What novel AI-driven approach can Dublin's traffic systems adopt to integrate weather data and improve daily congestion prediction beyond existing studies?

MSCDA_JANOL24_O

John Slattery
Student ID: x23265981

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	John Slattery
Student ID:	x23265981
Programme:	MSCDA_JANOL24_O **Year:** 2025
Module:	Research Project Report
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	11/12/2025
Project Title:	What novel AI-driven approach can Dublin's traffic systems adopt to integrate weather data and improve daily congestion prediction beyond existing studies?
Word Count:	5994 **Page Count** 20 (21 including references)

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 another author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………

Date: ………

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):	

What novel AI-driven approach can Dublin's traffic systems adopt to integrate weather data and improve daily congestion prediction beyond existing studies?

John Slattery
x23265981

Abstract

Existing traffic forecasting methods often rely solely on past traffic flow and do not account for environmental factors. This research instead focuses on strengthening Dublin's traffic management by forecasting short-term traffic flow using combined meteorological data and machine-learning algorithms. Data from the traffic flow in South Dublin County Council's Smart Dublin system over multiple years was combined with data from Met Éireann on rainfall, air temperature, and other factors.

A data processing pipeline has been implemented to clean and merge both data sources, followed by feature engineering that includes time- and season-related features. The three approaches are Random Forest models, Gradient Boosting models, and hybrids of both methods combined into a single model.

The research makes several contributions to forecasting accuracy while aligning with ethical considerations. The results show that the Hybrid Ensemble model improves prediction accuracy and demonstrates the value of integrating meteorological data into Dublin's traffic forecasting systems.

Introduction

Traffic congestion is a major challenge faced by cities worldwide. For Dublin, being one of the fastest-growing capitals in Europe, traffic congestion has become more pronounced. The rising number of commuters, together with the widespread pattern of residential areas, has resulted in high traffic density. The current traffic system is under strain, given today's levels. This has been highlighted in recent months particularly as the infrastructure around Dublin is reaching its maximum capacity.

Traditional traffic management methods, such as adaptive traffic control systems, route guidance systems, an incident response plans, address part of this problem but remain primarily based on anticipated regular patterns of traffic behaviour. Many of these traffic management methods rely on stable environmental assumptions and do not account for meteorological variations in traffic prediction or operation. This assumption seems less valid. Certain empirical studies verify strong inter relationships between traffic flow and a number of meteorological variables such as humidity, mean sea level pressure, precipitation, temperature, and visibility. Weather extremes may lead to decreasing visibility, varying levels of braking action, differing levels of caution among road-users, and changing modes of travel. All these factors affect traffic volume.

All such factors are of utmost importance to Dublin, a region with a maritime climate that experiences dramatic changes in weather. The pressure and wind systems demonstrate systematic changes in a relatively short space and time in such a region. Consequently, the non-linear variations in traffic patterns are not adequately accounted for in current forecasting models. However, the approaches used in practice often consist solely of autoregressive forecasting models based on traffic trend data. This creates both challenges and gaps in this approach, as it is ineffective at traffic forecasting under unusual or varied environmental conditions.

Traffic flow data is available from the Smart Dublin system administered by South Dublin County Council. This traffic information comprises data spanning several years on the daily use of major roads. Also, data from Met Éireann on atmospheric conditions are of high quality and include rainfall, temperature, humidity, vapour pressure, wet-bulb temperature, and mean sea-level pressure. All this data is anonymously available.

The current research builds on these strengths to address a specific research gap in the literature on traffic modelling in Ireland: the underuse of up-to-date meteorological data. While research in other countries has identified the successful use of meteorological data in traffic modelling, there appears to be limited research in the Dublin metropolitan region. To fill this gap, this project proposes and compares various machine-learning algorithms, including Random Forests and Gradient Boosting, that incorporate traffic and atmospheric data. Multiple ensemble algorithm selection would be beneficial in traffic forecasting and ensuring that the interpretation of the results remains clear. The methodology involves combining traffic data from multiple years and atmospheric data in feature engineering based on temporal and atmospheric aspects and testing the different traffic forecasting models. The interpretability of results, in terms of feature importance values and SHAP values, would demonstrate the proposed traffic forecasting system's interpretability and utility in the public domain.

Report Structure

The remainder of this report is structured as follows. Section 2 looks at current literature on traffic forecasting, weather integration, and AI techniques. Section 3 details the data outputs. Section 4 outlines the research methodology from data acquisition to ethical considerations. Section 5 outlines the design and implementation of the models from system overview to final system artifacts. Section 6 discusses results and evaluation of the work undertaken. Section 7 finishes with conclusion, discussion and future work.

Research Question:

What novel AI-driven approach can Dublin’s traffic systems adopt to integrate weather data and improve daily congestion prediction beyond existing studies?

Research Objectives

- This research aims to systematically gather, clean, and integrate multi-year daily traffic flow data from South Dublin County Council, along with meteorological data obtained from Met Éireann
- This research aims to focus on developing and evaluating machine learning models, specifically using techniques such as Random Forest and Gradient Boosting, as well as a hybrid ensemble approach, to improve the accuracy of daily traffic flow predictions
- This research aims to assess model performance using MAE, RMSE, and R^2 metrics.
- This research aims to design and implement a machine-learning framework capable of forecasting short-term traffic flow

Literature Survey

Introduction

This section reviews prior work on short-term traffic-flow forecasting, with a particular focus on:

- Traditional and hybrid predictive approaches
- Deep learning architectures
- Integration of weather and other external factors
- Emerging discussions on ethical and trustworthy AI in urban mobility

The aim is to position this research within current academic and practical developments and to identify the gap that motivates a weather-informed, interpretable ensemble model tailored to Dublin.

