A Graph Transformer Approach for Modeling Crash Occurrence at Intersections Using Telematics-Informed Road Networks

Paper 135

Simone Paradiso

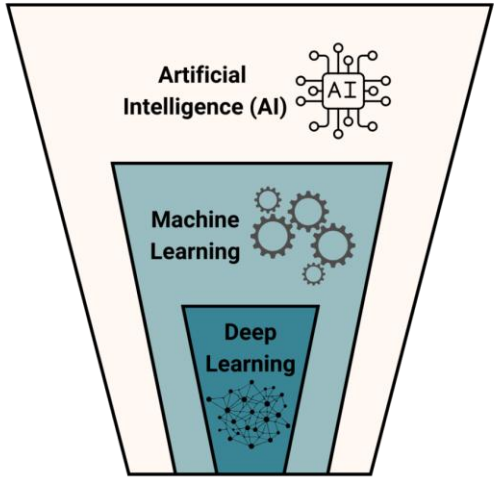
National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Iroon Polytechneiou Street, 15773, Athens, Greece
OSeven Single Member Private Company, 27B Chaimanta Street, 15234 Chalandri, Greece

simone_paradiso@mail.ntua.gr

Introduction

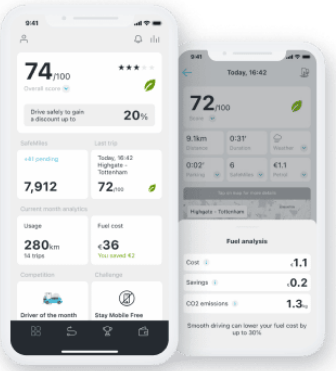
- ## Spatial Road Safety Analysis
  - **Spatial analysis** has been widely adopted in road safety to enable understanding of spatial patterns relevant to **crash analysis** and **road safety monitoring**.
  - While traditionally studied by employing econometrics and statistical approaches, advances in **Artificial Intelligence** (AI), particularly deep learning models have opened new avenues.

- ## Artificial Intelligence
  - **Tree-based models**, such as like Random Forest, CatBoost and XGBoost have shown strong performance.
  - The development of **Graph Neural Networks** (GNNs) enabled graph-structured information to be incorporated into modelling.
  - **XGBoost** → practical standard model operating on **traditional tabular data**.
  - **GNNs** → transition to topology-aware learning using **graph-structured data**.



Data Sources

- **OSeven Telematics Data**
  - **Trips** within the city center of Athens collected via a **smartphone application** at a **1 Hz** frequency collected during the last **four months of 2024**.
  - All data were **anonymized** and processed in compliance with **European data protection regulations** (GDPR).
  - Driving behavior metrics are generated using **proprietary machine learning algorithms**, validated against OBD data, on-road tests, and literature benchmarks.



- **OpenStreetMap (OSM)**
  - **Free and editable global map** released under an open-content license.

- **ELSTAT Database**
  - Historical road **crash data** for Athens, Greece (**2018–2022**), including crashes with injuries or fatalities.

Data Preprocessing

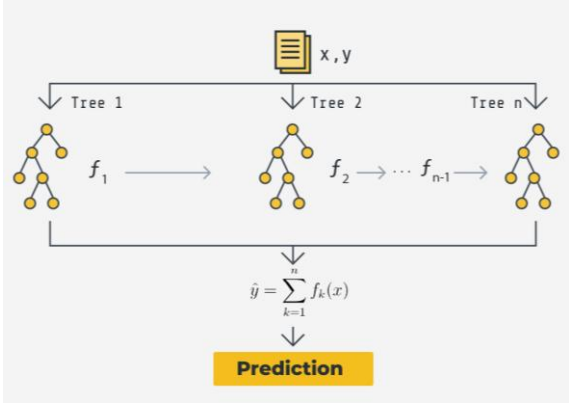
- **Telematics informed road network**
  - A **bounding box** was defined and used to extract a **graph** alongside road features from **OSM**.
    - Telematics observation were matched to the **nearest OSM edge**, and metrics were aggregated at the edge level.
    - Only telematics **observations on the edges connected to the corresponding node** within a **50-m radius buffer** area were considered for aggregation at the node level.
    - The process resulted in **9,615 telematics nodes** and **13,656 telematics edges**.

