STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, June 2026, pp 5389–5405.
Published online in International Academic Press (www.IAPress.org)

A CNN2D-LSTM Framework for Rule-Based Pedestrian-Vehicle Risk Scenario Detection

Oumaima Benkhadda*, Meriem Mandar

Department of Mathematics and Computer Science, HASSAN II University, Superior Normal School Casablanca, MOROCCO

Abstract In this work, we developed an approach for rule-based pedestrian-vehicle risk scenario detection from video sequences. The contributions lie in the classification of "risky" and "non-risky" situations automatically derived from behavioral and physical cues, which could improve accident prevention and intelligent driver-assistance systems. A two-dimensional convolutional neural network (CNN2D) is employed over the frames of the videos for visual features extraction, while the LSTM recurrent network models the temporal dynamics of the sequences. The data used in these experiments are sequences of video frames extracted from the JAAD dataset. Behavioral and physical variables include pedestrian crossing, gaze direction, vehicle action, proximity, which are used only for generating the risk labels. They are not provided as explicit input to the model. While these labels are heuristically generated and may not capture all possible risky scenarios, they provide a practical framework for model training and evaluation. The CNN2D extracts the spatial visual features from the frames, while the LSTM captures temporal dependencies, which permits the model to learn both in the spatial and temporal axes for the prediction of risk. Tests on the JAAD dataset composed of varied traffic conditions and images of pedestrian crossings report overall accuracy of 97% and class-wise precision 99.5% for the "No Risk" class and 92.2% for the "Risk" class. These results confirm the effectiveness of the suggested model and demonstrate the usefulness of fusing visual and temporal information collectively for automatic risk detection in difficult traffic environments.

Keywords Pedestrian-Vehicle Risk Detection, CNN2D-LSTM Architecture, Video-Based Risk Assessment, Intelligent Transportation Systems (ITS), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM)

DOI : 10.19139/soic-2310-5070-3458

1. Introduction

The threat of danger for human life is always present when people act as pedestrians, a recurring activity carried out on a daily basis. Evidence arising from traffic safety statistics confirms that crossing a roadway, an act regarded as mundane, is actually risky. During 2024, pedestrians were responsible for almost a quarter of road traffic deaths in Morocco, contributing 24.7% to the overall number of road traffic deaths. The importance of pedestrian road safety has become an increasingly significant challenge, for which many studies have recently addressed pedestrian safety using new AI advancements. The need for expertise has arisen due to the fact that pedestrians' reactions towards road danger have yet to be systematically investigated with the aim of designing systems that could improve the coexistence of pedestrians, vehicles, and consequently, decrease road deaths. These range from stand-alone pedestrian detection in traffic environments to the creation of intelligent systems that can anticipate high-risk situations and help prevent accidents. The work here presents a deep learning-based method for real-time prediction of the likelihood of accidents involving pedestrians and vehicles. The proposed approach depends on an advanced decision-support mechanism that offers better protection to vulnerable road users by contributing to accident prevention. It appears that the constructed system is based on a two-dimensional CNN integrated into an LSTM network. In fact, it seems that such a configuration is a natural fit given that the type

∗. Correspondence to: Oumaima Benkhadda (Email : oumaimab665@gmail.com). Department of Mathematics and Computer Science, HASSAN II University, Superior Normal School Casablanca, MOROCCO (20420)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

of data considered within this particular study is based on a video format. This naturally appears to have been incorporated as part of a task-specific CNN2D model that is primarily based around identifying possible signals of potential imminent danger within a spatial format provided in each of the frames within a video, and how this is then coupled within an LSTM framework which is based around identifying potential patterns within sequences of images captured within a more temporal format. It seems that such a powerful framework is then enabled around real-time predictive analysis of potential dangerous behavior during interactions of vehicles and pedestrians as captured within a video format. This combined system has the further ability to indirectly infer the implicit context and interaction in the data for things such as reversals for pedestrians, relative speed variation, as well as the impact of occlusions, which are typical indicators of potential threats in the context of traffic in an urban environment and are harder to determine by direct image extraction. Thus, the system described has the quality of anticipatory know-how in the estimation of risk through its capacity to use both spatial and temporal variation for the predictive estimation of dangerous incidents before they occur. It has even greater significance in a complicated environment such as the urban one where irregular pedestrian behavior and traffic arrangement demand early warning systems for the mitigation of threats posed by accidents and fatalities. Alongside this advancement in methodology, works in pedestrian intention estimation and prediction have gradually grown using combined structures of specialized data resources, sophisticated learning frameworks, risk estimation strategies, as well as competent perception modules. The PIE data set and baseline system proposed in [1] is a landmark achievement in this area, as it offers a large-scale benchmark solution to pedestrian intention estimation and trajectory prediction tasks, thereby allowing evaluation of cross/not cross behavior choices in real-world settings to be systematically conducted. Subsequent works have focused on more concrete feature characterization and more complex learning settings : while pedestrian intention prediction is defined in a multi-task learning framework to simultaneously predict intention along with other behaviorally relevant features in [2] , features are fused along with spatio-temporal attention mechanisms in [3]. Multimodal fusion strategies are further advanced by [4], who propose a hybrid attention mechanism to dynamically weight heterogeneous inputs, and by [5], who exploit multi-scale spatio-temporal representations derived from 3D joint information to explicitly encode body dynamics and motion cues relevant to pedestrian intent. Beyond predictive performance alone, increasing attention has been paid to transparency and interpretability : [6] provide experimental insights toward explainable and interpretable pedestrian crossing prediction, and [7] extend this line of work through a neuro-symbolic framework that combines learning-based models with symbolic reasoning to enhance interpretability in autonomous driving scenarios. Complementing intention-focused studies, a substantial body of literature addresses trajectory-based safety analysis and risk estimation using computer vision and advanced statistical modeling. [8] review advances and applications of computer vision techniques for vehicle trajectory generation and surrogate traffic safety indicators, demonstrating how vision-based pipelines can support proactive safety assessment. [9] introduce Bayesian generalized extreme value models for real-time pedestrian crash risk estimation at signalized intersections using AI-based video analytics, and apply extreme value theory to estimate crash risk during mandatory lane-changing in connected environments. These approaches are further reinforced by [10], who provide additional empirical insights into pedestrian crash risk analysis using extreme value models. At a finer interaction scale, [11] conduct a quantitative analysis of lane-based pedestrian-vehicle conflicts at non-signalized marked crosswalks, highlighting the importance of interaction geometry and dynamics. Alternative modeling perspectives are offered by [12], who present a comprehensive review and meta-analysis of fuzzy logic, neural networks, machine learning, and genetic algorithms for occupational risk assessment. As well as by [13], who propose fuzzy and intuitionistic fuzzy approaches for modeling pedestrian-vehicle risk exposure and adapt Gibson's theory to analyze distance perception dynamics between pedestrians and vehicles [14]. These higher-level modeling efforts are further supported by perception-oriented contributions, including the work of [15] on road traffic safety risk assessment using integrated radar-video sensor data, and that of [16], who address real-time processing constraints through an efficient FPGA implementation of multi-scale optical flow algorithms for high-resolution video streams. Collectively, these studies encompass datasets, learning-based intention prediction, explainability, trajectory and risk modeling, fuzzy and statistical approaches, and sensor-level perception, forming a comprehensive and coherent foundation for pedestrian intention and safety analysis. Nevertheless, despite the advances in pedestrian safety research, current solutions to pedestrian detection still have certain limitations. Methodologies addressing only pedestrian detection have in common an oversight on risk and, thus, do not

