

**Investigating a Deep Learning Approach To
Real-Time Air Quality Prediction and Visualisation On
UK Highways**

By

Taofeek Dolapo Akinosho

18970269



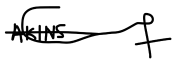
A thesis submitted in partial fulfilment of the requirements of the
University of the West of England, Bristol
for the degree of Doctor of Philosophy

June 2024

DECLARATION

I affirm that this research is my own and was conducted by me, excluding where due acknowledgement has been made in the text, and that it has not been submitted either in part or full for any other award than the degree of Doctor of Philosophy of the University of the West of England. Materials from other sources have been duly acknowledged and referenced in line with ethical standards, and the list of publications made from the thesis has been provided.

Signed: TAOFEEK AKINOSHO

Signature


Date 01/06/2024

COLLABORATION STATEMENT

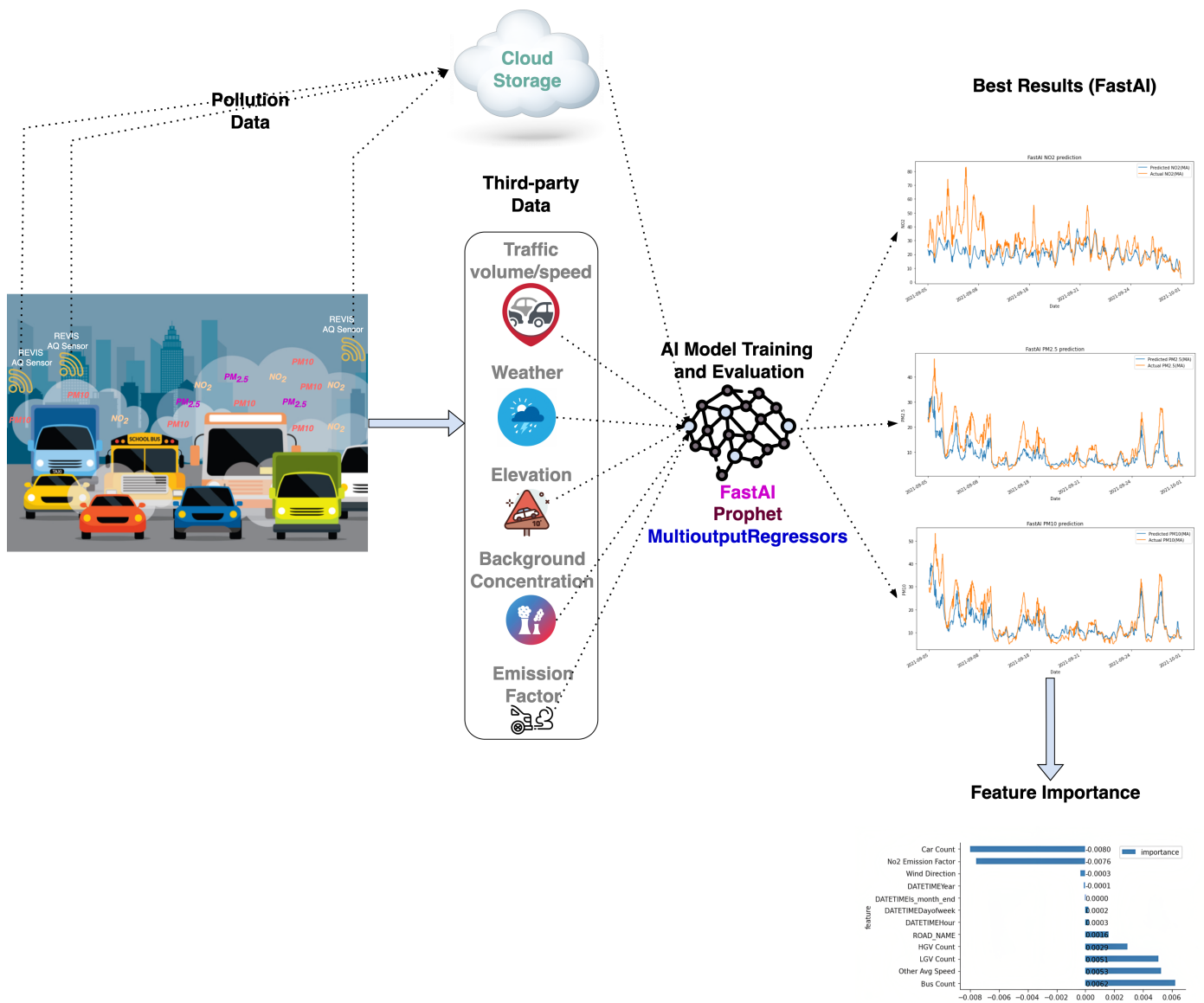
This research represents a collaborative effort involving two key industry partners, Costain PLC and TerOpta Ltd. TerOpta Ltd played a pivotal role by developing customised air quality sensors, essential for collecting real-time data on the specific pollutants central to this study. These bespoke sensing devices were crucial due to the limitations of commercially available air quality sensors. In parallel, Costain PLC facilitated this research by granting access to the designated case study highways, which served as the critical testing grounds for our experiments. While these collaborations were instrumental in the success of this project, it is important to emphasise that the author of this thesis, Taofeek Akinosho, remains the principal researcher behind the remainder of the research work. The entire body of work presented in this thesis is a direct product of his dedication, research findings, and endeavours.

ABSTRACT

The construction of intercity highways by the United Kingdom (UK) government has resulted in a progressive increase in vehicle emissions and pollution from noise, dust, and vibrations amid growing concerns about air pollution. Existing roadside pollution monitoring devices have faced limitations due to their fixed locations, limited sensitivity, and inability to capture the full spatial variability, which can result in less accurate measurements of transient and fine-scale pollutants like nitrogen oxides and particulate matter. Reports on regional highways across the country are based on a limited number of fixed monitoring stations that are sometimes located far from the highway. These periodic and coarse-grained measurements cause inefficient highway air quality reporting, leading to inaccurate air quality forecasts. Multi-target neural network is a type of machine learning algorithm that offers the advantage of simultaneously predicting multiple pollutants, enhancing predictive accuracy and efficiency by capturing complex interdependencies among various air quality parameters. The potentials of this and similar multi-target prediction techniques are yet to be fully exploited in the air quality space due to the unavailability of the right data set. To address these limitations, this doctoral thesis proposes and implements a framework which adopts cutting-edge digital technologies such as Internet of Things, Big Data and Deep Learning for a more efficient way of capturing and forecasting traffic related air pollution (TRAP). The empirical component of the study involves a detailed comparative analysis of advanced predictive models, incorporating an enriched dataset that includes road elevation, vehicle emission factors, and background maps, alongside traditional traffic flow, weather, and pollution data. The research adopts a multi-target regression approach to forecast concentrations of NO_2 , $PM_{2.5}$, and PM_{10} across multiple time steps. Various models were tested, with Fastai's tabular model, Prophet's time-series model, and scikit-learn's multioutput regressor being central to the experimentation. The Fastai model demonstrated superior performance, evidenced by its Root-Mean Square Error (RMSE) scores for each pollutant. Statistical analysis using the

Friedman and Wilcoxon tests confirmed the Fastai model's significance, further supported by an algorithmic audit that identified key features contributing to the model's predictive power. This doctoral thesis not only advances the methodology for air quality monitoring and forecasting along highways but also lays the groundwork for future research aimed at refining air quality assessment practices and enhancing environmental health standards.

GRAPHICAL ABSTRACT



DEDICATION

In memory of my late mother, whose passion for academic excellence continue to inspire me.
This dissertation is a tribute to your lasting influence.

ACKNOWLEDGEMENTS

I want to express my thanks to Almighty Allah for guiding me on this academic journey. I also want to extend my heartfelt thanks to the people who played a vital role in my success.

First, I am deeply thankful to my Dad - Alh. Idris Akinosho and my siblings - Khadijah Akinosho, Aisha Akinosho, and Idris Akinosho. Their unwavering support, love, and belief in my abilities have kept me motivated. Their faith in my goals has been the cornerstone of my determination, and I consider myself fortunate to have them as my family.

To my dear wife - Sahra Mohamed and son - Farhan Akinosho, your patience, understanding, and encouragement have been my driving force in pursuing my educational goals. Your unwavering belief in me, even during the most challenging times, has given me the strength to persevere. Your love and support have made every sacrifice worthwhile, and I look forward to a future filled with shared achievements. You are my inspiration.

I would also like to express my sincere appreciation to Professor Lukumon Oyedele, whose inspiration encouraged me to begin this research. My heartfelt thanks go to my Director of Studies, Professor Muhammad Bilal, and my supervisors, Professor Enda Hayes and Associate Professor Anuoluwapo Ajayi. Their guidance, expertise, and encouragement have been pivotal in shaping my academic journey. Their support and insightful feedback have been invaluable, and I feel fortunate to have had such mentors. Also to the other Professors in the Big Data Enterprise and Artificial Intelligence lab who have contributed in one way or another to the success of this journey, I am truly grateful for all your support and guidance.

