Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Low-complexity pedestrian intent prediction using contextual stacked ensemble learning

Chia-Yen Chiang

c.chiang@qmul.ac.uk

Queen Mary University of London

Yasmin Fathy

University of Cambridge

Gregory Slabaugh

Queen Mary University of London

Mona Jaber

Queen Mary University of London

Article

Keywords:

Posted Date: January 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8639560/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on September 18th, 2026. See the published version at https://doi.org/10.1038/s41598-026-70230-0.

Low-complexity pedestrian intent prediction using contextual stacked ensemble learning

Chia-Yen Chiang[1,*], **Yasmin Fathy**[2], **Gregory Slabaugh**[3], **and Mona Jaber**[1,3]

[1]The School of Electronic Engineering & Computer Science, Queen Mary University of London
[2]The Department of Engineering, University of Cambridge
[3]The Digital Environment Research Institute, Queen Mary University of London
[*]Corresponding author: c.chiang@qmul.ac.uk

ABSTRACT

Walking, as a form of active and sustainable mobility, plays a critical role in future smart transportation systems. Accurate prediction of pedestrian crossing intentions is essential for preventing collisions, particularly with the increasing deployment of autonomous vehicles. Existing approaches to near-miss prevention typically rely on computationally intensive computer vision and deep learning techniques. In contrast, this work proposes *CSE*, a lightweight contextual stacked ensemble-learning framework to efficiently predict pedestrian crossing intent. Pedestrians are first detected and their visual representation is compressed through skeletonization, and complementary pose, trajectory, and contextual cues are fused using a stacked ensemble model. Experimental results on multiple datasets demonstrate that the proposed approach achieves performance comparable to state-of-the-art pedestrian intent prediction methods while reducing computational complexity by at least 25 times. This reduction translates directly into a $25\times$ decrease in inference time, enabling deployment on resource-constrained edge devices without compromising accuracy and while avoiding the latency associated with cloud-based processing.

Introduction

Active travel, a mode of transportation in which people move from one place to another through physical activity such as walking or cycling, is essential for sustainable and efficient urban transportation systems. However, growing safety concerns for vulnerable road users (VRUs), particularly pedestrians, have hindered its widespread adoption. The latest data from the 8th UN Global Road Safety Week indicates that nearly 1.2 million people are killed and up to 50 million people are injured each year due to road accidents. Globally, more than one of every four deaths occurs among pedestrians or cyclists[1]. The future vision of smart and connected urban mobility involves multiple modes of transportation that share urban spaces and road networks, including VRUs and connected and autonomous vehicles within the Vehicle-to-Everything (V2X) paradigm[2]. Human drivers rely on implicit contextual reasoning to interpret scenes, anticipate VRU behavior, and communicate safety through cues such as eye contact when crossing[3]. Can autonomous vehicles be trained to develop a similar form of contextual reasoning to predict VRU behavior? Pedestrian intent prediction (PIP) has recently gained significant attention, driven by rapid advances in autonomous vehicles and smart mobility infrastructures.

Existing PIP methods often rely on technologies used in connected and autonomous vehicles and employ a range of sensors, including radar, LiDAR, optical cameras, and ultrasonic sensors. Data from these sensors is processed using machine learning (ML) techniques to automate decision-making and reduce collision risk[4]. Recent advances in computer vision have enhanced pedestrian detection and intent prediction accuracy. However, camera-based data are often complex and computationally intensive to process. In addition, such data can capture personal information from pedestrians, raising privacy concerns despite being largely irrelevant to the PIP task. Computer vision-enabled PIP generally follows two approaches. The first relies on roadside cameras with cloud-based processing, which requires reliable connectivity and raises privacy concerns when personal data are transmitted[5]. Edge computing mitigates these issues, but is constrained by limited computational power[6,7]. The second approach relies on vehicle-mounted cameras and sensors connected to the advanced driver-assistance systems (ADAS) and to other vehicles and objects within the V2X paradigm. With recent advances in 5G mobile networks and future 6G technologies, V2X could benefit from broadband connectivity for moving vehicles. However, this connectivity tends to be less reliable and more costly compared to the connectivity of fixed cameras[8]. Moreover, the robustness of the sensors and of the data fusion and artificial intelligence (AI) algorithms varies significantly between different vendors.

Across both paradigms, a common challenge is the substantial computational overhead of existing PIP models. Many state-of-the-art approaches rely on complex deep neural architectures, leading to high inference latency and energy consumption. These factors limit their suitability for deployment on resource-constrained edge devices and raise broader concerns regarding the sustainability of intelligent transportation systems. As a result, there is increasing interest in developing efficient, low-

complexity PIP models that balance predictive performance with computational and energy efficiency[9].

To address these challenges, this work proposes a low-complexity pedestrian intent prediction called Contextual Stacked Ensemble-learning (CSE), designed for data captured by front car-mounted cameras. CSE addresses the limitations found in the prior art by offering a lightweight PIP solution that is compatible with resource-constrained edge servers. Similar to previous studies[10,11], we employ a skeletonization step, which compresses each pedestrian's image into a compact set of 17 keypoints, thus reducing the volume of camera-based data by $4,336\times$. The same pedestrian image is mapped to location information relative to the ego-vehicle; this information is found to represent pedestrians' intent while being less than $10,000\times$ the size of images. As a result, CSE is an effective and resource-compliant edge-based solution for PIP that avoids high latency, privacy breaches, and reliance on the connectivity of cloud computing. A simplified version of the CSE approach was used for a PIP competition organized by the IEEE Intelligent Transportation Systems Society[12] and won the first prize, demonstrating its effectiveness beyond the scope of the research presented here.

The source code and trained models will be released upon acceptance of the paper.

Related work

Early research on pedestrian analysis largely focused on detection and tracking tasks[13]. Pedestrian intent prediction (PIP), however, requires forecasting future behavior, making it inherently more challenging due to the need to interpret subtle motion, pose, and contextual cues. To address this, recent PIP methods have employed pose-based methods using Graph Neural Networks (GNNs) with temporal modules[10,11] to model keypoints' dynamics and reduce input dimensionality, though many still depend on appearance or scene features, and therefore, increase computational cost. Other methods leverage recurrent or hybrid Convolutional Neural Networks (CNNs) and Long-Short-Term Memory (LSTM) architectures[14–16], and transformer-based approaches fuse multimodal cues for state-of-the-art performance[17–19], but these models remain computationally resource-intensive. Despite their accuracy, such models typically require processing high-dimensional visual inputs and rely on deep, parameter-heavy architectures, limiting their suitability for real-time or edge-based deployment.