represent pedestrian-vehicle encounter dynamics. Intention prediction models, although successful in predicting crossing intentions, involve only disconnected behavioral features and therefore fail to capture pedestrian-vehicle encounter dynamics. Models related to trajectory and risk, whether statistical or fuzzy, require strong assumptions or rely on rule-based methods, which are not flexible enough to cope with complex traffic scenarios. On the other hand, the proposed approach, the CNN2D-LSTM framework, not only overcomes these issues but can actually take advantages of the sequence of images in the video data. This approach can make use of the visual spatial information provided by the CNN part to extract high-level spatial features as well as the temporal dependencies within a sequence of images provided by the LSTM component. This paper intends to investigate the potential of spatio-temporal deep learning techniques for risk detection of pedestrian and vehicle interactions in video evidence. Instead of arguing for the readiness of the developed model for practical implementation in accident prediction, the research paper discusses the presentation of an engineering framework involving joint analysis of space and time for pedestrian and vehicle interactions using visual representations. This paper will make use of the available public benchmark dataset and emphasize the potential of its approach at its current stage.The remainder of this paper is organized as follows. Section 2 introduces the proposed CNN2D-LSTM spatio-temporal architecture. Section 3 details the dataset, preprocessing steps, and the heuristic risk annotation procedure. Section 4 presents the experimental results along with discussion and analysis. Section 5 concludes the paper and highlights directions for future work.

1.1. Proposed Model

In this work, we employ a hybrid CNN2D–LSTM model to jointly exploit the spatial and temporal information contained in video sequences. A sequence consists of consecutive frames showing the motion of pedestrians and vehicles over time. The risk level does not depend only on spatial positions at an isolated instant, but also on the temporal evolution of the scene.

The process begins with the extraction of successive frames from video sequences, followed by automatic pedestrian detection. Each frame is then processed by a CNN2D network to extract a compact visual feature representation. These frame-level feature vectors are arranged into a temporal sequence and provided to an LSTM, which models the temporal dependencies across consecutive frames. The final hidden representation is then passed to a fully connected layer to predict pedestrian–vehicle risk.

The proposed model has three main stages :

1. Frame-level spatial feature extraction using CNN2D :
   - Multiple convolutional layers with ReLU activation and adaptive pooling.
   - The goal is to extract compact visual representations from each frame.

2. Temporal modeling using LSTM :
   - The sequence of CNN-extracted feature vectors is used as input.
   - The LSTM captures temporal dependencies across consecutive frames.

3. Prediction :
   - The final temporal representation is fed into a fully connected layer.
   - The output layer produces the risk prediction.

It should be noted that, in the proposed CNN2D–LSTM architecture, the LSTM receives only the temporal sequence of CNN-extracted visual feature vectors. Handcrafted numerical variables, such as distance or speed, are not fused into this hybrid model. They are used separately only in the standalone LSTM baseline for comparison and for labeling the risk level of the sequences.

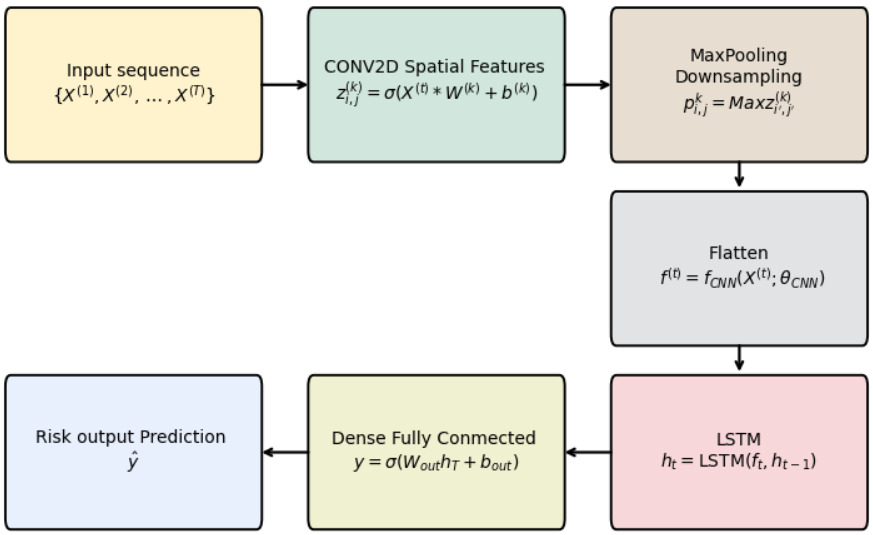


FIGURE 1. CNN2D-LSTM architecture for spatio-temporal risk prediction.

1.2. Convolutional Neural Network (CNN2D) Spatial Feature Extraction

The CNN2D is used to train spatial features from each frame, where with the dimensions being the height, width, and number of channels of the image, respectively. [17]

The output of a convolutional layer for the kk-th filter is given by :

$$z_{i,j}^{(k)} = \sigma\left(\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{c=1}^{C} X_{i+m,j+n,c}^{(t)} W_{m,n,c}^{(k)} + b^{(k)}\right) \tag{1}$$

Where :

— x is the pixel value at position (i, j) in channel c of the input frame,
— w is the convolution kernel weight for the k-th filter,
— b is the bias term,
— $M \times N$ is the filter size,
— $sigma(\cdot)$ is the activation function (commonly ReLU),
— $z_{i,j}$ is the resulting feature map value at location (i, j).

After convolution, a pooling operation reduces the spatial dimensions while retaining the most important features :

$$p_{i,j}^{(k)} = \text{pool}\left(z_{i',j'}^{(k)}\right), \qquad i', j' \in R_{i,j} \tag{2}$$

R : receptive field corresponding to the pooling operation.

Finally, the feature maps are flattened into a vector representing the spatial information extracted from frame t :

$$x_t = f_{\text{CNN}}\left(X^{(t)}; \theta_{\text{cnn}}\right) \tag{3}$$

where denotes the sequence of convolution, activation, pooling, and flattening operations parameterized by θ_{cnn}.

