International Symposium
“Navigating the Future of Traffic Management”
June 29 - July 3 2025, Athens, Greece

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

Simone Paradiso[1], Dimitrios Nikolaou[1], Apostolos Ziakopoulos[1], Alexis Aivaliotis[2], George Yannis[1]

[1] National Technical University of Athens, Department of Transportation Planning and Engineering, Athens, Greece
[2] OSeven Telematics, Chalandri, Greece

Introduction

Intersection crashes are one of the most common types of crash problems.

This study aims to investigate the factors contributing to crashes’ occurrence at intersections.

This study investigates crash factors at intersections—critical points in road networks—by applying both econometric and machine learning models. It also tackles data integration challenges caused by linking issues using advanced **Natural Language Processing** (NLP) techniques, which have proven effective for semantic understanding and knowledge fusion.

Figure 1: *NLP enables computers to understand text*

Objectives

This research maps geographic, transport network, and telematics attributes from **multi-source data** onto intersections and applies machine learning to investigate **factors** contributing to **crashes** at those locations.

NLP for street name matching

Police crashes report for the year 2021 in the Regional Unit of Central Athens shows some crashes’ attributes, geocodes for the zone and the street (or the streets, in case of intersections) where crashes occurred. **Intersection crashes** were selected and the **street names** identified by using these geocodes in the Hellenic Statistical Authority (ELSTAT) database.

OpenStreetMap (OSM), a free and editable map of the whole world, was used to retrieve the intersection geographic coordinates.

A **discrepancy** between the street names obtained through **geocoding** and those from **OSM** hindered the information from the two databases to be directly linked.

A Transformer-based model, **BERT**, was employed to generate vector representations (**embeddings**) of street names and to compute the **cosine similarity** between each street in the crash dataset and those in the OSM dataset. The most similar OSM street name was matched to each crash dataset street.

Coordinates and the **OSM IDs** for each intersection in the crashes report were retrieved, by finding common nodes shared by the corresponding street pairs in the edge GeoDataFrame.

