Modelling Pedestrian Signal Violations Using Video-Based Behavioural and Traffic Data

Stella Roussou*[1], Apostolos Ziakopoulos[2], George Yannis[3]

[1] Ph.D. Candidate, Research Associate, Department of Transportation Planning and Engineering, National Technical University of Athens, Athens, Greece

[2] Ph.D., Senior Associate, Department of Transportation Planning and Engineering, National Technical University of Athens, Athens, Greece

[3] Professor, Department of Transportation Planning and Engineering, National Technical University of Athens, Athens, Greece

SHORT SUMMARY

Pedestrian non-compliance at signalized intersections is a major safety concern in urban areas, particularly in locations with high pedestrian demand and complex traffic conditions. This study investigates the factors influencing pedestrian signal violations using video-based data collected at a signalized intersection in Athens, Greece. Pedestrian trajectories, vehicle trajectories, signal status, and interaction variables were extracted using a computer vision pipeline. Statistical analysis, including point-biserial correlation, binary logistic regression, and Random Forest classification, was applied to examine the relationship between traffic conditions and pedestrian compliance behaviour and to predict illegal crossings. The results indicate that pedestrian violations are strongly associated with traffic conditions such as vehicle speed, Time-to-Collision (TTC), signal phase, and vehicle presence. The predictive models achieved high classification accuracy, indicating that pedestrian non-compliance is not random but influenced by observable traffic and interaction conditions. The findings support proactive road safety analysis and the development of predictive safety assessment tools.

Keywords: Machine learning; Pedestrian behaviour; Signalized intersections; Traffic safety; Video analysis;

1. INTRODUCTION

Pedestrian safety is a significant issue in urban transportation infrastructures, especially at signalized intersections where interactions between pedestrians and vehicles are common and complex. A significant number of pedestrian crashes happen at intersections, and most of them are associated with non-compliant pedestrian behavior, such as crossing the intersection during the red signal or crossing outside the designated crossing area (Papadimitriou et al., 2013). Such non-compliant behavior of pedestrians may increase the probability of crashes involving pedestrians and vehicles.

Non-compliance of pedestrians may include crossing the intersection during the red signal, crossing the intersection during the flashing red signal, or crossing outside the designated area. Such non-compliance of pedestrians may be influenced by a number of factors, including waiting time, volume of traffic, speed of vehicles, characteristics of pedestrians, and environmental factors. Previous studies have showcased that signal indication is not the only factor of the pedestrians'

decision while crossing, but also the perceived risk and available gap in the traffic flow are key factors to that decision (Theofilatos et al., 2021; Tian et al., 2022;Wang et al., 2017). This behaviour suggests that pedestrian non-compliance is not random but influenced by traffic environment conditions and perceived safety margins

Understanding pedestrian behavior and the factors that influence non-compliance of pedestrians to the traffic signal is important for ensuring road safety and designing appropriate strategies for efficient management of traffic. Previous studies of pedestrian behavior are based on traditional methods, including observation, survey, and analysis of crash data. However, crash data may not be sufficient for analysis, as crashes are relatively less common events and do not provide sufficient information on the behaviors and interactions that happen before the crash happens. For this reason, surrogate safety measures and behavioural analysis have been increasingly used in road safety research to understand risky interactions before crashes occur (Saunier and Sayed, 2010; Tarko et al., 2008).

With the rapid development of computer vision and video analytics, it is now possible to extract detailed trajectory data and behavioural information from video recordings. Video-based data allow the analysis of pedestrian and vehicle movements, speeds, interactions, signal compliance, and conflict indicators such as Time-to-Collision (TTC) and Post-Encroachment Time (PET). These data provide valuable information for understanding pedestrian decision-making and interaction behaviour at signalized intersections (Ismail and Sayed, 2009).

The objective of this research is to examine the factors affecting pedestrian non-compliance in signalized intersections and to propose predictive models of illegal pedestrian crossings based on variables derived from video-based data sources. Statistical analysis and machine learning techniques are used to examine which variables are most significant in affecting pedestrian behavior and to assess the effectiveness of predictive models in distinguishing between legal and illegal pedestrian crossings. The research findings are expected to contribute to proactive road safety analysis and behavioral modeling in signalized intersections.