To my PhD colleagues and friends - Quayyum Gbadamosi, Habeeb Kusimo, Dimeji Olawale, Adekansola Labo-Popoola, Sofiat Abioye, Ummulyumn Tijani, Sameen Arshad, Rasheed Ojo, Kabir Kadiri, and Naimah Yakubu, your camaraderie, collaborative spirit, and shared enthusiasm for discovery made the lab an inspiring place to work. Our discussions and shared insights enriched my research.

Finally, I would like to extend my gratitude to the Director of Doctoral Research, College of Business and Law - Dr. Pawel Capik, and the UWE graduate school, especially Samantha Watts. Your assistance and support throughout this journey have been instrumental in my academic progress, and I'm thankful for your contributions to my success.

The knowledge, experiences, and friendships I've gained throughout this process are treasures that I will carry with me as I move forward in my academic and professional life. Thank you for being a part of this remarkable journey.

PUBLICATIONS

1. **Akinosho, Taofeek Dolapo**, Muhammad Bilal, Enda Thomas Hayes, Anuoluwapo Ajayi, Ashraf Ahmed, and Zaheer Khan. Deep learning-based multi-target regression for traffic-related air pollution forecasting. *Machine Learning with Applications*, page 100474, 2023b. URL <https://doi.org/10.1016/j.mlwa.2023.100474> - **In Thesis**
2. **Akinosho, Taofeek Dolapo**, Muhammad Bilal, Enda Thomas Hayes, and Anuoluwapo Ajayi. Algorithmic audit of a deep learning system for air pollution based journey planning: Towards healthier travelling on uk highways. Under Review, 2024 - **In Thesis**
3. **Akinosho, Taofeek D**, Lukumon O Oyedele, Muhammad Bilal, Ari Y Barrera-Animas, Abdul-Quayyum Gbadamosi, and Oladimeji A Olawale. A scalable deep learning system for monitoring and forecasting pollutant concentration levels on uk highways. *Eco- logical Informatics*, 69:101609, 2022. URL <https://doi.org/10.1016/j.ecoinf.2022.101609> - **In Thesis**
4. **Akinosho, Taofeek D**, Lukumon O Oyedele, Muhammad Bilal, Anuoluwapo O Ajayi, Manuel Davila Delgado, Olugbenga O Akinade, and Ashraf A Ahmed. Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, 32:101827, 2020. URL <https://doi.org/10.1016/j.jobeb.2020.101827> - **In Thesis**
5. Ari Yair Barrera-Animas, Lukumon O Oyedele, Muhammad Bilal, **Akinosho, Taofeek Dolapo**, Juan Manuel Davila Delgado, and Lukman Adewale Akanbi. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, 7:100204, 2022. URL <https://doi.org/10.1016/j.ecoinf.2022.101609>

Contents

Abstract	3
Graphical Abstract	5
Acknowledgements	7
1 Introduction	26
1.1 Background	26
1.1.1 Urban Highway Pollution And Prevalent Pollutants	29
1.1.2 Supporting Technologies For Highway Air Quality Monitoring and Forecasting	31
1.1.3 Concepts Of Scalability And Accuracy In Highway Air Qual- ity Forecasting	32
1.2 Research Problem	33
1.2.1 Coarse-Grained vs Fine-Grained Air Quality Measurements	34
1.2.2 Real-Time Air Quality Monitoring	35
1.2.3 Prediction Accuracy Of TRAP Forecasting Models	35
1.3 Justification For Study	37
1.4 Research Novelty	38
1.5 Research Questions	39
1.6 Research Aim and Objectives	39
1.7 Research Methodology And Workflow	40
1.8 Research Scope and Limitations	41
1.9 Expected Contribution To Knowledge	43

1.10 Thesis Layout	45
2 Review of Enabling Digital Technologies For TRAP Monitoring and Forecasting	47
2.1 Chapter Overview	47
2.2 Traffic-Related Air Pollution Monitoring - Enabling Technologies and Methods	48
2.2.1 Internet of Things And TRAP Monitoring	48
2.2.2 Existing Applications of IoT For TRAP Monitoring	49
2.2.3 Barriers To The Adoption of IoT For TRAP Monitoring	51
2.2.4 Big Data And TRAP Monitoring	53
2.2.5 Existing Applications of Big Data For TRAP Monitoring	54
2.2.6 Barriers To The Adoption of Big Data For TRAP Monitoring	56
2.3 Traffic-Related Air Pollution Forecasting - Enabling Technologies and Methods	58
2.3.1 Chemical Transport Models And TRAP Forecasting	58
2.3.2 Existing Applications Of Chemical Transport Models For TRAP Forecasting	59
2.3.3 Barriers To The Adoption of CTM Tools For TRAP Forecasting	61
2.3.4 Deep Learning And TRAP Forecasting	63
2.3.5 Deep Learning Architectures	63
2.3.5.1 Deep Neural Network (DNN)	64
2.3.5.2 Convolutional Neural Network (CNN)	64
2.3.5.3 Recurrent Neural Network (RNN)	65
2.3.5.4 Auto-Encoder (AE)	66
2.3.6 Existing Applications of Machine Learning Methods For TRAP Forecasting	67

2.3.7	Barriers To The Adoption of Machine Learning Methods For TRAP Forecasting	69
2.4	Chapter Summary	71
3	Research Methodology	72
3.1	Chapter Overview	72
3.2	Research Theory and Design Principles	72
3.2.1	Research Theory	72
3.2.1.1	Inductive Research Method	73
3.2.1.2	Deductive Research Method	73
3.2.1.3	Abductive Research Method	74
3.2.2	Research Design	74
3.2.2.1	Experimental Research Design	75
3.2.2.2	Comparative Research Design	75
3.2.2.3	Case-study Research Design	76
3.2.2.4	Cross-sectional Research Design	76
3.3	Research Philosophy	76
3.3.1	Epistemology	77
3.3.1.1	Objectivism	77
3.3.1.2	Constructivism	78
3.3.1.3	Subjectivism	78
3.3.2	Ontology	79
3.3.3	Axiology	79
3.4	Research Paradigms	80
3.4.1	Positivism	80
3.4.2	Interpretivism	81
3.4.3	Pragmatism	81
3.5	Data Collection Methods	82

3.5.1	Qualitative Research	83
3.5.2	Quantitative Research	83
3.6	Research Choices for This Study	84
3.6.1	Deductive Research Method	84
3.6.2	Experimental and Case Study Research Design	85
3.6.3	Epistemological Stance	86
3.6.4	Ontological Stance	87
3.6.5	Positivist Research Paradigm	88
3.6.6	Quantitative Data Collection Method	89
3.7	Sampling Method	89
3.8	Chapter Summary	93
4	A Scalable Framework for TRAP Monitoring and Forecasting	94
4.1	Chapter Overview	94
4.2	Framework Design Methodology	95
4.3	Framework Components	96
4.3.1	Hardware Layer	96
4.3.2	Data Storage Layer	98
4.3.3	Integration Layer	99
4.3.4	Analytics Layer	99
4.4	Development and Deployment of the REVIS System Prototype . .	100
4.4.1	REVIS Highway Monitoring Devices	100
4.4.1.1	REVIS Device Development and Evaluation	100
4.4.1.2	Device Deployment in Case Study Regions	103
4.4.2	Exploratory Analysis of Pollution and Weather Data	103
4.4.2.1	The Impact of Weather on $PM_{2.5}$, PM_{10} and NO_2 . .	106
4.4.2.2	The Impact of Region on $PM_{2.5}$, PM_{10} and NO_2 . . .	110
4.4.3	Forecasting Model Training and Evaluation	113

4.4.3.1	Data Description	113
4.4.3.2	Meteorology Data Integration and Data Set Pre-Processing	114
4.4.3.3	Validation Set Creation and Training Architecture .	116
4.4.3.4	Model Evaluation	117
4.4.4	Evaluating the Scalability Performance of the REVIS System	120
4.5	Chapter Summary	123
5	Multi-target Regression for TRAP forecasting	124
5.1	Chapter Overview	124
5.2	Monitoring Site and Integrated Third-Party Data	125
5.3	Data Description	125
5.3.1	Pollution Data	125
5.3.2	Traffic Data	126
5.3.3	Weather Data	127
5.3.4	Elevation Data	128
5.3.5	Emissions Factor Data	128
5.3.6	Background Air Pollution Concentration Data	130
5.4	Machine learning approach	131
5.4.1	Multi-target regression and RNNs	134
5.4.2	Fastai, prophet and multioutputregressor methods	135
5.4.3	Data preprocessing	136
5.5	Experimentation and Model Training	139
5.5.1	Experiment 1 - Comparing Fastai, Prophet and MultiOutputRegressor defaults	139
5.5.2	Hyperparameter tuning with optuna and gridsearchcv	141
5.5.3	Experiment 2 - exploring lagged dependent variables (LDVs)	142
5.6	Model Validation and Results	143
5.6.1	Performance Metrics	143