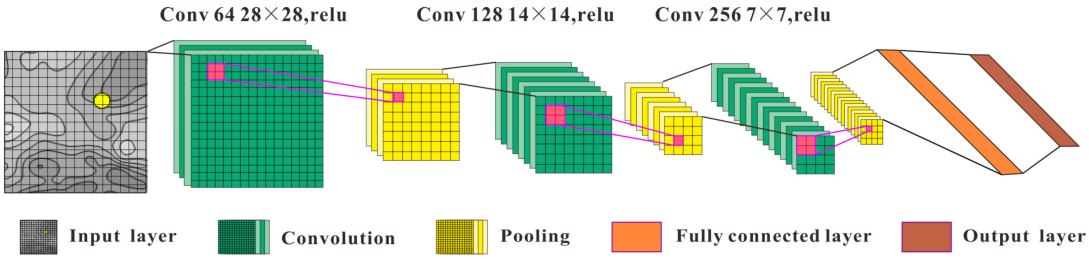


FIGURE 2. Detailed visualization of the CNN feature-extraction pipeline.[17]

1.3. Long Short-Term Memory (LSTM) – Temporal Modeling

The sequence of feature vectors is then fed into the LSTM to capture temporal dependencies. To avoid confusion with the forget gate , we put the CNN feature vector. [18], [19]

The update equations for the LSTM are :

$$C_t = \tanh\left(W_c x_t, h_{t-1} + b_c\right) \tag{1}$$

$$\tilde{C}_t = \sigma\left(W_f x_t, h_{t-1} + b_f\right) \times C_{t-1} + \sigma\left(W_i x_t, h_{t-1} + b_i\right) \times C_t \tag{2}$$

$$h_t = \sigma\left(W_o x_t, h_{t-1} + b_o\right) \times \tanh\left(C_t\right) \tag{3}$$

Where :
— $\tilde{c}_t$ is the candidate cell state,
— c_t is the updated cell state,
— h_t is the hidden state,
— W and b are the gates' weight matrices and biases,
— $\sigma(\cdot)$ is the sigmoid function.
Such integration of CNN and LSTM allows the model concatenate instantaneous spatial perception with temporal evolution in a coherent manner.

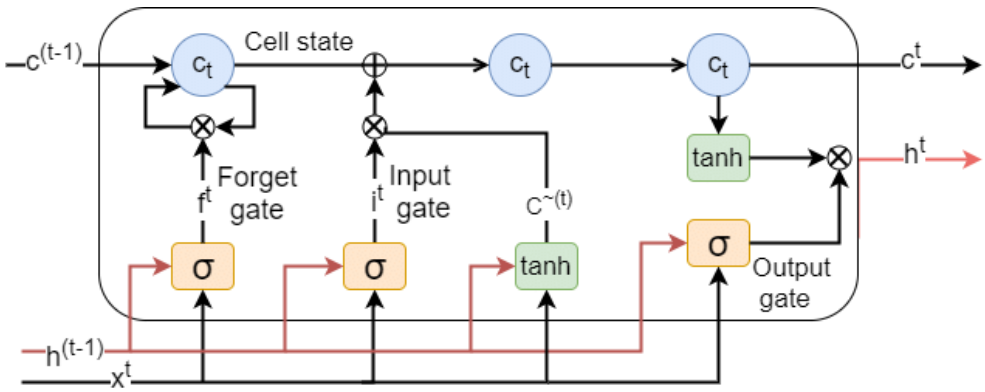


FIGURE 3. Internal structure of an LSTM cell [20]

1.4. Output Layer-Risk Classification

At the final time step, the LSTM hidden state is input into a fully connected for pedestrian-vehicle risk level prediction.

Binary classification (risk or no risk) :

$$y = \sigma\left(W_{out} h_T + b_{out}\right) \tag{4}$$

with binary cross-entropy loss :

$$L = y \log(y) + (1 - y) \log(1 - y) \tag{5}$$

2. Data

In this work, we apply a CNN2D-LSTM model on the JAAD dataset, after evaluating standalone CNN2D and standalone LSTM models to compare performance and identify the best model for capturing risk scenarios. While our evaluation is currently limited to the JAAD dataset and binary risk labels, future work will focus on assessing the generalizability of the model using continuous risk values on additional independent datasets, such as the PIE dataset. Moreover, we plan to analyze the model's performance across different sub-conditions within JAAD, including variations in weather, time of day, and pedestrian occlusion levels, to better understand its strengths and limitations under diverse real-world scenarios.

2.1. Data Preprocessing

For testing our method, we employed the JAAD (Joint Attention in Autonomous Driving) dataset, which comprises 346 annotated videos based on different behavioral characteristics. The JAAD dataset was partitioned at the video level to prevent any overlap of scenes or pedestrians between the training, validation, and test sets. Entire video sequences were randomly assigned to the training (60%), validation (20%), and testing (20%) subsets, and all frames extracted from a given video were kept within the same subset. This strategy ensures temporal independence between splits and avoids data leakage arising from shared contextual or behavioral information across sets.

The naturally imbalanced distribution between risky and non-risky situations was addressed during training by using a weighted binary cross-entropy loss, where class weights were computed exclusively from the training subset. This approach allows the model to account for class imbalance without altering the original data distribution. The proposed CNN2D-LSTM framework was implemented using PyTorch. The model takes RGB video frames as input, which are resized to a fixed resolution prior to training. Frames are sampled uniformly within each video, and temporal sequences are constructed using a sliding window mechanism. The length of the sequences (T) was varied between 5 and 16 frames, and the stride (S) between consecutive sequences was varied between 1 and 4 frames ; both parameters were optimized on the validation set. The CNN backbone extracts spatial features from individual frames, while an LSTM network models temporal dependencies with hidden state sizes ranging from 16 to 64 units. Optimization was performed using the Adam optimizer, with learning rates sampled log-uniformly between 10^{-4} and 10^{-2} and batch sizes of 16, 32, or 64. To handle class imbalance, a weighted binary cross-entropy loss was used.

2.2. Risk Annotation Strategy and Label Construction

Several descriptive variables are appended to every frame, including occlusion (whether the pedestrian is occluded or not), cross (whether the pedestrian crosses or not), look (whether the pedestrian looks or not for traffic), and ego-action (specifying the action of the vehicle : accelerating, decelerating, high speed, low speed, or stop) [1], [21]. We also included the distance between a car and a pedestrian as a second variable in the dataset. This was calculated using automatic vehicle-to-pedestrian detection with YOLOv11, as well as a formula involving the camera focal length and the perceived size of the pedestrian in the picture.

To generate the pedestrian-vehicle risk labels, we employ the YOLOv11 nano model for pedestrian detection in JAAD video frames [22]. YOLOv11 nano is a single-stage detector capable of performing real-time object detection with a lightweight architecture suitable for processing large video sequences. The model was trained for 36 epochs and achieved a mean Average Precision at IoU 0.5 (mAP@50) of 47.0% for pedestrian detection.

Although this mAP reflects a trade-off between computational efficiency and detection accuracy, it is sufficient to reliably extract pedestrian positions in diverse urban contexts, including crowded scenes, varying lighting conditions, and partial occlusions.

For each detected pedestrian, the model provides bounding box coordinates, which are subsequently used in a monocular geometric approach to estimate the distance between the vehicle and pedestrians. This estimation relies on the perspective projection principle :

$$\text{Distance} = \frac{f \cdot L_R}{L_V}$$

where f is the camera focal length in pixels, L_R is the real-world object height, and L_V is the apparent height in the image. This procedure allows robust approximation of the vehicle-pedestrian distance, which is then incorporated as a key variable for risk label generation.

