Predicting Pedestrian Movement in Unsignalized Crossings: A Contextual Cue-Based Approach

Kaliprasana Muduli
PhD Scholar, Civil Engineering Department
Indian Institute of Technology Roorkee
Roorkee, India
k_muduli@ce.iitr.ac.in

Vikas Sahu
M. Tech Student, Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology Roorkee
Roorkee, India
vikas_s@mfs.iitr.ac.in

Indrajit Ghosh
Associate Professor, Civil Engineering Department & Joint Faculty, Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology Roorkee
Roorkee, India
indrafce@iitr.ac.in

2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS) | 979-8-3503-8311-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/COMSNETS59351.2024.10427345

Abstract*— To ensure safe and secure coexistence between pedestrians and autonomous vehicles (AVs), AVs must be able to anticipate pedestrian behavior and respond to it. This research gathers video data from real traffic scenes to predict pedestrian crossing intentions at unsignalized crossings. Computer vision techniques such as YOLOv4, Deep SORT, and perspective transformation are employed for road user detection, tracking, and mapping image coordinates to world coordinates to prepare trajectory datasets. Using trajectory data, several features influencing pedestrian intention like walking speed, location in the road environment, count and direction of approaching traffic, speed and type of closest approaching vehicle upstream, etc., are extracted. The dataset for this study was obtained by analyzing 1,411 pedestrians, resulting in 223,136 samples. To predict pedestrian crossing intentions, LSTM and Bi-LSTM with an attention mechanism model were built and trained to anticipate the pedestrian crossing intention at unsignalized crossing. The proposed model successfully combined the characteristics and surrounding dynamics of pedestrians to produce accurate predictions, Bi-LSTM with an attention mechanism outperformed LSTM, achieving an AUC of 95.3%, 91.1%, 89.2%, 87.5%, and 84% on the testing dataset at unsignalized crossing on the 0.6 sec, 1.2 s, 1.8 s, 2.4 s, and 3 s time horizons. These outcomes can be used to improve Connected and Autonomous Vehicle (CAV) technologies, infrastructure-to-vehicle (I2V) connectivity, and driver assistance systems to enhance pedestrian safety while navigating through pedestrian crosswalks.

***Keywords*— Pedestrian Crossing Intention Prediction, Unsignalized Crossings, Computer vision, Attention.**

I. Introduction

Pedestrian safety is a crucial issue in modern transportation systems, and it has drawn the attention of policymakers, planners, and researchers worldwide. In recent years, the number of accidents involving pedestrians has increased, particularly at unsignalized crossings where pedestrian-vehicle interactions are largely unregulated. According to India's Ministry of Road Transport and Highway (MoRTH) report, 29,124 pedestrians died in 2021 which account for 18.9% of total road accident fatalities [1]. According to 2019 data, approximately 46% of the fatalities in Delhi, the capital of India, and 35% of those killed in traffic accidents in Chandigarh, India were pedestrians [2][3].

Predicting pedestrian intention is complex, and many dynamic road scenarios affect pedestrian behavior. As the world moves toward autonomous vehicles (AV), researchers have conducted in-depth surveys to identify the variables influencing pedestrian behavior. Kadali and Vedagiri [4] conducted a survey in Mumbai, India to examine pedestrian speed behavior in traffic and found that insufficient space between pedestrians and vehicles has a significant impact on the speed of pedestrian crossings, that medians lessen the frequency of changes in crossing behavior, that the size of the grouping of pedestrians reduces the variation in their crossing pace, and that there is a significant speed shift as faster-moving cars approach. In a survey conducted in Bengaluru, India, Nicholas N. Ferenchak [5] discovered a high correlation between age and gender in terms of pedestrian behavior at midblock crossings. Older persons were more careful, but male pedestrians were riskier than females. Waiting times, using crosswalks, and interactions with moving vehicles were all influenced by age and gender. Additionally, discovered that larger groups exhibit riskier behavior in comparison to smaller groups [6]. The types of vehicle and volume of traffic also affect pedestrian behavior as observed in the survey conducted by Asaithambi et al. [7] in Mangalore, India. There are a number of other features, including a pedestrian's pose, that help predict pedestrian intention. However, in countries like India, where non-lane-based heterogeneous traffic condition prevails, pedestrians' intention to cross the road largely depends on surrounding dynamics. This study uses contextual cues like closest vehicle speed, distance, vehicle count, etc., as input variables to predict pedestrians' intentions.

