Mapping Risk: Leveraging Telematics and Machine Learning to Analyze Crash Risks at Urban Intersections

Stelios Peithis (stelios_pithis@mail.ntua.gr), Paraskevi Koliou (evi_koliou@mail.ntua.gr) and George Yannis (geyannis@central.ntua.gr)

Abstract

This research explores the spatial and temporal correlations between unsafe driving behaviours ,characterised by harsh braking, rapid acceleration, and other telematics-based events, and traffic crashes at 439 urban intersections in central Athens. Utilizing a comprehensive dataset comprising telematics data from 2019, gathered via smartphone applications, and multi-year crash records obtained from police reports, this study integrates these sources to classify intersections by crash risk and delineate high-risk "danger zones." Employing advanced machine learning techniques, including Random Forest, XGBoost, Gradient Boosting Machines, and Logistic Regression, the analysis rigorously evaluates the predictive capacity of telematics-derived metrics in forecasting crash risks. The study leverages cutting-edge geospatial analytics to uncover critical patterns that link dynamic driving behaviours to elevated crash probabilities. These findings not only advance the methodological framework for crash risk prediction but also provide actionable insights for urban traffic safety management. By enabling proactive identification of high-risk intersections, this research contributes to the design of targeted, data-driven interventions aimed at mitigating crash risks and enhancing road safety in complex urban environments.

Keywords: *Urban Road Safety; Telematics; Crash Prediction; Machine Learning; Intersection Analysis; Geospatial Analytics; Smart Mobility*

1. Introduction

Road traffic collisions constitute a principal source of mortality and morbidity in urban environments, exacting profound human, economic, and societal costs. In densely inhabited cities such as Athens, the confluence of elevated vehicular volumes, restricted roadway capacity, and heterogeneous driver behaviours presents formidable challenges to conventional safety interventions. The advent of vehicular telematics and smartphone-based sensing technologies, along with the steep rise of Internet of Things technologies, affords granular insight into driver behaviour, specifically, instances of severe braking, rapid acceleration, and related events. Although extant literature has linked individual drivers' frequencies of harsh manoeuvres to their overall crash involvement, systematic analyses that spatially correlate telematics-derived indicators with intersection-level risk within an urban road network are scarce.

This paper addresses the correlation between incidents of intense driving events, such as harsh braking, and their proximity to intersections that tend to have frequent crashes. The main scope of this research is to investigate the efficacy of applications of machine learning algorithms in the proactive identification of road safety hotspots. Harnessing sophisticated geospatial analytics, this study aims to deepen the understanding of driving behaviours in relation to intersections of the road network, potentially

identifying metrics and patterns that conventional crash-count methodologies may not be able to pinpoint.

Urban road traffic accidents (URTAs) have emerged as a critical domain of study in public health and urban planning, given their profound human, social, and economic consequences. With the rapid pace of urbanisation and motorisation, cities have become epicentres of road traffic injuries and fatalities, reflecting a complex interplay of human, environmental, and infrastructural factors. Empirical research underscores that URTAs are often concentrated at hazardous locations, where deficiencies in road design, traffic management, and safety infrastructure converge to exacerbate accident risks.

The determinants of urban traffic accidents are multifaceted, encompassing behavioural factors such as speeding, impaired driving, and non-compliance with traffic regulations, as well as structural variables including road geometry, lighting conditions, and traffic flow dynamics. For instance, in Zagreb, Croatia, excessive speed, male gender, and night-time driving have been identified as key contributors to fatal accidents, with urban links proving more dangerous for fatalities. In contrast, urban junctions see a higher prevalence of non-fatal injuries (Vorko-Jović et al., 2006). Vulnerable road users, such as pedestrians and cyclists, face heightened risks due to insufficient safety measures, particularly in low- and middle-income countries, where the burden of RTAs is disproportionately high (Gopalakrishnan, 2012). (Greibe, 2003), Focusing on predictive accident models for urban road links and junctions, the study demonstrates that road width, lane configuration, and motor vehicle flow are significant predictors of accident frequency. (Obaidat & Ramadan, 2012) Further elucidate the role of geometric elements and environmental conditions in shaping accident patterns at hazardous locations, offering in-sights into effective mitigation strategies.

