CETRA2026 27-30 May 2026, Bol, Island Brač, Croatia
9th International Conference on Road and Rail Infrastructure

MACHINE LEARNING-BASED PEDESTRIAN INTENTION PREDICTION MODELS FOR COLLISION WARNING AT UNSIGNALIZED CROSSWALKS

Sami Haktan Cangut, Yalçın Alver
Ege University, Department of Civil Engineering, Turkey

Abstract

Pedestrian safety remains a critical challenge in urban environments and pedestrians account for approximately 21% of all traffic fatalities in Turkey. A major contributor to this risk is the uncertainty inherent in pedestrian–driver interactions at unsignalized crosswalks where drivers yield in only 29% of observed cases. Misinterpretation of pedestrian intentions in such ambiguous situations can lead to hazardous outcomes. This study addresses this problem by developing and comparing machine learning models that predict pedestrian crossing intention with high reliability. A comprehensive dataset containing 4, 835 pedestrian observations was created by combining the widely used JAAD, PIE, and PSI datasets. For each pedestrian, fifteen consecutive video frames were analyzed to extract kinematic features such as position and acceleration along with social interaction cues derived from nearby pedestrians. Three neural network architectures were evaluated using a two-stage transfer learning procedure. The models included a Bidirectional Long Short-Term Memory network (BiLSTM), a Social-LSTM incorporating social pooling mechanisms, and a Transformer model designed to capture complex temporal dependencies. Performance evaluation demonstrated that the Social-LSTM architecture achieved the highest overall performance with a precision of 85.02% and an F1-score of 84.57% by explicitly modeling the social context. Conversely, the Transformer model recorded an F1-score of 83.44% and suggested that architectural complexity does not guarantee superior performance when training data is limited. The BiLSTM provided a strong, kinematic-focused baseline with an F1-score of 84.06%. These findings confirm that sequential machine learning models can effectively anticipate pedestrian crossing intentions. The results highlight that incorporating social behavioral theories into engineering solutions is essential for developing robust collision warning systems aimed at reducing pedestrian–vehicle conflicts.

Keywords: pedestrian safety, intention prediction, machine learning, risk assessment, social-LSTM

1 Introduction

Urban transportation systems are undergoing a paradigm shift towards Vision Zero which aims to eliminate all traffic fatalities and severe injuries. Despite advancements in vehicle safety technologies and infrastructure design, pedestrians remain the most vulnerable component of the traffic ecosystem. According to the World Health Organization, pedestrians constitute approximately 23% of global road traffic deaths [1]. This highlights a critical need for robust safety interventions particularly in dense urban environments. The risk of pedestrian-vehicle conflict is particularly acute at unsignalized crosswalks where right-of-way rules are often ambiguous or disregarded.

In such environments, traffic safety depends not on signal phases but on the cognitive interpretation of behavioral cues which is a negotiation process described as joint attention [2]. Studies indicate that driver yielding rates at unsignalized crossings remain dangerously low in many regions, forcing pedestrians to make high-risk crossing decisions. To mitigate these risks, modern transport engineering must move beyond static infrastructure improvements to dynamic and intelligent solutions that can anticipate conflicts before they occur. Pedestrian Intention Prediction has emerged as a vital technology for Advanced Driver Assistance Systems and autonomous vehicles. Unlike traditional tracking methods that predict future positions based on Newtonian physics, intention prediction aims to estimate the cognitive state of the pedestrian before the physical action begins [3]. In this study, crossing intention is operationally defined as the probabilistic estimation of a pedestrian's cognitive commitment to enter the roadway within a future horizon of 0.5 to 1.0 seconds. This formulation distinguishes intention, which is the decision making phase inferred from behavioral cues such as head orientation and gait change, from the physical action of crossing. This distinction allows the system to generate alerts during the critical phase prior to movement. This capability provides a crucial safety buffer of 0.5 to 2.0 seconds allowing the vehicle to react smoothly and prevent collisions.

The literature on pedestrian behavior modeling has evolved from simple physical models to complex data-driven approaches. Early research relied on hand-crafted features and shallow learning methods. For instance, Völz et al. [4] utilized Support Vector Machines with features like optical flow to predict crossing probability. Similarly, Fang et al. [5] demonstrated the importance of body pose estimation as a precursor to movement. However, these methods often failed to capture the temporal evolution of human decision-making.