Recurrent neural networks (RNNs), in particular, have been studied as a potential deep learning model for predicting pedestrian intention by several academics. RNNs work by transferring outputs from the one-time frame in one layer to the next. RNNs, however, have problem when there is a large enough delay between the input and the unit [8][9]. A particular kind of RNN called LSTM was created to represent sequential data by capturing long-term dependencies [10]. Long short-term memory (LSTM) was used by Ghori et al. [11] to complete the sequence modeling challenge of pedestrian intention prediction. A modified form of LSTM called social LSTM was developed by Alahi et al. [12] to forecast walkers' trajectories based on their social interactions in congested environments. As seen by their successful applications in pedestrian intention and path prediction over the past few years, LSTM models have proven to be the most successful models for anticipating walkers' unexpected crossings, whether using posture sequences or historical trajectory data. Zhang Set al. [13] used LSTM for predicting pedestrian intention considering some contextual features, however as discussed earlier, many contextual features that influence pedestrian behaviors in developing countries like India have not been considered.

Jaywalking is an act when pedestrians cross the road without following the traffic rules–involves sudden and unpredictable movements. The presence and actions of nearby motor vehicle traffic greatly impact it. This behavior is especially common in developing countries like India, where it poses a serious risk to pedestrian safety when they engage with moving automobiles, particularly at unsignalized crossing locations. This risk is only likely to grow as connected and autonomous vehicles become more common. In low-trust situations (such as aggressive AV driving at unsignalized crosswalks), Jayaraman S et al. [14] discovered that using an explicit communication interface is one way to express the pedestrian intention to an AV. This can reduce uncertainty and increase AV trust. Nevertheless, despite this urgent issue, there has not been much focus on developing reliable techniques for forecasting the behavior of pedestrians. Numerous studies in the literature have aimed to forecast pedestrian crossing intentions by analyzing historical trajectory and pose sequences. However, in the Indian context, where non-lane-based heterogeneous traffic is prevalent, pedestrian intentions are significantly influenced by the surrounding context, which has largely been overlooked in many earlier studies. This study investigates the application of computer vision and deep learning methodologies to predict the future state of pedestrians while crossing the road, incorporating various critical parameters related to local dynamics.

II. DATA COLLECTION

A. Study Site

To select a suitable location for the study, certain prerequisites needed to be fulfilled, including a high volume of pedestrians, unsignalized crossing, and records of the accident. A stretch of an uncontrolled crossroads at Cheema Chowk, Panjab, India, was chosen for the study after consultation with local police stations. The location was chosen due to the high volume of pedestrians and the non-availability of proper signalized crossing. A study by Muduli et al. (2023) indicated that the study site was a hotspot for pedestrian crashes. Using an optical camera mounted on a nearby fly-over, video data were captured for 4 hours in the morning and evening hours of typical weekdays. The area under consideration consists of a dual carriageway with two lanes in each direction, separated by a physical median. The site's spatial map, which shows the study area and camera mounting location, is presented in Figure 1.

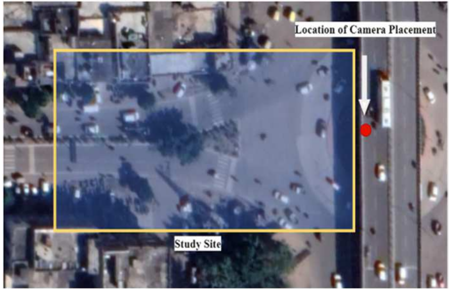


Figure 1. Satellite Image of the site of study.

