Degree project

UNIVERSITY OF SKÖVDE

Explainable Short-Horizon Prediction of Pedestrian Crossing-Corridor Entry Using Trajectory and Interaction Features

Master's degree Project in Informatics with a specialization in Data Science

Second Cycle 15 credits

Spring term 2026

Student: Valery Nkenguruke

Supervisor: Joe Steinhauer

Examiner: Göran Falkman

Abstract

Pedestrian–vehicle interactions in urban traffic are difficult to model because they unfold over short time scales, depend on local geometric context, and involve ambiguity between observed motion and latent intention. This thesis investigated whether short-horizon pedestrian crossing-corridor entry could be predicted from real-world trajectory data and how the conclusions depended on key design choices, especially the future horizon. Using stationary trajectory data from Viscando collected in an urban setting in Gothenburg, Sweden, the study defined an operational target as whether a pedestrian entered a predefined crossing corridor within a fixed future horizon. Horizons of 2, 3, and 5 seconds were compared, with 3 seconds selected as the main reporting horizon.

The approach combined interpretable feature engineering, logistic-regression baselines with staged interaction-feature ablation, a feed-forward MLP, and a compact GRU sequence-model comparison. Transparency was pursued both through interpretable model design and through post-hoc permutation-importance analysis applied to the strongest non-linear model. The intended application is a decision-support framework for traffic engineers, road-safety analysts, or ADAS developers requiring both accurate predictions and understandable model behavior.

Pedestrian-only motion and corridor-relative geometry provided a strong baseline. On the matched vehicle-context subset, richer interaction structure, especially relative vehicle geometry, was more informative than simple proximity. The interaction-aware MLP was the strongest overall point-estimate result, though the within-family pedestrian-only versus interaction-aware difference was not statistically confirmed. The compact GRU sequence branch did not outperform the engineered tabular MLP under the present representation and data conditions. Horizon choice materially affected class balance, performance, and which interaction signal was most useful, shifting from vehicle motion at 2 seconds, to relative geometry at 3 seconds, to local density at 5 seconds. Taken together, the findings show not only that short-horizon crossing-corridor entry can be predicted from stationary trajectory data, but also that target formulation, horizon choice, and feature representation materially shape what can be learned from the data.

Table of Contents

1 Introduction

Urban traffic environments contain frequent, short, and often ambiguous interactions between pedestrians, cyclists, and motor vehicles. These interactions are especially critical near crossings, roundabouts, and other shared urban spaces where vulnerable road users and vehicles move within limited space and time. In such settings, the ability to anticipate road-user behavior is important for proactive safety systems, traffic analysis, and the development of future intelligent transport systems, including Advanced Driver Assistance Systems (ADAS) and autonomous vehicles (Landry & Akhloufi, 2025; Mirzabagheri et al., 2025). Broader work in driver intention recognition and traffic-behavior prediction shows that short-horizon anticipation is both practically relevant and scientifically challenging, particularly because traffic behavior unfolds under uncertainty and is highly context-dependent (Vellenga et al., 2022).

ADAS and other predictive traffic technologies increasingly rely on anticipatory models rather than purely reactive rules, since reactive systems alone cannot provide the anticipation window necessary to prevent safety-critical events in complex urban settings (Mirzabagheri et al., 2025). Existing research confirms that recognizing imminent behavior can support safer and more informed responses but also highlights important open challenges: the distinction between latent intention and observable behavior, the heterogeneity of traffic contexts, the difficulty of transferring models between sites, and the tension between predictive power and model transparency (Vellenga et al., 2022). In interaction-heavy traffic settings, prior work has further shown that purely kinematic information may be insufficient and that contextual or behavioral cues can matter for prediction quality (Mohammadi et al., 2023).

For pedestrians specifically, one recurring challenge is that observed motion does not map cleanly onto true psychological intention. A pedestrian may approach a crossing area, slow down, wait, reorient, or move parallel to a crossing without actually entering it immediately (Rasouli & Tsotsos, 2020; Schneemann & Heinemann, 2016). This behavioral ambiguity means that practical modeling requires carefully defined operational targets rather than direct claims about unobservable inner states. This issue is closely related to a broader distinction in the literature between latent internal intention and externally observable short-term behavior, a distinction that is especially important when modeling relies only on trajectory data and has no access to richer embodied or contextual cues (Vellenga et al., 2022; Landry & Akhloufi, 2025).

This thesis therefore adopts a deliberately narrow and transparent formulation. The task is not to infer true psychological intention. Instead, it is to predict a short-horizon, operationally defined event: whether a pedestrian enters a predefined crossing corridor within a fixed future time window. This formulation is motivated by three considerations. First, it is observable and reproducible from trajectory data alone. Second, it is directly relevant for short-term anticipation in traffic-safety contexts. Third, it permits structured comparison between models with different levels of transparency, built from the same trajectory-based setting.

Two related but distinct concepts are used throughout this thesis, and it is important to distinguish them from the outset. Interpretability refers to the degree to which a model's structure and parameters are directly human-readable without additional tooling. Logistic

regression is interpretable in this sense: its coefficients can be read and compared directly. Explainability, by contrast, refers to post-hoc methods applied after training to approximate or summarize what a model has learned, even when the model itself is not inherently transparent. Permutation importance applied to a multilayer perceptron is an example of an explainability method. Both concepts matter in this thesis: interpretability is a core design principle pursued through logistic regression and staged ablation, while explainability is applied to the strongest non-linear model. The title's use of "explainable" covers both dimensions. This distinction is not merely terminological. Vellenga et al. (2022) identify interpretability and explainability as open challenges for deploying traffic-prediction systems in practice, and Castillo et al. (2025) specifically discuss the need to make pedestrian behavior prediction models transparent in autonomous driving contexts. Govindhan (2025) further illustrates the use of post-hoc explanation tools such as SHAP and LIME in a related traffic-prediction setting.

The predictor developed in this thesis is intended as a component of a traffic-safety analysis or decision-support system for practitioners such as traffic engineers, road-safety analysts, or developers of ADAS components. Knowing whether a pedestrian is likely to enter a crossing corridor within the next two to three seconds is directly relevant for triggering a warning, adjusting vehicle speed, or logging a near-miss event. The intended user is not a member of the general public but a technically informed operator or system designer who requires both predictive accuracy and an understanding of which features drive the model's outputs. This is precisely why transparency is a central design requirement: in safety-adjacent applications, model outputs must be auditable and understandable (Vellenga et al., 2022; Landry & Akhloufi, 2025; Mirzabagheri et al., 2025).

The study is based on stationary traffic observations provided by Viscando from an urban interaction setting in Gothenburg. The raw data include tracked road users, timestamps, positions, speeds, and actor types. From these trajectories, a crossing corridor and surrounding operational areas are defined, labels are constructed, and feature sets are derived. The thesis investigates not only whether short-horizon crossing-related behavior can be predicted, but also how the answer changes when important design choices change, especially the future horizon.

In many traffic-prediction studies, the future horizon is treated as a fixed technical parameter chosen for convenience. In the present work, it is treated as part of the scientific problem formulation. Three horizons are compared: 2 seconds, 3 seconds, and 5 seconds. These specific values were chosen for substantive reasons. A 2-second horizon represents the strictest operationalization of imminent behavior, capturing only pedestrians who are about to enter the corridor almost immediately. This is the most safety-critical interval but also the hardest prediction target due to sparse positive labels and the narrow behavioral window (Schneemann & Heinemann, 2016). A 5-second horizon substantially broadens the temporal window, producing more positive labels and generally easier prediction, but at the cost of blurring the distinction between a pedestrian who is genuinely committed to crossing and one who is still deciding (Rasouli & Tsotsos, 2020). The 3-second interval was selected as the primary reporting horizon because it offers a practically relevant anticipation window consistent with the reaction-time budgets used in ADAS design and traffic-safety research: Vellenga et al. (2022) note that safety-critical traffic interactions typically unfold within a few seconds, and Landry and Akhloufi (2025) document that the prevalent benchmark in pedestrian crossing-intention prediction uses a 1–2 second window, making 3 seconds a horizon that extends slightly beyond established practice while remaining within a safety-

relevant anticipation range. Comparing all three horizons makes it possible to study how changes in the target affect class balance, learnability, and the apparent usefulness of interaction features, transforming horizon choice from a hidden assumption into an empirical finding (Delgado, 2025; Govindhan, 2025; Mirkovic, 2025).

The work is positioned between two broader strands of literature. One strand emphasizes increasingly complex sequential and context-aware architectures for intention or risk prediction, using recurrent models and context-aware designs to capture temporal dependencies in interaction settings (Delgado, 2025; Govindhan, 2025; Mirkovic, 2025). Another strand emphasizes the need for transparent models, careful problem framing, and deployment realism, especially in safety-related domains where model outputs may affect real-world decision support and where opacity is a recognized barrier to practical adoption (Vellenga et al., 2022; Mirzabagheri et al., 2025). The present thesis contributes by combining these perspectives in a controlled way: it starts from an interpretable short-horizon formulation, then compares linear, non-linear, and sequential models without losing sight of operational clarity and transparency.

At the same time, the thesis remains clearly distinct from broader behavior-modeling approaches such as inverse reinforcement learning, where the aim is to recover the motivations or reward structures behind observed traffic behavior rather than to predict a narrowly defined future event (Amaya Corredor, 2021). The present work is more limited in scope, but also more explicit in what exactly is being predicted.

Rather than asking only which model performs best, the thesis asks a more structured question: what can be learned about short-horizon pedestrian crossing-related prediction when the target is carefully operationalized, the modeling pipeline remains interpretable, and the sensitivity to horizon choice is examined explicitly?

1.1 Problem definition

Road-user behavior prediction is scientifically relevant because it addresses a real anticipation problem under uncertainty, and practically relevant because it can support safer traffic systems. However, the problem must be delimited carefully. In the present setting, the data consist of externally observed trajectories rather than direct measurements of latent cognitive states. This constrains what can be claimed and motivates an operational definition of the target.

The central problem addressed in this thesis is therefore the following: how effectively can short-horizon pedestrian crossing-related behavior be predicted from real-world trajectory data, and how do key design choices, especially the future horizon and the type of model used, affect both predictive performance and the apparent usefulness of interaction context?

The work focuses on a single urban interaction setting and a specific binary prediction target: whether a pedestrian will enter a defined crossing corridor within a short future horizon. The main empirical question is not whether the most complex model can be built, but whether a carefully structured predictive pipeline can capture useful signal, how much is already explained by pedestrian-only motion and corridor-relative geometry, and under what conditions interaction-aware context adds value. The emphasis on transparent modeling is motivated not only by scientific principles but by the intended application: a decision-support

tool whose outputs need to be auditable and understandable to the practitioners who use it.

1.2 Aim

The aim of this thesis is to investigate how well short-horizon pedestrian crossing-corridor entry can be predicted from real-world urban trajectory data, and to evaluate how horizon choice, interaction-aware context, and model family affect predictive performance and model transparency. Transparency is pursued both through interpretable model design, in which the model structure itself reveals what drives predictions, and through post-hoc explainability methods applied to the strongest non-linear model. The intended output is a decision-support framework suitable for traffic engineers, road-safety analysts, or ADAS developers who need both accurate predictions and understandable model behavior.

1.3 Research questions

The thesis is guided by the following main research question:

How effectively can short-horizon crossing-corridor entry be predicted from pedestrian trajectory features, and how do horizon choice, interaction-aware context, and model family affect predictive performance and model transparency?

This question is addressed through four sub-questions:

1. How does the operational label behave across different future horizons (2, 3, and 5 seconds), and which horizon is most suitable as the main reporting setting? This question is motivated by Vellenga et al. (2022) and Landry & Akhloufi (2025), who identify inconsistent and underspecified horizon choice as a recurring limitation in traffic-prediction research.
2. How strong is a pedestrian-only baseline built from recent motion and corridor-relative geometry? This question is motivated by Rasouli & Tsotsos (2020) and Schneemann & Heinemann (2016), who show that observable pedestrian kinematics carry considerable predictive signal and serve as an important benchmark for evaluating the added value of interaction context.
3. Do interaction-aware features, that is, features derived from the local vehicle context surrounding the pedestrian, improve prediction beyond the pedestrian-only baseline, and if so, which interaction signals appear most useful across different horizons and model families? This is motivated by Mohammadi et al. (2023) and Zou et al. (2025), who demonstrate that spatiotemporal interaction context improves pedestrian behavior prediction, but leave open the question of which specific vehicle-context signals matter most.
4. How do an interpretable linear baseline, a modest non-linear MLP, and an exploratory sequence model compare in terms of predictive performance, and what do these comparisons reveal about the trade-off between performance and model transparency? Transparency here encompasses both interpretability-by-design (logistic regression) and post-hoc explainability (permutation importance applied to the MLP). This question is motivated by Vellenga et al. (2022) and Castillo et al. (2025), who identify the tension between model complexity and transparency as an open challenge in traffic-prediction research, and by Govindhan (2025), who demonstrates the value of post-hoc explanation tools in a related setting.

1.4 Delimitations

This study is delimited in several important ways. First, it uses one available dataset from one urban traffic setting. The results should therefore be interpreted as site-specific rather than universally generalizable. Second, the target is an operational, geometry-based proxy for short-horizon crossing-related behavior. Third, the thesis focuses on short-horizon supervised prediction rather than full trajectory generation, collision-risk estimation, simulation, or inverse reinforcement learning, although such approaches are relevant points of comparison in the broader literature (Amaya Corredor, 2021).

Fourth, the primary methodological emphasis remains on interpretability-first modeling and explicit feature analysis. Although the study includes a modest non-linear MLP and an exploratory GRU comparison, these are evaluated within the same operational framing rather than developed as a full architecture-optimization project (Delgado, 2025; Govindhan, 2025; Mirkovic, 2025). Fifth, interaction context is constructed from local scene matching based on rounded timestamps and nearby vehicles. This provides a practical operationalization, but not a complete representation of all interaction structure.

1.5 Structure of the report

The remainder of the report is structured as follows. Chapter 2 presents background and related research. Chapter 3 describes the method, including research design, horizon comparison strategy, validity, reliability, and ethical considerations. Chapter 4 details the implementation, including dataset processing, spatial operationalization, label construction, feature engineering, and the different model branches. Chapter 5 presents the results, beginning with horizon comparison and then focusing on the main 3-second analysis. Chapter 6 discusses the findings in relation to previous research, methodological choices, interpretability, and broader scientific, ethical, and societal issues. Chapter 7 concludes the thesis and outlines future work.

2 Background

2.1 Road-user intention and behavior prediction in traffic

Research on intention recognition and behavior prediction in traffic has grown alongside the development of Advanced Driver Assistance Systems (ADAS), autonomous driving, and proactive safety systems. Across this broader literature, the central aim is to anticipate what a road user is likely to do in the near future so that another agent or system can respond in time. This is especially important in short-horizon settings, because many safety-critical traffic interactions unfold within only a few seconds (Vellenga et al., 2022).

A central conceptual challenge in this field is the difference between latent intention and observable behavior. In a strict sense, intention refers to an internal mental state. Most traffic-prediction studies, however, do not observe such states directly. Instead, they rely on externally visible quantities such as trajectories, speeds, headings, relative positions, and scene context. For that reason, many practical systems do not truly infer intention in a deep psychological sense, but instead predict short-term behavior, imminent maneuvers, or other operational future events (Vellenga et al., 2022).

This distinction is especially important in pedestrian settings, where observable movement can be ambiguous. A pedestrian may approach a crossing area, slow down, wait, reorient, or continue in parallel without actually entering the road. Prior pedestrian research therefore often relies on operational proxies derived from motion and context rather than direct access to latent intention (Rasouli & Tsotsos, 2020; Schneemann & Heinemann, 2016; Landry & Akhloufi, 2025). This broader problem becomes even more pronounced in urban environments involving vulnerable road users, where local interaction context often shapes what an observed movement actually means.

2.2 Vulnerable road users and urban interaction settings

Pedestrians and cyclists are often described as vulnerable road users because they are more exposed to injury in traffic conflicts than occupants of motor vehicles. Their interactions with vehicles are especially complex in urban environments, where multiple actors move within shared or partially shared space, often under short temporal margins and limited predictability. Such settings include crossings, intersections, roundabouts, and other negotiation-intensive traffic spaces.

A central difficulty in these environments is that similar motion can have very different meanings depending on context. A pedestrian moving toward a crossing may enter immediately, hesitate, continue along the sidewalk, or change direction altogether. This means that useful prediction depends not only on the pedestrian's own motion, but also, at least potentially, on nearby vehicle context and the broader local interaction structure. In pedestrian crossing scenarios, previous work has shown that crossing-related prediction depends both on local context and on the precise temporal framing of the task (Rasouli & Tsotsos, 2020; Schneemann & Heinemann, 2016; Zou et al., 2025).

Naturalistic interaction research supports this broader contextual view. In cyclist–vehicle encounters at unsignalized intersections, Mohammadi et al. (2023) show that behavior is shaped not only by kinematics such as speed and position, but also by richer behavioral and contextual cues. Although the present thesis does not have access to comparable non-verbal or embodied information, that prior work still matters because it supports the broader claim that road-user behavior in urban interaction settings is fundamentally contextual. This leads naturally to the more specific question of how interaction-aware information should be represented in prediction models.

2.3 Interaction-aware prediction and behavioral cues

Prior work suggests that richer contextual and behavioral signals can improve prediction in interaction-heavy traffic settings. Mohammadi et al. (2023), for example, show that cyclist yielding behavior at unsignalized intersections is better explained when cues such as pedaling and head movement are considered alongside conventional motion variables. Even though the present thesis does not use such cues, the result is still informative: it suggests that the predictive problem is not purely geometric and that interaction-aware information can matter when available.

At the same time, the present thesis asks a more constrained question: what can be achieved using stationary trajectory data alone, supplemented only with modest local interaction variables such as nearest-vehicle distance, relative position, vehicle speed, vehicle type, or local density? This is an important practical question because many real-world traffic

datasets do not contain richer non-verbal indicators. It is also relevant in light of prior pedestrian work showing that crossing-related prediction can benefit from context-aware modeling even when the target remains an operational short-horizon event rather than a direct measure of intention (Schneemann & Heinemann, 2016).

The issue is therefore not simply whether context matters, but how it should be represented and modeled. That question connects the present thesis to broader modeling traditions in traffic prediction.

2.4 Related modeling traditions

The broader literature relevant to this thesis can be grouped into three overlapping traditions.

The first tradition focuses on driver intention recognition or maneuver prediction, often in settings such as lane changing, turning, or roundabout navigation. These studies compare machine-learning methods for predicting imminent maneuvers from vehicle dynamics, scene cues, and temporal observations. Review-based work shows that this field contains wide variation in maneuver types, feature spaces, and model families, but also substantial heterogeneity in task definitions and evaluation settings (Vellenga et al., 2022). This tradition is relevant here because it provides the broader methodological context for short-horizon traffic prediction, even though the present thesis addresses pedestrians rather than vehicles.