Beyond their immediate impact on human lives, road traffic accidents impose substantial societal costs, including loss of productivity, healthcare expenditures, and long-term rehabilitation needs. Vulnerable populations, particularly in urban areas, suffer disproportionately, with economically active age groups often constituting the majority of victims. Despite the significant progress achieved in high-income countries through advanced safety systems and strict law enforcement, developing regions continue to witness alarming trends in road traffic injuries, necessitating a multidisciplinary and data-driven approach to road safety.

The literature emphasises the necessity of adopting predictive models to identify high-risk areas and implementing targeted interventions. (Greibe, 2003) models, for example, account for more than 60% of the systematic variation in accidents on urban road links, underscoring the value of integrating geometric and traffic flow variables in safety assessments. Similarly, Obaidat & Ramadan (2012) propose specific infrastructural modifications, including enhanced pedestrian crossings, optimised horizontal curves, and improved lighting, as effective counter-measures for accident-prone urban locations.

This paper situates itself within the broader discourse on urban traffic safety by synthesising key findings from existing research and analysing critical factors that contribute to the prevalence of accidents in urban areas. It aligns with global initiatives, such as the World Health Organisation’s "Decade of Action for Road Safety," advocating for a holistic approach that integrates infrastructure improvements, behavioural interventions, and robust traffic law enforcement. By addressing the systemic challenges of urban road safety, this study aims to provide actionable insights to policymakers, urban planners, and public health professionals, fostering safer and more sustainable urban environments.

2. Main Text

2.1 Data Collection

The research uses two separate databases in order to construct a research framework that captures the complex relationship between driving characteristics and the risk of crashes in interventions. A database of crash data was deployed. This repository comprises all officially recorded traffic-collision incidents within the jurisdiction of the Greek Traffic Police for the calendar year 2019. Each entry includes the precise date and time of occurrence, as well as the names of the bounding streets as text that define the intersection at which the crash occurred. In order to fully leverage the information provided in this database, one key part of this research is the mapping process of crash locations recorded as plain street names in the database to actual geographical coordinates. This process is also referred to as geospatial mapping, and the OpenStreetMap database and OverPass API are used in order to extract the exact coordinates that shape each street as a path.

Figure 1. Database Framework

In addition to the crash database, a broad telematics data dataset was used to capture the driving characteristics near each intersection. Collected via a widely deployed smartphone-based telematics application, this dataset encompasses continuous trip records for vehicles operating throughout the broader Athens metropolitan area during 2019. Each record contains high-frequency GPS coordinates, timestamps, and derived behavioural metrics (e.g., harsh-braking and rapid-acceleration events). A feature table of aggregated measures of intense driving incidents is produced by spatially joining each telematics trip segment to the nearest intersection based on the proximity of latitude and longitude. These aggregated indicators then function as a feature table in the deployed machine learning models.

2.2 Descriptive Statistics

A comprehensive descriptive analysis of the crash dataset is conducted prior to predictive modelling. This section synthesizes key summary statistics that elucidate the distributional properties of crash occurrences across the 439 urban intersections, such as total crash counts and spatial dispersion by street segment.