Lightweight and Complexity Reduction Approaches to PIP

Compared with deep-learning-based methods, research explicitly targeting lightweight or computationally efficient PIP remains limited. One notable effort is the interpretable framework proposed by[20], which uses LIME and SHAP to identify influential features for binary crossing-intent classification. The study compares several traditional machine learning algorithms and reports strong performance from an XGBoost model. However, this work focuses primarily on feature interpretability rather than computational efficiency: it does not benchmark against established deep-learning PIP baselines, does not evaluate latency or resource usage, and employs a unique data-splitting strategy (âĂIJindividualâĂİ vs. âĂIJgroupâĂİ pedestrians) that complicates direct comparison with the existing PIP literature. Moreover, because these models operate on limited handcrafted features and do not capture temporal dynamics, the approach is not readily extendable to broader PIP settings where sequential motion cues are critical.

Model compression techniques, such as pruning, quantization, and knowledge distillation have been widely explored in related vision tasks to reduce computational cost. Recent studies have applied pruning to pedestrian image classification[21], dense crowd analysis[22], and traffic event detection[23], achieving substantial reductions in model size and inference time. However, such techniques have seen little application in PIP, despite its similar need for efficient on-edge deployment. Current PIP pipelines typically depend on high-dimensional visual inputs and sequential modeling, yet no prior work has systematically investigated compression strategies or assessed the accuracyâĂŞefficiency trade-offs under real-time constraints.

Ensemble learning for PIP related topics

Ensemble learning offers an alternative strategy for balancing robustness and efficiency by combining multiple lightweight base learners. In the context of pedestrian behavior analysis, ensemble methods have been explored primarily in fixed-camera or infrastructure-based settings. For example, a roadside cameraâĂŞbased system was proposed in[24], which predicts crossing intent for two predefined crossing options. The authors integrate multiple base learners into an ensemble that improves overall performance and demonstrate that incorporating human pose information provides a substantial accuracy gain. However, the method is constrained to a fixed junction and predefined crossing zones, limiting its applicability to moving-camera scenarios such as dashcam-based perception, where scene geometry and viewpoints vary continuously. Similarly, a stacking-based ensemble framework using LiDAR and high definition camera data was employed by[25] to classify pedestrian intent into three categories: walkingâĂŞwalking, stoppingâĂŞwalking, and walkingâĂŞstopping. The study further conducted statistical analysis to identify influential features and showed that the stacking framework outperformed individual learners. Yet, this approach also relies on a fixed-site sensing setup, making it unsuitable for dynamic urban driving environments captured by vehicle-mounted sensors. Although not directly predicting crossing intent,[26] developed an ensemble-based action recognition framework to classify pedestrian behavior (walk, run, stand). The approach consistently improved performance over its base

models on the same dataset used in our study. However, action recognition is a less challenging task than PIP, as it focuses on categorizing current actions rather than inferring future crossing decisions.

To address these limitations, we introduce CSE, a lightweight ensemble framework designed specifically for mobile, camera-based PIP. Unlike prior ensemble systems tied to fixed viewpoints from road side cameras, CSE leverages complementary feature streamsâĂŤsuch as pose dynamics, ego-vehicle motion, and scene context, and combines their predictions through a context-aware meta-classifier. This design enables the ensemble to adapt to varying viewpoints and environmental conditions while remaining computationally efficient. By decoupling feature extraction from the final prediction and using simple base learners, CSE achieves a balance between accuracy, robustness, and real-time feasibility, making it well-suited for deployment in on-vehicle PIP applications.

Research Methodology

In this work, PIP is formulated as a binary classification problem that aims to predict whether a pedestrian intends to cross the road (*Crossing*) or not (*Not-Crossing*), based on video clips captured by a front-mounted vehicle camera. Let $\mathbb{S}_v$, $v \in \{1,\ldots,V\}$ denote the vth video clip, where V is the total number of clip. Each video clip $\mathbb{S}_v$ consists of M_v consecutive frames denoted as F_i, $i \in \{1,\ldots,M_v\}$. Each clip contains a maximum of P_v bounding boxes (*bbx*), and each *bbx* bounds a single pedestrian with an index $j \in \{1,\ldots,P_v\}$ and may include other pedestrians in the background (i.e., not labeled). Video clips differ in length M_v and number of annotated pedestrians P_v. Thus, each clip includes the labeled behavior of P_v pedestrians with a unique index $j \in \{1,\ldots,P_v\}$. To model pedestrian motion over time, each video clip is partitioned into fixed-length, non-overlapping temporal segments using a sliding window of width $F_{max} = 32$ frames. Each segment is indexed by $s \in \{1,\ldots,M_v/F_{max}\}$ and describes the movement of a pedestrian with index j in a video clip with index v. Each pedestrian j in clip $\mathbb{S}_v$ is annotated with a label y; $y = 1$ *Crossing* if the pedestrian intends to cross a street at any point in $\mathbb{S}_v$ or $y = 0$ for (*Not-Crossing* or *Irrelevant*, i.e., without a behavioral label), otherwise. Pedestrians annotated as *Irrelevant* are treated as negative samples and merged with the *Not-Crossing* class, resulting in a binary classification setting. This clip-level label is inherited by all temporal segments s associated with pedestrian j.

The PIP problem can thus be formulated as predicting the label $\hat{y}_{v,j,s}$ associated with each segment of a pedestrian bounded by *bbx* with index j, in sliding window with index s of video clip with index v. The objective is to maximize the number of correctly predicted labels $\hat{y}_{v,j,s}$ for all video clips, all annotated pedestrians in each clip, and all segments of F_{max} frames, as formulated here:

$$\max \quad \sum_{v=1}^{V} \sum_{j=1}^{P_v} \sum_{s=1}^{\frac{M_v}{F_{max}}} (\hat{y}_{v,j,s} = y_{v,j,s}) \tag{1}$$

Towards solving the problem in (1), we extract features that characterize each pedestrian segment from two complementary perspectives. The first captures pedestrian motion by analyzing the evolution of the bounding box and its position relative to the image frame, yielding pose-based keypoints and trajectory features. The second captures contextual information at both global and local levels. Global context, shared across all pedestrians in a frame, includes ego-vehicle speed, traffic light status, and road type, while local context encodes pedestrian-specific cues such as whether the pedestrian is positioned at an intersection. These features are subsequently used to classify each segment as either *Crossing* or *Not-Crossing*.