The second tradition focuses on vulnerable-road-user interaction modeling, including cyclist–vehicle and pedestrian–vehicle encounters. In this line of work, the central questions often concern yielding, crossing, conflict development, or other short-term interaction outcomes. These studies are especially relevant because they emphasize that local context matters and that pedestrian and cyclist prediction differs from more isolated trajectory-forecasting problems (Mohammadi et al., 2023; Rasouli & Tsotsos, 2020; Schneemann & Heinemann, 2016).

The third tradition addresses broader behavior modeling, where the objective is not only to classify a future event but to approximate the logic or structure behind traffic behavior itself. One example is inverse reinforcement learning, which attempts to model human behavior in traffic interactions through inferred reward structures rather than event labels alone (Amaya Corredor, 2021).

The present thesis is closest to the second tradition, but it also intersects with the others through its concern with temporal prediction, operational target definition, and model-family comparison. Among these model classes, sequence-based approaches have recently been explored in closely related traffic-prediction settings, and their value relative to interpretable tabular baselines is one of the questions the present thesis directly addresses.

2.5 Sequence modeling in traffic prediction

Recent work in related traffic-prediction settings has explored recurrent neural architectures for maneuver and intention recognition, especially in turning-related or roundabout scenarios. Delgado (2025) compares recurrent neural models for vehicle turning-intention recognition at roundabouts, Mirkovic (2025) uses an LSTM for light-vehicle intention

recognition, and Govindhan (2025) develops a context-aware LSTM framework for cyclist–vehicle collision-related prediction. Together, these studies show that recurrent sequence models are relevant in settings where temporal dependencies are central and where the prediction problem can naturally be framed as sequence learning.

At the same time, these studies also illustrate that sequence models answer a somewhat different question from interpretable tabular baselines. A recurrent model is useful when the objective is to let the architecture learn temporal patterns directly from sequences of observations. By contrast, an interpretable tabular pipeline compresses recent temporal history into explicitly defined features and makes the model's behavior easier to inspect.

The present thesis therefore includes a compact GRU comparison rather than a broad multi-architecture sequence study. The purpose is not to optimize recurrent architectures extensively, but to test whether direct temporal modeling adds value beyond the main tabular logistic and MLP branches under the same operational problem framing. This comparison is useful, but it does not exhaust the broader space of behavior-modeling alternatives. The following section covers more ambitious modeling approaches that go beyond predicting a future event label and instead aim to reconstruct the underlying logic or motivation behind traffic behavior.

2.6 Behavior modeling beyond supervised classification

More ambitious approaches to traffic behavior modeling aim to recover the structure of behavior itself rather than only predict a future class label. Amaya Corredor (2021), for example, uses inverse reinforcement learning to model human behavior in traffic interactions. In that line of work, the aim is not merely to classify short-term events, but to infer behavioral structure that can support more realistic simulation or deeper interpretation of interaction dynamics.

This broader behavior-modeling perspective is useful for positioning the present study because it clarifies what the thesis does not attempt to do. The present work does not learn reward functions, generate realistic future trajectories, or simulate human-like policy. It remains deliberately narrower. Its contribution lies in the operationalization and empirical evaluation of a short-horizon crossing-related event, together with interpretable comparison across model families. Once the scope is delimited in this way, the next issue becomes not only what to model, but also how transparent those models should be.

2.7 Explainability in traffic prediction

Explainability is an important consideration in traffic prediction because model outputs may influence safety-related decisions, design choices, or future decision-support systems. In such settings, predictive performance alone is insufficient. It is also important to understand what information a model relies on and whether that reliance is meaningful in traffic terms (Vellenga et al., 2022; Castillo et al., 2025).

As established in Section 1, interpretability and explainability are related but distinct concepts. An interpretable model, such as logistic regression, reveals its reasoning through its parameters directly without requiring additional tooling: the sign and magnitude of coefficients indicate the direction and relative strength of each feature's contribution. An explainable model, by contrast, such as a neural network, requires an additional post-hoc

step to approximate what the model has learned, since the model's internal parameters are not directly human-readable.

Two widely used post-hoc explainability methods are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP assigns each input feature a contribution score based on game-theoretic Shapley values, quantifying how much each feature shifts the model output relative to a baseline across all possible feature subsets (Lundberg & Lee, 2017). LIME constructs a locally faithful linear approximation of a complex model around a specific prediction, making it possible to inspect which features mattered most for that particular instance (Ribeiro et al., 2016). Both methods can be applied to any model regardless of its internal architecture and are increasingly used in traffic-prediction research to make complex models more transparent (Govindhan, 2025; Castillo et al., 2025).

The need for transparency appears in two complementary ways in the relevant literature. First, review-based work in driver intention recognition identifies model transparency as a recurring challenge in moving from research prototypes to robust practical use (Vellenga et al., 2022). Second, more recent work in pedestrian-behavior prediction specifically frames explainability as a design requirement rather than an optional feature, arguing that models influencing safety decisions must be able to justify their outputs in terms that practitioners can evaluate (Castillo et al., 2025).

In the present thesis, transparency is not treated as a separate add-on but as a core design principle pursued throughout. Logistic regression provides coefficient-level interpretability and supports explicit staged ablation, allowing direct inspection of which feature groups contribute to performance. For the strongest non-linear branch, the MLP, transparency is addressed through permutation importance. Permutation importance is a model-agnostic post-hoc method that evaluates how much the model relies on each feature by measuring the loss in predictive performance when that feature is disrupted. Concretely, one feature at a time is randomly shuffled across all rows in the evaluation set, which breaks the real association between that feature and the target while leaving the rest of the data unchanged. The trained model is then re-evaluated on this permuted dataset using the same metric as in the main analysis, here PR-AUC. The decrease in PR-AUC relative to the unpermuted baseline is taken as the importance score for that feature. A large drop indicates that the model relied strongly on the feature, whereas a near-zero drop indicates that the model can perform almost as well without it. Repeating this procedure for all features produces a global ranking of feature relevance. The method is model-agnostic because it operates only on model inputs and outputs and does not require access to internal parameters or architectural details.

Permutation importance was chosen deliberately over SHAP and LIME because the scientific aim in this thesis is global rather than instance-level explanation. SHAP and LIME are especially useful when the objective is to explain why a model produced a specific prediction for a specific observation. In the present study, however, the central question is which features the MLP relies on overall across the held-out evaluation set, so that this global pattern can be compared with the coefficient and ablation evidence from logistic regression. For that purpose, permutation importance is the more direct tool. It was originally introduced in the context of random forests by Breiman (2001) and later formalized as a general model-agnostic feature-importance method by Fisher et al. (2019). It is also practically appropriate here: compared with SHAP and LIME, permutation importance is computationally lighter, avoids an additional layer of local approximation, and yields a performance-based importance measure that is easy to interpret in relation to the PR-AUC

metric used throughout the thesis. Together, these transparency considerations form part of the broader methodological landscape that the following section synthesizes at the field level.

2.8 Review-based field context

A broad review of driver intention recognition research by Vellenga et al. (2022) provides a useful field-level backdrop for the present thesis. Their review shows that the area includes many maneuver types, heterogeneous feature spaces, and a wide variety of AI methods. It also highlights several recurring open challenges: inconsistent task definitions, limited comparability across datasets, difficulties in real-world deployment, and the need for stronger interpretability and validation practices.

These observations matter here for two reasons. First, they confirm that short-horizon traffic prediction is an active and practically relevant area. Second, they reinforce one of the core themes of the present thesis: results depend strongly on how the prediction problem is defined. That includes what event is predicted, over what time horizon, with what sensor inputs, and under what evaluation procedure. In this sense, the review-based field context helps pull the discussion back from individual modeling strands to the broader methodological conditions under which traffic-prediction results should be interpreted.

With that broader context in place, the final step is to position the present thesis more explicitly within the literature and clarify what kind of contribution it aims to make.

2.9 Positioning of the present study

The literature discussed above supports three broad observations. First, short-horizon road-user prediction is a relevant and active area of traffic-safety research (Vellenga et al., 2022). Second, interaction-aware and context-aware information can matter, especially in vulnerable-road-user settings (Mohammadi et al., 2023; Rasouli & Tsotsos, 2020). Third, more ambitious sequence-based and behavior-modeling approaches exist, but they do not remove the need for carefully defined targets, transparent evaluation, and deployment realism (Amaya Corredor, 2021; Govindhan, 2025; Mirzabagheri et al., 2025). In this positioning, peer-reviewed review, journal, and conference publications serve as the main theoretical anchors, while student-thesis sources are used more narrowly as local methodological comparison points.

Within this landscape, the present study fills a specific gap. It does not ask whether the most complex traffic-prediction architecture can be built. Instead, it asks how short-horizon pedestrian crossing-corridor entry can be predicted from real-world trajectory data when the target is operationalized explicitly, the future horizon is varied systematically, and linear, non-linear, and sequential model families are compared under the same overall framing.

This leads to three specific contributions. First, the thesis develops a transparent operational formulation of crossing-related behavior as short-horizon corridor entry rather than latent intention. Second, it compares model families in a way that preserves interpretability while still testing whether non-linear or sequential models add value. Third, it shows that the empirical conclusions depend materially on horizon choice: changing the future horizon changes class balance, predictive performance, and the apparent usefulness of interaction features.

That final point is especially important. Many traffic-prediction studies treat the horizon mainly as a technical hyperparameter. In the present thesis, horizon choice becomes part of the substantive finding. The thesis therefore contributes not only a model comparison, but also a methodological argument about how target formulation shapes what can be learned from trajectory-based pedestrian prediction.

3 Method

3.1 Research design

This thesis adopts a quantitative, supervised-learning, comparative design. The overall workflow proceeds from raw trajectory inspection to geometric operationalization, label construction, feature engineering, model fitting, threshold selection, and comparative evaluation. The study is observational and retrospective, using previously recorded traffic trajectories rather than data collected through direct intervention.

The design is comparative in three connected senses. First, it compares the operational label across multiple future horizons in order to determine a suitable main prediction setting. Second, it compares pedestrian-only and interaction-aware feature formulations. Third, it compares model families with different levels of complexity and interpretability, namely logistic regression, a feed-forward multilayer perceptron (MLP), and exploratory recurrent sequence models. This makes it possible to study not only predictive performance, but also how the apparent usefulness of features changes depending on target design and model type.

The study is therefore not framed as a pure best-model-wins exercise. Instead, it is designed to answer a broader methodological question: how the conclusions depend on the way the prediction problem is operationalized, especially with respect to the chosen future horizon and the balance between transparency and modeling complexity.

3.2 Why an interpretable supervised approach was chosen

Once the study is framed as short-horizon prediction of an operationally defined event, the next methodological question is how that event should be modeled. More complex sequence-based and context-aware models are available in related traffic settings, and recent work has shown that recurrent architectures can perform strongly in roundabout and trajectory-prediction tasks (Delgado, 2025; Govindhan, 2025; Mirkovic, 2025). However, the main methodological path in this thesis was chosen to balance scientific relevance, scope, transparency, and comparability.

There are four reasons for this choice. First, the target is operational and short-horizon rather than a direct measure of psychological intention. This makes transparency especially important, because the interpretation must remain aligned with what is actually being predicted rather than what might be inferred psychologically. Second, the thesis aims not only to test whether interaction context helps, but also to understand which kinds of interaction signals help and under what conditions. This is easier to examine through explicit feature engineering and structured ablation than through a fully opaque end-to-end architecture, where the source of any performance gain would be difficult to attribute.

Third, the project scope favors a defensible and explainable comparison that can be fully

motivated and critically reflected upon within the given scope. This is consistent with wider concerns in driver intention recognition research, where transparency, comparability, and deployment realism remain open challenges (Vellenga et al., 2022; Mirzabagheri et al., 2025). Fourth, interpretable baselines remain scientifically valuable even when more flexible models exist. If a relatively simple and transparent model already captures most of the predictive signal, that is itself an important finding.

For these reasons, the thesis begins with logistic regression as an interpretable baseline, extends the comparison to a modest non-linear MLP, and then includes an exploratory GRU as a sequence-model reference point.

3.3 Unit of analysis and target formulation

The methodological design described above requires a clear unit of analysis and a clear target definition. In the tabular part of the study, the effective unit of analysis is a pedestrian observation row for which sufficient recent history and future visibility are available. Each such row contains the pedestrian's current state, short recent-history features, and, where available, matched local interaction context from nearby vehicles.

The target variable is a binary label indicating whether the pedestrian enters the predefined crossing corridor within a specified future horizon. In the implementation, this label is represented as cross_soon, but conceptually it is treated throughout the thesis as short-horizon crossing-corridor entry.

For the sequence-model branch, the unit of analysis changes from a single row to a fixed-length sequence window of pedestrian observations. However, the target remains anchored to the last timestep in the sequence and uses the same operational crossing-corridor-entry definition. This allows the sequence models to be compared with the tabular models under the same overall problem framing.

3.4 Horizon selection strategy

Because short-horizon prediction is central to the thesis, the label is examined across multiple future horizons rather than fixed at a single setting from the outset. The study compares 2-second, 3-second, and 5-second horizons. The rationale for these values is discussed in Section 1, where each horizon is motivated by its semantic properties and by prior work in pedestrian crossing-intention prediction (Schneemann & Heinemann, 2016; Vellenga et al., 2022; Landry & Akhloufi, 2025).

The purpose of this comparison is not merely descriptive. Different horizons change the positive rate, the semantic meaning of the label, and the difficulty of the prediction task. A very short horizon preserves immediacy but may produce a sparse and difficult label. A longer horizon produces more positive cases and often stronger apparent performance, but it also broadens the target to include pedestrians who are still deliberating and may not cross at all.

For that reason, horizon selection is treated as part of the research design rather than as a minor parameter. The three horizons are compared in terms of class balance, interaction coverage, and resulting predictive behavior. On this basis, 3 seconds is selected as the main reporting horizon because it represents the best empirically supported compromise between

immediacy and learnability in this dataset, while 2 seconds is retained as a stricter sensitivity setting and 5 seconds as a broader one.

3.5 Feature strategy

The horizon comparison defines what is being predicted. The next methodological question is what information should be made available to the models. For that reason, the thesis uses two main feature perspectives: pedestrian-only and interaction-aware.

The pedestrian-only features are designed to capture the pedestrian's current state and recent motion. Corridor-relative geometry, specifically the pedestrian's position relative to the crossing corridor boundary, is included because spatial proximity to the target zone is the most direct kinematic predictor of imminent entry. Displacement, heading, and speed are included because they capture the pedestrian's directional commitment and movement pace, both of which are known to signal crossing intent in prior work (Schneemann & Heinemann, 2016; Rasouli & Tsotsos, 2020). Speed change is included to capture deceleration or hesitation patterns that often precede crossing decisions. Short-window motion summaries over recent timesteps are added to compress recent temporal history into a tabular form compatible with logistic regression and the MLP, since these models do not natively process sequences.

The interaction-aware features are designed to capture the local vehicle context surrounding the pedestrian. Nearest-vehicle distance is included as the most direct measure of whether a vehicle is close enough to plausibly affect pedestrian behavior. Nearest-vehicle speed and relative vehicle geometry, specifically the lateral offset between the pedestrian and the nearest vehicle, are included because they capture whether the vehicle is decelerating or approaching the crossing zone, which prior work suggests is relevant for pedestrian yielding decisions (Mohammadi et al., 2023; Zou et al., 2025). Vehicle type is included because heavy vehicles and motorcycles may elicit different behavioral responses than passenger cars. Local density counts of nearby vehicles are included to capture the overall traffic pressure in the scene, which may influence whether a pedestrian judges it safe to cross.

This two-perspective design supports two complementary goals. First, it allows structured testing of what the pedestrian-only signal already contains before introducing any vehicle context. Second, it enables the staged ablation analysis described in Section 3.6.2, which examines which specific interaction feature groups add predictive value rather than assuming all vehicle context is uniformly beneficial.

3.6 Model strategy

Once the target and feature perspectives are fixed, the next step is to determine how the prediction problem should be modeled. The model strategy proceeds in three layers of increasing complexity, organized so that each layer builds on the previous one and can be compared under a consistent operational framing.

3.6.1 Logistic regression as interpretable baseline

The main baseline model is logistic regression with standardized features and class-weight balancing. Logistic regression was chosen because it provides a strong and directly interpretable baseline: its coefficients can be read and compared without additional tooling,

making it possible to discuss which features the model relies on and in what direction. It is therefore both a performance benchmark and a primary transparency tool. Logistic regression is a well-established method for binary classification with tabular features (Hastie et al., 2009) and is particularly appropriate here because the operational target is a binary event label.

3.6.2 Interaction-feature ablation

To evaluate which interaction features add predictive value, the thesis uses a staged, cumulative ablation design. Starting from the pedestrian-only baseline, interaction feature groups are added one at a time in a fixed sequence: proximity, vehicle motion, relative geometry, vehicle type, and local density. Each stage adds the new group to all features already present in the previous stage, that is, stage D includes all features from stages A through C plus relative geometry. This cumulative structure was chosen over an exhaustive combinatorial search across all possible feature subsets, because the cumulative design tests a theoretically motivated ordering from simpler to richer interaction representations, and because the total number of feature combinations grows exponentially with the number of groups, making exhaustive search impractical within the project scope. The result is a structured characterization of which interaction information adds value at the margin, conditional on what is already in the model.

3.6.3 Multilayer perceptron (MLP)

A feed-forward MLP with two hidden layers and ReLU activations is included as a modest non-linear extension of the tabular modeling setup (Goodfellow et al., 2016). The MLP uses the same main feature tables as the logistic model, which makes it possible to test whether additional non-linear capacity improves performance without changing the basic data representation. This comparison is methodologically important because it distinguishes between performance gains attributable to richer model capacity and gains attributable to richer feature design. Since the MLP is not interpretable by design, its behavior is examined post-hoc using permutation importance, as described in Section 3.7.

3.6.4 A sequence model

The thesis also includes a sequence-model branch using a Gated Recurrent Unit (GRU), a recurrent neural network architecture designed to model temporal dependencies in sequential data through learned gating mechanisms that control information flow across timesteps (Cho et al., 2014). The GRU operates on fixed-length temporal windows of pedestrian observations rather than row-level tabular inputs. It is included as an informative reference point for the main tabular models, allowing the study to assess whether direct temporal modeling adds value beyond hand-crafted temporal summaries under the same operational problem framing. The sequence branch is secondary and exploratory, but its inclusion serves a clear methodological purpose: it tests the assumption that compressing temporal history into tabular features does not lose important predictive structure.

3.7 Explainability strategy

Because the thesis is concerned not only with predictive performance but also with model transparency, explainability is treated as a core methodological component rather than a final

add-on.

For logistic regression, explainability is built directly into the model through coefficient inspection, grouped feature interpretation, and structured staged ablation. These tools make it possible to discuss which features drive the logistic model and how that changes as interaction feature groups are added.

For the MLP, explainability is addressed through permutation importance, which provides a global summary of which input features the model relies on most across the full evaluation set. This is especially important because the MLP may capture non-linear patterns that the logistic model misses, and permutation importance makes those patterns partially visible without requiring a fully transparent model architecture.

For the GRU sequence model, no separate post-hoc explainability block is included. This choice is not made because the GRU performs worse than the MLP. Rather, it reflects a principled scope decision: the GRU input representation is a raw temporal sequence, and the relationship between individual timestep features and the final prediction is substantially harder to attribute using standard post-hoc tools.

