# Combining diverse data sources for intersection crash analyses based on incomplete records

**Simone Paradiso**
PhD Candidate
Department of Transportation Planning and Engineering
National Technical University of Athens, 5 Heroon Polytechniou Str., GR-15773 Athens, Greece
Email: simone_paradiso@mail.ntua.gr

**Dimitrios Nikolaou**
Senior Researcher
Department of Transportation Planning and Engineering
National Technical University of Athens, 5 Heroon Polytechniou Str., GR-15773 Athens, Greece
Email: dnikolaou@mail.ntua.gr

**Apostolos Ziakopoulos**
Senior Researcher
Department of Transportation Planning and Engineering
National Technical University of Athens, 5 Heroon Polytechniou Str., GR-15773 Athens, Greece
Email: apziak@central.ntua.gr

**Alexis Aivaliotis**
R&D Project Manager
OSeven Telematics, 27B Chaimanta Str., 15234 Chalandri, Greece
Email: aaivaliotis@oseven.io

**George Yannis**
Professor
Department of Transportation Planning and Engineering
National Technical University of Athens, 5 Heroon Polytechniou Str., GR-15773 Athens, Greece
Email: geyannis@central.ntua.gr

Word Count: 4736 words (excl. reference list)

*Submitted [15/12/2024]*

**ABSTRACT**
Road safety research has raised concerns about the quality and reliability of police reports, which often lack pivotal features such as crash coordinates, even in developed countries. To explore possible solutions, this study aims to link geospatial and telematics data to the intersections of a specific study area in Athens and identify and locate intersection crashes registered in the police crashes database. Using the geocodes from the Hellenic Statistical Authority (ELSTAT), which features police datasets, the street names forming the intersection in the police report were identified. Subsequently, the intersections' coordinates have been retrieved from OpenStreetMap (OSM) for the study area by using the street names. Telematics data for the study area were also obtained. To resolve discrepancies between the street names obtained through geocoding and those from OSM, a Natural Language Processing (NLP) technique was applied. Specifically, a string similarity task which involves encoding strings into contextual embeddings and calculating their similarity. Nonetheless, some data losses occurred while attempting to retrieve the names, primarily due to the limited coverage of the telematics dataset which defined the study area excluding some of the crashes reported by the police. By creating a buffer around each node within the study area, each node has been characterized by the road centroids' attributes falling within the buffer, from a multi-source dataset containing geographic, transport network and telematics attributes per road. Therefore, the intersections recorded by the police were successfully flagged. The created datasets are analyzed with spatial models and relevant insights are obtained for intersection crash occurrences. Speeding and road angle were identified as significant factors highly correlated with the intersection crash occurring as well as a crash index.

**Keywords:** Road Safety, Telematics, NLP, Intersections.

## INTRODUCTION

Road crashes and their consequences continue to be a serious global societal concern, supported by estimates provided by the World Health organization (WHO) which state that 1.19 million road users lost their lives in 2021 due to road crashes, ranking road crashes as the 12th leading cause of death globally (World Health Organization, 2023). Regarding the respective figures in the European Union, 20,640 road users lost their lives due to road crashes in 2022. In Greece, where the analyses of the present study were conducted, there were 60 road fatalities per million population recorded in 2023, ranking the country in the 22nd position among the 27 countries of the European Union (European Commission, 2023). Road crashes are a complex phenomenon affected by several parameters that can be categorized into three distinct aspects: (i) road users (drivers, riders, passengers and pedestrians); (ii) vehicles; and (iii) road infrastructure and environment. It is possible to represent the latter by means of a graph. In the graph, the nodes will be the intersections inside the infrastructure. Intersection crashes are one of the most common types of crash problems, particularly in urban areas. In rural areas, or where vehicle speeds are high, the consequence of collisions at intersections can be particularly severe (iRAP, International Road Assessment Programme).

This study aims to investigate the factors contributing to crashes' occurrence at intersections. The work contributes to the existent literature by analyzing various econometrics and machine learning models. Machine Learning algorithms have been widely used in crashes analysis to overcome the limitations of traditional statistical models (Ghandour, Hammoud, & Al-Hajj, 2020), or to allow transferability of the models among different areas (Ziakopoulos, Vlahogianni, Antoniou, & Yannis, 2022). A general overview on the most used models is provided in (Ziakopoulos & Yannis, 2020) and more specific advances and challenges are highlighted in (Ali, Hussain, & Haque, 2024) for crash occurrence, crash frequency and injury severity. The study also addresses the challenge of integrating information from various databases with linking issues by means of an advanced Natural Language Processing technique. These techniques have proven highly significant when it comes to applications such as semantic text understanding and knowledge fusion (Yin, et al., 2019), which this study aims to tackle.

The structure of the paper is as follows: in METHODS the processing of the data is shown along with the chosen models. In RESULTS the models are analyzed, and performances are compared. In DISCUSSION the findings are interpreted, and an overview of the results is provided. Finally, the study is summarized, and some takeaways are highlighted in the CONCLUSIONS.

## METHODS

Road safety researchers have highlighted how the quality and reliability of police reports is not at the top when it comes to analyzing them. In this study, the police crashes report for the year 2021 in the Regional Unit of Central Athens has been analyzed.

