Leveraging Deep Learning for Pedestrian and Cycle Flow Prediction in Urban Environments

MSc Research Project
Data Analytics

Carlos Manuel Benitez Martinez
Student ID: 24104515

School of Computing
National College of Ireland

Supervisor: Mr. John Kelly

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Carlos Manuel Benitez Martinez
Student ID:	24104515
Programme:	Master in Data Analytics **Year:** 2025
Module:	MSc Research Project
Supervisor:	Mr. John Kelly
Submission Due Date:	11/08/2025
Project Title:	Leveraging Deep Learning for Pedestrian and Cycle Flow Prediction in Urban Environments
Word Count:	6961 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Manuel Benitez

Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Leveraging Deep Learning for Pedestrian and Cycle Flow Prediction in Urban Environments

Carlos Manuel Benitez Martinez
Student ID: x24104515

Abstract

The accurate prediction of urban mobility is essential for developing sustainable and responsive smart cities. Monitoring and predicting urban mobility remain a challenge. However, previous work integrates spatiotemporal data as static variables, and spatial data available is limited. This study focuses on predicting pedestrian and cyclist flows in Dublin city on an hourly basis with the integration of data from multiple sources, including data from footfall (13 locations) and cyclists (6 locations), with three different features as input, dynamically varying time (12 features), weather (12 conditions) and spatial (2 features). This study proposes the usage of advanced deep learning architectures, Long Short-Term Memory (LSTM) and hybrid Convolutional Neural Network combined with Long Short-Term Memory (CNN-LSTM) models, with capabilities to capture complex spatiotemporal information and understand temporal context in urban mobility patterns. By comparing these models against traditional machine learning baselines such as Random Forest and XGBoost, the study demonstrates superior performance of deep learning in handling volatile pedestrian and cyclist flows. The work indicates that a scalable pipeline can improve open-source cloud resources to achieve reproducibility in data-driven urban planning. The findings noted how integrating dynamic weather data and temporal features, excluding spatial features shows significant improvements in error reduction of more than 50% for pedestrians and 30% for cyclist flows. The results show that LSTM and hybrid CNN-LSTM model without spatial features outperform Random Forest and XGBoost models according to three metrics commonly used in regression: mean absolute error, root mean squared error and coefficient of determination.

Keywords: Heterogeneous data integration, pedestrian prediction, cycle flow prediction, deep learning, CNN, LSTM.

1 Introduction

Modern cities face different challenges when population grows and demands infrastructure and sustainability for mobility systems, which leads to understanding and managing the urban mobility. Two indicators help understand urban mobility patterns: pedestrian footfall, and cycle flow. While pedestrian footfall refers to the number of pedestrians in a specific location during a specific period of time, cycle flow refers to the count of cyclists. Predicting both relies on identifying temporal trends on an hourly, daily, and monthly basis, but these approaches remain insufficient to accurately predict traffic flow, and other factors become relevant in studying their patterns.

After reviewing prior studies on pedestrian and traffic predictions, it is noticeable that many approaches have been developed, using different models that can be categorised as baselines, such as ARIMA, Random Forest, eXtreme Gradient Boosting (Komar & James, 2024), and deep learning, including a variety of neural networks such as GAN, LSTM, CNN-

LSTM, GCN-LSTM, and CNN-BiLSTM. By focusing on traffic speed or traffic flow predictions, machine learning or deep learning models have been employed including temporal, spatial, weather and event-related features, where XGBoost has proven to accurately predict speed, hybrids CNN-LSTM (Wang & Susanto, 2023) and CNN-BiLSTM (Méndez, G. Merayo, & Núñez, 2023), or graph neural network-based models (Fang, Pan, Chen, Du, & Gao, 2021) particularly have proven to be effectively predict traffic. Although spatiotemporal data has not been involved much, and its addition is believed to improve the performance of predictions results.

1.1 Motivation

The motivation of this study lies in the difficulties of existing models to balance predictive accuracy with computational efficiency and the need for models to predict accurately including not only temporal features but weather conditions to understand hidden patterns, as Dublin City experiences a significant growth in both population and bicycle usage (Sustrans, 2024).

1.2 Research question

The value of this research is to identify: **How deep learning techniques can be effectively leveraged to predict pedestrian and cycle flows on an hourly basis in urban environments for Dublin City by integrating data from multiple sources such as footfall counts, cycle counts, weather conditions, temporal and spatial features?**

1.3 Objectives

Considering the integration of external factors can help to discover patterns that can effectively improve predictive capacity, the specific research objectives are:

1. Evaluate how deep learning models can integrate a framework that leverages the pedestrian and cycle predictions in different Dublin city locations.
2. Integrate multi-source data, especially weather as a dynamic influencing factor and temporal encodings to help capture complex spatiotemporal patterns.
3. Assess each model predictive performance through different evaluation metrics.
4. Compare the predictive performance in urban environments of baseline machine learning models such as Random Forest and eXtreme Gradient Boost (XGBoost), with deep learning architectures, such as LSTM and CNN-LSTM.

1.4 Contribution

The main contributions of this research are summarised as follows:

- A novel architecture for time series forecasting employing LSTM and a hybrid of a CNN and LSTM neural networks.
- A data preparation pipeline that provides a clear data preparation framework for incomplete and inconsistent data.
- Implementation of four different models to thirteen footfall locations and six cyclists; deep learning architectures include a variant exclusively including spatial features, with careful fine-tuning of considered hyperparameters.

- To demonstrate the usefulness of models through the comparison of deep learning architectures over powerful baseline models used for time series forecasting.

The rest of the paper is organised as follows. Section 2, a brief review of related work. Section 3 explains research methodology followed by data collection, preprocessing, integration, and transformation process. Section 4 justifies model selection. Section 5 describes models implementation. Section 6 explains and discuss the results of implementation. Finally, the conclusion and future work of this paper are located in Section 7.

2 Related Work

2.1 MDTP: A Multi-source Deep Traffic Prediction Framework over Spatio-Temporal Trajectory Data.

The MDTP framework was introduced by Fang et al. (2021) to address the limitations of single-source traffic models like Autoregressive Integrated Moving Average (ARIMA) and XGBoost. Traditional models excel in small datasets but seem limited with nonlinear dynamics and spatiotemporal dependencies. MDTP integrates multiple traffic data streams (i.e., bikes and taxis), using improved Graph Convolutional Networks (GCNs) and Long Short-Term Memory (LSTM) networks. These components allow the model to dynamically adjust to temporal shifts and spatial relationships, such as peak vs off-peak travel patterns. By incorporating Sum Connect and Concat Connect mechanisms, MDTP effectively bridges latent features across sources, significantly improving the accuracy in predictions. Furthermore, the system designed by Fang et al. (2021) achieved scalability, processing five months of NYC taxi data in under ten seconds far outperforming benchmarks like Spatial-Temporal Dynamic Network (STDN) that took beyond two minutes and provided unrealistic predictions for real-life scenarios. Its MDTP+ visualization tool provides planners with heatmaps and congestion trends, bridging predictive analytics with real-time urban mobility decisions. This advancement strengthens intelligent transportation systems, particularly for large-scale, dynamic cities.

2.2 Forecasting of Bicycle and Pedestrian Traffic Using Flexible and Efficient Hybrid Deep Learning Approach.

Harrou et al. (2022) pioneered a hybrid deep learning framework, such as a Guided-Attention Hybrid Deep Learning architecture for Variational Autoencoders (GAHD-VAE), which combines variational autoencoders (VAE), long short-term memory networks (LSTM), and multi-stage attention mechanisms to forecast bicycle and pedestrian traffic flows. Their primary objectives were to capture the temporal dependencies in traffic data through LSTM-combined with VAEs, improve accuracy in predictions through the selection of attention-driven features, and validate the model across six urban datasets with diverse traffic patterns. The contribution of the study lies in its improvement in RMSE over baseline models like standalone LSTM and Gated Recurrent Unit (GRU). However, it fell into the limitations in the field, contextualizing environmental factors, and weather variables were treated as static inputs (i.e., binary rain indicators), ignoring real-time fluctuations. This gap highlights the importance of developing models that consider weather as a dynamic, time-varying input rather than a static covariate.