Feature Extraction

To reduce model complexity and avoid reliance on high-dimensional visual representations, we propose a context-aware feature representation that extracts the cues relating to PIP from the sequence of frames in a video footage. The selected features are designed to be compact, complementary, and computationally efficient. The first relates to the gradual change of a pedestrian's skeleton, represented as a connected graph of keypoints; this is the input to Model M1. The second consists of the dynamic changes in the semantic information describing the surroundings of the pedestrian as they move during the video clip; this is the input of Model M2. The third is the varying position of the pedestrian with respect to the ego-vehicle; this is the input to Model M3. Together, these representations capture complementary aspects of pedestrian intent, including pose dynamics, contextual cues, and relative motion, without requiring full image-based processing. The models are described in the subsequent Model Overview section. The influence of each feature type on PIP accuracy is examined in the Results and Discussion Section.

Skeletonization

The first data representation aims to capture the pose of a pedestrian by examining the essential keypoints of the human body following a skeletonization approach. The reasoning behind this representation is multifaceted. First, it reduces input dimensionality while preserving motion- and pose-related cues that are relevant to pedestrian intent prediction. Second, it avoids the capturing of personal information (e.g., face, hair, clothing, etc.) while retaining the essential information for pose

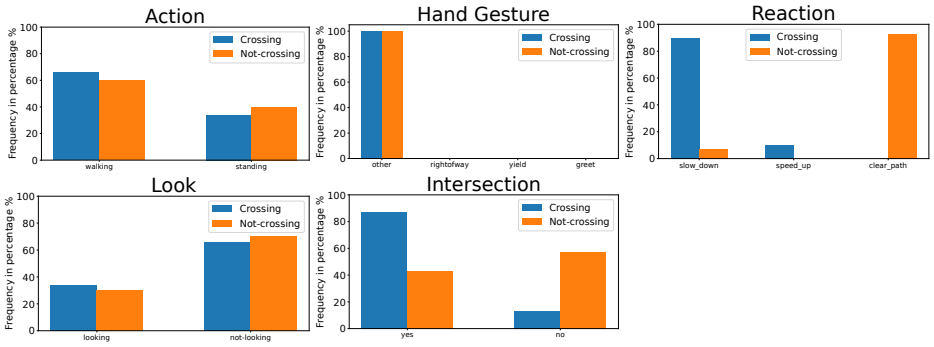

Figure 1. Pedestrian Behavior comparison between *Crossing* and *Not-Crossing* before crossing action. Intersection and Reaction have the most distinguishable patterns in subcategories for both classes.

analysis and, therefore, action prediction. In this work, we first apply the MMpose[27] algorithm which returns 17 key points. Next, we combine the keypoints representing the left eye, right eye, and nose into a single keypoint that represents the head. Similarly, we replace the keypoints representing the left and right shoulders with a single point in the middle that represents the thorax. Thus our M1 model is based on 14 keypoints instead of 17. Skeletons are modelled as connected graphs where the vertices are the K_{max} keypoints representing a target pedestrian with index $j \in \{1,\ldots,P_v\}$ bounded by a *bbx* in the sliding window with index $s \in \{1,\ldots,M_v/F_{max}\}$ of clip v and maximum number of indexed pedestrians P_v. Each graph encodes the spatial configuration of the pedestrianâĂŹs body at a given time step. For simplicity, in the description of M1, we drop the indices related to the video clip and to the pedestrian and focus on the index that tracks the movement of the target pedestrian, i.e., the index of the frame. For notational simplicity, we further omit the video and pedestrian indices and retain only the sliding-window index s, since all skeletonization representations are at the segment level. For a given target pedestrian bounded by a *bbx*, each vertex (i.e., keypoint) in each frame F_i within the sliding window is depicted by a matrix $\mathscr{K}_i$, which is obtained through the skeletonization process. In this context, $k_\kappa = \{w_\kappa, h_\kappa\}$ denotes the relative position of a keypoint with index κ, and $\mathbb{K}_s$ represents the sequence of F_{max} matrices $\mathscr{K}_i$ within a sliding window, as detailed below:

$$\mathscr{K}_i = [k_1,\ldots,k_\kappa], \kappa = \{1,\ldots K_{max}\} \tag{2}$$

$$k_\kappa = \{w_\kappa, h_\kappa\} \mid 0 < w_\kappa < W; 0 < h_\kappa < H \tag{3}$$

$$\mathbb{K}_s = \{\mathscr{K}_i : 1 \leq i \leq F_{max}\} \tag{4}$$

These keypoints (vertices) are connected with a fixed adjacency matrix $\mathscr{A}$ defined to represent the connectivity between any two different keypoints k_o and k_q in a skeleton, such that $A(o,q) = 1$ when these are connected (e.g. head and neck)[28].

$$\begin{aligned}\mathscr{A} = [&\{A(1,1),\ldots,A(1,K_{max})\},\ldots,\\ &\{A(\kappa,1),\ldots,A(\kappa,K_{max})\},\ldots,\\ &\{A(K_{max},1),\ldots,A(K_{max},K_{max})\}]\end{aligned} \tag{5}$$

Contextual data

V2X connects vehicles and road sensors with location-based services, thus, providing a wealth of information to solve the PIP problem. This section explores different contextual data included in the JAAD annotation, including three categories of attributes:

- Pedestrian Behavior shown in Figure 1. It includes **Action**={*Walking, Standing*}, **Hand Gesture**={*Greet, Yield, Rightofway, other*}, **Reaction**={*Slow-down, Speed-up, Clear-path*}, Look={*Yes, No*}, and **Intersection**={*Yes, No*}. All except **Intersection** are time-series data.

- Traffic Background shown in Figure 2 and includes **Crosswalk** (or Pedestrian Crossing)={*Yes, No*}, **Traffic Light**={*Green, Red*}, and **Road Type**={*Garage, Parking lot, Street*}). All except **Road Type** are time-series data.

- Vehicle speed before pedestrian's crossing action shown in Figure 2 (d).

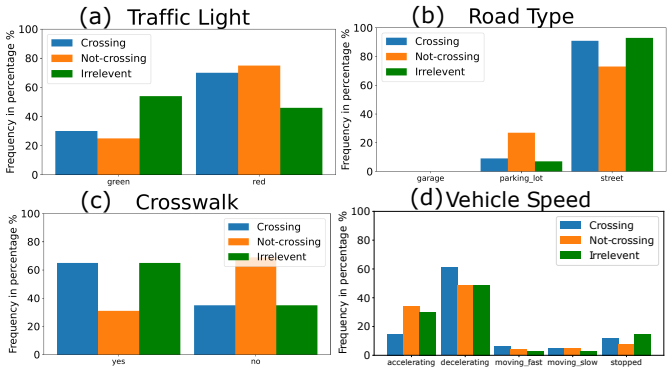

Figure 2. Traffic and Vehicle Speed Annotation comparison between *Crossing*, *Not-Crossing* and *Irrelevant*. It is clear that *Irrelevant* has similar patterns with *Crossing* in Road type (b) and Crosswalk (c), whereas, Traffic Light (a) cannot differentiate *Crossing* and *Not-crossing*. In (d), vehicles tend to accelerate in *Not-Crossing* and *Irrelevant* but decelerate in *Crossing*.