With the advent of deep learning, Recurrent Neural Networks and Long Short-Term Memory units became the standard for modeling temporal traffic data. Saleh et al. [6] and Rasouli et al. [7] showed that stacked LSTMs could effectively fuse visual context with motion history. Yet, a major limitation remained as pedestrians do not move in a vacuum. Their behavior is heavily influenced by the presence and movement of other road users which is a concept rooted in the Social Force Model originally proposed by Helbing and Molnar for crowd dynamics [8]. Addressing this, Alahi et al. [9] introduced the Social-LSTM which integrates a social pooling layer to share information between neighboring pedestrians thereby mathematically modeling the group interactions and avoidance behaviors observed in dense urban centers. Recently, the Transformer architecture has challenged LSTM-based dominance in sequence modeling [10]. Transformers offer the theoretical advantage of capturing long-range dependencies in complex traffic scenes. However, their application in safety-critical transportation tasks raises questions regarding data efficiency and robustness. Unlike LSTMs which process time sequentially, Transformers process sequences in parallel which may require significantly larger datasets to learn the causal relationships inherent in traffic safety scenarios [11].

Distinct from existing studies that predominantly focus on accuracy metrics within sanitized benchmarks, this work contributes to the literature by empirically identifying a critical safety gap in real-world deployment. We hypothesize that explicit modeling of social interactions provides greater safety benefits than purely kinematic or attention-based approaches for pedestrian intention prediction at unsignalized crosswalks. To test this, we constructed a large-scale and heterogeneous dataset by integrating three benchmark datasets totaling 4, 835 pedestrian samples. We systematically compare three architectures including BiLSTM, Social-LSTM, and Transformer by evaluating them on their ability to minimize false negatives in collision-prone scenarios.

2 Method

The proposed framework establishes a proactive safety assessment mechanism for unsignalized crosswalks using deep learning techniques. The methodology is structured to translate raw video data from urban traffic scenes into actionable intention predictions. This process involves three key stages: (1) the aggregation of heterogeneous traffic datasets to represent diverse conflict scenarios, (2) the formulation of sequential models to capture pedestrian kinematics and social interactions, and (3) a transfer learning strategy designed to calibrate these models for high-uncertainty environments.

2.1 Data acquisition and traffic scenario representation

To ensure the proposed models are robust against the stochastic nature of mixed traffic flows, this study integrates data from three diverse sources: JAAD [7], PIE [12], and PSI [13]. Unlike controlled laboratory experiments, these datasets capture naturalistic road user behavior in dense urban environments across North America and Europe. The combined dataset comprises 4, 835 unique pedestrian trajectories, specifically filtered to focus on critical interaction zones at unsignalized mid-block crosswalks and intersections. It is important to note that while these datasets provide diverse scenarios, they predominantly feature clear weather and optimal lighting. Following prior literature [15, 16], we acknowledge that adverse environmental conditions (e.g. rain, low light) introduce additional sensory noise and behavioral shifts such as increased walking speeds which remain a critical challenge for the robustness of vision-based systems.
The input state for each pedestrian at a given time step is defined by a kinematic feature vector that includes horizontal and vertical position, velocity, and acceleration components. These features are obtained from the centroids of pedestrian bounding boxes extracted from consecutive video frames. To account for scale variations resulting from different camera perspectives, such as dashboard-mounted cameras and fixed roadside cameras, all spatial coordinates are normalized with respect to the video frame width and height. This normalization ensures that the models learn relative motion patterns that are indicative of pedestrian crossing intention, such as deceleration before reaching the curb, rather than overfitting to absolute pixel coordinates.

2.2 Modeling pedestrian dynamics

Pedestrian crossing is not merely a physical action, but a complex decision-making process influenced by internal goals and external social forces. We evaluated three architectures to model these dynamics.

2.2.1 Kinematic modeling: Bidirectional LSTM

The Bidirectional Long Short-Term Memory (BiLSTM) network serves as the baseline, modeling the pedestrian's internal state based purely on motion history. In transportation dynamics, a pedestrian's future action is strongly correlated with their immediate past trajectory. The LSTM unit mitigates the vanishing gradient problem common in standard RNNs, allowing the model to learn temporal dependencies over the 15-frame observation window (approximately 0.5 seconds). This window size was specifically selected to balance the need for sufficient temporal context with the requirement for low latency inference in real time collision avoidance systems. The bidirectional nature allows the network to preserve context from both the start and end of the observed sequence, effectively smoothing noisy data inherent in frame-by-frame tracking data [6].