B. Video Processing

To process the gathered video data from the study site, deep learning-based computer vision techniques are utilized. The main objectives of this processing are to extract road users' trajectories. The computer vision techniques used for this purpose include object detection, object tracking, and perspective transformation, similar to the approaches used by Muduli & Ghosh (2023).

Object detection: In this work, YOLOv4[17] algorithm was used for road user detection. YOLOv4 is a cutting-edge object detection algorithm that achieves high accuracy and speed in detecting multiple objects in real time. YOLO is faster than two-stage object detection algorithms because it uses a single-stage detector that directly predicts object bounding boxes and class probabilities in a single forward pass through the network without the need for an extra region proposal network (RPN) stage. In contrast, two-stage detectors, such as Faster R-CNN, require a separate RPN stage to generate region proposals, which are then fed into the object detection network for final detection [18][19].

Object tracking: The process of tracking the movements of objects is called Multiple Object Tracking (MOT), and it has been widely used in various applications. In this study, the Deep SORT algorithm [20][21] was employed for object tracking, which has demonstrated high performance on the MOT16 Challenge benchmark [22]. The Deep SORT algorithm utilizes a combination of deep neural networks for both object detection and tracking. Initially, the YOLOv4 algorithm is used to detect and classify road users, and then the Deep SORT algorithm takes the detection boxes as input to track the movement of each road user.

Perspective Transformation: In this work, the perspective transformation technique has been used to convert the location of an object in an image from image coordinates to world coordinates. This conversion is accomplished using a homography matrix (h), which relates the image plane to the world plane. As shown in (Equation 1), the homography matrix contains nine values, which are used to transform the image coordinates (u , v) to their corresponding world coordinates (X, Y). The homography matrix is calculated by finding the transformation that maps a set of points in the image plane to their corresponding points in the world plane. These points are typically defined by a set of image coordinates (u_i, v_i) and their corresponding world coordinates (X_i, Y_i), (which are obtained using a GPS). The process of calculating the homography matrix involves using a linear least squares method [23][24] and singular value decomposition. Once the homography matrix has been calculated, it can be used to transform all image coordinates to their corresponding world coordinates (Equation 1).

$$\begin{bmatrix} \mathrm{u} \\ \mathrm{v} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathrm{h}_1 & \mathrm{h}_2 & \mathrm{h}_3 \\ \mathrm{h}_4 & \mathrm{h}_5 & \mathrm{h}_6 \\ \mathrm{h}_7 & \mathrm{h}_8 & \mathrm{h}_9 \end{bmatrix} \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ 1 \end{bmatrix} = \mathrm{h} \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ 1 \end{bmatrix}$$

$$=> \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ 1 \end{bmatrix} = \mathrm{h}^{-1} \begin{bmatrix} \mathrm{u} \\ \mathrm{v} \\ 1 \end{bmatrix} \quad (1)$$

(u_i, v_i): Coordinate of each point on the image plane
(X_i, Y_i): Coordinate of each point on the world plane

The haversine formula is used for calculating the speed of moving road users between time t_1 and t_2 (Equation 2).

$$Speed = \frac{haversine\ ((X_1, Y_1), (X_2, Y_2))}{t_2 - t_1} \quad (2)$$

The flowchart of the data extraction process is shown in Figure 2.