1. Traditional and Hybrid Traffic Forecasting Approaches

Traditionally, traffic prediction techniques rely on time series analysis and statistical methods such as ARIMA and Kalman filters, which require traffic variables to be modelled as a stationery process with linear relationships among variables. This model performs satisfactorily when traffic variables are linear and have less variations; however, this model does not perform satisfactorily when traffic variables are nonlinear.

However, owing to the increased availability of data and improvements in computing capacity, there have also been many advancements in traffic forecasting models from conventional methods towards more intelligent approaches using advanced machine learning models. Recent literature reviews regarding AI-based traffic flow prediction highlight the importance of tree-based and ensemble learning techniques such as Random Forest and Gradient Boosting Regressor for short-term traffic prediction tasks (Sayed et al., 2023). Such methods are most appropriate for dealing with complex relationships existing in traffic data. The Random Forest model and the Gradient Boosting Regressor are renowned ensemble learning methods that can help establish complex relationships in data with fewer assumptions compared to other machine learning algorithms. Also, RF and GBR algorithms score the feature importances, which are helpful in identifying variables that are responsible for traffic trends.

From a literature perspective, a study by Hou et al. (2021) has shown that blending various factors such as atmospheric humidity, wind speed, and pressure while predicting results in a more accurate forecast when using prediction models. This was also explained in a similar study by Chai et al. (2022) regarding using Multi Feature Fusion Model.

Despite this progress, most of the existing literature focuses either on urban areas or other cities outside of Ireland. Also, a major research gap appears in literature regarding ensemble models with explicit consideration for weather variables in Dublin or similar maritime climate cities. Additionally, a relatively low focus appears on techniques that prioritize transparency and relevance for public sectors. This points towards a need for more research and development in this particular domain.

2. Deep Learning Architectures for Traffic Prediction

Deep learning solutions have recently become more prevalent in traffic prediction. Among them are Recurrent Neural Networks (RNNs), such as Long Short-Term Memory Networks

(LSTMs) and Gated Recurrent Units (GRUs), both of which have made significant contributions to traffic prediction. Ma et al. (2015) have shown that LSTMs are more effective than traditional forecasting methods for modelling traffic data. Zhang and Kabuka (2018) later improved this accuracy using GRU networks, considering atmospheric conditions. The use of the combination of modelling strategies has proved successful. Zhuang and Cao (2022) developed a CNN-BiLSTM method that combines convolutional neural networks with bidirectional LSTMs in one structure in which bidirectional LSTMs play a crucial role in the analysis of data over time. Afandizadeh et al. (2024) argue that comparison studies have shown that a deep network outperforms a standard network in this regard, as long as sufficient samples have been obtained.

Also, Graph Neural Networks (GNNs) are another prominent advancement in this regard, as they help better represent road network data. This fact can be observed in research such as Qiu & Zhao (2024). However, several challenges exist when using deep learning techniques in public organizations regarding requirements such as a large amount of data, extensive hyperparameters adjustment, and great computational power requirements. Additionally, a major drawback of using deep learning techniques is a low level of interpretability. This leads to questions about model transparency. Therefore, this research focuses on tree-based ensemble techniques that may be more interpretable while improving prediction capability for traffic prediction..

3. Weather and External-Factor Integration

There is strong evidence supporting that meteorological factors affect urban traffic conditions. Researchers Mitsakis et al. (2014) indicated that intense storm conditions, such as heavy rainfall, affect urban traffic conditions in terms of lower speeds and route deviation. This indicates that extreme storm conditions affect traffic conditions. Wu and Zhang (2024) clearly explained and indicated that increased temperatures reduce congestion delay, while rainfall, wind speed, and levels of congestion yield adverse effects for congestion delay. A nonlinear inverse-U association was identified for congestion delay and precipitation for 98 Chinese cities. A Study conducted by Zhang and Kabuka (2018) indicated a better level of prediction with effective use of deep learning methods considering meteorological factors compared to those considering traffic factors only. A better prediction level was achieved with a combination of traffic factors and meteorological factors according to Hou et al. (2021). This series of studies indicates that meteorological factors should be considered, especially for varying climatic conditions.

Despite this integration, some challenges remain. The data resolution may not match in this integration. Traffic data are measured in one-minute intervals. The other variables measure either hourly or daily. Traffic and other variables may form a nonlinear association. Additionally, when several events occur together, this association may become more complex. Although research such as the TomTom Traffic Index demonstrates variations in traffic congestion driven by local conditions, there are currently no studies in Ireland on this topic. This research aims to fill this gap with a free source of meteorology data available via Met Éireann along with traffic data from South Dublin County Council. Information about methodologies used in conducting this study can be found in the Methodology.

4. Ethical AI and Governance in Urban Mobility

As more and more of urban infrastructure gets integrated with artificial intelligence (AI), questions of equity, accountability, and transparency have come forward as major themes for studies. Sawhney (2023) writes that urban intelligent systems need to be optimized for efficiency while maintaining standards for rights and responsibilities in mobility systems. Traffic prediction systems need to be user-friendly and traceable with low levels of congestion and emissions.