Overall, the combination of YOLOv11 detection and monocular distance estimation provides reliable spatial information, supporting the quality of the automatically generated risk labels used in our study. Future work will include more extensive quantitative validation of the YOLOv11-based distance estimation on JAAD and other datasets, such as PIE, to further confirm the precision and recall of this approach.

To construct the ground-truth labels, a heuristic rule-based annotation strategy was adopted to identify observable configurations associated with potentially dangerous pedestrian-vehicle interactions. This approach was motivated by the absence of explicit accident or near-miss annotations in the JAAD dataset, which necessitates the definition of operational criteria for supervised learning. Ultimately, we developed an annotation rule that classifies each image as either "risk" or "no risk" [23].

These variables were preprocessed and encoded as categorical variables solely to construct the risk annotation rules. The labeling algorithm relies on the following conditions : to label pedestrian-vehicle risk, we considered the pedestrian's state, visual behavior, distance relative to the vehicle, and the vehicle's action. Occlusion was also taken into account, that is, situations where the pedestrian may be partially or fully obscured by other objects, which increases the level of risk. The labeling algorithm operates as follows :

- If the pedestrian is crossing (cross = True) and is looking away from the vehicle (look = False or not looking), and if the distance between the vehicle and the pedestrian is below a critical threshold $(distance_{threshold})$, then :
  - If the vehicle is moving fast (moving-fast) or accelerating (accelerating), the situation is considered high-risk (risk).
  - Otherwise, the situation is considered safe (no risk).
- In all other cases, the situation is labeled as no risk.

The critical distance threshold was chosen based on [14], which demonstrates that situations with a proximity below this threshold are significantly more dangerous.

It should be noted that this labeling approach is heuristic and may not capture all possible real-world risk scenarios. While providing a practical way to generate labels for training, it has limitations : some dangerous situations might not satisfy the rule conditions. These labels serve solely as a tool for model supervision ; the CNN2D-LSTM model does not use these variables directly as input. Future work will explore human-annotated labels to better represent complex risk dynamics in traffic environments.

2.3. Early stopping

To limit overfitting and improve generalization, we employed an early stopping strategy, which consists of halting training when the performance on the validation set does not improve for several consecutive epochs. This approach allows selection of the best-performing model in validation without overtraining. In our case, training was capped at a maximum of 40 epochs, but early stopping after epoch 22, i.e., the validation loss had converged. This approach guaranteed that the final model captures the best trade-off between training error and generalization capability and thus steers clear of overfitting.

2.4. Optuna

For further model performance tuning, we employed Optuna, a library for automated hyperparameter optimization based on Bayesian optimization. This method efficiently traverses the hyperparameter space to identify settings that maximize model performance over the validation set, with minimal search time compared to exhaustive or manual searching. Optuna tries different values of the parameters and iteratively reduces the area under investigation to the most promising regions of the hyperparameter space, enabling us to pinpoint the best configuration of the model.[24]

Hyperparameters	**Optimal value**
Learning rate	2.13
Hidden size	46
Batch size	64
Sequence lenght	6
stride	3

TABLE 1. Hyperparameters and optimal values

3. Results and Discussion

To evaluate pedestrian-vehicle risk detection, three model configurations were compared : an LSTM-only model, a CNN-only model, and the proposed hybrid CNN2D–LSTM model. The standalone LSTM model takes as input the temporal sequence of numerical features describing pedestrian–vehicle interactions, whereas the CNN-only model operates directly on image frames and extracts spatial features only. In the hybrid CNN2D–LSTM model, each frame is first processed by the CNN2D backbone to extract a visual feature vector, and the resulting sequence of feature vectors is then provided to the LSTM for temporal modeling. Table 2 and the corresponding confusion matrices (Figure 4) highlight the strengths and limitations of each approach.

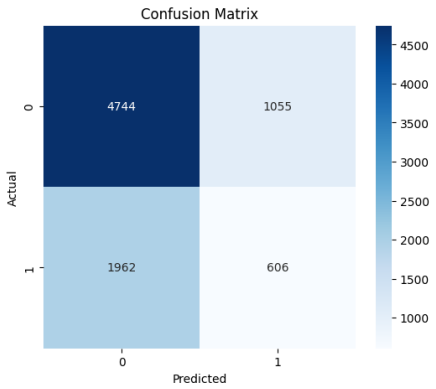
The CNN-only model, which relies exclusively on spatial feature extraction, struggles to accurately identify risky events. This limitation is reflected in its confusion matrix, which shows a high number of false negatives (1962) and false positives (1055). Consequently, the model exhibits low recall (0.26) and precision (0.36), indicating that purely spatial representations are insufficient to capture the temporal dynamics that characterize pedestrian–vehicle interactions.

In contrast, the LSTM-only model, designed to capture temporal dependencies from sequential numerical descriptors, demonstrates significantly improved performance. Although it achieves high precision and recall, its confusion matrix still reveals 153 false negatives. In risk prediction, missing a true risk (false negative) is more critical than issuing a false alarm (false positive), as undetected dangerous situations may lead to severe consequences. Therefore, minimizing false negatives is a primary objective in safety-oriented applications. This limitation motivated the integration of spatial visual modeling to further enhance detection reliability.

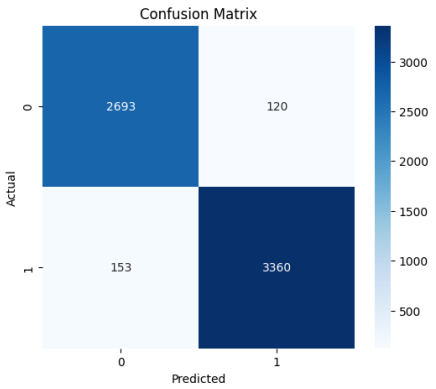
These considerations led to the proposed hybrid CNN2D–LSTM approach, which combines convolutional spatial feature extraction from video frames with temporal sequence modeling of the extracted visual representations. Its confusion matrix indicates substantially reduced misclassifications, with only 214 false positives and 26 false negatives. This reduction in false negatives directly aligns with the risk detection objective. The hybrid model achieves superior overall performance (accuracy : 0.971, precision : 0.922, recall : 0.99, F1-score : 0.955). By jointly leveraging spatial and temporal information, the proposed framework ensures robust detection of risky events, effectively minimizing missed detections while maintaining a controlled level of false alarms, which is essential for proactive pedestrian–vehicle safety systems.

Model	Accuracy	Precision	Recall	F1-score
LSTM-ONLY	0.95	0.96	0.95	0.96
CNN-ONLY	0.64	0.36	0.26	0.29
CNN2D-LSTM	**0.971**	**0.922**	**0.99**	**0.955**

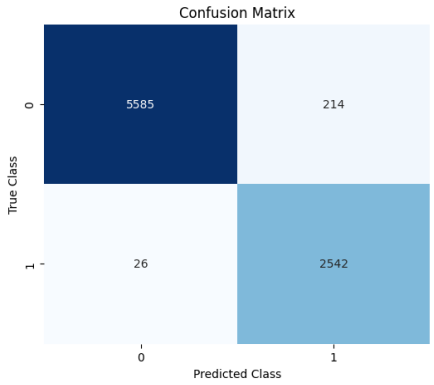
TABLE 2. Performance comparison of CNN2D–LSTM, CNN-only, and LSTM-only models.