2.3 Review of Pedestrian Trajectory Prediction Methods: Comparing Deep Learning and Knowledge-based Approaches.

In this study, the contrast between knowledge-based (KB) and deep learning (DL) approaches for the prediction of pedestrian trajectory, Korbmacher and Tordeux (2022) highlight the complementary strengths and limitations. KB models, such as the Social Force Model (SFM), excel in simulating large-crowd dynamics through predefined rules, offering scalability and interpretability for general behaviours but it struggles with complex localized interactions like bottleneck navigation. In contrast, DL methods, such as LSTMs, CNNs, and GANs, achieve high accuracy in low-density environments and enable multimodal path prediction (i.e., GANs generating diverse trajectories), however, they still suffer from "black box" limitations and reliance on extensive training data. To address these gaps, hybrid models integrate the interpretability of KB with the adaptability of DL: neural networks refine KB parameters (i.e., calibrating SFM forces), while KB-generated synthetic data trains DL models for rare scenarios. The authors systematically evaluate input structures (trajectory vs. grid data), use cases (evacuation planning vs. autonomous vehicles), and trade-offs (computational cost vs. accuracy), proposing a framework that balances real-time decision-making with strategic urban mobility forecasting. This synthesis addresses critical gaps in dynamic environmental modelling, directly informing Dublin’s need for hourly pedestrian/cycle flow predictions using multi-source data integration.

2.4 Regression Analysis for Prediction of Road Travel Speed.

As urban mobility research leverages in machine learning and deep learning methods, different methods to forecast traffic speed, flow, and related variables in a complex city environment. Rubasinghe and Hettiarachchi (2023) conducted a comprehensive study of road travel speed prediction in Gothenburg, Sweden, using regression analysis with advanced models such as XGBoost and LightGBM. The model was built using temporal variables, route information, and weather conditions to develop robust predictive models, handling complex and multi-source datasets, considering the advantages of ensemble-based approaches. The evaluation metrics included RMSE, MAE, MSE, and R^2, emphasising the importance of data cleaning, alignment of temporal features, and integration of external data sources like weather and holiday information to improve accuracy. This methodology aligns with the approach taken in this study, by focusing on integrating multi-source data (temporal, spatial and weather features), and emphasises the importance of the preprocessing and feature engineering phases, and the application of advanced regression models for urban mobility prediction. They highlight, the use of tree-based ensemble models (Random Forest, XGBoost) and the model performance evaluation through cross-validation and residual analysis. Additionally, it recognises the value of analysing feature importance and the impact of external features (such as weather and holidays) on predictive outcomes.

2.5 Traffic Flow Prediction with Heterogenous Data Using a Hybrid CNN-LSTM Model

Using traffic flow data from 2016 to 2019, this study by Wang et al. (2023) highlights the frequent use and success of Neural Network methods for time series prediction. Moreover, it describes its limitations, i.e. LSTM using a simple neural network unable to capture complex characteristic of traffic flow. The proposed method is a combination CNN-LSTM compared to LSTM, BILSTM, and LSTM-BILSTM (Wang & Susanto, 2023, p. 3104). It employs a robust hyperparameter tuning with scalable different values to identify the combination of values for the best model in terms of accuracy and performance. It also, identify step-by-step the overall improvement of the model through the feature addition to evaluate how each set of features contributes to the accuracy of the model. However, Wang et al. (2023) still considers weather as a static feature based on historical data and descriptive temporal data only. The current study considers the strength of addition of different features type for the improvement seen by CNN-LSTM compared to other deep learning techniques.

2.6 Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model.

The hybrid model by Méndez et al. (2023), implemented using CNN and BiLSTM designed for long-term traffic flow prediction in urban environment in Madrid city. The model demonstrated that the integration of convolutional layers (for feature extraction and noise filtering) with bidirectional LSTM layers (for capturing temporal dependencies) outperform traditional machine learning approaches and single-architecture deep learning techniques. Their approach leverages multi-source data, including traffic count, meteorological variables, and temporal context, all of which were aligned and pre-processed for deep learning. Also, after carefully tuning the architecture and training process of the model, among the metrics evaluation it is recognised the disadvantage of using MAPE when actual values of traffic due to different factors are close to zero generating large percentage errors and make emphasis of the use of MAE. In total, eight baselines models that typically are used in time series problems are employed, meanwhile the current study employs only four of them which are Random Forest, XGBoost, LSTM and a hybrid of CNN-LSTM, with a variant in the deep learning approaches that will not include geographical features.

2.7 Orchestrating Urban Footfall Prediction: Leveraging AI and batch-oriented workflow for Smart City Application

This study by Komar & James (2024) advances urban mobility analytics by addressing the trade-off between predictive power and computational efficiency . This study employs classical machine learning algorithms like Random Forest and Gradient Boosting, using desktop-grade hardware. Their system prioritizes modularity and interoperability, employing an open-source stack (Scikit-learn, PostgreSQL) allows cities to integrate new data streams without restructuring core components. Furthermore, the study addresses edge cases such as low-footfall zones by implementing dynamic model thresholds. The ability of the system to handle incremental learning ensures that predictions remain responsive to changing urban patterns. This work contributes a replicable blueprint for smart cities seeking scalable, real-time predictive tools, particularly in regions with limited technological resources. However, while the system demonstrates operational efficiency and adaptability, it does not incorporate weather variables (a key variable for Dublin city), where hourly pedestrian and cycle flows are heavily influenced by environmental conditions.

2.8 What amenities drive footfall in UK town centres? A machine learning approach using OpenStreetMap data.

A study by Ntounis & Taecharungroj (2024), using OpenStreetMap dataset, applied random forest model to identify important amenities driving footfall in UK town centres. Their objectives included mapping 222,299 unique polygonal areas corresponding to 132 predictors variables (Ntounis & Taecharungroj, 2024, p. 2), and proposing an amenity classification (primary/secondary/tertiary). The study made significant contributions by identifying hotels and pedestrian paths as top predictors of footfall, followed by retail amenities like clothing shops, with a relative importance of 100%, 67%, and 59%, respectively. However, it entirely excluded weather variables, despite the well-documented influence of weather on amenity usage. This contrasts with Komar & James (2024) idea of weather-induced modal shifts in footfall but lacks granular amenity data. This omission implicates a critical blind spot: the attractiveness of amenities and its weather-importance, but models tend to isolate these factors like Ntounis & Taecharungroj (2024), while the static treatment of environmental conditions limits it from real-world applicability. This limitation aligns with the reliance on static weather inputs of Harrou et al. (2022), reflecting a broader oversight of dynamic bidirectional relationships between environmental conditions and urban systems according to Ntounis & Taecharungroj (2024).

2.9 Research Gap

In summary, while significant advancements have been made in traffic prediction and urban mobility analytics, a critical gap remains in the integration of dynamic environmental factors, particularly weather, into predictive models. Existing studies, such as Harrou et al. (2022) often treat weather as static inputs or exclude it entirely, overlooking its proven influence on pedestrian and cyclist behaviour. Other works, including Méndez et al. (2023) and, Rubasinghe and Hettiarachchi (2023), emphasize the importance of data integration and feature engineering. Through this approach, the study explores a novel and actionable methodology for urban mobility prediction that advances beyond existing literature. Unlike earlier studies that simply treated weather as static, this research adopts a more realistic approach by capturing weather as a dynamically changing, hour-by-hour phenomenon. By modelling weather as a time-varying feature, it uncovers how shifting conditions influence the way people move through the city, whether walking or cycling. Building on these ideas, this work introduces a flexible pipeline that merges dynamic weather data, the unique context of each counter location, and the power of advanced deep learning models. This approach not only improves prediction accuracy but also offers insights that city planners and decision-makers can truly use, helping to create smarter, more responsive urban mobility systems.

