# A Geo-Spatial Analysis of Unsafe Traffic Events and Crash Occurrence at Urban Intersections: Insights from Telematics Data and Machine Learning

**Stelios Peithis**
Ph.D. Student, Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: stelios_pithis@mail.ntua.gr

**Dr Paraskevi Koliou, corresponding author**
Research Associate, Senior Research Engineer
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: evi_koliou@mail.ntua.gr

**George Yannis**
Professor
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece
Email: geyannis@central.ntua.gr

Word Count: 5499 words + 2 table (500 words per table) = 5,999 words

*Submitted [13/12/2024]*

# ABSTRACT

Accurately assessing and predicting traffic crashes is essential for improving urban traffic safety and implementing targeted interventions. Traditional crash analysis relies heavily on historical crash records, which, while valuable, often face limitations such as underreporting, delays in data collection, and insufficient granularity. These challenges highlight the pressing need for innovative approaches to traffic safety analysis that are more dynamic, granular, and predictive.

This study explores the relationship between unsafe traffic events—specifically harsh braking and acceleration—and crash occurrences at 478 intersections in central Athens. Using telematics data from 2019 obtained through smartphone applications and crash records sourced from traffic police reports, the analysis integrates these datasets to examine the spatial dimensions of unsafe traffic events and their potential to classify intersections based on crash risk. The research utilises patterns and observations derived from telematics data to identify "dangerous" hotspots and evaluate the predictive capabilities of unsafe traffic event data.

Advanced statistical methods, geospatial tools, and machine learning models were employed to uncover factors influencing crash occurrences. Machine learning techniques were particularly valuable in extracting complex patterns and relationships, improving the understanding of crash risk factors at both primary and secondary road intersections. By leveraging unsafe traffic events as a supplementary resource to traditional crash records, this study highlights the value of telematics-based data for proactive safety analysis.

The findings demonstrate how telematics and smartphone applications can complement traditional crash data, enabling data-driven traffic safety management with actionable urban planning and policymaking insights to enhance road safety.

**Keywords:** Traffic safety, Crash prediction, Telematics data, Machine learning, Urban intersections

## INTRODUCTION

Road safety is a critical global challenge, representing a significant public health and socio-economic burden. Traffic crashes continue to cause substantial mortality and morbidity annually, with over 1.19 million fatalities and millions of serious injuries reported worldwide in 2021 [1]. In Europe, road safety trends reveal substantial disparities among member states, with countries like Sweden and Denmark achieving the lowest fatality rates while others like Romania and Bulgaria report significantly higher rates [2], [3]. Despite efforts to reduce road fatalities, including the European Union's Vision Zero strategy aiming to halve road deaths by 2030, progress remains slow.

Traditional traffic safety studies, till recent years, rely on historical crash data to assess risks and identify hazardous locations. While crash records provide valuable insights, their retrospective nature and limitations, such as underreporting, spatial inaccuracies, and insufficient granularity, hinder timely and effective safety interventions. Crashes are rare and inherently random events, making it challenging to collect sufficient data over short periods for robust statistical analysis [4]. Furthermore, the reliance on crash data necessitates a reactive approach to traffic safety, requiring crashes to occur before safety measures can be implemented. This underscores the pressing need for innovative, proactive approaches to traffic safety analysis.

Advancements in telematics and smartphone technologies offer a transformative opportunity to address these challenges. By capturing high-resolution data on unsafe traffic events—such as harsh braking, rapid acceleration, and near misses—smartphone applications provide a dynamic and granular perspective on driving behaviour and traffic conditions. These events occur far more frequently than crashes, serving as valuable proxies for predicting crash risk and enabling a shift from reactive to proactive safety management [5], [6].  Moreover, this real-time data can inform targeted interventions, enhance road safety planning, and support data-driven policymaking.

A unique contribution of this research lies in its integration of diverse unsafe traffic events, such as near misses and harsh acceleration/deceleration events, within a unified analytical framework. The study further explores how varying spatiotemporal resolutions impact the predictive accuracy of crash models, addressing critical gaps in the literature regarding the optimal granularity and minimum data requirements for effective traffic safety analysis. Machine learning techniques, combined with traditional statistical methods, enhance the predictive capabilities of the analysis by uncovering complex, non-linear relationships and enabling more accurate classification of high-risk intersections.

Additionally, the insights derived from unsafe traffic events extend beyond traffic safety research, with implications for industries such as car insurance. For example, programs like Pay-As-You-Drive (PAYD) promote responsible driving by linking insurance premiums to driving behaviours, such as harsh braking or acceleration, which are known precursors to crashes [7] . These advancements demonstrate the potential of telematics data to influence both public safety and commercial applications.

The study of the relationship between unsafe traffic events and crash occurrences has emerged as a critical area of research in the pursuit of improved road safety [8]. Advances in technology, particularly the widespread adoption of smartphones, have created unparalleled opportunities to collect and analyse traffic-related data in real-time [9], [10]. Smartphone applications now capture a wealth of information, including GPS locations, speed, acceleration, braking patterns, and even driver behavior metrics, such as phone usage while driving. This extensive dataset provides a granular perspective on driving habits and traffic conditions, offering valuable insights that traditional traffic studies—reliant on police reports and crash statistics—often overlook.

Leveraging smartphone data to understand unsafe traffic events is essential for identifying risky driving behaviors before they result in crashes. Indicators such as sudden braking, rapid acceleration, and sharp turns are strongly associated with aggressive driving, a known precursor to accidents [11], [12]. Analyzing these events enables the development of predictive models that can identify high-risk locations and times, facilitating proactive interventions [13]. Furthermore, real-time data can be used to educate drivers on safer driving practices, support targeted enforcement campaigns, and inform the design of more effective road safety strategies.

Road safety remains a global challenge, with traffic accidents contributing significantly to mortality and morbidity rates each year. Understanding the underlying factors behind crashes and unsafe driving behaviors is crucial for devising effective policies and interventions aimed at reducing traffic-related injuries and fatalities. Traditional approaches to traffic safety research have relied mainly on post-accident analyses based on police reports, crash statistics, and infrastructure assessments [14]. While these methods provide valuable insights, they are inherently retrospective and often constrained by data availability.