5.6.2	Test and Validation Data	144
5.6.3	Experiment 1 Results	145
5.6.4	Experiment 2 Results	151
5.6.5	Statistical significance of results	152
5.6.6	Results comparison with related work	156
5.6.7	Experiment 3 - verify model's performance on missing data .	157
5.7	Feature importance on best model results	163
5.7.1	Experiment 4 - fewer features, same accuracy	163
5.7.2	Experiment 5 - features ablation test	165
5.8	Chapter Summary	167
6	Algorithmic Audit and Model Integration In Journey Planner	168
6.1	Chapter Overview	168
6.2	Baseline Model Evaluation	168
6.2.1	Multi-target regression deep learning model	168
6.2.2	Identifying outliers from original data set	169
6.2.3	Model's performance on outliers	170
6.2.4	Synthetic data generation from outliers	171
6.3	Experimentation	174
6.3.1	MTR-1E Model Training and Validation	174
6.3.2	Model Performance Evaluation and Results	175
6.4	Model Deployment and Journey Planner Integration	178
6.4.1	Oracle Cloud Deployment	178
6.4.2	Development of REVIS Travel Planner (RTP)	179
6.4.3	Integration of Mapbox Software Development Kit	180
6.4.4	Integrating Model Forecasts into RTP	181
6.5	Chapter Summary	182

7	Conclusion and Recommendations	183
7.1	Chapter Overview	183
7.2	Summary Of The Study	183
7.3	Reflections on the Quantitative Results	185
7.4	Challenges	187
7.5	Implication For Practice	189
7.6	Directions For Future Research	191
	References	201

List of Figures

1.1	The current situation of highway AQ monitoring in the UK. (House of Commons 2024)	29
1.2	Research workflow for addressing the objectives defined for this study	42
1.3	Thesis Layout	46
2.1	Deep Neural Networks(DNNs) Architecture	64
2.2	Convolutional Neural Networks(CNNs) Architecture	65
2.3	Recurrent Neural Networks (RNNs) Architecture	66
2.4	Autoencoders (AEs) Architecture	66
3.1	Philosophical Perspective For This Study	84
3.2	Sampling Techniques, Source: (Thornhill et al. 2009)	93
4.1	Pipeline diagram for framework design and validation	95
4.2	Scalable Framework for Highway Air Quality Monitoring and Prediction	97

4.3	Calibrated NO_2 and $PM_{2.5}$ readings from field. Vertical units are in $\mu g/m^3$ for $PM_{2.5}$ and ppb for NO_2 . Even with the calibration, NO_2 readings sometimes record negative readings.	102
4.4	Maps showing the distribution of 14 REVIS devices across four UK regions: Newport (1), Chepstow (1), Lewisham (6), and Southwark (6). For the Southwark and Lewisha locations in London, devices captured readings from the A302 and A2209 highways, while those in Newport and Chepstow were deployed near the M4 and A48 highways, respectively.	105
4.5	Total monthly readings captured by deployed sensors between November 2020 and August 2021. These plots illustrate the amount of missing data in the first two months when some devices were offline. Chepstow had the lowest monitored readings overall.	107
4.6	Distance matrix of weather parameters using Pearson’s correlation. A strong correlation can be noticed between “temp”, “temp_min”, “temp_max”, “wind_speed”, “wind_degree” and “feels_like”. There is also a discernible correlation between “clouds_all” and “humidity”/“windspeed”.	107
4.7	The seasonal trends for temperature in Newport, Southwark, Lewisham and Chepstow. Newport has the lowest temperature of $8.6^\circ C$ in winter as there was also no reading recorded for Chepstow, as illustrated in plot (a). Chepstow had the highest average temperature of $19.76^\circ C$ in spring and $21.78^\circ C$ in summer, as shown in plots (b) and (c)	109
4.8	The seasonal trends for humidity in Newport, Southwark, Lewisham and Chepstow. Similar to temperature and pressure, no reading was captured for Chepstow in winter. However, the region recorded the least humidity of 53.95% in spring, as illustrated in plot (b). Newport had the highest average humidity of 90.96% in winter and 73.54% in spring, as shown in plots (a) and (b) . . .	111

4.9	Plots highlighting the varying monthly averages for the three monitored pollutants. These averages varied significantly and are an indication that some influential factors may have affected the concentration levels	112
4.10	Auto-SQL generation to pre-process the data set. An SQL command which generates 3-hour and 6-hour pollutant averages from the preceding readings is depicted.	115
4.11	Sample of integrated weather dataset after pre-processing.	115
4.12	The model's training loss against the learning rate to determine the appropriate learning rate. The learning rate was fixed at the point where the plot started dipping (i.e., 10^{-4})	117
4.13	A plot showing the model's training and validation losses against the number of epochs. It is worth noting that there was a gradual decrease in both losses as the training epochs increased which indicates that the model was learning. Further training beyond 20000 epochs would have either resulted in overfitting or no further drop in both losses	119
4.14	An illustration of captured NO_2 pollutant readings (blue highlight) and the deep learning model predictions (red highlight). These results were derived from an evaluation using the validation data set. It should be pointed out that the model's predictions are not too far off the actual readings.	119
4.15	Plots of bare metal vs GPU instance as number of devices increased	121
4.16	Plots of scalability metrics showing database performance as the number of devices increased	122
5.1	Snapshot of pollution data.	126
5.2	Snapshot of traffic data.	127
5.3	Snapshot of weather data.	127
5.4	Snapshot of elevation data.	128
5.5	Snapshot of emission factor data.	129

5.6	Snapshot of background concentration data.	131
5.7	Multi-target model training architecture using the newly curated data set. Feature engineering steps including normalisation and log transformation were carried out before training on three different algorithms used for experimentation.	133
5.8	Multi-target vs single-target neural networks.	135
5.9	Data distribution for all three pollutants.	139
5.10	Summary of experiments carried out in this study.	140
5.11	Training and validation losses on Fastai after 1500 and 3000 epochs for exper- iments 1 and 2 respectively.	147
5.12	Experiment 1 - Fastai's model predictions.	148
5.13	Experiment 1 - MultiOutputRegressor's model predictions.	149
5.14	Experiment 1 - Prophet's model predictions.	150
5.15	Experiment 2 - Fastai MTR predictions for $NO_2, PM_{2.5}$ and PM_{10}	153
5.16	Experiment 2 - MultiOutputRegressor's MTR predictions for $NO_2, PM_{2.5}$ and PM_{10}	154
5.17	Experiment 2 - Prophet's model predictions.	155
5.18	Fastai model's performance when missing traffic data.	158
5.19	Fastai model's performance when missing weather data	159
5.20	Fastai model's performance when missing elevation data	160
5.21	Fastai model's performance when missing emissions factor data	161
5.22	Fastai model's performance when missing background concentration data	162
5.23	Feature importance from experiment 2. Traffic features including 'LGV count' and 'car count', 'average speed' were in the top list with the hour of the day, 'wind direction', 'PM emission factor' and ' NO_2 emission factor' also part of this list. Some of the least influential parameters were 'bike count', minute of the day and similar date parameters.	164

5.24	Feature importance after retraining on the top twelve features from experiment 2. All the traffic features except ‘car count’ maintained the top spot while ‘wind direction’ and ‘ NO_2 emission factor’ dropped further down the importance list.	164
5.25	Feature ablation test to reveal features with the most impact on fastai model’s predictions. The x-axis contains the feature list with each tick representing the feature that was removed when the model was retrained and RMSE score recalculated. The RMSE scores are represented on the y-axis. This chart indicates the importance of traffic and weather data as the RMSE scores increased when these features were removed from the data set.	166
6.1	Outliers detected within target pollutants: A total of 447 outliers detected, comprising 168 NO_2 outliers, 150 $PM_{2.5}$ outliers, and 129 PM_{10} outliers. . .	171
6.2	Plots of MTR-1 model’s performance on predicting $NO_2, PM_{2.5}$ and PM_{10} outliers	172
6.3	Distribution of outlier data sets (a) and generated training data set using the GaussianCopula method (b). Both plots depict normal distributions for both data sets, a key characteristic and prerequisite for employing the Gaussian-Copula data generation technique. Notably, a similarity is also evident in the distributions of both data sets	173
6.4	Training and validation loss after 1300 epochs training a model with the synthetic data set.	175
6.5	Plots of MTR-1E model’s performance on predicting $NO_2, PM_{2.5}$ and PM_{10} outliers	177
6.6	Screenshots demonstrating key features of the RTP iOS and Android app: user location detection, travel modes, emission-based route planning, and colour-coded route display on maps	180