The Guidelines for Trustworthy AI issued by the European Union contains criteria to assess AI systems from various angles: human control, technical integrity and interpretability, privacy, transparency, avoiding discrimination, and alignment with socially valuable outcomes. The realization of this guideline in the context of transport forecasting would involve, among other things, the use of anonymized or aggregated data if possible in preference to identifiable data, ensuring the existence of a clear and traceable path from inputs through to outputs, recording assumptions underlying modeling choices, and ensuring that traceability and explainability are possible. The use of black box approaches in AI applications in the public sector would be problematic if the approaches are not interpretable but are highly accurate. The lack of being able to determine the reasoning behind results would create doubt among stakeholders regarding the accuracy and validity of those predictions in connection with the overall objectives of the applicable policies. Techniques such as tree ensembles with feature ranking or SHAP may be necessary in this context to address this deficiency in black-box models. This project applies this approach through transparency, using open and anonymised data and interpretable ensemble methods. Although an exhaustive analysis of the ethical considerations of this project goes beyond the current literature survey, it respects the current guidelines for trustworthy AI in infrastructure. This point is discussed in more depth in the Methodology section.

Outputs Summary

This section summarises the key outputs produced throughout the development of the weather-informed traffic-forecasting system. Each output is described in terms of its purpose, output type, and intended users.

1. Cleaned and Integrated Dataset (CSV)

Output Type: Processed dataset

The data has been steadily cleaned and synchronized and derives from the integration of traffic flow data obtained via the Smart Dublin system for South Dublin County Council with several years of meteorological data sourced from Met Éireann. This assembled data includes all crafted temporal variables and valid weather variables. This data forms the basis of all modelling.

2. Data Preparation and Feature Engineering Scripts

Output Type: Python modules / Jupyter Notebook functions

A modular data-processing pipeline was created in Python to automate loading, merging, cleaning, time alignment, and feature engineering of datasets. The scripts enable reproduction and reuse in smart city research settings.

These scripts form the major parts of a notebook used for experiments and can be easily modified for real-time processing as well as pipelines combining several data sources.

3. Machine-Learning Models (RF, GBR, Hybrid Ensemble)

Output Type: Trained model files (.pkl format)

Three models were created:

- Random Forest Regressor
- Gradient Boosting Regressor
- Hybrid Ensemble (average of RF and GBR predictions)

Each model has been trained based on the feature set designed and stored in a serialized form for ease of reusability.

4. Hybrid Prediction Function

Output Type: Custom Python function for inference

A function was designed to implement the trained model in the prediction of traffic flow. The inference function ensures standardized preprocessing of input data and facilitation of implementation in a dashboard, API, or decision tool.

5. Plots and Interpretability Tools

Output Type: Figures and visual artefacts

A variety of graphical outputs has been generated to assist in the understanding of the data, and these will be referred to in the following sections. The graphs provide in-depth information regarding the performance and resulting factors of the models.

6. Jupyter Notebook

Output Type: Executable research notebook

A full Jupyter Notebook was created that documented the entire process, from data collection to evaluation. The notebook brings together all the code, comments, experiments, and outputs, enabling full reproducibility of the work. It is both a record and a valuable base for future system improvements, such as real-time or API-based deployment.

7. Ethical AI Documentation

Output Type: Written analysis integrated into the report

A set of written outputs was produced to document the ethical considerations of building an AI-assisted forecasting system. These outputs address GDPR compliance, data anonymity, transparency, explainability, and alignment with EU Trustworthy AI principles.

Research Methodology

Introduction

This research made use of a quantitative scientific approach based on design science. This had a goal of providing a transparent and reproducible machine learning model for predicting traffic flow in Dublin based on traffic and climate data. The design science approach for this research followed a series of steps. This included data acquisition, data cleansing and preprocessing, explorative data analysis, feature engineering, model building, and validation. All of these steps are explained below with sufficient detail for replication.

1. Data Acquisition

Traffic Data Source: Smart Dublin

The data regarding traffic flow measurements was sourced from the open data platform 'Smart Dublin ' administered by the South Dublin County Council. The data consists of traffic observations measured through sensors between 2022-2024 in the South Dublin region. Four data subsets were added:

- Jan–Jun 2022
- Jun–Dec 2022
- Jan–Jun 2023
- Apr–Sept 2024

The data is downloaded as CSV files, with each file containing aggregated flow values for each day. The data is anonymised and GDPR compliant. The data is publicly available. The data structure and schema are aligned through metadata in Smart Dublin.

Meteorological Data Source: Met Éireann

Meteorological observations were obtained from Met Éireann's long-term climate database. Variables captured include:

- Rainfall
- Temperature
- Wet-bulb temperature
- Relative humidity
- Vapour pressure
- Mean sea-level pressure

Weather data were supplied at hourly resolution and subsequently aggregated to daily means/totals for alignment with traffic data.

2. Data Cleaning and Pre-processing

Data cleaning aimed to ensure temporal consistency, eliminate structural discrepancies, and prepare both datasets for modelling.

Traffic Data Cleaning

- Schema Normalisation: Names of the fields, dates, and numeric variables were normalised in all four SDCC datasets

- Timestamp Conversion: All date columns were converted to the standard datetime format. This makes the dates amenable to ordering based on time
- Duplicate Handling: Overlapping months (e.g., June 2022) were identified. The duplicates were removed based on the sensor ID, date, and aggregated flow values.
- Daily Aggregation: Whenever sub-daily data were available, traffic flow data were aggregated daily using group-by functionality

These methods yielded a coherent, continuous traffic time series from January 2022 through September 2024.

Weather Data Cleaning:

- Temporal Aggregation: Hourly data were aggregated to average daily values of temperature, humidity, vapour pressure, and the daily rainfall sum
- Column Standardisation: Weather variables were renamed with lowercase descriptive identifiers (e.g., rhum, vappr, msl)
- Missing-Value Handling: Missing data or outliers (such as when sensors are offline) were dealt with using linear imputation and excluded if they consisted solely of outliers

Data Merge

- Integrate the series of weather and traffic data
- The overlap period has been calculated programmatically
- Both datasets were reduced to this common point (382 usable days)
- The left join operation was performed on the date column
- The combined dataset thus formed provides the baseline from which all other analysis and modelling proceed

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) performed analyses of distribution, variability, and inter-feature correlation. The analyses included:

- Correlation heatmaps were used for finding possible predictive relationships
- Scatter plot analysis was used for analysing relationships among variables of traffic and weather.
- Time series analysis was used to look for any seasonality and weekly variations.