Recent advancements in big data analytics and telematics have significantly improved the ability to predict crashes by identifying patterns in risky driving behaviours. Employing methodologies such as machine learning and spatial analysis, researchers can develop predictive models to pinpoint danger zones and times, enabling proactive and timely interventions [15]. Geographic Information Systems (GIS) and temporal trend analysis further enhance this approach by visualising crash hotspots and revealing underlying patterns, providing a more comprehensive understanding of crash causation.

The integration of smartphone data into traffic safety research signifies a shift from traditional, reactive methods to a more proactive strategy [9]. Real-time monitoring of driving behaviours through smartphone applications facilitates the detection of risky activities, such as sudden braking, rapid acceleration, and distracted driving, which are common precursors to accidents. Addressing these behaviours early allows for the implementation of targeted interventions, the creation of safer traffic environments, and the development of more effective road safety measures. Additionally, this proactive approach supports driver education on safe practices, ultimately reducing the likelihood of accidents and contributing to long-term road safety improvements.

This study focuses on central Athens, an urban environment with dense road networks and complex traffic dynamics. It integrates telematics data from smartphone applications with police-reported crash records to investigate the spatial and temporal dimensions of unsafe traffic events and their relationship with crash occurrences at 478 intersections. The research uses advanced statistical techniques and machine learning models to classify intersections based on crash risk, identify critical hotspots, and explore the factors influencing crash likelihood.

Finally, the research contributes to a deeper understanding of the factors influencing crash risk by leveraging smartphone-based telematics data and advanced analytical methods [16], [17]. The findings are expected to inform urban planning, traffic management, and road safety strategies, ultimately improving road safety outcomes and reducing the socio-economic burden of traffic crashes. This proactive approach, grounded in real-time monitoring and predictive analytics, aligns with global goals to enhance road safety and mitigate the impact of traffic incidents on public health and well-being.

## METHODS

This study adopts a comprehensive and multi-faceted methodological framework to investigate the relationship between unsafe driving events and crash occurrences. By leveraging telematics data collected from a smartphone application and integrating it with crash data, the analysis focuses on 478 intersections across central Athens. The methodologies employed include clustering analysis, local spatial analysis, feature importance evaluation using machine learning techniques, multicollinearity assessment, and dimensionality reduction. These approaches enable a detailed examination of the spatial and temporal dynamics of unsafe traffic events and their correlation with crash risk. The integration of telematics and crash datasets provides a robust foundation for uncovering complex patterns, enhancing the understanding of crash risk factors, and identifying critical intersections within a dense urban network.

### Study Area and Data Sources

The study area encompasses the broader region of Athens, Greece, characterised by a dense urban road network. Road intersections, defined as points where two or more roads meet, cross, or diverge, serve as the focal points of the analysis. Data processing was conducted using OpenStreetMap (OSM) graph networks (see **Figure 1**), where nodes represent points (e.g., intersections or other significant features), and edges represent connections between nodes (e.g., roads).

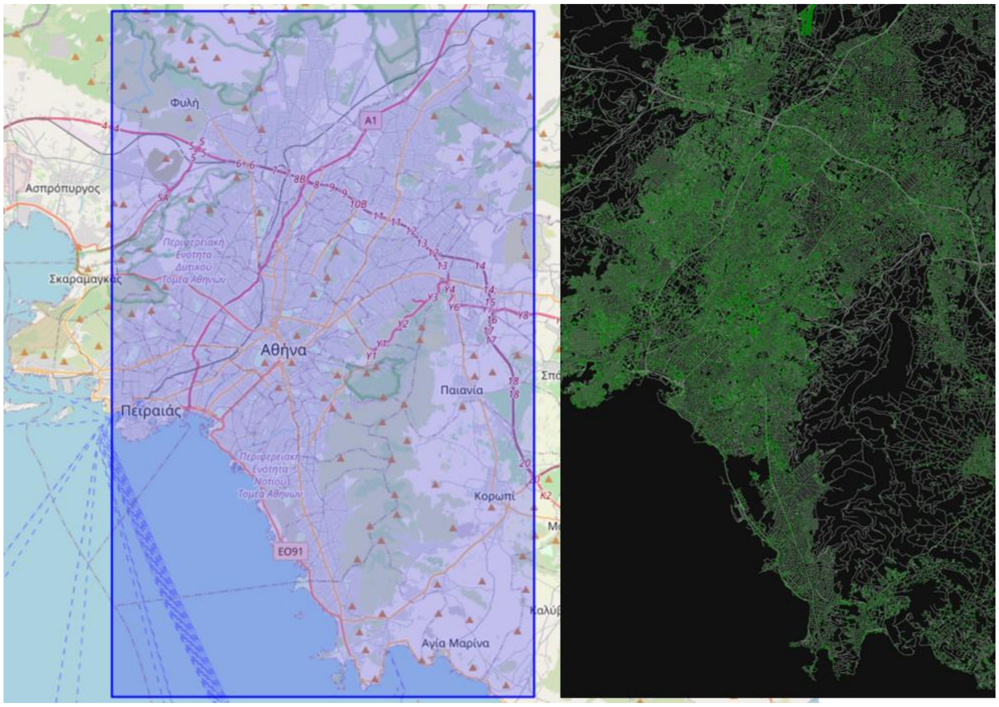

**Figure 1: Study network of Athens, used for OSM data.**

To obtain spatial data for this study, the bounding box of the Athens region was defined using latitude and longitude coordinates. This bounding box served as the area of interest for retrieving road network data through the OverPass API in conjunction with the OSMnx library. The graph_from_bbox function, configured with the network_type parameter set to “all,” was used to retrieve a comprehensive graph network of the area.

## Intersection Identification

The retrieved graph network was processed to isolate road intersections. Initially, the dataset included all nodes and edges in the specified area, including nodes representing road characteristics changes rather than actual intersections. To refine this dataset, intersections were identified using filtering methods, explaination of which can be seen in **Figure *2*** to extract the exact locations of intersections:

1. Crash Database Integration: The Greek Traffic Police Crash Database was used to identify intersections with recorded crashes. This database includes detailed information about crash occurrences, such as the names of the two streets involved. To ensure focus on unique intersections, duplicate entries based on street names were removed, resulting in 739 unique intersections.
2. Filtering for Intersection Points: Using the two street names associated with each recorded crash, the edge dataset derived from OSMnx was filtered to identify the corresponding geometries for each street (e.g., Streetname_1 and Streetname_2). GeoDataFrames were created for both streets, containing their respective geometries.
3. Spatial Overlay for Intersection Extraction: Spatial overlays were performed to determine the intersection points between the geometries of Streetname_1 and Streetname_2. The overlay function from the GeoPandas library was used for this purpose, generating a new GeoDataFrame containing the geometries of intersection points.
4. Clustering and Intersection Point Refinement: In cases where multiple geometries were identified for a single intersection, the clustering algorithm DBSCAN was applied to group the points and compute centroids representing the precise intersection locations. This ensured that a single point described

each intersection. This ensured an accurate representation of intersection locations, ultimately yielding 530 intersections.