3 Research Methodology

The methodology used for this study is the Knowledge Discovery in Databases (KDD) explained by Fayyad et al. (1996) to investigate how deep learning techniques can be effectively leveraged to predict pedestrian and cycle flow in urban environments by integrating multi-source data (footfall counts, cycle counts, and weather conditions), refer to **Figure 1**.

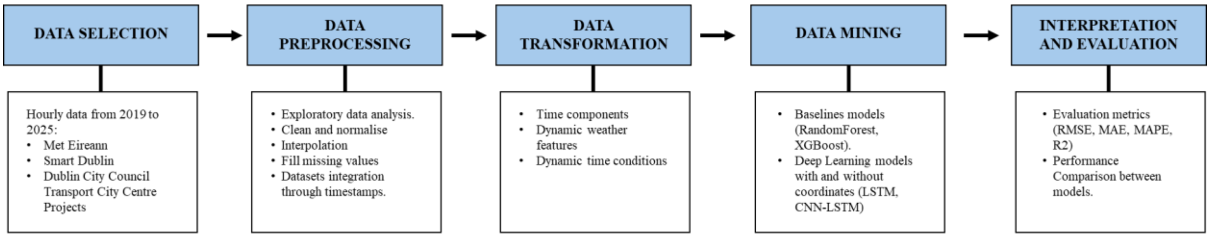

Figure 1: Methodology approach.

3.1 Data Selection

The data collection process involved gathering datasets of pedestrian footfall count and cycle count from Dublinked Open DataStore published by the Dublin City Council (DCC) Transport City Centre Projects (2020, 2025). It also included the counter locations by Smart Dublin (2025), and weather conditions data from © Met Éireann (n.d.), all of them published under the Creative Commons Attribution 4.0 International (CC BY 4.0).

3.2 Data Preparation

From a total of nineteen counters considered in this study, the plots in **Figure 2** show the seven busiest counters and it suggests the most frequent location for pedestrians to commute is Henry Street Merged, located near the shopping centre and public spaces followed Aston Quay Merged. These two locations show a stable trend throughout the time of study. On the other hand, for cyclists, Grove Road Totem stands out with very sharp peaks at around 8 AM and 5 PM, at these times, the counter reaches an average of more than three hundred cyclists, indicating this location as the major commuter route.

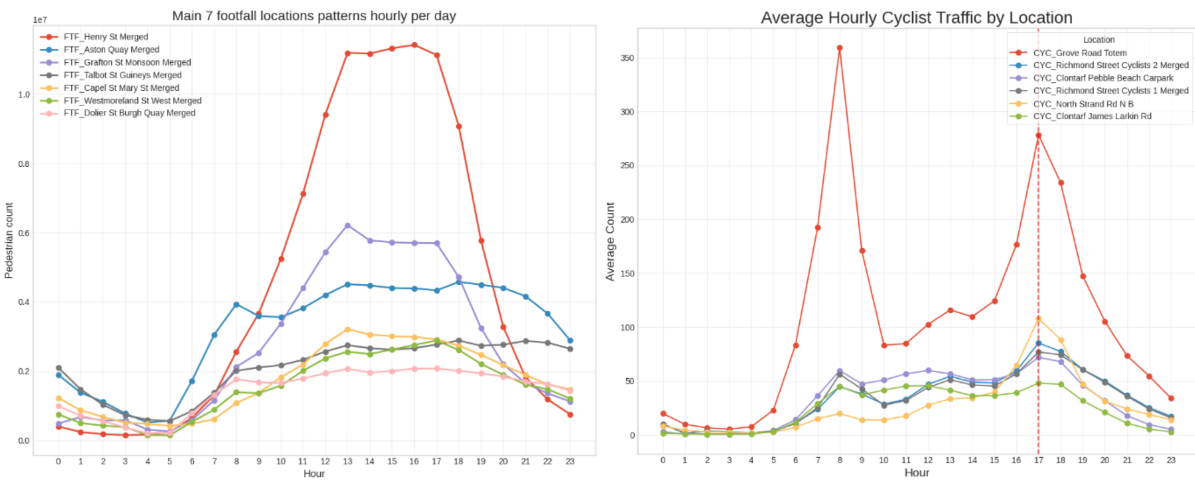

Figure 2: **Hourly traffic for cyclists and main seven footfall locations.**

The plot in **Figure 3** indicates the existence of a diurnal cycle where the highest temperature occurs between June and August. Per hour and month, the highest temperatures occur between 14:00 and 17:00 reaching the highest peak around 15:00 during the month of July.

Also, the map shows there are multiple locations (refer to **Figure 4**), in total there are eighty-five counters registered (cyclists, pedestrians, cyclist/pedestrians) but only nineteen were selected for this study after a robust selection process based on the quality of the available data.

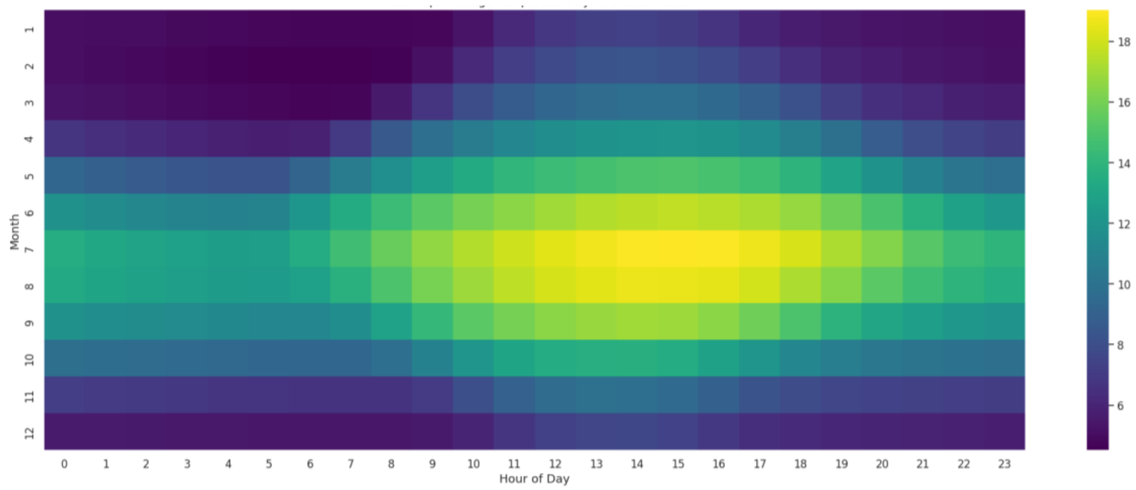

Figure 3: Average temperature by hour and month heatmap.

Figure 4: Counter locations in Dublin City.

3.3 Data Integration

The data was stored in Azure Blob storage and imported to a Python environment with the aid of Google Colab notebook. The feature extraction mechanism involved weather interaction, temporal features and time components of urban mobility (refer to **Table 1**).

Data integration was based on shared timestamps, which enabled feature alignment across sources for the period 2019-2025 in an hourly basis. Manual intervention was included as the name inconsistencies involved up to four different names for a single counter. A counter locations dataset from the Dublin City Council Transport City Centre Projects (2025) contained the different names and observation point one single counter could have and captures data from. As a result, a total of twenty counters were merged for footfall, and only two for cyclists (refer to **Table 2**).