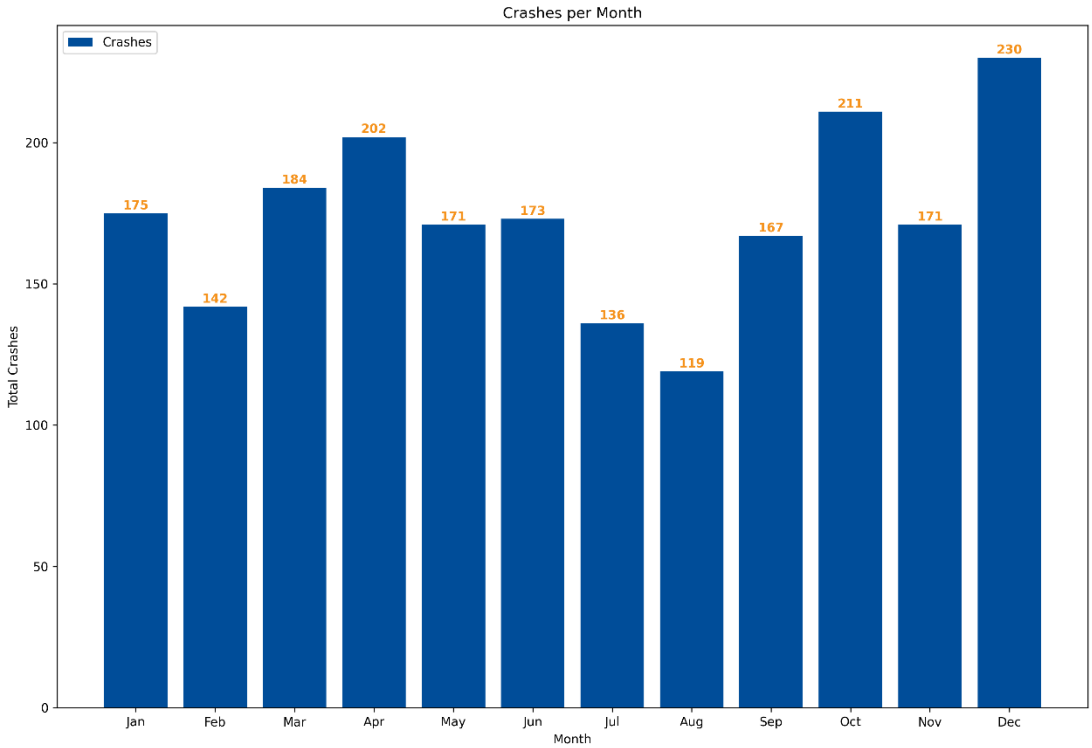

Figure 2. Monthly distribution of intersection crashes in the Athens area.

Figure 2 illustrates the monthly distribution of recorded collisions across the 439 study intersections during 2019. A pronounced seasonal pattern is evident: crash incidence is lowest in the summer months, reaching an all-year low in August, and increases sharply thereafter, peaking in December. Notably, the first quarter demonstrates an upward trend from February to April, followed by a mid-year decline. This cyclicality may reflect variations in traffic volume, weather conditions, daylight hours, and holiday-related travel behaviours. The elevated crash counts observed in October and December suggest heightened risk during increased traffic flow and potentially adverse meteorological conditions. These findings underscore the importance of accounting for temporal heterogeneity in descriptive and predictive urban crash risk analyses.

Table 1: Descriptive Statistics for the Crash sample

Statistics	Crash_Count
count	1837.00
mean	1.13
std	0.42
min	1.00
25%	1.00
50%	1.00
75%	1.00
max	6.00