The identified intersections were visualised on a Folium map to verify their accuracy and alignment with the crash records; see **Figure *2*** .

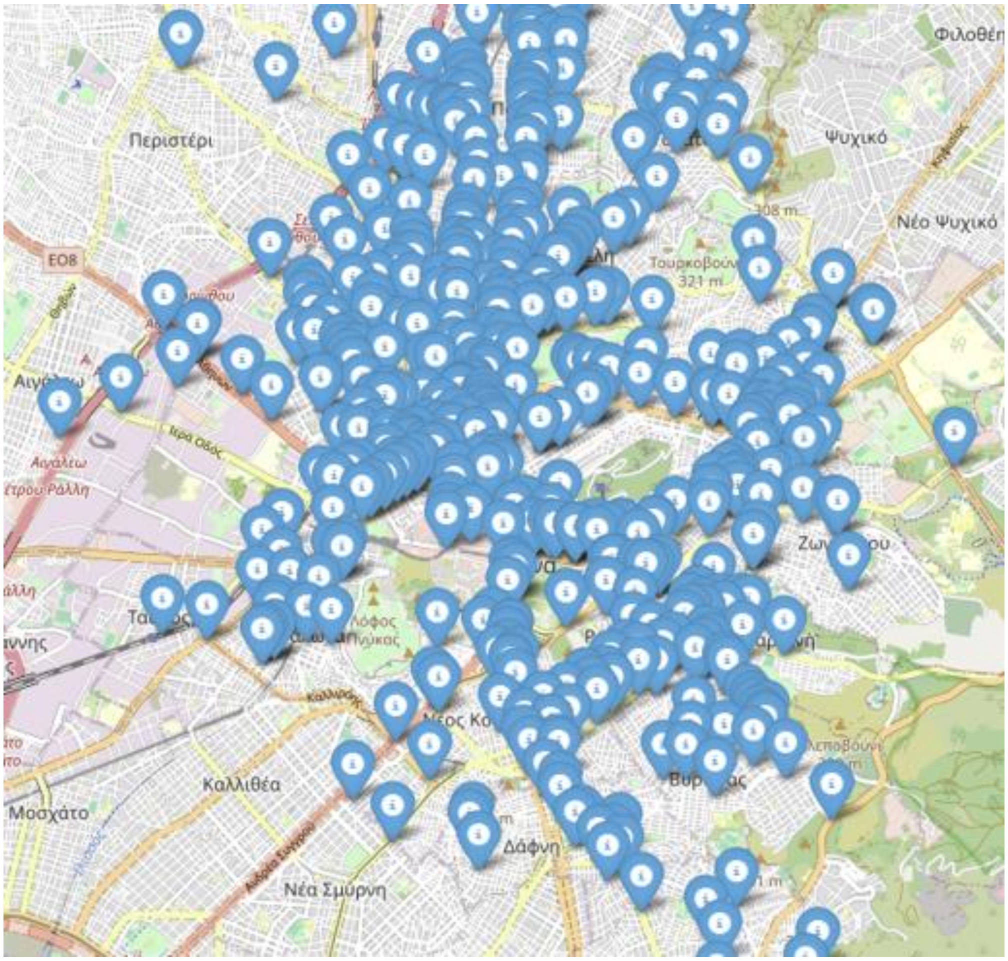

**Figure 2: Plot of Crashes that occurred in intersections on a Folium Map.**

**Telematics Data**

Telematics data were collected from a smartphone application that monitors driver’s behaviour. This dataset included:

- Metrics Recorded: Harsh braking, harsh acceleration, and speeding events were recorded for each driver's trip.
- Dataset Overview: The data comprised 257 unique drivers completing 2,615 trips over two months, resulting in over 3.5 million data points. Most recorded events were non-harsh braking or non-speeding, reflecting the traffic conditions of central Athens.

The telematics data provided high-resolution spatial insights into driving behaviours, offering a granular view of unsafe driving events (***Figure 3***).