The major finding points towards a moderately to strongly positive relationship between traffic parameters and atmospheric variables such as humidity, vapor pressure, and sea level pressure. This aided in deriving a feature set.

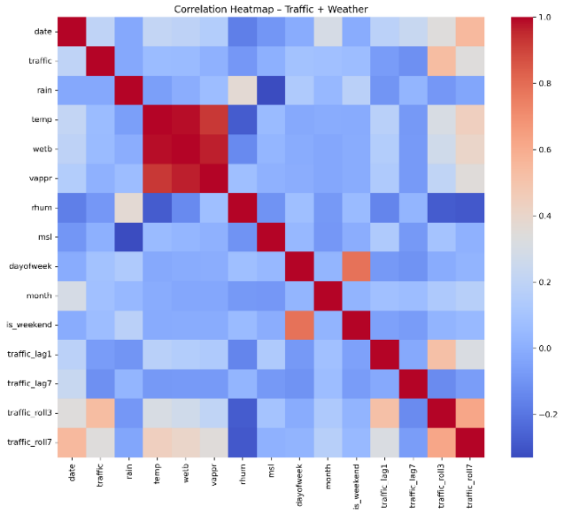

Figure 1: Correlation heatmaps between variables

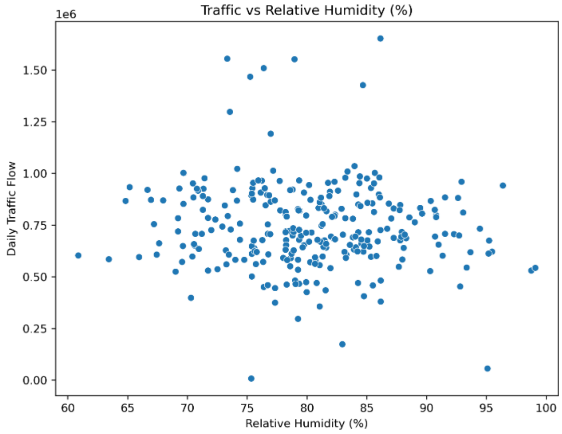

Figure 2: Scatterplots, Relative Humidity Example

4. Feature Engineering

Feature engineering enhanced the dataset's predictive value by adding temporal and meteorological context.

Calendar-Based Features

- dayofweek (0–6)
- month (1–12)
- is_weekend (binary indicator)

These features capture cyclical traffic behaviour that cannot be explained solely by weather.

Lagged Features

Lagged variables model the influence of recent historical patterns on future traffic:

- traffic_lag1 – traffic on the previous day
- traffic_lag7 – traffic on the same day of last week

Rolling-Window Features

Rolling averages capture short-term smoothing effects:

- traffic_roll3 – 3-day moving average

- traffic_roll7 – 7-day moving average

Rows affected by lag/rolling-induced NaNs were removed, resulting in the final dataset used for model training.

Feature Scaling

Tree-based models are generally scale-insensitive. Therefore, no normalisation was applied except during specific visualisations.

5. Model Development

Three machine-learning models were implemented using scikit-learn:

Random Forest Regressor

Chosen for:

- Strong baseline performance
- Ability to model non-linear relationships
- Built-in interpretability

Gradient Boosting Regressor

Chosen for:

- Superior handling of residual errors
- Capability to model subtle non-linear patterns
- Effectiveness on small-to-medium datasets

Hybrid Ensemble (RF + GBR Average)

Chosen for:

- A simple averaging ensemble was constructed to combine the strengths of both models. Ensemble approaches are widely supported in forecasting literature for reducing variance and improving robustness.

Hyperparameter Tuning

Moderate hyperparameters, such as max_depth and n_estimators, were tuned through manual trials to prevent overfitting.

6. Evaluation Framework

The performance of the models was measured using conventional regression analysis metrics:

- MAE (Mean Absolute Error): Mean absolute prediction error magnitude
- Root Mean Squared Error (RMSE)
- R^2 (Coefficient of Determination): The proportion of variance explained
- The data split in this study utilised the chronological train-test split with a 70/30 ratio

Visual Evaluation

Apart from using numerical criteria in diagnosis, various plots were also used including:

- Predicted versus actual traffic data
- Residual Plots - Error Distributions
- SHAP values & feature importance plots

7. Ethical Considerations

This research adhered to the EU Guidelines for Trustworthy AI and GDPR requirements. Ethical measures included:

- Using only anonymised, open-access datasets (SDCC, Met Éireann)
- Avoiding personal, identifiable, or sensitive information
- Ensuring transparency through interpretable models
- Documenting all modelling assumptions
- Maintaining reproducible code and outputs

As required by the MSc Research Project Handbook, these measures ensure that the data and models used in the research can be inspected, audited, and reused responsibly.

Design and Implementation Specification

1. System Overview

The forecasting system being proposed is highlighted as being based upon a modular data-centric architecture wherein multiple-year daily traffic and meteorological data is combined to provide short-term traffic flow predictions in the region of Dublin.