Developing a dedicated sequence-level explanation for the GRU would constitute a separate methodological contribution and is therefore identified as future work rather than included here.

3.8 Data splitting and evaluation logic

To avoid leakage between training, validation, and test data, the study uses track-level splitting by pedestrian ID rather than row-level random splitting. This is essential because adjacent rows within the same trajectory are highly correlated, and row-level splitting would produce artificially optimistic results by allowing the model to see near-identical observations from the same pedestrian in both training and evaluation.

The data are divided into three non-overlapping subsets. The training set is used exclusively for model fitting. The validation set is used for model comparison, threshold selection, and all intermediate design decisions made during development. The test set is locked and used only once, at the end, for final evaluation of each chosen model configuration. This design is important because validation results are already part of the iterative development process and therefore provide an optimistic estimate of generalization; the locked test results provide the clearest evidence of how each model performs on genuinely unseen data. For interaction-aware evaluation, a fair interaction-only subset is reconstructed from the locked split IDs so that pedestrian-only and interaction-aware models are compared on exactly the same rows.

3.9 Threshold selection

Because the prediction task involves class imbalance as crossing events are relatively rare compared to non-crossing observations, and because false negatives and false positives have different practical implications in a safety-adjacent setting, model evaluation is not restricted to the default decision threshold of 0.5. A threshold of 0.5 applied to an imbalanced problem will typically suppress positive predictions and produce misleadingly high accuracy while missing a disproportionate share of true positives.

Instead, threshold sweeps are performed on the validation set across a range of candidate values, and final thresholds are selected using recall-aware criteria that reflect the anticipated cost asymmetry between missed crossings and false alarms. The selected threshold for each model is then fixed and applied once to the locked test set. This procedure provides a more operationally meaningful evaluation than default thresholding. It also allows a clear distinction between two aspects of model quality: ranking quality, reflected in area-under-curve metrics such as ROC-AUC and PR-AUC which are threshold-independent, and thresholded classification behavior, reflected in precision, recall, and F1 at the chosen operating point.

3.10 Main and sensitivity analyses

The thesis is organized around one main analysis and two sensitivity analyses, all using the same methodological pipeline.

The main analysis uses the 3-second horizon and constitutes the primary basis for all reported conclusions. It includes pedestrian-only and interaction-aware logistic regression, the structured interaction-feature ablation, MLP comparison, exploratory GRU sequence-model comparison, and MLP permutation importance analysis.

The sensitivity analyses use the 2-second and 5-second horizons respectively. These are not treated as separate full thesis pipelines. Instead, they serve as methodological stress tests: they examine how changes in the temporal definition of the target affect class balance, interaction coverage, model performance, and the apparent usefulness of interaction features. This design is central to the thesis's methodological contribution. It shows that important predictive conclusions are not model-family-specific in isolation; they also depend on how the prediction problem is framed temporally. A finding that holds across all three horizons carries more evidential weight than one that holds only at the main reporting setting.

3.11 Validity and reliability

Validity and reliability are especially important in this thesis because the task is anchored to an operational label rather than a direct ground-truth psychological construct.

3.11.1 Construct validity

The study does not claim to measure true crossing intention as a psychological state. It measures entry into a geometrically defined crossing corridor within a fixed future time window. This operationalization strengthens clarity and reproducibility but limits direct psychological interpretation. The thesis is therefore strongest when interpreted as a study of short-horizon crossing-related behavior prediction rather than deep intention recognition. This is consistent with broader guidance in the traffic-prediction literature that cautions against conflating observable short-term events with latent mental states (Vellenga et al., 2022; Rasouli & Tsotsos, 2020).

3.11.2 Internal validity

Internal validity is strengthened by several explicit design choices: geometric definitions and eligibility rules are specified before analysis; track-level train/validation/test splitting

prevents leakage; the test set is locked and used only once; and the interaction-only comparison is reconstructed from the same split IDs, ensuring that interaction-aware and pedestrian-only models face identical evaluation rows. A remaining internal-validity concern is that some geometric boundary choices and preprocessing decisions were partly informed by exploratory data inspection, which is common in applied trajectory analysis but means that those choices could in principle have been shaped by the data distribution rather than purely by prior theoretical considerations. A further consideration concerns the selected group split. Several candidate split seeds were inspected to avoid strongly imbalanced train, validation, and test partitions. This was done to avoid unstable evaluation caused by pathological group splits, not to optimize model performance. Still, the reported results remain conditional on one selected balanced split. Repeated group-split evaluation or group cross-validation would provide a stronger estimate of split-level robustness.

3.11.3 External validity

External validity is limited. All results derive from one recorded site, one stationary sensing setup, one traffic culture, and one operationalization of crossing-related behavior. Generalization to other locations, infrastructures, sensor configurations, or pedestrian populations should be treated with caution and is identified as a priority for future work (Mirzabagheri et al., 2025).

3.11.4 Reliability

The implementation is based on explicit preprocessing rules, fixed feature definitions, reproducible splits anchored to pedestrian IDs, and structured model comparisons. These choices improve repeatability. Statistical reliability is further strengthened through track-cluster bootstrap confidence intervals, which resample pedestrian trajectory IDs rather than individual rows and therefore better reflect the grouped structure of the held-out data. McNemar tests are used as paired row-level diagnostics for thresholded model comparisons on the same held-out examples, but their p-values are interpreted cautiously because neighbouring rows from the same pedestrian track are not fully independent. Exact reproducibility also depends on implementation details and execution order, particularly in notebook-based workflows, so the report and notebook are designed to document the pipeline as explicitly as possible.

3.12 Research ethics

Although the thesis uses previously collected trajectory data rather than direct experimentation on human participants, it still raises ethical questions.

First, stationary sensing and traffic surveillance involve questions of privacy, governance, and responsible data handling. Even when the final model operates on de-identified trajectories, behavior in public space is still being monitored and abstracted into predictive systems.

Second, predictive traffic models may influence future safety-system design or decision-support tools. This requires caution. A model that performs well in one site-specific setting may not generalize elsewhere, and errors may be distributed unevenly across situations.

Third, false positives and false negatives have different practical implications. Excessive false

positives can undermine trust and burden users or systems, while false negatives may fail to flag important situations. For this reason, the thesis evaluates multiple metrics rather than relying on accuracy alone.

Finally, the study avoids overstating what the model knows. Describing the target honestly as an operational short-horizon behavior proxy rather than latent intention is not only scientifically appropriate, but also ethically important.

The trajectory dataset used in this thesis was obtained through access granted by Viscando. Use of the data was subject to a non-disclosure agreement (NDA), which regulated data handling, storage, and reporting. This has influenced the presentation of the material in the thesis: only information necessary for scientific transparency and reproducibility is reported, while confidential details about the dataset, collection setup, or company-specific information are not disclosed.

The main executed thesis pipeline in this notebook starts from one canonical merged raw trajectory file. This is deliberate. The aim is to derive the pedestrian crossing-related analysis transparently from a single primary raw source rather than mixing multiple inputs or relying on prefiltered intermediate artefacts. Some additional files were also received as part of the broader data delivery, but they are not used in the main thesis pipeline presented here. The notebook therefore treats the canonical merged raw trajectory file as the sole primary input for the reported analysis.

4 Implementation

This chapter describes how the methodological design outlined in Chapter 3 was concretely instantiated in the dataset and notebook pipeline. It moves from the raw trajectory data through spatial operationalization, label construction, feature engineering, split construction, and model-branch execution. The purpose of the chapter is not to re-justify the analytical choices, but to show how they were implemented in practice.

4.1 Dataset

The raw dataset consisted of approximately 1.98 million observations with seven primary columns: object identifier, timestamp, X position, Y position, speed, actor type, and an indicator of whether the position was estimated. The observed actor types included pedestrians, bicycles, light vehicles, and heavy vehicles. The data were collected through stationary traffic sensing in an urban interaction setting in Gothenburg and provided through Viscando.

Access to the dataset was granted under a non-disclosure agreement (NDA). For that reason, the thesis reports the information necessary for scientific transparency and interpretation while omitting unnecessary confidential details about the source material and collection setup.

The raw observations were not used directly for modelling. Instead, they were transformed through the implementation pipeline described in the remainder of this chapter, including spatial filtering, label construction, feature engineering, split construction, and model evaluation. The complete executable implementation, including preprocessing code, feature

engineering, threshold selection, sensitivity analyses, and supplementary outputs, is provided in Appendix A.

4.2 Preliminary trajectory characteristics

Before constructing the modelling tables, the raw trajectories were inspected through per-track summary statistics computed separately for each actor type. These checks included mean duration, mean speed, and the distribution of inter-observation time intervals.

Pedestrian tracks were generally longer in duration and substantially lower in speed than vehicle tracks. The median inter-observation interval was centered around approximately 0.25 seconds for most actor types, indicating an effective sampling frequency of roughly 4 Hz, although the sampling was not perfectly uniform across all tracks.

These checks served as implementation diagnostics rather than substantive results. They confirmed that the dataset was structurally usable and informed several downstream design choices. The approximate 0.25-second interval was used to define short-history windows in feature engineering, to motivate the 500-millisecond scene-time bin used for interaction matching, and to confirm that the chosen GRU sequence windows captured meaningful recent motion history for most eligible pedestrian tracks.

4.3 Spatial operationalization

The spatial operationalization translated the methodological target formulation into three concrete regions: the broad interaction region, the crossing corridor, and the approach band. These regions were identified through a structured process combining site knowledge provided by Viscando with quantitative inspection of the trajectory data.

The broad interaction region was identified first by overlaying all pedestrian and vehicle trajectories on the coordinate space and selecting the contiguous area where both pedestrian and vehicle density were non-negligible. This region enclosed the part of the scene where meaningful pedestrian–vehicle co-presence occurred and was used as a coarse spatial filter for subsequent steps.

Within that broader region, the crossing corridor was defined as the rectangular area corresponding to the marked pedestrian crossing visible in the site layout. Its boundaries were aligned with the structural geometry of the crossing as reflected in the trajectory density pattern: the area where pedestrian trajectories transitioned from pre-crossing approach to active crossing movement formed a visually distinct corridor-shaped ridge.

The approach band was then defined as the region immediately adjacent to the crossing corridor on the pre-entry side. Its depth was chosen to be large enough to include pedestrians with a realistic possibility of entering the corridor within the longest studied horizon; while remaining narrow enough to exclude pedestrians whose trajectories showed no plausible progression toward the corridor.

The exact coordinate bounds of all three regions are documented in Appendix A. These spatial choices are revisited as a methodological sensitivity in Section 6.3.

4.4 Operational label construction and horizon comparison

After the spatial regions had been defined, the binary prediction target was constructed from pedestrian observations in three steps.

First, pedestrians were filtered to operationally relevant rows using the spatial setup described above. The purpose was to retain rows representing meaningful pre-entry context rather than rows already well inside the corridor or clearly unrelated to the crossing process.

Second, each candidate row was checked for sufficient future visibility. Because the target depends on what happens after the current observation, rows without enough future trajectory information could not be labelled reliably and were therefore excluded.

Third, a row was labelled positive if the same pedestrian entered the crossing corridor within the specified future horizon. Otherwise, the row was labelled negative.

This procedure produced a binary operational label representing short-horizon crossing-corridor entry. In the implementation, this label was represented as cross_soon, but conceptually it should be understood as an operational behaviour proxy rather than a direct measure of latent intention.

4.4.1 Horizon comparison

Rather than fixing a single horizon from the start, the implementation generated horizon-specific summary tables for 2, 3, and 5 seconds. For each horizon, the tables reported the number of eligible observations, positive and negative counts, positive rates, and interaction-context coverage.

This comparison served as the implementation basis for the later methodological and empirical choice of the main horizon. The 2-second setting preserved the strictest imminent interpretation but produced a sparser task, whereas the 5-second setting increased the positive rate and made the task easier while broadening the meaning of the label. On this basis, 3 seconds was selected as the main horizon for the full modelling pipeline, while 2 and 5 seconds were retained for later sensitivity analysis.

4.5 Pedestrian feature engineering

Once the labelled pedestrian table had been constructed, pedestrian-only features were extracted to represent the pedestrian's current state, recent short-history movement, and spatial relation to the crossing target.

The table below describes each implemented pedestrian-only feature, its internal name in the implementation, its unit where applicable, and its behavioral rationale:

Feature name	Unit	Meaning and rationale
speed	m/s	Current speed of the pedestrian. A key indicator of movement intensity; very low speed may

		signal hesitation or stopping.
dist_to_corridor	m	Euclidean distance from the pedestrian's current position to the nearest boundary of the crossing corridor. The most direct spatial predictor of imminent corridor entry.
estimated	binary (0/1)	Whether the current position is tracker-estimated rather than directly observed. Retained because tracking uncertainty may correlate with specific spatial situations near the corridor.
dx_past_1s	m	X-axis displacement over the past 1 second. Captures the pedestrian's lateral movement direction and pace.
dy_past_1s	m	Y-axis displacement over the past 1 second. Captures the pedestrian's longitudinal movement direction and pace.
dt_past_1s	s	Actual elapsed time over the window used to compute past-1s features. Included because the trajectory is not perfectly uniformly sampled; normalizing by actual elapsed time improves consistency.
displacement_past_1s	m	Euclidean displacement magnitude over the past 1 second. Summarizes overall movement regardless of direction.
heading_past_1s	radians	Movement heading computed over the past 1 second. Captures directional orientation toward or away from the corridor.
speed_change_past_1s	m/s	Change in speed over the past 1 second. Captures deceleration or acceleration, both of which may signal a transition in crossing intent (Schneemann & Heinemann, 2016).
mean_speed_past_2s	m/s	Mean speed over the past 2 seconds. A slightly longer-window summary of overall pace that smooths out noise in the instantaneous speed.
std_speed_past_2s	m/s	Standard deviation of speed over the past 2 seconds. Captures

		variability in movement pace, which may indicate hesitation or irregular gait.
estimated_frac_past_2s	proportion (0–1)	Fraction of observations in the past 2 seconds that were tracker-estimated. A summary measure of recent tracking quality.

Table 1. Pedestrian-only features used in the modelling pipeline. Each feature is listed with its internal implementation name, unit, and behavioral rationale. Features were extracted track by track in strict temporal order to prevent leakage.

These features were extracted track by track in strict temporal order to avoid leakage: for each pedestrian observation row, only trajectory information from the same track and from timesteps strictly prior to the current observation was used to compute historical features. Rows for which sufficient prior history was unavailable, specifically rows occurring too early in a track to populate the 2-second lookback window, were excluded from the modelling table. The resulting pedestrian modelling table was then checked for missingness and target balance before use.

4.6 Interaction feature engineering

To construct interaction-aware features, vehicle observations were first filtered to the broad interaction region and grouped by rounded scene-time bin. For each pedestrian observation row, the nearest vehicle within the same scene-time bin was identified and used to derive the nearest-vehicle interaction variables. Where no vehicle was present in the same bin, the row was recorded as having no matched vehicle context and was excluded from the interaction-only subset used for fair comparison.

4.6.1 Scene-time matching

A scene-time matching procedure was implemented by rounding timestamps to the nearest 500 milliseconds and treating pedestrian and vehicle observations falling in the same rounded bin as belonging to the same approximate local scene moment. This choice was motivated by the approximately 0.25-second inter-observation interval established in Section 4.2: a 500-millisecond bin corresponds to roughly two raw observation steps and minimizes the risk of failing to match genuinely co-occurring actors due to minor timestamp asynchrony between pedestrian and vehicle tracks. A narrower bin would have increased missed matches due to natural timestamp jitter, while a broader bin would have risked matching vehicles that no longer belonged to the same local traffic moment.

4.6.2 Derived interaction variables

For each pedestrian row with matched vehicle context, the following interaction-aware variables were derived:

Feature name	Unit	Meaning and rationale
dist_to_nearest_vehicle	m	Euclidean distance to the nearest vehicle in the scene-time bin. The most direct measure of vehicle proximity; shorter distances may

		inhibit or accelerate a pedestrian's crossing decision (Mohammadi et al., 2023).
speed_nearest_vehicle	m/s	Speed of the nearest vehicle. A decelerating or slow-moving nearest vehicle signals that the vehicle may be yielding, which can influence pedestrian crossing behavior.
rel_x_nearest_vehicle	m	Relative X position of the nearest vehicle with respect to the pedestrian. Encodes lateral geometry rather than pure scalar distance, distinguishing between vehicles to the left and right of the pedestrian.
rel_y_nearest_vehicle	m	Relative Y position of the nearest vehicle with respect to the pedestrian. Encodes longitudinal geometry, distinguishing between vehicles ahead of and behind the pedestrian relative to the crossing direction.
type_nearest_vehicle	categorical	Actor type of the nearest vehicle (light vehicle, heavy vehicle, bicycle). Different vehicle types may elicit different behavioral responses from pedestrians.
is_heavy_nearest_vehicle	binary (0/1)	Derived indicator of whether the nearest vehicle is classified as heavy. Heavy vehicles have larger physical profiles and may present a stronger deterrent to crossing.
vehicle_count_r1	count	Number of vehicles within a fixed inner radius of the pedestrian. Captures immediate local traffic density.
vehicle_count_r2	count	Number of vehicles within a larger outer radius of the pedestrian. Captures broader scene-level traffic pressure.

Table 2. Interaction-aware features derived from matched nearest-vehicle context. Each feature captures a distinct dimension of the local vehicle scene: proximity, vehicle dynamics, relative spatial geometry, vehicle type, and local density.

These variables were designed to represent distinct and complementary dimensions of the local interaction context: immediate proximity, vehicle behavior, relative spatial geometry, vehicle character, and overall scene density. This multi-dimensional design supported the staged ablation study described in Section 4.8, where each dimension was tested separately rather than treating all vehicle context as a single undifferentiated block.

4.6.3 Interaction coverage

Not all eligible pedestrian rows had matched vehicle context at the same scene time. This produced two distinct tables used throughout the modelling pipeline: the full pedestrian modelling table containing all eligible rows regardless of vehicle context, and the interaction-only subset containing only rows where matched vehicle context existed. This distinction was preserved throughout all subsequent comparisons and is reflected in the split and evaluation setup described in Section 4.7.

4.7 Data splitting and evaluation setup

The tabular data were split at pedestrian-track level using ID-based partitioning. The target split size was approximately 70% training, 15% validation, and 15% test. To reduce imbalance across partitions, 300 random seeds were tested and the seed minimizing the largest positive-rate deviation across train, validation, and test was selected. This seed-selection procedure was used only to avoid a pathological group split with strongly distorted class balance across train, validation, and test partitions. The seed was not selected based on model performance, validation scores, test scores, or downstream conclusions. Nevertheless, because the final split was selected after inspecting label balance, the test partition should be understood as a balanced held-out group split rather than as a purely arbitrary random draw. This choice improves comparability between model branches, but it also means that the results remain conditional on one selected partition. A more extensive evaluation using repeated group splits or group cross-validation would provide a stronger estimate of split-level robustness. The resulting split was then locked and reused throughout the notebook.