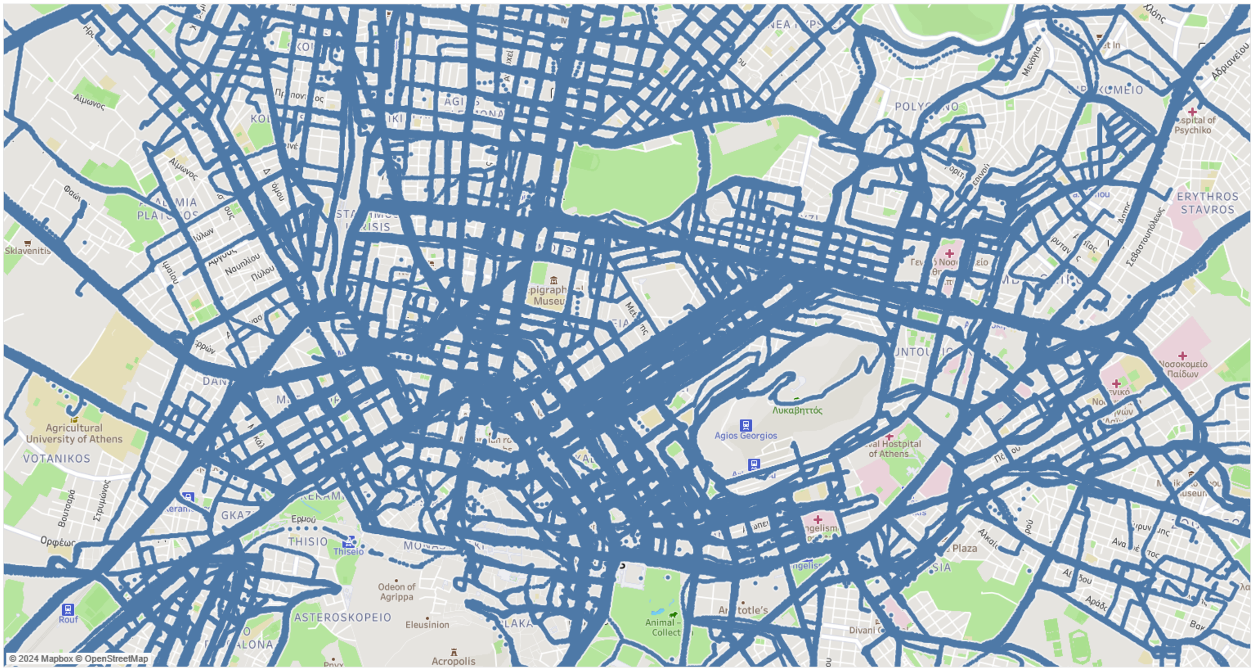

**Figure 3: Telematics data collected in Athens urban roads.**

### Assignment of Telematics Data to Intersections

The dataset was further enhanced by integrating crash records with telematics data derived from smartphone applications. This telematics data included unsafe traffic events, such as harsh braking and acceleration, recorded with high spatial resolution. By associating unsafe traffic event data with identified intersections, the study enabled spatial analyses of crash occurrences and their correlation with unsafe driving behaviours.

To correlate unsafe driving events with crash occurrences, telematics data were assigned to intersections (see **Figure 4**):

- Intersection Influence Area: Each intersection was represented by its centroid, with a 50-meter radius defined as its influence area. This radius was chosen based on best practices in the literature.
- Data Aggregation: Telematics events falling within the defined radius of each intersection were aggregated and linked to the corresponding intersection. This facilitated evaluating how unsafe driving behaviours correspond to crash counts at specific locations.

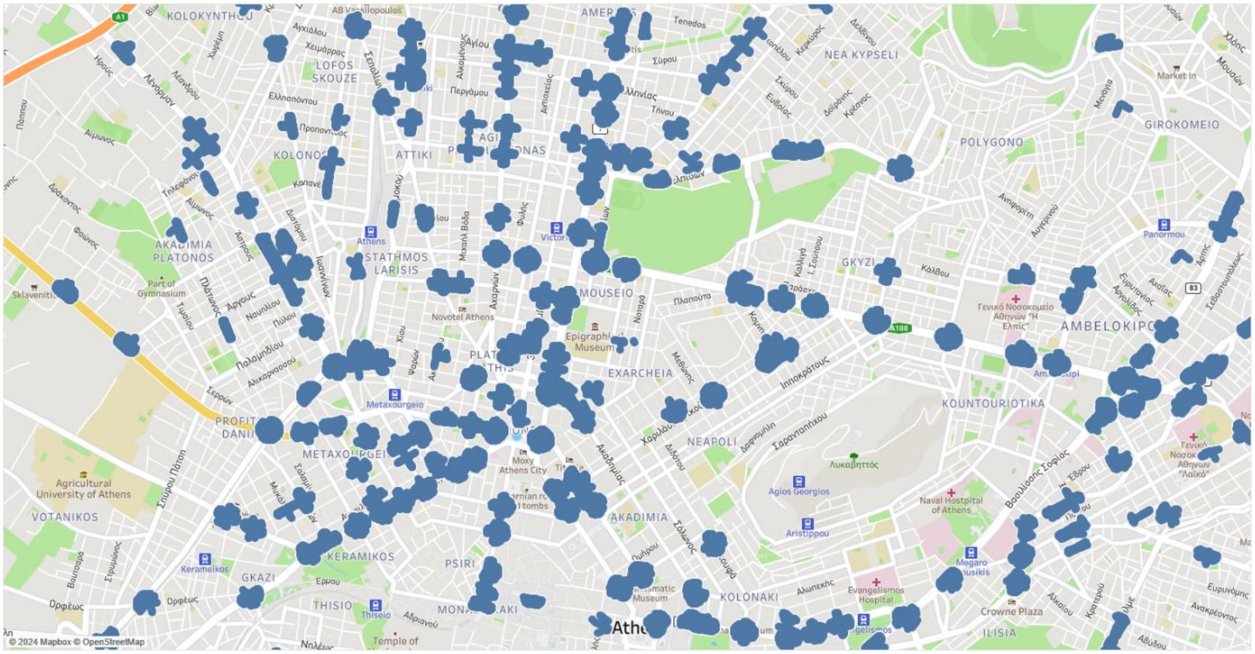

**Figure 4: Telematics data assigned to intersection with centroids.**

## Statistical and Machine Learning Analysis

A combination of statistical and machine learning methods was employed to analyse the data:

- Statistical Analysis: Descriptive statistics were computed to summarise the spatial distributions of crashes and unsafe driving events. Correlations between telematics metrics and crash occurrences were also examined.
- Geospatial Analysis: Spatial patterns of crashes and unsafe driving events were analysed using local spatial statistics to identify high-risk areas and "dangerous" hotspots.
- Machine Learning Techniques:
  - Feature Importance: Algorithms such as Random Forests and Gradient Boosting were used to evaluate the importance of telematics features (e.g., harsh braking, acceleration) in predicting crash risk.
  - Dimensionality Reduction: Principal Component Analysis (PCA) was applied to reduce multicollinearity among features and streamline the dataset for predictive modelling.
  - Clustering: DBSCAN and other clustering techniques were used to group high-risk intersections based on similar patterns of unsafe driving behaviours and crash occurrences.

The combination of statistical and machine learning approaches allowed for identifying complex relationships and developing predictive models for crash risk at intersections.

## Model Evaluation

The predictive models were evaluated using standard metrics such as accuracy, precision, recall, and F1-score. These metrics ensured the reliability of the models in classifying intersections based on crash risk. Insights from the models were used to identify key factors influencing crash occurrences and to validate the role of unsafe driving events as predictors.

This multi-layered methodology ensured a robust analysis of the relationship between unsafe driving events and crash occurrences, leveraging cutting-edge tools and techniques to provide actionable insights for traffic safety management.

## RESULTS

### Data Processing and Summary Statistics

After integrating telematics data with crash records, a comprehensive dataset comprising 929,543 data points was created. These data points were spatially joined to intersections using a 50-meter buffer radius around each intersection's centroid. This spatial association ensured that the telematics events were accurately linked to their corresponding intersections, enabling more precise analysis of unsafe driving behaviours in proximity to crash-prone areas.