Raw data from footfall and cyclist counters contained missing values. Linear interpolation was attempted but was not successful due to the large volume of gaps with missing values, resulting in an inaccurate ability to recognise footfall patterns precisely and damage to the data integrity. The formula applied for linear interpolation is the following:

$$y_2 = y(x_2) = y_1 + \frac{(x_2 - x_1)}{(x_3 - x_1)} . (y_3 - y_1)$$

Equation 1: Linear interpolation formula

Filling missing values with zeros was found to be a more reliable option, considering there is no record of pedestrians and cyclists in the time and location specified. Therefore, the large proportion of missing values led to the exclusion of several counters from this study to ensure quality in the analysis of patterns and behavioural trends. The percentages of missing or zero values accepted for this study were 45% for cyclists, leaving six counters available, and 40% for footfall, with thirteen counters available. Previous work (Méndez, G. Merayo, & Núñez, 2023) used only four locations, furthermore nineteen is considered reasonable for this study to evaluate different patterns around Dublin city.

Table 1. Feature information dictionary

Weather		Time	
Name	**Description**	**Name**	**Description**
WTH_rain	Precipitation Amount (rain)	**TEMP_year**	Year
WTH_temp	Air Temperature (C)	**TEMP_holiday**	Holiday
WTH_wetb	Wet Bulb Temperature (C)	**TEMP_hour_sin**	Hour sine
WTH_dewpt	Dew Point Temperature (C)	**TEMP_hour_cos**	Hour cosine
WTH_vappr	Vapour Pressure (hPa)	**TEMP_day_sin**	Day sine
WTH_rhum	Relative Humidity (%)	**TEMP_day_cos**	Day cosine
WTH_msl	Mean Sea Level Pressure (hPa)	**TEMP_month_sin**	Month sine
WTH_temp_3h	Rolling average temperature last three hours (C)	**TEMP_month_cos**	Month cosine
WTH_dew_spread	Dew Spread point (C)	**TEMP_weekend**	Weekend
WTH_rain_int	Rain intensity threshold	**TEMP_mor_peak**	Morning peak
WTH_prob_cond	Probability of condensation	**TEMP_eve_peak**	Evening peak
WTH_severe	Weather severity	**TEMP_bus_hours**	Business hours

Table 2: Examples of Counter data integration by name

Type	Final input variable	Original input variables
Footfall	**dawson street merged**	- dawson street - dawson street replacement - dawson street Molesworth - dawson street molesworth pedestrian
Cycle	**richmond street cyclists one merged**	- richmond street cyclists one - richmond street inbound

3.4 Data Transformation

This research directly addresses these gaps by developing a deep learning framework that models weather as a dynamic (using rolling averages, severity thresholds) and time-varying input, systematically quantifies its impact on mobility, rigorously compares models with and without spatial features, and includes time components, peak hours, sine and cosine calculations for temporal features. Feature engineering included features based on weather and temporal conditions, these were created using different formulas (refer to **Table 3**).

Table 3: Feature engineering description with formulas.

Feature	Description	Formula
Rolling averages	Average of temperatures within the last three hours (ACCA Global, n.d.).	$WTH_{\_temp_{3h(t)}} = \frac{1}{3}\sum_{i=0}^{2} WTH\_temp(t-i)$
Condensation probability through dew point spread	The closer the temperature and dew point temperature are the probability of condensation or fog increases according to National Weather Service (n.d.), and Fiveable (2025).	$WTH\_dew\_spread = WTH\_temp - WTH\_dewpt$ According to: $WTH_{\_prob\_cond} = \begin{cases} 1 & if\ WTH_{\_rhum} \geq 90\ and\ WTH_{\_dew\_spread} \leq 2 \\ 0 & if\ other\ conditions\ occur \end{cases}$
Weather severity thresholds	Considering rough weather conditions as low temperature and super heavy rain, inspired by the National Center for Hydrology and Meteorology (NCHM).	$WTH_{\_rain_{int}} = \begin{cases} 0 & if\ WTH_{\_rain} = 0\ (no\_rain) \\ 1 & if\ 0 < WTH_{\_rain} \leq 10\ (slight) \\ 2 & if\ 10 < WTH_{\_rain} \leq 30\ (moderate) \\ 3 & if\ 31 < WTH_{\_rain}\ 70\ (heavy) \\ 4 & if\ WTH_{\_rain} > 70\ (very\_heavy) \end{cases}$
Rainfall intensity classification	The rain intensity has been classified according to the Rates of rainfall of U.S. Geological Survey.	$WTH_{\_severe} = \begin{cases} +1 & if\ WTH_{\_temp} < 5\,°C` \\ +1 & if\ WTH_{\_rain} > 4\ mm/h \end{cases}$
Time components	Cyclical codification of time for hour, day and month through sine and cosine according to Amat Rodrigo & Escobar Ortiz (n.d.)	$TEMP_{\_hour\_sin} = \sin\left(\frac{2\pi TEMP\_hour}{24}\right)$, $TEMP_{\_hour\_cos} = cos\left(\frac{2\pi TEMP\_hour}{24}\right)$ $TEMP_{\_day\_sin} = \sin\left(\frac{2\pi TEMP\_weekday}{7}\right)$, $TEMP_{\_day\_cos} = \cos\left(\frac{2\pi TEMP\_weekday}{7}\right)$ $TEMP_{\_month\_sin} = \sin\left(\frac{2\pi TEMP\_month}{12}\right)$, $TEMP_{\_month\_cos} = cos\left(\frac{2\pi TEMP\_month}{12}\right)$
Urban mobility indicators	Detection of weekends, peak hours during morning and evening peaks, and business hours.	$TEMP\_weekend = \begin{cases} 1 & if\ TEMP_{weekday} \in \{5,6\} \\ 0 & if\ TEMP_{weekday}\ is\ other\ number \end{cases}$ $TEMP\_mor\_peak = \begin{cases} 1 & if\ 7 \leq TEMP_{hour} \leq 9\ and\ TEMP_{weekday} < 5 \\ 0 & if\ any\ other\ time\ and\ day \end{cases}$ $TEMP\_eve\_peak = \begin{cases} 1 & if\ 17 \leq TEMP_{hour} \leq 19\ and\ TEMP_{weekday} < 5 \\ 0 & if\ any\ other\ time\ and\ day \end{cases}$ $TEMP\_bus\_hours = \begin{cases} 1 & if\ 9 \leq TEMP_{hour} \leq 17\ and\ TEMP_{weekday} < 5 \\ 0 & if\ any\ other\ time\ and\ day \end{cases}$

4 Design Specification

Random Forest and XGBoost are selected as baseline machine learning models for this research, due to their interpretability and low demand for computational resources (Komar & James, 2024; Ntounis & Taecharungroj, 2024). They were preferred over ARIMA and support vector machine (SVM), as they can model regular trends but demonstrate difficulty to handle complex nonlinearity or data from multiple sources and be more sensitive to noise or disturbances in the data (Wang & Susanto, 2023). For deep learning architectures, LSTM and a hybrid CNN-LSTM are selected due to their proven reliability to capture temporal dependencies and traffic patterns to forecast spatiotemporal traffic and footfall in urban environments (Korbmacher & Tordeux, 2022; Wang & Susanto, 2023).

Models like a hybrid CNN-BiLSTM and GCN-LSTM demonstrate reliable performance; however they were not selected as they are computationally intensive and complexity increases, as they capture past and future context (Méndez, G. Merayo, & Núñez, 2023), and model explicit spatial relationships in traffic networks (Fang, Pan, Chen, Du, & Gao, 2021), respectively. Transformers models were also not considered due to large-scale datasets requirement, which exhibits a high computation cost (Harrou, Dairi, Zeroual, & Sun, 2022).