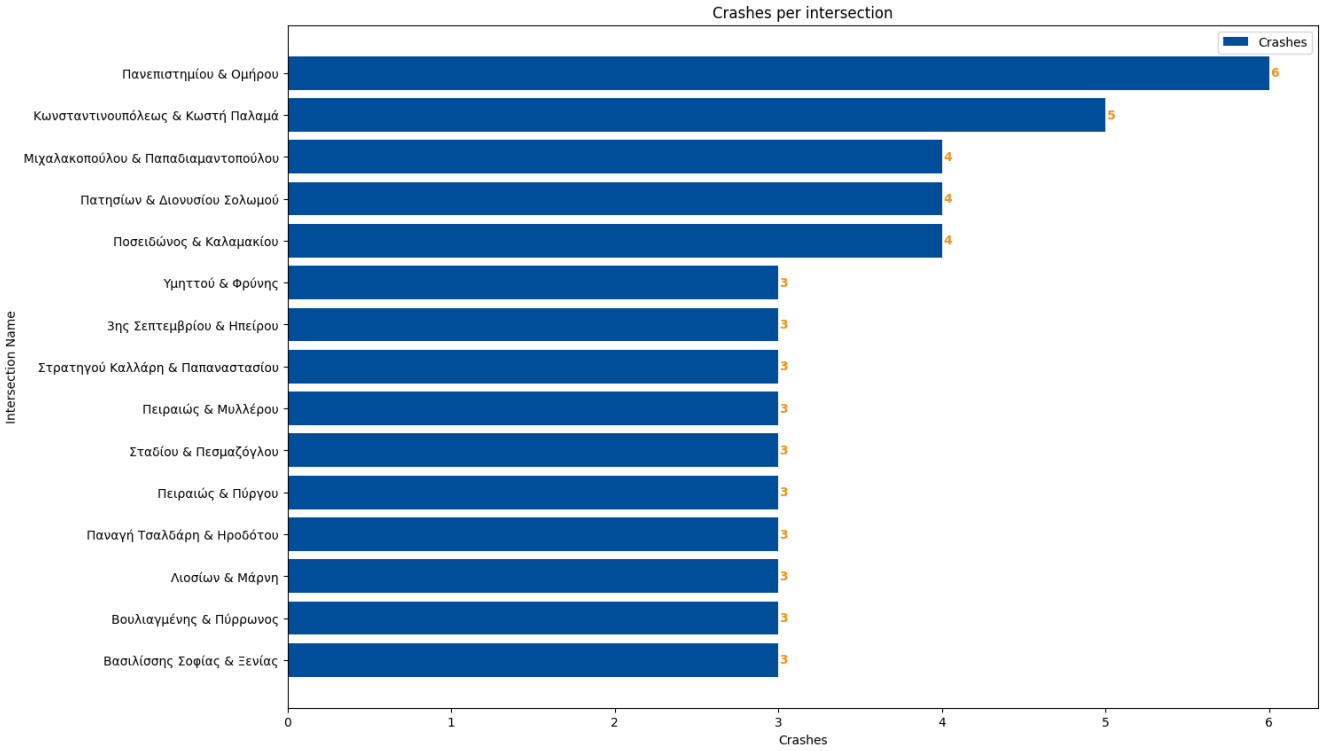

Figure 3. Top 15 Intersections based on crash count.

Figure 3 Illustrates the distribution of crash counts among the most crash-prone intersections in central Athens during 2019, set against the broader descriptive statistics for all 1,837 intersection-year observations. Although the vast majority of intersections experienced just one crash, a small subset exhibits markedly higher counts. Given that the mean crash count barely exceeds one event per intersection, these outliers, recording three to six incidents, stand out as clear “hotspots. The relatively low overall variability ($\sigma = 0.42$) underscores these high-count sites' exceptionalness within the network. Their disproportionate contribution to aggregate crash totals suggests persistent safety deficiencies that warrant targeted investigation, whether owing to traffic volumes, geometric complexity, signal timing, or visibility constraints. This study aims to examine the extent to which analysis of telematics data can predict which intersections are more prone to exhibiting a high frequency of crashes, since driving metrics such as harsh braking and harsh acceleration can be representative of traffic conditions on each intersection.