List of Tables

4.1	Sensor Specifications and Accuracy	100
4.2	Regression analysis of weather parameters vs pollutant concentration	108
4.3	Pollutant summary statistics based on region	113
4.4	Descriptive statistics for the dataset	114
4.5	Hardware specifications of the two oracle cloud instances used to test scalability	120
5.1	Pollutant background concentration for the four regions of interest in the year 2020 and 2021	130
5.2	Descriptive statistics of the pollutants data	138
5.3	Hyperparameters used for experiment 1 - default configurations	140
5.4	Details of Hyperparameters optimised using Optuna and GridSearchCV	142
5.5	Experiment 1 results of MTR models prediction for different timesteps	145
5.6	Experiment 2 results of MTR models prediction for different timesteps	151
5.7	Statistical significance and model evaluation using Wilcoxon signed rank test	152
5.8	Comparison of prediction results with existing studies based on RMSE score	157
6.1	Boundary values for outlier detection in the target pollutants	169
6.2	Comparing results of MTR-1 vs MTR-1E's performance on Test set A and B	176

Appendices

Appendix A: Data summary for pollutant estimation before processing	193
Appendix B: List of Attributes After Processing, Including Classification as Categorical, Continuous, Independent, and Dependent Variables	195
Appendix C: List of attributes captured for MTR pollutant concentration forecasting	198

List of Acronyms

ADMS Atmospheric Dispersion Modeling System

AE Auto-Encoder

AI Artificial Intelligence

AQI Air Quality Index

AURN Automatic Urban and Rural Network

BPTT Back Propagation Through Time

CAP Credit Assignment Path

CCC Climate Change Committee

CD Continuous Deployment

CERC Cambridge Environmental Research Consultants

CI Continuous Integration

CNN Convolutional Neural Network

CO Carbon monoxide

DAQI Daily Air Quality Index

DBN Deep Belief Network

DEFRA Department for Environment, Food and Rural Affairs

DL Deep Learning

DNN Deep Neural Network

DRL Detection Range Limits

EFT Emission Factor Toolkit

ETL Extract, Transform and Load

GAN Generative Adversarial Network

GDP Gross Domestic Product

GHG Greenhouse Gas

GFLSM General Finite Line Source Model

GPU Graphics Processing Units

GRU Gated Recurrent Unit Networks

HGV Heavy Goods Vehicle

ITS Intelligent Transportation Systems

LDV Lagged Dependent Variables

LGV Light Goods Vehicle

LOCF Last Observation Carried Forward

LSTM Long Short-Term Memory

LR Linear Regression

MAE Mean Absolute Error

M2M Machine To Machine

ML Machine Learning

MRF Markov Random Field

MSE Mean Squared Error

MTR Multi-Target Regression

NN Neural Network

NO2 Nitrogen dioxide

PCM Pollution Climate Model

PM Particulate Matter

PWA Programmable Web Apps

RCP Royal College of Physicians

RF Random Forest

RFID Radio Frequency Identification

ReLU Rectified Linear Units

REVIS Real-time Emission Visualisation System

RMSE Root Mean Square Error

RNN Recurrent Neural Network

RSU Road Side Unit

RTP REVIS Travel Planner

SAE Stacked Autoencoders

SVM Support Vector Machine

TMU Traffic Monitoring Unit

TPU Tensor Processing Units

TRAP Traffic Related Air Pollution

UFP Ultrafine Particle

UK United Kingdom

UPS Unified Prediction Service

VOC Volatile Organic Compound

XGB XGBoost

Chapter 1

Introduction

1.1 Background

The United Kingdom (UK) has struggled in recent years to manage the impact of the greenhouse gas (GHG) emissions emanating from its key sectors such as transportation, waste management, agriculture, amongst others. According to CCC (2017), despite a 5% reduction in emission rate between 2015 and 2016, the country was unable to meet its second carbon budget and GHG reduction targets in 2017. Studies associated the missed target to the uptake in road construction projects that led to an increased number of cars and emissions on major highways (Sloman et al. 2017). With the transportation sector being a significant contributor to a substantial amount of GHG emissions (27%), it is no surprise the government has invested about £100 million to proactively tackle air quality challenges in a bid to protect its citizens' health and support clean air initiatives (DEFRA 2019). The impact of traffic-related air pollution on human health cannot be overemphasised. Its long and short term effects have been linked with many life-threatening health conditions and diseases (Pascal et al. 2013, Wu, Shaowei et al. 2016). Evidence suggests that currently, the UK records an average of about 40,000 deaths yearly as a result of NO_2 pollution alone (RCP 2016). The recorded annual death by countries, according to EEA (2016), ranked UK as the second country with the highest number of fatalities from NO_2 , second only to Italy.

NO_2 and other harmful pollutants like $PM_{2.5}$ and PM_{10} have been associated with ailments such as cancer, asthma and heart diseases; and have resulted in enormous treatment costs for people suffering from such conditions.

According to World Bank (2022), the global cost of the adverse health effects associated with exposure to air pollution is \$8.1 trillion, equivalent to 6.1 per cent of global GDP. It is, therefore, surprising that a substantial fraction of the UK populace (particularly those that commute to their various destinations via highways) are still susceptible to the adverse health effects of air pollutants along the UK highways (Vohra et al. 2021). Due to exposure to motor vehicle exhaust emissions, non-exhaust related pollution from brake and tyre wear, and particles from highway construction (Barikayeva et al. 2018), commuters are constantly at risk of high concentrations of air pollutants. Some of these pollutants, especially $PM_{2.5}$, PM_{10} , NO_2 are the most life-threatening road pollutants that have been linked to cardiovascular and respiratory illnesses (Mabahwi et al. 2014, Alvanchi et al. 2020). The study of Public Health England (2019) estimates that between 2017 and 2025, these air pollutants would have costed the NHS and social care system in England a total of £1.6 billion.

There is therefore a pressing and cogent need to find innovative and sustainable ways to monitor air pollutants and curb their devastating effects on health and human capital, as well as associated GDP losses (DEFRA 2020). According to (Alvanchi et al. 2020), monitoring particulate matter ($PM_{2.5}$, PM_{10}) and other highway pollutants like NO_2 is not a straightforward task because pollutants tend to decay and diffuse into the background concentration within 200m from the source. Highway speed limits and traffic congestion complicate things further as they result in varying driving patterns such as sudden slow-downs and speedups, which elevate these pollution levels or limit their dispersion (Karner et al. 2010, Zhang & Batterman 2013). The most affected are the commuters or residents living close to roads since they are constantly exposed to numerous pollutants. An average

commuter will spend an average of 4%–7% of their day on or close to a major road.

According to Barthwal & Acharya (2018), most countries monitor air pollution using stationary monitoring stations operated by government authorities. Figure 1.1 illustrates how the UK currently monitors highways to come up with its ultra low emission policies. Highways are monitored by Highways England (a government-owned company charged with operating, maintaining, and improving motorways in England) via its automatic urban and rural network (AURN), which collects sparse air pollutant data. However, evidence suggests that these air quality analysers are relatively heavy and expensive to install or maintain (Carullo et al. 2007, Barthwal & Acharya 2018). Therefore, it is impracticable for Highways England’s monitoring stations to be deployed across the UK to capture pollutant concentration levels and improve air quality. Data captured from these AURN stations are instrumental for the UK government in monitoring long-term pollution trends, assessing the impact of policy initiatives, and ensuring compliance with health-based air quality standards. This data supports the UK’s air quality forecasting system, developed by the Met Office, aimed at mitigating health risks associated with traffic-related pollution. Despite these advances, real-time traffic-related air pollution (TRAP) forecasting remains a challenge, primarily due to the complex interplay of variables like weather conditions and traffic flow (Barrera-Animas et al. 2022, Sun et al. 2021). Over the past decade, there has been a significant increase in research efforts aimed at overcoming the hurdles of precise forecasting, with scholars delving into novel methodologies to tackle this complex issue. These studies have explored a range of innovative approaches, including the application of advanced machine learning algorithms, the integration of big data analytics, and the utilisation of real-time data streams to enhance prediction accuracy. Despite these advancements, the field continues to face several persistent constraints, such as the dynamic nature of the variables involved, the need for vast and high-quality datasets, and the challenges in integrating interdisciplinary knowledge for comprehensive forecasting models. Additionally, the rapid evolution of technology and changing

environmental conditions introduce further complexity, requiring ongoing adaptation and refinement of forecasting methodologies