For the pedestrian-only modelling table, the locked split produced separate training, validation, and test partitions containing non-overlapping pedestrian tracks. For interaction-aware evaluation, the same locked pedestrian IDs were used to reconstruct the interaction-only subset so that pedestrian-only and interaction-aware models could later be compared on the same underlying tracks. All model fitting was performed on the training partition. The validation partition was used for threshold selection and intermediate model comparison. The test partition remained locked until final evaluation.

4.8 Logistic-regression branch

The logistic-regression branch served as the main interpretable model family in the thesis.

Pedestrian-only baseline
A logistic-regression baseline was trained on the pedestrian-only feature set using standardized features and class balancing. This model represented the core interpretable benchmark against which more complex feature sets and model families were later compared.

Interaction-aware comparison
A fair interaction-aware comparison was then conducted on the interaction-only subset. Here, the pedestrian-only logistic model was compared with an interaction-aware logistic model using the same rows and splits. This step was important because it separated the question of whether interaction context helps when available from the different question of

how well the pedestrian-only model performs on the larger overall table.

Structured interaction-feature ablation
The implementation included a staged, cumulative ablation study in which interaction feature groups were added one at a time to all features already present in the previous stage. The stages were as follows: stage A included pedestrian-only features; stage B added proximity features (nearest-vehicle distance); stage C added vehicle motion features (nearest-vehicle speed) to stage B; stage D added relative geometry features (rel_x, rel_y) to stage C; stage E added vehicle type and heavy-vehicle indicator to stage D; and stage F added local density counts to stage E. Each stage therefore includes all features from all previous stages plus the new group.

Threshold tuning and row-level comparison
Threshold sweeps were performed on the validation set, and final thresholds were chosen according to recall-aware criteria. These selected thresholds were then applied once to the locked test set. This produced a cleaner estimate of thresholded test performance than relying only on default thresholding.

4.9 Horizon-sensitivity implementation

After the main 3-second modelling branches had been established, the implementation returned to the 2-second and 5-second horizons in a dedicated sensitivity section. For each horizon, the pedestrian and interaction modelling tables were rebuilt using the same operational labeling procedure described in Section 4.4. The previously locked pedestrian-track IDs were then reused to reconstruct training, validation, and test partitions, ensuring consistency with the main-horizon splits.

On these horizon-specific tables, the same comparison pipeline was applied in compact form: the horizon-specific label-balance and interaction-coverage summary, the logistic pedestrian-only versus interaction-aware comparison, the staged cumulative interaction-feature ablation, and the MLP pedestrian-only versus interaction-aware comparison. The outputs were collected into horizon-level summary tables used in the Results chapter.

4.10 MLP branch

The MLP branch extended the tabular modelling pipeline with a modest non-linear model. A feed-forward multilayer perceptron was trained on the same tabular feature tables as the logistic models so that any performance difference could be interpreted relative to the same underlying feature representation.

At the main 3-second horizon, the implementation compared two MLP variants:

- a pedestrian-only MLP
- a full interaction-aware MLP

Both models were evaluated under the same split logic, threshold-selection procedure, and metric suite as the logistic branch.

4.10.1 MLP explainability

After the main-horizon MLP comparison had identified the stronger branch, permutation importance was computed for that model on the held-out evaluation set. Feature importance was summarized by the mean drop in predictive performance when each feature was randomly permuted. This provided a global ranking of which input variables contributed most strongly to the selected MLP branch.

4.11 GRU sequence-model branch

The sequence-model branch was implemented as a compact extension rather than a broad architecture search across multiple recurrent families. Unlike the tabular branches, the GRU operated on fixed-length windows of consecutive pedestrian observations. Each sequence example was anchored at the last timestep in the window, and the same short-horizon crossing-corridor-entry label was applied relative to that endpoint.

At the main 3-second horizon, the implementation compared:

- a pedestrian-only GRU
- a full interaction-aware GRU

The GRU branch provided a sequence-model comparison under the same overall operational problem framing as the logistic and MLP branches. However, the comparison is not a pure test of sequence models against tabular models in general. The GRU used raw fixed-length observation windows, whereas the tabular logistic and MLP models used engineered temporal summaries such as recent displacement, heading, speed change, and mean speed. The branch should therefore be interpreted as a comparison between this compact raw-sequence GRU implementation and the engineered tabular representation, not as a general verdict on recurrent architectures.

No dedicated post-hoc explainability block was implemented for the GRU. This reflects a scope decision rather than a judgment about the legitimacy of sequence-model interpretation. Sequence-level attribution for recurrent architectures is methodologically distinct from tabular feature importance and would require a separate explanation framework beyond the scope of the present study.

4.12 Cross-model synthesis

After the logistic, MLP, and GRU branches had been evaluated at the main 3-second horizon, the implementation combined the main representatives from each family into a compact cross-model synthesis. This synthesis summarized the locked-test results side by side and enabled direct comparison across:

- interpretable linear models
- modest non-linear tabular models
- compact recurrent sequence models

The resulting table provided a direct locked-test comparison across the three model families under the same overall problem framing.

4.13 Uncertainty and paired comparison checks

After the main 3-second cross-model comparison had been established, the implementation

added an uncertainty section consisting of track-cluster bootstrap confidence intervals for the main held-out metrics and paired diagnostic tests for thresholded tabular-model comparisons on the same test examples.

The bootstrap analysis resampled pedestrian trajectory IDs rather than individual observation rows. For each sampled track ID, all held-out rows belonging to that trajectory were included in the bootstrap sample. This was done because multiple rows can originate from the same pedestrian track and are therefore not fully independent. The paired McNemar tests were retained as row-level diagnostics of whether thresholded models made different errors on the same examples, but their p-values are interpreted cautiously because they do not fully model within-track dependence. The notebook also reports track-cluster paired bootstrap intervals for selected tabular metric differences.

4.14 Performance measures

Before presenting the results, the performance measures used in this chapter are briefly defined. Six metrics are reported for each model: accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC.

Accuracy is the proportion of correct predictions, but it is not a strong primary metric in imbalanced classification because a model can achieve high accuracy by overpredicting the majority class (Saito & Rehmsmeier, 2015). Precision measures the proportion of predicted positive cases that are correct, while recall measures the proportion of true positive cases that are correctly identified. F1-score is the harmonic mean of precision and recall and provides a single summary of thresholded classification performance.

ROC-AUC is the area under the receiver operating characteristic curve and summarizes threshold-independent class separability across all possible thresholds. It is retained because it remains useful for broad model comparison. However, in imbalanced settings it can appear overly favorable, since the false positive rate is normalized by the large number of negative cases (Saito & Rehmsmeier, 2015). For that reason, ROC-AUC is treated here as a complementary metric rather than the main basis for judging model quality.

PR-AUC is the area under the precision-recall curve and is more sensitive to minority-class performance. It is therefore more informative when the positive class is relatively rare and when correct identification of positive cases is the main concern (Saito & Rehmsmeier, 2015; Davis & Goadrich, 2006). For this reason, PR-AUC is treated as the primary ranking metric in this thesis, while precision, recall, and F1-score are used to evaluate thresholded behaviour at the chosen operating point.

Confusion matrices are reported alongside the metric tables to show the raw counts of true positives, false positives, true negatives, and false negatives at the selected threshold. This makes it possible to inspect how each model distributes its errors in absolute terms, which is particularly useful in an imbalanced setting where small threshold changes can substantially alter the error profile.

5 Results

This chapter reports the empirical findings in three stages. It begins with the spatial audit, horizon comparison, and main-horizon dataset characteristics. It then presents the main 3-

second model results for logistic regression, MLP, and GRU, followed by cross-model synthesis and uncertainty checks. The chapter ends with the horizon-sensitivity analysis and a summary of the main findings.

5.1 Spatial audit and operational geometry

The first stage of the analysis was a spatial audit of the scene in order to identify the parts of the environment most relevant to pedestrian–vehicle interaction. Within the broad interaction region, the dataset contained 1,635 pedestrian tracks, 10,863 bicycle tracks, 37,166 light-vehicle tracks, and 1,539 heavy-vehicle tracks. In row terms, this corresponded to 44,210 pedestrian observations, 160,000 bicycle observations, 412,131 light-vehicle observations, and 18,259 heavy-vehicle observations.

The operational geometry was then refined into a crossing corridor and a candidate approach band. The crossing corridor contained 919 pedestrian tracks, 8,136 bicycle tracks, 33,834 light-vehicle tracks, and 1,336 heavy-vehicle tracks. The candidate approach band was broader and contained 1,208 pedestrian tracks, 9,652 bicycle tracks, 34,904 light-vehicle tracks, and 1,448 heavy-vehicle tracks. Among pedestrians entering the crossing corridor, 179 were approximately classified as left-to-right traversals and 115 as right-to-left traversals, while 625 were classified as other, partial, or ambiguous trajectories. This confirms that the corridor captures a substantial amount of crossing-related pedestrian activity, but also that many trajectories remain operationally ambiguous.

The coordinate layout of the full roundabout site, including the relative position of the crossing corridor within the broader intersection geometry, is documented in Appendix A together with the trajectory density heatmaps used to guide the spatial operationalization. Readers unfamiliar with the site are encouraged to consult those figures before proceeding to the label and feature results.

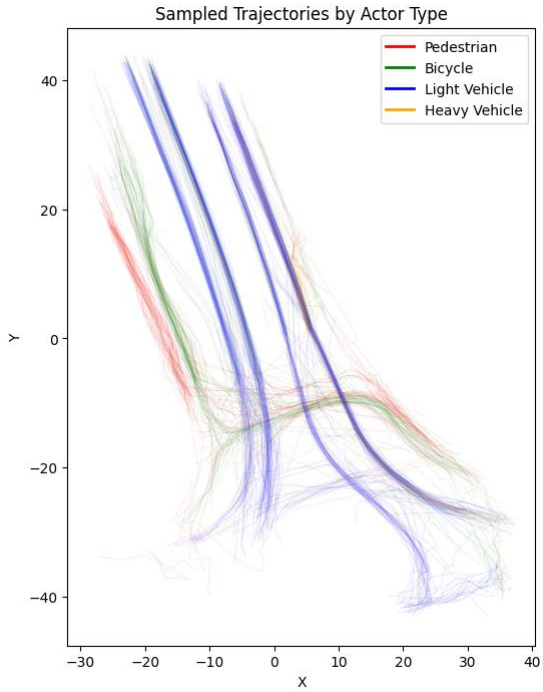

Figure 1. Sampled trajectories in the broad interaction region. The figure shows the spatial distribution of pedestrians, bicycles, light vehicles, and heavy vehicles within the broad operational scene and motivates the restriction to a crossing-focused analysis area.

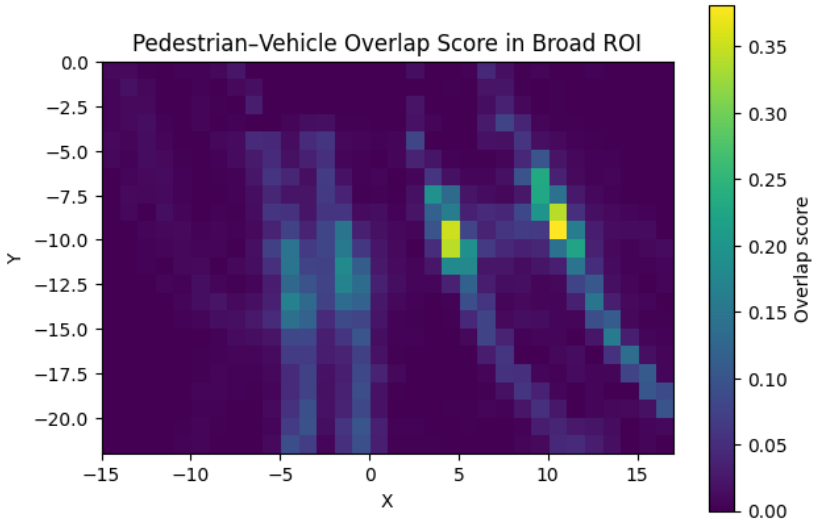

Figure 2. Pedestrian–vehicle overlap heatmap. The figure visualizes the densest overlap between pedestrian and vehicle movement and supports the definition of the final crossing-related geometry.

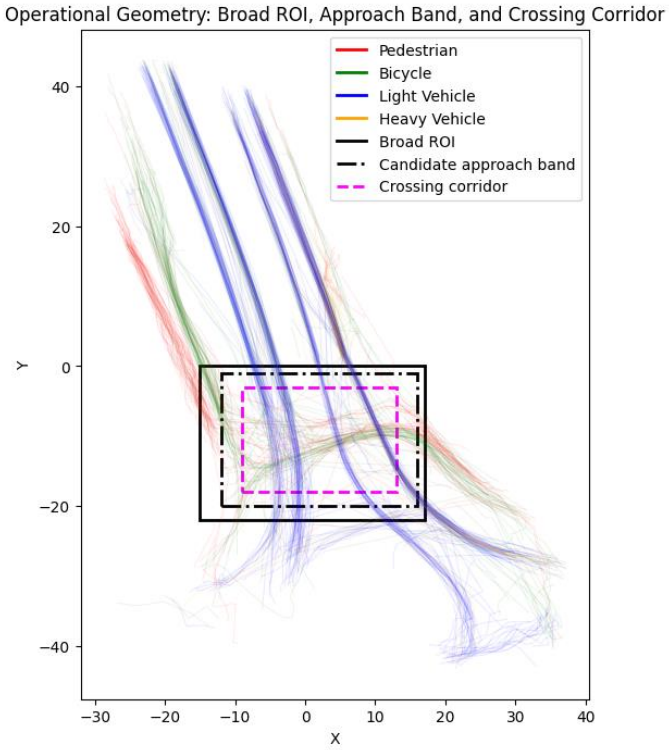

Figure 3. Final operational geometry: broad interaction region, candidate approach band, and crossing corridor. This figure summarizes the spatial setup used later for label construction and feature engineering.

5.2 Horizon comparison and main-horizon choice

Before committing to a single main prediction horizon, the operational label was compared across 2-second, 3-second, and 5-second settings. The resulting label statistics are summarized in Table 3. As the table shows, increasing the horizon increased the proportion of positive cases while reducing the number of valid observations, a pattern that reflects the broader semantic shift in what the label captures as the future window expands. This confirms that horizon choice affects both the meaning of the target and the resulting class balance. The horizon comparison confirmed that 3 seconds offered the empirically best-supported compromise between immediacy and learnability in this dataset: it retained a

positive rate large enough to produce stable model training while preserving a short enough horizon to remain behaviorally meaningful, as anticipated in Section 3.4, and was therefore selected as the main executed setting.

	horizon_sec	n_valid	n_positive	n_negative	positive_rate
0	2	9044	2414	6630	0.266917
1	3	8537	2856	5681	0.334544
2	5	7760	2932	4828	0.377835

Table 3. Horizon comparison: valid observations, positive and negative counts, and positive rates across 2, 3, and 5 seconds.

5.3 Pedestrian-only modelling table and interaction coverage at the 3-second horizon

After pedestrian-only feature extraction at the main 3-second horizon, the final pedestrian modelling table contained 7,530 observations and 27 columns. The target distribution at this stage was 5,274 negative and 2,256 positive observations, corresponding to a positive rate of 0.2996. No missing values remained in the pedestrian feature columns used for modelling.

When vehicle context was added through scene-time matching, 3,628 observations had matched interaction context and 3,902 observations did not. This means that 48.2% of the pedestrian rows had usable vehicle context. The resulting interaction-only subset contained 3,628 observations, of which 1,069 were positive and 2,559 negatives, corresponding to a positive rate of 0.2947. A further diagnostic of the 500 ms scene-time matching procedure showed that the match rate was similar for positive and negative classes. Among matched rows, 3.6% had at least two vehicles within 5 m, while 29.1% had at least two vehicles within 10 m. This suggests that very dense immediate multi-vehicle scenes were uncommon, while somewhat broader local multi-vehicle scenes were more frequent.

5.4 Train/validation/test split

In the pedestrian-only table, the final split contained 440 training IDs, 94 validation IDs, and 95 test IDs. In row terms, this corresponded to:

- Training: 5,134 rows, positive rate 0.3081
- Validation: 887 rows, positive rate 0.3315
- Test: 1,509 rows, positive rate 0.2518

For the interaction-only subset, the split contained 298 training IDs, 63 validation IDs, and 65 test IDs, corresponding to:

- Training: 2,471 rows, positive rate 0.3116
- Validation: 441 rows, positive rate 0.2993
- Test: 716 rows, positive rate 0.2332

This confirms that the interaction-aware models were trained and evaluated on a substantially smaller subset than the full pedestrian-only models.

5.5 Logistic-regression results at the 3-second horizon

5.5.1 Initial baseline comparison

The initial logistic comparison at the main 3-second horizon contrasted a pedestrian-only baseline with an interaction-aware version on the full modelling table.

On validation data, the pedestrian-only logistic model achieved:

- Accuracy = 0.717
- Precision = 0.555
- Recall = 0.735
- F1 = 0.633
- ROC-AUC = 0.760
- PR-AUC = 0.509

The corresponding interaction-aware logistic model achieved:

- Accuracy = 0.748
- Precision = 0.559
- Recall = 0.750
- F1 = 0.641
- ROC-AUC = 0.769
- PR-AUC = 0.505

This first comparison indicated that interaction-aware information was associated with slightly improved thresholded validation performance, although the PR-AUC difference was small (0.509 vs. 0.505) and should be interpreted as a preliminary signal rather than a confirmed advantage. The question of whether these differences are statistically meaningful is addressed in Section 5.12, where paired McNemar tests and bootstrap confidence intervals are reported. The confusion matrix (Figure 4) shows the distribution of true positives, false positives, true negatives, and false negatives at the default threshold for each model, making it possible to see that both models produce a similar error profile at this stage, with the interaction-aware model recovering slightly more true positives at the cost of marginally more false positives.

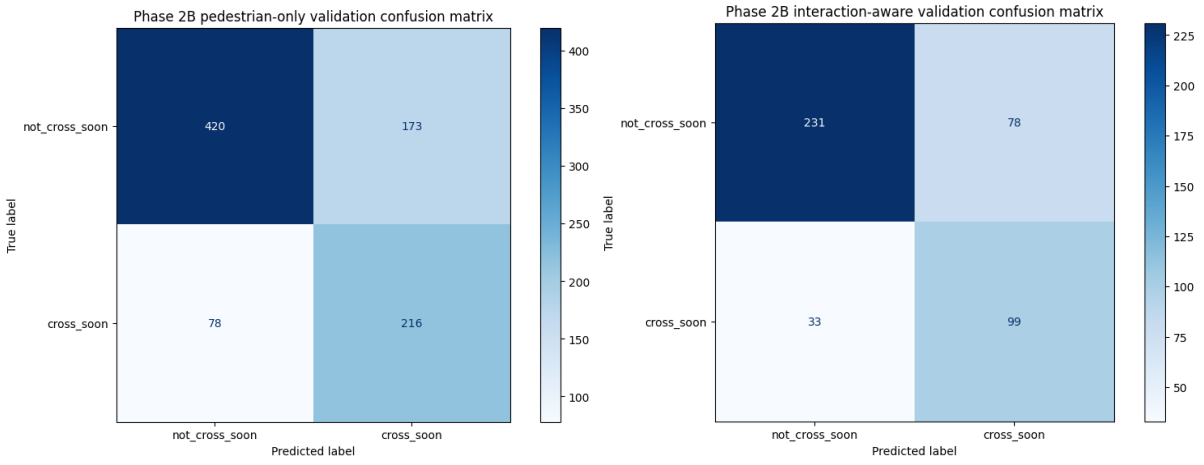

Figure 4. Confusion matrices for the pedestrian-only and for Interaction-aware logistic baselines at the 3-second horizon.

