INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 1, PP. 1–10
Manuscript received December 18, 2025; revised February, 2026. doi: 10.24425/ijet.2026.157869

Context-aware uncertainty modeling for pedestrian intention detection in urban environments

Yusuf Yesilyurt, and Marek Woda

***Abstract*—The present study investigates the application of uncertainty modelling for the purpose of detecting pedestrian intentions in contexts pertaining to autonomous driving. The proposed framework integrates two mechanisms: threshold-modulation networks for aleatoric uncertainty and cost-sensitive learning for risk-aware decision making.**

Experiments on the PIE dataset with ResNet50, VGG16, and AlexNet demonstrate that cost-sensitive learning enhances F1-scores marginally (0.05-0.58 percentage points) by prioritising recall for crossing pedestrians. ResNet50 demonstrates the strongest performance (98.30% accuracy, 96.35% F1-score), significantly outperforming more elementary architectures. Threshold networks have been observed to introduce computational overhead, resulting in approximately a doubling of training time, accompanied by slight performance reductions.

The study provides empirical evidence for the trade-offs between uncertainty modelling complexity and classification performance in pedestrian intention detection, offering insights for designing safety-oriented perception systems with appropriate computational constraints.

***Keywords*—Pedestrian intention detection; uncertainty modelling; aleatoric uncertainty; epistemic uncertainty; cost-sensitive learning; autonomous driving; PIE dataset**

I. Introduction

THE ability to anticipate the intentions of pedestrians represents a seminal challenge in the development of safe autonomous driving systems. It has been demonstrated that human drivers utilise subtle contextual cues, including but not limited to body posture, gaze direction and movement patterns, in order to predict whether a pedestrian is about to cross the road. However, the translation of this intuitive reasoning into a reliable computer vision framework remains an open research problem, particularly given the inherent uncertainty of complex urban environments. It is imperative to acknowledge the potential for significant disruptions to the reliability and confidence of models in safety-critical decision-making processes. Such disruptions can be attributed to rapid changes in lighting, the presence of obstructions, variations in pedestrian behaviour, and sensor noise.

Recent deep learning–based pedestrian intention detection (PID) systems primarily employ convolutional and recurrent architectures to map sequences of video frames onto discrete crossing decisions. Whilst these models achieve classification accuracies that can be considered impressive on benchmark datasets, they typically produce outputs that are deterministic and lack information regarding prediction confidence. In the context of autonomous driving, the presence of uncalibrated probabilities can lead to a situation where overconfident misclassifications occur, which can potentially result in dangerous actions being taken. Consequently, incorporating uncertainty estimation mechanisms into PID frameworks is a crucial step towards trustworthy perception.

The present study introduces a *context-aware uncertainty modelling framework* for predicting pedestrian behaviour in urban traffic scenarios. The proposed system unifies two complementary perspectives on uncertainty:

- **Aleatoric uncertainty**, which represents variability in observations, is addressed through a *threshold-modulation network*. This network dynamically adjusts the decision boundary according to contextual cues, such as vehicle speed, pedestrian gaze and gestures, the presence of crosswalks, the state of traffic lights, and the level of occlusion.
- **Epistemic uncertainty**, which originates from model limitations or data scarcity, is mitigated through a *cost-sensitive learning scheme* that penalises false negatives more severely than false positives. This encourages risk-aware behaviour in ambiguous cases.

Contrary to the focus of preceding studies, which chiefly concentrated on the evaluation of prediction accuracy, this study adopts a *design-oriented perspective*, emphasising the operational integration of uncertainty into conventional deep-learning pipelines. The framework has been validated using the publicly available Pedestrian Intention Estimation (PIE) dataset, which provides a wealth of multimodal context, including ego-vehicle signals, bounding-box annotations and environmental attributes. In this study, three convolutional neural network architectures – ResNet50, VGG16 and AlexNet – are adapted to this framework in order to systematically assess its influence on model calibration, robustness and computational cost.

The main contributions of this paper can be summarised as follows:

1) A unified architecture that combines a context-aware thresholding module with a cost-sensitive training formulation in order to capture both aleatoric and epistemic uncertainties simultaneously.

Y. Yesilyurt (e-mail: yusuf.yesilyurt.ml@gmail.com).
M. Woda is with Wrocław University of Science and Technology, Wrocław, Poland (e-mail: marek.woda@pwr.edu.pl).

2) A detailed pre-processing strategy for multi-modal data fusion using visual and contextual features from the PIE dataset.
3) An implementation-level evaluation of the effects of uncertainty integration on accuracy and F1-score across multiple network backbones;
4) Practical insights are provided for deploying uncertainty-aware PID systems, highlighting the trade-offs between reliability, computational complexity and safety performance.

These contributions shift the focus of the study from achieving high accuracy to developing calibrated, interpretable and safety-aligned decision systems for autonomous vehicles. The findings presented herein are intended to inform both academic research and the practical deployment of frameworks for detecting pedestrian intentions that are aware of uncertainty.

II. Related Work and Background

Pedestrian intention detection (PID) has evolved from traditional vision-based behaviour classification to become a multifaceted learning problem integrating spatial, temporal and contextual reasoning. The relevant literature on this topic can be categorised into three themes: (i) vision-based intention prediction; (ii) uncertainty modelling in deep learning; and (iii) model calibration and risk-aware inference. Together, these strands form the conceptual basis of the proposed framework.