(a) CNN2D-only



(b) LSTM-only



(c) Hybrid CNN2D-LSTM

FIGURE 4. Confusion matrices of the evaluated models.

Beyond the comparative evaluation, the proposed CNN2D–LSTM model was analyzed independently to better assess its intrinsic predictive capability. Figure 5 presents the global performance metrics of the model, with an F1-score of 0.976, an accuracy of 0.971, and a recall of 0.99. The high F1-score indicates a well-balanced trade-off between precision and recall, reflecting the model's overall reliability in distinguishing risky from non-risky situations. Similarly, the strong accuracy value confirms the consistency of the classification performance across the entire dataset. More importantly, the very high recall demonstrates the model's effectiveness in detecting nearly all true risk events, which is a critical requirement in safety-oriented pedestrian-vehicle interaction analysis.

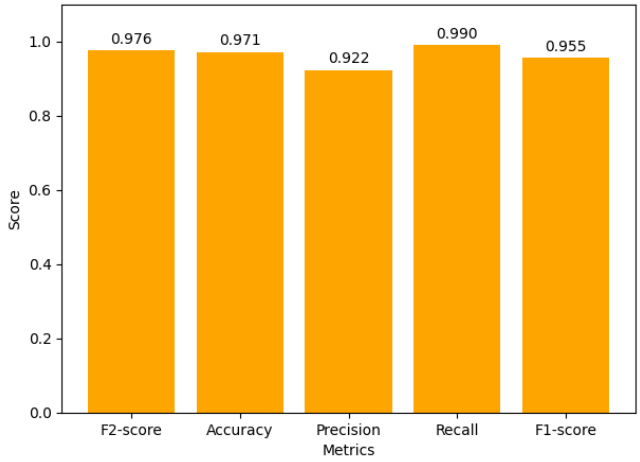


FIGURE 5. Performance metrics and confusion matrix of the proposed CNN2D–LSTM model.

In addition, Table 3 reports the detailed classification performance for each class.

— **Class 0 (non-risk) :** The model achieves very high precision (0.9954), indicating a very low rate of false positives. The recall (0.9631) confirms that most non-risk situations are correctly identified, although a small proportion are misclassified as risk cases.

— **Class 1 (risk) :** The precision is slightly lower (0.9224), while the recall remains extremely high (0.9899), indicating that nearly all truly risky situations are correctly detected. This characteristic is particularly desirable in safety-critical contexts, where failing to detect a risk (false negative) may lead to severe consequences.

Class	Precision	Recall	F1-score	Support
Class 0	0.9954	0.9631	0.9789	5799
Class 1	0.9224	0.9899	0.9549	2568

TABLE 3. Classification performance per class.

These results are particularly significant in imbalanced classification scenarios, where hazardous events are less frequent but more critical. The relatively small difference between the F1-scores of both classes indicates that the model effectively handles class imbalance without compromising overall reliability. More importantly, the performance demonstrates the model's ability to capture subtle contextual cues and adapt to dynamic pedestrian behaviors, such as trajectory changes, proximity to vehicles, and partial occlusions, which are essential for accurate risk assessment.

The remaining misclassifications mainly occur in borderline or ambiguous situations, including complex pedestrian-vehicle interactions or visually challenging scenes. This reflects the intrinsic difficulty of understanding real-world traffic environments and provides insights for potential model improvements.

The discriminative capability of the model is further supported by the ROC and Precision-Recall analyses shown in Figure 6.

The near-perfect ROC-AUC and high average precision values indicate that the model maintains strong sensitivity and specificity across a wide range of decision thresholds. This performance highlights the complementary effect of CNN-based spatial feature extraction and LSTM-based temporal modeling, enabling the system to detect subtle indicators of potential risk in pedestrian motion and vehicle interactions. Furthermore,

stable performance across thresholds demonstrates the robustness and adaptability of the proposed framework under varying safety requirements and tolerance levels for false alarms.

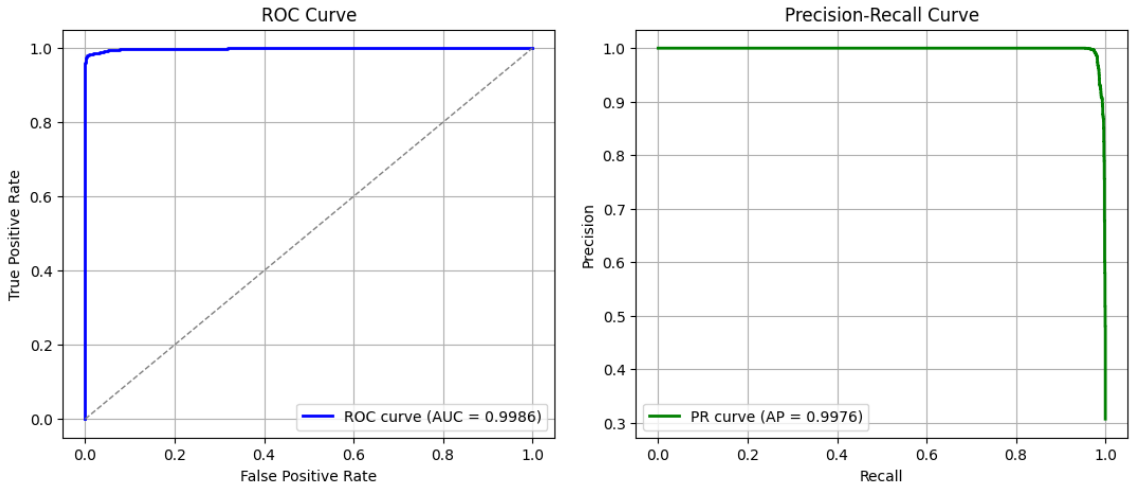


FIGURE 6. ROC and Precision-Recall curves of the proposed model.

Although direct quantitative comparison with state-of-the-art models is not reported in this study, several strong approaches have been evaluated on the JAAD dataset for closely related tasks such as pedestrian crossing intention and crossing action prediction. For example, [3] propose a spatio-temporal attention architecture with feature fusion to improve pedestrian crossing intention prediction on JAAD. More recently, [6] investigate an explainable and interpretable crossing prediction framework combining deep learning with fuzzy logic, and report extensive experimental insights on JAAD/PIE settings. These works demonstrate the effectiveness of spatio-temporal modeling (and, increasingly, attention/interpretability mechanisms) for pedestrian behavior understanding.

To provide context relative to existing approaches, we present an indicative comparison of our CNN2D-LSTM model with state-of-the-art JAAD models evaluated for event risk assessment, including SFGRU (Spatio-Functional GRU), PCPA (Pedestrian-Centric Prediction Architecture), BiPed (Bidirectional Pedestrian Modeling Network), and PedFormer (Pedestrian Transformer-based Model). Our model produces binary risk labels (risk / no risk) using heuristic annotation rules, whereas the referenced models output continuous risk scores. Direct comparison is inherently limited, as discussed in [25], because model performance varies across tasks and metrics, and models trained for intention estimation or action prediction are not directly comparable to binary risk detection.

