Health Organization, approximately 1.35 million people die annually due to road traffic collisions, with many more sustaining serious injuries [1]. Rapid urbanisation and increasing vehicle ownership have intensified accident risks in metropolitan regions, particularly in developing cities where infrastructure expansion often lags behind traffic growth. Understanding the environmental and behavioural factors that contribute to accident occurrence and severity is therefore critical for designing effective road safety interventions.

Bangkok, Thailand, has consistently recorded high road accident rates, remaining of of the highest in Asia and upper-middle-income countries. Complex road networks, heterogeneous traffic composition, and high-density urban activity make accident prevention particularly challenging. Data-driven approaches that leverage accident records and road-network attributes offer an opportunity to identify high-risk environments and prioritise targeted safety measures.

Recent advances in data analytics and machine learning have enabled more accurate prediction and interpretation of accident severity. Traditional statistical models such as Poisson and logistic regression remain widely used for analysing accident frequency and risk factors. However, machine learning models, particularly methods such as Random Forests, have demonstrated superior predictive performance while also providing interpretable feature-importance measures. These techniques allow identification of the most influential infrastructure and contextual factors contributing to severe accident outcomes.

This study applies a structured data analytics pipeline to accident data from Bangkok to investigate (i) how road class influences accident frequency, and (ii) how road-environment and contextual factors influence accident severity. The analysis integrates exploratory data analysis, statistical regression, and machine learning classification to derive robust and interpretable insights into accident risk patterns. The findings aim to support evidence-based prioritisation of road safety interventions in urban environments.

II. Background and Related Work

Accident frequency and severity modelling has long been an important research topic in transportation safety analysis. Early studies relied primarily on statistical regression techniques, particularly Poisson and Negative Binomial models, to relate accident counts to road and traffic characteristics (such as road segment or intersections). Lord and Mannering [2] provide a comprehensive overview of statistical methods for crash-frequency analysis, highlighting the suitability of count-based models for accident occurrence data.

In the same vein, severity analysis has traditionally employed discrete outcome models such as multinomial logit and ordered probit regression to examine how driver behaviour, vehicle type, and road characteristics influence injury severity. Savolainen et al. [3] present a detailed review of accident severity modelling methods and emphasise the growing interest in incorporating roadway and environmental factors.

Recent work increasingly applies machine learning to traffic accident prediction, with tree–based ensemble models outperforming many traditional approaches on both accuracy and

flexibility. Studies comparing algorithms for severity prediction report that Random Forests and Gradient Boosting Trees capture non-linear interactions between roadway, environmental, and temporal factors more effectively than classical regression models, while remaining interpretable through feature importance and SHAP-based explanations [4]. For example, recent applications on highway and urban crash datasets show Random Forests achieving higher accuracy than logistic or ordered probit models in injury–severity classification, with model diagnostics highlighting weather, location type, and road conditions as dominant predictors [5].

Beyond infrastructure, contextual attributes such as weather, vehicle type, lighting conditions, and detailed road geometry further shape crash outcomes. Empirical studies in both high-income and middle-income countries demonstrate significant effects of precipitation, visibility, vehicle mix, and temporal patterns on injury severity and crash risk, although the relative importance of individual variables varies by context and dataset. Work comparing high-dimensional machine-learning models with reduced feature sets also cautions that including too many weak or highly categorical predictors can increase complexity and overfitting without improving generalisation performance, reinforcing the value of staged model building, feature selection, and explicit assessment of variable contribution using importance metrics or SHAP values [6].

Building on this literature, the present study adopts a hybrid analytical framework that combines traditional statistical modelling with machine-learning methods to examine both accident frequency and severity on Bangkok's road network. Using OpenStreetMap's road-environment attributes together with national accident records, the analysis compares road-only Random Forest models against extended models that incorporate contextual features such as weather and temporal factors, in order to assess whether these additional variables provide measurable predictive gains beyond core infrastructure characteristics.