5.5.2 Fair interaction-only comparison

Because the full pedestrian-only table and the interaction-only subset represent different data regimes, the latter containing only rows where matched vehicle context was available, a

fairer comparison was conducted by evaluating both models on the same interaction-only validation subset. This is the primary logistic-regression comparison reported in this thesis.

On this fair subset, the pedestrian-only logistic model achieved:

- Accuracy = 0.737
- Precision = 0.548
- Recall = 0.697
- F1 = 0.613
- ROC-AUC = 0.750
- PR-AUC = 0.480

The interaction-aware logistic model achieved:

- Accuracy = 0.748
- Precision = 0.559
- Recall = 0.750
- F1 = 0.641
- ROC-AUC = 0.769
- PR-AUC = 0.505

The confusion matrices (Figure 5) illustrate how the error profiles shift when both models are evaluated on the same rows. The interaction-aware model recovers more true positives (higher recall) while maintaining comparable precision, which accounts for the higher F1. These results feed directly into the ablation study in Section 5.6, which examines which specific interaction feature groups are responsible for this improvement.

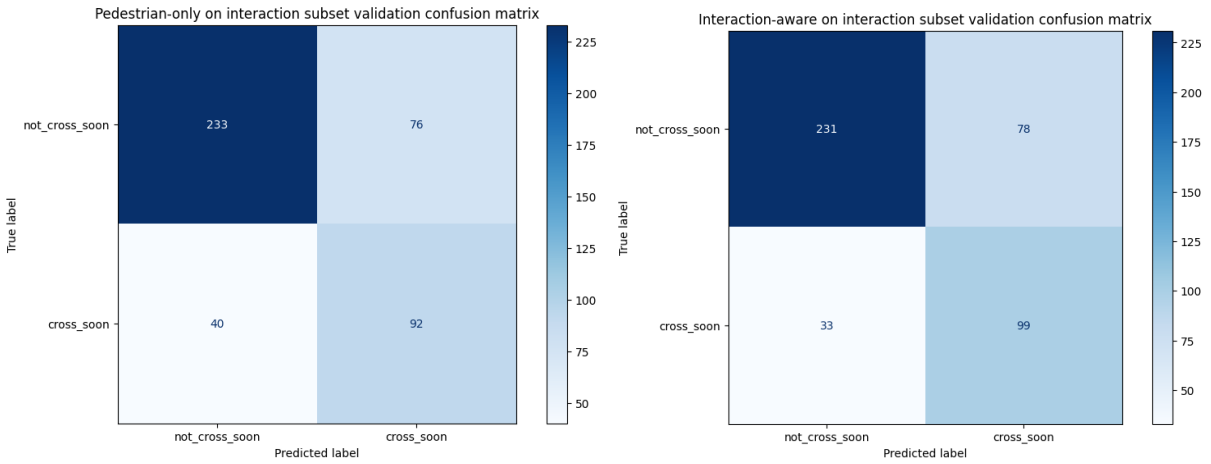

Figure 5. Confusion matrices for the pedestrian-only and for Interaction-aware logistic baselines trained on the same interaction-only subset at the 3-second horizon.

5.5.3 Threshold tuning and locked test evaluation

Using the validation-selected recall-aware thresholds described in Section 3.9, the logistic models were then evaluated once on the locked interaction-only test set. The selected validation thresholds were 0.33 for the pedestrian-only logistic model and 0.31 for the interaction-aware logistic model.

On the locked interaction-only test set, the pedestrian-only logistic model achieved:

- Accuracy = 0.719

- Precision = 0.443
- Recall = 0.784
- F1 = 0.566
- ROC-AUC = 0.840
- PR-AUC = 0.632

The interaction-aware logistic model achieved:

- Accuracy = 0.728
- Precision = 0.456
- Recall = 0.862
- F1 = 0.596
- ROC-AUC = 0.860
- PR-AUC = 0.668

At the main 3-second horizon, the interaction-aware logistic branch outperformed the pedestrian-only logistic branch on all reported locked-test metrics, with the largest numerical differences appearing in recall and PR-AUC. These point-estimate differences are directionally consistent with the validation results. However, the question of whether the pedestrian-only versus interaction-aware logistic difference reaches statistical significance is examined in Section 5.12: as reported there, the McNemar test for this pair does not reach significance after Holm correction ($p = 0.60$), which means the within-family logistic comparison should be interpreted as a consistent directional pattern rather than a confirmed statistically significant advantage. The confusion matrix (Figure 6) shows the distribution of errors at the recall-aware threshold for each model. Note that these results are directly comparable to ablation stage F reported in Section 5.6, which uses the full interaction-aware feature set; the two sets of results should be consistent, and any apparent discrepancy reflects the fact that 5.5.3 reports locked test results whereas the ablation in 5.6 reports validation results.

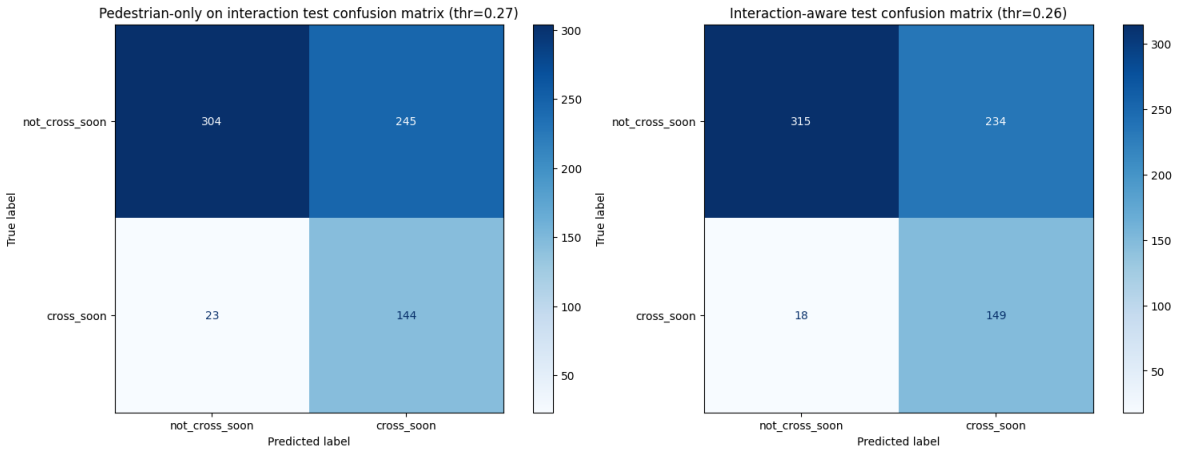

Figure 6. Locked-test confusion matrices for the threshold-tuned pedestrian-only and interaction-aware logistic models at the 3-second horizon.

5.6 Logistic coefficient inspection and interaction-feature ablation

Inspection of the standardized logistic coefficients revealed that both the pedestrian-only and interaction-aware models were driven mainly by pedestrian motion and corridor-relative

geometry. In the pedestrian-only model, the strongest coefficients were associated with dx_past_1s, mean_speed_past_2s, heading_past_1s, estimated_frac_past_2s, and dist_to_corridor. In the interaction-aware model, strong coefficients again included pedestrian kinematic features, but also nearest_vehicle_speed, while nearest_vehicle_distance itself had a comparatively modest coefficient magnitude. The coefficient plots are shown in Figure 7 (pedestrian-only model and interaction-aware model), which allow direct visual inspection of the direction and relative magnitude of each feature's contribution.

The staged cumulative ablation results on the interaction-only validation subset are summarized in Table 4 below. Each stage adds the listed feature group to all features already present in the previous stage: stage B adds proximity to stage A (pedestrian-only); stage C adds vehicle motion to stage B; stage D adds relative geometry to stage C; stage E adds vehicle type to stage D; and stage F adds local density to stage E.

Stage	**Features added**	**F1**	**PR-AUC**
A	Pedestrian-only	0.613	0.480
B	+ Proximity	0.609	0.481
C	+ Vehicle motion	0.627	0.498
D	+ Relative geometry	0.628	0.507
E	+ Vehicle type	0.641	0.505
F	+ Local density	0.626	0.506

Table 4. Staged cumulative ablation results on the interaction-only validation subset at the 3-second horizon.

The ablation results show that simple proximity alone (stage B) did not improve the baseline and even reduced F1 marginally. Meaningful improvement began with vehicle motion (stage C), which added the nearest vehicle's speed to the model. Stage D added relative geometry (rel_x and rel_y of the nearest vehicle), which produced the best PR-AUC of the ablation series (0.507). Notably, stage D's PR-AUC (0.507) exceeds stage A's (0.480), meaning that relative geometry adds value relative to the pedestrian-only baseline, but the PR-AUC at stage D is slightly lower than at stage E (0.505) and F (0.506) on F1, which reflects the non-monotone pattern typical of cumulative ablations where adding features can interact with the logistic model's regularization in non-linear ways. The apparent inconsistency between the interaction-aware logistic results in Section 5.5.3 and ablation stage F here is explained by the fact that Section 5.5.3 reports locked test set results at a recall-aware threshold, whereas Table 4 reports validation set results at the default threshold; these are different evaluation settings and differences between them are expected. The best validation F1 was obtained at stage E (vehicle type), while the best validation PR-AUC was obtained at stage D (relative geometry), indicating that richer interaction structure beyond simple proximity was consistently more informative at the main 3-second horizon. Figure 8 shows the full ablation profile visually.

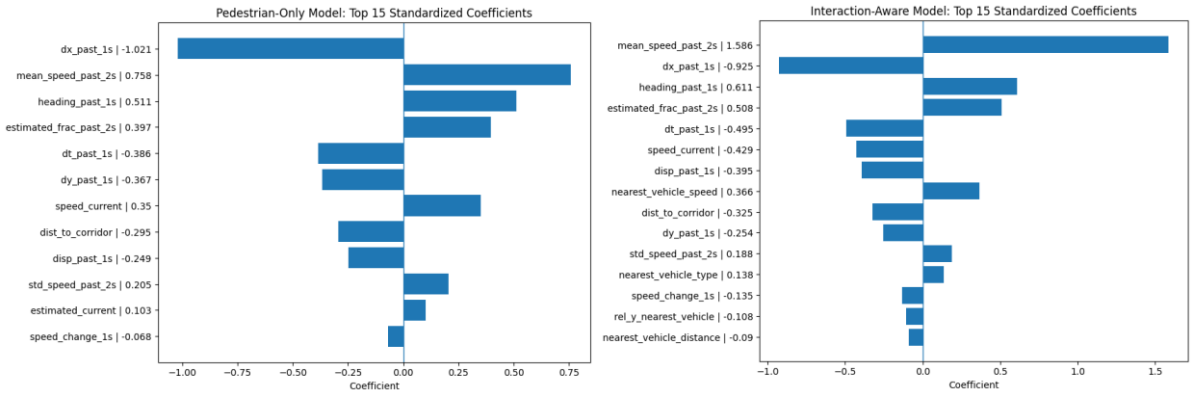

Figure 7. Standardized coefficient plots for the pedestrian-only and interaction-aware logistic models at the 3-second horizon.

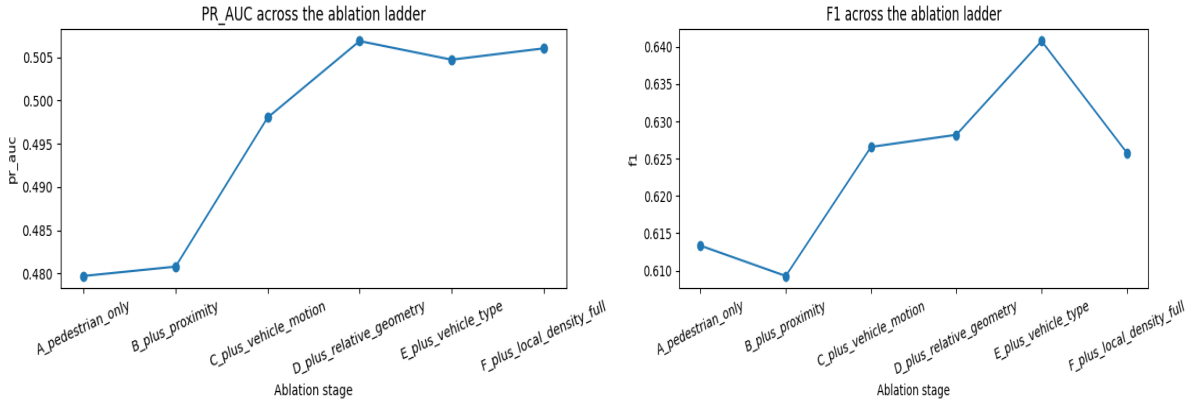

Figure 8. Validation ablation results across stages A–F at the 3-second horizon. The figure shows F1 and PR-AUC for the staged interaction-feature additions.

5.7 MLP results at the 3-second horizon

A small feed-forward MLP was trained as a non-linear comparison using the same main 3-second tabular feature views.

On validation data at threshold 0.50, the pedestrian-only MLP achieved:

- Accuracy = 0.812
- Precision = 0.652
- Recall = 0.795
- F1 = 0.717
- ROC-AUC = 0.888
- PR-AUC = 0.794

The interaction-aware MLP achieved:

- Accuracy = 0.832
- Precision = 0.681
- Recall = 0.826
- F1 = 0.747
- ROC-AUC = 0.904

- PR-AUC = 0.770

After recall-constrained threshold selection, the locked-test thresholds were:

- Pedestrian-only MLP: threshold = 0.36
- Interaction-aware MLP: threshold = 0.32

On the locked test set, the pedestrian-only MLP achieved:

- Accuracy = 0.860
- Precision = 0.660
- Recall = 0.826
- F1 = 0.734
- ROC-AUC = 0.931
- PR-AUC = 0.806

The interaction-aware MLP achieved:

- Accuracy = 0.876
- Precision = 0.695
- Recall = 0.832
- F1 = 0.757
- ROC-AUC = 0.949
- PR-AUC = 0.881

This is the strongest result in the main 3-second comparison. The interaction-aware MLP outperformed the pedestrian-only MLP across all locked-test metrics, and both MLP branches clearly outperformed the logistic models. This means that the main performance gain at the 3-second horizon came from non-linearity, not from the added simple proximity feature. Full output tables and supplementary diagnostics are provided in Appendix A.

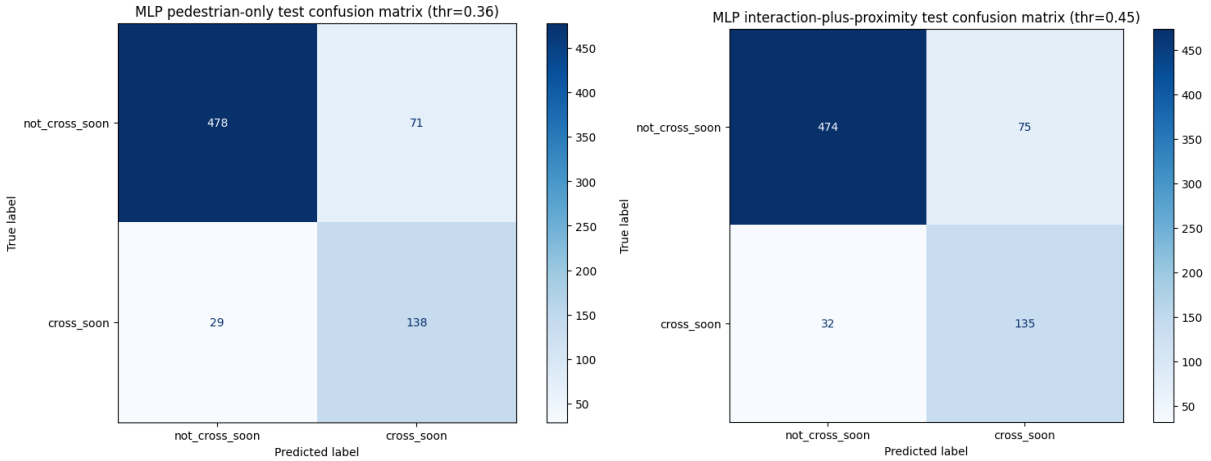

Figure 9. Locked-test confusion matrices for the pedestrian-only and interaction-plus-proximity MLP branches at the 3-second horizon.

5.8 MLP explainability

Permutation importance was computed for the stronger MLP branch, which was the interaction-aware MLP.

The most important features by mean permutation importance were:

- rel_x_nearest_vehicle: 0.187
- dx_past_1s: 0.154
- speed_current: 0.133
- dist_to_corridor: 0.074
- heading_past_1s: 0.068

Other positive contributors included:

- std_speed_past_2s
- dy_past_1s
- estimated_frac_past_2s
- speed_change_1s
- nearest_vehicle_speed
- nearest_vehicle_distance

These results suggests that the stronger MLP branch relied on a combination of pedestrian motion, corridor-relative geometry, and selected interaction terms, with relative vehicle geometry playing the most prominent role. The dominance of rel_x_nearest_vehicle is especially notable because it is consistent with the broader pattern that richer interaction structure appears more informative than simple proximity alone.

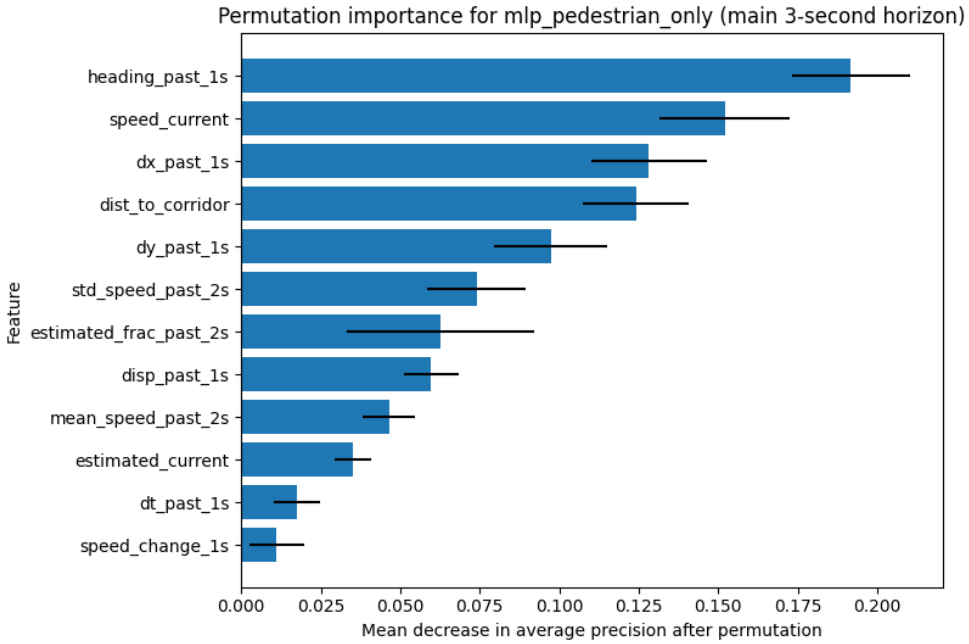

Figure 10. Permutation-importance ranking for the interaction-aware MLP branch at the 3-second horizon.