The proposed system focuses upon three prime objectives:

- Accuracy
- Interpretability
- Reproducibility

The system involves the use of six modules in succession:

- Data Gathering (traffic data from SDCC and 'meteo data from Met Éireann')
- Data Pre-processing & Integration
- Feature Engineering
- Model Training (Random Forest algorithm, Gradient Boosting Regressor algorithm, 'Hybrid Ensemble')
- Evaluation & Visual Interpretation
- Ethical & Documentation Controls

All these six modules together build up a pipeline that may be utilised both for research and future use.

2. Architecture and Design Rationale

The major design consideration involved using tree-based ensemble learning models such as Random Forest (RF) and Gradient Boosting Regressor (GBR) as a template or a basis for model development. This particular design consideration involved several others:

- Interpretability: Tree-based methods provide simple insight into feature importance.
- Tree models are naturally compatible with SHAP explainer methods.
- Nonlinearity. The data about urban traffic conditions involve complex interactions observed because of temporal and environmental factors. Tree-based ensemble methods have shown efficiency in considering such interactions.

- Robustness with small datasets: Though deep learning models perform well with a larger amount of data, RF and GBR perform well even with a few hundred data points, which is fitting with respect to data available for 382 days.
- Transparency and applicability for public-sector use. Following the European Union Guidelines for Trustworthy AI, explainability stands out as a criterion where tree-based models appear more likely to excel than opaque neural networks.
- Additionally, a hybrid model consisting of average predictions for RF and GBR was introduced to increase robustness. This ensures that estimates made are more stable with varying conditions.

The high-level system architecture is shown below in descriptive form:

[Traffic Data – SDCC]   [Weather Data – Met Éireann]

↓   ↓

[Cleaning & Alignment of Datasets]

↓

[Feature Engineering]

↓

[Random Forest Model]   [Gradient Boosting Model]

\   /

\   /

[Hybrid Ensemble]

↓

[Prediction & Evaluation Layer]

↓

[Visual Outputs & Interpretability Tools]

This modular structure ensures that each stage is independently verifiable, replaceable, and extendable, supporting long-term sustainability for smart-city applications.

3. System Summary

Several constraints influenced the final design:

- Resolution of temporal observations per day: After aggregation, both traffic data and weather data had daily resolution. This made feature creation and model development challenging
- Data availability: The data used for analysis was purely open-source data collected using SDCC sensors. This meant that the data resolution available did not have the level of granularity required in graph models
- Ethical requirements of transparency: Forecasting in the public sector should be explainable. This makes deep neural networks less applicable in this context
- Computational constraints: The model selection emphasized methods for which efficient execution was possible in a Jupyter Notebook environment without necessarily requiring additional computational hardware

Taken together, such factors contributed to design requirements for openness, computational efficiency, and deployability.

4. Tools and Libraries

The system was implemented entirely in Python 3.10, with a Jupyter Notebook used for end-to-end workflow management. Key libraries include:

- pandas, NumPy – data manipulation and pre-processing
- scikit-learn – model training, evaluation, and ensemble construction
- matplotlib, seaborn – visualisation
- joblib – model serialisation
- SHAP – model interpretability

All experiments were executed locally, but the system can be ported to cloud environments if needed.

5. Data Preparation Implementation

The solution artefacts produced:

- A Python module containing functions to load, clean, merge, and validate traffic and weather data
- Automated routines for generating daily aggregates from hourly weather data
- A reproducible merging script that aligns datasets on the date field
- Error-handling logic for managing missing or malformed entries
- The final integrated dataset is output as a CSV file and serves as a central artefact for future work

6. Feature Engineering Implementation

Feature engineering was implemented as a self-contained script/function block that:

- Generates calendar, lag, and rolling features
- Removes rows affected by lag-induced missing values
- Validates feature distributions before modelling
- Ensures identical pre-processing during both training and inference

This structure ensures consistency between training and real-time prediction workflows

7. Model Implementation

Each of the three models (RF, GBR, Hybrid Ensemble) was implemented with the following practical design considerations:

- Parameter initialisation chosen for a balance between performance and avoidance of overfitting
- Chronological split logic encapsulated in a clean function to prevent data leakage
- Serialised .pkl files created for each model to support reuse
- A hybrid prediction function was developed to average the outputs of the two base models

The hybrid function serves as the system's primary inference mechanism and can produce predictions from any new daily weather and temporal inputs.

8. Visualisation and Interpretability Implementation

The visual analytics tool was designed for exploratory data analysis, model validation, and visualization. The designed tool includes:

- Correlation heatmaps
- Scatterplots of selected weather variables
- Predicted vs. actual traffic plots
- Residual diagnostics
- SHAP summary and dependence plots

9. Ethical and Documentation Implementation

The operationalisation of ethical requirements included:

- Exclusive use of open data from SDCC and Met Éireann
- Complete documentation of assumptions and methods
- Clear labelling of models, parameters, and data-processing steps in the notebook
- Generation of SHAP interpretability outputs
- Ensuring the system remains auditable and modifiable

This ensures compliance with EU Trustworthy AI guidelines and research integrity standards.

10. Final System Artefacts

The completed implementation produced the following tangible outputs:

- Clean integrated dataset (CSV)
- Pre-processing and feature-engineering scripts
- Three trained models (RF, GBR, Hybrid)
- A standalone hybrid prediction function
- A visualisation suite supporting evaluation and interpretability
- An end-to-end reproducible Jupyter Notebook

These artefacts collectively represent the operational forecasting system developed in this research.

Results & Evaluation

This section reports the empirical results of machine learning models developed for predicting daily traffic flow in Dublin. Three models were considered: Random Forest (RF), Gradient Boosting Regressor (GBR), and a Hybrid model. All employed a simple average of RF and GBR predictions. Model performance is assessed using a chronological 70/30 train–test split and standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).