III. Datasets

Two datasets were primarily used to observe the relationship between road accidents in Bangkok and road types.

A. Thailand Road Accident [2019-2022]

The dataset used in this study contains comprehensive records of road traffic accidents in Thailand, covering the period from approximately 2019 to 2022. The data originates from the Office of the Permanent Secretary, Ministry of Transport, and is publicly available through Kaggle with a total of 81,735 accidents recorded during these 3 years, already cleaned and ready to use. Each record in the dataset represents a reported road accident and includes a range of attributes describing temporal, geographic, environmental, and situational factors. Key variables of interest include the date of the incident, geographical coordinates, vehicle type, fatalities and injuries.

B. Thailand Roads

To support spatial analysis of road accident distribution, a road network dataset for Thailand was obtained from OpenStreetMap (OSM). OpenStreetMap is a collaborative, open-source geographic database that provides detailed mapping of infrastructure worldwide. The extracted dataset contains all mapped road features in Thailand tagged with highway attributes, representing the national road network. According to OpenStreetMap's regional summary statistics, the dataset contains approximately 999.7 thousand kilometres of roads, covering an estimated 76% of the total road length in the region based on AI-mapped completeness estimates. The average age of the mapped data is approximately four years, with around 4% of road segments added or updated within the last six months, indicating a relatively current and actively maintained dataset. Each feature in the dataset includes spatial geometry (latitude and longitude coordinates) and classification tags describing road type (e.g., motorway, primary, secondary, residential). This dataset adds an infrastructure context to the road accidents in Thailand, mapping accident points onto roads for analysis.

IV. Software and Libraries

All data processing and analysis in this study were conducted using Python as the primary programming language. Python was chosen due to its extensive capability to handle data files, visualization, and spatial processing.

Pandas was used for general data manipulation, cleaning, and statistical analysis of the road accident dataset. For spatial data processing, GeoPandas was employed to handle georeferenced datasets, while Shapely was used to construct and manage geometric objects such as coordinate-based point features representing accident locations. The OpenStreetMap road network data was processed and analysed within this geospatial framework.

Data visualization was carried out using Matplotlib and Seaborn to generate statistical plots and spatial distribution figures. Additionally, NumPy was used for numerical operations and array-based computations. For density-based spatial analysis, the Gaussian kernel density estimation function from SciPy was applied to identify accident concentration hotspots.

V. Hypothesis

With the understanding that road infrastructure plays a central role in traffic safety, it can be assumed that a relationship may exist between road-environment characteristics and accident outcomes. In particular, high-capacity and high-speed road types may be associated with greater accident frequency and increased accident severity.

As such, the following hypotheses are proposed:

1) Road classes with higher traffic capacity, such as motorways and major arterial roads, may exhibit higher accident frequencies than minor road types.
2) Accidents occurring on major roads and at intersections may have an increased likelihood of resulting in severe

outcomes compared to those on minor roads or non-intersection locations.

3) Road-environment characteristics may serve as effective predictors for accident severity classification.

Furthermore, it may be expected that additional contextual factors, such as vehicle type, weather conditions, and road geometry, contribute to accident severity. However, it is anticipated that road-environment features remain the dominant predictors, and that incorporating broader contextual variables may provide limited improvement over models based solely on road infrastructure attributes.

VI. Preprocessing

The first step in preprocessing was to filter the accident records to the year 2022. This year was selected to represent a post-pandemic period in which road usage had largely returned to normal conditions, avoiding distortions in traffic patterns caused by COVID-19 restrictions. After filtering, the dataset was reduced from 81,735 nationwide accident records to 21,032 accidents occurring in 2022.