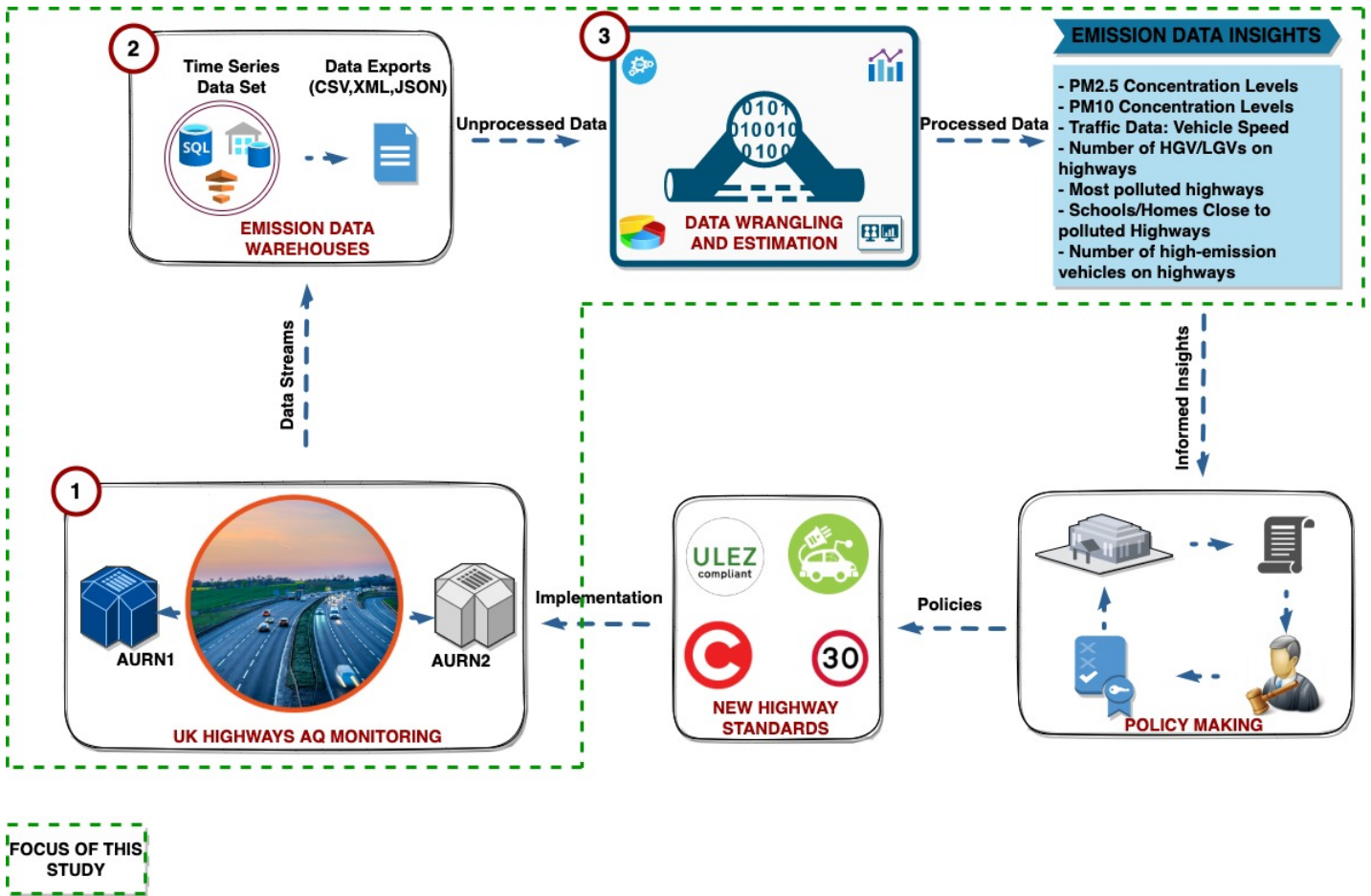


Figure 1.1: The current situation of highway AQ monitoring in the UK. (House of Commons 2024)

1.1.1 Urban Highway Pollution And Prevalent Pollutants

Highway traffic pollution in urban areas is a pressing concern with far-reaching implications for both public health and the environment. This issue is primarily driven by the emissions originating from a wide range of vehicles, including cars, trucks, and buses, which rely predominantly on fossil fuels for propulsion. The combustion of these fuels generates a complex mixture of pollutants released into the urban atmosphere. Promoting public transportation and reducing traffic congestion have been integral parts of urban planning and policy ini-

tiatives aimed at minimising the number of vehicles on the road and ultimately mitigating pollution. Various strategies, including congestion pricing, carpooling incentives, and the development of cycling and pedestrian-friendly infrastructure, are employed to manage traffic more effectively Kuss & Nicholas (2022). The complexities of urban highway pollution are intricate and multifaceted, transcending the simple consideration of health impacts and specific pollutant types. This multifaceted issue presents itself with a plethora of intricacies, one of which is the diverse range of pollution sources that contribute to the overall problem. While vehicle emissions are a significant contributor, they coexist with emissions from industrial and construction activities near urban highways, as well as emissions arising from road surfaces themselves when inadequately maintained. Each of these sources has unique characteristics, necessitating tailored and multifaceted approaches to address the issue effectively. Geographical and meteorological factors add another layer of complexity. The dispersion and concentration of pollutants are influenced by local topography, wind patterns, and atmospheric conditions, creating a dynamic and challenging environment in which pollution levels can vary significantly across different locations within the same urban area (Amato et al. 2014).

Traffic-related pollutants comprise a complex blend of harmful substances, with key pollutants including particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), and carbon monoxide (CO). $PM_{2.5}$ and PM_{10} are of particular concern due to their ability to remain suspended in the air and penetrate the respiratory system, causing health issues such as asthma, bronchitis, and cardiovascular problems Wang et al. (2009). Nitrogen oxides, primarily produced from vehicle combustion processes, contribute to ground-level ozone formation and can exacerbate respiratory conditions and cardiovascular diseases. Carbon monoxide, resulting from the incomplete combustion of fossil fuels, impairs the blood's ability to carry oxygen, leading to various health issues, including headaches, dizziness, and in severe cases, asphyxiation (Yli-Tuomi et al. 2005). Mitigating these pollutants involves comprehensive monitoring,

stringent emissions controls, and innovative pollution management strategies. Effective mitigation strategies necessitate a nuanced understanding of the geographical intricacies and their integration into comprehensive approaches, as well as leveraging advancements in technology to monitor and reduce pollution levels dynamically. By addressing these multifaceted challenges, urban areas can work towards significantly improving air quality and public health.

1.1.2 Supporting Technologies For Highway Air Quality Monitoring and Forecasting

The Internet of Things (IoT) has become transformative by connecting everyday objects to the internet, enabling comprehensive data collection and sharing. This technology bridges the physical and digital worlds and has evolved from machine-to-machine (M2M) communication in the late 20th century. IoT applications span various domains, including transportation, where it supports connected vehicles, intelligent traffic management, and enhances transportation safety Arthurs et al. (2021). In smart cities, IoT optimises infrastructure such as traffic management, waste collection, and street lighting Ramírez-Moreno et al. (2021). In healthcare, IoT enables remote patient monitoring and real-time health tracking through wearable devices Yuehong et al. (2016). In this modern era, the term “big data” has become increasingly prominent, signifying a revolutionary shift in the way we handle and derive value from data. Big data represents a fundamental transformation in the scale, speed, and diversity of data generated in our interconnected world. This shift has given rise to a new era of possibilities and challenges, shaping how businesses, organisations, and society as a whole collect, process, and utilise information (Johnson et al. 2023). At its core, big data is a term used to describe extraordinarily large and complex datasets. These datasets are often too massive to be effectively managed and analysed using traditional data processing tools and methods. Managing and analysing big data requires sophisticated tools like Apache Hadoop and Apache Spark for distributed processing and tools like Apache Kafka and Apache Flink for real-time data stream processing.

Deep learning is a sub-field of machine learning that has witnessed significant advancements, particularly in the last two decades. These developments have been propelled by a combination of factors, including the exponential growth in computational power, the availability of vast datasets, and innovative optimisation algorithms. Innovations like rectified linear units (ReLU) have addressed issues like the “vanishing gradient” problem, allowing for deeper networks. The recent surge in deep learning has also led to the development of ever-larger neural networks, including models with hundreds of millions or even billions of parameters. These models, often referred to as “transformers”, have demonstrated remarkable capabilities in various tasks, particularly in natural language processing. Deep learning, however, is not without challenges. Model interpretability remains a concern, as the internal workings of deep neural networks can be difficult to comprehend (Lisboa et al. 2023). Ethical considerations, such as bias in models and data, are also areas of active research and scrutiny. Researchers are working to make deep learning more transparent, fair, and accountable. As deep learning continues to advance, it holds the promise of further breakthroughs in artificial intelligence, from improved language models to more capable autonomous systems. Its potential is virtually limitless, and its influence is manifesting across diverse domains, from healthcare and finance to creative arts and environmental sciences Hatcher & Yu (2018). These technologies have the potential of collectively enhancing highway air quality monitoring and forecasting by enabling real-time data collection, processing, and analysis, leading to more accurate and timely insights for improving air quality.

1.1.3 Concepts Of Scalability And Accuracy In Highway Air Quality Forecasting

Scalability is a critical consideration in air quality forecasting, as it determines the ability of a forecasting system to handle increasing amounts of data and expand to cover larger geographical areas or more parameters without a significant drop in performance. Scalable air quality forecasting systems must be capable of integrating diverse data sources, such as IoT sensors, satellite imagery, and meteorological data, and processing this information in

real-time to provide timely and relevant forecasts Kaginalkar et al. (2021). The use of cloud computing platforms and high-performance computing resources is essential for scalability Jackson et al. (2010). These technologies allow for the deployment of distributed computing frameworks that can manage vast datasets and complex models efficiently. Scalable systems ensure that air quality forecasts remain accurate and reliable as more sensors are deployed and as the system is expanded to cover larger urban areas or regions.