The selected models will be assessed with metrics like root mean square error, mean absolute error, coefficient of determination, and mean absolute percentage error. Therefore, performance and predictions output visualisations are performed.

5 Implementation

The implementation employed Python 3 using two separate configurations in the Google Colab Pro+ environment. The first environment was configured to train baseline machine learning models and handling results DataFrame with a CPU accelerator with high RAM mode (up to 51 GB system RAM). The second environment employed an NVIDIA GA100 accelerator (40 GB VRAM) and up to 83.5 GB system RAM for training advanced machine learning and deep learning architectures. This environment was set up for the computational power necessary to train complex sequences. Microsoft Azure was employed for data storage, organisation, and access to all datasets during the research process. The use of Azure guaranteed scalability and reliability made data integration and retrieval more efficient with the computational environment implemented during training and evaluation phases.

5.1 Random Forest and Extreme Gradient Boost (XGBoost)

A robust and modular pipeline was implemented according to the complexity of datasets, and the need of interpretable and comparable results (refer to **Figure 5**). A base configuration was established and the hyperparameters defined, such as number of trees, maximum depth, and random seed to ensure the reproducibility of the experiment; including a standard scaler to predictor variables normalisation with dictionaries to store the models and results obtained. Kavlakoglu & Russi (n.d) suggest that the Extreme Gradient Boost pipeline, unlike Random Forest, improves gradient boosting optimisations through a setup that includes a learning rate determining how quickly the algorithm learns from each iteration.

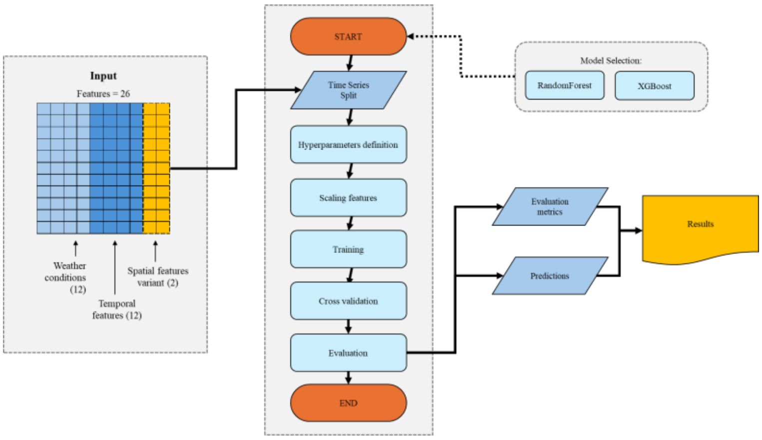

Figure 5: RandomForest and XGBoost models implementation.

The training stage included scaling all input features and training a Random Forest and XGBoost model for each target series independently, capturing the characteristics of each location type, whether it is a pedestrian or cyclist counter and most importantly, predicting whole numbers as predictions of pedestrians and cyclists. Meanwhile, TimeSeriesSplit was employed to obtain the evaluation dataset to apply a cross-validation method according to Amat Rodrigo, J. & Escobar Ortiz, J. (n.d.) to preserve the chronological time-ordered sequence of the data and prevent future data leakage into the training process. In which, each fold of validation, calculated key evaluation metrics.

Additional functions were developed to organise and compare results between categories of pedestrians and cyclists and generate reports of recent predictions to assess the capacity of the model behaviour in recent scenarios.

5.2 Long Short-Term Memory network (LSTM)

The pipeline designed for LSTM neural networks model, follows a structure to manage multiple targets time-series able to forecast a sequence of twenty-four-time hourly pattern for multiple counters (cyclist and pedestrian). Initially, the feature engineering is applied, where interaction features are combined with temporal features and weather variables. The pipeline includes normalising, generating sequences, and scaling independent variables for each target series. Two variants of the LSTM model were developed for comparison purposes, in which only one includes geographic coordinates of the counter location (latitude, and longitude), that allows the model to learn both spatial and temporal patterns; meanwhile the other exclude geographic coordinates (refer to **Figure 6**).

The LSTM model architecture was designed with two layers of LSTM, with 64 and 32 units, respectively, each uses dropout regularisation to prevent overfitting, followed by a dense layer and a single output node. The training phase included the Adam optimiser and mean squared error (MSE) loss. Additionally, early stop and learning rate reduction techniques were applied automatically in the training process to make it more efficient and avoid data overfitting, and a separate LSTM model was trained for each counter using the distribution strategy of TensorFlow ensuring GPU usage. After training was completed, the predictions were deescalated and rounded to whole numbers as the nature of the count is people. The model performance was assessed and predictions compared against real values.

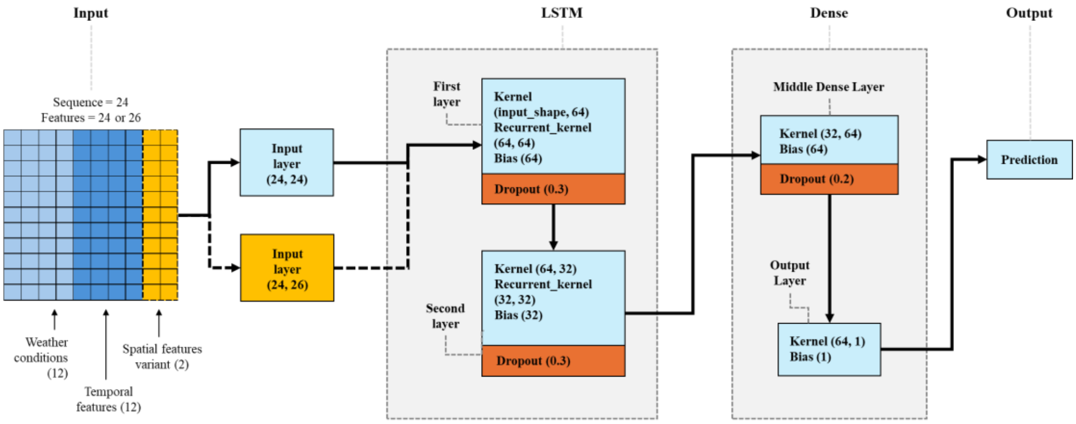

Figure 6: LSTM architecture flow diagram.

5.3 Convolutional neural network and long short-term memory (CNN-LSTM)

To forecast multiple time series, a model that combines CNN, and LSTM networks has been designed (refer to **Figure 7**). It creates sequences constructed from the normalised features, in which a model variant includes spatial features (latitude and longitude) added to each target. The sequences are created in a fixed sliding window of twenty-four steps, creating three-dimensional inputs (samples, twenty-four time-steps per sample, features per step). Then, the data is split into training, validation and testing test while preserving the temporal sequence. The architecture begins with 1D convolutional layers to extract short-term patterns, followed by LSTM layers that capture long-term dependencies. To prevent overfitting, the model incorporates L2 regularisation, batch normalisation, and dropout. Training is performed using the Huber loss function and the Adam optimiser, including callbacks for early stopping and an adaptive learning rate reduction.