Next, both the accident and road network datasets were spatially restricted to the Bangkok metropolitan area. This was achieved using a geographic bounding box defined by latitude 13.5–14.2 and longitude 100.3–100.9. Applying this spatial filter reduced the accident dataset to 3,311 records within Bangkok. Figure 1 illustrates the resulting accident point distribution and the corresponding kernel density heatmap, which provides an initial visual indication of spatial clustering. Following spatial filtering, the accident records were converted into a GeoDataFrame by transforming latitude and longitude coordinates into point geometries. The OpenStreetMap road network data was similarly loaded as a GeoDataFrame containing line geometries representing road segments. Both datasets were re-projected to a common projected coordinate reference system (EPSG:3857) to enable accurate distance-based spatial operations. This projection step is essential because spatial joins and buffer-based calculations require metric units rather than angular degrees. To associate each accident with its surrounding road infrastructure, a nearest-neighbour spatial join was performed, assigning every accident point to the closest road segment in the OpenStreetMap network. This process enriched each accident record with road classification attributes, enabling subsequent analysis of accident frequency and severity across different road types. Additionally, an intersection indicator variable was derived from the road network geometry. Road segment endpoints were extracted as proxy junction locations and buffered by 20 metres to account for GPS positional uncertainty. Accident points falling within these buffered regions were classified as intersection-related. This produced a binary feature distinguishing intersection and non-intersection accidents for later comparative analysis. Finally, an accident severity score was constructed from the recorded number of fatalities and injuries. Accidents involving fatalities were assigned a severity score of 3, those involving injuries a score of 2, and damage-only incidents a score of 1. This additional variable enabled quantitative assessment of how infrastructure characteristics influence accident outcomes.

Fig. 1. Accident Point Density - Bangkok (2022)

VII. Data Analysis

To understand the overall spatial pattern of accidents in Bangkok, accident point locations were plotted over the city's road network. A kernel density estimation (KDE) heatmap was then generated to highlight areas of concentrated accident activity.

Figure 2 shows that accidents are not randomly distributed but cluster strongly along major arterial road corridors and dense urban zones. This initial visual exploration suggests a relationship between accident occurrence and road network structure, motivating further infrastructure-based analysis.

A. Exploratory Data Analysis

To establish baseline trends in the dataset, accident rates per kilometre were computed for each road class. Figure 3 illustrates the accident frequency normalised by road length for Bangkok in 2022.

The results show that motorways exhibit the highest accident rate at 3.12 accidents per kilometre, substantially exceeding all other road categories. Trunk and primary roads also display elevated accident frequencies, while residential and minor road types exhibit comparatively low rates. This trend aligns with expectations that higher-speed, higher-capacity roads present greater exposure to severe collisions.

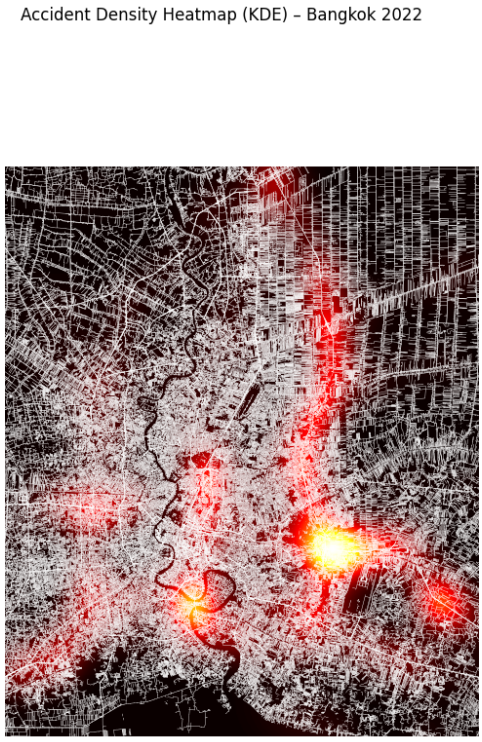

Fig. 2. Accident Density Heatmap (KDE) - Bangkok (2022)