The telematics data included detailed driver behaviour metrics such as harsh braking, harsh acceleration, and speeding events. To facilitate analysis, the dataset was aggregated at the intersection level. Key aggregated metrics were computed for each intersection, including the total number of trips, the frequency of harsh events, speeding flags, and the number of lanes.

Three key ratios were derived to quantify unsafe driving behaviours at intersections:

- Harsh Braking Ratio: This metric was calculated as the ratio of harsh braking events to the total number of trips recorded at an intersection, reflecting the prevalence of this behaviour in the area.
- Harsh Acceleration Ratio: Similarly, this ratio represented the proportion of trips with harsh acceleration events, providing insights into aggressive driving patterns.
- Speeding Ratio: This metric was computed as the ratio of speeding flags to total trips, highlighting intersections where drivers frequently exceeded speed limits.

To enrich the dataset, crash data from the Greek Traffic Police database were integrated. This data included the count of crashes for each intersection, which was subsequently categorised into three risk levels: Low, Medium, and High. This categorisation provided a clear framework for analysing crash risk and its relationship with unsafe driving behaviours at intersections.

The resulting dataset was the foundation for further analysis, enabling a comprehensive understanding of how telematics-based driving behaviours correlate with crash occurrences across urban intersections.

### Feature Distributions

The distributions of key features provided valuable insights into driving behaviour and intersection characteristics. The Harsh Braking Ratio, which represents the proportion of trips involving harsh braking events, showed that most intersections exhibited low ratios. This indicates that harsh braking was not a widespread behaviour across the majority of intersections. However, a small subset of intersections displayed higher harsh braking ratios, suggesting localised issues potentially linked to road design, traffic flow, or driver habits.

The Speeding Ratio, calculated as the proportion of trips involving speeding events, revealed a similar pattern. Speeding events were generally uncommon across intersections, with the majority of intersections showing low speeding ratios. However, some intersections showed higher concentrations of speeding, possibly indicating areas where road characteristics or speed limits were less effective at curbing speeding behaviours.

Finally, the Number of Lanes distribution highlighted that intersections with one or two lanes were the most prevalent. Intersections with three or more lanes were less common, reflecting the urban structure of central Athens, where narrow roads and intersections exist in the majority. These insights into the number of lanes help contextualise driving behaviour metrics, as wider intersections might accommodate higher speeds or more complex traffic patterns, influencing unsafe driving behaviours.

Overall, these feature distributions provide a foundational understanding of the variability in driving behaviours and infrastructure characteristics across intersections, setting the stage for further analysis of their relationship with crash occurrences. The descriptive results of which can be seen in the **Figure 5** below.

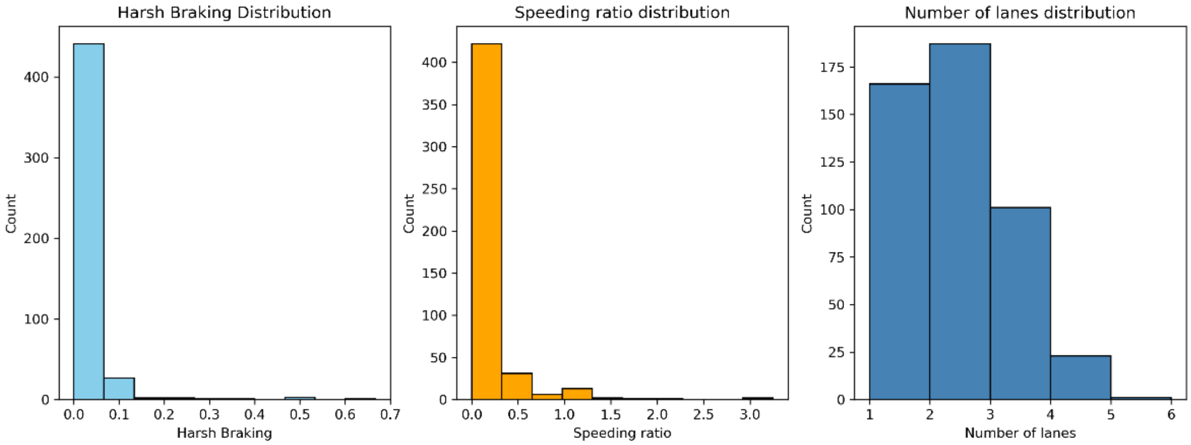

**Figure 5: Features Distribution without oversampling**

### Crash Risk Classification

The crash risk classification process categorized intersections into three distinct levels based on the number of recorded crashes: **Low Risk**, **Medium Risk**, and **High Risk**. Intersections with only one recorded crash were classified as **Low Risk**, making up the majority of the dataset with 398 intersections. **Medium Risk** intersections, defined as those with 2–3 crashes, accounted for 70 intersections. Lastly, **High Risk** intersections, with 4 or more crashes, were the least frequent, comprising only 10 intersections. This classification reflects the natural imbalance in crash data, with most intersections experiencing few crashes and a small subset identified as critical hotspots.

The initial class distribution revealed significant imbalances across the three risk categories, as shown in the left panel of **Figure *6***. Low-risk intersections were outnumbered the dataset, while high-risk intersections were underrepresented. This imbalance created a challenge for predictive modelling, as machine learning algorithms typically perform poorly with skewed class distributions.

To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset. SMOTE oversamples the minority classes (Medium and High Risk) by generating synthetic examples based on their feature space, resulting in an even distribution across all categories. The right panel of Figure 6 illustrates the balanced class distribution after applying SMOTE, where the number of intersections in each risk category is equalised. This step ensured that the predictive models were trained on a balanced dataset, improving their ability to classify intersections accurately across all risk levels. This balanced dataset was subsequently used in the modelling phase, enabling the development of robust models capable of identifying intersections at higher risk of crashes.