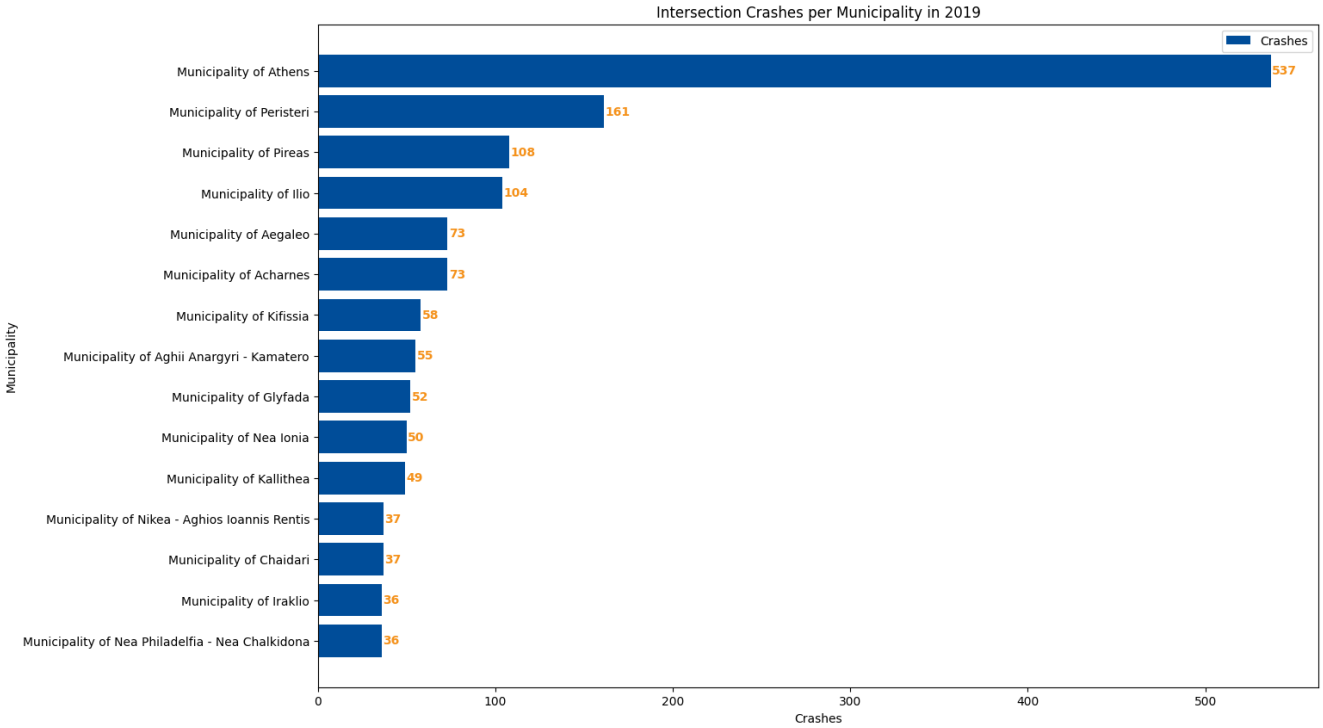

Figure 4. Intersection Crashes per Municipality in 2019.

Figure 4 depicts the spatial distribution of intersection-level crash counts aggregated by municipality for 2019. The Municipality of Athens overwhelmingly dominates the crash totals, with 537 recorded collisions, more than half of all events in the study area, reflecting its high density of intersections, traffic volumes, and complex road geometry. This pronounced gradient, from the central municipality to the outer suburbs, underscores the influence of urban form, vehicular demand, and land use intensity on collision risk. However, variations in the number of signalised intersections and traffic exposure across municipalities must be considered when interpreting these aggregate counts.

2.3 Methodology

This study combines telematics-based driving behaviour indicators with intersection-level crash occurrence data to forecast road safety risks in urban settings. The methodology framework comprises five principal phases: data preparation, spatial joining, feature aggregation, crash categorisation, and supervised classification.

Data Preprocessing and Spatial Alignment

Telematics travel data and crash reports were initially standardized inside a unified geographical reference system (EPSG:2100). The crash database included textual descriptions of intersecting street names for each occurrence, which were geocoded utilising OpenStreetMap and the Overpass API to get accurate intersection coordinates. The intersections were buffered with a 35-meter radius to encompass adjacent telematics occurrences spatially. The telematics data, gathered from a smartphone-based sensing device, were transformed into geographical points and reprojected to align with the crash data.

Spatial Joining and Feature Engineering

A spatial join technique was utilized to link each telematics point with its respective intersection buffer.