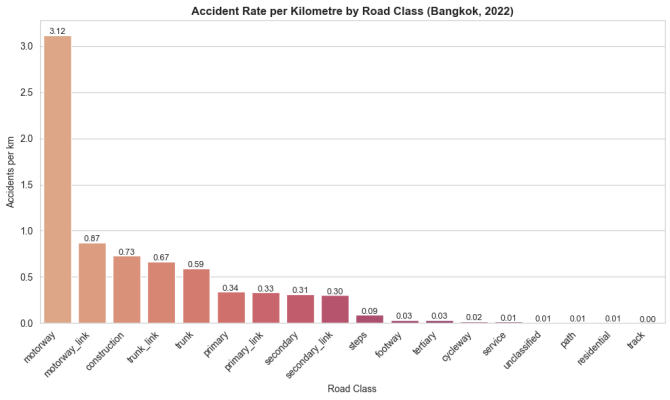

Fig. 3. Accident Rate per Kilometer by Road Class (Bangkok, 2022)

This descriptive analysis provides an empirical foundation for the subsequent modelling stage. It indicates that road class is likely to be a strong predictor of accident severity, justifying its inclusion as a key feature in the classification model.

Notably, the Random Forest feature importance analysis later confirms this pattern: motorway, primary, and secondary road indicators emerge as the most influential predictors of accident severity. The consistency between observed accident rates and model-derived importance scores suggests that the model captures meaningful real-world relationships rather than spurious correlations.

B. Statistical Modelling: Poisson Regression

To quantify the relationship between road class and accident frequency, a Poisson Generalised Linear Model (GLM) was fitted using accident counts as the dependent variable and road class as a categorical predictor.

The regression results show that all road-class coefficients are statistically significant ($p < 0.001$). Using motorway roads as the reference category, all other road classes exhibit large negative coefficients, indicating substantially lower expected accident counts relative to motorways. For example:

- Residential roads show the largest negative coefficient (6.394), reflecting extremely low accident rates.
- Tertiary roads also show low expected accident frequencies (4.548).
- Trunk and primary roads show moderately lower rates relative to motorways.

These findings quantitatively confirm the EDA observation that motorway environments carry significantly higher accident risk. The Poisson model therefore provides a statistically grounded baseline confirming that road class is strongly associated with accident occurrence frequency.

C. Random Forest

While the Poisson model quantified associations with accident counts, a Random Forest classifier was trained to predict accident severity classes based on road-environment features. In addition to predictive performance, Random Forests provide feature importance scores, representing each feature's contribution to reducing classification impurity across the ensemble.

The extracted importance scores reveal that:

- Road_motorway is the most influential predictor (importance ≈ 0.384).
- Road_primary, road_secondary, and intersection presence also show substantial importance.
- Residential and tertiary roads contribute minimally to severity prediction.

Figure 4 presents the feature importance scores for the road-environment Random Forest model. Visualization highlights the dominance of highway segment in predicting severity, followed by primary and secondary road classes and intersection presence, while residential and tertiary roads exhibit comparatively low influence.

The confusion matrix in Figure 5 shows the baseline Random Forest model demonstrating strong predictive performance for low-severity accidents but poor discrimination for medium- and high-severity cases. In particular, the model fails to correctly identify any high-severity accidents, misclassifying all severe cases as lower severity. This behaviour reflects the strong class imbalance in the dataset, where severe accidents represent a small proportion of the total sample. As a result, while the model achieves reasonable overall accuracy, it lacks sensitivity to severe accident outcomes, limiting its practical applicability for identifying high-risk scenarios.