Model	Accuracy	Precision	F1-score
SFGRU	0.42	0.24	0.22
PCPA	0.37	0.15	0.11
BiPed	0.39	0.21	0.20
PedFormer	0.53	0.43	0.41
CNN2D-LSTM	**0.971**	**0.922**	**0.955**

TABLE 4. Indicative comparison of global CNN2D-LSTM performance with state-of-the-art JAAD-based models.

Despite methodological differences, Table 4 provides an indicative comparison of Accuracy, Precision, and F1-score to contextualize the performance of our CNN2D-LSTM model relative to existing state-of-the-art JAAD-based models. While previous models such as SFGRU, PCPA, BiPed, and PedFormer output continuous risk scores,

our approach relies on heuristic binary labels to detect early-stage pedestrian-vehicle risk. Notably, our CNN2D-LSTM achieves a global Accuracy of 0.971, Precision of 0.922, and F1-score of 0.955, markedly surpassing the compared architectures.

These results indicate that the proposed lightweight spatio-temporal model is highly effective at capturing critical interactions and predicting high-risk situations, even in complex urban scenarios with occlusions, varying illumination, and dynamic backgrounds. The substantial performance gap highlights the advantage of combining 2D convolutional feature extraction with temporal modeling through LSTM, enabling the system to detect emerging risks rapidly while maintaining robustness across diverse conditions.

Overall, this comparison underscores the practical potential of our CNN2D-LSTM framework as a reliable and computationally efficient solution for early-stage pedestrian risk detection, offering a meaningful reference point for future work on interpretable and deployable traffic safety models.

In contrast, the present work targets a complementary formulation-heuristic early-stage risk proxy detection-using a lightweight CNN2D-LSTM design, prioritizing feasibility and spatio-temporal fusion rather than competing directly with specialized intention-prediction architectures.

A qualitative assessment was also conducted through sequence-level visualizations, providing deeper insight into the model's behavior in realistic urban environments. Across a diverse set of non-risk and risk scenarios, the CNN2D-LSTM architecture consistently demonstrates its ability to capture temporal dependencies and evolving pedestrian-vehicle interactions.

For non-risk sequences (Figure 8), the model consistently produces low-risk predictions, demonstrating stability even in the presence of significant visual variability, such as changes in illumination, partial occlusions, diverse scene layouts, and the presence of other dynamic objects like vehicles or cyclists. This indicates that the model is able to reliably distinguish safe interactions from potentially hazardous situations, without being misled by environmental noise or background motion.

Conversely, for risk scenarios (Figure 7), the model successfully detects hazardous situations across a variety of contexts and lighting conditions, assigning high risk probabilities that often exceed 0.8. These results highlight the model's ability to capture different types of pedestrian-vehicle interactions, including challenging cases where visual cues are subtle or partially occluded.

Furthermore, the qualitative results reveal that the model is robust to diverse viewpoints, motion blur, crowded scenes, and dynamic backgrounds. It reliably tracks pedestrian trajectories and contextual cues, such as approach direction, trajectory curvature, vehicle deceleration patterns, and scene geometry. In some sequences, the model anticipates pedestrian crossing behavior before it becomes visually apparent, suggesting a form of implicit understanding of interaction dynamics.

Collectively, these visual analyses complement and reinforce the quantitative findings. They confirm that the proposed system achieves high predictive accuracy while also generalizing to varied urban traffic contexts, including complex intersections, occluded crossings, and rapidly changing scenes. Overall, the CNN2D-LSTM architecture proves to be a strong candidate for real-time pedestrian safety support in intelligent transportation systems, offering both early-warning capabilities and robustness to realistic traffic variability.

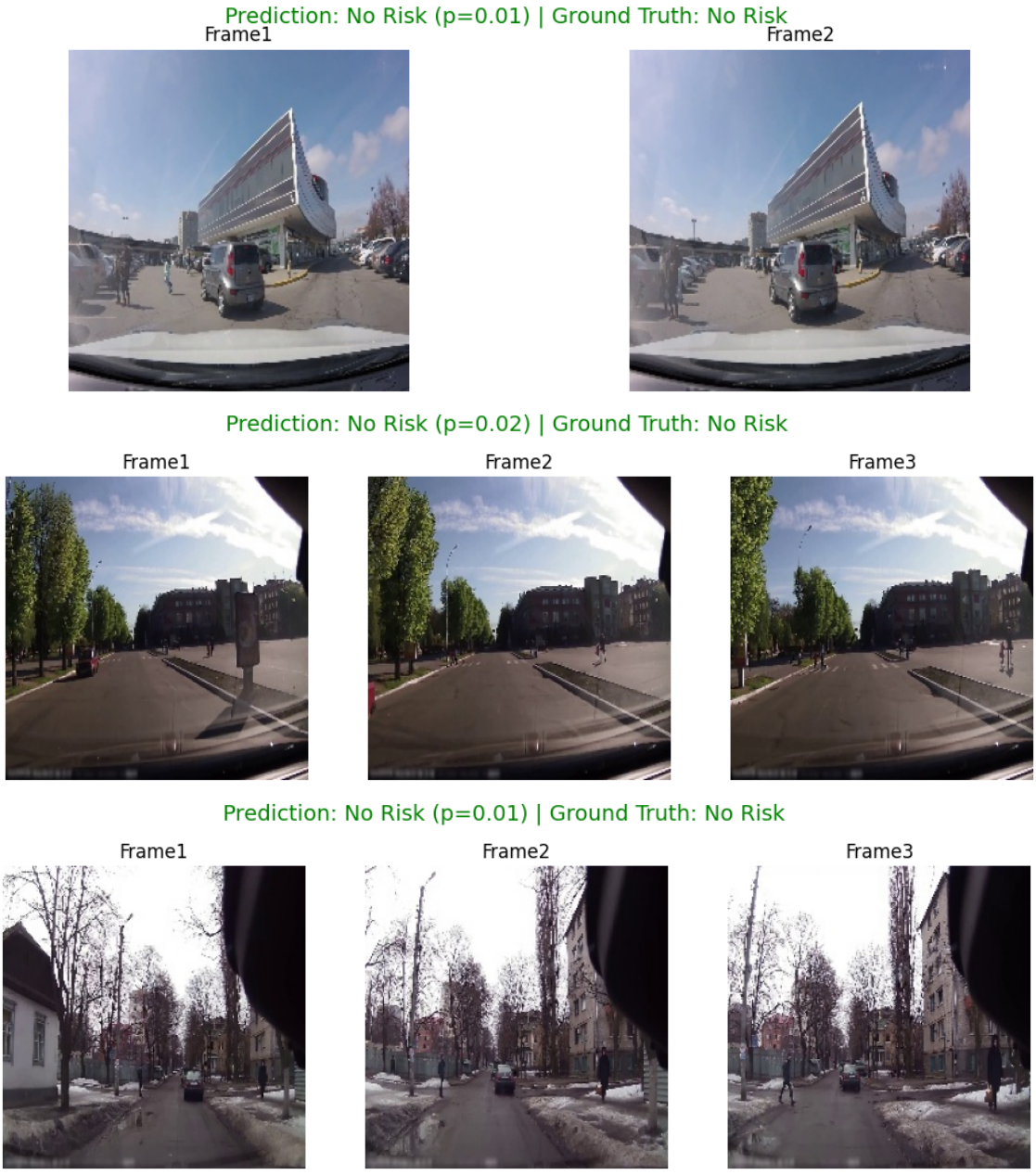


FIGURE 7. Non-Risk sequences where the model correctly produces low risk probabilities.