2. METHODOLOGY

This study is based on video data collected at a signalized intersection in Athens, Greece. The video recordings were processed using a computer vision pipeline in order to extract pedestrian and vehicle trajectories, signal status information, and interaction variables. Video-based trajectory extraction has been widely used in traffic safety studies to analyse interactions and surrogate safety measures (Saunier and Sayed, 2010).

The computer vision pipeline included object detection, object tracking, trajectory extraction, and signal phase detection (Ventura et al., 2025). From the processed video data, a dataset was created including variables related to pedestrian behaviour, traffic conditions, and pedestrian–vehicle interactions. The dataset included variables such as pedestrian speed, vehicle speed, signal status, Time-to-Collision (TTC), vehicle presence, and crossing location. TTC is commonly used as a surrogate safety indicator representing the time remaining before a potential collision if the trajectories of two road users remain unchanged (Ismail and Sayed, 2009).

Object Detection and Feature Extraction

The first step in processing is the detection of objects within the video frames. The YOLOv8 model is employed for this purpose, and it formulates object detection as a single regression problem. For an input image extracted as a frame from the video, YOLOv8 (Ultralytics, 2024) outputs a set of bounding boxes B=($B1, B2, \dots, B_n$), where each bounding box B_i is defined by:

$$B_i = (x_i, y_i, w_i, h_{i,} c_i) \quad \text{(Eq. 1)}$$

Where (x_i, y_i) are the coordinates of the center of the box, $w_i, and\ h_i$ are the width and height of the bounding box, and c_i is the class probability score indicating whether the object in the box is a pedestrian or vehicle.

Once the objects are detected, ResNet-50 is employed for feature extraction. The feature vector F_i corresponding to each detected object B_i is extracted from the output of ResNet-50, represented as:

$$F_i = ResNet - 50(B_i) \quad \text{(Eq. 2)}$$

These then form the basis for tracking and re-identification. For pedestrian tracking, the system relies on the application of Kalman filters (Kalman, 1960). A Kalman filter has two main steps: Prediction and Update. The Kalman Gain balances the uncertainties in the prediction and the measurement. In addition, it also makes use of the Hungarian algorithm for data association. This algorithm solves optimally the association between the predicted object locations and new detections to get a minimal cost function regarding object-to-track association.

Once objects are detected and tracked, the system analyzes pedestrian behavior in relation to traffic signal phases. The pedestrian behavior is classified as either legal or illegal depending on whether they cross the pedestrian line during a red pedestrian light. The behavior status of tracked pedestrian is $S_{ped}(t)$, where:

$$S_{ped}(t) = \begin{cases} illegal & \text{if } S_{light}(t) = red \\ legal & \text{if } S_{light}(t) = green \\ unknown & \text{if } S_{light}(t) = intergreen \end{cases}; \quad (3)$$

Each illegal crossing event is logged, and the number of violations is counted. It then generates annotated outputs in video and data file annotations after the detection of illegal crossings by pedestrians. This includes the video output with marked illegal crossings, along with time stamps, bounding boxes, and phases of the traffic signal corresponding to every detected event. The data files include illegal crossing counts along with violation times and locations.