The report shows some crashes' attributes such as the time and the day when they occurred, codes for the zone and the street (or the streets, in case of intersections) where they occurred. By using these geocodes in the Hellenic Statistical Authority (ELSTAT) database, the street names forming the intersections in the police report were identified. The key used in both databases to merge the information is built by merging the district code and the street code.

In the police report, only the crashes occurred at the intersections having been selected, by simply filtering out the rows which did not show the street code for the second street or they displayed the code 9999 (unknown).

Some crashes happened at the same intersections, since the goal of the study was to infer on the crash occurrence, the duplicated events will be dropped out from the analysis, leading to an equivalent of 704 intersection crashes occurred in the Regional Unit of Central Athens in 2021.

The aim of the present reserach is to map attributes from a multi-source dataset containing geographic, transport network and telematics attributes per road, onto the intersections within a study area. The study area has been defined considering the coverage of telematics data. Telematics data comes from an innovative smartphone application developed by OSeven (OSeven), aiming to record driver behavior using the hardware sensors of the smartphone device. Furthermore, a variety of APIs are

exploited to read sensor data and temporarily store them to the smartphone's database before transmitting them to the central (back-end) database. The data collected are highly disaggregated in space and time. Once stored in the backend cloud server, they are converted into meaningful driving behavior and safety indicators, using signal processing, Machine Learning (ML) algorithms, Data fusion and Big Data algorithms. This is achieved by using state-of-the-art technologies and procedures, which operate in compliance with standing Greek and European personal data protection legislation (GDPR).

The study also made use of data from OpenStreetMap (OSM), which is a free, editable map of the whole world that is being built by volunteers largely from scratch and released with an open-content license (OpenStreetMap). A Python library termed OSMnx has been used, which is a tool that easily downloads and analyzes street networks for anywhere in the world. OSMnx contributes five primary capabilities for researchers and practitioners. First, it enables automated and on-demand downloading of political boundary geometries, building footprints, and elevations. Second, it can automate and customize the downloading of street networks from OpenStreetMap and construct them into multidigraphs. Third, it can correct and simplify network topology. Fourth, it can save/load street networks to/from disk in various file formats. Fifth and finally, OSMnx has built-in functions to analyze street networks, calculate routes, project and visualize networks, and quickly and consistently calculate various metric and topological measures. These measures include those common in urban design and transportation studies, as well as advanced measures of the structure and topology of the network (Boeing, 2017).

A bounding box was defined using minimum and maximum latitude and longitude values from the telematics dataset. Within this bounding box, the road network was extracted using the Overpass API, specifying the pre-defined network type 'drive'. The graph has been converted into node and edge GeoDataFrames for futher analysis.

**Leveraging Natural Language Processing (NLP) for street name matching between crash data and OpenStreetMap (OSM) data**

It has been observed a discrepancy between the street names obtained through geocoding and those from OSM. This discrepancy does not allow the information from the two databases to be directly linked.

In order to resolve it, a language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers (Kenton & Toutanova, 2019) has been used. BERT is designed to pre-trained deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. Specifically, since the area includes Greek names, a Greek pre-trained BERT has been chosen for this study case (Koutsikakis, Chalkidis, Malakasiotis, & Androutsopoulos, 2020). The model shows a Micro F1 = 85.7 ± 1.00 for Named Entity Recognition task with Greek NER dataset, Micro F1 is a metric used to evaluate performance in multi-class tasks by aggregating true positives, false positives, and false negatives across all classes before computing precision and recall.

The names from the crash reports and OpenStreetMap (OSM) data were standardized using a combination of a custom normalization function and the one recommended for the Greek BERT from the Hugging Face community. A numerical high dimensional vector representation (embedding) of the given street names has been obtained through BERT, this representation has been used to calculate the cosine similarity, which is a measure of the similarity between two vectors. It calculates the cosine of the angle between them in a multi-dimensional space. Finally, for each street name in the crashes report, the street name with the highest similarity score in the OSM dataset was retrieved.

Once the names between the two datasets have been aligned, for each intersection in the crashes report the coordinates and the OSM IDs (unique identifier for the nodes in OSM) were retrieved for each intersection in the crashes report. The process involved checking on the common nodes shared by the streets forming each intersection in the edge GeoDataFrame and linking the retrieved nodes to the coordinates from the node GeoDataFrame.

*Telematics coverage*

Some data losses occurred while attempting to retrieve the names, primarily due to the limited coverage of the telematics dataset which defines the study area excluding some of the crashes reported by the police. By using (Nominatim), a geocoding software that powers the official OSM site, geographical coordinates for a representative point for most streets in the crashes report have been fetched, some of them located far outside the study area, suggesting potential errors. Despite these inaccuracies the points were plotted to validate the occurred loss of data, showing a significant number of points outside the telematics coverage.