This facilitated the compilation of driving behavior characteristics at each intersection. The subsequent metrics were calculated for each intersection:

- Total Trips: Unique trip IDs intersecting the buffer zone.
- Harsh Braking Events
- Rapid Acceleration Events
- Speeding Flags

Normalised ratios—hard braking per trip, severe acceleration per trip, and speeding incidents per trip—were calculated to account for variations in traffic exposure.

Crash Risk Classification

Intersections were categorised into binary classes to simplify the modelling task:

- **Safe (Class 0)**: Intersections with one or zero crashes in 2019.
- **Unsafe (Class 1)**: Intersections with more than one crash.

This dichotomous labelling mitigates the effects of skewness in crash counts and aligns with the study's aim of hotspot prediction.

Model Training and Evaluation

An XGBoost classifier was chosen for its ability to manage non-linear relationships and its proven efficacy in imbalanced classification scenarios. The feature set included three standardized telematics metrics. To address class imbalance, with dangerous intersections as the minority class, the Synthetic Minority Oversampling Technique (SMOTE) was utilized during model training.

The dataset was split into training (80%) and testing (20%) subsets, maintaining class distribution using stratified sampling. The model's performance was assessed by accuracy, precision, recall, and F1-score. Emphasis was placed on recall for the hazardous category, due to its practical significance in risk mitigation techniques.

3. Results

The results of the modeling process provide insights into the relationship between telematics-based driving behaviors and crash risk at urban intersections in Athens. After filtering for sufficient traffic exposure (minimum 10 telematics-equipped trips per intersection), a total of 439 intersections were retained for analysis. Of these, 76 were classified as **unsafe** (more than one recorded crash in 2019), while the remaining 363 were labeled **safe**.

3.1 Model Performance

The XGBoost classifier, trained on synthetic-balanced data via SMOTE and tested on a holdout set (20% of the dataset), achieved an **overall accuracy of 74%**. The detailed performance metrics are shown in Table 1.

Table 1. XGBoost Classifier Performance Metrics

Metric	Safe Class	Unsafe Class
Precision	0.83	0.39
Recall	0.84	0.37
F1-Score	0.83	0.38

The classifier demonstrated strong performance for identifying **safe** intersections, with both precision and recall exceeding 0.80. However, performance on the **unsafe** class was more limited, with an F1-score of 0.38. This reflects the model's challenge in capturing minority-class patterns, a common issue in imbalanced urban crash datasets. Despite these limitations, the **weighted average F1-score of 0.74** indicates reliable overall predictive ability, especially given the sparsity of crashes at most intersections.