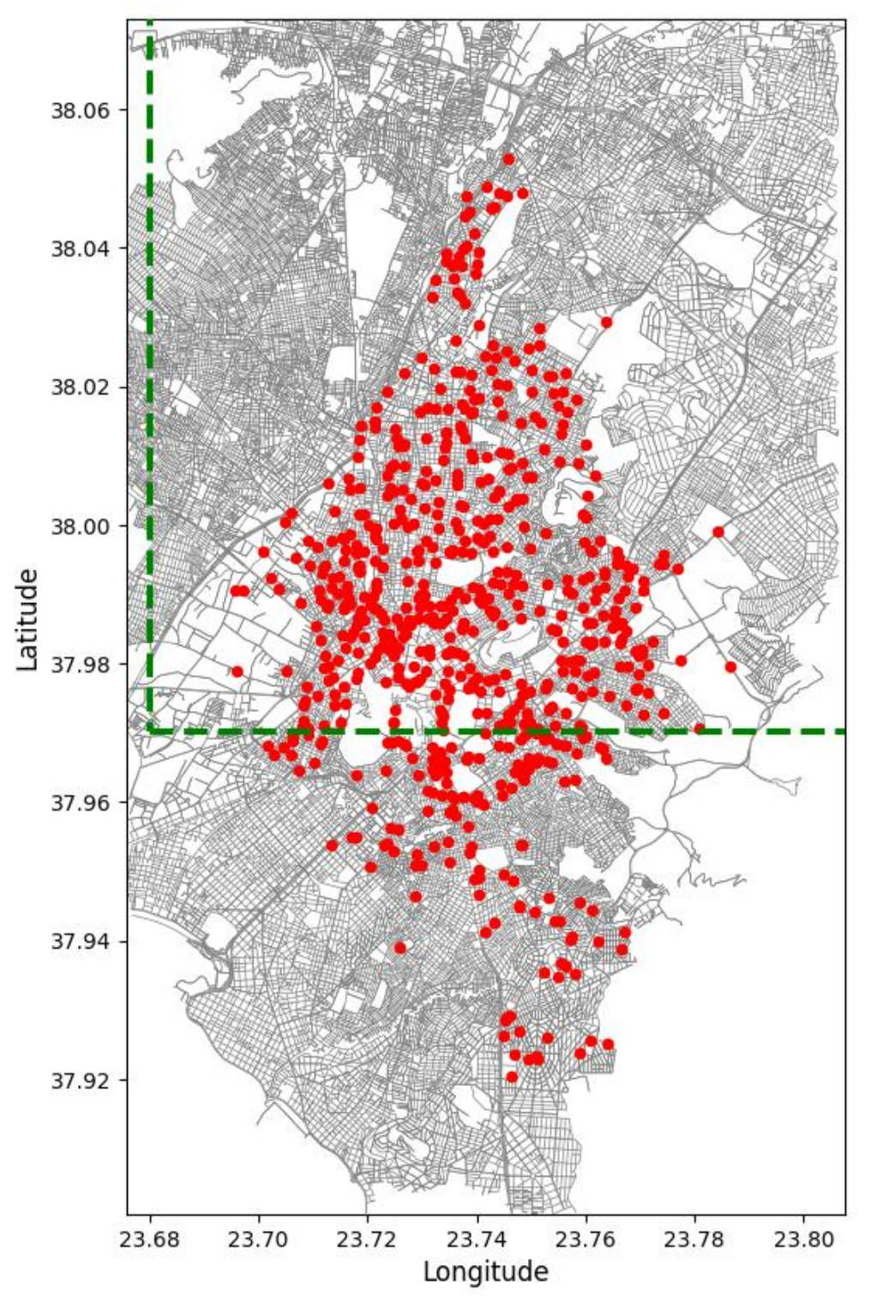

Figure 2: *Intersection crashes and Telematics coverage*

Methodology

- ❑ The study area was defined by telematics data coverage from OSeven’s innovative smartphone app. A bounding box (Figure 2, dashed line) was used to extract the road network via the **OSMnx** Python library. The resulting graph includes node and edge GeoDataFrames for further analysis.
- ❑ After aggregating data per node, **6716 nodes** were enriched with attributes from multiple sources. Matching OSM node IDs with crash data flagged **100 crash sites**, resulting in a highly imbalanced dataset.
- ❑ High **kurtosis** in some predictors prompted robust outlier removal using **DBSCAN**. After cleaning, 100 crash cases remained out of **6686 nodes**. Highly correlated and redundant features were dropped using **correlation** and **VIF** analysis to reduce multicollinearity.
- ❑ Four models have been trained comparatively: A Penalized **L1-Logistic Regression**, k-Nearest Neighbors (**k-NN**), **Random Forest**, and Support Vector Classifier (**SVC**).
- ❑ **Stratified K-Fold** ensured class balance across splits. **SMOTE** and **RandomUnderSampler** were applied to address class imbalance in each model.

Results

Models and **feature importance** were analyzed and compared:

- ❑ Logistic Regression with LASSO (L1) highlighted features like the number of **connected streets** and **phone usage** as increasing crash risk, while speeding and harsh acceleration reduced it.