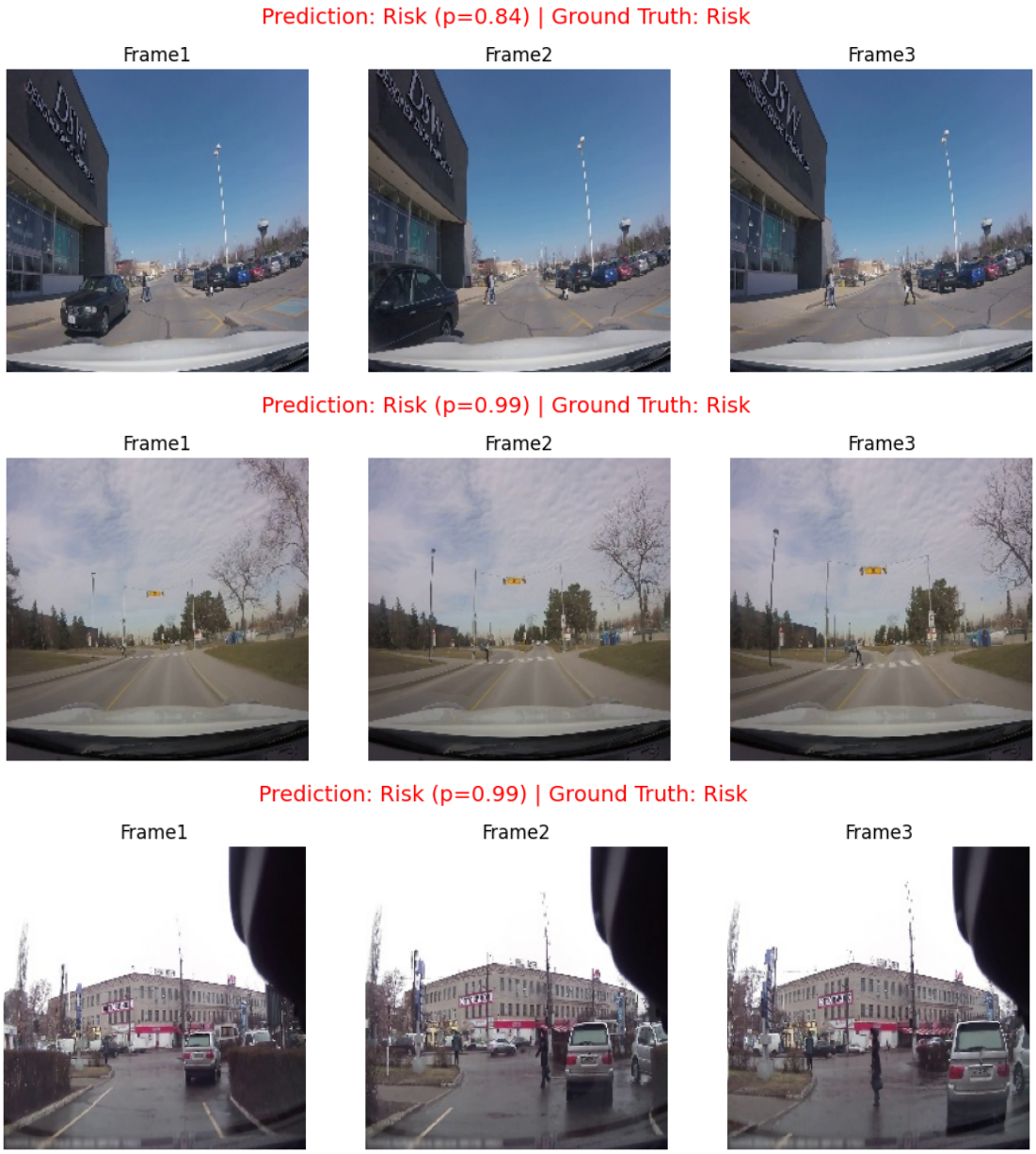


FIGURE 8. Risk sequences where the model correctly outputs high risk probabilities.

On the other hand, Figure 9 highlights false negative cases, in which the proposed model misclassifies hazardous situations as safe. An analysis of these failure cases suggests that the model is particularly challenged by sudden pedestrian appearances within the interaction area. In Figure 9a, the pedestrian is initially occluded by the vehicles ahead and only becomes visible abruptly, making early risk recognition difficult; the model assigns a probability of 0.20 to this situation. In Figure 9b, the vehicle turns near a crosswalk, while the pedestrians are initially outside the camera's field of view and become visible only later in the sequence; here, the model assigns a probability of 0.39. Although these scenarios are classified as "safe" under a binary labeling scheme, the non-negligible probabilities suggest that the concept of binary risk is somewhat limiting and motivate the exploration of continuous risk variables.

FIGURE 9. Risk sequences where the model failed to correctly detect hazardous situations.

A closer examination of these misclassifications reveals a connection to a key methodological limitation of the present study : the heuristic construction of the ground-truth risk labels. While the proposed rule-based framework enables reproducible and interpretable annotation in the absence of explicit accident-level data, it inherently simplifies the multidimensional nature of real-world traffic risk.

Indeed, pedestrian-vehicle risk is a continuous and context-dependent phenomenon influenced by a wide range of factors, including behavioral uncertainty, environmental conditions, vehicle dynamics, and scene geometry. The deterministic labeling rule adopted in this work captures only a subset of critical configurations, primarily those that can be reliably inferred from observable visual cues.

As a consequence, certain genuine high-risk scenarios may not be represented in the training data, particularly situations involving sudden pedestrian motion, ambiguous crossing intentions, or complex vehicle maneuvers. This may lead to a form of definition-induced false negatives, where dangerous interactions are not labeled as risk despite their potential severity. Additionally, although the variables used in the heuristic rule are not explicitly provided as structured inputs, they remain visually encoded in the image data. This creates a potential source of implicit supervision bias, whereby the model may learn to associate specific visual patterns with the hand-engineered labeling logic. Such bias may contribute to high performance within the dataset while limiting generalization to more diverse real-world environments.

These considerations motivate future research directions : developing probabilistic, continuous, or fuzzy logic-based formulations of pedestrian risk, supported by human-annotated safety assessments and multimodal

contextual information. In particular, validating heuristic labels against expert judgment and measuring inter-annotator agreement would provide stronger evidence regarding the reliability of operational risk definitions.