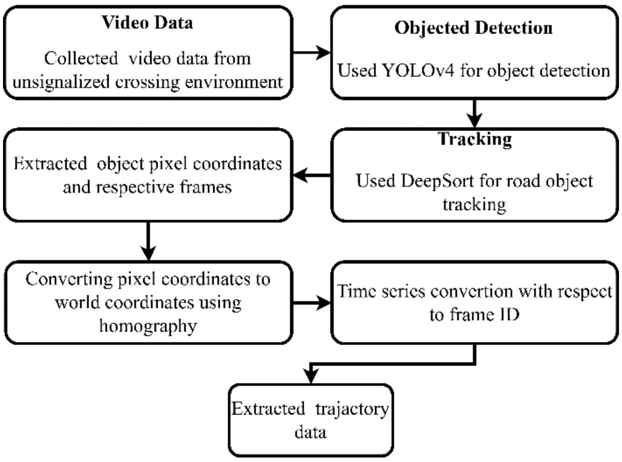

Video Data
Collected video data from unsignalized crossing environment
Objected Detection
Used YOLOv4 for object detection
Extracted object pixel coordinates and respective frames
Tracking
Used DeepSort for road object tracking
Converting pixel coordinates to world coordinates using homography
Time series convertion with respect to frame ID
Extracted trajectory data

Figure 2. Flowchart of data extraction process

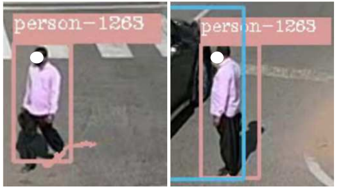
III. METHODOLOGY

A. *Input Variable*

Pedestrian crossing intention can be influenced by various parameters, such as the speed of pedestrians, the direction of oncoming traffic, the location of pedestrians relative to the road environment, gender, grouping, closest approaching vehicle description, etc. Five categories are used to categorize the pedestrian's position in relation to the road environment: closer to the beginning curb, closer to the first centerline, closer to the median, closer to the second centerline, and closer to the finishing curb. This classification is used to maintain flexibility across roads with varying numbers of lanes rather than relying on lane numbers as a variable.

B. Pedestrian crossing intention labeling

A time-series dataset was created from the trajectories of pedestrians, and the dependent variable Y was labeled to indicate whether the pedestrian intended to cross the crossing. The behavior of pedestrians were observed in two stages, namely waiting and crossing, and the dependent variable Y was labeled based on their behavior, as shown in Figure 3.



(a) Crossing (b) Waiting

Figure 3. Pedestrian intention labeling

For prediction, a prediction horizon of 0.6s, 1.2s, 1.8s, 2.4s, and 3s were used. Hence, annotated labels were advanced by respective time to prepare the data set. Table I summarizes all of the independent and dependent variables along with their descriptions.

TABLE I. INDEPENDENT AND DEPENDENT VARIABLES

Variables		Variable Description	Type
Independent variable	Speed of pedestrian	Speed of pedestrian m/s	Continuous
	The direction of Approaching Vehicle	Vehicle approaching direction with respect to pedestrian, for right side (0), for left side (1)	Nominal categorical
	Pedestrian location in the road environment	Nearer to Starting curb (0), nearer to the first centerline (1), nearer to the median (2), nearer to the second centerline (3), nearer to the ending curb (4)	Nominal categorical
	Speed of closest approaching vehicle	Speed of the nearest approaching vehicle in m/s.	Continuous
	Closest Vehicle Distance	Distance in (m)	Continuous
	Type of closest approaching vehicle	No vehicle (0), Bicycle (1), Motorbike (2), Car, van, tuk-tuk (3), Light Truck (4), Truck (5), Bus (6)	Nominal categorical
	Approaching vehicle count	Total number of approaching vehicles	Numeric
	Gender	Male (0), Female (1)	Nominal categorical
	Grouping	Whether the pedestrian is walking in a group, no grouping (0), group of two (1), group of three (2), group of four or more (3),	Nominal categorical
	Age group	Child (0), Teen (1), Mid-age (2), Old-age (3)	Nominal categorical
Dependent	Crossing intention	Whether the pedestrian will cross. Crossing (1), Waiting (0)	Nominal categorical

C. *Model Development*

In this research recurrent neural networks (RNNs) based models, Long Short-Term Memory (LSTM), and Bi-directional Long Short-Term Memory (Bi-LSTM) with attention model used to predict pedestrian crossing intentions based on the characteristics of the pedestrians and surrounding dynamics at the unsignalized pedestrian crossing. Since pedestrian trajectories involve time-series data, these RNN-based models can effectively capture the temporal dependencies of pedestrian trajectories, crucial for intention prediction.