Feature	Description
street_count	Number of streets connected to the intersection
smoothenedSpeed	Average speed near this intersection
SpeedingFlag_per_trip	Count of speeding events occurred near the intersection per trip
mobile_usage_per_trip	Count of mobile phone usage events near the intersection per trip
harsh_acc_per_trip	Count of harsh acceleration events near the intersection per trip
harsh_brk_per_trip	Count of harsh braking events near the intersection per trip
event_intensity	Average intensity score of harsh events occurring nearby the intersection
speed_std	Speed standard deviation near the intersection

- **Intersection-level crash mapping**
  - Crash records lacked **geographic coordinates**.
  - Street names were normalized and geocoded using the **Google Maps Geocoding API**.
  - **572 of 9,615 telematics** nodes were matched to **at least one crash**.
  - Mapping accuracy was limited by **poor data quality** and missing street numbers.

Modelling Approaches

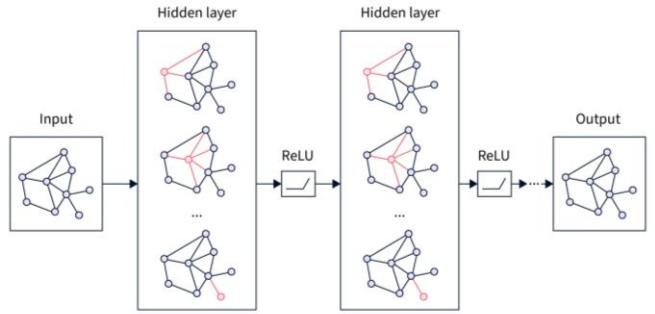
- **XGBoost**
  - **Gradient boosting algorithm** that builds an ensemble of **decision trees** sequentially, with each tree correcting errors from previous ones.
  - **Key hyperparameters**: *n_estimators* (number of trees), *learning_rate* (learning rate), *max_depth* (tree complexity), *colsample_bytree* (feature sampling), *subsample* (training sample fraction).



- **Graph Neural Network**
  - Processes graph-structured data by updating each node's representation (h_i') using information from its neighboring nodes through **message passing** and **aggregation**.

$$h_i' = f\left(h_i, AGGREGATE\left(\{h_j \mid j \in \mathcal{N}_i\}\right)\right)$$

Where i is the node being updated, j is a neighboring node, h_i and h_j are the respective current representations, and $\mathcal{N}_i$ is the set of neighbors of node i.



  - **Key hyperparameters**: *hid1*, *hid2* (hidden layer sizes), *dropout* (regularization), and *lr* (learning rate).

Explainable Artificial Intelligence (XAI)