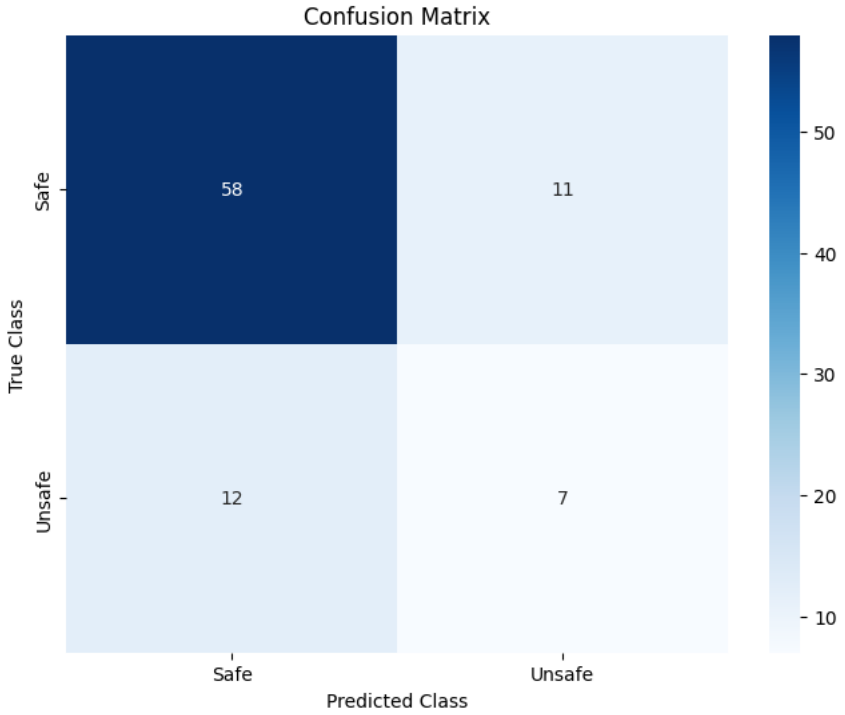

Figure 5: Confusion Matrix.

3.2 Feature Importance

The model identified Speeding Ratio, Harsh Acceleration Ratio, and Harsh Braking Ratio as the most influential predictors of intersection-level crash risk. Speeding Ratio emerged as the dominant factor, indicating that intersections with a higher prevalence of speeding behaviour are significantly more susceptible to crashes. Both harsh acceleration and harsh braking also showed meaningful contributions, likely reflecting moments of driver instability or abrupt decision-making, conditions often symptomatic of elevated traffic stress or poor intersection design.

3.3 Spatial Patterns

Visual analysis of the mapped intersection buffers revealed notable spatial clustering of high-risk zones in central Athens. Intersections with elevated ratios of harsh braking were predominantly located in the city's dense core, where constrained geometry, mixed land use, and high vehicular volumes coincide.
Dynamic mapping of telematics indicators (Figure 7) highlighted zones where traditional crash counts alone may underrepresent underlying risk. In several cases, the model flagged intersections with few crashes but high telematics event ratios as potentially hazardous, underscoring the value of integrating behavioural data into proactive safety analysis.