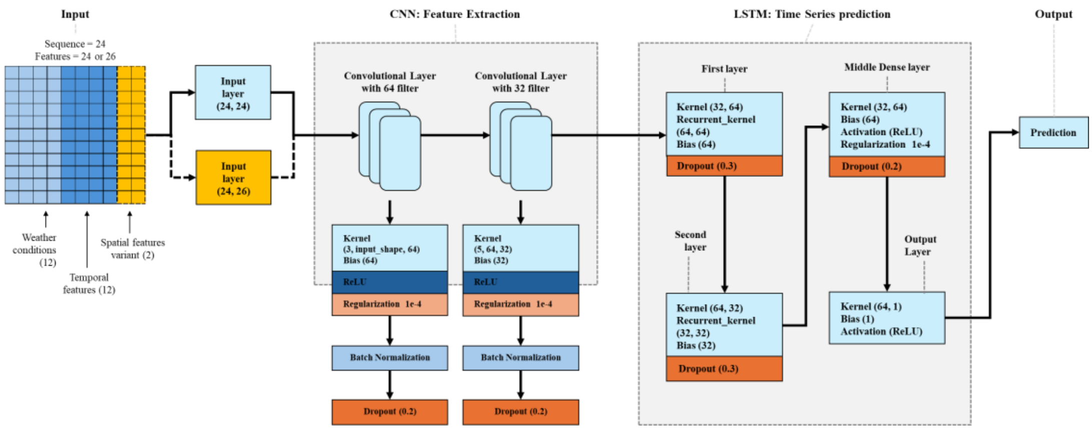

Figure 7: CNN-LSTM architecture flow diagram.

6 Evaluation

In this section, the models proposed are evaluated using four metrics to each model (refer to **Table 4**), comparing baselines models against deep learning techniques. For LSTM and CNN-LSTM experiments, data was temporally split into 70% for training, 15% for testing and 15% for validation ensuring sequential and non-overlapping splits. Meanwhile for baseline, the split employs a five-fold TimeSeriesSplit cross validation to have different fold combinations of train/test and evaluate the stability of metrics with no specific validation split percentage. Improvement of the model is indicated by lower values of each except the R^2, for which improvement is indicated by higher values.

Table 4: Evaluation metrics formulas

Metric	Formula
MAE	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
R^2	$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}$ where results are 1=perfect fit, 0= no predictive power
MAPE	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$

Where:

$\hat{y}_i = predicted\ value\ of\ y$

$\bar{y}_i = mean\ value\ of\ y$

$n = total\ number\ of\ observatioxns$

It should be noted that MAPE is a typically used metric in regression and results were obtained. However, Méndez et al. (2023) highlight its disadvantage to yield in extremely large percentage errors or producing undefined errors which was detected and obtained with the current data, resulting in MAPE being discarded from this study. Also, for visualisation purposes some counters are not included.

6.1 Mean Absolute Error (MAE) results.

Overall, MAE values for pedestrians are higher than cyclists and this is constant, across all modelling frameworks implemented. The significant differences in data magnitudes between cyclists and pedestrians impact the error produced, as for pedestrians an average error is hundreds, meanwhile for cyclists it is reduced to less than hundred units (refer to **Figure 2** and **Figure 8**). Random Forest and XGBoost seem to be outperformed by LSTM and CNN-LSTM on pedestrian counters, however they show similar predicting power for cyclists.

LSTM without geographical coordinates demonstrates the best improvement over baselines models for pedestrians in Grafton St Monsoon Merged by improving MAE values by 59.18% (-813.49) and 59.87% (-837.3) compared to Random Forest and XGBoost, respectively; followed by Westmoreland St West Merged, employing the same model, which demonstrates an improvement of 55.16% (-310.69) and 59.11% (-364.97) against Random Forest and XGBoost, respectively. Which means Random Forest and XGBoost predictions will be at least two times higher than using LSTM without geographical coordinates. CNN-LSTM with spatial features achieves a low MAE value of 83.80 for Grafton St Nassau St Suffolk St. On the other hand, for cyclists, the most suitable improvement for average errors is detected in Clontarf James Larkin Rd, where the higher improvement is seen applying LSTM and CNN-LSTM, both without geographical coordinates over Random Forest, with improvements of 30.99% (-5.6) and 35.14% (-6.35) respectively. Also, Grove Road Totem suggest using LSTM with geographical coordinates rather than without them, achieves a reasonable MAE decrease of 49.00 (53.89%).

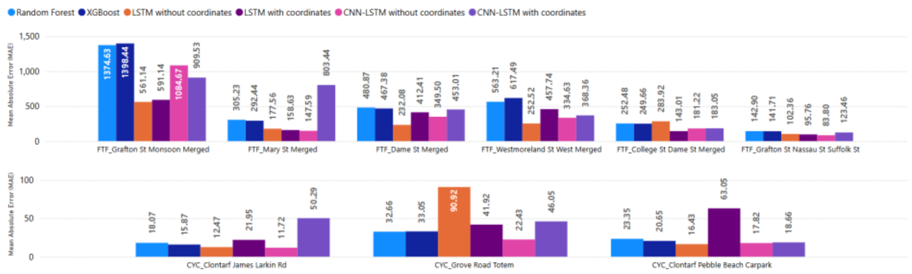

Figure 8: Mean Absolute Error (MAE) results.

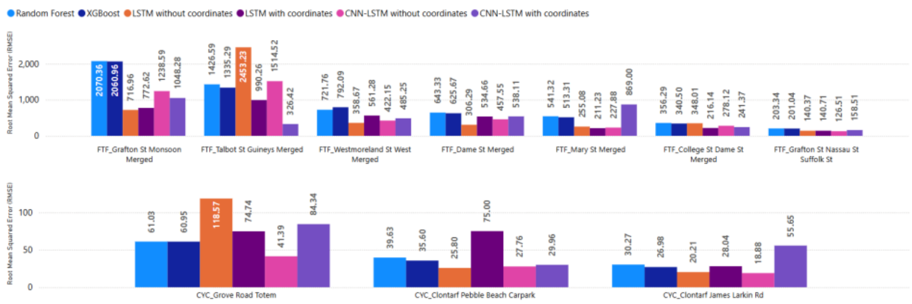

Figure 9: Root Mean Squared Error (RMSE) comparison results by model.

6.2 Root Mean Squared Error (RMSE) results.

Generally, RMSE values for pedestrians are higher than cyclists, as a magnitude of errors in predictions (refer to **Figure 2** and **Figure 9**). Grafton St Monsoon Merged and Talbot St Guineys Merged registers a high value of pedestrian count, which leads to a higher error using Random Forest and XGBoost, but these are outperformed in most scenarios by deep learning techniques. LSTM and CNN-LSTM, mostly without coordinates, indicate lower RMSE in its application over the inclusion of coordinates, i.e. for Westmoreland St West Merged, applying LSTM without geographical coordinates compared to XGBoost improves RMSE in 54.72% (-433.42) rather than coordinates that improves 29.14% (-230.81) only. Meanwhile, Colle St Dame St Merged and Grafton St Nassau St Suffolk St achieves a constant increasing performance by reducing the error with deep learning over baseline models.

Furthermore, for cyclists, Clontarf James Larkin Rd using LSTM and CNN-LSTM, both without geographical coordinates, both indicate overall a great performance over Random Forest by achieving incredibly low RMSE values, with the highest improvement values of 33.23% (-10.06) and 37.63% (-11.39), respectively.

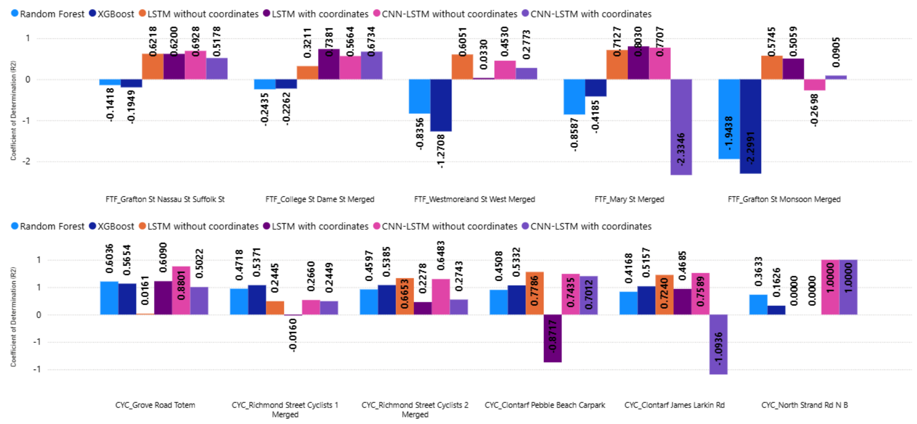

Figure 10: Coefficient of Determination (R^2) comparison results by model.

