Predicting pedestrian illegal crossing using vehicle dynamics with LightGBM

Roussou S., Ziakopoulos A., Yannis G.

I. INTRODUCTION

Understanding pedestrian behavior in urban traffic has paramount significance in the field of road safety, especially in the cases of intersections where the pedestrian–vehicle interaction is frequent. The objective of this study is to investigate whether vehicle movement features can predict pedestrian non-compliance (illegal crossings) using a LightGBM classifier, which is a gradient boosting framework based on decision trees, known for its efficiency and accuracy in structured data classification tasks, contributing to the development of intelligent crosswalk monitoring systems.

Recent studies have highlighted the concern over pedestrian safety. Infrastructure, environmental, and driver-related features are primary determinants of pedestrian crash patterns, which can be effectively uncovered through machine learning and econometric modelling [3]. Factors like vehicle speed, arrival order, age, and crossing delays can be modeled in prediction models and have shown their influence on pedestrian crossing behavior [2] [4]. Moreover, illegal crossings lead to more abrupt vehicle reactions and reduced driver yielding behavior [1]. These findings support the hypothesis that vehicle dynamics may detect early indicators of hazardous pedestrian behavior independently.

Although there is growing interest in using machine learning to model pedestrian behavior, most research relies on pedestrian-side factors such as age and decision context. There has been little investigation of whether data from vehicle movements alone, independent of fine-grained pedestrian input, can forecast illegal pedestrian crossings. This provides a gap for developing lightweight, vehicle-centric models deployable in smart vehicles or infrastructure with limited sensing. Our study aims to do this by validating the predictive ability of vehicle dynamics (i.e., speed, position) to predict pedestrian non-compliance using a LightGBM classifier.

II. METHODS

To create the dataset used in behavioral classification, a computer vision algorithm with YOLOv8 (You Only Look Once) and ResNet-50 models, with Kalman Filtering and homography transformations mapping image coordinates to ground plane positions, was used [5]. This algorithm was applied to a video recording from the Omonoia location in the city center of Athens. Through its detection and tracking, real-world trajectories of pedestrians and vehicles were acquired as depicted in Figure 1 below.

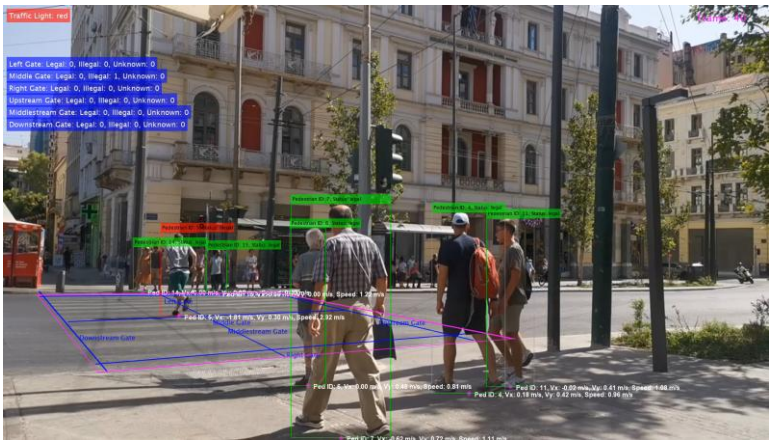

Figure 1: Algorithm Example of Processing

A dedicated preprocessing step was conducted to combine pedestrian and vehicle data, but only temporarily through the feature of timestamp (i.e., for the purpose of aligning vehicle and pedestrian movements at the same moment without long-term trajectory merging). Each record was then labeled as 'legal' or 'illegal' based on compliance of pedestrians with the traffic light signal. The final dataset included vehicle properties such as position coordinates (x, y), velocity vectors (vx, vy), speed magnitude, and direction of movement that were

related to pedestrian behavior at the moment in time, and it was used as ground truth for training a LightGBM classifier to predict pedestrian non-compliance solely based on vehicle behavior.