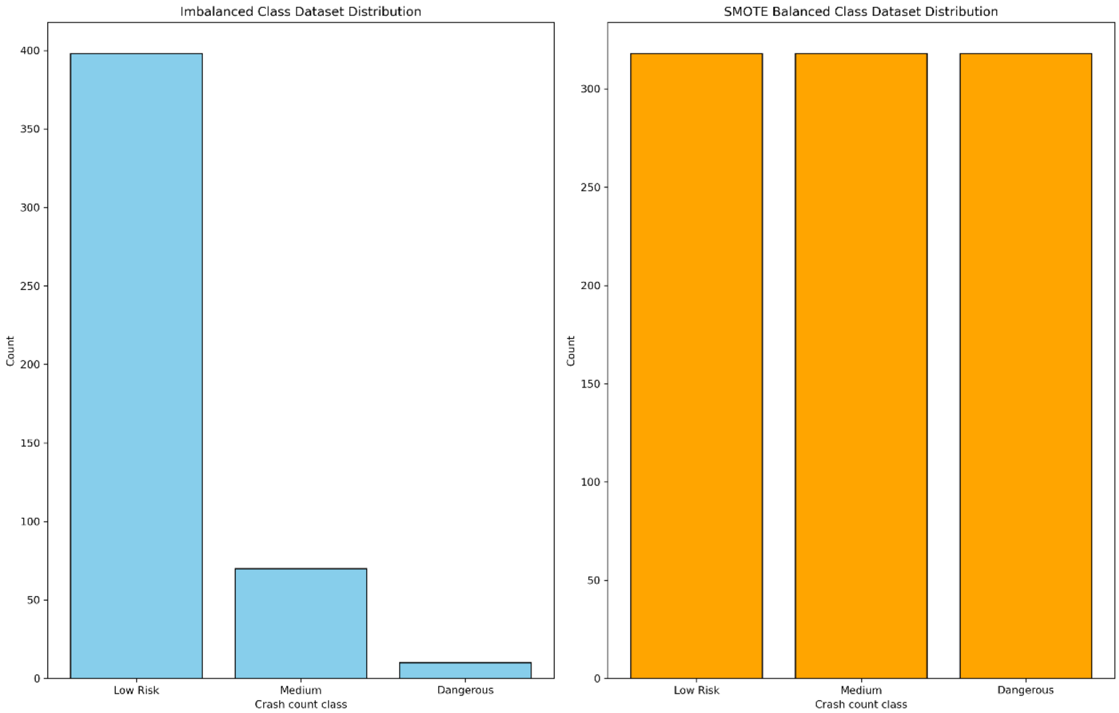

**Figure 6: Class of Risk Distribution before and after SMOTE.**

## Predictive Modelling

*Random Forest Classifier*

The Random Forest classifier was employed to predict crash risk categories for intersections using key features such as the harsh braking ratio, speeding ratio, and the number of lanes. This model demonstrated a robust overall performance with an accuracy of 76%. However, the performance varied across the three crash risk categories due to the inherent imbalance in the dataset, which persisted despite applying SMOTE.

For Low-Risk intersections, the model achieved a high precision of 93% and a recall of 80%, reflecting its ability to correctly classify most intersections with minimal crash risk. In contrast, the performance for Medium Risk intersections was moderate, with a precision of 40% and a recall of 57%, indicating challenges in accurately identifying intersections with 2–3 crashes. The model struggled the most with High-Risk intersections, achieving a precision of only 14% and a recall of 50%. This limitation highlights the difficulty of accurately predicting intersections with 4 or more crashes, even with oversampling techniques.

**Table 1: Random Forest Classifier Performance Summary**

| **Crash Risk Category** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Low Risk | 93% | 80% | High |
| Medium Risk | 40% | 57% | Moderate |
| High Risk | 14% | 50% | Low |

The confusion matrix, see **Figure 7**, revealed that the model performed exceptionally well in correctly classifying low-risk intersections but frequently misclassified medium and high-risk intersections, potentially due to overlapping feature distributions and the relatively small size of the high-risk class

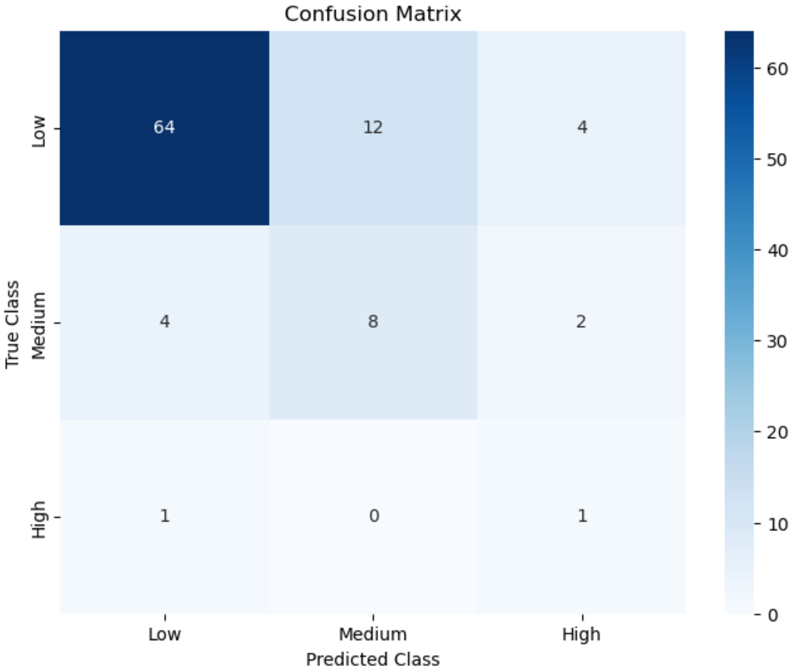

**Figure 7: Confusion matrix for the model performance, with Random Forest Classifier**

This analysis underscores the importance of further refining the feature set and exploring additional modelling techniques to improve the classification of medium- and high-risk intersections. Despite these challenges, the Random Forest classifier provided valuable insights into the key factors influencing crash risk at intersections.

*XGBoost Classifier*

The XGBoost classifier was implemented to enhance the predictive performance of crash risk classification. This advanced model improved overall accuracy by 80%, outperforming the Random Forest classifier. The enhanced performance highlights the strength of XGBoost in handling complex patterns and relationships within the dataset, particularly for the high-risk category.

For Low-Risk intersections, the XGBoost classifier achieved a precision of 94% and a recall of 84%, demonstrating its reliability in accurately classifying intersections with minimal crash risk. The performance for Medium Risk intersections remained moderate, with a precision of 44% and a recall of 57%, similar to the Random Forest classifier. However, the most significant improvement was observed for High-Risk intersections. The XGBoost model achieved a precision of 29% and a recall of 100%, successfully identifying all high-risk intersections in the test dataset.