6.3 Coefficient of Determination (R^2) results

The values obtained for pedestrian varies depending on the locations, which suggest the models in a moderate way, are able to explain the variations observed in data, and for those in negative values, it indicates it could predict far from the real value (refer to **Figure 10**). Most LSTM and CNN-LSTM models without coordinates suggest better performance with values over 60% such as Grafton St Nassau St Suffolk St, College St Dame St Merged and Mary St Merged. In contrast, Random Forest and XGBoost registered high negative R^2 values in most of pedestrian locations which indicates these cannot explain the variability through the features included in the models. In contrast, for cyclists, only North Strand Rd N B is considered overfitted with a value of 1.00, and it could be related to the quality of the data and not learning real patterns in it. Grove Road Totem provides the highest value registered with 88.01% suggesting the features included in CNN-LSTM without coordinates explain its variability. And despite having a negative R^2, Clontarf Pebble Beach Carpark and Clontarf James Larkin Rd indicate one negative R^2 value, most of the results are constant in positive values for the rest of the models applied, especially for LSTM without geographical.

The highest improvement over baselines models is observed in pedestrians in Grafton St Monsoon Merged, after baselines models R2 values are -1.9438 and 2.2991 for Random Forest and XGBoost respectively, while LSTM values are above 0.50. The inclusion of geographical features may not be sufficient to explain the variability of the models, therefore, results with them are lower overall.

6.4 Mean Absolute Percentage Error (MAPE) results.

Overall, MAPE values obtained are extremely high. The highest values are calculated using Random Forest and XGBoost models, for both pedestrians and cyclist. Therefore, the highest value registered belongs to Henry St Merged as 3.14 E+20 % (Random Forest), followed by its application with XGBoost (2.65 E+20 %). Also, reasonable low values belong to deep learning techniques applications, such as CNN-LSTM without geographical coordinates in College St Dame St Merged (61.17%) and Mary St and Merged (71.26%). For cyclist counters, employing LSTM without coordinates generates MAPE values that remain low such as Grove Road Totem (47.37%). On the other hand, Richmond Street Cyclist 2 Merged achieves 46.22%

and 54.68%, using LSTM and CNN-LSTM, both without geographical features considering it one of the lowest MAPE registered (refer to **Figure 11**).

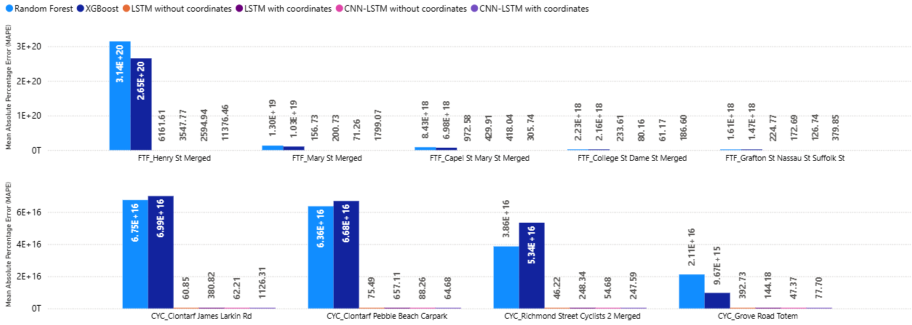

Figure 11: Mean Absolute Percentage Error (MAPE) comparison results by model.

6.5 Forecast values.

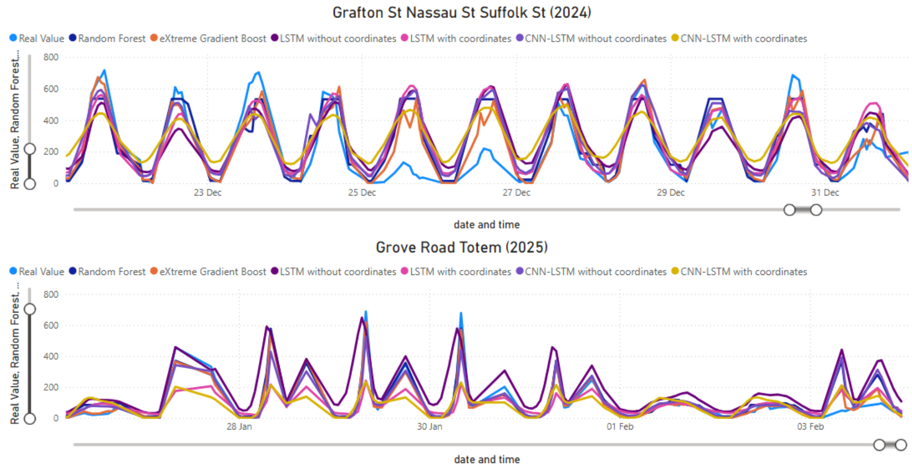

Figure 12: Real and forecasted values from models.

Taking a sample of around seven days of prediction across different years suggests that the models follow the shifting trend of the real value. However, considering Grafton St Nassau St Suffolk St had constantly low values in evaluation metrics, when using CNN-LSTM with geographical coordinates tends soften and predict more than average values in the period between December 25th and 27th (refer to **Figure 12**), forecasting values closer to the mean, which suggests difficulties in capturing non-ordinary events or sudden drops.

On the other hand, for cyclist, most models achieve values close to the real values overall, but they show limitations with forecasts and tends to smooth significant or extreme variations.

6.6 Discussion

This study is exploratory, and it has advanced the integration of dynamic weather and time features for urban mobility prediction. The models developed were tuned to the Dublin setting with its distinctive highly shifting meteorological conditions.

Prior studies, such as Fang et al. (2021) with the MDTP framework and Wang & Susanto (2023) with their hybrid CNN-LSTM model, provided evidence that the results from the current study confirm that deep learning models as LSTM and CNN-LSTM in particular, are capable of outperforming traditional models like Random Forest and XGBoost to forecast complex time series that include abrupt and shifting patterns, that is very frequent in pedestrian counts. The main improvement in MAE and RMSE occurs applying CNN-LSTM without geographical coordinates in Mary St Merged and Grafton St Nassau St Suffolk St (refer to **Figure 8**, **Figure 9** and **Figure 9**), locations with a high volume of pedestrians, where the evaluation metrics values shows reduction for MAE and RMSE, and a reasonable increase for R2, compared to baselines models, in which also noted by Méndez et al. (2023), hybrid neural networks have demonstrated the ability to capture the complexity of spatiotemporal dynamics. Geographical coordinates could represent noise when only latitude and longitude are available and considered insufficient to capture spatial relationships as mentioned by Fang et al. (2021).

The efficiency of the implementation of cloud-based computational resources (Azure and Google Colab), like MDTP framework by Fang et al. (2021), and the use of GPU acceleration and parallelisation allowed the training and evaluation of complex models, supporting the idea of replicability by Komar & James (2024) for cities looking for urban mobility solutions.

This study aligns with Komar & James (2024) in identifying some locations have insufficient or no footfall data, which is inadequate to train models, and incomplete data generates difficulties. Consequently, data interpolation was avoided to prevent bias and considered feature engineering to include more explanatory features instead. Furthermore, the trend of extremely high MAPE values was confirmed, especially for baselines and in some cases, for deep learning models, as highlighted by Méndez et al. (2023). However, unlike Komar & James (2024), this study found higher MAPE value is produced in high-flow footfall locations, rather than low-flow ones. This supports Rubasinghe & Hettiarachchi (2023) recommendation to use metrics such as MAE and RMSE to evaluate the model performance.