Pedestrians that are not labeled as *Crossing* or *Not-Crossing* have been marked as *Irrelevant* in Figures 1 and 2. Overall, some attributes play a vital role in predicting labels that signify a pedestrian's intention to either *Crossing* or *Not-Crossing*. In the following section, we explore how these features contribute to label prediction.

It may be counter-intuitive at first that some of the **Pedestrian Behavior** attributes shown in Figure 1 do not have a distinct pattern for *Crossing* or *Not-Crossing*. This is particularly evident in the case of pedestrians' intentions, where those who are looking and those who are walking show a near-equal likelihood to *Crossing* or *Not-Crossing*. However, careful analysis indicates that pedestrians who *stand* are more likely to intend to cross, and those who *look* may be standing-by. *Reaction* shows that *Crossing* pedestrians are more likely to slow down while *Not-Crossing* are more likely to clear path. *Intersection* shows that *Crossing* pedestrians are likely to stand on the intersection.

While examining the **Traffic Annotation** attributes, it seems *road type* and *crosswalk* are similar for both the *Crossing* and *Irrelevant* labels, which indicates that they carry weak representation of pedestrian intent. Moreover, pedestrians labeled as *Irrelevant* are more likely to be seen in the traffic scene when the traffic light is green than the other two classes. Following a careful investigation of such occurrences, we noted in most cases, pedestrians are not in the same lane as the ego-vehicle. Instead, they could be on the zebra crossing parallel to the ego-vehicle's lane or on the sidewalks.

Further examining the **Traffic Annotation** attributes in Figure 2 (a), we can see *Crossing* pedestrians are more likely to appear in the traffic scene with *Crosswalk* (Pedestrian Crossing) and on the *Street* rather than in *Garage*, comparing to *Not-Crossing*. As for **Vehicle Speed** attribute in Figure 2 (d), it is intuitive to note that vehicles *Decelerate* at the sight of pedestrians irrelevant of what their PIP. However, they are more likely to *Decelerate* in front of *Crossing* pedestrians and likely to *Stop* when facing *Crossing* and *Irrelevant* pedestrians.

Pedestrian trajectory

The trajectory of the pedestrian within each given sliding window carries contextual information that can be useful in solving the PIP problem. Rather than relying on full image content, pedestrian motion is represented using compact, geometry-based cues derived from bounding box information. The location of the pedestrian is equated to the relative position of *bbx* with respect to the frame. In this work, we calculate the relative position of the central *bbx* pixel located at $[W_c = W/2, H_c = H/2]$, where W and H are the width and height of the *bbx*. The pedestrian's depth, Z, is approximated from the size of the *bbx* where a small value indicates that the *bbx* is far away from the camera and vice-versa. While this depth estimate is coarse, it provides a lightweight proxy for relative distance without requiring explicit depth sensors or camera calibration. Knowing that the absolute coordinates (W_c, H_c) and the size Z can be biased by different frame sizes, we normalize the values such that the pedestrian normalized location is $(W_c/W_F, H_c/H_F)$ and the normalized size is $Z/(H_F \times W_F)$, where H_F is the height of the frame and W_F is the width. In addition, we calculate the change of normalized size over time $\Delta Z/(H_F \times W_F)$ as well as the speed of movement, which was estimated by getting the ratio of the normalized traveled distance over time.

For each sliding window with index s, the normalized positioning vector $q_{s,i} = [W_{ci}/W_F, H_{ci}/H_F, Z_i/(H_F \times W_F), \Delta Z/(H_F \times W_F), D/t]$ of each *bbx* containing a pedestrian in *bbx* F_i $(i = [0, \ldots, F_{max}])$ is calculated. Thus, the trajectory of a pedestrian in sliding window with index with index s is represented by $\tau_s = \{q_{s,0}, \ldots, q_{s,F_{max}}\}$.

Model Overview

Different methods have been proposed to address PIP in[11,28–30] however, many of these methods rely on computationally intensive architectures. Motivated by recent work on low-complexity intent prediction frameworks for resource-constrained deployment[31], we propose CSE, a novel ensemble learning system that builds on three individual models which we introduce in this section. The first model, M1, consists of a graph neural network (GNN) model for classifying skeletonized data as *Crossing/Not-Crossing*. The second model, M2, employs stacked gated recurrent network units (GRUs) for solving PIP using time-series and categorical data from JAAD's vehicle and traffic annotations. The third model, M3, employs a one-dimensional convolutional neural network (1D-CNN) for solving PIP based on a pedestrian's trajectory. The proposed CSE framework combines these two or three models as shown in Figure 4 (right). Each of models M1, M2, M3, and CSE model is structured to represent the PIP of a target pedestrian during a sliding window of F_{max} frames. For comparison purposes, two models are also implemented using a similar structure. The overall ensemble learning system is thus structured as follows.

M1: Graph Neural Network

The GNN model (M1) is used to classify a sequence of graphs $\mathscr{K}_i$ that represent a segment of index i of a video clip (index of video clip is dropped here for simplicity) as *Crossing/Not-Crossing*. The model consists of a Convolutional Graph Neural Network layer (GCN layer), a Relu+Dropout, a skip connection, a Flatten operation and a GRU, as shown in Figure 3. The input keypoints are first multiplied by a normalized adjacency matrix, which aggregates features from adjacent nodes in the skeleton graph. The aggregated node features are embedded into a convolutional operation in the GCN layer to extract higher-level spatial representations. To prevent shrinking feature maps, the convolutional operation must use the "same" padding. Relu and Dropout (0.5) layers are then used to encourage model generalization.

The skip connection operation adds input keypoints to the dropout output. By preserving early-layer information and facilitating gradient flow, the model can keep both low and high level representations, thus improves feature learning. A 1x1 convolution on the skip connection branch adjusts the channel dimension to enable the addition to the features output by the GCN layer.