LSTM: Long Short-Term Memory (LSTM) unit provides a method to selectively allow information to pass through by deleting or adding some information through the gate structure. This is the key to this network. To retain and update the cell state, the LSTM unit has three gate structures (input gate, forget gate, and output gate). The three gate structures and the memory cell state corresponding to time t are represented here by the symbols i_t , f_t , o_t , and c_t respectively.

Forget gate (f_t) eliminates some data from the memory cells, input gate (i_t) determines which new information should be kept in the cell state, output gate (o_t) collects the output information in the end. information flow in LSTM is regulated by these three gates.

$$f_t = \sigma\left(w_f[h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{4}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{5}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{7}$$

$$y_t = w_y \cdot h_{t-1} + b_y \tag{8}$$

σ = Sigmoid Activation function.
w = Weight matrices
b = Bias
x_t = input vector
h_t = Hidden output vector
y_t = output vector

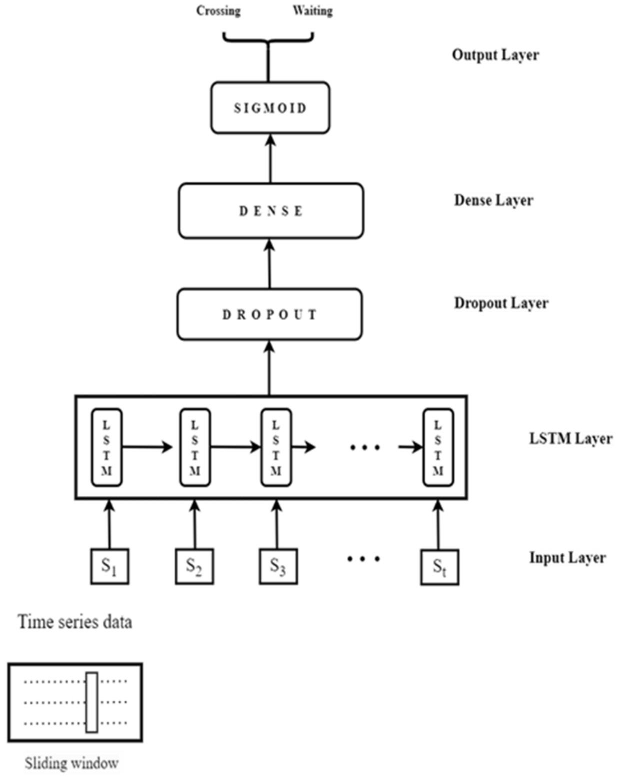


Figure 4: Architecture of LSTM model

Output vector (y_t) is computed using equation (3) to equation (8). Figure 4 shows the LSTM model architecture that was applied in this work. The model consists of an input layer, a stacked LSTM layer, a dense (fully linked) layer, and an output neuron that represents the classification result. Additionally, the dropout layer is included to avoid overfitting. For optimization, the Adam function is used. The final output is obtained using the Sigmoid activation function.

Bi-LSTM with attention: Bi-LSTM (Bi-directional LSTM) neural network model can effectively capture the temporal dependencies of pedestrian trajectories by processing the input sequence in both forward and backward directions. An attention mechanism is added to the Bi-LSTM to improve its capacity to concentrate on important parts of the input sequence. The model can give varying weights to the input sequence. As a result, the model can focus on the portions of the input that are most important for making predictions. In order to determine the relative significance of each time step's hidden state for the outcome prediction, attention weights are computed for each time step. The learning function f (equation 9), which is implemented using a feed-forward network, is used to compute the alignment score. The alignment scores are normalized using the softmax function (equation 10) to get a distribution of weights that adds up to 1. The model can concentrate on the most relevant portions of the time series by taking a weighted sum of the hidden states using the attention weights. The weighted sum of the hidden states is also known as the context vector c (equation 11).

$$e_t = f(h_t) \tag{9}$$

$$a_t = \frac{\exp(e_t)}{\sum_{j=1}^{T_x} \exp(e_j)} \tag{10}$$

$$\mathrm{c} = \sum_{t=1}^{T_x} a_t h_t \tag{11}$$

were,
e_t : Alignment score at time t.
a_t : Weight at time t.
c : Context vector.