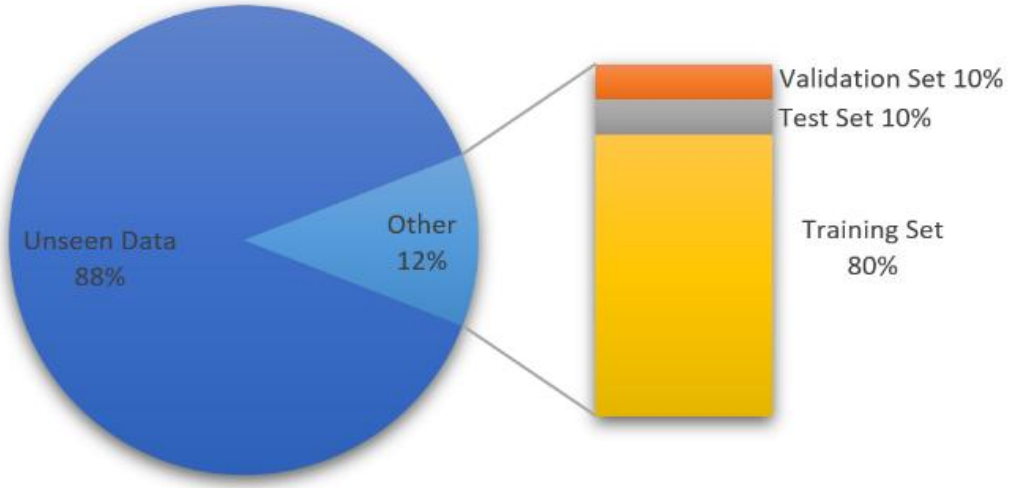
- **XAI technique termed Integrated Gradients (IG)**
  - IG measures how much **each input feature** contributes to a **GNN's prediction**.
  - By gradually **nudging** the input from a **baseline** (typically all zeros) toward the **actual feature values**, the method interpolates along this path and **accumulates the gradients** to capture how the **output changes**.
  - The attribution for feature i is computed as:

$$IG_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} \, d\alpha$$

Where F is the GNN model, x is the feature vector of a single node, x' is the baseline which is often a vector of zeros and α is a scalar parameter that ranges from 0 to 1.

Training Dataset Preparation

- **Data Split**
  - **572 of 9,615 telematics** nodes were **labeled as positive** (matched to at least one crash).
  - All 572 positive nodes were retained, and an equal number of negative nodes were randomly sampled, forming a **balanced dataset of 1,144 –nodes** (~12%; undersampling).
  - A first **stratified split** was then applied to maintain balanced class representation, dividing the **constructed dataset** into training (80%), validation (10%), and test (10%) sets.



Hyperparameter Optimization and Model Evaluation

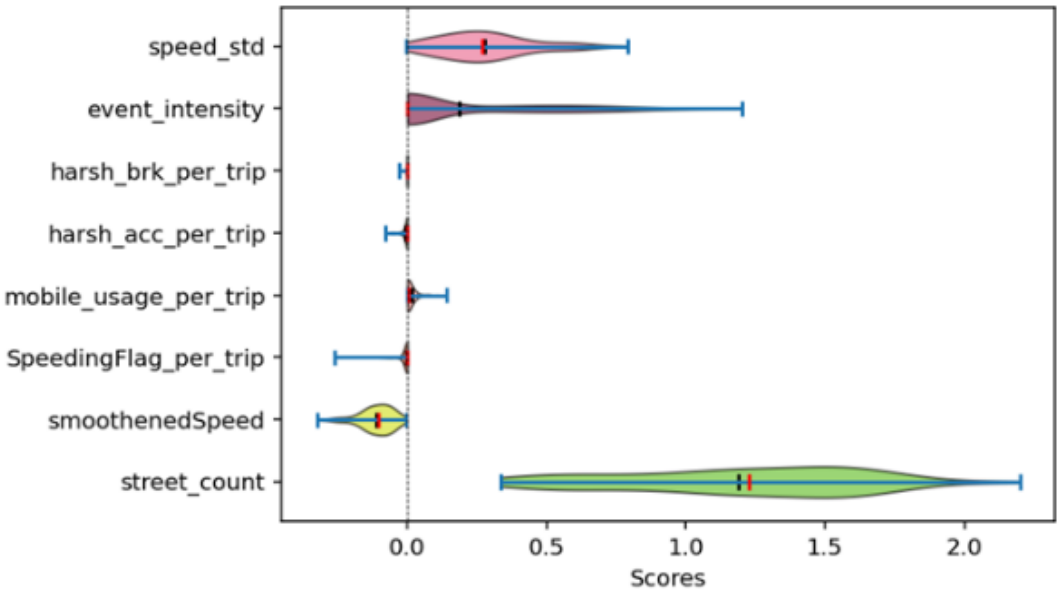
- **XGBoost**
  - 5-fold **stratified cross-validation** on combined training + validation set
  - **Min-max scaling** applied (fit on training set only to prevent data leakage)
  - **324 hyperparameter** configurations evaluated; best model achieved **71.8% cross-validated accuracy**.
- **GNN (3× (TransformerConv + ReLU + Dropout) + Linear head)**
  - **Min-max scaling** applied (fit on training set only to prevent data leakage)
  - **54 hyperparameter** configurations evaluated
  - Trained up to **150 epochs** with **early stopping** (based on validation set); best model achieved **78.9% validation accuracy**.