2.2.2 Interaction modeling: Social-LSTM

In crowded urban spaces, pedestrian velocity and movement direction are constrained by surrounding agents. This phenomenon is classically described by Helbing's Social Force Model [8], in which interactions between pedestrians are conceptualized as repulsive or attractive forces that shape collective motion behavior. The Social-LSTM architecture [9] operationalizes this theoretical concept within a deep learning framework through the use of a social pooling mechanism. For a given target pedestrian, the model identifies all neighboring pedestrians located within a spatial radius of five meters. This threshold is consistent with the interaction horizons defined in the Social-LSTM framework [9] and represents the critical perceptual zone in which pedestrians typically respond to surrounding motion dynamics in urban traffic environments. Information from pedestrians outside this radius is assumed to have a negligible influence on short-term intention prediction within a prediction horizon of 0.5 to 1.0 seconds.
The internal hidden representations of the neighboring pedestrians from the previous time step are aggregated through the social pooling mechanism to construct a representation of the local social context. This social context representation is then combined with the target pedestrian's own kinematic features and provided as input to the recurrent network. This mechanism enables the model to implicitly learn group-level behaviors, such as a pedestrian stopping or slowing down in response to the actions of a leading pedestrian, thereby capturing the collective intelligence and interaction dynamics of pedestrian crowds.

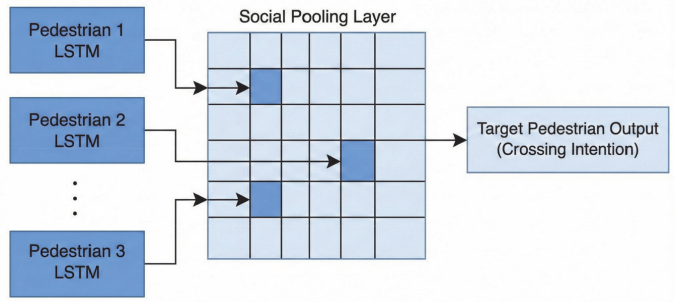

Figure 1 Schematic representation of the Social-LSTM architecture and the social pooling mechanism used to capture pedestrian-to-pedestrian interactions

2.2.3 Global context modeling: Transformer

While LSTMs process data sequentially, the Transformer architecture [11] utilizes a self-attention mechanism to process the entire temporal sequence simultaneously. This allows the model to assign different "weights" to different time steps, theoretically enabling it to focus on critical moments (e.g. the exact moment a pedestrian turns their head) regardless of when they occur in the sequence. The core operation is the Scaled Dot-Product Attention:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

However, unlike the Social-LSTM which has a built-in inductive bias for spatial interactions (neighbors affect neighbors), standard Transformers treat all inputs with equal potential relevance, requiring significantly larger datasets to learn the specific causal rules of traffic safety.

2.3 Implementation and transfer learning strategy

A critical challenge in developing safety-critical ADAS is the scarcity of high-quality "intention" labels. To address this, we employed a two-stage transfer learning protocol. Although the number of unique pedestrians is limited, the sliding window generation process resulted in a total of 539, 285 training sequences. This volume provided sufficient sample density for model convergence, which was further stabilized by the two stage transfer learning protocol. In stage 1 (Pre-training), the models were trained on the large-scale JAAD and PIE datasets to learn fundamental pedestrian motion primitives (e.g. walking, standing, starting). In stage 2 (fine-tuning), the weights were transferred and optimized on the PSI dataset, which contains more complex, socially situated crossing scenarios. This strategy mimics the engineering practice of calibrating a general traffic model to local site conditions, ensuring that the final predictions are both robust and site-specific.

3 Results and discussion

The quantitative evaluation of the deep learning architectures was conducted using the PSI dataset to assess their ability to anticipate pedestrian crossing intentions at unsignalized crosswalks. Given the safety-critical nature of the problem, the analysis focused on Precision, Recall, and F1-score metrics in order to balance the trade-off between false alarms and missed detections. Table 1 presents the comparative performance metrics of the evaluated models following the fine-tuning phase.