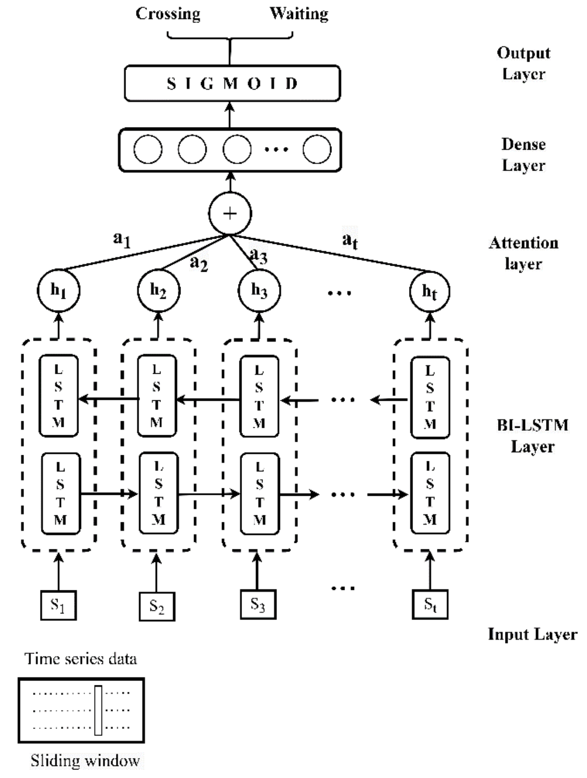


Figure 5: Architecture of Bi-LSTM with the attention mechanism

The architecture of Bi-LSTM with the attention mechanism used in this study, is shown in Figure 5. The architecture consists of multiple layers. First, the input layer accommodates the time-series data. Second, two layers of Bi-LSTM are applied to the input sequence to process the sequence. Third, the attention layer is utilized to concentrate on the relevant part of the sequence, then the attention output is flattened to prepare for further processing. Finally, a dense layer with a sigmoid activation function is used to produce binary classification output, indicating pedestrian future crossing intention.

IV. EXPERIMENTS AND RESULTS

The Bi-LSTM with attention mechanism model was built to predict pedestrian intention at unsignalized crossings environment, The hyperparameters of the model are tuned such that it will achieve the best performance.

A. *Evaluating metrics*

The following evaluation matrices were used to check the performance of the proposed model:

Accuracy: The proportion of correctly classified instances out of the total number of instances, as shown in Equation (6).

$$Accuracy \;= \frac{True\ Positive+True\ Negative}{True\ Positve+False\ Positive+False\ Negative+True\ Negative} \qquad (6)$$

Precision: The proportion of accurately identified samples among positively classified samples, as shown in Equation (7).

$$Precision = \frac{True\ Positive}{True\ Positve+False\ Positive} \qquad (7)$$

Recall (also known as sensitivity or true positive rate): The proportion of correctly identified samples among genuine positive samples, as shown in Equation (8).

$$Recall = \frac{True\ Positive}{True\ Positve+False\ Negative} \qquad (8)$$

F1-score: The harmonic mean of precision and Recall, as shown in Equation (9).

$$F1score = \frac{2*Precision*Recall}{Precision+Recall} \qquad (9)$$

AUC (Area Under the ROC Curve): The performance metric used to evaluate the effectiveness of a binary classification model. It represents the total area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold settings.

B. *Experiment Results*

In this discussion, 4 hours of video data were collected from the study site, which has data of 1,411 pedestrians. The pedestrian data are split into training sets (80%) and test sets (20%) sets in order to train and test the performance of both proposed models. There was a total of 223,236 instances, out of which 202,753 were positive and 20,383 were negative instances.