A. Vision-Based Pedestrian Intention Prediction

In the early days of PID studies, the focus was on hand-crafted features such as optical flow, pose trajectories and scene geometry. These features were used to infer motion intent [1], [2]. While these methods were capable of interpretation, they were limited in their ability to capture high-level semantic cues such as gaze direction or interaction with vehicles. The advent of deep convolutional networks signified a major shift towards data-driven learning, with subsequent research introducing multi-stream and temporal architectures that enable joint modelling of visual and contextual information.

A plethora of public datasets, including *Joint Attention for Autonomous Driving* (JAAD) [3] and *Pedestrian Intention Estimation* (PIE) [4] have emerged as instrumental resources in the field, offering extensive, annotated benchmarks encompassing both pedestrian behaviours and vehicle dynamics. In the field of multimodal learning, PIE has been identified as a particularly valuable system due to its ability to synchronise video, bounding boxes, GPS, and on-board diagnostics. Nevertheless, the dataset exhibits significant class imbalance and natural ambiguity in labels – two factors that motivate the inclusion of uncertainty modelling in this work.

Recent research in the field of PID has investigated a range of advanced techniques, including attention mechanisms, graph-based reasoning, and trajectory forecasting. Whilst the efficacy of these approaches is evident, they generally output deterministic probability scores without explicit confidence estimation, which limits interpretability and reliability in safety-critical deployment.

B. Uncertainty Modeling in Deep Neural Networks

The quantification of uncertainty in deep learning has become an active area of research. Foundational distinctions have been made between *aleatoric* uncertainty, which stems from noise in the observations, and *epistemic* uncertainty, which arises from limited knowledge of model parameters [5]–[7]. Aleatoric uncertainty is frequently modelled via heteroscedastic likelihoods or auxiliary branches that learn data-dependent variance. The concept of epistemic uncertainty can be captured through Bayesian approximations, including Monte-Carlo dropout [8], deep ensembles [9], or evidential networks [10]. Monte-Carlo dropout provides uncertainty estimates by performing multiple stochastic forward passes during inference, while deep ensembles aggregate predictions from independently trained networks to capture model disagreement. Evidential learning offers a single-pass alternative by placing priors over predictive distributions. While these approaches have demonstrated effectiveness in various domains, their application to pedestrian intention detection remains limited, motivating the exploration of computationally lighter alternatives such as the cost-sensitive framework proposed in this study.

In the domain of autonomous driving, uncertainty estimation has been integrated into various perception tasks [7]. Nevertheless, the explicit integration of both uncertainty types within pedestrian intention detection remains limited. Existing PID works rarely adjust decision boundaries or training objectives based on uncertainty cues, resulting in a significant gap between theoretical advances and applied safety systems. The present study addresses this gap by embedding uncertainty handling directly into both the architecture (through context-aware thresholding) and the loss function (via cost-sensitive training).

C. Summary and Identified Research Gap

A review of the extant literature reveals three observations that are converging. Firstly, contemporary PID models attain high nominal accuracy; however, they frequently neglect uncertainty quantification and calibration. Secondly, while uncertainty estimation techniques have been well established in other domains, they have not been systematically adapted to the pedestrian intention context, particularly where multimodal cues influence decision thresholds. Thirdly, safety-critical applications necessitate mechanisms that integrate uncertainty estimation with asymmetric risk management.

Drawing upon these insights, this paper proposes a unified framework that operationalises uncertainty in both architectural design and training dynamics. The integration of contextual threshold modulation for aleatoric uncertainty and cost-sensitive optimization for epistemic uncertainty serves to bridge the theoretical reliability principles with the practical implementation in real-world pedestrian intention prediction.

III. Dataset and Pre-processing

The experiments in this study are conducted using the *Pedestrian Intention Estimation (PIE)* dataset, which is a large-scale benchmark explicitly designed for studying pedestrian–vehicle interactions in real urban environments [4]. PIE provides high-resolution video sequences recorded from an ego-vehicle, synchronised with GPS and on-board diagnostics (OBD) data, and enriched with detailed pedestrian annotations. Each annotated instance comprises bounding-box coordinates, behavioural labels, and contextual descriptors such as pedestrian gaze, head orientation, gesture, traffic light status, and crosswalk presence. These features enable comprehensive modelling of pedestrian intention as a function of both visual appearance and surrounding context.

A. Dataset Composition

PIE comprises over six hours of driving footage, with 911,000 video frames, of which 293,000 are annotated with pedestrian information. The dataset under consideration contains approximately 1,800 unique pedestrian tracks with 740,000 bounding-box annotations. Pedestrian samples are categorised into two primary intention states: The terms *crossing* and *not crossing* are employed to denote the presence or absence of a particular phenomenon. Pedestrians are tracked over time, thereby providing temporal coherence, which is an essential component of sequential modelling. The dataset also contains scene-level information, including lane topology, weather, and illumination conditions, offering opportunities for multi-domain learning.

The dataset under scrutiny is characterised by a marked class imbalance, with instances that do not intersect significantly outnumbering those that do. This imbalance is indicative of the underlying statistical distribution of traffic, yet it has the capacity to influence conventional classifiers, favouring the majority class. Consequently, this can lead to an elevated prevalence of false-negative outcomes. Consequently, the learning pipeline integrates cost-sensitive weighting strategies to preserve recall for the minority (crossing) class.