4. Conclusion

This work investigated the feasibility of spatio-temporal deep learning for early-stage pedestrian-vehicle risk detection from video sequences. A CNN2D-LSTM framework was proposed to jointly model spatial visual cues and their temporal evolution in pedestrian-vehicle interactions. Experiments conducted on the JAAD dataset show that the proposed approach achieves an overall accuracy of 97%, with a recall of 98.9% for the risk class and a precision of 92.2%. These results indicate that the model is able to detect potentially hazardous situations with a very low false-negative rate, which is a critical requirement for safety-related applications. Beyond raw performance, the main scientific contribution of this work lies in demonstrating that explicit temporal modeling across short video sequences provides measurable benefits for capturing interaction dynamics that are not accessible through frame-based analysis alone. The results empirically support the relevance of spatio-temporal fusion for heuristic risk proxy detection in complex urban traffic environments.
Future work will focus on integrating fuzzification techniques to better manage uncertainty and improve interpretability, as well as developing a mobile or embedded risk-prevention system capable of real-time detection and early warning for next-generation intelligent mobility.

References

[1] A. Rasouli, I. Kotseruba et J. K. Tsotsos, « It's Not All About Size : On the Role of Data Properties in Pedestrian Detection, » en, in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé et S. Roth, éd., t. 11129, Series Title : Lecture Notes in Computer Science, Cham : Springer International Publishing, 2019, p. 210-225.

[2] S. A. Bouhsain, S. Saadatnejad et A. Alahi, *Pedestrian Intention Prediction : A Multi-task Perspective*, arXiv :2010.10270 [cs], mai 2021.

[3] D. Yang, H. Zhang, E. Yurtsever, K. Redmill et Ü. Özgüner, *Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention*, arXiv :2104.05485 [cs], oct. 2021.

[4] J. Guo, Y. Ding et A. Tian, « Multimodal feature fusion for pedestrian crossing intention prediction based on hybrid attention mechanism, » in *2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)*, juill. 2024, p. 70-74.

[5] S. Ahmed, A. A. Bazi, C. Saha, S. Rajbhandari et M. N. Huda, « Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation, » *Expert Systems with Applications*, t. 225, p. 120 077, sept. 2023.

[6] A. N. Melo, C. Salinas et M. A. Sotelo, *Experimental Insights Towards Explainable and Interpretable Pedestrian Crossing Prediction*, arXiv :2312.02872 [cs], déc. 2023.

[7] A. N. Melo Castillo, C. Salinas Maldonado et M. Á. Sotelo, « Towards Explainable Pedestrian Behavior Prediction : A Neuro-Symbolic Framework for Autonomous Driving, » en, *Applied Sciences*, t. 15, n° 11, p. 6283, jan. 2025, Number : 11.

[8] M. Abdel-Aty, Z. Wang, O. Zheng et A. Abdelraouf, « Advances and Applications of Computer Vision Techniques in Vehicle Trajectory Generation and Surrogate Traffic Safety Indicators, » 2023, Version Number : 2.

[9] Y. Ali, M. M. Haque et F. Mannering, « A Bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics, » en, *Analytic Methods in Accident Research*, t. 38, p. 100 264, juin 2023.

[10] A. ANKUNDA, Y. ALI et M. MOHANTY, « Pedestrian crash risk analysis using extreme value models : New insights and evidence, » en, *Accident Analysis & Prevention*, t. 203, p. 107 633, août 2024.

[11] R. ALMODFER, S. XIONG, Z. FANG, X. KONG et S. ZHENG, « Quantitative analysis of lane-based pedestrian-vehicle conflict at a non-signalized marked crosswalk, » en, *Transportation Research Part F : Traffic Psychology and Behaviour*, t. 42, p. 468-478, oct. 2016.

[12] C. MITRAKAS, A. XANTHOPOULOS et D. KOULOURIOTIS, « Techniques and Models for Addressing Occupational Risk Using Fuzzy Logic, Neural Networks, Machine Learning, and Genetic Algorithms : A Review and Meta-Analysis, » en, *Applied Sciences*, t. 15, n° 4, p. 1909, jan. 2025.

[13] O. BENKHADDA et M. MANDAR, « Modeling Risk Exposure : Fuzzy and Fuzzy Intuitionistic Approaches to Pedestrian and Vehicle Interaction, » *International Journal of Computing*, p. 155-162, mars 2025.

[14] O. BENKHADDA, M. MANDAR et N. MELLOULI, « Distance Perception Dynamics Between Pedestrians and Vehicles : Adaptation of Gibson's Theory, » en, in *Advances on Intelligent Computing and Data Science II*, F. SAEED, F. MOHAMMED, E. MOHAMMED, S. BASURRA et M. AL-SAREM, éd., t. 255, Series Title : Lecture Notes on Data Engineering and Communications Technologies, Cham : Springer Nature Switzerland, 2025, p. 166-177.

[15] X. CAI, Z. LI, W. QIAO, X. CHENG, B. PENG et D. ZHANG, « Research on Road Traffic Safety Risk Assessment Based on the Data of Radar Video Integrated Sensors, » en, *Promet - Traffic&Transportation*, t. 37, n° 2, p. 523-545, mars 2025.

[16] K. BLACHUT et T. KRYJAK, « Real-Time Efficient FPGA Implementation of the Multi-Scale Lucas-Kanade and Horn-Schunck Optical Flow Algorithms for a 4K Video Stream, » en, *Sensors*, t. 22, n° 13, p. 5017, juill. 2022.

[17] K. DING, L. XUE, X. RAN, J. WANG et Q. YAN, « CNN2D-SENet-Based Prospecting Prediction Method : A Case Study from the Cu Deposits in the Zhunuo Mineral Concentrate Area in Tibet, » en, *Minerals*, t. 13, n° 6, p. 730, mai 2023.

[18] F. A. GERS, J. SCHMIDHUBER et F. CUMMINS, « Learning to Forget : Continual Prediction with LSTM, » en, *Neural Computation*, t. 12, n° 10, p. 2451-2471, oct. 2000.

[19] S. HOCHREITER et J. SCHMIDHUBER, « Long Short-Term Memory, » en, *Neural Computation*, t. 9, n° 8, p. 1735-1780, nov. 1997.

[20] C. KIŞMIROĞLU et O. ISIK, « Temperature Prediction Using Transformer–LSTM Deep Learning Models and Sarimax from a Signal Processing Perspective, » en, *Applied Sciences*, t. 15, n° 17, p. 9372, août 2025.

[21] A. RASOULI, I. KOTSERUBA et J. K. TSOTSOS, « Are They Going to Cross ? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior, » in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy : IEEE, oct. 2017, p. 206-213.

[22] G. JOCHER et J. QIU, *Ultralytics YOLO11*, 2024.

[23] A. ROSEBROCK, *Find distance from camera to object/marker using Python and OpenCV*, jan. 2015.

[24] T. AKIBA, S. SANO, T. YANASE, T. OHTA et M. KOYAMA, « Optuna : A Next-generation Hyperparameter Optimization Framework, » en, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA : ACM, juill. 2019, p. 2623-2631.

[25] A. RASOULI et I. KOTSERUBA, *Diving Deeper Into Pedestrian Behavior Understanding : Intention Estimation, Action Prediction, and Event Risk Assessment*, Version Number : 1, 2024.