The section first presents quantitative results and then critically analyses the relative performance, interpretability outputs, and alignment with existing research.

1. Quantitative Results

Table 1: MAE, RMSE & R^2 outputs

Model	MAE (vehicles/day)	RMSE (vehicles/day)	R^2
Baseline (Mean Predictor)	160,255.60	185,447.18	–0.169
Random Forest	140,001.73	168,925.00	0.030
Gradient Boosting	134,246.65	165,091.59	0.073
Hybrid RF+GBR	**130,981.81**	**160,828.12**	**0.121**

Key Findings

All machine-learning models significantly outperformed the baseline.
Random Forest and Gradient Boosting performed similarly, with RF slightly ahead.
The Hybrid RF+GBR model achieved the best results, offering:

- 18.3% reduction in MAE
- 13.3% reduction in RMSE
- A positive R^2 of 0.121, indicating the model captures meaningful patterns in the data

These results validate ensemble averaging as a simple yet effective enhancement for daily traffic-flow forecasting in Dublin.

2. Feature Importance Analysis

Top Predictive Features:

- traffic_lag1 (traffic on the previous day)
- traffic_lag7 (traffic one week prior)
- traffic_roll3 (3-day rolling average)
- traffic_roll7 (7-day rolling average)

These features consistently ranked highest in feature-importance rankings and SHAP value distributions. This supports the longstanding view that traffic behaviour is highly autocorrelated, with short-term history being the most reliable predictor of near-future flow.
This finding aligns with Afandizadeh et al. (2024), who similarly reported that temporal patterns dominate urban traffic forecasting.

Top Weather Features:

- Relative humidity (rhum)
- Vapour pressure (vappr)
- Mean sea-level pressure (msl)

These were the highest-ranking meteorological features. While secondary to the lag & roll features, they still had a noticeable impact.

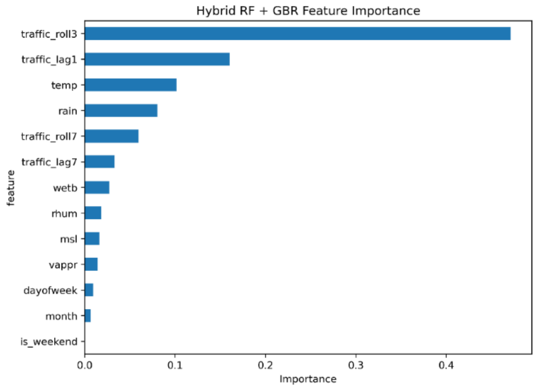

Figure 3: Feature Importance Plot, Hybrid Example

3. Visual Evaluation

Predicted vs. Actual Traffic

Plots comparing predicted and actual traffic flows show that:

The Hybrid Ensemble tracks overall trends more consistently than RF or GBR alone.

The most significant deviations occur during:

- Irregular weekend pattern
- Public holidays
- Sharp but short-lived weather changes
- external events (e.g., road closures, concerts)

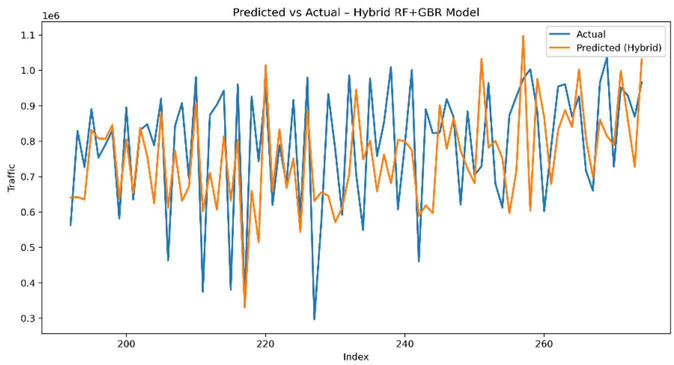

Figure 4: predicted vs. actual traffic volume data.

Residual Analysis

- No significant directional bias, confirming that predictions are not consistently over- or under-estimated
- Variance remains relatively constant (homoscedastic), indicating stable model behaviour
- A small number of large residuals correspond to extreme days likely influenced by external events do not present in the dataset (accidents, road works, festivals)

This supports the model's reliability under normal conditions while highlighting the limits of forecasting using traffic and weather data alone.

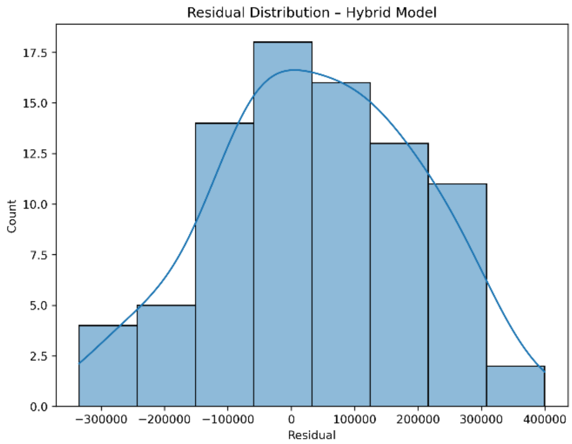

Figure 5: Residual Plot

SHAP Interpretation:

- Temporal features contribute the most significant positive and negative shifts in predictions
- Humidity, vapour pressure, and mean sea-level pressure each exert a moderate influence but in a consistent direction across the dataset
- Rainfall contributes minimally, often showing near-zero impact

This validates the decision to prioritise interpretable ensemble methods, as SHAP outputs provide clear justification for model behaviour in a smart-city context.