B. Data Partitioning

In order to guarantee impartial evaluation and circumvent the occurrence of overlap between training and testing identities, pedestrians are segmented in accordance with the established PIE protocol [4]. Pedestrians who appear entirely within a given split are included in the study to avoid data leakage. During the preprocessing stage, successive image frames pertaining to each pedestrian are grouped into temporal snippets comprising ten frames. This approach captures short-term motion patterns while remaining computationally tractable.

C. Image Preprocessing

It is imperative to resize all video frames to 224×224 pixels, in order to ensure compatibility with ImageNet-pretrained convolutional backbones. Normalization is achieved through the application of ImageNet channel statistics (mean subtraction of [0.485, 0.456, 0.406] and standard deviation scaling of [0.229, 0.224, 0.225] for RGB channels), facilitating the transfer of pretrained weights without distortion. Each pedestrian instance is subject to cropping based on bounding-box coordinates and temporally ordered to form an image sequence tensor of shape $(10, 3, 224, 224)$ for subsequent input to the network.

D. Contextual Feature Extraction

A defining characteristic of PIE is the availability of contextual cues that extend beyond visual appearance. For each temporal sequence, the following contextual variables are extracted and temporally aligned with pedestrian frames:

- **Ego-vehicle dynamics:** OBD speed, GPS speed, heading angle, and gyroscope readings;
- **Environmental semantics:** crosswalk availability, traffic light state and type, and traffic sign presence;
- **Pedestrian attributes:** actions, head orientation, gaze direction, hand gesture, and occlusion ratio.

Each categorical variable is encoded using one-hot encoding, while continuous features such as speed are standardized to have a zero mean and unit variance. The resulting context vector is then concatenated to the high-level feature representation extracted from the visual backbone. This fusion enables the subsequent threshold-modulation network (introduced in Section IV) to adapt decision boundaries based on scene-specific variability.

E. Sequence Labeling and Temporal Alignment

Each temporal sequence is assigned a binary intention label (crossing or not crossing) based on the annotations provided in the PIE dataset. The resulting dataset provides temporally consistent input–output pairs, with each pair consisting of an image-sequence tensor, a synchronised context vector, and a binary intention label.

F. Challenges and Considerations

The dataset under consideration poses two practical challenges. Firstly, a significant proportion of pedestrians are partially occluded or appear at a reduced scale, thereby increasing aleatoric uncertainty in the visual domain. Secondly, it should be noted that certain contextual variables (e.g. gaze or gesture) are occasionally absent due to annotation gaps. The management of missing attributes is achieved through the utilisation of two methodologies: the initialisation of default values and the exclusion of samples when critical information is unavailable. The efficacy of these preprocessing steps is predicated on their ability to ensure that the learning pipeline receives a complete yet realistic representation of the urban scene

The proposed preprocessing pipeline standardises multimodal inputs and aligns temporal, contextual and visual dimensions into a coherent representation, thereby establishing the foundation for the uncertainty-aware framework that will be presented in the following section.

IV. Framework and Methods

The proposed framework introduces uncertainty modelling into pedestrian intention detection (PID) through a three-phase experimental design: The following three steps are to be taken in order to achieve the desired result: firstly, baseline CNN models must be established; secondly, context-aware threshold networks must be integrated to capture aleatoric uncertainty; and thirdly, cost-sensitive learning must be incorporated to account for epistemic uncertainty. This progressive approach facilitates a systematic evaluation of each uncertainty modelling component.

A. System Overview

The framework utilises three convolutional neural networks as visual backbones: ResNet50, VGG16, and AlexNet. Each of these is pretrained on ImageNet and fine-tuned on the PIE dataset. Each backbone is responsible for processing a 10-frame image sequence in order to extract high-level spatiotemporal features. Concurrently, contextual information, encompassing vehicle speed, pedestrian attributes, and environmental cues, is encoded and integrated with visual features to inform the decision process.

B. Phase 1: Baseline Models

For an input image sequence $\mathbf{I} = \{I_1, I_2, \ldots, I_T\}$ with $T = 10$ frames, each CNN backbone extracts spatial features and produces a binary classification output through a softmax layer:

$$p_{cross} = \frac{e^{z_{cross}}}{e^{z_{cross}} + e^{z_{not}}}, \tag{1}$$

where z_{cross} and z_{not} denote the final-layer logits. The baseline models are trained using standard cross-entropy loss:

$$\mathcal{L}_{ce} = -y \log(p_{cross}) - (1 - y) \log(p_{not}), \tag{2}$$

where $y \in \{0, 1\}$ is the ground truth label. These baseline results establish reference performance metrics for subsequent uncertainty modeling phases.

C. Phase 2: Threshold Networks for Aleatoric Uncertainty

Aleatoric uncertainty is attributable to inherent variability in observations, including but not limited to lighting changes, occlusions, and ambiguous pedestrian behaviours. In order to model this uncertainty, threshold networks are integrated as auxiliary modules that process contextual features to generate adaptive decision boundaries.

The threshold network is responsible for the processing of contextual information, including:

- Vehicle OBD speed and dynamics
- Pedestrian gaze direction, gestures, and actions
- Traffic light state and crosswalk presence
- Occlusion levels

These contextual cues are encoded into a feature vector that informs the threshold network's output. The network is implemented as a small multilayer perceptron that adjusts the model's confidence requirements based on scene-specific conditions. During the training phase, the threshold network is optimised in conjunction with the base CNN, thereby facilitating the acquisition of skills that enable the identification of scenarios necessitating more conservative or confident predictions.