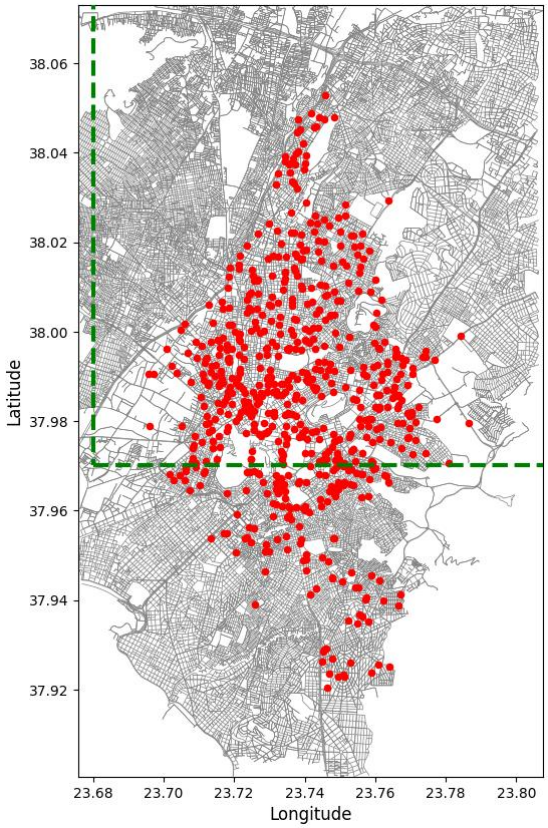


**Figure 1: Coordinates from the crash reports after processing and the bounding box from the telematics coverage area.**

**Figure 1** shows the telematics coverage (green dashed box) area excludes some streets contained in the crashes report.

**Obtaining the final intersection dataset with flagged crashes**

At this stage, it has been decided to zoom in the map, along with the nodes dataset and the multi-source dataset described in (Nikolaou, Ziakopoulos, Kontaxi, Theofilatos, & Yannis, 2025) by defining the bounding box based on the minimum and maximum latitude and longitude values from the

intersection crashes dataset. The approach aims to minimize bias introduced by suburban areas where crashes are not observed. As illustrated below in **Figure 2**.

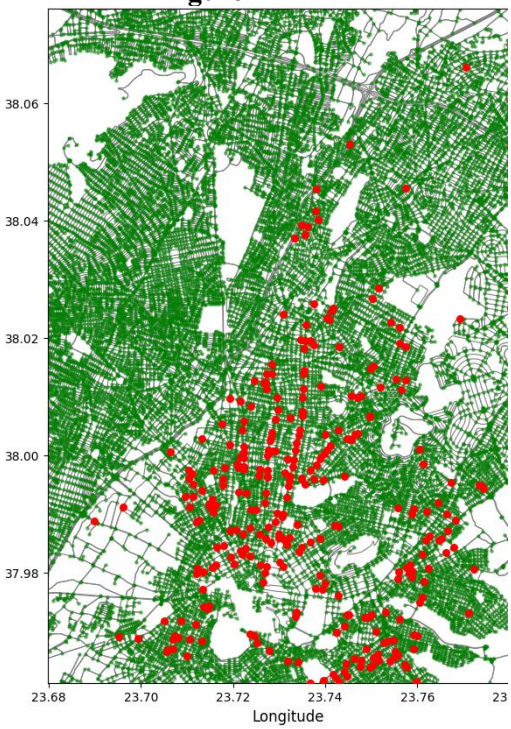


**Figure 2: Edges, nodes and intersection crashes within zoomed-in the study area.**

A buffer has been created around each node from the node GeoDataFrame within the defined study area. A 20-meter radius was chosen for the buffer, as it aligns with established practices in the literature, researchers in (Lee, Park, & Kim, 2016) drew a buffer with a 15 meters radius to define features such as the sum of perceived risk point at the intersection and the number of pedestrian-involved crashes at the intersection.

Each row of the multi-source dataset is referred to a road centroid with geographic, transport network and telematics attributes per road. Each node from the node GeoDataFrame has been characterized by the road centroids' attributes falling within the created buffer.

Road centroids might fall in multiple buffers and a buffer can include multiple centroids or none at all. An inner join has been performed to include only nodes characterized by at least one road centroid contained by its buffer.

**TABLE 1** shows all the columns of the joined dataset along with a brief explanation of each and describes how they have been handled during the aggregation at the node level.

**TABLE 1 Columns in the joined dataset**

| Column's name | Explanation | Action related to node aggregation |
|---|---|---|
| length | Road length | Dropped |
| highway_left | Road function and importance | Dropped |
| Centroidlon | Centroid longitude | Dropped |
| Centroidlat | Centroid latitude | Dropped |
| geometry_left | Centroid geometry column | Dropped |
| highway_right | Node function and importance | Dropped |

| ref | Reference number or code | Dropped |
|---|---|---|
| Crashes2016-2020 | Municipal area index calculated using Inverse Distance Weighting (IDW) (Lu & Wong, 2008) for crashes between 2016 and 2020 | Sum |
| Fat2016-2020 | Municipal area index calculated using IDW for fatalities between 2016 and 2020 | Sum |
| KSI2016-2020 | Municipal area index calculated using IDW for 'Fatalities + Serious injuries' between 2016 and 2020 | Sum |
| speeding_count | Number of speeding flags per road centroid | Sum |
| mobile_usage_count | Number of mobile usage flags per road centroid | Sum |
| harsh_acc_count | Number of harsh acceleration flags per road centroid | Sum |
| harsh_braking_count | Number of harsh braking flags per road centroid | Sum |
| trip_count | Number of times the road had been chosen | Sum |
| tot_angle | The sum of angles formed by consecutive segments of a road | Mean |
| efficiency | The ratio of the direct (Euclidean) distance to the actual length of the road. | Mean |
| angle_rate | The ratio of tot_angle to the direct (Euclidean) distance between the start and end points of the road. | Mean |
| slope | Slope or gradient of the roads. | Mean |
| PC_D_Seatbelt_Yes% | Percentage Drivers with Seatbelt | Mean |
| PTW_D_Helmet_Yes% | Percentage Drivers with Helmet | Mean |
| PC_D_Phone_No% | Percentage Drivers without Distraction | Mean |
| PC_Speeding_No% | Percentage Cars without Speeding | Mean |
| speeding_rate | speeding_count normalized on trip_count | Mean |