5.9 GRU results at the 3-second horizon

A GRU was trained as the compact sequence-model comparison. The sequence dataset had the following structure for both views:

- sequence length = 8
- pedestrian-only features per timestep = 12
- interaction-aware features per timestep = 17
- training sequences = 1,083
- validation sequences = 157

- test sequences = 457

The positive rates were 0.115 in training, 0.197 in validation, and 0.098 in test, indicating a substantially smaller and more imbalanced setting than the tabular branches.

At threshold 0.50 on validation data, the interaction-aware GRU outperformed the pedestrian-only GRU, achieving:

- Accuracy = 0.688
- Precision = 0.375
- Recall = 0.871
- F1 = 0.524
- ROC-AUC = 0.871
- PR-AUC = 0.611

However, after recall-aware threshold selection and locked test evaluation, this pattern did not hold. The pedestrian-only GRU achieved threshold:

- Accuracy = 0.732
- Precision = 0.415
- Recall = 0.871
- F1 = 0.563
- ROC-AUC = 0.875
- PR-AUC = 0.683

On the locked test set, after threshold selection, the pedestrian-only GRU achieved:

- Threshold = 0.52
- Accuracy = 0.880
- Precision = 0.436
- Recall = 0.756
- F1 = 0.553
- ROC-AUC = 0.923
- PR-AUC = 0.605

Whereas the interaction-aware GRU achieved:

- Threshold = 0.63
- Accuracy = 0.864
- Precision = 0.387
- Recall = 0.644
- F1 = 0.483
- ROC-AUC = 0.895
- PR-AUC = 0.565

This suggests that the GRU models were informative, but that the validation-selected threshold transferred less stably than in the tabular branches, especially for the interaction-aware GRU. For the pedestrian-only GRU, threshold selection traded some recall for improved precision and a slightly higher F1. For the interaction-aware GRU, by contrast, the locked threshold reduced recall substantially without sufficient precision gain. Thus, the GRU branch remained a useful sequence-model comparison, but it did not outperform the MLP, and its operating-point stability was weaker than that of the tabular models. Full output tables and training diagnostics are available in Appendix A.

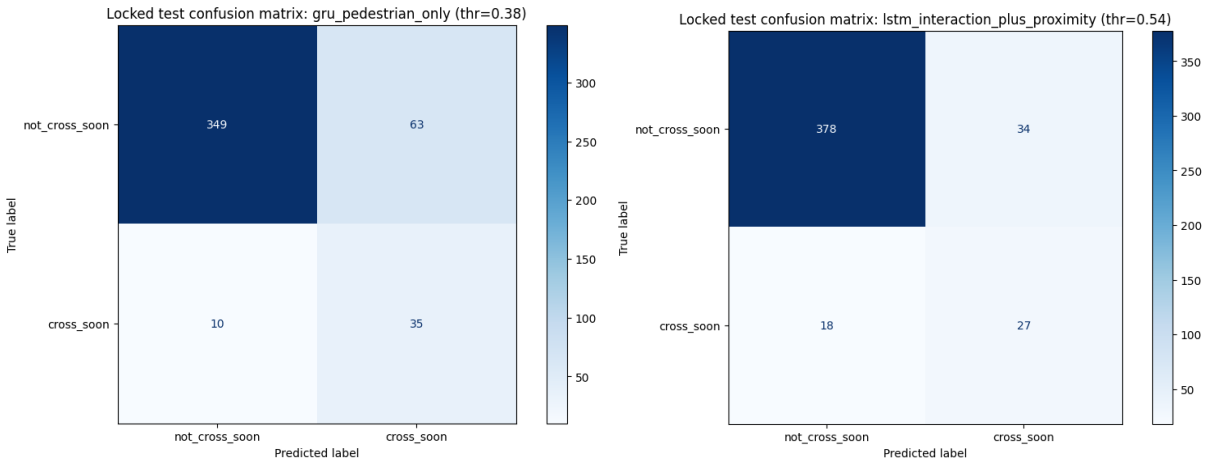

Figure 11. Locked-test confusion matrices for the pedestrian-only GRU and interaction-aware GRU at the 3-second horizon.

5.10 Cross-model synthesis at the 3-second horizon

To compare the three model families directly, the main locked-test results were synthesized side by side.

The strongest results were:

- MLP, interaction-aware, F1 = 0.757, ROC-AUC = 0.949, PR-AUC = 0.881
- MLP, pedestrian-only, F1 = 0.734, ROC-AUC = 0.931, PR-AUC = 0.806
- Logistic, interaction-aware, F1 = 0.596, ROC-AUC = 0.860, PR-AUC = 0.668
- Logistic, pedestrian-only, F1 = 0.566, ROC-AUC = 0.840, PR-AUC = 0.632
- GRU, pedestrian-only, F1 = 0.553, ROC-AUC = 0.923, PR-AUC = 0.605
- GRU, interaction-aware, F1 = 0.483, ROC-AUC = 0.895, PR-AUC = 0.565

The strongest main-horizon result was therefore the interaction-aware MLP. The MLP family outperformed both logistic regression and GRU on the locked 3-second comparison, while the GRU branch did not surpass even the stronger logistic interaction-aware result on F1 or PR-AUC.

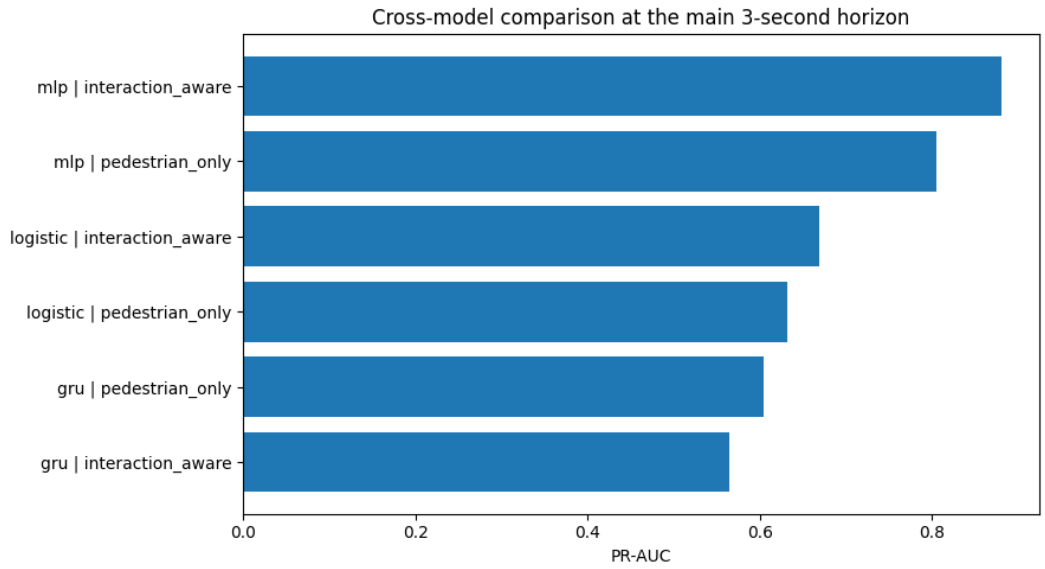

Figure 12. Locked-test cross-model summary at the main 3-second horizon

5.11 Horizon sensitivity

The horizon-sensitivity analysis was included to examine how the main conclusions changed when the target was tightened to 2 seconds or broadened to 5 seconds, relative to the main executed 3-second setting. The analysis reports changes in class balance and interaction coverage, the locked logistic comparison, the strongest interaction-feature ablation stage, and the locked MLP comparison. The purpose is to show how the empirical conclusions vary across horizons rather than to re-justify the horizon strategy itself.

5.11.1 Label balance and interaction coverage by horizon

The data properties changed systematically across horizons, the pedestrian and interaction tables changed as follows:

- 2 seconds: pedestrian rows = 7,979, pedestrian positive rate = 0.235; interaction rows = 3,869, interaction positive rate = 0.224; interaction coverage = 0.485
- 3 seconds: pedestrian rows = 7,530, pedestrian positive rate = 0.300; interaction rows = 3,628, interaction positive rate = 0.295; interaction coverage = 0.482
- 5 seconds: pedestrian rows = 6,852, pedestrian positive rate = 0.339; interaction rows = 3,226, interaction positive rate = 0.341; interaction coverage = 0.471

These quantities show two important trends. First, the positive rate increased as the horizon became longer, which is expected because a broader future window captures more eventual corridor-entry events. Second, interaction coverage declined slightly as the horizon increased, meaning that longer-horizon tasks were based on somewhat fewer matched interaction rows. Thus, horizon choice changed both the semantic meaning of the target and the structure of the available modelling tables.

5.11.2 Logistic comparison across horizons

For the fair locked-test logistic comparison, the relationship between the pedestrian-only and interaction-aware views changed across horizons rather than pointing to one uniformly dominant setting.

At 2 seconds, the pedestrian-only logistic model achieved:

- F1 = 0.572
- ROC-AUC = 0.868
- PR-AUC = 0.564,

while the interaction-aware model achieved:

- F1 = 0.558
- ROC-AUC = 0.871
- PR-AUC = 0.592.

Thus, at the strictest horizon, interaction-aware context improved the ranking metrics slightly, especially PR-AUC, but did not improve F1.

At 3 seconds, the pedestrian-only logistic model achieved:

- F1 = 0.566
- ROC-AUC = 0.840
- PR-AUC = 0.632

whereas the interaction-aware model achieved:

- F1 = 0.596
- ROC-AUC = 0.860
- PR-AUC = 0.668

This is the horizon at which the interaction-aware logistic view showed its clearest overall advantage, improving both thresholded and ranking-based performance relative to the pedestrian-only version.

At 5 seconds, the pedestrian-only logistic model achieved:

- F1 = 0.665
- ROC-AUC = 0.838
- PR-AUC = 0.659

while the interaction-aware model achieved:

- F1 = 0.650,
- ROC-AUC = 0.863
- PR-AUC = 0.697

Here the interaction-aware model again retained an advantage in the ranking metrics, but the pedestrian-only model achieved the slightly higher F1.

Taken together, these results show that the effect of interaction-aware context in logistic regression is not uniform across horizons. The interaction-aware view consistently improves PR-AUC and usually improves ROC-AUC, but its effect on F1 depends on the horizon. The clearest thresholded advantage appears at 3 seconds, whereas the 2-second and 5-second comparisons are more mixed. This supports the broader interpretation that horizon choice changes not only the difficulty of the task, but also the apparent value of interaction-aware information.

5.11.3 Best ablation stage by horizon

The best validation ablation stage also changed across horizons.

At 2 seconds, the best-performing stage was C (+ vehicle motion) with F1 = 0.473 and PR-AUC = 0.363. At 3 seconds, the best stage was D (+ relative geometry) with F1 = 0.628 and PR-AUC = 0.507. At 5 seconds, the best stage was F (+ local density) with F1 = 0.660 and PR-AUC = 0.615.

This pattern is one of the clearest methodological findings of the thesis. It shows that the apparent usefulness of interaction features depends on how far ahead the target event is defined. At the shortest horizon, vehicle motion is most useful, suggesting that immediate dynamic context matters most when the event is defined very tightly. At the main 3-second horizon, relative geometry becomes the strongest interaction addition, indicating that spatial relation between pedestrian and vehicle is especially informative at this intermediate timescale. At 5 seconds, the best stage includes local density, suggesting that a broader interaction summary becomes more valuable when the target is extended further into the future.

The ablation results therefore reinforce the same main conclusion as the broader horizon comparison: the role of interaction-aware information is horizon-dependent, and different temporal formulations of the task highlight different aspects of the local scene.

5.11.4 MLP comparison across horizons

The locked-test MLP comparison showed a different pattern from the logistic models.

At 2 seconds, the pedestrian-only MLP achieved:

- F1 = 0.665
- ROC-AUC = 0.933
- PR-AUC = 0.745

while the interaction-aware MLP achieved:

- F1 = 0.765
- ROC-AUC = 0.959
- PR-AUC = 0.852.

At 3 seconds, the pedestrian-only MLP achieved:

- F1 = 0.734
- ROC-AUC = 0.931
- PR-AUC = 0.806

while the interaction-aware MLP achieved:

- F1 = 0.757
- ROC-AUC = 0.949
- PR-AUC = 0.881.

At 5 seconds, the pedestrian-only MLP achieved:

- F1 = 0.718
- ROC-AUC = 0.876
- PR-AUC = 0.780

whereas the interaction-aware MLP achieved:

- F1 = 0.823
- ROC-AUC = 0.959
- PR-AUC = 0.913

Unlike the logistic branch, the MLP results show a more consistent point-estimate advantage for the interaction-aware view across all three horizons on the matched vehicle-context subsets. The strongest overall MLP performance appears at 5 seconds, where the interaction-aware MLP reaches the highest values for F1, ROC-AUC, and PR-AUC. However, these interaction-aware comparisons apply to observations for which matched vehicle context was available, not to all pedestrian observations. The 3-second horizon remains important because it preserves a shorter-horizon interpretation of the task while still producing strong MLP performance. In this sense, the horizon comparison does not undermine the use of 3 seconds as the main reporting setting, but it does show that a longer horizon yields an easier and more predictable task for the non-linear tabular model.

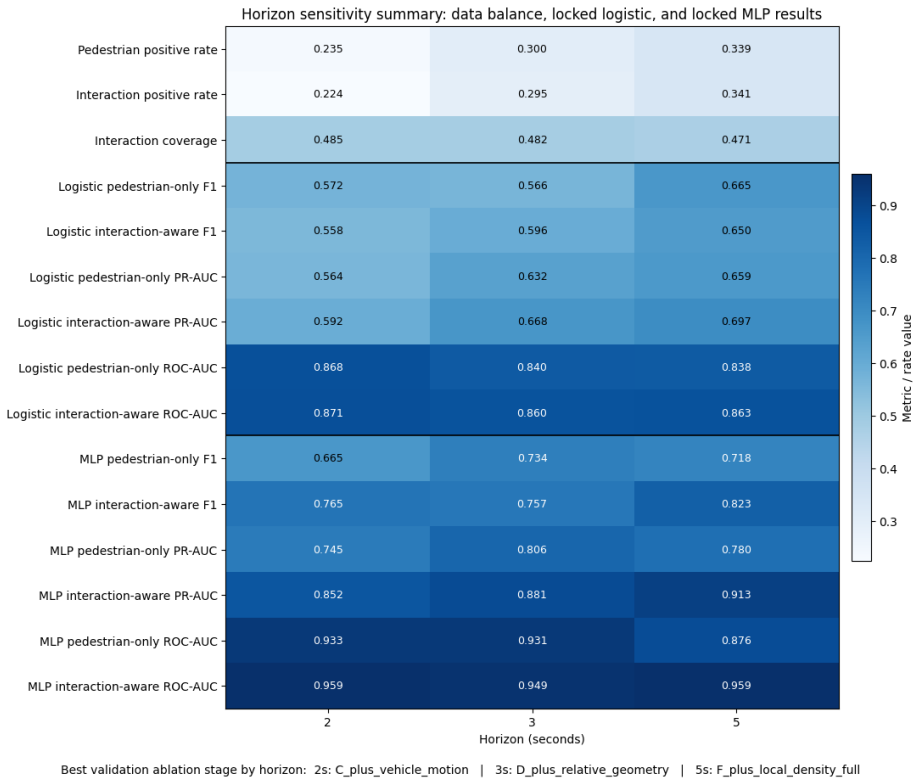

Figure 13. Summary of the full horizon sensitivity section

5.11.5 Summary of the horizon-sensitivity findings

Across the sensitivity analysis, three conclusions stand out. First, horizon choice materially changed class balance and slightly changed interaction coverage. Second, the effect of interaction-aware information was both model-dependent and horizon-dependent: the logistic branch showed a mixed pattern across horizons, whereas the MLP branch benefited from interaction-aware context more consistently. Third, the strongest interaction-feature group shifted with horizon, moving from vehicle motion at 2 seconds to relative geometry at 3 seconds and local density at 5 seconds.

These findings support one of the main methodological claims of the thesis: the future horizon is not just a technical hyperparameter, but part of the substantive problem formulation. Changing the horizon changes what is being predicted, how difficult the task is, and which forms of contextual information appear most useful.

5.12 Uncertainty and paired comparison checks

Track-cluster bootstrap confidence intervals were computed for the main locked-test metrics, F1, ROC-AUC, and PR-AUC, in order to assess how stable the reported model performance was when the grouped trajectory structure was respected. Instead of resampling individual rows, the bootstrap resampled pedestrian track IDs and included all held-out rows belonging to each sampled track. The point-estimate ranking remained unchanged: the interaction-aware MLP achieved the strongest locked-test performance at the main 3-second horizon, with F1 = 0.757, ROC-AUC = 0.949, and PR-AUC = 0.881.

The cluster-bootstrap intervals were substantially wider than row-level bootstrap intervals, especially for F1, PR-AUC, and the GRU branches. This indicates that uncertainty is larger

when within-track dependence is respected. The two MLP branches still maintained the strongest overall profile in point estimates, while the GRU branches were more variable, partly because the GRU test set contained fewer independent trajectory clusters. The interaction-aware MLP's strongest advantage over the pedestrian-only MLP was in PR-AUC, but the interval overlap means that this within-family difference should be interpreted as a point-estimate advantage rather than as a statistically definitive improvement.

To complement the confidence intervals, exact McNemar tests with Holm correction were used as paired row-level diagnostics for thresholded tabular-model predictions on the same held-out interaction-only test rows. These diagnostics showed that every logistic-versus-MLP comparison was significant after correction, whereas the within-family pedestrian-only versus interaction-aware comparisons were not. Because multiple rows can originate from the same pedestrian trajectory, the McNemar p-values are interpreted as diagnostic evidence about paired error patterns rather than as fully cluster-independent significance tests.

Taken together, these uncertainty checks strengthen the main model-family conclusion more than the within-family interaction comparison. The strongest evidence is that the MLP family differs meaningfully from the logistic baseline on the same held-out rows. The smaller pedestrian-only versus interaction-aware differences, especially within the MLP family, should be interpreted more cautiously. Full output tables, including track-cluster paired bootstrap differences for F1, ROC-AUC, and PR-AUC, are provided in Appendix A.