D. Phase 3: Cost-Sensitive Learning for Epistemic Uncertainty

Epistemic uncertainty is defined as the limitations in model knowledge due to insufficient data or parameter uncertainty. In the context of pedestrian intention detection, the failure to predict an actual crossing, known as false negatives, poses a significant safety hazard. In order to address this issue, the third phase incorporates cost-sensitive learning.

The cost-sensitive approach involves the modification of the training objective through the implementation of higher penalties for false negatives in comparison to false positives. This is achieved by assigning a greater numerical value to the loss function, thereby ensuring that the recall for the crossing class is prioritised.

$$\mathcal{L}_{cs} = -\alpha\, y \log(p_{cross}) - \beta\, (1 - y) \log(p_{not}), \tag{3}$$

where $\alpha > \beta$ enforces stricter penalties for missed crossing predictions. This asymmetric loss encourages the model to adopt risk-aware behavior, erring on the side of caution when uncertainty is high.

E. Training Configuration

All models undergo training for 10 epochs using the Adam optimiser with an initial learning rate of 10^{-3} and batch size of 32. Each phase is built upon the preceding one:

- **Phase 1:** CNNs were initially trained using a standard cross-entropy loss function.
- **Phase 2:** Threshold networks were incorporated and trained in conjunction with CNNs.
- **Phase 3:** The application of cost-sensitive loss to models from Phase 2 is imperative.

This progressive integration facilitates a systematic evaluation of the influence of each uncertainty modelling component on prediction accuracy, F1-score, and computational requirements.

F. Interpretation and Model Behavior

The three-phase design offers insights into the various aspects of uncertainty present in PID systems. Threshold networks facilitate the adjustment of confidence in a context-dependent manner, while cost-sensitive learning encodes safety priorities directly into the optimization objective. Collectively, these mechanisms transform standard CNNs into uncertainty-aware systems that are better suited to safety-critical autonomous driving applications.

V. Implementation and Experimental Setup

The implementation follows a three-phase experimental design to systematically evaluate the impact of uncertainty modelling on pedestrian intention detection. The following section delineates the architectural choices, training protocol, evaluation metrics, and computational considerations.

A. Network Architectures

Three convolutional neural network architectures serve as visual backbones: ResNet50, VGG16, and AlexNet. Each network is initialized with ImageNet-pretrained weights and fine-tuned on the PIE dataset. All models process temporal sequences of 10 consecutive frames, with each frame resized to 224×224 pixels and normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], standard deviation: [0.229, 0.224, 0.225] for RGB channels).

ResNet50 [11] employs residual connections across 50 layers, thereby facilitating effective gradient flow and robust feature extraction for complex pedestrian behaviours.

VGG16 [12] employs a uniform architecture comprising 16 layers, each consisting of 3x3 convolutional filters, providing a straightforward baseline for comparison.

AlexNet [13] is a model of reduced complexity, with eight layers, and thus offers a lower computational burden as a baseline.

B. Training Configuration

All models are trained using the Adam optimiser, with an initial learning rate of 1×10^{-3} and a batch size of 32. In the interest of avoiding overfitting, the training process is constrained to a maximum of 10 epochs. The models have been implemented in PyTorch and trained on GPU-accelerated hardware.

In the context of this study, contextual features – including vehicle speed, pedestrian attributes (gaze, gesture, actions), environmental conditions (e.g. crosswalk presence, traffic light state) and occlusion levels – are extracted from the PIE dataset annotations and integrated with visual features during training.

C. Evaluation Metrics

The evaluation of model performance employs two primary metrics:

Accuracy is defined as the proportion of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

F1-Score is a metric that balances precision and recall, thus providing a more robust measure given the class imbalance present in the PIE dataset.

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

The metrics are computed on the test set for each experimental phase, thus enabling a systematic comparison of baseline and uncertainty-aware models.

D. Experimental Phases

The experiments are organised into three sequential phases:

Phase 1. The initial phase of the project involved the creation of baseline models. The three CNN architectures are trained on the PIE dataset using standard cross-entropy loss. This phase establishes reference performance metrics for pedestrian intention detection without uncertainty modelling.

Phase 2. The second phase of the process is Threshold Network Integration. Threshold networks are integrated as auxiliary modules that process contextual information to generate adaptive decision boundaries. These networks are trained in conjunction with the base CNNs to model aleatoric uncertainty arising from data variability. The threshold networks adjust the model's confidence requirements based on contextual cues such as vehicle speed, pedestrian behaviour, and environmental conditions.

Phase 3. The third phase of the process is cost-sensitive learning integration. Cost-sensitive learning is a methodology employed to address epistemic uncertainty by means of penalising false negatives more severely than false positives. This phase involves modifying the loss function to prioritise recall for the crossing class, thereby encouraging risk-aware predictions in safety-critical scenarios.

Each phase is built upon the preceding one, thereby enabling a systematic evaluation of how each uncertainty modelling component affects prediction accuracy, F1-score, and computational requirements.

E. Training Times and Computational Considerations

Training times for ResNet50 across the three phases are approximately:

- Phase 1 (Baseline): 122 minutes
- Phase 2 (Threshold Networks): 205 minutes
- Phase 3 (Cost-Sensitive Learning): 248 minutes

A comparison of VGG16 and AlexNet reveals analogous trends, with training times rising in proportion to the incorporation of uncertainty modelling components. The computational overhead is indicative of the augmented complexity engendered by threshold networks and modified loss functions.