An imbalanced distribution of data will cause poor performance of the model. The synthetic minority over-sampling method (SMOTE)[25] is utilized to artificially produce additional data on the minority class in order to address the data imbalance in this work. SMOTE is a well-known machine learning approach to handling unbalanced data. To balance the class distribution, it operates by generating artificial cases of the minority class in a dataset. SMOTE operates by locating minority class instances and interpolating between them to produce new synthetic instances. The proposed Long Short-Term Memory (LSTM) and Bi-directional Long Short-Term Memory (Bi-LSTM) with attention model was evaluated in the time horizon of 0.6s, 1.2s, 1.8s, 2.4s, and 3s . The evaluation matrices' results are shown in Table II. The performance metrics (AUC, Precision, Accuracy, and F1-Score) often fall with an increase in prediction horizon, as can be seen for both models. This implies that the accuracy of the model's predictions declines with increasing prediction horizons. Predicting pedestrian crossing intentions accurately at shorter time horizons is somewhat expected due to the immediacy of the decision-making process involved. The predictions are more accurate because the model, at a shorter time frame, covers the most immediate, deliberative behaviors of pedestrians when deciding to cross. Although both models perform satisfactorily, bi-LSTM with attention outperforms LSTM in all metrics at all time horizons proving its robustness for handling longer time sequences. The Bi-LSTM's attention mechanism enables the model to concentrate on more important temporal data, improving effectiveness in capturing pedestrians' behavior.

TABLE II. EVALUATION MATRICES' RESULTS

Time Horizon	Model	Accuracy	Recall	Precision	F1-Score	AUC
0.6 sec	LSTM	0.873	0.938	0.834	0.883	0.934
	Bi-LSTM with Attention	0.919	0.948	0.897	0.922	0.953
1.2 sec	LSTM	0.833	0.823	0.845	0.834	0.875
	Bi-LSTM with Attention	0.854	0.836	0.868	0.851	0.911
1.8 sec	LSTM	0.814	0.880	0.775	0.828	0.864
	Bi-LSTM with Attention	0.854	0.909	0.820	0.862	0.892
2.4 sec	LSTM	0.819	0.897	0.777	0.833	0.839
	Bi-LSTM with Attention	0.844	0.917	0.799	0.854	0.875
3 sec	LSTM	0.814	0.866	0.785	0.823	0.811
	Bi-LSTM with Attention	0.854	0.887	0.830	0.858	0.840

V. CONCLUSION

This study presents a deep-learning approach using LSTM and Bi-LSTM with an attention mechanism for predicting pedestrian crossing intentions at an unsignalized crossing in

Ludhiana, Punjab, India. Utilizing real-world traffic video data, the research captures not only motion parameters (such as speed and distance of vehicles and pedestrians) but also contextual features like gender, age, and location on the road, summing up to 10 different features that influence pedestrian behavior in heterogeneous traffic.

The Bi-LSTM with attention model outshines the standard LSTM in accuracy, providing more reliable predictions at intervals ranging from 0.6 to 3 seconds into the future. The model exhibits high recall values, crucial for safety-critical applications like ADAS in autonomous vehicles, ensuring most pedestrians with the intention to cross are correctly identified. With over 80% accuracy at all prediction horizons, the proposed system could significantly enhance pedestrian safety by offering earlier warnings to drivers or autonomous systems.

The study acknowledges the complexity of pedestrian behavior and suggests that future research should focus on more sophisticated deep-learning models and broader datasets, possibly integrating multi-modal sensor data to refine the accuracy and reliability of prediction systems. The proposed model could be deployed at unsignalized crossings to facilitate I2V communication, alerting vehicles about potential pedestrian movements, thereby preventing accidents and enhancing pedestrian safety.

REFERENCES

[1] "Road Accidents in India Road Accidents in India Road Accidents in India" 2021. [Online]. Available: www.morth.nic.in

[2] DTP. *Road Accidents in Delhi*. Delhi Traffic Police, Government of Delhi, India, 2019.