Finally, after flattening the merged features from the dropout layer and skip connection, a GRU layer is used to model the evolution of keypoint features in the time domain. The hidden size is set to 16. The last hidden output is then fed to a fully-connected layer for *Crossing/Not-Crossing* classification.

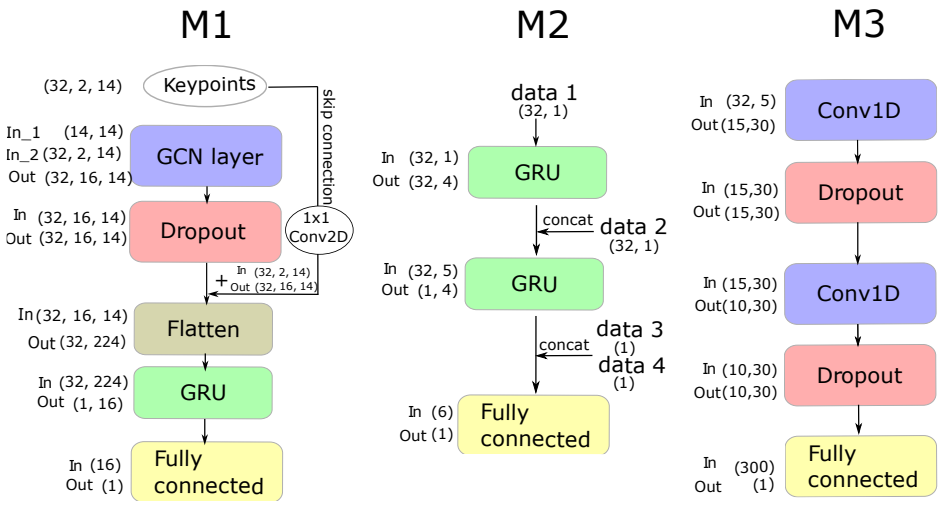

Figure 3. M1, M2, M3: The figure shows the submodel structure with input and output dimensions in each model block. Note that the shape of Keypoints at the input of M1 is determined by $F_{max} = 32$, 2D coordinates, and $K_{max} = 14$.

M2: Stacked Gated Recurrent Unit Network

Inspired by the stacked GRU in[32], we apply the stacked GRU in modelling time-series and categorical contextual information from the vehicle and traffic background as shown in Figure 3. The stacked GRUs model (M2) iterates each time-series data and feeds it to a GRU with only one layer to produce all hidden outputs from frame F_0 to F_{max}. Before feeding all hidden outputs to

the next GRU, a next time-series data is merged with the hidden outputs. The output of the last GRU is the last hidden output instead of a full sequence of hidden outputs.

Specifically, the categorical data used for contextual information in M2 consists of the following data (Figure 3):

- Data 1 (time-series data over the duration of F_{max} frames) - vehicle speed {*Stopped, Slow, Fast, Accelerating, Decelerating*}: A pedestrian is less likely to cross if they notice an upcoming vehicle that is speeding.
- Data 2 (time-series data over the duration of F_{max} frames) - Traffic signal status {*Red, Green*}: A pedestrian is less likely to cross the street if the traffic signal controlling the passing vehicles is green; in which case the pedestrian signal would be red.
- Data 3 (categorical with one value for each *bbx*)- Is the pedestrian standing at an intersection? {*Yes, No*}: A pedestrian waiting at an intersection is likely to cross the street.
- Data 4 (categorical with one value for each video)- Type of road {*Garage, Parking lot, Street*}: It is more likely that a pedestrian seen on the street might cross than a pedestrian standing/walking in a parking lot.

Our proposed approach is different from[32] in the following way. The last hidden output of stacked GRUs in the proposed M2 is concatenated with categorical data. The combined features then are projected to a fully connected layer before *Crossing/Not-Crossing* classification (see Figure 3).

M3: 1D convolutional Neural Network

For modelling the relationship between a pedestrian's trajectory within a segment with index s, τ_s, and PIP decision, a shallow 1D-CNN (M3) is proposed as a lightweight temporal feature extractor. (Note: the video clip index and pedestrian index are dropped here for simplicity). 1D-CNN is often used to predict a future trajectory given historical data[33,34] while[35] uses both RNN and CNN to predict future classes of an encoded trajectory. Trajectories possess not only pedestrian speed, direction, and pedestrian depth information but are also strongly related to PIP decisions. We therefore adopt a 1D-CNN for M3 to solve PIP based on past trajectory as contextual information. Our proposed 1D-CNN model contains only three layers: a convolutional layer for extracting 1D time-series features, a dropout layer for preventing overfitting, and a fully connected layer for PIP classification, as shown in Figure 3.

CSE: Contextual Stacked Ensemble

Similar to the approach used in[36] for multimodal feature fusion, we propose a stacked ensemble learning method to fuse features from two or three models defined above. Each of the base models M1, M2, and M3 is trained by a five-fold stratified cross-validation approach with a fixed sampling seed, ensuring each fold preserves the same class ratio. The last fully connected layer of each of the three models M1, M2, and M3 produces an output S1, S2, or S3, respectively. Using out-of-fold predictions prevents information leakage between training and ensemble stages. The outputs S1, S2, S3 (or a subset) are concatenated to form the aggregated features that are then fed into a CSE classifier. This meta-classifier learns to combine complementary decision cues from the individual models, thereby improving robustness and overall prediction performance. A sigmoid and argmax is used to generate the final *Crossing/Not-Crossing* classification results (see CSE model structure in Figure 4).

Results and Discussion

We conducted the experiments of our proposed method on two different labeled versions of the Joint Attention in Autonomous Driving (JAAD) dataset[37]: JAADbeh and JAADall. We further validate the method using the PSI dataset[38] to prove the generalizability of the model.

Dataset description

We first use the JAAD dataset which is widely used for PIP evaluation. The dataset is richly annotated and contains $V = 346$ video clips (each is around five to ten seconds long) extracted from over 240 hours of driving footage from dash cameras positioned at the front of the car. The dataset provides frame-wise, distinct labels including pedestrian reaction toward drivers, pedestrians' demographic attributes, and traffic conditions, which are not found in other datasets, such as the Multiple Object Tracking (MOT) Challenge[39]. There are two versions of the same dataset, JAADbeh and JAADall, where JAADbeh is a subset of JAADall, and the latter includes additional annotations of pedestrians in the background. JAADall dataset incorporates all non-labeled pedestrians in JAADbeh by manually annotating these to *Not-Crossing* class, the same as other comparable state-of-the-art works[16,40].