Accuracy in air quality forecasting is paramount, as it directly impacts the reliability of the forecasts and the ability to make informed decisions. Accurate models can predict pollution levels with high precision, enabling authorities to issue timely warnings and take appropriate measures to mitigate the impact of poor air quality on public health. Deep learning advancements, particularly the development of sophisticated neural network architectures such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have significantly improved the accuracy of air quality forecasting models Liao et al. (2020). These models can capture complex spatial and temporal patterns in the data, leading to more precise predictions. Techniques like transfer learning and fine-tuning of pre-trained models further enhance accuracy by leveraging existing knowledge and adapting it to specific datasets. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), the coefficient of determination (R^2), and the correlation coefficient (r) are used to assess model performance. These metrics provide quantitative measures of accuracy and reliability, helping to refine and improve forecasting models to ensure they deliver dependable air quality predictions.

1.2 Research Problem

The current monitoring strategy employed by most EU member states and the UK involves the use of fixed monitoring stations like the AURN stations which are expensive to acquire and maintain (Borrego et al. 2015). This section introduces some of the prevalent challenges

of this strategy to better understand the importance of the proposed intervention in this study.

1.2.1 Coarse-Grained vs Fine-Grained Air Quality Measurements

Coarse-grained and fine-grained air quality measurements refer to the scale and detail at which air quality data is collected and analysed. Coarse-grained air quality measurements provide data at a broader scale, typically over larger geographical areas or longer time periods. These measurements might give an overall picture of air quality trends or average conditions but can miss local variations and short-term fluctuations. They're often used for general assessments, policy planning, and compliance with long-term air quality standards. Coarse-grained data can be sufficient for understanding regional trends or making comparisons between different large areas. AURN monitoring stations produce a limited spatial resolution of air quality (due to the distance between stations) which is needed for comprehensive spatiotemporal mapping and assessment (Yang et al. 2018). The expensive cost of acquiring these stations encourages sparse deployment, which in turn results in these relatively coarse-grained measurements (Rai et al. 2017). As highlighted in the study of Devarakonda et al. (2013), this form of raw data still requires the adoption of dispersion models to infer the concentration of pollution. Contrarily, a fine-grained measuring technique is spatially-dense and allows scalability to provide information on short-term concentration changes when needed (Baron & Saffell 2017). This type of measurement can detect short-term changes and are essential for understanding the dynamic nature of air pollutants and for making informed decisions that protect public health at a local level. Studies indicate that this sort of detailed pollutant information is currently restricted or completely non-existent (Kumar et al. 2015, Ahlers et al. 2016, Carabetta 2019).

1.2.2 Real-Time Air Quality Monitoring

While fine-grained air quality measurements are crucial for informed decision-making in environmental planning and policy, the need for immediate data access is equally imperative (Kadri et al. 2013). Advancements in wireless communication and sensing technologies have enabled companies like Aeroqual and Membrapor to develop sensors for real-time, comprehensive monitoring. However, these devices have inherent limitations. They often require multiple sensors to track various pollutants due to their specific detection capabilities, potentially complicating and escalating the costs of monitoring systems (Lewis et al. 2016). The accuracy and sensitivity of these sensors can vary, influencing the precision of pollutant level readings. External factors such as temperature and humidity can also affect sensor performance. Furthermore, regular calibration and maintenance are essential to ensure sustained accuracy, introducing further challenges in sensor deployment and data consistency. These factors are vital in the effective application of sensors in air quality monitoring strategies. Additionally, the cost-benefit balance regarding sensor longevity and measurement accuracy remains a significant challenge, as many low-cost sensors fail to provide accurate measurements over long periods, primarily due to the power required for continuous operation (Rai & Kumar 2017).

1.2.3 Prediction Accuracy Of TRAP Forecasting Models

Asides from inefficient highway air quality monitoring, another major challenge rests on the issue of how data disparity and isolated data sets affect the accurate prediction of pollutant concentration levels. Traditional models are mostly trained on traffic flow, meteorological and historic pollution data collected over many years. Other highway and traffic-related data such as background air pollution concentrations, vehicle emission factor and highway topography are often ignored because of their unavailability. Consequently, many of the machine learning models only excel on the often limited data sets upon which they have been trained. While it might seem intuitive that providing a real-time quality dataset to an air pollution forecasting

model would enhance its prediction accuracy, this is not always the case. The integration of real-time data into forecasting models is influenced by multiple factors, such as the model’s architecture, data complexity, and the model’s capability to assimilate and adapt to real-time inputs. Real-time data can sometimes introduce noise, complicating the learning process and not always enhancing model accuracy. Hence, its impact on forecasting must be assessed individually, considering the model and data’s unique attributes.

The study of Fong et al. (2020) for example, could only make next day predictions and struggled with periods shorter than a day or even several days ahead. While the proposed approach was able to demonstrate the use of transfer learning techniques in conjunction with LSTM recurrent neural networks, the authors linked the difficulty of making predictions over various timescales to data availability. Another important limitation of the study is the inability of the proposed approach to simultaneously and accurately predict multiple pollutants and the impact of contributing variables. The impacts of contributing variables like weather parameters, traffic flow, and historical pollution data on air pollution forecasting are significant and multifaceted. Weather conditions, including temperature, wind speed, and humidity, directly influence pollutant dispersion and concentration, while traffic flow data reflect emission levels, particularly in urban settings. Historical pollution trends also provide valuable insights into periodic variations, enabling models to predict future air quality. Predictions from machine learning algorithms like linear regression depend on the linear dependency between different highway parameters and pollutants. However, these relationships are complex and non-linear, thereby making multi-target predictions even more difficult (Masmoudi et al. 2020). Also, most of the developed models do not offer pragmatic solutions that can be deployed in a real-world scenario. Rigorous validation of these models in these kinds of scenarios is almost non-existent.

1.3 Justification For Study

The importance of tracking highway air pollution and implementing effective mitigation strategies cannot be overstated, especially in today's context of escalating environmental concerns globally. Air pollution, particularly from traffic-related sources, poses significant health risks, contributing to respiratory and cardiovascular diseases among populations exposed to pollutants. The UK, with its dense traffic networks, serves as a critical area of study but also provides insights applicable to Europe and other regions worldwide, where similar challenges persist.

The ability to accurately forecast and monitor Traffic-Related Air Pollution (TRAP) is essential for developing strategies that can effectively reduce exposure and mitigate the adverse effects on public health and the environment. By enhancing our understanding and prediction capabilities, policymakers can make informed decisions that lead to more sustainable and healthier urban environments. This includes implementing traffic management strategies, urban planning reforms, and pollution control technologies that align with both national and international environmental goals. Moreover, as climate change continues to impact environmental dynamics globally, the need for robust, scalable solutions to monitor and predict air pollution becomes even more critical. These solutions are pivotal not only for immediate health concerns but also for long-term sustainability efforts. Implementing advanced forecasting and monitoring frameworks can therefore provide a blueprint for other regions struggling with similar issues, making the research not only relevant to the UK but also to the global effort in combating air pollution. This global relevance underscores the necessity of this research at a time when both environmental awareness and the urgency to act are at their peak.

1.4 Research Novelty

There is a predominant gap in knowledge within the body of air quality literature, especially concerning the real-time capturing and forecasting of TRAP pollutants, as argued in preceding subsections. Existing studies have not yet explored an integrated approach like the one proposed in this study, which could improve the accuracy of forecasting models. This lack of exploration is due to the complexities associated with integrating various digital technologies, the substantial resources needed, and challenges related to data accessibility. On the back of significant advancements in scalable machine learning (ML) approaches such as deep learning, this study, therefore, proposes and implements a scalable monitoring and forecasting framework for real-time capturing and estimation of pollutant concentration levels on UK highways. This framework leverages internet of things (IoT) sensors for real-time monitoring, graphics processing units (GPUs) for parallel computing, big data for scalable storage and deep learning for forecasting highway pollutant concentration.

In addition, this study takes a different approach and models the prediction of traffic pollutant concentration as a multi-target regression problem with additional highway data such as background air pollution concentrations from the UK Pollution Climate Model (PCM), vehicle emissions factor and terrain data added to the conventional weather and historic pollution data. A range of available datasets was thoroughly examined to pinpoint the most pertinent ones, especially those offering detailed and comprehensive data that aligns with the objectives of the research. The PCM was chosen for its detailed, high-resolution depiction of air pollution, crucial for the study’s analytical depth while the vehicle emissions and terrain data allows for a nuanced analysis that acknowledges the complex interplay of factors influencing traffic pollutant levels. While Multi-Target Regression (MTR) permits the simultaneous prediction of multiple dependent variables, its real-world application still poses numerous challenges due to the complexity of some domains (Borchani et al. 2015). Previous investigations into using MTR for forecasting pollutant concentrations have encoun-

tered issues with accuracy or feature selection. Notably, no prior research has assessed the unique dataset combination presented in this study. On a related note, many studies confirm that deep learning algorithms can help models learn the fundamental relationships between variables in a dataset (Guo & Berkhahn 2016, Shrestha & Mahmood 2019, Akinosho et al. 2020). Nevertheless, there's ongoing discussion among researchers about the effectiveness of these algorithms with tabular data (Fayaz et al. 2022). Therefore, this study also seeks to examine the applicability of deep learning for tabular data and to explore the dynamics of variable relationships through an extensive audit of the developed forecasting models.