[3] CTP. *Road Accidents in Chandigarh*. Chandigarh Traffic Police, India, 2019.

[4] B. R. Kadali and P. Vedagiri, "Evaluation of pedestrian crossing speed change patterns at unprotected mid-block crosswalks in India," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 7, no. 6, pp. 832–842, Dec. 2020, doi: 10.1016/j.jtte.2018.10.010.

[5] N. N. Ferenchak, "Pedestrian age and gender in relation to crossing behavior at midblock crossings in India," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 3, no. 4, pp. 345–351, Aug. 2016, doi: 10.1016/j.jtte.2015.12.001.

[6] "Pedestrian Crossing Behavior in Relation to Grouping and Gender in a Developing Country Context," *Journal of Global Epidemiology and Environmental Health*, pp. 37–45, Dec. 2017, doi: 10.29199/geeh.101018.

[7] G. Asaithambi, M. O. Kuttan, and S. Chandra, "Pedestrian Road Crossing Behavior Under Mixed Traffic Conditions: A Comparative Study of an Intersection Before and After Implementing Control Measures," *Transportation in Developing Economies*, vol. 2, no. 2, Oct. 2016, doi: 10.1007/s40890-016-0018-5.

[8] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Trans Neural Netw*, vol. 5, no. 2, pp. 157–166, 1994, doi: 10.1109/72.279181.

[9] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training Recurrent Neural Networks," Nov. 2012, [Online]. Available: http://arxiv.org/abs/1211.5063

[10] M. L. , H. P. , Y. K. , J.-S. P. , G.-J. J. J.-H. K. Donghyun Lee, "Long_short-term_memory_recurrent_neural_network-based_acoustic_model_using_connectionist_temporal_classification_on_a_large-scale_training_corpus".

[11] O. Ghori *et al.*, "Learning to Forecast Pedestrian Intention from Pose Dynamics," in *IEEE Intelligent Vehicles Symposium, Proceedings*, Institute of Electrical and Electronics Engineers Inc., Oct. 2018, pp. 1277–1284. doi: 10.1109/IVS.2018.8500657.

[12] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 961–971. doi: 10.1109/CVPR.2016.110.

[13] S. Zhang, M. Abdel-Aty, J. Yuan, and P. Li, "Prediction of Pedestrian Crossing Intentions at Intersections Based on Long Short-Term Memory Recurrent Neural Network," *Transp Res Rec*, vol. 2674, no. 4, pp. 57–65, Apr. 2020, doi: 10.1177/0361198120912422.

[14] S. K. Jayaraman *et al.*, "Pedestrian Trust in Automated Vehicles: Role of Traffic Signal and AV Driving Behavior," *Front Robot AI*, vol. 6, Nov. 2019, doi: 10.3389/frobt.2019.00117.

[15] Muduli, K., Sahu, D., & Ghosh, I. (2023, July 20). A GIS-based framework for identification and prioritization of traffic crash hotspots. Paper presented at the World Conference on Transport Research - WCTR 2023, Montreal.

[16] K. Muduli and I. Ghosh, "Prediction of the Future State of Pedestrians While Jaywalking Under Non-Lane-Based Heterogeneous Traffic Conditions," Transportation Research Record: Journal of the Transportation Research Board, p. 036119812311616, Apr. 2023, doi: 10.1177/03611981231161619.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.10934

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[19] R. Girshick, "Fast R-CNN," Apr. 2015, [Online]. Available: http://arxiv.org/abs/1504.08083

[20] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.07402

[21] N. Wojke and A. Bewley, "Deep Cosine Metric Learning for Person Re-Identification," Dec. 2018, doi: 10.1109/WACV.2018.00087.

[22] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.00831

[23] J. Jakubˇ, J. Vojtěch Bartl, and R. Juránek, "Vehicle Re-Identification and Multi-Camera Tracking in Challenging City-Scale Environment." [Online]. Available: https://medusa.fit.vutbr.cz/traffic/

[24] Z. Tang *et al.*, "CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.09254

[25] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.