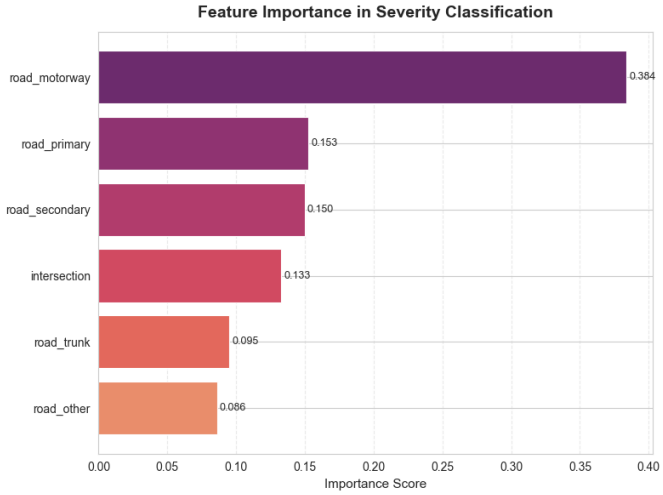

Fig. 4. Feature Importance in Severity Classification

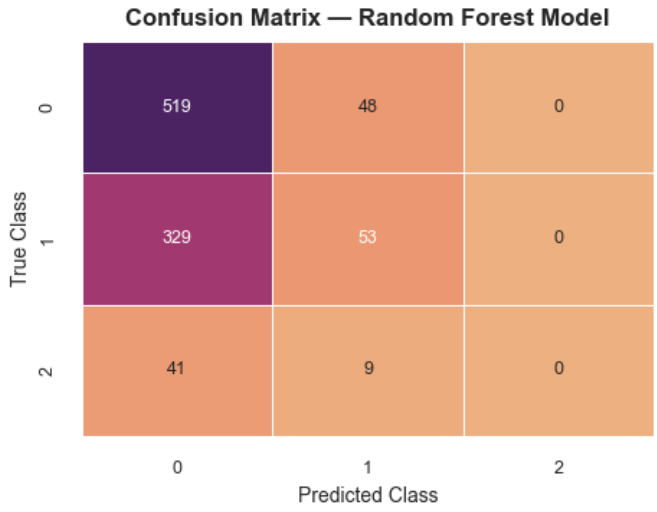

Fig. 5. Confusion Matrix from the Random Forest Model trained

D. Extended Feature Random Forest Model

To evaluate whether broader contextual factors improve severity prediction, an extended Random Forest model was trained incorporating additional attributes beyond road environment features. Specifically, weather condition, vehicle type, road geometry (road_description), and slope description were added alongside road class and intersection presence. All categorical variables were one-hot encoded and the model was trained using the same cross-validation procedure as the baseline road-only model.

We first look at severity distributions across vehicle type, weather condition, road geometry, and road slope (Figures 6, 7, 8, 9). Passenger cars account for most low-severity accidents, while motorcycles show a higher proportion of severe outcomes. Most accidents occur under clear weather, though rainy and dark conditions exhibit relatively greater severity. Straight roads dominate accident counts, but curved