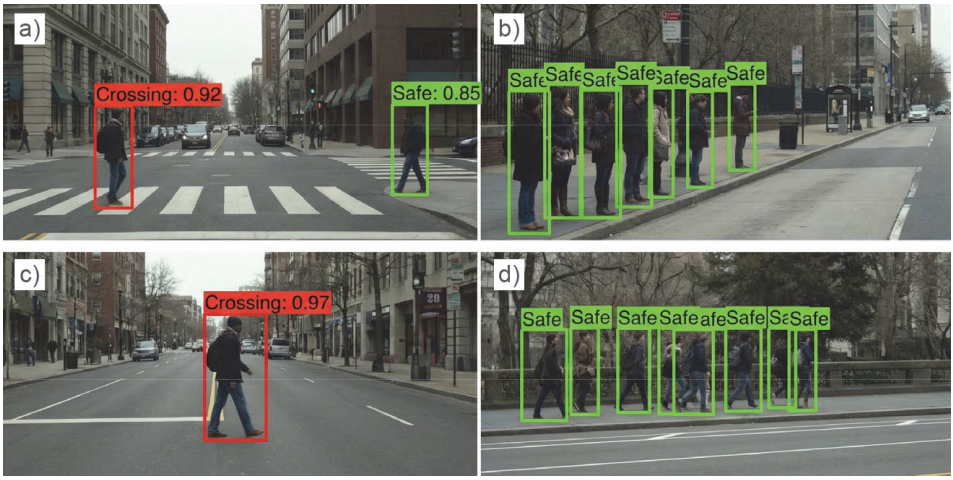

Figure 2 Qualitative prediction results, green bounding boxes indicate "Safe / Not Crossing" predictions, while red bounding boxes indicate "Crossing" intentions predicted by the Social-LSTM model

The results show that the Social-LSTM architecture achieved the highest overall reliability, with an F1-score of 84.57 percent and a precision of 85.02 percent. This superior performance demonstrates that explicitly modeling interactions between pedestrians provides a significant advantage in crowded urban environments. The findings indicate that the social pooling mechanism effectively captures group dynamics and collective decision-making behaviors described in traffic flow theory, such as the herd effect frequently observed at mid-block crossings. In contrast, the BiLSTM baseline achieved the highest recall value of 88.78 percent but exhibited a lower precision of 79.82 percent.

This outcome suggests that while kinematic features such as velocity and acceleration are highly sensitive indicators of movement initiation, they are insufficient for filtering false positives in ambiguous situations where pedestrians may remain waiting at the curb despite showing motion cues. Although high recall is desirable from a safety perspective, reduced precision may lead to driver desensitization due to frequent false warnings. The Transformer-based model achieved an F1-score of 83.44 percent, underperforming relative to the LSTM-based approaches. Despite its architectural complexity and ability to model long-range temporal dependencies, the Transformer appears to lack the domain-specific inductive biases required to generalize effectively from limited traffic safety datasets. Additionally, the impact of the training strategy was analyzed. It was found that the two-stage transfer learning approach – pre-training on JAAD and PIE before fine-tuning on PSI – Improved the model stability significantly. The analysis showed that without this transfer learning step, the models struggled to converge due to noise in the target dataset. This confirms that learning fundamental motion primitives from larger datasets is a prerequisite for accurate intention prediction in specific local environments.

Table 1 Comparative performance metrics of intention prediction models

Model Architecture	Precision [%]	Recall [%]	F1-Score [%]
Social-LSTM	85.02	84.13	84.57
BiLSTM	79.82	88.78	84.06
Transformer	77.86	86.28	83.44

Overall, the results of the study confirm that integrating social context into intention prediction algorithms significantly reduces false warnings, thereby enhancing the trustworthiness of collision warning systems in intelligent transportation networks.