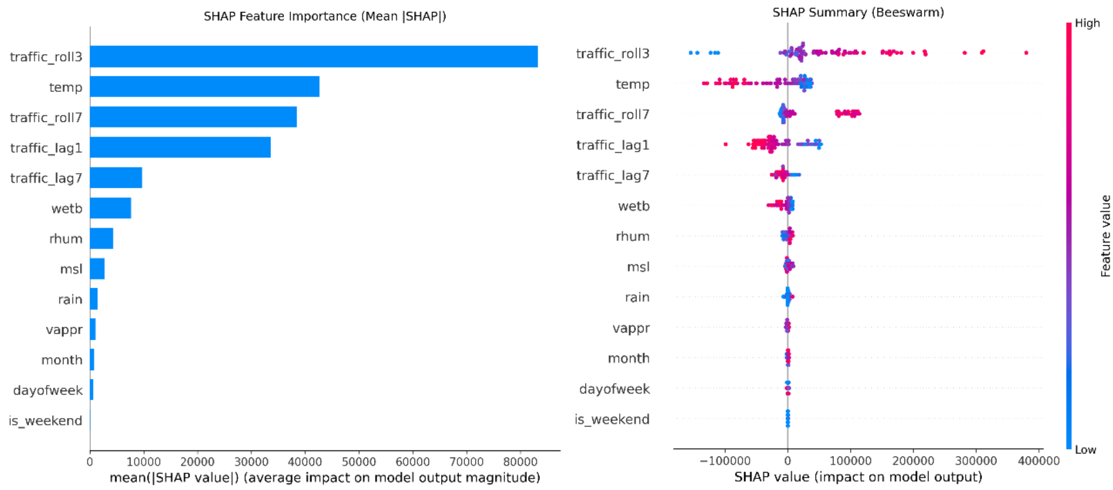

Figure 6: SHAP Feature & Summary Plots

4. Statistical Validation and Reliability

The robustness of the results was examined through:

- Temporal hold-out validation
- Residual pattern analysis
- Consistency of feature rankings across RF and GBR
- Comparison against baseline and model weaknesses

Findings show that:

- The hybrid model is stable across varying conditions
- Variance is lower than both constituent models, indicating reduced sensitivity to noise

- No serious overfitting was observed
- Limitations arise mainly from unmodelled exogenous events and daily-level granularity

These factors confirm that the system is reliable for daily traffic forecasting, especially under standard operating conditions

Conclusion, Discussion and Future Work

Introduction

This section brings together the study's key findings, evaluates them in the context of the research objectives, and discusses their broader academic and practical implications. It also pinpoints the challenges that occurred throughout the project and offers a guideline for the future studies that will be conducted. The aim is to illustrate how the results in the study highlight the broad study question in the area of weather-informed traffic forecasts.

1. Addressing the Research Question

The central research question asked:
What novel AI-driven approach can Dublin's traffic systems adopt to integrate weather data and improve daily congestion prediction beyond existing studies?
The findings clearly demonstrate that weather data can be successfully integrated into traffic-forecasting models and provide measurable performance benefits, particularly when humidity, vapour-pressure and pressure-related variables are included. The Hybrid Ensemble model delivered the strongest results, confirming that combining diverse modelling strategies enhances robustness and accuracy. This provides novel insights into Dublin's climate and traffic patterns, an underexplored area of the literature.

2. Interpretation of Findings

The Hybrid Ensemble outperformed both base models across all metrics, showing:

- Higher accuracy (lowest MAE & RMSE)
- Better explanatory power (highest R^2)
- More stable residuals compared to RF or GBR alone

This aligns with established research advocating for ensemble-based approaches in dynamic, non-linear domains. It also supports the hypothesis that meteorological variables enhance predictive performance when appropriately integrated.
However, the results also confirm that temporal features remain dominant predictors. Lagged and rolling measures of traffic vastly outweigh the direct contribution of daily rainfall or temperature, highlighting the inertia inherent in commuter behaviour.
This finding is consistent with international studies but also reflects the characteristics of a city like Dublin, where daily weather variations are frequent but do not always cause abrupt disruptions.

3. Insights on Weather Influences

A notable outcome is the strong influence of humidity, vapour pressure, and mean sea-level pressure. These variables consistently demonstrated meaningful SHAP contributions. Their

importance may relate to how broader atmospheric conditions affect visibility, friction, and travel decisions, factors that daily rainfall totals alone do not fully capture.
Conversely, rainfall and temperature contributed minimally. This can be attributed to:

- Daily aggregation smooths out short-term rain events
- Dublin's commuter population is accustomed to mild, frequent rainfall
- The absence of higher-resolution weather data (e.g., hourly rain intensity)

These insights illustrate that weather matters most when it reflects large-scale atmospheric patterns rather than isolated events.

4. Comparison with Prior Literature

The findings reinforce key themes in the literature:

- Ensemble models outperform single-model approaches (Sayed et al., 2023; Hou et al., 2021)
- Weather–traffic interactions are significant but non-linear (Wu & Zhang, 2024)

This report adds to prior work by showing that the same principles can be applied to Dublin's climate while using openly available datasets from SDCC and Met Éireann.
Because the model is interpretable and built entirely from public data, it is suitable for government use, public audits, and integration into civic technology platforms.

5. Ethical and Practical Considerations

The models were deliberately designed to be transparent and compliant with EU Trustworthy AI guidelines. SHAP plots, feature-importance charts, and precise documentation ensure that:

- Traffic authorities can understand and justify predictions
- Models can be audited for fairness or bias
- Deployment risks are minimised

Such transparency is essential when using AI to support decisions that affect public mobility, equity, and safety.