**Table 2: XGBoost Classifier Performance Summary**

| **Crash Risk Category** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Low Risk | 94% | 84% | High |
| Medium Risk | 44% | 57% | Moderate |
| High Risk | 29% | 100% | Moderate |

The confusion matrix revealed that the XGBoost classifier made significant strides in correctly identifying high-risk intersections, addressing a critical limitation observed in the Random Forest model. However, the classification of medium-risk intersections continued to present challenges, likely due to feature overlaps and class complexity.

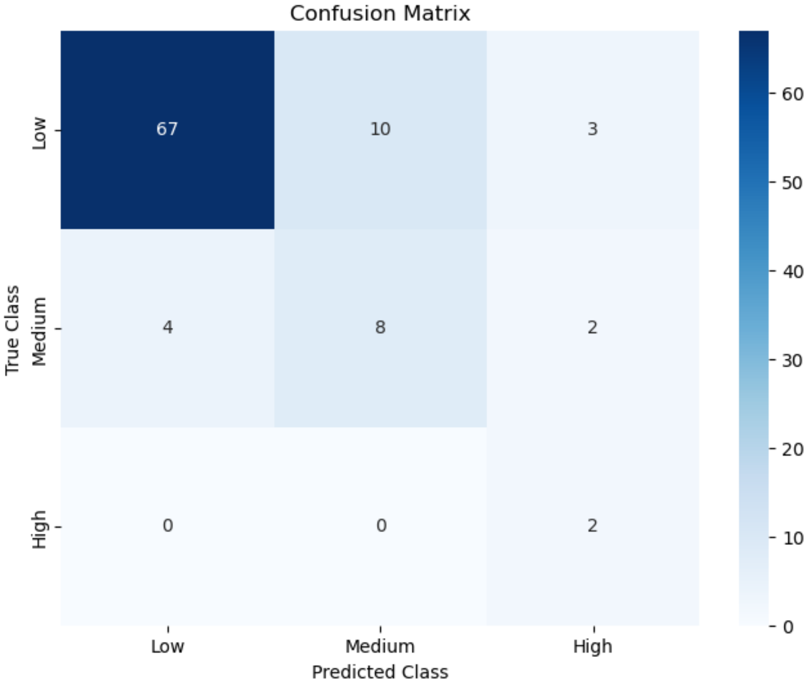

**Figure 8: Confusion matrix for the model performance, with XGBoost classifier.**

The XGBoost classifier improved overall accuracy and class-specific performance, particularly for high-risk intersections. These results underline its suitability for crash risk prediction and potential to identify critical hotspots for proactive safety interventions.

### Feature Importance

The feature-important analysis provided valuable insights into the contribution of various predictors to crash risk classification. Among the features examined, the Harsh Braking Ratio emerged as the most significant predictor of crash risk. This highlights the strong correlation between frequent harsh braking events and crash-prone intersections, reinforcing the importance of this unsafe driving behaviour in identifying high-risk locations.

The Speeding Ratio was identified as the second most important feature in the analysis. While less critical than harsh braking, speeding was particularly relevant for high-speed intersections, where drivers are likelier to exceed speed limits. This behaviour increases the likelihood of severe crashes, making speeding an essential factor for targeted safety interventions.

The Number of Lanes was found to be less predictive overall but still played a meaningful role in distinguishing low-risk intersections. Wider intersections with more lanes might experience greater traffic complexity. Still, they do not necessarily correlate directly with crash risk in the same way that driver behaviours like harsh braking and speeding do.

### Model Evaluation

The evaluation of the classification models highlighted the potential of telematics-based data in enhancing the understanding of crash risk factors at urban intersections. Integrating telematics data, such as harsh braking and speeding events, provided a more dynamic and granular approach to analysing crash risks compared to traditional crash data alone.

Between the models, the XGBoost classifier demonstrated superior performance, achieving higher overall accuracy and excelling in classifying high-risk intersections. The model's ability to correctly identify all high-risk intersections reflects its effectiveness in capturing complex patterns and relationships within the dataset. By comparison, the Random Forest model showed strong performance for low-risk intersections but struggled with medium- and high-risk categories.

Despite these advances, challenges remained with imbalanced class distributions, particularly for medium-risk intersections. This issue suggests the need for further techniques, such as advanced oversampling methods, enhanced feature engineering, or ensemble modelling, to improve classification performance for this category. These findings underscore the critical role of telematics data in proactive crash risk management and highlight areas for future research and methodological improvement.

## DISCUSSION

This study demonstrates the value of integrating telematics data with crash records to advance urban traffic safety analysis. The results highlight the importance of unsafe driving behaviors, particularly harsh braking, as critical predictors of crash risk. By leveraging telematics data, this study offers a proactive approach to understanding and mitigating traffic risks, moving beyond the traditional reliance on historical crash data.

The findings underscore the capability of telematics data to provide a more granular and dynamic view of driving behaviours at urban intersections. The spatial assignment of telematics events, such as harsh braking and speeding, enabled the identification of patterns that correlate strongly with crash occurrences. The Harsh Braking Ratio was consistently the most significant predictor of crash risk, emphasising the need for targeted interventions at intersections with frequent braking events. Meanwhile, speeding, although less prevalent, played an important role in identifying high-risk intersections, particularly those with higher speed limits.

Applying machine learning techniques, particularly XGBoost, proved effective in uncovering complex relationships between unsafe driving behaviours and crash risks. The model's ability to correctly identify all high-risk intersections demonstrates the value of advanced modelling techniques in traffic safety research. However, the challenges encountered with medium-risk intersections suggest that further refinement is needed in feature selection and class balancing techniques.

The imbalance in crash risk categories, even after applying SMOTE, presented difficulties in accurately classifying medium-risk intersections. This reflects a broader challenge in traffic safety research, where high variability in crash data can obscure patterns for less frequent but significant categories. Addressing this issue requires further exploration of advanced data augmentation techniques or ensemble models to improve classification performance.

This research also has implications beyond traffic safety. The insights into unsafe driving behaviours have potential applications in urban planning, traffic management, and commercial sectors like car insurance. For instance, programs linking insurance premiums to driving behaviours could benefit from the granular metrics provided by telematics data.