| | | |
|---|---|---|
| mobile_usage_rate | mobile_usage_rate normalized on trip_count | Mean |
| harsh_acc_rate | harsh_acc_rate normalized on trip_count | Mean |
| harsh_braking_rate | harsh_braking_rate normalized on trip_count | Mean |
| index_right | Node OSM ID | Pick the first (just one) (Dropped during the modelling phase) |
| x | Node longitude | Pick the first (just one) (Dropped during the modelling phase) |
| y | Node latitude | Pick the first (just one) (Dropped during the modelling phase) |
| geometry_right | Buffer geometry column | Pick the first (just one) (Dropped during the modelling phase) |
| street_count | Number of streets connected to the node | Pick the first (just one) |

After the aggregation per node, a dataset containing the nodes within the chosen area and their attributes mapped from the multi-source dataset has been built.

By iterating over this dataset and checking on whenever the node OSM ID appeared in the intersection crashes dataset, the intersections recorded by the police were successfully flagged. This process leads to 100 nodes being flagged with a crash occurrence out of 6716. Resulting in a heavily imbalanced dataset, as already found in this research field (Cai, Abdel-Aty, M., Lee, & Wu, 2020).

The observed reduction in the number of crashes in the final dataset can be attributed to the data preparation process. Specifically, intersections were omitted if no road centroids fell within the defined buffer zone, avoiding the association of telematics data with these nodes. Consequently, the analysis is limited to a subset of nodes where telematics data could be reliably linked by using the buffer. Expanding the buffer size would increase the number of nodes characterized with telematics data, potentially capturing more crash-related nodes. However, this adjustment may not necessarily improve the proportion of crash to non-crash nodes in the dataset, since non-crash nodes would be added as well.

The created dataset is analyzed with spatial models and relevant insights are obtained for intersection crash occurrences.

**Spatial models**

The initial analysis of the dataset, through the description of the features, suggests the need for further cleaning. The high kurtosis observed in some predictors indicates heavy tails which make complicated to work with the usual outliers' detection formulas. Hence DBSCAN has been employed to further clean the dataset from outliers, providing a robust way to do it. DBSCAN (Khan, Rehman, Aziz, Fong, & Sarasvady, 2014) is a clustering algorithm working with a density-based approach; therefore, it identifies outliers based on their isolation. The algorithm relies on two parameters: the *eps* chosen by plotting the k-distance graph and the *min_samples* chosen by the rule of thumb *min_samples = dimensions + 1*. After selecting the parameters, the clustering was applied and the elements labelled with the cluster *-1* were identified as outliers, following the convention used by the algorithm.
The final dataset consists of 100 crash examples out of 6686 intersections.

A correlation matrix at this stage was generated. Specific columns referring to crashes (*Fat2016-2020 & KSI2016-2020)* were removed due to their high correlation with the *Crashes2016-2020* column. Similarly, from the columns referring to angle, only one was retained due to multicollinearity.

The variance inflation factor (VIF) for each predictor has been evaluated, leading to remove some of them which were characterized by a high VIF value and maitining the ones considered useful for the analysis. The remained columns in the dataset are the following ones:

**TABLE 2 Columns in the final dataset**

| Columns | Explanation |
|---|---|
| Efficiency | Average efficiency, straightness of a road, for centroids within the node buffer. |
| Slope | Average slope for centroids within the node buffer. |
| Crashes2016-2020 | Sum of municipal crashes (2016-2020) indexes for all road centroids within the node buffer. |
| Speeding_count | Sum of speedings per road centroids within the node buffer. |
| Mobile_usage_count | Sum of mobile usage per road centroids within the node buffer. |
| Harsh_acc_count | Sum of harsh accelerations per road centroids within the node buffer. |
| Harsh_braking_count | Sum of harsh brakings per road centroids within the node buffer. |
| Trip_count | Sum of trips per road centroids within the node buffer. |
| Speeding_rate | Average speeding rate for centroids within the node buffer. |
| Mobile_usage_rate | Average mobile usage rate for centroids within the node buffer. |
| Harsh_acc_rate | Average harsh accelaration rate for centroids within the node buffer. |
| Harsh_braking_rate | Average harsh braking rate for centroids within the node buffer. |
| Street_count | Number of streets connected at the node |
| Crash Occurrence (1/0) | Target column |

Four models have been trained comparatively: A Penalized L1-Logistic Regression, k-Nearest Neighbors (k-NN), Random Forest, and Support Vector Classifier (SVC).

For all the models a Stratified K-Fold validation has been used to train them, maintaining the same proportion of the classes in every split. The SMOTE (Synthetic Minority Over-sampling Technique) and the RandomUnderSampler have also been applied to each model to address the issue of class imbalance in the dataset. They are two popular methods used for handling imbalanced datasets, especially in classification tasks where one class is underrepresented (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), (Wongvorachan, He, & Bulut, 2023).

In the case of SVC, k-NN and penalized Logistic Regression the data has been scaled to not mislead the model by the different magnitude of the variables.