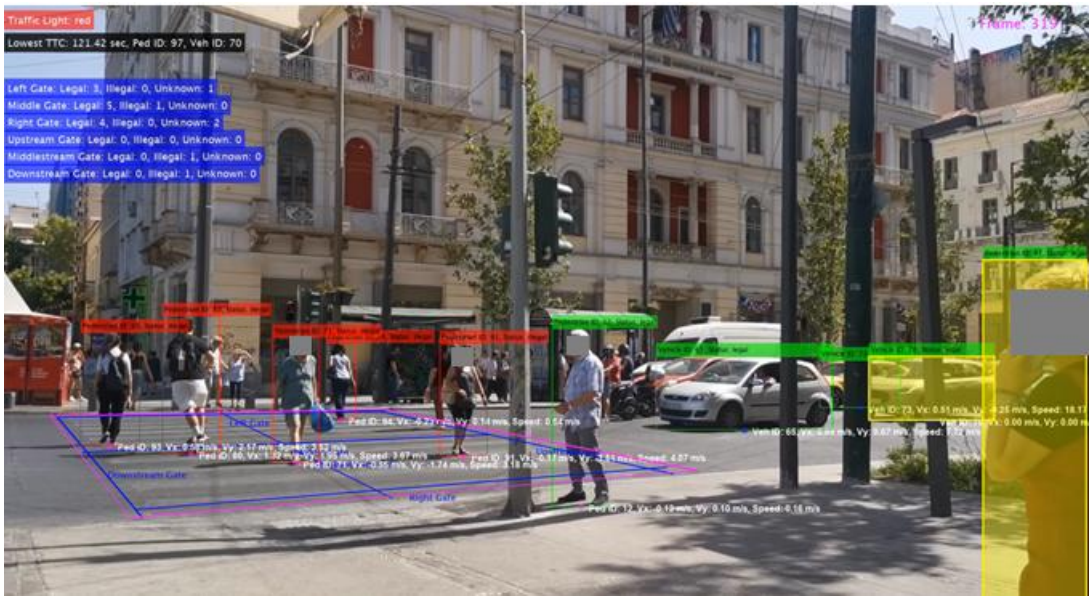


Figure 1: Omonoia Square in Athens

Two binary dependent variables were defined: a) Whether the pedestrian crossed legally or illegally, and b) Whether the pedestrian crossed during the pedestrian green phase or not.

To investigate the relationship between traffic variables and pedestrian behaviour, point-biserial correlation analysis was first applied to examine the relationship between continuous variables and binary outcomes. Then, binary logistic regression models were developed to estimate the probability of pedestrian violations based on traffic and interaction variables. Logistic regression is commonly used in behavioural and safety studies to model binary outcomes such as crash occurrence or violation behaviour (Papadimitriou et al., 2013).

In addition to logistic regression, Random Forest classification models were developed to predict pedestrian violations. Random Forest is a machine learning method that combines multiple decision trees and can capture non-linear relationships between variables and complex interactions between traffic variables (Zhang et al., 2019). The Random Forest model was also used to estimate the relative importance of each variable in predicting pedestrian non-compliance.

Random Forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and combines their outputs to improve predictive performance. Its robustness to noise and overfitting makes it particularly suited for datasets with imbalanced classes. Key features of the RF implementation include:

- Bagging (Bootstrap Aggregating): By training each tree on a different random subset of data, RF reduces variance and enhances generalization.
- Feature Importance Ranking: RF inherently ranks features by importance, aiding interpretability and highlighting key predictors, such as pedestrian speed and traffic light phases.
- Optimization: Hyperparameters, including the number of trees, maximum tree depth, and minimum samples per split, were tuned using grid search. Class weights were adjusted to counter class imbalance, ensuring that minority classes (e.g., illegal crossings) were not overlooked.

Model performance was evaluated using classification accuracy and confusion matrices. The comparison between statistical and machine learning models allowed the identification of the most important behavioural and traffic variables influencing pedestrian violations.

3. RESULTS AND DISCUSSION

Based on the results of the analysis, it is clear that the non-compliance of pedestrians in signalized intersections is significantly influenced by factors such as traffic condition variables and interaction variables, but not by randomness. The point-biserial correlation analysis revealed significant correlations between pedestrian violations and variables such as vehicle speed, TTC, signal status, and vehicle presence. These findings are consistent with those of previous studies, which concluded that pedestrian decisions to cross the road depend on gap acceptance and perceived risk, rather than signal indications (Wang et al., 2017).

The results of the logistic regression analysis showed that the probability of illegal pedestrian crossing is higher if the speed of the vehicle is lower, if the value of TTC is higher, and if there are no vehicles approaching the area. This implies that pedestrians are more likely to cross the street illegally if they think that there is sufficient time for them to cross the street safely. In contrast, when vehicles are close to the crossing area and TTC values are low, pedestrians are more likely to wait for the green signal.