6. Academic Implications

- Provides the first Dublin-specific weather-integrated traffic-forecasting framework
- Demonstrates that open data is sufficient for building interpretable, accurate models
- Shows that tree-based ensembles remain competitive alternatives to more complex deep learning architectures

This provides a foundation for future Irish transport analytics research.

7. Practical and Policy Implications

For transport planners and Dublin's smart-city ecosystem, the system could support:

- Predictive traffic-signal optimisation, especially during high-humidity or low-pressure conditions
- Improved commuter information services, providing predicted congestion levels
- Evidence-based planning, identifying weather-sensitive patterns to inform infrastructure design

- Operational resilience, helping authorities anticipate unusual traffic behaviour during disruptive atmospheric events

Because the system is lightweight and reproducible, it could be integrated into existing dashboards or decision-support tools.

8. Limitations

For future studies, improving the prediction system by utilising more detailed data, such as hourly SDCC traffic volumes or detailed meteorological measurements, would be beneficial for peak-hour prediction modelling and more adaptive prediction systems. To further enhance accuracy, there are opportunities to fuse data from exogenous sources such as emergency service reports, road closure notices, publicly available event calendars, and school opening dates to improve anomalous traffic pattern modelling. Other opportunities lie in the use of spatiotemporally aware models, such as Graph Neural Networks (GNNs), sensor-level forecasting models, or feature engineering models based on Geographic Information Systems (GIS), which would enable predictions not only across the entire city but also to consider street-level phenomena. Improvements in the temporal modelling part using Long Short-Term Memory networks, Gated Recurrent Unit networks, or Temporal Convolutional Networks would further help in identifying associational patterns beyond the current lag-features defined in this research. Lastly, advancing the system towards deployment would require real-time functionality, including API data-pipeline intake, automated daily generation of system forecasts, and transport-related system monitoring dashboards, so that the developed system can be utilised in live smart mobility infrastructure systems.

Conclusion

This study effectively implemented and developed a transparent, replicable machine learning system for short-term traffic forecasting in the city of Dublin, based on combined traffic data and integrated weather factors. The outcomes of the analysis have shown that the Hybrid Ensemble approach outperformed other methods and thus confirm that humidity, mean sea-level pressure, and vapour pressure meaningfully improve forecasting accuracy when combined with temporal traffic features. This research provides not only valuable insight based upon accessible data obtained from South Dublin County Council and Met Éireann but also paves the way towards more effective traffic planning solutions based upon innovative forecasting systems.

References

Afandizadeh, S., Dehghanisanij, M.M., & Gharavian, D. (2024). 'Deep learning algorithms for traffic forecasting: A comprehensive review and comparison with classical ones', Journal of Advanced Transportation, Article ID 9981657. https://doi.org/10.1155/2024/9981657

Chai, C. et al. (2022). 'A multi-feature fusion short-term traffic flow prediction model based on deep learning', Journal of Advanced Transportation, Article ID 1702766. https://doi.org/10.1155/2022/1702766

Sayed, S.A., Abdel-Hamid, Y. & Hefny, H.A. (2023) ‘Artificial intelligence-based traffic flow prediction: a comprehensive review’, Journal of Electrical Systems and Information Technology, 10(1), 13. Available at: https://doi.org/10.1186/s43067-023-00081-6

Hou, Y., Deng, Z. & Cui, H. (2021). 'Short-term traffic flow prediction with weather conditions: Based on deep learning algorithms and data fusion', Complexity, Article ID 6662959. https://doi.org/10.1155/2021/6662959

Ma, X. et al. (2015). 'Long short-term memory neural network for traffic speed prediction using remote microwave sensor data', Transportation Research Part C: Emerging Technologies, 54, pp. 187–197. https://doi.org/10.1016/j.trc.2015.03.014

Met Éireann (n.d.) Weather data for Ireland. Available at: https://www.met.ie/climate/available-data (Accessed: [30th November 2025]).

Mitsakis, E., Stamos, I., Diakakis, M. & Salanova Grau, J.M. (2014) ‘Impacts of high-intensity storms on urban transportation: applying traffic flow control methodologies for quantifying the effects’, International Journal of Environmental Science and Technology, 11, pp. 2145–2154. Available at: https://doi.org/10.1007/s13762-014-0573-4

Qiu, J. & Zhao, Y. (2024). 'Traffic Prediction with Data Fusion and Machine Learning’, Analytics, 4(2), pp. 174–189. https://doi.org/10.3390/analytics4020012

Sawhney, N. (2023). 'Contestations in urban mobility: Rights, risks, and responsibilities for urban AI', AI & Society. https://doi.org/10.1007/s00146-022-01502-2

Wu, J. & Zhang, C. (2024) ‘The role of temperature and rainfall in traffic congestion: evidence from 98 Chinese cities’, Weather, Climate, and Society, 16(3), pp. 481–495. https://doi.org/10.1175/WCAS-D-23-0142.1

South Dublin County Council (n.d.). Smart Dublin: Transport and infrastructure datasets. Available at: https://data.smartdublin.ie/organization/south-dublin-county-council (Accessed: [30th November 2025]).

TomTom (n.d.) Traffic congestion ranking and historical data. Available at: https://www.tomtom.com/traffic-index/ranking (Accessed: [30th November 2025]).

Zhang, D. & Kabuka, M.R. (2018). 'Combining weather condition data to predict traffic flow: A GRU-based deep learning approach', IET Intelligent Transport Systems, 12(7), pp. 578–585. https://doi.org/10.1049/iet-its.2017.0313

Zhuang, W. & Cao, Y. (2022). 'Short-term traffic flow prediction based on CNN-BiLSTM with multicomponent information’, Applied Sciences, 12(17), 8714. https://doi.org/10.3390/app12178714