## RESULTS

The models are presented below along with their feature importance, followed by a comparison on the performances.

*Logistic Regression*

Logistic Regression is a parametric statistical learning model for binary classification tasks.

LASSO regularization (L1) (Muthukrishnan & Rohini, 2016) was chosen because it encourages sparsity in the model by shrinking less important feature coefficients to zero. This helps in identifying the most significant variables, making the model interpretable and performing feature selection.

Analysing the odds ratios in **Figure 3** it has been seen how variables such as the number of streets connected to the intersection and phone usage around the intersection are associated with a higher likelihood of a crash occurring at the node, as well as the efficiency—which measures the average absence of curvature in the roads connected to the intersection. Whereas variables such as the amount of speeding near the intersection and the harsh acceleration are linked to a reduced likelihood of crashes.

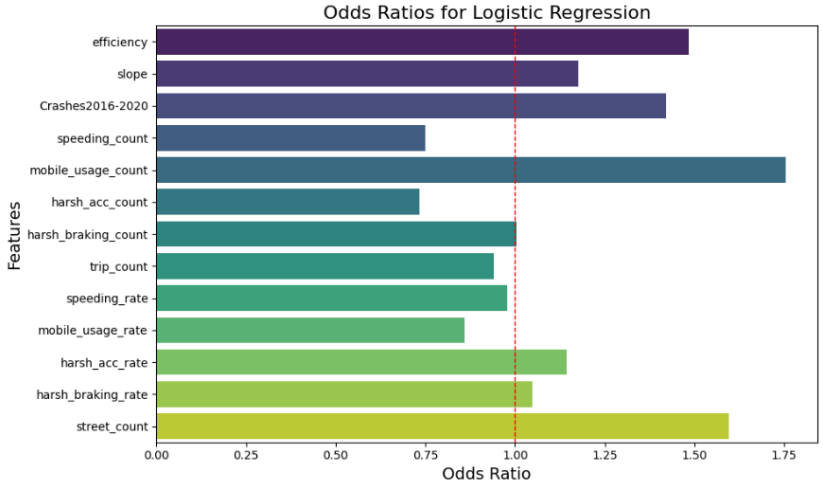


**Figure 3: Odds Ratios for L1 Logistic Regression**

The ROC curve (**Figure 4**) has been plotted, and the Area Under the Curve (AUC) is widely considered as the most comprehensive performance measure. This is because it evaluates the model's performance across all possible thresholds, making it threshold-independent.

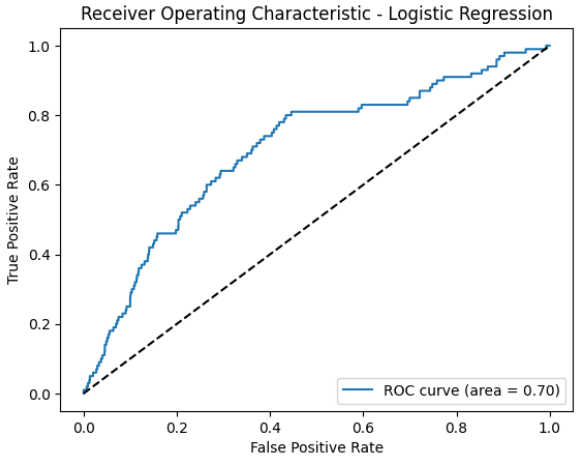


**Figure 4: ROC-AUC Logistic Regression**

*Random Forest*

A Random Forest (Biau & Scornet, 2016) is a popular ensemble machine learning algorithm which builds multiple decision trees during training and aggregates their outputs for improved accuracy, robustness, and generalization.

The Random Forest for our task has been tuned and SHAP values (Salih, et al., 2024) on the whole initial dataset have been calculated to interpret the model, they have recently become involved in road safety analysis (Ziakopoulos A. , 2024) and shown in **Figure 5**. Shapley values are a widely used approach from cooperative game theory that come with desirable properties, they are used to explain machine learning models. The results are coherent from what it has been seen through the odds ratios in the L1 Logistic Regression: variables such as the number of streets connected to the intersection, the index calculated using IDW for crashes between 2016 and 2020, the phone usage around the intersection and the efficiency are associated with a higher likelihood of a crash occurring at the node.

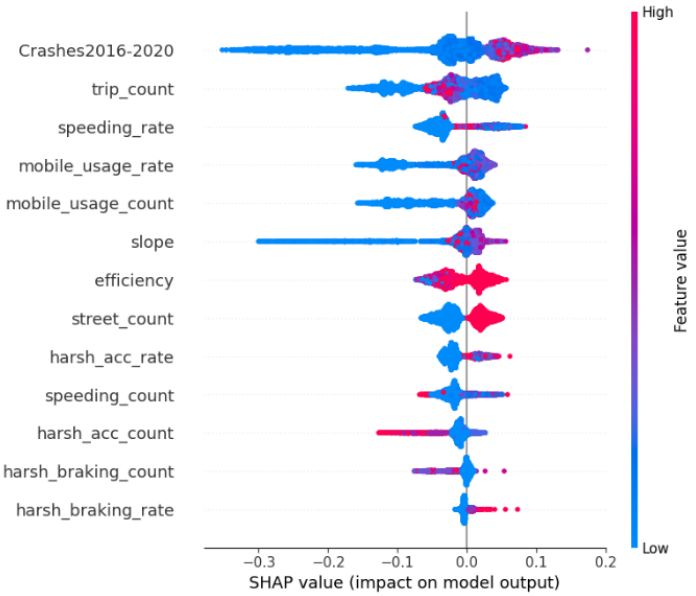


**Figure 5: SHAP values Random Forest**

The ROC curve for model evaluation is presented in **Figure 6**.