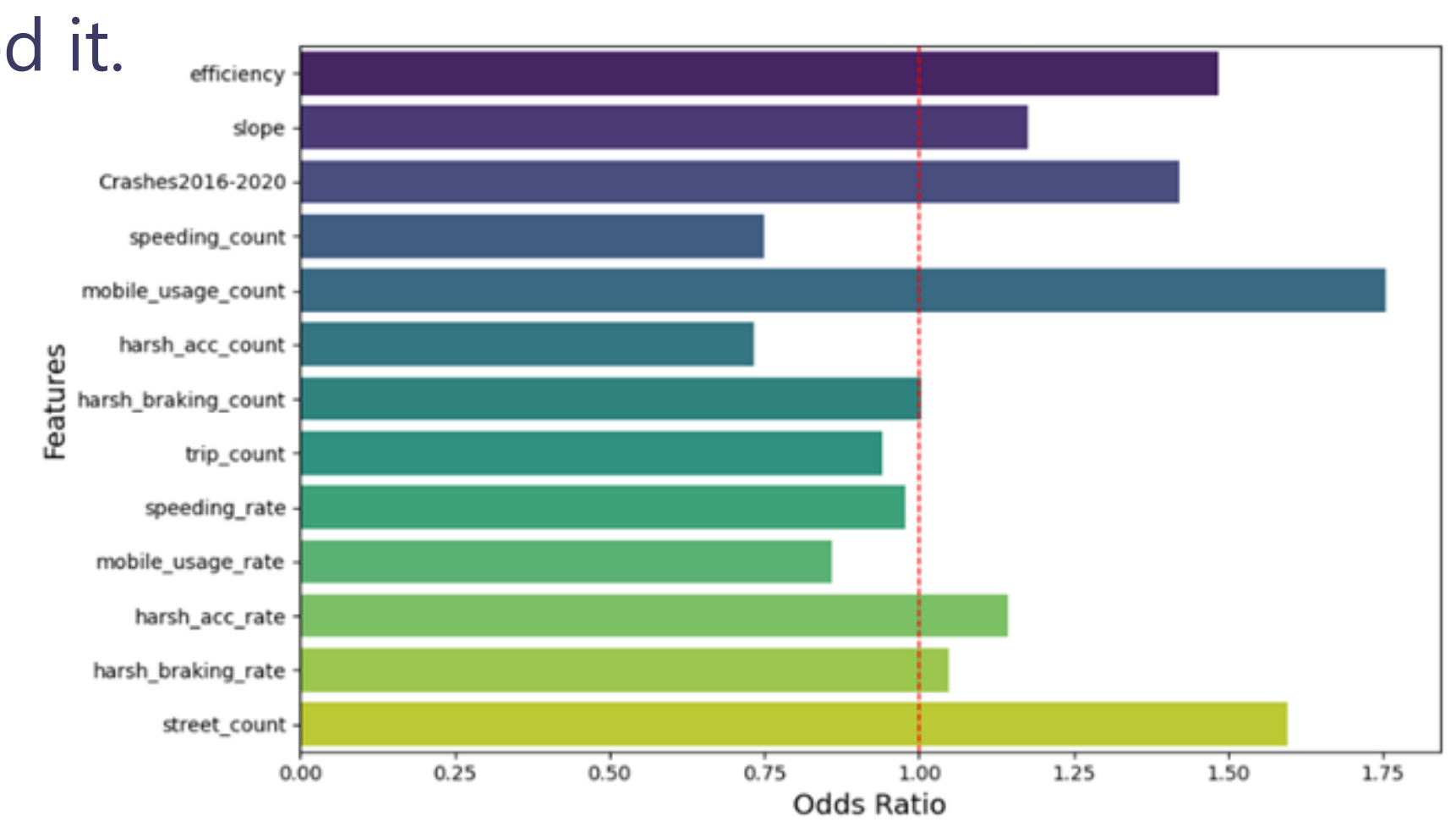

Figure 3: *Odds Ratios for L1 Logistic Regression*

- ❑ **SHAP values** for Random Forest confirms **similar important features** and crash risk patterns.