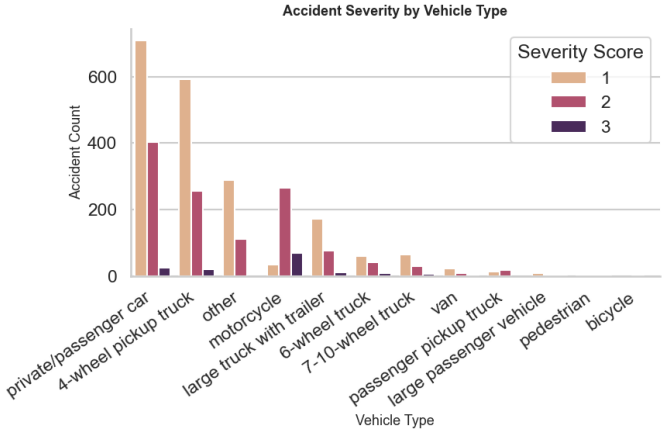

Fig. 6. Enter Accident severity by Vehicle Type

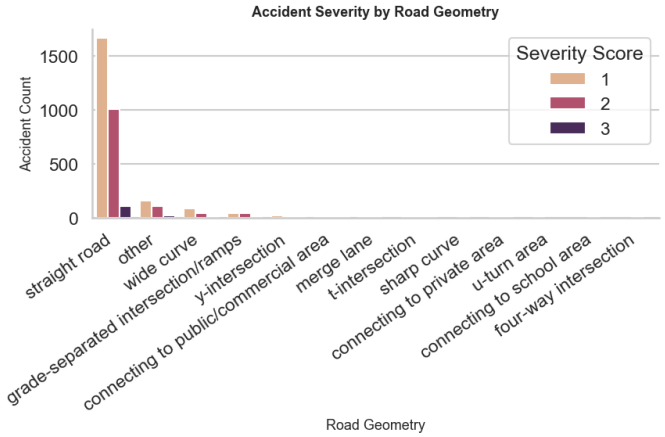

Fig. 7. Accident Severity by Road Geometry

and intersection-related geometries display varied severity patterns. Finally, while most accidents occur on flat segments, sloped roads show a slightly higher relative proportion of severe accidents. These observations suggest that contextual attributes may influence accident severity and justify their inclusion in the extended severity prediction model.

Figure 10 presents the feature importance scores of the extended model. The most influential predictors are vehicle_type_motorcycle and road_class_motorway, indicating that accident severity is strongly associated with vulnerable vehicle types and high-capacity roads. Other meaningful contributors include vehicle_type_private/passenger car, road_class_secondary, and intersection presence. Weather conditions and road geometry descriptors appear with lower importance scores, suggesting weaker predictive contribution to severity classification within this dataset.

The confusion matrix of the extended model (Figure 11) shows that the classifier maintains reasonable discrimination for lower-severity accidents but to struggle with correctly identifying high-severity cases. This behaviour reflects the persistent class imbalance, where severe accidents form a relatively small proportion of the dataset.