The experimental configuration provides a controlled framework for the evaluation of uncertainty modelling in pedestrian intention detection. The three-phase design facilitates a systematic analysis of the influence of threshold networks and cost-sensitive learning on model performance, robustness, and computational efficiency. The subsequent section is devoted to the presentation of the empirical results obtained from these experiments.

VI. Results and Analysis

This section presents the empirical findings from the three experimental phases described in Section V. The analysis focuses on two key performance indicators: classification performance and computational efficiency. It is important to note that all results presented herein have been obtained from the test set of the PIE dataset.

A. Quantitative Performance

As illustrated in Table I, the core metrics of accuracy and F1-score are summarised across the three model backbones and experimental phases. ResNet50 has been shown to outperform VGG16 and AlexNet on the PIE dataset, demonstrating its superior representational capacity for pedestrian intention detection.

TABLE I
CLASSIFICATION PERFORMANCE ACROSS EXPERIMENTAL PHASES

Model	**Phase 1 (Baseline)**	**Phase 2 (Threshold)**	**Phase 3 (Cost-Sens.)**
	Accuracy (%) / F1-Score (%)		
ResNet50	98.23 / 96.30	98.20 / 96.25	**98.30 / 96.35**
VGG16	69.60 / 67.54	69.50 / 67.40	**69.65 / 67.60**
AlexNet	68.39 / 67.13	68.10 / 66.90	**68.40 / 67.20**

B. Performance Analysis Across Phases

Phase 1 – Baseline Models. The baseline models establish reference performance for pedestrian intention detection without uncertainty modeling. ResNet50 achieves the highest accuracy (98.23%) and F1-score (96.30%), while VGG16 and AlexNet achieve more modest performance around 69% accuracy.

The efficacy of ResNet50 can be attributed to its advanced architecture and residual connections, which facilitate the extraction of features from the PIE dataset's intricate urban scenarios with greater efficiency.

Phase 2 – Threshold Network Integration.

The integration of threshold networks for aleatoric uncertainty modelling has been demonstrated to result in a slight decrease in accuracy across all models. ResNet50's accuracy has been observed to decrease marginally from 98.23% to 98.20%, while its F1-score has decreased from 96.30% to 96.25%. Analogous trends are observed for VGG16 and AlexNet. This decline in performance can be attributed to the augmented complexity engendered by the threshold-modulation mechanism, in conjunction with the challenge of concurrently optimising the threshold network and the base classifier.

Notwithstanding the slight numerical decrease, threshold networks provide value by enabling context-dependent decision boundaries. The models have the capacity to adapt their confidence requirements in accordance with environmental conditions, thereby potentially enhancing robustness in ambiguous scenarios. However, it should be noted that this behaviour is not fully captured by aggregate accuracy metrics.

Phase 3 – Cost-Sensitive Learning Integration. The incorporation of cost-sensitive learning to address epistemic uncertainty has been demonstrated to yield performance improvements across all models. ResNet50 demonstrates a 98.30% accuracy rate and an 96.35% F1-score, thus surpassing both the baseline and Phase 2 results. VGG16 exhibits an enhancement in accuracy, attaining 69.65%, while AlexNet achieves 68.40%. The F1-score for VGG16 is 67.60%, indicating a slight decline in performance metrics when compared to AlexNet, which attains 67.20% accuracy and an F1-score of 67.60%.

The enhancements are most evident in the F1-score, suggesting that cost-sensitive learning effectively enhances the model's capacity to accurately identify crossing pedestrians. By imposing a greater penalty on false negative outcomes than on false positive outcomes, the cost-sensitive loss motivates the model to prioritise recall for the crossing class, a critical aspect for ensuring safety in autonomous driving applications.

C. Model Architecture Comparison

Across all three phases, ResNet50 demonstrates substantially superior performance in comparison to VGG16 and AlexNet. The performance disparity (approximately 28-30 percentage points in accuracy) remains consistent across phases, suggesting that architectural depth and residual connections provide fundamental advantages for this task, irrespective of the uncertainty modelling approach employed.

The VGG16 and AlexNet models demonstrate comparable performance across the experimental trials, with VGG16 exhibiting a marginal superiority of 1-2 percentage points. It is evident that both less complex architectures encounter challenges in processing the intricacy inherent in the PIE dataset. This underscores the significance of employing more intricate networks for the purpose of accurately capturing the nuanced visual and contextual cues indispensable for effective pedestrian intention prediction.

D. Computational Overhead

As illustrated in Table II, the training times for the ResNet50 backbone are documented across the three experimental phases. The duration of training is shown to increase in a progressive manner as components pertaining to uncertainty modelling are incorporated, thereby reflecting the augmented computational intricacy of threshold networks and modified loss functions.

TABLE II
TRAINING TIMES FOR RESNET50 BACKBONE ACROSS PHASES

Experimental Phase	**Training Time [min]**
Phase 1 – Baseline	122
Phase 2 – Threshold Networks	205
Phase 3 – Cost-Sensitive Learning	248

The threshold network integration (Phase 2) increases the duration of training by approximately 68%, in comparison with the baseline, whilst the complete uncertainty-aware system (Phase 3) necessitates approximately double the duration of the baseline training. This computational overhead represents a significant trade-off that must be considered in practical deployments.