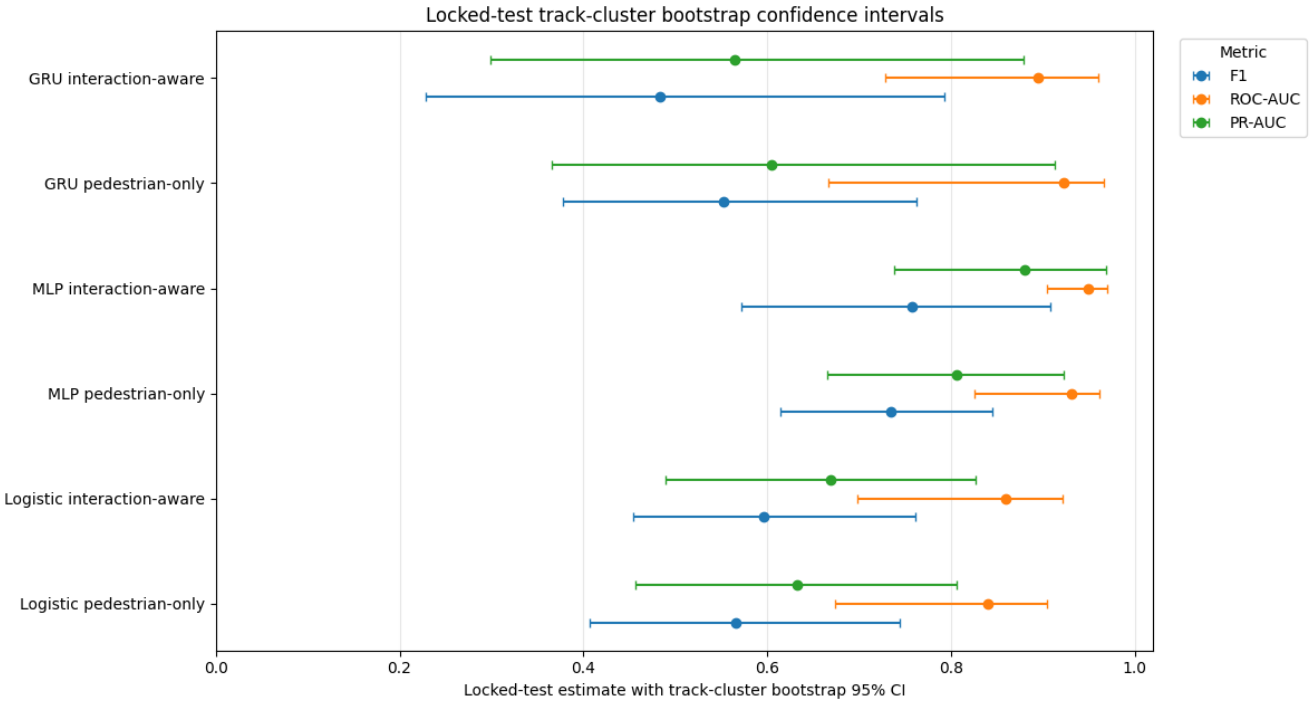

Figure 14. Track-cluster bootstrap confidence intervals for the main locked-test metrics.

	model_a	model_b	a_only_correct	b_only_correct	raw_p_value	holm_corrected_p_value	significant_after_holm
0	logistic_pedestrian_only	mlp_interaction_aware	25	137	6.592736e-20	3.955641e-19	True
1	logistic_pedestrian_only	mlp_pedestrian_only	25	126	2.039779e-17	1.019890e-16	True
2	logistic_interacti	mlp_interaction	32	138	6.898772	2.759509e-16	True

	model_a	model_b	a_only_correct	b_only_correct	raw_p_value	holm_corrected_p_value	significant_after_holm
	on_aware	_aware			e-17		
3	logistic_interaction_aware	mlp_pedestrian_only	32	127	1.350878e-14	4.052633e-14	True
4	mlp_pedestrian_only	mlp_interaction_aware	32	43	2.480457e-01	4.960915e-01	False
5	logistic_pedestrian_only	logistic_interaction_aware	43	49	6.024123e-01	6.024123e-01	False

Table 5. Exact McNemar row-level diagnostics for paired thresholded tabular-model comparisons.

5.13 Summary of the main findings

Five main findings emerged from the results.

First, the 3-second horizon provided a meaningful main analysis setting. It did not dominate every metric at every comparison, but it offered the clearest empirically supported compromise between behavioral immediacy, dataset size, and learnability. The sensitivity analysis showed that 2 seconds produced a stricter and sparser task, while 5 seconds produced a broader and generally easier task. This supports the use of 3 seconds as the main reporting horizon while confirming that horizon choice materially affects both performance and interpretation.

Second, the strongest overall main-horizon point estimate came from the interaction-aware MLP on the matched vehicle-context subset. At the locked 3-second test evaluation, this model achieved the highest reported values among all compared models and also exceeded the pedestrian-only MLP in point estimates. However, the row-level McNemar diagnostic for the pedestrian-only versus interaction-aware MLP comparison did not reach statistical significance after Holm correction ($p = 0.50$), and the track-cluster bootstrap intervals showed substantial uncertainty around the size of this within-family advantage. The MLP interaction effect should therefore be interpreted as a consistent point-estimate pattern rather than as a statistically confirmed improvement. The strongest confirmed result is that both MLP branches differed clearly from both logistic branches, indicating that the clearest performance gain came from non-linear model capacity applied to the tabular feature representation.

Third, pedestrian-only motion and corridor-relative geometry already provided a substantial baseline. Both the logistic and MLP branches showed that the pedestrian-only feature set contained strong predictive signal. This means that a large share of the short-horizon prediction problem can be addressed from the pedestrian trajectory itself, even before local vehicle context is added.

Fourth, the usefulness of interaction-aware context was visible in point estimates, but remained horizon-dependent, model-dependent, subset-dependent, and not statistically confirmed at the decision level. The interaction-aware analyses apply to observations with matched vehicle context, not to all pedestrian rows. In the logistic branch, interaction-aware information consistently increased the ranking metrics (ROC-AUC and PR-AUC) across

horizons, but its effect on thresholded F1 was mixed, and the paired McNemar diagnostic for the pedestrian-only versus interaction-aware logistic comparison was not significant ($p = 0.60$). In the ablation study, relative geometry was more informative than simple proximity at 3 seconds, while the best interaction stage shifted from vehicle motion at 2 seconds to local density at 5 seconds. This shows that different vehicle-context signals become relevant under different temporal formulations of the task.

Fifth, the GRU remained informative as a compact sequence-model comparison, but it did not outperform the strongest engineered tabular MLP branch under the present representation and data conditions. Its ranking metrics were competitive in some settings, but its locked thresholded performance was less stable, especially for the interaction-aware view. Because the GRU used raw fixed-length sequences while the tabular models used hand-engineered temporal summaries, this result should not be read as a general claim that recurrent sequence models are inferior for crossing prediction. It shows only that this compact GRU implementation did not add sufficient value in the present setup to surpass the non-linear tabular approach.

Taken together, these findings show that short-horizon crossing-corridor entry can be predicted from stationary trajectory data, but also that the conclusions depend materially on target design, horizon choice, feature representation, and model family.

6 Discussion

This chapter interprets the results in relation to previous research, evaluates the main methodological choices, reflects on validity and reliability, and discusses broader scientific, practical, ethical, and societal implications. The purpose is not only to summarize the findings, but to clarify what they support, what they do not support, and why the operational design choices matter so much in this thesis.

6.1 Relation to previous research

A broad body of research has shown that short-horizon road-user prediction is relevant for ADAS, proactive safety, and intelligent transport systems. Review-based work in driver intention recognition emphasizes both the promise of anticipatory models and the field's remaining challenges, including contextual heterogeneity, methodological inconsistency, uncertainty representation, and the difficulty of deploying reliable systems in practice (Vellenga et al., 2022). One particularly important point in that review is that the usefulness of a prediction system depends strongly on how early and under what temporal conditions behavior can be recognized, which makes horizon choice a substantive issue rather than a minor technical parameter. This positioning is also consistent with pedestrian-specific prediction research, where crossing-related behavior has often been treated as an operational short-horizon event derived from observable motion and context rather than as direct access to latent intention (Rasouli & Tsotsos, 2020; Schneemann & Heinemann, 2016).

The present thesis is consistent with that literature in one central respect: short-term traffic behavior is meaningfully predictable, but only when the problem is carefully delimited. At the same time, the thesis differs from much previous work in its explicit target formulation. The goal here is not to infer latent psychological intention in a strong cognitive sense, nor to build a full behavioral simulation model. Instead, the study predicts a clearly defined operational

event: entry into a predefined crossing corridor within a fixed future horizon. This narrower formulation reduces psychological ambition but improves interpretive clarity.

The thesis also connects with prior research on vulnerable-road-user interaction, where context has been shown to matter. Mohammadi et al. (2023) found that cyclist–vehicle yielding behavior at unsignalized intersections is better explained when richer behavioral and contextual signals are considered alongside more conventional motion information. That work is important because it shows that interaction outcomes are not purely kinematic. The present thesis aligns with that broader view, but under more constrained data conditions. Since only stationary trajectory data were available here, the study asked a more modest question: how far short-horizon prediction can go using explicit trajectory-derived features and relatively simple interaction context rather than richer non-verbal cues. Unlike studies that address pedestrian crossing prediction with richer sensor context or more direct crossing-intention formulations (Rasouli & Tsotsos, 2020; Schneemann & Heinemann, 2016), the present thesis asks how far stationary trajectory data alone can go when the target is defined operationally and the comparison remains strongly interpretable.

The thesis also relates to recent traffic studies using recurrent neural architectures for maneuver or intention recognition. These studies show that temporal sequence models can perform strongly when the problem is naturally framed as sequence learning and when temporal dependencies are central to the task (Delgado, 2025; Govindhan, 2025; Mirkovic, 2025). However, the present findings suggest that greater temporal complexity is not automatically beneficial in every setting. In the current problem, sequence models were informative but did not outperform the strongest tabular MLP branch. This highlights an important contrast with more architecture-driven studies: added model complexity does not guarantee better performance when the target is highly operational, the data are limited to trajectory signals, and the tabular feature engineering is already strong.

Finally, the thesis differs from broader behavior-modeling approaches such as inverse reinforcement learning, where the goal is to infer reward structures or more general principles underlying traffic behavior rather than classify a narrowly defined future event. Compared with that line of work, the present thesis is narrower and less ambitious in behavioral scope, but stronger in operational transparency. It does not claim to recover motivation. It predicts a clearly specified short-horizon event under explicit spatial and temporal assumptions (Amaya Corredor, 2021).

Taken together, the literature suggests that the present thesis contributes at a useful middle level. It does not introduce a new traffic-prediction architecture, but it does provide a carefully delimited and empirically grounded answer to a practical question: what can be learned about short-horizon pedestrian crossing-related prediction when horizon choice, interaction features, and model family are examined explicitly within the same operational framework?

6.2 Interpretation of the main findings

The first major finding is that pedestrian-only motion and corridor-relative geometry already provide a strong baseline. This is important because it shows that a substantial share of the predictive signal is already present in the pedestrian's own recent movement relative to the crossing corridor. Before vehicle context is considered at all, recent pedestrian motion

already carries considerable information about likely short-horizon crossing-related behavior.

This baseline result has two implications. First, it confirms the relevance of short-history kinematic and geometric features for the present task. Second, it raises the bar for interaction-aware extensions. If the pedestrian-only baseline is already strong, added context must contribute genuinely complementary information rather than merely increasing feature count or apparent realism.

The second major finding is that interaction effects were visible but not uniform across feature representations, horizons, or subsets. The staged ablation showed that nearest-vehicle proximity alone was not the strongest interaction addition at the 3-second horizon. Simple proximity slightly changed some metrics but did not consistently improve F1 or PR-AUC. Stronger gains appeared once richer interaction structure was included: relative geometry was the best stage on PR-AUC, and vehicle type was the best stage on F1. This means the interaction question should not be framed too narrowly as whether distance to the nearest vehicle helps. A better interpretation is that interaction information can matter when matched vehicle context is available, but its value depends strongly on how that information is represented. Simple scalar distance is not the same as spatial relationship.

The third major finding is that the interaction-aware MLP produced the strongest overall point estimate at the main 3-second horizon. At locked test evaluation it achieved F1 = 0.757, ROC-AUC = 0.949, and PR-AUC = 0.881, outperforming both logistic branches and the pedestrian-only MLP (F1 = 0.734, ROC-AUC = 0.931, PR-AUC = 0.806). This result is important because it suggests that two sources of improvement were present simultaneously: a strong gain from richer model capacity, reflected in the MLP-versus-logistic differences, and a higher point-estimate result for the interaction-aware MLP relative to the pedestrian-only MLP. The first of these is strongly supported by the paired row-level McNemar diagnostics; every logistic-versus-MLP comparison was significant after correction. The second is visible in the point estimates, especially PR-AUC, but should be interpreted with caution because the within-family MLP comparison was not significant in the row-level McNemar diagnostic and the track-cluster bootstrap intervals were wide. The permutation-importance results nevertheless provide partial support for the interaction-aware interpretation, since the most influential feature in the interaction-aware MLP was rel_x_nearest_vehicle. This indicates that relative geometry, rather than raw proximity alone, was the main interaction signal the model learned to use.

The fourth major finding is that the compact GRU sequence model did not outperform the strongest engineered tabular MLP under the present setup, even though recurrent architectures are often attractive in traffic-prediction settings. This does not mean that sequence models are unhelpful in general. Rather, it suggests that with short fixed windows of 8 steps, a limited sequence-level training set, raw sequence inputs, and already informative hand-crafted tabular temporal features, recurrent modeling did not add enough value to displace the non-linear tabular baseline. The GRU also showed less stable threshold transfer on the locked test set than the tabular models, which further limits confidence in its operating point. This is a useful empirical finding, but it should be interpreted as implementation-specific rather than as a general model-family verdict.

The fifth major finding is methodological: the conclusions changed across horizons, and the pattern differed between model families. For the logistic branch, interaction-aware

information consistently improved ranking metrics such as PR-AUC and ROC-AUC, but its effect on F1 was mixed and horizon-dependent. The best ablation stage shifted from vehicle motion at 2 seconds, to relative geometry at 3 seconds, to local density at 5 seconds. This shows that the most useful vehicle-context signal depends on how far ahead the target event is defined. For the MLP branch, by contrast, the interaction-aware model showed a consistent point-estimate advantage over the pedestrian-only model across horizons, with larger gaps at 2 and 5 seconds than at 3 seconds: the F1 gap was 0.100 at 2 seconds, 0.023 at 3 seconds, and 0.105 at 5 seconds. Horizon choice therefore does not only change class balance and task difficulty. It also shapes which forms of contextual information the models can exploit, and to what degree.

Taken together, these findings suggest that the most important contribution of the thesis is not simply that one model performed best, but that the empirical answer changed when the task was formulated differently. In that sense, the thesis contributes a methodological result as much as a predictive one: in short-horizon pedestrian prediction, what can be learned depends strongly on how the future event is operationalized, how context is represented, and which model family is used.

6.3 Methodological reflections

6.3.1 Operationalization of the target

The most important methodological choice in the thesis was the decision to operationalize the target as crossing-corridor entry within a fixed future horizon. This was a defensible choice because it produced a measurable and reproducible label. However, it also created the thesis's main interpretive boundary.

The label is not equivalent to psychological intention. A pedestrian may intend to cross and still not enter the corridor within the next few seconds. Conversely, a row may satisfy the operational label without representing a rich internal commitment state. For that reason, the thesis is strongest when read as a study of short-horizon crossing-related behavior prediction, not deep intention recognition.

This limitation does not invalidate the work. On the contrary, it is one of the strengths of the thesis that this distinction is kept explicit. The problem is defined in terms of what the data can actually support.

6.3.2 Geometric definitions

The broad interaction region, crossing corridor, and approach band were central to the final analysis. These definitions were empirically guided by trajectory density and overlap patterns, but they also involved analytic judgment. Different but still reasonable boundaries could have produced somewhat different row inclusion, class balance, and feature behavior.

This affects construct validity. The target was only as good as the geometry used to define it. At the same time, the explicitness of the geometry is an advantage. The spatial setup is visible, criticizable, and reproducible, which is preferable to hidden or implicit labeling choices.

6.3.3 Horizon choice as a scientific variable

One of the strongest methodological points in the thesis is that horizon choice should not be treated as a minor hyperparameter. Vellenga et al. (2022) emphasize that traffic-prediction performance needs to be understood in relation to the time horizon at which behavior is to be recognized, because the required anticipation time depends on the complexity of the scenario and the practical action window available to a safety system. The current results are fully consistent with that view: changing the horizon changed the class balance, the learnability of the task, and the apparent usefulness of interaction features.

The choice of 3 seconds as the main reporting horizon appears methodologically appropriate. A 2-second horizon preserved a stricter immediate interpretation but created a harder and sparser target. A 5-second horizon produced stronger predictive performance but broadened the semantics of the event. The 3-second horizon therefore provided the best compromise between immediacy and learnability. Importantly, however, the sensitivity analyses show that no single horizon is universally correct; rather, each horizon answers a slightly different version of the prediction question.

The same reasoning that motivated treating the horizon as a scientific variable rather than a fixed parameter could, in principle, be applied to other design choices, most notably the scene-time bin size used for interaction feature construction. Just as different horizons produce different class balances and highlight different interaction signals, different bin sizes would produce different interaction coverage rates and potentially different feature quality. The choice of 500 ms was operationally motivated in Section 4.6.1, but a systematic comparison across bin sizes, for example 250 ms, 500 ms, and 1000 ms, would have provided a more rigorous sensitivity check on this assumption. This is acknowledged as a limitation of the current implementation and is identified as a direction for future work.

6.3.4 Scene-time matching and local vehicle context

The 500 ms scene-time bin was a practical choice for temporal synchronization between pedestrian and vehicle observations. It was not selected because 500 ms represents a full interaction episode or a behavioral time constant. It was selected as an engineering compromise: narrow enough to preserve local temporal relevance, but broad enough to tolerate asynchronous sensing.
This choice made interaction-feature construction feasible, but it also limited it. Not all pedestrian rows had matched vehicle context, and the available interaction representation remained local and simplified. In particular, nearest-vehicle features collapse multi-agent scenes into a single closest-vehicle summary, while density features only partially compensate for this simplification. Therefore, the interaction-aware results should be interpreted as applying to vehicle-context-available observations and modest local interaction context, not to all pedestrian observations or to full multi-agent interaction structure.

6.3.5 Model-family choice

The choice to begin with logistic regression was methodologically appropriate. Logistic regression provided a transparent baseline, direct coefficient interpretation, structured ablation, and a controllable threshold-tuning workflow. This made it especially valuable for

understanding feature effects.

The MLP comparison was also methodologically justified. Because it used the same tabular inputs, it tested whether non-linearity itself improved predictive performance without changing the operational target or the feature representation. Extending the MLP to both pedestrian-only and interaction-aware views then made it possible to separate model-capacity gains from feature-richness gains within the same architecture.

The sequence-model branch was useful as an exploratory comparison. It demonstrated that sequence modeling was relevant enough to test, but not strong enough under the present raw-sequence setup to displace the engineered tabular MLP as the best-performing model family. Because the tabular models received hand-crafted temporal summaries while the GRU received raw fixed-length sequences, the comparison should be read as representation-specific rather than as a general conclusion about recurrent architectures.