The success of LSTM lies in its ability to learn temporal relationships over twenty-four timesteps from low memory loss in short-term sequences. However, the network may also memorise noise when variables are highly correlated. The length of the time series plays a key role reducing overfitting as longer sequences require more data to generalisation, which could explain the error increases with the inclusion of geographical coordinates in the models. In the long term, the inclusion of sequences with different lengths and higher temporal granularity could improve denoising the data and improve the effective memory of the model.

Although it shows powerful predictive power, CNN-LSTM in some cases is outperformed by LSTM, which indicates limitations in hybridisation. Overall, the computational resources are more intensive, including more RAM and training time that limits its applicability to large-scale data. While the CNN detects local patterns in the timesteps, long-term trends could be skipped as they are outside of temporal coverage that the LSTM kernel design allows (refer to **Figure 7)**. LSTM sensitivity to high correlation variables means LSTM layers could receive noise input and not reconstruct long term relationship, as the CNN has already processed it.

In summary, Random Forest and XGBoost resulted useful for scenarios with less temporal complexity and low volume of data, meanwhile deep learning architectures as LSTM and CNN-LSTM excel in capturing highly volatile dynamics in urban environments.

7 Conclusion and Future Work

The primary research objective was to evaluate how deep learning models, integrated with weather data, could improve the prediction of pedestrian and cycle flows in Dublin city over established baselines. This study provides a novel contribution to the limited previous work

(Harrou, Dairi, Zeroual, & Sun, 2022; Ntounis & Taecharungroj, 2024), modelling dynamic meteorological variables and time-varying inputs instead of static features and succeeds in the error reduction for hourly pedestrian and cyclist flow predictions of deep learning models implemented for many counters located in Dublin city.

Deep learning architectures, specifically LSTM and CNN-LSTM excluding geographical coordinates, demonstrate superior capability to forecast pedestrian and cyclist counts for volatile urban environments. However, geographical features, such as latitude and longitude, do not necessarily increase the predictive performance, indicating urban setting and temporal dynamics are predominant and suggesting the need for more spatial availability.

Multi-source data integration with urban data is feasible employing open-source and cloud-based solutions and can support a reproducible and scalable deployment for city authorities looking for smart mobility solutions.

Recognising methodological and practical limitations of the current research creates directions for future work:

- Strengthen data completeness to improve data integrity for footfall and cyclists counts, using sophisticated imputation methods for missing values to evaluate more set of locations that were not accessible in this study.
- The inclusion of real-time event data, i.e. concerts, sports, construction; and increase granularity of data to strengthen model predictions to handle abrupt increase/decrease shifts in mobility.
- Generalisation of the model from Dublin to other cities, but the success of the implementation will depend on the integrity of local data collection.
- Examine deeper the spatial dimension through advanced methods or graph-based neural networks as suggested in related literature. In the current study, the impact of spatial features inclusion leads to noise in data and performance decreases.
- Explore the application of pedestrian footfall and meteorological conditions on businesses consumption and sales influence in locations close to counters.

Continued research in this area reinforces the importance of data-driven urban planning which creates more efficient, sustainable, and responsive cities suited for needs of the population.

8 References

ACCA Global. (n.d.). *Time series and moving averages*. Retrieved June 8, 2025, from https://www.accaglobal.com/gb/en/student/exam-support-resources/fundamentals-exams-study-resources/f5/technical-articles/time-series.html

Amat Rodrigo, J., & Escobar Ortiz, J. (n.d.). *Cyclical features in time series forecasting*. Retrieved June 2, 2025, from https://skforecast.org/0.9.0/faq/cyclical-features-time-series

Dublin City Council Transport City Centre Projects. (2020, 8 14). *Pedestrian Footfall DCC*. Retrieved Feb 18, 2025, from https://data.smartdublin.ie/dataset/dublin-city-centre-footfall-counters

Dublin City Council Transport City Centre Projects. (2025, 2 5). *Cycle Counts DCC*. Retrieved Feb 7, 2025, from https://data.smartdublin.ie/dataset/dublin-city-centre-cycle-counts

Fang, Z., Pan, L., Chen, L., Du, Y., & Gao, Y. (2021). MDTP: A Multi-source Deep Traffic Prediction Framework over Spatio-Temporal Trajectory Data. *Proceedings of the VLDB Endowment, 14*(8), 1289 - 1297.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).* Portland, Oregon. Retrieved from https://cdn.aaai.org/KDD/1996/KDD96-014.pdf

Fiveable; Bahr, Becky. (2024). *5.4 Dew point and relative humidity – Meteorology.* Retrieved June 2, 2025, from https://library.fiveable.me/meteorology/unit-5/dew-point-relative-humidity/study-guide/RfxicYTAmshEsWlH

Harrou, F., Dairi, A., Zeroual, A., & Sun, Y. (2022, April 28). Forecasting of Bicycle and Pedestrian Traffic Using Flexible and. *Applied Sciences*, pp. 1-21.

Komar, T., & James, P. (2024). Orchestrating Urban Footfall Prediction: Leveraging AI and batch-oriented workflow for Smart City Application. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48*.

Korbmacher, R., & Tordeux, A. (2022, 9 9). Review of Pedestrian Trajectory Prediction Methods: Comparing Deep Learning and Knowledge-based Approaches. *IEEE Transactions on Intelligent Transportation Systems, 23*(12), 24126-24144. Retrieved Mar 30, 2025

Kumar, U., & Sharma, N. (2024). A Scientific Analysis of Weather Forecasting Techniques and its Classification. *Journal of Management World, 4*(10), 917-923.

Méndez, M., G. Merayo, M., & Núñez, M. (2023). Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model. *Engineering Applications of Artificial Intelligence, 121*(106041), 1-15.

Met Éireann. (n.d.). *Historical Data - Display and Download Historical Data from current stations*. Retrieved Feb 7, 2025, from https://www.met.ie/climate/available-data/historical-data

National Center for Hydrology and Meteorology (NCHM). (n.d.). *Rainfall intensity classification.* Retrieved June 2, 2025, from https://www.nchm.gov.bt/attachment/ckfinder/userfiles/files/Rainfall%20intensity%20classification.pdf

National Weather Service. (n.d.). *What is Relative Humidity?* Retrieved June 2, 2025, from https://www.weather.gov/lmk/humidity#:~:text=Relative%20humidity%20(RH)%20

Ntounis, N., & Taecharungroj, V. (2024, Oct). What amenities drive footfall in UK town centres? A machine learning approach using OpenStreetMap data. *Urban Analytics and City Science*, pp. 1-19.

Rubasinghe, N., & Hettiarachchi, P. (2023). *Regression Analysis for Prediction of Road Travel Speed.* Malmö: Malmö University.

Smart Dublin. (2025, Mar 10). *Eco Counter Locations.* Retrieved July 2, 2025, from https://data.smartdublin.ie/dataset/pedestrian-and-cycle-counter-api-for-dublin-region/resource/f80d1ec3-e915-4ef0-988f-e3304b97dd37?view_id=71a929c0-b6f8-48ac-86bf-a2d58bc0eee7

Sustrans. (2024). *Dublin Metropolitan Area Walking and Cycling Index 2023.* Dublin: National Transport Authority.

Wang, J.-D., & Susanto, C. O. (2023). Traffic Flow Prediction with Heterogenous Data Using a Hybrid CNN-LSTM. *Computers, Materials & Continua, 76*(3), 3097-3112. Retrieved from https://doi.org/10.32604/cmc.2023.040914