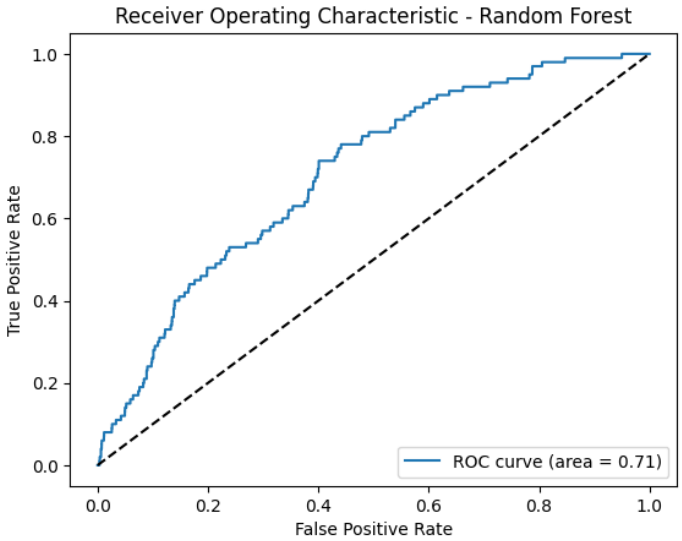


**Figure 6: ROC-AUC Random Forest**

*k-Nearest Neighbors (k-NN)*

The k-NN model is a simple, non-parametric supervised learning algorithm used for both classification and regression tasks. For this task the Number of Neighbors has been set to 8.

The ROC curve for model evaluation is presented in **Figure 7**.

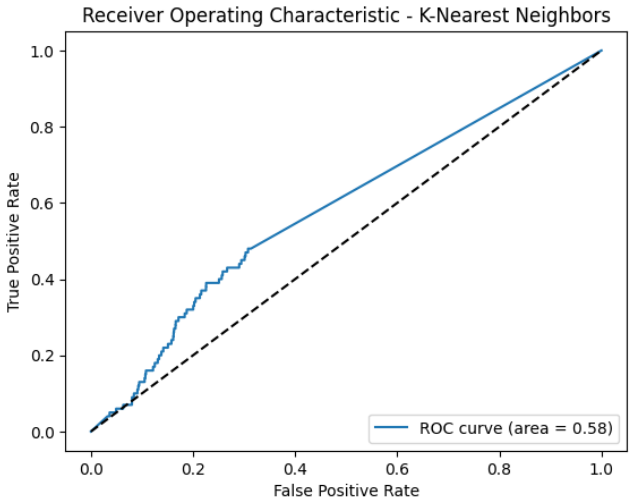


**Figure 7: ROC-AUC k-NN**

*Support Vector Classifier*

The Support Vector Classifier (SVC) is a supervised machine learning algorithm based on Support Vector Machines (SVM) (Kecman, 2005), particularly effective in high-dimensional spaces.

The SVC has been tuned with a Gaussian kernel, and with a regularization by setting the parameter C=0.1 (the default value is 1), which is the parameter that controls the trade-off between maximizing the margin and minimizing classification errors on the training data. The probability for each element to belong to one class has been obtained as well.

SHAP values on a subset of the initial dataset have been calculated to interpret the model and they are shown in **Figure 8**. A background dataset was used to reduce the computational cost, while estimating the expected contribution of each feature. Additionally, only a defined amount of first *n* rows of the data have been analyzed to further limit computation time. The feature Crashes2016-2020 emerges as the most important variable although with different results from the Random Forest likely due to the sampling to explain the features. High values of slope feature are associated with higher probability of crashes, which aligns with the Random Forest and Logistic Regression models, the same applies to the mobile usage count feature.

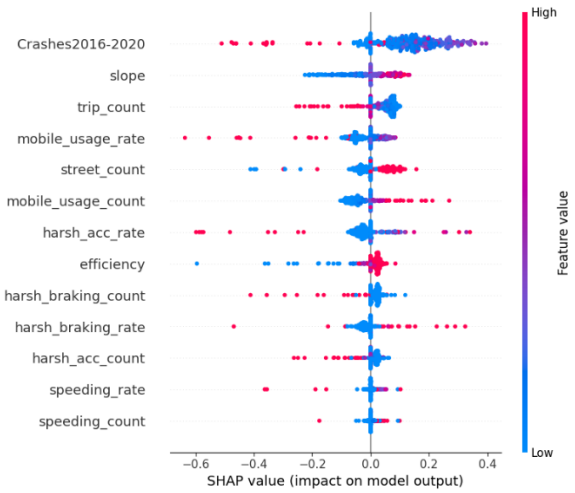


**Figure 8: SHAP values SVC**

In **Figure 9** the ROC curve is displayed.