III. INITIAL FINDINGS

The LightGBM Classifier was trained on a pedestrian behavior dataset (legal vs. illegal crossing) and a vehicle dynamics dataset (position and velocity vectors). The model was validated with an 80/20 train-test split. The dataset consisted of approximately 66% legal and 34% illegal crossing samples, leading to moderate class imbalance. To mitigate this before training, we applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes and improve model sensitivity to illegal events. A two-stage modeling strategy was applied to refine predictions on high-risk cases, those including illegal crossings and close vehicle interaction, increasing the possibility of a crash.

	Precision	Recall	F1-Score	Support
0 (Legal)	0.84	0.89	0.86	3705
1 (Illegal)	0.76	0.66	0.71	1910
Accuracy		0.82		5615
Macro avg	0.80	0.78	0.79	5615
Weighted avg	0.81	0.82	0.81	5615

Table 1: Classification Report

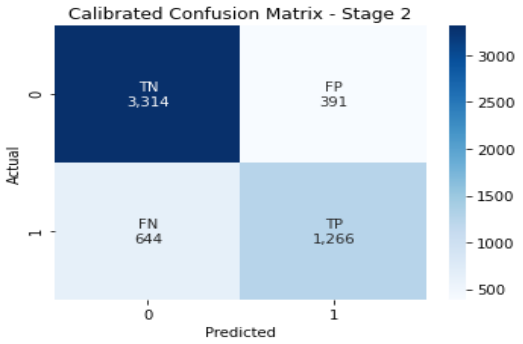

Figure 2: Confusion Matrix

The results indicate that vehicle speed and coordinates are good predictors of pedestrian crossing action, even without pedestrian-side features. The model performed particularly well in identifying legal crossings (high precision and recall), while maintaining moderate recall for illegal events. The absolute value of velocity was the most important feature, meaning that oncoming speed is the most important factor in the probability of non-compliance by pedestrians.

IV. DISCUSSION

The initial findings confirm the viability of vehicle-based feature prediction of pedestrian illegal crossings. The LightGBM classifier achieved satisfactory overall accuracy (82%), high precision for illegal cases, and satisfactory recall through a second-stage targeted refinement. While the overall accuracy is 82%, the lower recall for illegal crossings indicates that the model remains more conservative in flagging non-compliant behavior, potentially due to residual bias toward the majority class. This confirms the hypothesis that patterns of vehicle movement can convey indirect information about pedestrian behavior, even in the absence of direct pedestrian-side data.

This outcome fills a crucial knowledge gap in the literature, as many previous models have required high-density pedestrian inputs such as age, stance, or intent to cross. Our approach makes it possible for light and scalable solutions to be used in intelligent infrastructure or on-board systems within networked cars with limited sensing abilities.

However, some of these limitations remain. Data comes from only one intersection in Athens (Omonoia) and therefore, generalization across other urban areas is not yet clear. Moreover, while timestamp alignment allowed synchronization between pedestrian and vehicle data, the model may remain sensitive to temporal granularity. Future work will have to try the model with different intersections and examine more sophisticated fusion techniques between pedestrian and vehicle trajectories. Future work will also include statistical profiling of illegal crossing instances by analyzing associated surrogate safety measures (SSMs), such as time-to-collision and vehicle proximity, to offer behavioral insights that complement the black-box nature of machine learning models

V. REFERENCES

[1] Bella et al., Transportation Research Procedia, 2020.
[2] Cai J et al., Sensors, 2024.
[3] Rella et al., Sustainability, 2022.
[4] Singh et al., Multimodal User Interfaces, 2024.
[5] Ventura et al., Transportation Research Interdisciplinary Perspectives,2024.

The present research was carried out within the project “PHOEBE - Predictive Approaches for Safer Urban Environment”, which has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101076963.
Stella Roussou is a PhD Candidate in Road Safety at the National Technical University of Athens (Department of Transportation Planning and Engineering - 5, Iroon Polytechniou - GR-15773 Athens, Greece). Apostolos Ziakopoulos is a Senior Postdoctoral Researcher in Road Safety at the National Technical University of Athens (Dpt. as first author). George Yannis is Professor of Road Safety at the National Technical University of Athens (Dpt. as first author).