We address the problem of PIP for each given input sliding window sampled from each video clip $\mathbb{S}_v$. The predicted label $\hat{y}$ is a binary variable, where $\hat{y} = 1$ for *Crossing* if a pedestrian is predicted to intend to cross a street or $\hat{y} = 0$ for (*Not-Crossing*).

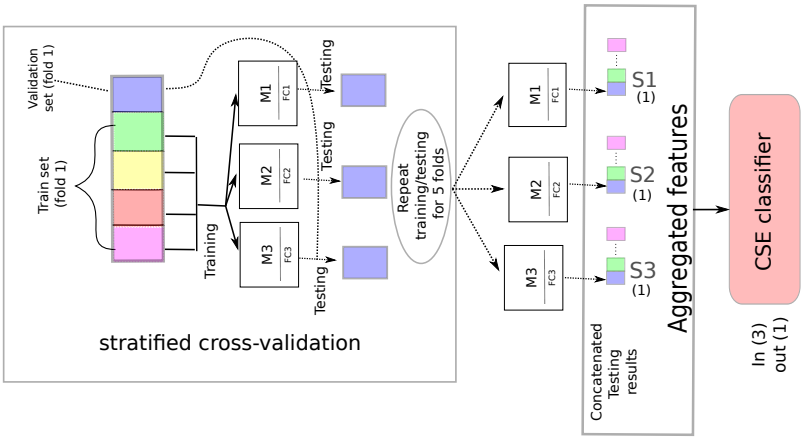

Figure 4. CSE: Model structure of Contextual Stacked Ensemble learning. In stratified cross-validation, we first train models with five-fold training datasets. An example is shown in Fold 1, where three models are trained on four blocks (see the bottom-to-top 4 blocks marked train set Fold 1 above). In this case, the validation block (top purple) is used to test each of the three models. This process is repeated with the iteration of folds; for instance, the next fold would use the green block for validation and train using the remaining blocks. The testing results of all five folds are concatenated to form S1, S2, and S3 (or subet) as an input of the CSE classifier.

In preparation for the model training and testing, we split the dataset into a training, validation, and test subsets based on JAAD's suggestion[41]. The JAAD's train and validation were then merged before splitting again for stratified 5-fold cross-validation[42]. For each fold, the method preserves the percentage of *Crossing* and *Not-Crossing* data points instead of random sampling on the positive ($y = 1$, *Crossing*) and negative ($y = 0$, *Not-Crossing*, *Irrelevant*) samples pool. This approach ensures that the training performance of each fold is not biased by a random ratio of classes and preserves all information in the samples. The stratified 5-fold crossing validation was used only in training models on JAADbeh. This method preserves the class ratio and therefore, ensures the training stability, which works well for JAADbeh. However, with an extreme class distribution, such as JAADall, we find that models perform better when classes are balanced. Therefore, we create a customized 5-fold dataset in which each fold has all positive samples from the merged data and randomly selected negative samples (without replacement) from the merged data to match the number of positive samples. Nonetheless, the test results are derived using[41] for both JAADbeh and JAADall in order to enable fair comparison with prior art.

Each of the three data streams is used independently to train three different models, M1, M2, and M3. The CSE approach shown in Figure 4 is applied to each combination of two models (2-model-based CSE) and to the three models (3-model-based CSE). We first present the results of each independent model in addition to the different CSE options in Table 1 and compare the performance and complexity of these to the state-of-the-art models.

As seen in Table 1, our proposed CSE using M1, M2, and M3 delivers the best F1 score, outperforming the next best models, Pedgraph+[11] and Faster PCPNet[10], by 3% at a much lower computational complexity. Our model requires 331 times less FLOPS than Pedgraph+ and 25 times less than Faster-PCPNet; this directly translates to faster inference time, as indicated. Moreover, our model consists of 5.2 times less trainable parameters than Pedgraph+ and 2.2 times less than Faster-PCPNet. While the table lists the performance of other models[16,40], their performance is under-par while being three orders of magnitude more complex that the proposed CSE model.

Limitations and promising directions

The results presented in Table 1 clearly support the claim that the proposed CSE approach is orders of magnitude less computationally expensive than prior art. Moreover, CSE with JAADall outperforms the best next model[10] by 0.03 points in F1-score, thus demonstrating its superiority in balancing between both *Crossing* and *Not-Crossing* predictions with 25 times faster inference. Nonetheless, the model's performance with JAADbeh is under-par compared to the other two baselines[16,43]. Indeed, the closest performance in this case is achieved with M2 alone as the added features obtained with M1 and M3 seem to mislead the CSE model. There are two interpretations to this limitation. The first is related to the inconsistency in class distribution between the train and test dataset in JAADbeh: 56 : 44 for training but 4 : 6 for testing. The model is thus trained and tested on opposite skewed-ness in the datasets which is not the best practice in data science. We still opted to use the standard test dataset[41] in this case to enable the comparison with the prior art and note that the next best model[10] is not tested on JAADbeh, possibly due to the same issue. We posit that using JAADall addresses this limitation as the training dataset is

Model \Dataset	*JAADbeh*					*JAADall*					Complexity		Inference
	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	FLOPs(K)	Params(K)	Time (ms)
M1	58	65	58	61	61	81	85	62	60	78	1.54×10^3	12.5	0.308
M2	66	**76**	66	72	70	64	64	48	53	62	6.92	0.22	0.0014
M3	61	69	61	65	64	68	79	52	56	72	24.6	0.7	0.005
CSE(M1+M2)	59	67	58	62	62	82	84	63	61	78	1.55×10^3	12.72	0.31
CSE(M1+M3)	60	70	59	65	63	83	87	63	61	79	1.56×10^3	13.2	0.312
CSE(M2+M3)	65	72	65	70	69	75	81	57	57	74	31.52	0.92	0.0064
CSE(M1+M2+M3)	60	70	60	64	63	88	87	**68**	65	78	$\mathbf{1.57 \times 10^3}$	**13.42**	**0.314**
Global PCPA[40]	-	-	-	-	-	83	86	63	51	**82**	154.3×10^6	60.9×10^3	104
Pedgraph+[11]	**70**	70	**76**	**77**	75	86	**88**	65	58	75	520×10^3	70.3	30,800
Faster-PCPNet[10]	-	-	-	-	-	**89**	77	65	**73**	58	40×10^3	30	8
Group-CNNLSTM[16]	61	55	74	72	**76**	82	81	62	52	78	300×10^3	3×10^3	-

Table 1. Performance and complexity of proposed models using JAADbeh and JAADall. Note: Values shown for Faster-PCPNet, Global PCPA, PedGraph+, and Group-CNNLSTM are not reproduced; these are copied from[10,11,16,40] instead. The best performance is shown in **bold** and the next best performance is underlined.

equally balanced and significantly larger than JAADbeh. Indeed, for JAADbeh the number of pedestrians with labeled intention is $P_v = 464$ while JAADall has $P_v = 2564$. When applied to JAADall, the proposed CSE model outperforms all prior models with an increase of 4.6% in F1-score. The second interpretation relates to the high correlation between M2 and M3's false predictions; we suspect that in such a situation, when both models are merged (CSE(M2+M3)), the performance is likely to get worse compared to pre-merged performances. (Note: we compare M2 to CSE(M2+M3) as the former is our best performing and the latter is the second best.)