VGG16 and AlexNet demonstrate analogous trends, with training times of 110, 196, and 231 minutes for VGG16, and 75, 114, and 117 minutes for AlexNet across the three phases, respectively. The shallower AlexNet architecture incurs the smallest absolute overhead, although the relative increase remains substantial.

E. Summary of Findings

The experimental results demonstrate that uncertainty modelling components provide a quantifiable yet limited enhancement in classification performance. In the third phase, the cost-sensitive learning algorithm demonstrated a successful recovery and slight enhancement in baseline performance. This was evidenced by improvements in the F1-score, which indicated an enhancement in recall for the crossing class. The introduction of threshold networks (Phase 2) has been shown to engender a degree of complexity that can temporarily result in a decline in performance. However, these networks are also capable of facilitating context-aware decision-making processes, a capability that has the potential to enhance the robustness of systems in scenarios that are not fully captured by aggregate metrics.

The significant enhancement in performance exhibited by ResNet50 in comparison to less complex architectures underscores the pivotal role of intricate, meticulously designed networks in the domain of pedestrian intention detection. The computational burden imposed by uncertainty modelling, marked by a near-doubling of training time, constitutes a pragmatic constraint that must be weighed against the negligible performance enhancements and the theoretical safety advantages.

VII. Discussion

The experimental results presented in Section VI demonstrate that uncertainty modelling provides a quantifiable yet limited enhancement in pedestrian intention detection performance. The subsequent section is dedicated to the interpretation of these findings, the discussion of their implications, and the identification of their limitations and future research directions.

A. Performance and Trade-offs

The experimental design, which is comprised of three phases, reveals distinct effects of aleatoric and epistemic uncertainty modeling components. The integration of threshold networks (Phase 2) resulted in slight performance decreases across all models, with accuracy and F1-score dropping marginally compared to the baseline. This reduction can be attributed to the additional complexity introduced by the threshold-modulation mechanism and the challenge of jointly optimizing the threshold network alongside the base classifier.

Conversely, cost-sensitive learning (Phase 3) demonstrated a successful recovery and subsequent enhancement in performance, surpassing the baseline level. All three models demonstrated enhancements in both accuracy and F1-score, with ResNet50 attaining 98.30% accuracy and 96.35% F1-score. The enhancements in the F1-score are especially salient, indicating that cost-sensitive learning augmented the model's capacity to accurately identify crossing pedestrians by prioritizing recall for this safety-critical category.

The performance advantage of ResNet50 over VGG16 and AlexNet remained consistent across all three phases, with a gap of approximately 28-30 percentage points in accuracy. This substantial difference underscores the importance of deep architectures with residual connections for capturing the complex visual and contextual cues necessary for accurate pedestrian intention prediction in urban scenarios.

B. Computational Considerations

The computational burden imposed by uncertainty modelling constitutes a substantial practical constraint. The training time increased almost twofold from the baseline (122 minutes) to the complete uncertainty-aware system (248 minutes) for ResNet50. Analogous relative increases were observed for VGG16 and AlexNet, indicating that the computational cost scales with the addition of uncertainty modelling components, irrespective of the underlying architecture.

Regarding inference latency, the threshold network introduces a lightweight MLP that processes the context vector in parallel with the classification layer, adding minimal computational overhead per prediction. However, explicit inference time measurements were not conducted in this study. For real-time autonomous driving deployment, where perception systems typically require sub-100ms response times, quantifying this overhead represents an essential validation step that should be addressed in future work.

The majority of this overhead can be attributed to the threshold network integration, which increases training time by approximately 68 percent compared to the baseline. The incorporation of cost-sensitive learning has been demonstrated to augment the duration of the process by approximately 21%, in comparison to Phase 2. This augmentation is observed without the introduction of additional parameters, a phenomenon that is presumably attributable to the modification of loss computation and gradient flow characteristics.

It is imperative that these computational demands are meticulously evaluated in relation to the negligible performance enhancements that have been observed. In environments or applications where resources are limited, or where models need to be updated quickly, there is often a trade-off between how reliable the models are and how efficiently they can be trained. In these cases, simpler baseline approaches or alternative optimization strategies may be favoured.

C. Implications for Pedestrian Intention Detection

The results of the study highlight several important considerations for developing uncertainty-aware pedestrian intention detection systems:

Architecture Selection: The marked disparity in performance between ResNet50 and less complex architectures (VGG16, AlexNet) remains consistent, irrespective of the adopted uncertainty modelling approach. This finding indicates that architectural depth and design continue to be the predominant factors influencing performance in PID tasks, with uncertainty modelling offering only marginal enhancements rather than substantial improvements.

Safety-Oriented Learning: The enhancements in the F1-score in Phase 3 suggest that cost-sensitive learning effectively directs the model towards higher recall for the crossing class. By imposing a greater penalty on false negatives, the approach encourages a more cautious prediction of imminent crossings by pedestrians, aligning with safety priorities in autonomous driving applications. However, the magnitude of improvement (approximately 0.05-0.58 percentage points in F1-score across models) is modest.

Threshold Modulation Challenges: The decline in performance observed in Phase 2 indicates that threshold networks, despite their conceptual appeal in modelling aleatoric uncertainty, pose significant integration challenges. The added complexity may require more sophisticated training strategies, larger datasets, or different architectural designs to realise their potential benefits. It is recommended that future research endeavours explore a range of alternative approaches to the adaptation of context-aware decision boundaries.