1.5 Research Questions

Listed below are the identified research questions of this study :

- What are the limitations of existing highway pollution forecasting methods?
- What are the computational challenges in implementing a scalable solution for highway pollution forecasting, and how can they be overcome?
- Can the integration of supplementary highway data, including background air pollution levels, vehicle emission factors, and terrain information, with historical pollution and meteorological records enhance the accuracy of forecasting models?
- Can the application of deep learning techniques, such as categorical embeddings, enhance the current accuracy levels?
- What is the most effective way to communicate monitored and forecasted highway pollution information to the general public?

1.6 Research Aim and Objectives

The overall aim of this research is to investigate a better performing and scalable method of monitoring and forecasting highway pollution using big data, IoT and deep learning tech-

nologies. Hence, the following objectives have been identified as being critical to realising the primary aim.

- Evaluate existing TRAP forecasting methods, identify limitations, and identify appropriate deep learning algorithms.
- Develop and implement a scalable forecasting framework to address and resolve computational challenges in highway pollution modelling.
- Develop an extensive training dataset incorporating relevant highway features to enhance the accuracy of pollutant forecasting and demonstrate the complexities of data integration.
- Examine and compare the accuracies of different multi-target forecasting models for NO_2 , $PM_{2.5}$, and PM_{10} concentrations, with and without the use of categorical embeddings, against prominent machine learning algorithms, and identify influential features using the model that performs best.
- Design a prototype application to demonstrate model integration and real-time air quality visualisation and journey planning on UK highways.

1.7 Research Methodology And Workflow

To fulfil the research objectives of this study, a combination of experimental design and case-study methodology was employed, each serving distinct but interrelated purposes. The rationale behind this methodological approach stems from the need to achieve two primary objectives. Firstly, this study recognises the importance of system implementation to vividly demonstrate the practicality of the proposed framework through the development of a fully operational system. This approach inherently aligns with the principles of experimental design in research, emphasising the importance of systematic and structured processes (Harrison et al. 2017). On the other hand, the adoption of a case-study strategy was motivated

by the intention to rigorously test the results of the developed system in diverse real-life settings. This pragmatic approach has proven highly effective and has been a cornerstone in related research endeavours featuring multiple case studies, as exemplified in studies by Saide et al. (2016) and Chauhan et al. (2021). The use of a case-study methodology is crucial for assessing the system’s effectiveness in various contextual situations, greatly contributing to the study’s thoroughness. Moreover, the experimental design approach was key in shaping the layered structure of the framework to be proposed later, directly aligning with the second research objective. Layering within the framework represents a widely adopted application design technique, facilitating the systematic dissection of intricate software systems into manageable modules. These layers encompass vital components such as libraries, programming languages, and services, which are indispensable for the system’s effective monitoring and forecasting. This architectural structure enhances the organisation of the system, making it more modular, scalable, and amenable to maintenance and expansion. Figure 1.2 depicts a high-level workflow defined for achieving the objectives of this study.

1.8 Research Scope and Limitations

In this research, addressing the “who”, “what”, “why”, and “how” is crucial to defining the study’s scope clearly and ensuring its relevance and applicability. The “who” focuses on the key stakeholders impacted by Traffic-Related Air Pollution (TRAP), including environmental researchers, urban planners, policymakers, and the UK public, especially those near case-study highways. The “what” entails the development and implementation of a sophisticated monitoring and forecasting framework for TRAP, leveraging IoT sensors, GPUs, and deep learning within the context of these highways. The “why” underscores the urgency to enhance TRAP understanding and predictions, driven by the imperative to mitigate health and environmental risks. Lastly, the “how” details the methodological approach, utilising real-time data collection, advanced computational processes, and predictive analytics to extract meaningful insights. Addressing these questions ensures a holistic understanding of the

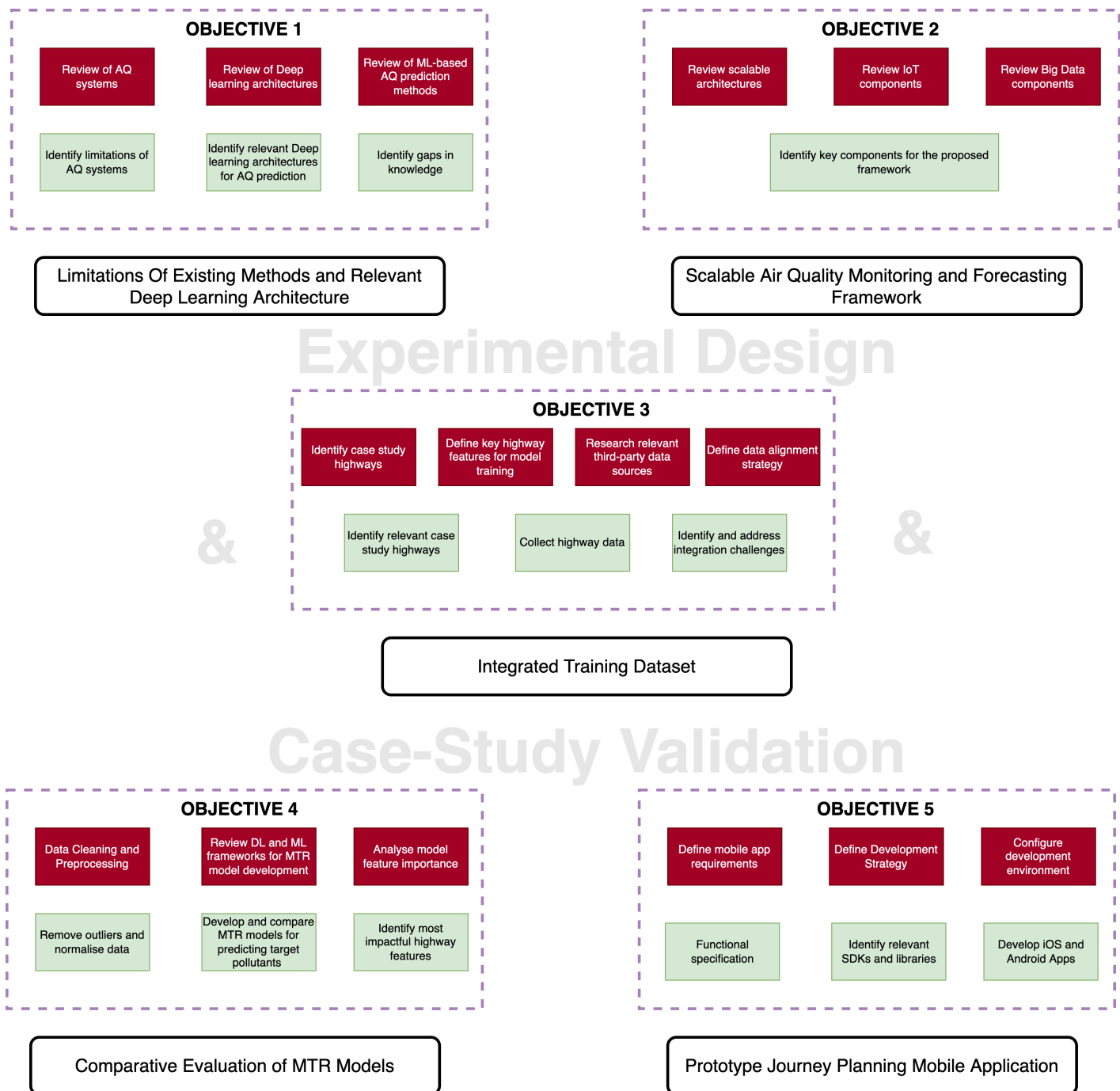


Figure 1.2: Research workflow for addressing the objectives defined for this study

study’s intent, methodology, and potential impact, providing a comprehensive framework that guides the research towards meaningful outcomes.

While the research aims to make a substantial contribution to the field of TRAP monitoring and forecasting, it recognises the limitations inherent in its scope. The selection of only four UK highways may not encompass the full variability of highway environments across the country, potentially limiting the generalisability of the findings. The study’s reliance on advanced technologies and specific data sets means that the results are contingent on the current state of these tools and the data’s accuracy and completeness. Moreover, the focus on the UK context necessitates careful consideration when applying the findings to other geographic settings, as local traffic, environmental, and policy conditions may differ significantly. Despite these constraints, the research offers significant insights into TRAP monitoring and forecasting, with potential implications for policy and practice within and possibly beyond the UK context.