4 Conclusion

In this study, a comparative analysis of deep learning architectures was conducted to address the critical challenge of pedestrian intention prediction at unsignalized crosswalks. By integrating three major datasets including JAAD, PIE, and PSI into a unified framework, the study evaluated the efficacy of kinematic-based BiLSTM, interaction-aware Social-LSTM, and attention-based Transformer models under a safety-critical lens. The results indicate that modeling the social interactions between pedestrians is fundamental for reliable prediction in urban environments. The Social-LSTM model outperformed the other architectures and achieved the highest F1-score of 84.57%. This finding supports the hypothesis that pedestrian crossing decisions are not isolated events but are significantly influenced by the collective behavior of the surrounding crowd, aligning with the Social Force theory in traffic engineering. Furthermore, the study revealed that architectural complexity does not inherently guarantee superior performance in data-constrained safety applications. The Transformer model underperformed compared to the LSTM-based approaches due to the lack of domain-specific inductive biases required to learn from limited traffic data despite its capability to capture long-range dependencies. Additionally, the effectiveness of the two-stage transfer learning strategy was confirmed, as pre-training on large-scale datasets established robust motion primitives and allowed the models to adapt to specific and high-uncertainty scenarios. However, several limitations of this study must be acknowledged. First, the dataset primarily consists of daytime scenarios with clear visibility. The performance of the models under adverse weather conditions such as heavy rain or snow and low-light nighttime environments remains to be fully validated. Second, the evaluation was conducted offline.

Real-time implementation on embedded hardware may introduce latency challenges that were not addressed in this study. For vehicle based deployment, the proposed architecture is designed to operate downstream of a real time object detector such as YOLO to extract pedestrian sequences. The relatively low parameter count of the Social-LSTM makes it suitable for standard embedded ADAS hardware like NVIDIA Jetson, where a calibrated probability threshold can be set to trigger collision warnings. Finally, the study focused on pedestrian-to-pedestrian interactions while the influence of vehicle dynamics on pedestrian decision-making was implicitly learned but not explicitly modeled as a separate interaction force. Future work will focus on addressing these limitations by fusing computer vision data with LiDAR sensors to improve robustness under adverse weather and lighting conditions. Additionally, the integration of V2X (Vehicle-to-Everything) communication technologies will be explored to enable cooperative intention sharing between connected pedestrians and autonomous vehicles, further contributing to the Vision Zero goal in intelligent transportation systems.

References

[1] World Health Organization: Global status report on road safety, Geneva, 2018.

[2] Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Agreeing to cross: How drivers and pedestrians communicate, IEEE Intelligent Vehicles Symposium (IV), pp. 264–269, 2017.

[3] Keller, C.G., Gavrila, D.M.: Will the pedestrian cross? A study on pedestrian path prediction, IEEE Transactions on Intelligent Transportation Systems, 15 (2014) 2, pp. 494–506

[4] Völz, B., Mielenz, H., Agamennoni, G., Siegwart, R.: Feature relevance estimation for learning pedestrian behavior at crosswalks, IEEE 18th International Conference on Intelligent Transportation Systems, pp. 854–860, Gran Canaria, Spain, 15-18 September 2015.

[5] Fang, Z., López, A.M.: Is the pedestrian going to cross? Answering by 2D pose estimation, IEEE Intelligent Vehicles Symposium (IV), pp. 1271–1276, Jiangsu Province, China, 2018.

[6] Saleh, K., Hossny, M., Nahavandi, S.: Intent prediction of vulnerable road users from motion trajectories using stacked LSTM network, IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 327–332, Yokohama, Japan, 16-19 October 2017.

[7] Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior, IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 206–213, Venice, Italy, 22-29 October 2017.

[8] Helbing, D., Molnar, P.: Social force model for pedestrian dynamics, Physical Review E, 51 (1995) 5, 4282

[9] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: Human trajectory prediction in crowded spaces, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–971, Las Vegas, NV, USA, 27-30 June 2016.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[11] Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting, International Conference on Pattern Recognition (ICPR), pp. 10335–10342, 2021.

[12] Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction, IEEE International Conference on Computer Vision (ICCV), pp. 6262–6271, Seoul, Korea, 27 October–2 November 2019.

[13] Rasouli, A., Kotseruba, I.: PSI: A Pedestrian Behavior Dataset for Socially Intelligent Autonomous Car, preprint, 2021.

[14] Pan, S.J., Yang, Q.: A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering, 22 (2010) 10, pp. 1345–1359

[15] Rasouli, A., Tsotsos, J.K.: Autonomous vehicles that interact with pedestrians: A survey of theory and practice, IEEE Transactions on Intelligent Transportation Systems, 21 (2020) 3, pp. 900–918

[16] Meftah, L.H., Cherif, A., Braham, R.: Improving autonomous vehicles maneuverability and collision avoidance in adverse weather conditions using generative adversarial networks, IEEE Access, 12 (2024), pp. 89679–89690