Our proposed approach relies on M2 which learns the intent of pedestrians based on features that are particular to the JAAD dataset, i.e., Pedestrian Behavior, Traffic Background, and Vehicle Speed. It follows that the model's applicability may be deemed as conditional on such annotations which are not widely available. We argue that a main strength in the proposed approach is its agility to adapt to available data; indeed, our method using CSE(M1+M3) and excluding M2 secured the first prize in the IEEE Intelligent Transportation Systems Society student competition in Pedestrian Behavior that used the PSI dataset[38]. We trained our model based on a combination of JAAD and PSI data and the winning test results were obtained using a test dataset later released by the competition organizers. These results demonstrate the generalizeability of the CSE approach with or without M2-based features.

Since the main claim of this work is the low computational complexity of the proposed method, it is important to explore pruned versions of prior art to offer a fair comparison. It is not feasible to prune complex deep learning models such as those listed in Table 1 without full access to the architecture of the models and an understanding of the purpose of each of their layers. Instead, we adopt an empirical approach based on[44] which shows that dropping the last few fully connected layers can achieve the desired pruning effect. Furthermore, we assume the best-case scenario (which is unlikely) that the models' performance would not deteriorate due to this pruning. When applied to[40], this approach results in a negligible 0.013% reduction in trainable parameters, while for[43] a change in hyperparameters might halve the number of trainable parameters. Given the complexity of[10] it is not possible to estimate the effect of pruning on the model's computational cost; assuming it can be halved, it would still be higher than the proposed CSE with lower performance. The same empirical approach does not apply to the model in[16] where the only fully connected layer is used for classification. Nevertheless, we note that the proposed CSE method requires 191× fewer FLOPs and 223× fewer trainable parameters than[16]; this difference, exceeding two orders of magnitude, cannot reasonably be reduced through pruning alone.

Despite these limitations, this work presents a promising method for predicting the intent of pedestrians that leverages contextual information and three representative streams for feature extraction. There are various ways we intend to improve on this method which include joint training of these streams (including combinations with two models) to avoid the pitfalls faced with JAADbeh where both models learn the same mistakes. To further analyse the generalizeability of the model, we propose expanding the study to other datasets including PIE[45] and STIP[46].

Conclusion

This work examines the critical problem of pedestrian intent prediction (PIP), which raises safety concerns for vulnerable road users (VRU) in the presence of driver-assisted and driver-less vehicles. Ensuring the safety of VRUs when these types of

vehicles are present is crucial, as it impacts the adoption of active travel in urban areas. An increase in the uptake of active travel accelerates the transition to sustainable mobility and transportation solutions.

With growing concerns about the carbon footprint of artificial intelligence and the need for low inference latency, we propose a Contextual Stacked Ensemble-learning (CSE) approach to analyze visual data collected from ego-vehicles. Compared to the best next model[10], our method requires $> 25\times$ less inference with $> 2.2\times$ less trainable parameters while achieving the best performance. The effectiveness of our approach is built on three main pillars: the first focuses on data compression through skeleton-ization; the second emphasizes contextualization based on a thorough analysis of attributes inferred from visual data; and the third involves trajectory analysis of VRUs.

We highlight the limitations of the current work, mostly visible with JAADbeh, and propose joint model training and broader dataset validation in our future work to address these and further prove the generalizability of the CSE approach.

Funding

This research received no external funding.