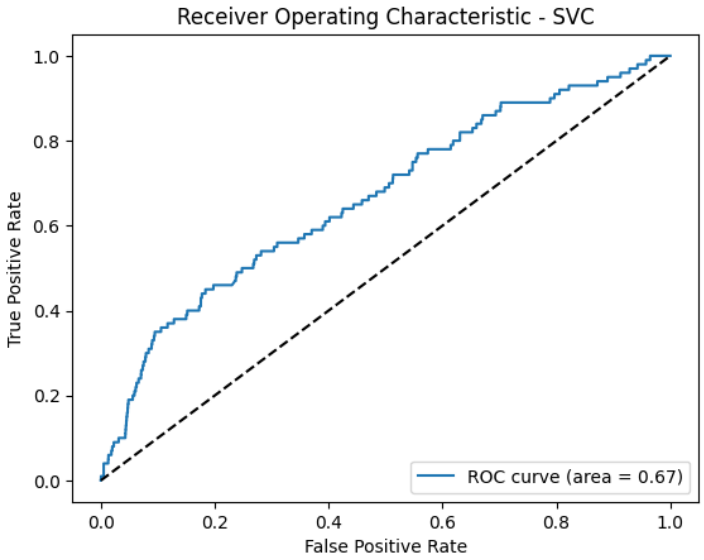


**Figure 9: ROC-AUC SVC**

**Figure 10** shows all the models show a very low precision to predict a crash which is a rare event, the nature of the dataset being highly imbalanced makes diffucult for the model to capture the nature of the crashes. Random Forest and k-NN show very low values in recall for the crashes as well, while L1-Logistic Regression model and SVC give modest values of recall. At the same time, the latter show lower recall for the non-crash events, but they seem to provide a more balanced performance compared to Random Forest and k-NN.

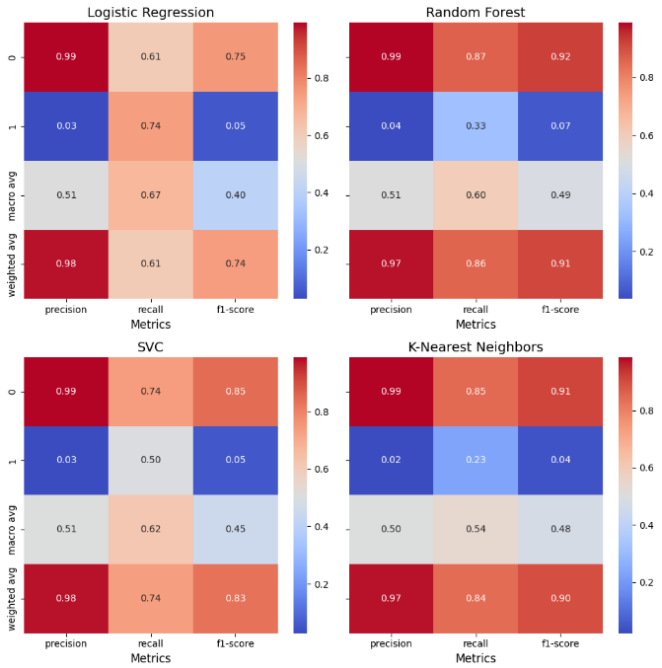


**Figure 10: Classification report per each model**

Taking a look at the macro average and weighted average could provide a better perspective on the models.

## DISCUSSION

The objective of this study was to analyze intersection crashes combining diverse data sources with incomplete records. When dealing with different datasets or multi-source datasets, issues related to how to link the information across them can arise. This study provides an approach by means of Natural Language Processing to overcome the lack of alphanumeric keys normally used to merge datasets and work on names of the streets which can have different formats among different sources.
After having processed the data and cleaned them from outliers, four models have been employed to analyze the intersection crashes. The dataset being highly imbalanced, leads the models to favor the majority class to improve the overall metrics and ignore the minority class (crashes).

For the minority class the models produce many false positives, however considering the topic of crashes, this might be still accurate in a conservative way. The threshold determines the cutoff probability for assigning positive or negative labels in a classification model. Lowering it leads toward higher recall but lower precision, an ideal approach, where failing to detect an event is costlier than many false positives. Increasing it leads toward lower recall but higher preocision, ideal when false positives are problematic.

Hence, for this study the AUC has been chosen to evaluate the model performance, since it measures the model's ability to classify instances across all possible threshold values. Looking at the AUC, the L1 Logistic Regression and the Random Forest slightly outperform the other models showing better performances in class separation, having an AUC of 0.70 and 0.71 respectively. Whereas k-NN is performing poorly in the classification, with an AUC of 0.58.

The Random Forest and the Logistic Regression have shown similar findings, suggesting that the increase of variables such as the number of the streets connected to an intersection, the index for crashes between 2016 and 2020 aggregated at the specific intersection, the phone usage around the node lead to higher probability of crashing at the intersection. High values of average efficiency, indicating the

straightness of a road, per node are associated with a higher probability of crashing, likely due to the greater vibility encouraging people to speed and reducing reaction time.

Dataset imbalance was addressed using resampling techniques, however surrogate safety measures (SSMs) might be investigated. The main idea behind them is that near-crash traffic events can be used as surrogates for real crashes (Johnsson, Laureshyn, & Dágostino, 2021) given the advantages of being more frequent than crashes. They are meant to be an alternative to crash data enabling a proactive approach to traffic safety, but they may also address the imbalance in crash datasets by supplementing crash data with more abundant near-crash data.

Additionally, the single node was characterized by road centroids from a multi-source dataset with aggregated telematics data, further work can be done with the raw data.