Another factor that was identified as significant in influencing pedestrian behavior was the signal status. It was also noticed that pedestrians were crossing illegally during the red signal, especially if the speed of vehicles was low and there was a gap in the flow of vehicles. This shows that pedestrians are dynamic in assessing the situation of the traffic.

Table 1: Results of Logistic Regression Models

Pedestrian presence in the crossing area					**Legal or illegal pedestrian crossing**			
Model:	Logit	Method:	MLE		Model:	Logit:	Method:	MLE
Dependent Variable:	IsInCrossing Area	Pseudo R-squared:	0.643		Dependent Variable:	Pedestrian Status	Pseudo R-squared:	0.907
Date:	2025-08-10 07:54	AIC:	7084.329		Date:	2025-08-04 10:56	AIC:	1977.827
No. Observations:	14278	BIC:	7152.427		No. Observations:	15236	BIC:	2046.510
		Log-Likelihood:	-3533.2		Model:		Log-Likelihood:	-979.91
		LL-Null:	-9896.8			BIC:	LL-Null:	-10561.
		LLR p-value:	0.000			Log-Likelihood:	LLR p-value:	0.000

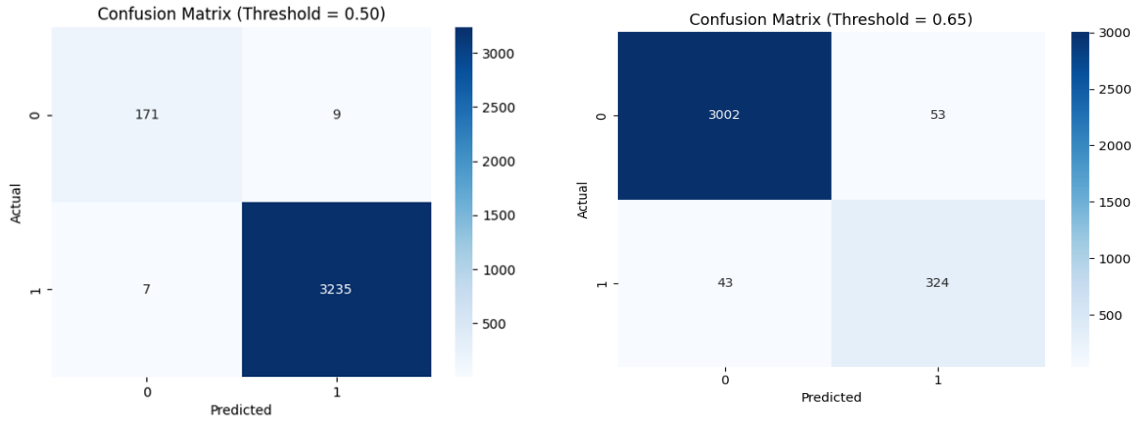
The Random Forest model showed high classification accuracy in predicting pedestrian violations, indicating that pedestrian behaviour can be predicted using traffic and interaction variables. The variable importance analysis showed that signal status, TTC, vehicle speed, and vehicle presence were among the most important variables influencing pedestrian non-compliance. Machine learning methods have been increasingly used in traffic safety studies for behaviour prediction and risk modelling due to their ability to handle complex and non-linear relationships between variables (Zhang et al., 2019).

These results suggest that pedestrian violations are influenced by perceived safety gaps and traffic conditions rather than random behaviour. Pedestrians appear to make risk-based decisions, crossing illegally when they perceive sufficient time and space to cross safely. This behaviour has important implications for traffic safety analysis and signal timing design.