References

1. Pan American Health Organization. https://www.paho.org/en/campaigns/8th-global-road-safety-week (2025). Accessed: 2025-12-15.
2. Zohir, H. M., Ismael, I. M., El-Gendy, E. M. & Saafan, M. M. Advancements in accident-aware traffic management: A comprehensive review of v2x-based route optimization. *Sci. Reports* **15**, DOI: 10.1038/s41598-025-20878-x (2025).
3. Onkhar, V., Bazilinskyy, P., Dodou, D. & de Winter, J. The effect of driversâĂŹ eye contact on pedestriansâĂŹ perceived safety. *Transp. Res. Part F: Traffic Psychol. Behav.* **84**, 194–210, DOI: https://doi.org/10.1016/j.trf.2021.10.017 (2022).
4. Ravindran, R., Santora, M. J. & Jamali, M. M. Multi-object detection and tracking, based on dnn, for autonomous vehicles: A review. *IEEE Sensors J.* **21**, 5668–5677, DOI: 10.1109/JSEN.2020.3041615 (2021).
5. Zhang, H. *et al.* Research on a pedestrian crossing intention recognition model based on natural observation data. *Sensors* **20** (2020).
6. Xu, L., Ge, M. & Wu, W. Edge server deployment scheme of blockchain in iovs. *IEEE Transactions on Reliab.* **71**, 500–509 (2022).
7. Safavifar, Z., Mechalikh, C., Xie, J. & Golpayegani, F. Enhancing VRUs Safety Through Mobility-Aware Workload Orchestration with Trajectory Prediction using Reinforcement Learning. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 6132–6137, DOI: 10.1109/ITSC57777.2023.10421846 (2023).
8. Mahmoodi, T. 5G networks enable automated control of train traffic. *Nature* **617**, 474–475, DOI: 10.1038/d41586-023-01401- (2023).
9. Gaur, L., Afaq, A., Arora, G. K. & Khan, N. Artificial intelligence for carbon emissions using system of systems theory. *Ecol. Informatics* **76**, 102165, DOI: https://doi.org/10.1016/j.ecoinf.2023.102165 (2023).
10. Yang, B. *et al.* Faster pedestrian crossing intention prediction based on efficient fusion of diverse intention influencing factors. *IEEE Transactions on Transp. Electrification* (2024).
11. Cadena, P. R. G., Qian, Y., Wang, C. & Yang, M. Pedestrian Graph +: A Fast Pedestrian Crossing Prediction Model Based on Graph Convolutional Networks. *IEEE Transactions on Intell. Transp. Syst.* **23**, 21050–21061, DOI: 10.1109/TITS.2022.3173537 (2022).
12. IEEE ITSS. IEEE ITSS Student Competition in Pedestrian Behaviour Prediction. https://psi-intention2022.github.io/ (2023).
13. Zhang, C. & Berger, C. Pedestrian behavior prediction using deep learning methods for urban scenarios: A review. *IEEE Transactions on Intell. Transp. Syst.* **24**, 10279–10301 (2023).
14. Ahmed, S., Al Bazi, A., Saha, C., Rajbhandari, S. & Huda, M. N. Multi-scale pedestrian intent prediction using 3d joint information as spatio-temporal representation. *Expert. Syst. With Appl.* **225**, 120077 (2023).
15. Gazzeh, S., Lo Presti, L., Douik, A. & La Cascia, M. Context-ped: multi-modal context fusion for pedestrian crossing intention prediction. *Mach. Vis. Appl.* **36**, 1–16 (2025).
16. Zou, H. *et al.* A pedestrian group crossing intention prediction model integrating spatiotemporal features. *Sci. Reports* **15**, 20675 (2025).
17. Wang, T.-W., Lai, S.-H. *et al.* Multi-modal pedestrian crossing intention prediction with transformer-based model. *APSIPA Transactions on Signal Inf. Process.* **13** (2024).
18. Xie, C., Lin, C., Zheng, X., Gong, B. & López, A. M. Gtranspdm: a graph-embedded transformer with positional decoupling for pedestrian crossing intention prediction. *IEEE Signal Process. Lett.* (2025).
19. Zhang, Z., Tian, R. & Ding, Z. Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. *Proc. AAAI Conf. on Artif. Intell.* **37**, 3534–3542, DOI: 10.1609/aaai.v37i3.25463 (2023).
20. Li, H., Jin, Y. & Ren, G. Interpretable prediction of pedestrian crossing intention: Fusion of human skeletal information in natural driving scenarios. *IEEE Transactions on Intell. Transp. Syst.* (2024).
21. Alajlan, N. N., Alhujaylan, A. I. & Ibrahim, D. M. Prdtinyml: deep learning-based tinyml-based pedestrian detection model in autonomous vehicles for smart cities. *Indonesian J. Electr. Eng. Comput. Sci.* **39**, 283–309 (2025).
22. Chen, Z., Xie, X., Qiu, T. & Yao, L. Dense-stream yolov8n: a lightweight framework for real-time crowd monitoring in smart libraries. *Sci. Reports* **15**, 11618 (2025).
23. Gehrig, D. & Scaramuzza, D. Low-latency automotive vision with event cameras. *Nature* **629**, 1034–1040 (2024).

24. Abdelrahman, A. S., Islam, Z. & Abdel-Aty, M. Vrucrosssafe for crossing intention prediction of vulnerable road users for improving safe crossing at intersections. *npj Sustain. Mobil. Transp.* **2**, 20 (2025).

25. Zhang, H., Gao, S. & Wang, P. Stacking-based ensemble learning method for the recognition of the pedestrian crossing intention. *J. Adv. Transp.* **2022**, 6601014 (2022).

26. Lu, J., Chen, H., Bai, Y., Hu, C. & Zhang, X. Recognition of pedestriansâĂŹ street-crossing intentions based on skeleton features. *J. Shanghai Jiaotong Univ. (Science)* 1–14 (2024).

27. OpenMMLab . https://github.com/open-mmlab/mmpose/tree/main. Accessed: 2025-12-15.

28. Cadena, P. R. G., Yang, M., Qian, Y. & Wang, C. Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2000–2005 (IEEE, 2019).

29. Rasouli, A., Kotseruba, I. & Tsotsos, J. K. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC* (2019).

30. Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A. & ÃŰzgÃijner, Ã. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intell. Veh.* **7**, 221–230, DOI: 10.1109/TIV.2022.3162719 (2022).

31. Zhong, R., Chiang, C.-Y., Fathy, Y., Slagaugh, G. & Jaber, M. Safe intelligent vehicles: Intent is what we need. *IEEE intelligent transportation Syst. magazine* (2026).

32. Rasouli, A., Kotseruba, I. & Tsotsos, J. K. Pedestrian action anticipation using contextual feature fusion in stacked RNNs. *arXiv preprint arXiv:2005.06582* (2020).

33. Zamboni, S., Kefato, Z. T., Girdzijauskas, S., NorÃl'n, C. & Dal Col, L. Pedestrian trajectory prediction with convolutional neural networks. *Pattern Recognit.* **121**, 108252, DOI: https://doi.org/10.1016/j.patcog.2021.108252 (2022).

34. Nikhil, N. & Tran Morris, B. Convolutional neural network for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018).

35. Abu Farha, Y., Richard, A. & Gall, J. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5343–5352 (2018).

36. Zhou, Q., Liang, H., Lin, Z. & Xu, K. Multimodal feature fusion for video advertisements tagging via stacking ensemble. *arXiv preprint arXiv:2108.00679* (2021).

37. Rasouli, A., Kotseruba, I. & Tsotsos, J. K. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 206–213 (2017).

38. Chen, T. & et al. PSI: A Pedestrian Behavior Dataset for Socially Intelligent Autonomous Car (2022). 2112.02604.

39. Leal-Taixé, L., Milan, A., Reid, I., Roth, S. & Schindler, K. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]* (2015). ArXiv: 1504.01942.

40. Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A. & Özgüner, Ü. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intell. Veh.* **7**, 221–230 (2022).

41. ykotseruba. Jaad: default split ids dataset âĂŤ jaad_2.0 (split_ids/default). https://github.com/ykotseruba/JAAD/tree/JAAD_2.0/split_ids/default (2025). [Source code, GitHub repository].

42. Pedregosa, F. & et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

43. Cadena, P. R. G., Qian, Y., Wang, C. & Yang, M. Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intell. Transp. Syst.* **23**, 21050–21061, DOI: 10.1109/TITS.2022.3173537 (2022).

44. Han, S., Mao, H. & Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

45. Rasouli, A., Kotseruba, I., Kunic, T. & Tsotsos, J. K. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6262–6271 (2019).

46. Liu, B. *et al.* Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics Autom. Lett.* **5**, 3485–3492 (2020).

Author contributions statement

The ideation for this work was done by C.C, Y.F, and M.J while the implementation was led by C.C. All authors C.C, Y.F, G.S., and M.J contributed to the analysis of the results, the write-up and paper review.

Correspondence and requests for materials should be addressed to C. C..

Competing interests

The authors declare no competing interests.