D. Limitations

However, it is important to note that the interpretation and generalizability of these findings are constrained by several limitations:

The following dataset constraints must be observed: The PIE dataset, while comprehensive, exhibits significant class imbalance and occasional annotation ambiguities (see Section III for further details). These characteristics may have had an impact on the training and evaluation of the model, particularly with regard to the components designed to handle ambiguous cases in uncertainty modeling.

The utilisation of restricted evaluation metrics is a key consideration. The evaluation process concentrated on two key metrics: accuracy and the F1-score. These metrics offer a comprehensive evaluation of the classification performance, yet they do not fully capture all the aspects that are pertinent to uncertainty-aware systems. It is suggested that metrics such as calibration error, confidence distributions, and threshold sensitivity should be employed in order to provide additional insights into the effectiveness of uncertainty modelling. Specifically, calibration metrics such as Expected Calibration Error (ECE) would quantify whether predicted probabilities align with actual outcomes—a critical property for safety-critical systems where overconfident predictions can lead to dangerous decisions. Similarly, analysis of confidence distributions and precision-recall curves would provide deeper insight into model behaviour across different operating thresholds. The absence of these metrics limits the assessment of whether the uncertainty modelling components genuinely improve prediction reliability beyond aggregate accuracy measures.

Single Dataset Evaluation: It is important to note that all experiments were conducted exclusively on the PIE dataset. In order to assess the generalisability of the uncertainty modelling approach across different environments and data collection conditions, it would be necessary to validate this approach on additional pedestrian datasets (such as JAAD) or real-world deployment scenarios.

The implementation process is characterised by a high degree of complexity. The precise mechanisms by which threshold networks influence decision boundaries and how contextual features are weighted in practice remain underspecified in the current implementation. In order to achieve a comprehensive understanding of these components and to facilitate their optimisation, it is necessary to undertake a more detailed analysis.

E. Future Research Directions

In light of the study's findings and its inherent limitations, several avenues merit further exploration through dedicated investigation.

Alternative Uncertainty Techniques: Exploration of Bayesian neural networks, deep ensembles, or evidential learning approaches may offer a more effective means of quantifying uncertainty, with the potential advantage of reduced computational overhead in comparison to the threshold network approach that has been employed in this study.

Advanced Architectures: The investigation of transformer-based models and attention mechanisms has the potential to enhance baseline performance and uncertainty modelling capabilities, especially with regard to the capture of long-range temporal dependencies and context relationships.

Comprehensive Evaluation: The expansion of the evaluation framework to encompass calibration metrics, confidence distributions, confusion matrices and safety-specific measures would facilitate a more comprehensive assessment of uncertainty-aware PID systems.

Cross-Dataset Validation: In order to ascertain the robustness and transferability of the proposed uncertainty modelling approach, it is necessary to test the framework on multiple pedestrian datasets and real-world scenarios.

Hyperparameter Optimization: A systematic investigation of cost-sensitive loss weights, threshold network architectures, and training strategies could identify configurations that better balance performance gains against computational costs.

F. Concluding Remarks

The present study demonstrates that integrating uncertainty modelling into pedestrian intention detection systems can provide measurable improvements in classification performance, particularly through cost-sensitive learning that prioritises safety-critical errors. Nevertheless, it should be noted that these benefits are accompanied by a significant computational overhead and implementation complexity. The modest magnitude of performance improvements suggests that uncertainty modelling should be regarded as one component of a comprehensive safety strategy rather than a transformative solution. The substantial performance advantage of deep architectures (ResNet50) over simpler models remains the dominant factor in determining the effectiveness of PID systems. Future endeavours should concentrate on achieving equilibrium between the conflicting imperatives of performance, computational efficiency, and uncertainty quantification to formulate pragmatic, implementable systems for autonomous vehicles navigating intricate urban environments.

VIII. Conclusion

The present paper set out to investigate the application of uncertainty modelling in the context of pedestrian intention detection within the complex environment of urban traffic. The present study explored the integration of two complementary mechanisms—threshold networks for aleatoric uncertainty and cost-sensitive learning for epistemic uncertainty—into conventional CNN architectures through a three-phase experimental design.

Experiments on the PIE dataset utilising three CNN backbones (ResNet50, VGG16, and AlexNet) have revealed that uncertainty modelling components exert distinct effects on classification performance. The integration of the Threshold network (Phase 2) introduced a degree of complexity, resulting in a slight decline in performance. Specifically, the accuracy of ResNet50 decreased from 98.23% to 98.20%, and the F1-score declined from 96.30% to 96.25%. In the third phase, the implementation of cost-sensitive learning successfully restored baseline performance and achieved modest enhancements. Notably, ResNet50 attained an accuracy of 98.30%, along with an F1-score of 96.35%.

The study demonstrated that deep architectures, particularly ResNet50 with its residual connections, substantially outperform simpler networks for pedestrian intention detection, regardless of uncertainty modelling approach. The performance disparity of approximately 28-30 percentage points between ResNet50 and simpler architectures (VGG16, AlexNet) remained consistent across all experimental phases.

A. *Key Findings*

The research provides several important insights:

The following paper sets out to explore the potential benefits of cost-sensitive learning. The efficacy of cost-sensitive learning in addressing epistemic uncertainty was demonstrated by its ability to bias the model towards enhanced recall for the crossing class. The modest F1-score improvements (0.05-0.58 percentage points across models) indicate enhanced sensitivity to safety-critical crossing events, though the magnitude of improvement is limited.