- **Evaluation using AUC**

Model	Balanced Test Set – AUC (best model)	Extended Test Set – AUC (best model)	Hyperparameter Ranges Evaluated	Optimal Hyperparameters
XGBoost	0.67	0.72	• n_estimators: [50, 100, 200], • learning_rate: [0.01, 0.1, 0.2, 0.3], • max_depth: [3, 5, 7], • colsample_bytree: [0.5, 0.7, 1], • subsample: [0.5, 0.7, 1]	• n_estimators= 50, • learning_rate = 0.01, • max_depth = 5, • colsample_bytree = 1, • subsample = 0.5
GNN	0.79	0.78	• hid1: [14, 21], • hid2: [21, 28, 35], • lr: [0.001, 0.0005, 0.0001], • dropout: [0.2, 0.3, 0.4]	• hid1 =14 • hid2 = 35, • lr = 0.001, • dropout = 0.2

Explaining Model Predictions with Integrated Gradients

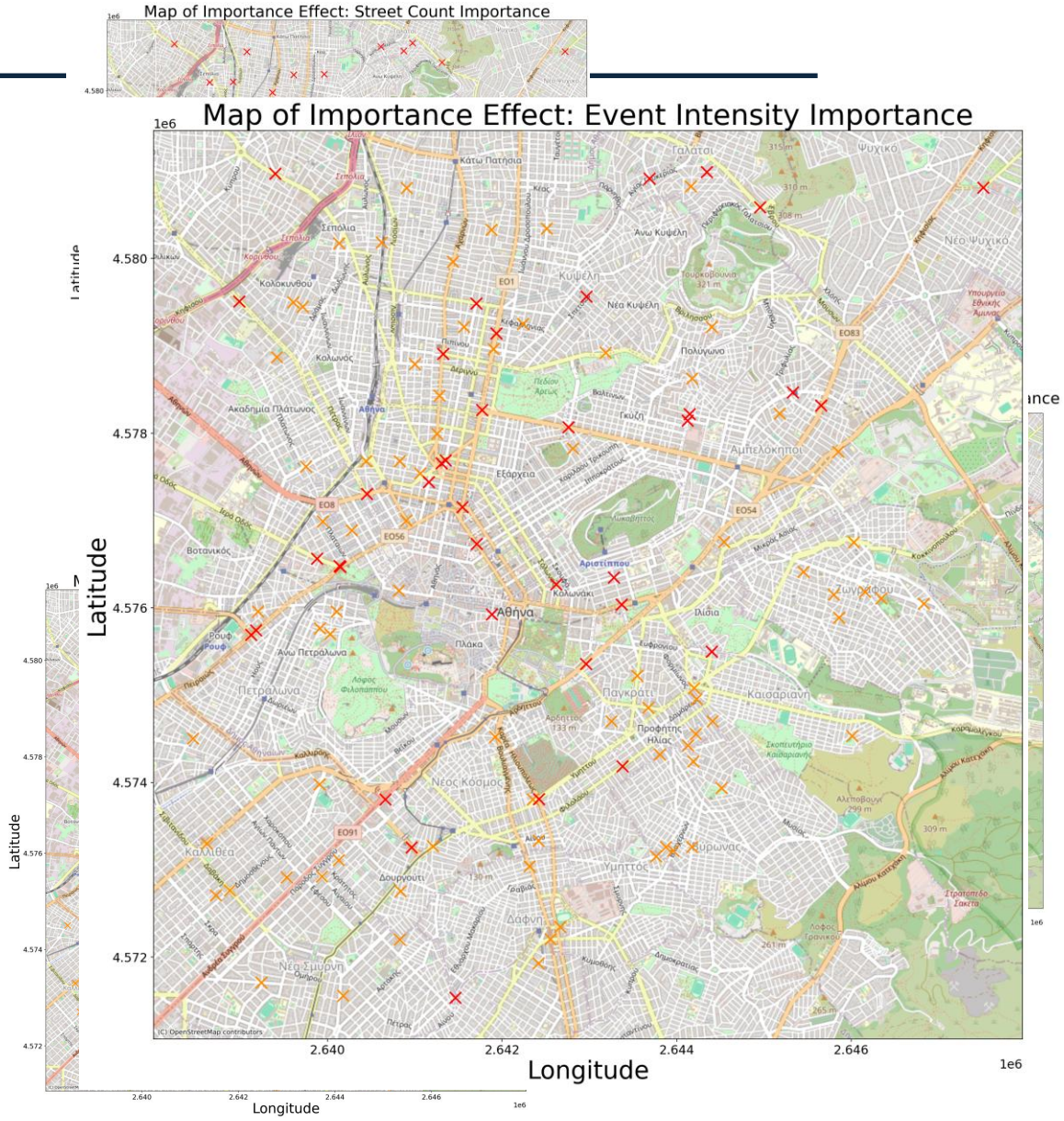
- $IG_i(x)$ **scores** computed for **each test node** to quantify how each input feature affects the **model's predictions on new unseen data**.
- Min-max normalization ensures the **zero baseline** represents the **absence of a feature**.
- Node-level attributions were aggregated into **violin plots** to provide a global view of feature importance.
- Analysis was performed with and without **speed standard deviation**; its inclusion did not significantly alter model performance nor its reasoning.