Table 1: Random Forest Feature Importance of the variables

Pedestrian presence in the crossing area		**Legal or illegal pedestrian crossing**	
Feature	Importance	Feature	Importance
GroundPlaneCentroid_2	0.234	IsInCrossingArea	0.579
PedestrianStatus	0.181	ResolvedTrafficLightStatus	0.136
TrafficLightStatus	0.106	MinTTC_wins_log	0.034
MinTTC_wins_log	0.08	magnitude_veh	0.028
Confidence_veh	0.017	VehicleStatus	0.0002
IsInCrossingArea_veh	0.01		
VehicleStatus	0.01		

Figure 2: Confusion Matrix of Random Forest (a) Pedestrian presence in the crossing area (b) Legal or illegal pedestrian crossing



(a) (b)

From a road safety perspective, the results indicate that signal timing, vehicle speed management, and pedestrian waiting time may influence pedestrian compliance behaviour. Long waiting times and low vehicle speeds may encourage illegal crossings, while higher vehicle speeds and shorter waiting times may increase pedestrian compliance but may also increase crash severity in case of conflicts. Therefore, signal timing and traffic management measures should consider both pedestrian behaviour and safety outcomes.

The results also demonstrate the potential of video-based data and predictive modelling for proactive safety analysis. Predictive models of pedestrian behaviour could be used to identify high-risk locations, evaluate signal timing strategies, and support the development of intelligent traffic management systems.

4. CONCLUSIONS

The objective of the present study is to investigate pedestrian non-compliance behavior at signalized intersections using video-derived behavioral and traffic data statistically and through machine learning approaches to develop models for predicting pedestrian non-compliance behavior at signalized intersections.

It is concluded from the results of the present study that pedestrian non-compliance behavior is significantly influenced by traffic environment conditions such as vehicle speed, Time-to-Collision, signal status, and vehicle presence. The results of the present study revealed that pedestrian non-compliance behavior is not random and that pedestrians make tactical decisions in crossing signalized intersections. The results of the present study revealed that pedestrian non-compliance behavior can be predicted with high accuracy using logistic regression and random forest models.

It is concluded that pedestrian non-compliance behavior should be regarded as behavior influenced by traffic environment conditions and not random behavior. The study also concluded that video-derived data and machine learning models are valuable tools for proactive road safety analysis and modeling. The study also recommended that pedestrian non-compliance behavior at signalized intersections should be further examined through additional intersections and behavioral variables. Future research should examine the pedestrian non-compliance behavior in terms of its relationship with surrogate safety measures.

ACKNOWLEDGEMENTS

The present research was carried out within the research 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.

REFERENCES

Ismail, K., Sayed, T. [2009]. Risk-optimal hot spot identification using automated video analysis. Accident Analysis & Prevention.

Kalman, R.E. A new approach to linear filtering and prediction problems. J. Basic Eng. 1960, 82, 35–45.

Papadimitriou, E., Yannis, G., Evgenikos, P. [2013]. Road safety performance indicators: A literature review. Accident Analysis & Prevention.

PHOEBE Project. PHOEBE Library. https://phoebe-project.eu/phoebe-library/. Accessed July 1, 2024.

Saunier, N., Sayed, T. [2010]. Automated analysis of road safety using video data. Transportation Research Record.

Tarko, A. P. [2018]. Surrogate measures of safety.

Theofilatos, A., Ziakopoulos, A., Oviedo-Trespalacios, O., & Timmis, A. (2021). To cross or not to cross? Review and meta-analysis of pedestrian gap acceptance decisions at midblock street crossings. Journal of Transport & Health, 22, 101108.
Tian, K., Markkula, G., Wei, C., Lee, Y. M., Madigan, R., Merat, N., & Romano, R. (2022). Explaining unsafe pedestrian road crossing behaviours using a psychophysics-based gap acceptance model. Safety science, 154, 105837.
Ultralytics. (2024). *Ultralytics YOLOv8* [GitHub repository]. GitHub. https://github.com/ultralytics/ultralytics (Accessed on September 27, 2024).
Wang, W., Guo, H., Gao, Z. [2017]. Pedestrian crossing behaviour modelling at signalized intersections. Transportation Research Record.
Zhang, J., Wang, F., Wang, K. [2019]. Machine learning for traffic safety prediction. Accident Analysis & Prevention.
Ultralytics. (2024). *Ultralytics YOLOv8* [GitHub repository]. GitHub. https://github.com/ultralytics/ultralytics (Accessed on September 27, 2024).