Threshold Network Challenges: The integration of threshold networks to model aleatoric uncertainty proved to be more challenging than had been anticipated. The decline in performance observed in Phase 2 indicates that the added complexity resulting from dynamic thresholding necessitates more advanced training strategies or architectural modifications to achieve its full potential.

Computational Trade-offs: The implementation of uncertainty modelling has been demonstrated to engender a considerable computational overhead. The duration of training for ResNet50 increased from 122 minutes (the baseline) to 248 minutes (completion of the entire framework), representing a 103% increase. When considering practical deployments, this computational cost must be weighed against the modest performance improvements.

Architecture Dominance: The selection of backbone architecture is the primary factor determining the performance of PID systems. Uncertainty modelling provides incremental refinements rather than transformative improvements, with ResNet50's architectural advantages persisting across all experimental phases.

B. *Contributions*

The present study contributes to the field of pedestrian intention detection research in three ways:

Firstly, it provides a systematic experimental evaluation of uncertainty modelling components within PID systems, isolating the effects of aleatoric and epistemic uncertainty handling through a structured three-phase design.

Secondly, it demonstrates that cost-sensitive learning offers a parameter-free approach to encoding safety priorities directly into network training, achieving modest improvements in F1-score without the need for architectural modifications.

Thirdly, it quantifies the computational costs associated with uncertainty modelling in PID systems, thereby establishing baseline measurements for training time overhead across multiple architectures.

C. *Limitations*

However, it is important to note that the findings are constrained by several limitations, which restrict both the scope and generalizability of the results. The evaluation was conducted exclusively on the PIE dataset, employing solely accuracy and F1-score metrics. A comprehensive uncertainty assessment would require additional metrics such as calibration error, confidence distributions, and threshold sensitivity analysis. The study did not include measurements of inference time, embedded hardware testing, or cross-dataset validation, which are essential for assessing real-world deployment feasibility.

D. *Future Directions*

It is recommended that future research address the limitations identified and explore several promising avenues for further study.

- **A comprehensive evaluation of the metrics is required**. The incorporation of calibration metrics (ECE), confidence distributions, precision-recall analysis, and confusion matrices is imperative in order to provide a comprehensive assessment of uncertainty-aware systems.
- **Alternative uncertainty techniques:** The investigation encompasses Bayesian neural networks, deep ensembles, and evidential learning approaches, with the objective of ascertaining whether these methods can offer enhanced uncertainty quantification with reduced computational overhead.
- **Temporal modeling:** The framework is to be extended with recurrent or transformer-based temporal encoders with a view to capturing long-term pedestrian motion dynamics and temporal uncertainty evolution.
- **Cross-dataset validation:** The approach is to be tested on additional benchmarks (JAAD, TITAN) in order to assess generalisation across diverse visual domains and data collection conditions.
- **Real-world deployment:** The following three steps are to be taken in order to establish the practical feasibility of autonomous vehicle applications: firstly, inference time measurements must be conducted; secondly, embedded hardware testing must be carried out; and thirdly, real-world validation must be undertaken.
- **The process of threshold network refinement is outlined as follows**: The present study investigates alternative architectures, training strategies and contextual feature selection with a view to realising the potential benefits of adaptive thresholding.

E. Closing Remarks

This study demonstrates that the application of uncertainty modelling in pedestrian intention detection systems offers both opportunities and challenges. Whilst the cost-sensitive learning mechanism offers a straightforward method for encoding safety priorities, yielding modest performance benefits, the integration of threshold networks for aleatoric uncertainty has proven to be more complex than initially anticipated. The substantial computational overhead and modest performance improvements suggest that uncertainty modelling should be carefully evaluated against specific deployment requirements rather than being adopted universally.

The research emphasises that architectural design, specifically the utilisation of deep networks with residual connections, remains the predominant factor influencing the performance of PID systems. Uncertainty modelling offers incremental refinements that may be valuable in safety-critical contexts, where even minor improvements in recall for crossing events justify additional computational investment.

Future endeavours should concentrate on the development of more efficient uncertainty estimation techniques, conducting comprehensive evaluations including calibration and confidence metrics, and validating approaches across multiple datasets and real-world deployment scenarios. By addressing these challenges, the research community can work towards building pedestrian intention detection systems that effectively balance performance, reliability, computational efficiency, and safety requirements for autonomous vehicle applications.

References

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012. [Online]. Available: https://doi.org/10.1109/TPAMI.2011.155

[2] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2016, pp. 2243–2248. [Online]. Available: https://doi.org/10.1109/IROS.2016.7759351

[3] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2017, pp. 206–213. [Online]. Available: https://doi.org/10.1109/ICCVW.2017.33

[4] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6262–6271. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00636

[5] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, 2009. [Online]. Available: https://doi.org/10.1016/j.strusafe.2008.06.020

[6] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021. [Online]. Available: https://doi.org/10.1007/s10994-021-05946-3

[7] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5574–5584. [Online]. Available: https://doi.org/10.48550/arXiv.1703.04977

[8] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059. [Online]. Available: https://doi.org/10.48550/arXiv.1506.02142

[9] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6402–6413. [Online]. Available: https://doi.org/10.48550/arXiv.1612.01474

[10] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 3179–3189. [Online]. Available: https://doi.org/10.48550/arXiv.1806.01768

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1409.1556

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105. [Online]. Available: https://doi.org/10.1145/3065386