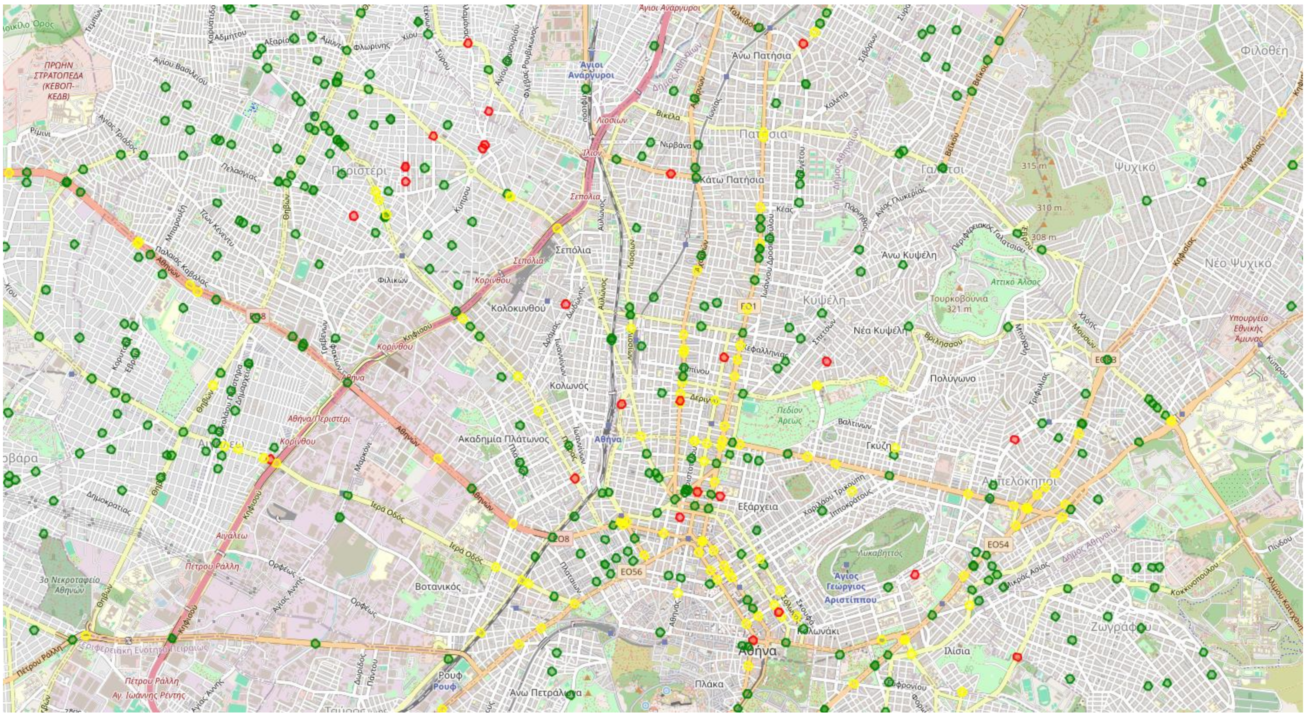

Figure 6: Spatial Distribution of harsh braking.

4. Conclusion

This study demonstrates the potential of integrating telematics-based behavioural data with machine learning techniques to identify crash-prone intersections in dense urban environments proactively. The analysis moves beyond traditional reactive safety models that rely solely on historical crash counts by leveraging a novel dataset combining high-resolution GPS driving records with official police-reported crashes.

The results show that specific telematics indicators—most notably the speeding ratio—are strongly associated with elevated crash risk. This reinforces the hypothesis that behavioural data captured from in-motion vehicles can serve as dynamic proxies for latent intersection hazards such as poor visibility, confusing signalization, or abrupt geometry changes.

While the XGBoost classifier achieved promising accuracy (74%), the recall for high-risk intersections (37%) reveals challenges inherent in modelling rare events. The limited number of crash-prone sites introduces a class imbalance that complicates prediction. Nevertheless, even partial identification of unsafe intersections can support traffic engineers in prioritizing safety audits and interventions.

The findings also suggest that urban traffic risk is not uniformly distributed but concentrated at intersections with aggressive driving signatures. The model highlights latent risks that may precede actual collisions by mapping intersections with elevated harsh braking or acceleration ratios, even in the absence of crashes. This forward-looking perspective aligns with the broader shift in road safety planning from reactive to preventive strategies.

This study offers both methodological and practical contributions to the field of traffic safety research. Methodologically, it presents a scalable framework that integrates telematics-derived behavioral data with crash records using geospatial analytics and supervised machine learning. This approach enhances

the precision and depth of intersection-level risk assessment. Practically, the framework provides urban policymakers with a proactive toolset for identifying high-risk intersections before crash frequencies escalate, supporting more timely, targeted, and cost-effective safety interventions.

Despite its contributions, the study faces several limitations. The telematics dataset may be biased toward specific user demographics or driving styles, and some intersections lacked sufficient trip density for robust modelling. Future research should expand to multi-year datasets, incorporate vehicular volumes or signal phase data, and explore ensemble modelling for greater robustness.
In conclusion, this research affirms that telematics-derived driving behaviour, when paired with machine learning, offers a powerful lens through which urban traffic safety can be understood and proactively managed.

5. References-Bibliography

Gopalakrishnan, S. (2012). A Public Health Perspective of Road Traffic Accidents. *Journal of Family Medicine and Primary Care*, *1*(2), 144. https://doi.org/10.4103/2249-4863.104987

Greibe, P. (2003). Accident prediction models for urban roads. In *Accident Analysis and Prevention* (Vol. 35). http://www.dtf.dk

Obaidat, M. T., & Ramadan, T. M. (2012). Traffic Accidents at Hazardous Locations of Urban Roads. In *Jordan Journal of Civil Engineering* (Vol. 6, Issue 4).

Vorko-Jović, A., Kern, J., & Biloglav, Z. (2006). Risk factors in urban road traffic accidents. *Journal of Safety Research*, *37*(1), 93–98. https://doi.org/10.1016/j.jsr.2005.08.009