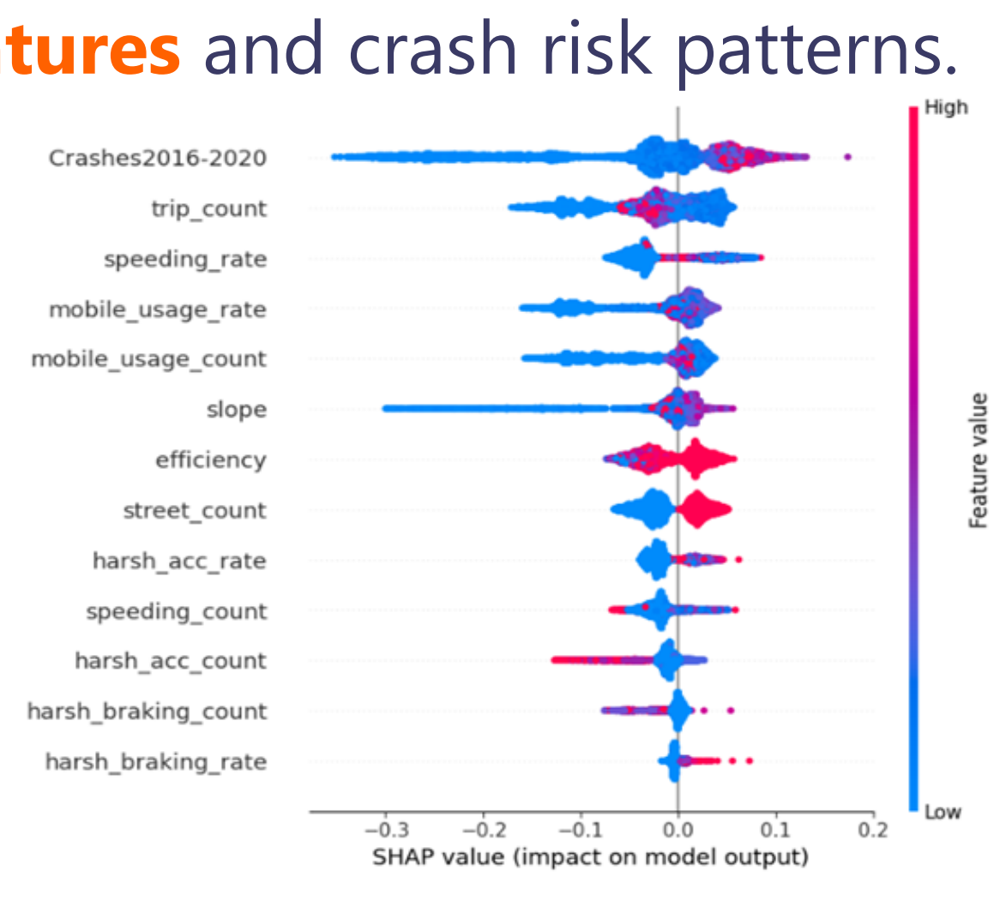

Figure 4: *Random Forest SHAP values*

- ❑ k-NN (k=8) performance was the **poorest**.
- ❑ **SHAP values** calculated on a **subset** of the initial dataset for SVC with **Gaussian kernel** and **regularization parameter** C=0.1 identifies **historical crashes** and **slope** as key predictors.

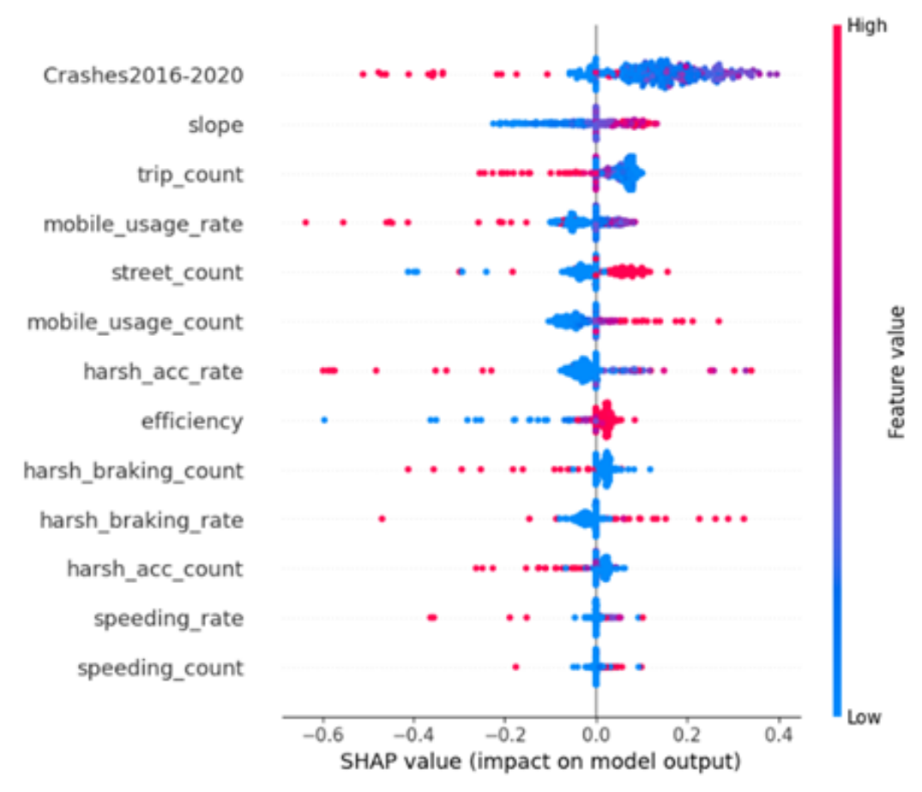

Figure 5: *Support Vector Classifier SHAP values*

Classification Results

The high class imbalance in the dataset causes models to favor the **majority class** to optimize overall performance metrics, often at the expense of the **minority class** (crashes). As a result, many false positives are produced for the crash class. However, given the context of crash prediction, this **conservative bias** might still be considered acceptable. This trend is observed across all models, as illustrated in **Figure 6**.

Lowering the **threshold** for the binary classification leads toward higher recall but lower precision, which is an ideal approach, where **failing to detect an event** is costlier than many false positives. Increasing it leads toward lower recall but higher precision, ideal when **false positives are problematic**.

The **AUC** was chosen to evaluate the model performance, since it measures the model’s ability to classify instances across all possible threshold values. L1 Logistic Regression and the Random Forest slightly outperform the other models with an AUC of 0.70 and 0.71 respectively. Whereas k-NN is **performing poorly** in the classification, with an AUC of 0.58.