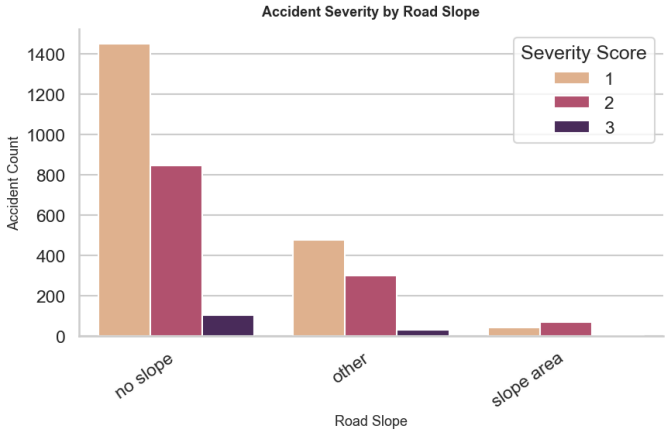

Fig. 8. Accident Severity by Road Slope

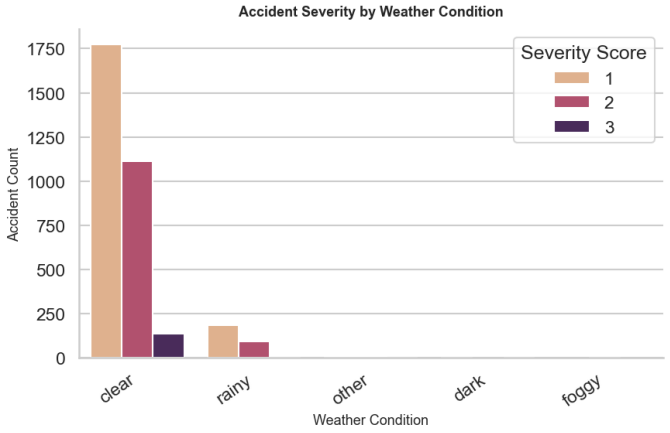

Fig. 9. Accident Severity by Weather Condition

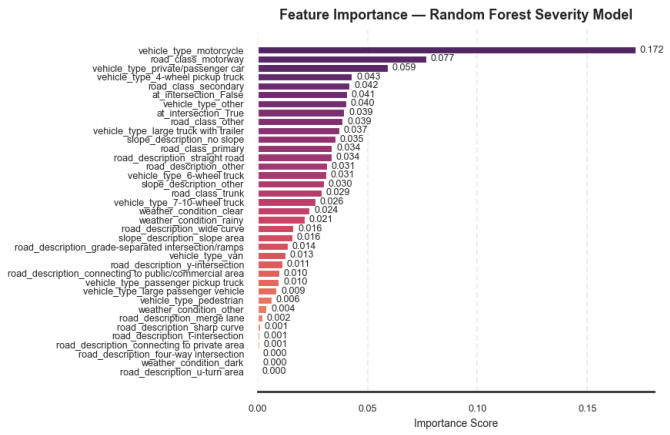

Fig. 10. Feature Importance - Extended Forest Model

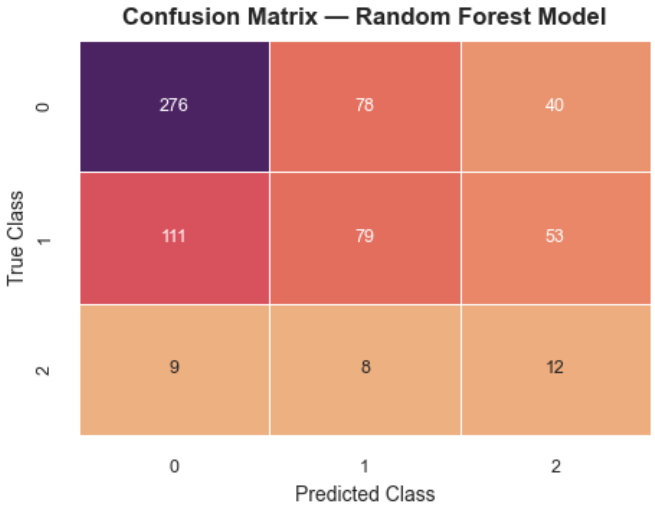

Fig. 11. Confusion Matrix for the Extended Feature Model

E. Model Performance Comparison

To assess whether the additional contextual features improved predictive performance, cross-validation accuracy was compared between the baseline road-only model and the extended model.

- Road-only model cross-validation accuracy: 0.595
- Extended model cross-validation accuracy: 0.548

Contrary to expectations, the extended model performs slightly worse than the simpler road-only model. This indicates that while vehicle type and weather conditions introduce additional contextual information, they also increase feature dimensionality and noise, leading to reduced generalisation performance on unseen data.

This result suggests that road environment features alone are strong predictors of accident severity in this dataset, and that adding further categorical variables does not necessarily improve model performance. It is likely that the extended model suffers from sparse category representation, particularly in rarely occurring weather and road geometry classes, which limits the model's ability to learn robust patterns.

VIII. Conclusions

This study investigated accident frequency and severity patterns in Bangkok using a structured data analytics pipeline combining exploratory analysis, statistical modelling, and machine learning techniques. It was hypothesised that high-capacity road classes and intersection locations would exhibit higher accident frequencies and increased likelihood of severe outcomes, and that road-environment features would serve as effective predictors for accident severity classification. The results consistently support these hypotheses.

While the overall classification accuracies of 0.595 for the road-only model and 0.548 for the extended model indicate moderate predictive performance, these values should be interpreted in the context of strong class imbalance within the

dataset. Low-severity accidents represent the majority class and are classified reliably, whereas high-severity cases form a small proportion of the data and are frequently misclassified. As a result, overall accuracy is driven largely by the model's ability to predict common low-severity events, while sensitivity to severe accidents remains limited. This highlights that, although the models successfully capture dominant structural patterns in the data, further refinement is required before such classifiers could be reliably applied for real-world identification of high-severity accident risk.