While this study successfully demonstrates the potential of telematics data for crash risk prediction, several limitations must be addressed in future research. The reliance on data from a single urban area (Athens) limits the generalizability of the findings. Expanding the analysis to other cities with varying road infrastructures and traffic conditions would provide a more comprehensive understanding of the relationship between telematics events and crash risks. Additionally, incorporating other contextual factors, such as weather conditions and traffic volume, could enhance the predictive power of the models.

In conclusion, this study highlights the transformative potential of telematics data for proactive traffic safety management. By integrating high-resolution behavioural data with crash records, this research offers actionable insights for reducing crash risks and improving urban road safety. Future work should focus on refining predictive models and exploring broader applications of telematics data in transportation and safety planning.

## CONCLUSIONS

This study highlights the transformative potential of integrating telematics data with traditional crash records to enhance urban traffic safety analysis. By focusing on unsafe driving behaviours, such as harsh braking and speeding, and their spatial association with crash occurrences, the research provides a proactive framework for identifying high-risk intersections and informing targeted interventions.

The findings underline the significant role of telematics-derived metrics, particularly the Harsh Braking Ratio, as critical predictors of crash risk. The incorporation of advanced machine learning models, such as XGBoost, demonstrated strong performance in classifying intersections by crash risk, particularly for high-risk categories. However, challenges with medium-risk intersections emphasise the need for further methodological advancements to handle class imbalances and refine feature selection.

This research has important implications for traffic management, urban planning, and road safety strategies. By leveraging granular behavioural data, policymakers and practitioners can implement more effective safety interventions, design safer road environments, and enhance traffic monitoring systems. Moreover, the insights derived from telematics data extend beyond traffic safety, offering potential applications in fields such as insurance and driver education.

Despite its contributions, this study is not without limitations. The analysis was confined to intersections in central Athens, which may limit the generalizability of the findings. Future research should explore the application of this framework in diverse urban and rural settings and incorporate additional contextual factors, such as weather conditions, traffic density, and road design characteristics.

In conclusion, integrating telematics data and crash records represents a significant advancement in proactive traffic safety management. This study lays the groundwork for data-driven approaches to reducing crash risks and improving road safety in urban environments by providing actionable insights and identifying key risk factors. Future efforts should focus on expanding the scope of analysis and further refining predictive models to maximise their utility in real-world applications.

## ACKNOWLEDGMENTS

Funded by the Basic Research Program PEVE 2021 of the National Technical University of Athens.

## AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: study conception and design: Stelios Peithis, Paraskevi Koliou, and George Yannis; Data analysis and programming: Stelios Peithis; Interpretation of results: Paraskevi Koliou; Draft manuscript preparation: Paraskevi Koliou and Stelios Peithis. All authors reviewed the results and approved the final version of the manuscript.

# REFERENCES

[1] "World Health Organization. (2023). Global status report on road safety 2023. World Health Organization."

[2] "European Commission. (2023). Road safety: 20,640 people died in a road crash last year – progress remains too slow. Available online: https://transport.ec.europa.eu/news-events/news/road-safety-20640-people-died-road-crash-last-year-progress-remains-too-slow-2023-10-19_en. (Accessed on 07 November 2023).".

[3] "European Transport Safety Council. (2021). 15th Annual Road Safety Performance Index (PIN) Report; ETSC: Brussels, Belgium, 2021."

[4] "Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. Accident Analysis & Prevention, 130, 151-159.".

[5] "Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. Accident Analysis & Prevention, 135, 105323.".

[6] "Tarko, A. P. (2018). Surrogate measures of safety. In Safe mobility: challenges, methodology and solutions (Vol. 11, pp. 383-405). Emerald Publishing Limited.".

[7] "Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis & Prevention, 98, 139-148.".

[8] "Shah, S. A. R., & Ahmad, N. (2020). Accident risk analysis based on motorway exposure: an application of benchmarking technique for human safety. International journal of injury control and safety promotion, 27(3), 308-318.".

[9] V. Petraki, A. Ziakopoulos, and G. Yannis, "Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data," *Accid Anal Prev*, vol. 144, Sep. 2020, doi: 10.1016/j.aap.2020.105657.

[10] T. E. Glasgow, H. T. K. Le, E. Scott Geller, Y. Fan, and S. Hankey, "How transport modes, the built and natural environments, and activities influence mood: A GPS smartphone app study," *J Environ Psychol*, vol. 66, Dec. 2019, doi: 10.1016/j.jenvp.2019.101345.

[11] E. Papadimitriou, A. Argyropoulou, D. I. Tselentis, and G. Yannis, "Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving," *Saf Sci*, vol. 119, pp. 91–97, Nov. 2019, doi: 10.1016/j.ssci.2019.05.059.

[12] A. Botzer, O. Musicant, and Y. Mama, "Relationship between hazard-perception-test scores and proportion of hard-braking events during on-road driving – An investigation using a range of thresholds for hard-braking," *Accid Anal Prev*, vol. 132, Nov. 2019, doi: 10.1016/j.aap.2019.105267.

[13] J. Stipancic, L. Miranda-Moreno, and N. Saunier, "Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers," *Accid Anal Prev*, vol. 115, pp. 160–169, Jun. 2018, doi: 10.1016/j.aap.2018.03.005.

[14] D. Nikolaou, A. Ziakopoulos, and G. Yannis, “A Review of Surrogate Safety Measures Uses in Historical Crash Investigations,” May 01, 2023, *MDPI*. doi: 10.3390/su15097580.

[15] J. Rane, Ö. Kaya, S. K. Mallick, and N. L. Rane, “Artificial intelligence-powered spatial analysis and ChatGPT-driven interpretation of remote sensing and GIS data,” in *Generative Artificial Intelligence in Agriculture, Education, and Business*, Deep Science Publishing, 2024. doi: 10.70593/978-81-981271-7-4_5.

[16] R. Tamakloe, K. Zhang, A. Hossain, I. Kim, and S. H. Park, “Critical risk factors associated with fatal/severe crash outcomes in personal mobility device rider at-fault crashes: A two-step inter-cluster rule mining technique,” *Accid Anal Prev*, vol. 199, May 2024, doi: 10.1016/j.aap.2024.107527.

[17] J. H. R. Adrian Sandt, Haitham Al-Deek, “Identifying Wrong-Way Driving Hotspots by Modeling Crash Risk and Analyzing Traffic Management Center Response Times,” *Transp Res Rec*, no. No. 17-05009., 2017.