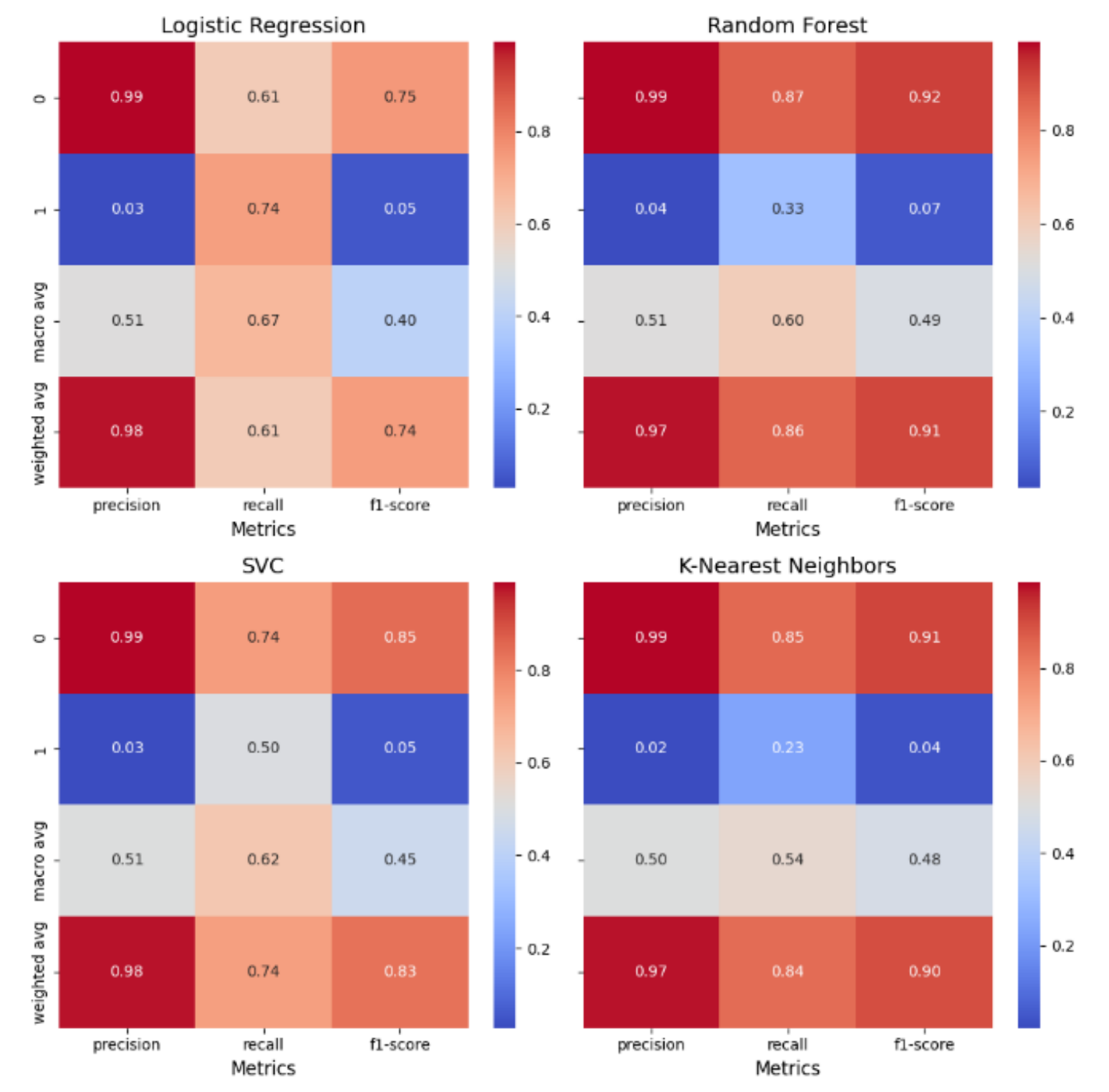

Figure 6: *Classification report per each model*

Conclusions

- ➢ Concerns have been raised about the quality and reliability of **police reports**, along with challenges in **merging data** from different sources.
- ➢ NLP has proven valuable to work with **textual data** and use it to overcome the lack of alphanumeric keys normally used to merge datasets and work on names of the street.
- ➢ While **resampling techniques** can address class imbalance, using **surrogate safety measures** offers a more natural solution, as near-crash events are more frequent, reducing the need for synthetic data..
- ➢ Future studies could explore using the **count of crashes** rather than crash occurrences for a more detailed analysis.

Acknowledgments

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101119590

Contact Information:
Simone Paradiso, IVORY PhD Candidate and Researcher
Department of Transportation Planning and Engineering
Email: simone_paradiso@mail.ntua.gr
Website: https://linktr.ee/simone_paradiso