- **More streets at an intersection** → higher crash risk due to increased conflict points.
- **Higher intensity of harsh driving events** → associated with greater crash risk.
- **Greater speed variability** → indicates unstable traffic flow leading to increased crash likelihood.

Identifying Priority Intersections

- **Node-level feature importance scores $IG_i(x)$**
  - Based on the violin plot distributions, a **threshold of $\pm$ 0. 02** was selected to classify **feature effects** at **each node**:
    - 🟢 $IG_i < -0.02$ feature i **reduces crash likelihood**.
    - 🟠 $-0.02 < IG_i < +0.02$ feature i shows **negligible attribution to predicted crash risk**, suggesting the node may be deprioritized for interventions on that feature.
    - 🔴 $IG_i > +0.02$ feature i **increases crash likelihood**, suggesting locations where interventions may be beneficial.



Conclusion

- **Key Findings**
  - **GNN outperformed XGBoost** for spatial crash occurrence prediction.
  - **Integrated Gradients (IG)** enabled node-level feature attribution and visualization on the road network.
  - The proposed framework provides **actionable insights into infrastructure**, helping to understand factors associated with road safety and **to identify and prioritize potentially unsafe intersections for targeted interventions**.
- **Limitations & Future Work**
  - Although **324 hyperparameter configurations** were evaluated for XGBoost, **only 54 were tested for the GNN**. More extensive tuning and additional computational resources could further improve GNN performance and enable analysis over larger study areas with richer topological relationships.
  - The **optimal binning thresholds** require additional investigation.
  - The **temporal mismatch** between crash and telematics data should be addressed in future studies.

A Graph Transformer Approach for Modeling Crash Occurrence at Intersections Using Telematics-Informed Road Networks

Paper 135

Thank you for your attention!

simone_paradiso@mail.ntua.gr