A. Limitations

Several limitations should be acknowledged. First, severe accidents constitute only a small share of recorded crashes, creating substantial class imbalance and limiting model sensitivity for high-severity outcomes, a challenge widely reported in crash-severity modelling with machine-learning methods. Second, key categorical predictors such as detailed weather condition, road geometry, and slope include sparse or infrequent categories, which can hinder robust pattern learning and increase the risk of overfitting in tree-based models when sample sizes are not as big. Third, the analysis relies solely on reported accident records without explicit exposure metrics such as traffic volume, speed, or flow, an omission known to bias risk estimates and complicate comparison across facilities or time periods. Finally, spatial and temporal dependencies were not explicitly modelled, despite evidence that crash risk exhibits strong spatial autocorrelation and time-of-day effects, which may limit the generalisability of the results to other locations or time horizons.

B. Applications

The results of this study highlight major roads and intersections as high-risk environments for severe accidents in Bangkok, indicating clear targets for safety-focused infrastructure interventions. Improvements such as enhanced intersection design, better signage, and speed-control measures on major arterial roads could help reduce accident severity in these locations. The modelling framework developed in this work provides a data-driven method for identifying high-risk road segments and intersections, supporting more effective prioritisation of safety investments.

Recent advances in intelligent transportation systems offer further opportunities to address these risks. Smart traffic signal systems that adapt to real-time traffic conditions have been shown to significantly reduce congestion in Bangkok, which may in turn lower collision risk at intersections [8]. Integrating accident-risk prediction models with such smart traffic management systems could enable proactive monitoring and targeted interventions in the most vulnerable areas.

In addition, the strong association between motorcycle involvement and severe accidents highlights the importance of targeted rider-safety initiatives in Thailand, where motorcyclists account for a large proportion of road traffic fatalities [9]. Factors such as high-risk riding behaviour, inadequate training, and challenging road conditions contribute significantly to motorcycle accident risk. Measures such as dedicated motorcycle lanes on major roads, stricter enforcement of traffic regulations, and focused rider-safety awareness campaigns could therefore play an important role in reducing severe accident outcomes. These findings illustrate how data-driven accident analysis can inform targeted interventions to support safer urban mobility planning.

C. Future Work

Future research could address current limitations by incorporating direct exposure measures such as traffic volume, speed, and temporal variables (e.g. time of day, day of week, seasonality) to move from crash counts to exposure-adjusted risk indicators, in line with recent methodological frameworks for severity and risk modelling [7]. Advanced techniques such as resampling, cost-sensitive learning, or rare-event and anomaly-detection models may improve the identification of severe but infrequent crashes within highly imbalanced datasets. Extending the analysis to explicitly spatial and network-based models—using GIS, spatial autocorrelation measures, and road-network risk mapping—could yield finer-grained insight into localised hotspots and corridor-level patterns of risk. Finally, integrating infrastructure data with behavioural, enforcement, and vehicle-fleet information (for example, data on speeding violations, helmet and seat-belt use, and vehicle inspections) would enable more comprehensive, quasi-causal analyses of the mechanisms driving crash occurrence and severity in Bangkok and comparable urban settings.

REFERENCES

[1] World Health Organization, *Global Status Report on Road Safety 2018*, WHO Press, Geneva, 2018.

[2] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Research Part A*

[3] P. Savolainen, F. Mannering, D. Lord, and M. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident Analysis & Prevention*

[4] M. AlHashmi, "Using Machine Learning for Road Accident Severity Prediction and Optimal Rescue Pathways

[5] J.P.S Shashiprbha, R.M. Kelumi, and H.R. Psindu, "Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers,"

[6] H. Khanum,, A. garg and, M.I. Faheem, R.M. Kelum, "A methodological framework for road accident severity prediction for indian highways using machine learning models,"

[7] X. Wang, Y. Su and, Z. Zheng, "Prediction and interpretive of motor vehicle traffic crashes severity based on random forest optimised by meta-heuristic algorithm,"

[8] "Bangkok reveals smart traffic signals reduce congestion by up to 41%"

[9] "Motorbike Accidents in Thailand: A Look at the Statistics, Causes, and Prevention Measures"