## CONCLUSIONS

Crashes documented in police reports often lack details on crash severity or type, highlighting the need for more investigations into road safety at intersections. Concerns have been raised about the quality and reliability of police reports, along with challenges in merging data from different sources. The present study focused on linking geospatial and telematics data to the intersections of a study area in Athens, aiming to identify and locate intersection crashes registered in the police crashes database. Using the geocodes from the Hellenic Statistical Authority (ELSTAT), which features police datasets, the street names forming the intersection in the police report were identified. Subsequently, the intersections' coordinates have been retrieved from OpenStreetMap (OSM) for the study area by using the street names. Telematics data for the study area were also obtained.

In the present work, NLP has proven valuable to work with textual data and use it to address these limitations. The limitation due to the imbalance of the dataset remains, especially when it comes to rare events such as crashes. This limitation was addressed by using resampling techniques but approaches based on surrogate safety measures might reduce the data imbalance and reducing the amount of synthetic data generated. In case of suboptimal improvements, further conceptualization work can be conducted towards understing the role of the thresold for the specific binary classification might enhance our ability to reach meaningful insights about road crashes which continue to be a serious global societal concern. Additionally, future studies could explore using the count of crashes rather than crash occurrences for a more detailed analysis.

## ACKNOWLEDGMENTS

This research has been conducted within the IVORY project. The project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590.

## AUTHOR CONTRIBUTIONS

**Simone Paradiso**: Conceptualization, Data curation, Methodology, Software, Formal Analysis, Writing-original draft, Writing-review & editing.
**Dimitrios Nikolaou**: Data curation, Methodology, Writing-review & editing.
**Apostolos Ziakopoulos**: Conceptualization, Methodology, Software, Supervision, Writing-original draft, Writing-review & editing.
**Alexis Aivaliotis**: Data collection, Writing-review & editing.
**George Yannis**: Conceptualization, Supervision.

**REERENCES**

Ali, Y., Hussain, F., & Haque, M. M. (2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention, 194*, 107378.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*, 197-227.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, environment and urban systems, 65*, 126-139.

Cai, Q., Abdel-Aty, M., Y. J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. *Transportation research part C: emerging technologies, 117*, 102697.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

European Commission. (2023). *Road safety in the EU: Fatalities below pre-pandemic.* European Commission. Retrieved from https://ec.europa.eu/commission/presscorner/detail/e%20n/ip_23_953

Ghandour, A. J., Hammoud, H., & Al-Hajj, S. (2020). Analyzing Factors Associated with Fatal Road Crashes: A Machine Learning Approach. *International Journal of Environmental Research and Public Health, 17*(11), 4111. doi:https://doi.org/10.3390/ijerph17114111

iRAP, International Road Assessment Programme. (n.d.). *Crash type: Intersections.* iRAP. Retrieved from https://toolkit.irap.org/crash-type/intersections/

Johnsson, C., Laureshyn, A., & Dágostino, C. (2021). A relative approach to the validation of surrogate measures of safety. *Accident Analysis & Prevention, 161*, 106350.

Kecman, V. (2005). Support vector machines–an introduction. *Support vector machines: theory and applications*, 1-47. Retrieved from https://www.ibm.com/topics/support-vector-machine

Kenton, J. D., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of naacL-HLT, 1*, 2.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *The fifth international conference on the applications of digital information and web technologies, ICADIWT 2014*, 232-238.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). Greek-bert: The greeks visiting sesame street. *11th Hellenic conference on artificial intelligence*, 110-117. Retrieved from https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1

Lee, G., Park, Y., & Kim, J. C. (2016). Association between intersection characteristics and perceived crash risk among school-aged children. *Accident Analysis & Prevention, 97*, 111-121.

Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & geosciences, 34*(9), 1044-1055.

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. *IEEE international conference on advances in computer applications, ICACA*, 18-20.

Nikolaou, D., Ziakopoulos, A., Kontaxi, A., Theofilatos, A., & Yannis, G. (2025). Spatial analysis of telematics-based surrogate safety measures. *Journal of Safety Research, 92*, 98-108.

(n.d.). *Nominatim.* Retrieved from https://nominatim.org/

OpenStreetMap. (n.d.). Retrieved from https://wiki.openstreetmap.org/wiki/About_OpenStreetMap

OSeven. (n.d.). Retrieved from www.oseven.io

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, 2400304.

Tanimoto, A., Yamada, S., Takenouchi, T., Sugiyama, M., & Kashima, H. (2022). Improving imbalanced classification using near-miss instances. *Expert Systems with Applications, 201*, 117130.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information, 14*(1), 54.

World Health Organization. (2023). *Global status report on road safety.* World Health Organization. Retrieved from https://www.who.int/publications/i/item/9789240086517

Yin, X., Huang, Y., Zhou, B., Li, A., Lan, L., & Jia, Y. (2019). Deep entity linking via eliminating semantic ambiguity with BERT. *IEEE Access, 7*, 169434-169445.

Ziakopoulos, A. (2024). Analysis of harsh braking and harsh acceleration occurrence via explainable imbalanced machine learning using high-resolution smartphone telematics and traffic data. *Accident Analysis & Prevention, 207*, 107743.

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

Ziakopoulos, A., Vlahogianni, E., Antoniou, C., & Yannis, G. (2022). Spatial predictions of harsh driving events using statistical and machine learning methods. *Safety science, 150*, 105722.