6.3.6 Explainability choices

The transparency design was appropriate given the scope of the thesis. It encompassed both interpretability, pursued through logistic regression coefficients and staged ablation, where the model structure itself is directly readable, and post-hoc explainability, pursued through permutation importance applied to the MLP, where the model's behavior requires an additional analytical step to summarize. Both dimensions, as defined in Sections 1 and 2.7, contributed to the thesis's overall transparency goal.

For logistic regression, interpretability was implemented through standardized coefficient inspection and staged cumulative ablation. These tools made it possible to examine not only whether the model performed well, but also which feature groups contributed to that performance and in what direction. This was especially valuable for the interaction analysis, because it allowed the gradual effect of proximity, vehicle motion, relative geometry, vehicle type, and local density to be inspected explicitly rather than inferred indirectly from a single opaque model.

For the MLP, explainability was addressed through permutation importance. This choice was methodologically appropriate because the goal was not to explain isolated individual predictions, but to obtain a global summary of which features the model relied on most across the held-out evaluation set. Both MLP branches were examined in this way, with the interaction-aware MLP selected for primary interpretation because its point-estimate performance was strongest at the main horizon, even though the difference from the pedestrian-only MLP did not reach statistical significance after Holm correction. The resulting importance ranking showed that the model relied most strongly on rel_x_nearest_vehicle, recent pedestrian displacement, current speed, corridor distance, and heading. This pattern is consistent with interpretable traffic-relevant structure rather than arbitrary complexity.

No separate post-hoc explainability block was included for the GRU. This was not simply because the GRU performed worse than the MLP. Rather, it reflected a principled scope decision: the GRU input representation consisted of raw temporal sequences, and the relationship between individual timestep features and the final prediction was substantially harder to attribute using standard post-hoc tools. Since this compact GRU did not

outperform the engineered tabular MLP and showed less stable threshold behavior, omitting a separate sequence-model explainability analysis was a reasonable scope choice, although GRU-specific explanation methods remain a relevant direction for future work.

6.4 Validity and reliability

Construct validity
Construct validity is limited by the fact that the target is an operational proxy rather than a ground-truth measure of latent intention. This is addressed explicitly throughout the thesis, which strengthens rather than weakens the work. Still, any statement about "intention" must be understood within this operational boundary.

Internal validity
Internal validity is strengthened by the leakage-safe track-ID split, the locked evaluation protocol, and the consistent threshold-selection procedure across all model families. However, some geometric choices and preprocessing decisions were partly informed by exploratory data inspection, which creates a mild risk that certain design decisions are implicitly optimized for this particular dataset. This risk is mitigated by the explicit documentation of all choices and the transparency of the spatial geometry.

External validity
External validity is the weakest dimension of the thesis. The study is based on a single site, namely one crossing on Slottskogsgatan using one sensor system under the specific spatial, traffic, and pedestrian-flow conditions of that location. The conclusions hold for this setting. How well they generalize to different crossing types, sensor positions, or pedestrian populations is unknown. This is not a flaw but a boundary that must be stated clearly.

Reliability
Reproducibility is supported by fixed random seeds, explicit feature definitions, leakage-controlled splits, and a notebook that documents the full pipeline. The track-cluster bootstrap confidence intervals provide a more conservative uncertainty estimate than row-level resampling because they resample pedestrian trajectories rather than individual observations. These intervals preserve the main point-estimate ranking, but they are wide enough to caution against overinterpreting smaller within-family differences, especially the pedestrian-only versus interaction-aware MLP gap. The McNemar diagnostics are more decisive for logistic-versus-MLP comparisons than for within-family comparisons, but because they operate at row level they should be interpreted as paired diagnostic evidence rather than fully cluster-independent tests.

6.5 Implications for theory and practice

Scientific implications
The thesis demonstrates that short-horizon crossing-corridor entry prediction is feasible from trajectory data using interpretable features and moderate model complexity. The most important theoretical contribution is not the performance levels themselves, which reflect a single site and a specific operational setup, but the framework for examining how horizon choice, feature representation, and model family interact. Treating the horizon as a substantive design variable rather than a fixed parameter is a transferable methodological lesson. So is the demonstration that interaction effects are representation-dependent:

proximity and relative geometry are not interchangeable operationalizations of “vehicle context.”

Practical implications
The pedestrian-only MLP and the interaction-aware MLP represent two practically relevant operating points. A system that has access only to pedestrian trajectories can already achieve strong ranking performance and stable threshold behavior. A system with matched vehicle context may improve on this, particularly on PR-AUC and F1, but that conclusion applies to situations where such vehicle context is available and reliably synchronized. In safety-critical deployment, the difference between these two operating points may matter less than the stability of the threshold behavior, which favors the tabular MLP family over the compact GRU branch in the present setup.

The choice of recall as the primary threshold-tuning objective is appropriate for safety applications where missed crossings carry higher cost than false alarms. However, the McNemar results also show that no within-family interaction comparison reached significance, which means the practical case for the added complexity of interaction-aware features over pedestrian-only features rests on point estimates rather than statistically confirmed decision-level differences.

6.6 Ethical and societal aspects

The data used in this thesis consist of anonymized trajectory coordinates from a public crossing site, collected by a roadside sensor system. No faces, identifiers, or personally distinguishable features were recorded or used. The data were provided for academic research purposes under terms that preclude re-identification.

From a societal perspective, systems that can anticipate pedestrian crossing behavior have genuine potential to improve road safety, particularly for vulnerable road users. However, deployment of such systems requires careful attention to several risks. False negatives i.e. missed crossings may create unwarranted confidence and delay protective action. False positives i.e. flagged crossings that do not occur may cause alarm fatigue or disruptive braking responses. Both error types have safety consequences, and their relative costs are scenario-dependent.

There is also a surveillance dimension. Persistent trajectory monitoring of pedestrians in public space raises questions about proportionality and consent, even when the data are anonymized. As systems of this kind become more capable, the governance frameworks around their deployment will matter as much as their predictive accuracy.

Finally, the interpretability of the model is ethically relevant. A logistic model whose threshold behavior can be audited and explained is preferable to an opaque system in accountability contexts. The thesis's emphasis on interpretable features and explainability methods directly addresses this concern.

6.7 Additional knowledge needed

Several directions would meaningfully extend the findings of this thesis.

First, multi-site validation is the most pressing need. Generalizing beyond Slottskogsgatan requires data from crossings that differ in geometry, traffic volume, pedestrian density, and urban context.

Second, richer interaction representations such as multi-agent scene encoding, vehicle trajectory prediction, or social force terms would allow a stronger test of whether interaction context can improve beyond what relative geometry provides in the current single-nearest-vehicle setup.

Third, calibration-aware evaluation would be valuable for safety applications. Point estimates and McNemar tests address rank and decision accuracy, but the practical deployment of a warning system also depends on whether predicted probabilities are well-calibrated under distribution shift.

Fourth, a longer observation window for the sequence models would address one of the current GRU's main constraints. With only 8 timesteps, the model has limited opportunity to learn approach dynamics that unfold over several seconds.

Fifth, testing whether the operational label can be improved, for example by combining corridor entry with deceleration patterns or gaze proxies, would sharpen the connection between the proxy target and the underlying behavioral construct.

7 Conclusion

This thesis set out to examine whether short-horizon pedestrian crossing-corridor entry could be predicted from stationary trajectory data using interpretable features and a structured model comparison. The aim was not to recover latent psychological intention, but to predict a clearly defined operational event, namely the entry into a predefined spatial corridor within a fixed future horizon in a way that made the modelling decisions transparent and the conclusions methodologically grounded.

7.1 Answers to the research sub-questions

- **Sub-question 1: Can a reliable and interpretable binary label for short-horizon crossing-corridor entry be constructed from trajectory data?**

A reproducible and geometrically explicit binary label was successfully constructed. For each eligible pedestrian observation, the label captured whether the pedestrian's track entered the operational crossing corridor within the defined future horizon. The three-second horizon provided a class balance and dataset size that supported stable modelling, while the 2-second and 5-second settings confirmed that the label's properties vary predictably with horizon choice. The key interpretive boundary, that this is a proxy for observable short-horizon behavior rather than a direct measure of intention, was maintained throughout and is reflected in all subsequent analysis.

- **Sub-question 2: Do pedestrian trajectory features provide a meaningful baseline for prediction, and which features are most informative?**

Yes. Pedestrian trajectory features did provide a strong and consistent baseline. At the main 3-second horizon, the pedestrian-only logistic model achieved PR-AUC = 0.632 and ROC-AUC = 0.840 on the locked interaction-only test set, establishing a feature-driven baseline before any vehicle context was added. Coefficient inspection and permutation importance converged on a consistent set of informative features: distance to the corridor, recent lateral and longitudinal displacement, current speed, heading over the past second, and the fraction of recent observations that were estimated. These features reflect the pedestrian's spatial commitment to crossing and the dynamics of their approach, which is consistent with what prior work suggests should matter for short-horizon prediction (Schneemann & Heinemann, 2016; Rasouli & Tsotsos, 2020).

Sub-question 3: Does vehicle-interaction context improve prediction beyond the pedestrian-only baseline, and if so, under what conditions?

The results show evidence that some forms of vehicle-interaction context are informative when matched vehicle context is available, but the answer depends on the type of interaction feature, the model family, and the horizon. At 3 seconds, the structured ablation showed that the effect of interaction features was not uniform across feature types. Simple nearest-vehicle proximity added little. The stronger contributions came from relative geometry (relative x-position of the nearest vehicle) and vehicle type, which produced the strongest validation PR-AUC and F1 respectively. On the locked 3-second interaction-only test set, the logistic interaction-aware model exceeded the pedestrian-only logistic model on all reported metrics, with the largest differences in recall and PR-AUC. However, this advantage was not confirmed by the paired row-level McNemar diagnostic after Holm correction, so the decision-level benefit of added interaction features should be interpreted cautiously. In the MLP branch, the interaction-aware model produced stronger point estimates across all three tested horizons, and the interaction-aware MLP was the strongest locked-test result in the entire comparison. Taken together, the evidence supports a cautious conclusion: interaction context can be useful on vehicle-context-available observations, but its value is selective rather than uniform and is stronger in point estimates than in statistically confirmed paired decision-level differences.

Sub-question 4: Does the choice of future horizon materially affect the conclusions?

Yes. Horizon choice materially affected the conclusions across every dimension examined. Changing the horizon from 2 to 3 to 5 seconds changed the class balance, interaction coverage, the best-performing ablation stage, and the relative advantage of interaction-aware features. The best interaction group shifted from vehicle motion at 2 seconds, to relative geometry at 3 seconds, to local density at 5 seconds. For the logistic branch, the effect of interaction-aware context on F1 was mixed across horizons, while PR-AUC and ROC-AUC showed a more consistent advantage. For the MLP, the interaction-aware advantage was consistent and grew with the horizon. The horizon is therefore a substantive part of the problem formulation, not a technical hyperparameter, and this finding is one of the thesis's core methodological contributions.

7.2 Main contributions

The thesis makes four contributions.

First, it provides a reproducible operational framework for short-horizon pedestrian crossing-corridor entry prediction from trajectory data. The spatial geometry, label-construction procedure, feature-engineering pipeline, and evaluation protocol are fully documented and could be applied to different sites.

Second, it demonstrates that pedestrian-only trajectory features constitute a strong and interpretable baseline. This is practically relevant because it shows how much of the short-horizon prediction task can be addressed from pedestrian motion and corridor-relative geometry alone, before any vehicle-side context is introduced.

Third, it shows that interaction effects in trajectory-based crossing prediction are representation-dependent, horizon-dependent, and conditional on vehicle-context availability. Simple proximity is not equivalent to relative geometry, and the strongest interaction signal changes across horizons. This has implications for how future studies should operationalize vehicle context rather than treating all interaction information as interchangeable.

Fourth, it provides a structured model-family comparison across logistic regression, MLP, and GRU under consistent evaluation conditions, supplemented by track-cluster bootstrap confidence intervals and paired McNemar row-level diagnostics. The strongest comparative finding is that the MLP family differs meaningfully from logistic regression, while within-family pedestrian-only versus interaction-aware differences do not reach statistical significance at the decision level. This provides a more careful calibration of which claims can and cannot be made confidently from this type of data.

7.3 Limitations

The main limitations are single-site generalizability, the operational nature of the target label, the dependence of the label on chosen spatial boundaries, the simplified nearest-vehicle interaction representation, the fact that interaction-aware conclusions apply only to matched vehicle-context observations, and the small sequence-model training and test sets that constrained the GRU comparison. No systematic sensitivity analysis was performed over alternative corridor widths, corridor placements, approach-band definitions, or scene-time bin sizes. Because the target label is geometry-derived, small changes in the corridor definition could change the positive class near the boundary. The uncertainty analysis partially addresses within-track dependence through a track-cluster bootstrap, but the final results remain conditional on one selected balanced group split. These limitations are stated explicitly because they define the scope within which the thesis findings should be interpreted.

7.4 Future work

Several directions follow naturally from this work.

First, future studies should test the same operationalization across multiple sites and infrastructures. Since the present study uses one recorded urban traffic setting, broader validation is needed before stronger claims can be made about transferability. This is especially important in light of wider traffic-prediction research, where context and site differences often limit direct comparability across studies (Vellenga et al., 2022).

Second, future work should repeat the full modelling pipeline across several group-level train/validation/test splits, or use group cross-validation, to evaluate whether the observed model ranking and interaction-feature conclusions remain stable across alternative partitions.

Third, future work should perform spatial sensitivity analyses. The corridor, approach band, and interaction region should be varied across several plausible boundary definitions to test whether the label balance, model ranking, and feature-importance patterns remain stable when the geometry-derived target is changed slightly.

Fourth, future work should continue to explore horizon sensitivity more explicitly. One of the clearest findings of this thesis is that horizon choice changes the class balance, model difficulty, and apparent usefulness of interaction features. Related traffic work has also shown that predictive performance can vary substantially with temporal distance to the target event, which suggests that horizon should remain a primary design variable rather than a fixed convention (Mirkovic, 2025).

Fifth, richer forms of interaction representation should be investigated. The current interaction features were constructed from local scene matching, nearest-vehicle summaries, relative geometry, vehicle type, and density counts. These provide a useful starting point, but they do not fully represent multi-agent interaction structure. Future work could investigate improved temporal synchronization, more robust local-scene construction, richer multi-vehicle representations, and systematic comparisons of alternative scene-time bin sizes. It would also be valuable to test whether interaction features become more useful when combined with more expressive but still interpretable scene encodings.

Sixth, future work should examine richer behavioural cues where available. Prior research on cyclist-vehicle interaction suggests that contextual and behavioural signals beyond simple trajectories can improve prediction in interaction-heavy settings (Mohammadi et al., 2023). If analogous cues become available for pedestrians, such as orientation, hesitation, or other embodied movement indicators, they may help distinguish between corridor-adjacent movement and truly imminent crossing commitment.

Seventh, future work should develop a stronger and more deliberate sequence-modelling branch if recurrent architectures are to be evaluated more fully. In the present thesis, the sequence models were informative but secondary. A more extensive comparison would require further work on sequence representation, window construction, feature channels, hyperparameter search, larger effective training sets, and fairer comparisons between raw sequence inputs and engineered temporal tabular features. Studies explicitly designed around RNN, GRU, and LSTM comparison show that such models can be relevant when temporal dependencies are central and the sequence setup is optimized for the task (Delgado, 2025; Mirkovic, 2025).

Finally, future work should examine cost-sensitive and calibration-aware evaluation more directly while continuing to maintain a clear distinction between operational behavior proxies and latent psychological intention. In practical traffic-assistance settings, false positives and false negatives do not have identical consequences, and the value of future predictive systems will depend not only on stronger metrics, but also on whether their targets, assumptions, uncertainties, and limitations remain transparent.

References

Amaya Corredor, S. E. (2021). *Modelling of human behaviour in traffic interactions using inverse reinforcement learning* [Master's thesis, Chalmers University of Technology and University of Gothenburg]. Chalmers Open Digital Repository. https://hdl.handle.net/20.500.12380/304092

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Castillo, A. N. M., Salinas Maldonado, C., & Sotelo, M. Á. (2025). Towards explainable pedestrian behavior prediction: A neuro-symbolic framework for autonomous driving. *Applied Sciences*, *15*(11), 6283. https://doi.org/10.3390/app15116283

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. https://doi.org/10.1145/1143844.1143874

Delgado, A. (2025). *Comparative analysis of recurrent neural networks for vehicle turning intention recognition at roundabouts* [Master's degree project, University of Skövde]. https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-25587

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81. https://doi.org/10.48550/arXiv.1801.01489

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org

Govindhan, A. (2025). *Context-aware trajectory prediction and collision detection: A case study using NAS-optimized context-aware LSTM for cyclist-vehicle collision detection at roundabouts* [Master's degree project, University of Skövde]. https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-25597

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Landry, F.-G., & Akhloufi, M. A. (2025). Predicting pedestrian crossing intention in autonomous vehicles: A review. *Neurocomputing*, *618*, 129105. https://doi.org/10.1016/j.neucom.2024.129105

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1705.07874

Mirzabagheri, A., Ahmadi, M., Zhang, N., Alirezaee, R., Mozaffari, S., & Alirezaee, S. (2025). Navigating uncertainty: Advanced techniques in pedestrian intention prediction for autonomous vehicles — A comprehensive review. *Vehicles*, *7*(2), 57. https://doi.org/10.3390/vehicles7020057

Mirkovic, D. (2025). *Intention recognition of light vehicles in traffic by using LSTM deep learning model* [Master's degree project, University of Skövde]. https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-25897

Mohammadi, A., Bianchi Piccinini, G., & Dozza, M. (2023). How do cyclists interact with motorized vehicles at unsignalized intersections? Modeling cyclists' yielding behavior using naturalistic data. *Accident Analysis & Prevention*, *190*, 107156. https://doi.org/10.1016/j.aap.2023.107156

Rasouli, A., & Tsotsos, J. K. (2020). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems*, *21*(3), 900–918. https://doi.org/10.1109/TITS.2020.3006767

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Schneemann, F., & Heinemann, P. (2016). Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2243–2248. https://doi.org/10.1109/IROS.2016.7759351

Vellenga, K., Steinhauer, H. J., Karlsson, A., Falkman, G., Rhodin, A., & Koppisetty, A. C. (2022). Driver intention recognition: State-of-the-art review. *IEEE Open Journal of Intelligent Transportation Systems*, *3*, 602–629. https://doi.org/10.1109/OJITS.2022.3197296

Zou, H., Guo, Y., Wei, F., Guo, D., Li, Q., & Pirov, J. (2025). A pedestrian group crossing intention prediction model integrating spatiotemporal features. *Scientific Reports*, *15*, 20675. https://doi.org/10.1038/s41598-025-05128-4

Appendix A: Executable notebook and supplementary outputs

The complete Jupyter notebook submitted alongside this report contains the full implementation, including data preprocessing, spatial operationalization, feature engineering, model training, threshold selection, sensitivity analyses, explainability analyses, and supplementary outputs not reproduced in the main text. The notebook and the accompanying data are publicly available in the following GitHub repository:
https://github.com/valnke03/Master-Thesis