1.9 Expected Contribution To Knowledge

This doctoral research makes a substantial contribution to the field of air quality management and data-driven forecasting. One key aspect it addresses is the intricate process of data integration. The research explores the intricacies of data integration, elucidating the process of combining data from varied sources to create a unified, more coherent dataset. This process is not without its challenges, as highlighted in the study, which also emphasises the critical need to obtain the necessary authorisations to access specific datasets. Gaining insights from a real-world perspective is invaluable, offering researchers and practitioners critical understanding and practical knowledge that can guide their work, enhancing the accuracy and depth of their analyses, and facilitating more informed decision-making in their respective fields. Another fundamental component of this research is its exploration of data format disparities among the data sets. It introduces the innovative concept of data

integration maps, which are schematic representations that illustrate how data from various sources is combined, transformed, and loaded into a target system, acting as a bridge between data sources with differing formats. This approach offers a practical solution for harmonising heterogeneous data and streamlining the integration process. The study’s focus on data format standardisation adds depth to the understanding of data integration complexities, benefiting those involved in working with diverse data sources.

Similarly, this study builds on existing research knowledge through the comparative analysis of deep learning against other prominent machine learning algorithms in the context of air quality forecasting. The study underscores the importance of algorithm selection and hyperparameter tuning. It reveals that, regardless of the data’s quality or type, effective model training requires careful consideration and customisation. These findings serve as a valuable resource for researchers and practitioners seeking to enhance the accuracy of air quality forecasting, emphasising the need for adaptable machine learning techniques in this field.

Finally, the study underscores the social, economic, and technological significance of accurate air quality forecasting. It emphasises the potential to mitigate traffic-related pollution risks and addresses the pressing issue of environmental injustice in developed countries. Additionally, it recognises the economic implications of air pollution, making the case for more informed decision-making in air quality management systems. From a technological perspective, the research exemplifies the productionisation of air quality models for traffic-related pollution, offering a streamlined approach to deploying models for real-world use cases while addressing the challenge of model drift. These comprehensive insights contribute significantly to the knowledge base in air quality management and environmental sustainability.

1.10 Thesis Layout

This thesis is structured into chapters, with each chapter dedicated to addressing one or two research objectives. Each chapter follows a consistent format: it begins with a brief introduction, proceeds to the main body, and concludes with a chapter summary. Following this introductory chapter, Chapter Two undertakes a review of the existing body of literature in the realm of air quality monitoring and forecasting. This chapter emphasises the significant constraints identified in studies attempting to utilise the three primary technologies, namely Big Data, the Internet of Things, and Deep Learning, all of which are central to this study. In addition, this chapter offers an overview of existing deep learning architectures, serving to fulfil the first research objective. Chapter Three delves into the philosophical underpinnings of this study, expounding upon the principles guiding research design, data collection methods, and sampling techniques employed throughout the study. In Chapter Four, a scalable framework designed for city-wide and national-scale air quality monitoring and prediction is presented. This chapter provides insights into the framework’s architectural layers and their key components. Moreover, a practical implementation is included to validate the framework’s functionality and demonstrate its real-world applicability. Chapter Five extends the baseline model developed during the framework validation process in Chapter Four. This extension involves the integration of additional highway data into the training data set, following the framework’s recommendations. The augmented data set is then used to train an improved model, with comprehensive details of the training process and comparisons with other machine learning models presented in this chapter. In Chapter Six, an algorithmic audit is conducted on the best-performing model to identify outliers that may impede its performance. This chapter outlines the results of this audit and implements rectification strategies. Finally, Chapter Seven serves as the conclusion of this study. It offers a summary of research findings, addresses the challenges encountered, and provides recommendations for future research endeavours as shown in figure 1.3 below.

1	Chapter 1: Introduction Background, Research Problem, Aim and Objectives and Research Methodology
2	Chapter 2: Review of Enabling Digital Technologies for TRAP Monitoring/Forecasting TRAP Monitoring/Forecasting Technologies, Big Data/IoT and CTM Overview, Barriers to Adoption
3	Chapter 3: Research Methodology Research Theory and Design Principles, Research Philosophy, Research Paradigms, Data Collection Methods, Sampling Method
4	Chapter 5: A Scalable Framework For TRAP Monitoring and Forecasting Framework Components, Development and Deployment of REVIS System Prototype
5	Chapter 5: Multi-Target Regression For TRAP Forecasting Monitoring Site and Integrated Third-Party Data, Data Description, Model Training and Validation Results, Feature Importance
6	Chapter 6: Algorithmic Audit and Model Integration In Journey Planner Baseline Model Evaluation, Experimentation, Model Deployment and Journey Planner Integration
7	Chapter 7: Conclusion and Recommendations Challenges, Implication For Practice, Directions For Future Research

Figure 1.3: Thesis Layout

Chapter 2

Review of Enabling Digital Technologies For TRAP Monitoring and Forecasting

2.1 Chapter Overview

This chapter explores the digital technologies pivotal in managing traffic pollution, building upon the challenges outlined in the preceding chapter. While previous research has offered solutions, they often tackle the issues in isolation. The chapter reviews technologies such as the Internet of Things, Big Data, and Machine Learning, recognised in the literature as key components in traffic pollution management. It examines the specific roles these technologies play in monitoring and forecasting traffic-related pollution, discussing their application in existing studies and implementation. In addition, the chapter addresses the obstacles to the widespread adoption of these technologies in the field.

2.2 Traffic-Related Air Pollution Monitoring - Enabling Technologies and Methods

This section explores the applications of IoT and Big Data in TRAP monitoring, highlighting their transformative impact on urban environmental management. It also discusses the barriers to their adoption, including technical challenges, data privacy concerns, and the need for robust infrastructure and regulatory frameworks.

2.2.1 Internet of Things And TRAP Monitoring

The relevance of IoT to TRAP monitoring cannot be overstated in today's increasingly urbanised and technologically advanced world. IoT's role in TRAP monitoring is a game-changer, offering innovative solutions to the challenges posed by urban air pollution. IoT's significance lies in its ability to create a network of interconnected sensors and devices that continuously collect data on various environmental parameters, including air quality. These sensors are strategically placed throughout urban areas, providing a real-time, granular view of air pollution levels. This vast amount of data is instrumental in understanding the complex dynamics of TRAP. By amalgamating diverse data streams, researchers and policymakers can identify trends and patterns that would be impossible to discern from traditional monitoring alone. IoT's geospatial capabilities are equally relevant to TRAP monitoring. It provides a fine-grained understanding of pollution hotspots and vulnerable areas. This information is invaluable in designing targeted mitigation strategies, such as the placement of green infrastructure or the rerouting of traffic away from sensitive locations. By harnessing the geospatial aspects of IoT, cities can optimise resource allocation and focus efforts where they are most needed. IoT goes beyond data collection; it encourages citizen engagement and crowd-sourced data collection. Mobile apps and online platforms enable residents to report pollution incidents, contributing real-time data for monitoring. This citizen science approach not only improves data accuracy but also fosters community awareness and participation in

pollution management.

The interconnectedness of IoT systems allows for collaborative efforts between cities, regions, and even countries. Shared data and experiences can lead to more effective regional pollution control strategies and the standardisation of monitoring and reporting practices. This global perspective is vital in addressing transboundary pollution issues and striving for more sustainable, cleaner urban environments. In addition, IoT contributes to transparency and accountability in pollution management. By making pollution data accessible to the public, it empowers citizens to hold authorities and industries accountable for their pollution levels. This transparency can stimulate action and encourage the adoption of cleaner technologies and practices, ultimately driving a positive change in urban pollution dynamics. In summary, IoT's relevance to TRAP monitoring lies in its ability to provide real-time, data-driven insights, enable predictive analysis, support targeted mitigation strategies, engage citizens, facilitate collaboration, and promote transparency, all of which are pivotal in addressing the challenges of urban air pollution.

2.2.2 Existing Applications of IoT For TRAP Monitoring

With the recent application of IoT in several sectors, various applications in developing roadside and traffic-related air quality monitoring systems emerged. For instance, Martín-Baos et al. (2022) introduced a cost-effective IoT system that combined video processing and machine learning to monitor traffic flow and the Air Quality Index (AQI). The real-time traffic flow computation using motion vectors and the AQI estimation through machine learning regression models signified a significant advancement in precise air quality monitoring. Their approach not only overcame calibration complexities but also proved to be a valuable solution for diverse traffic and climate scenarios. Manna et al. (2014) tackled the issue of air quality degradation due to increasing urbanisation and vehicle emissions. Their innovative use of Wireless Sensor Networks, Electrochemical Toxic Gas Sensors, and Radio Frequency Identifi-

fication (RFID) offered real-time vehicle pollution monitoring. This approach was especially valuable for identifying and controlling sources of pollution exceeding specified limits. The combination of these technologies opened new possibilities for reducing the impact of vehicle pollution on urban environments. Rana et al. (2022) aligned their research with the United Nations' agenda for improved pollution detection and sustainable urban living. Their IoT architecture, complete with sensor nodes, gateways, and Long-Range communication, brought a new level of sophistication to air quality monitoring. By customising hardware and reducing data redundancy, this system offered real-time monitoring of air quality indicators. The evaluation metrics, including bit rate, receiver sensitivity, and time on air, ensured data accuracy. The study highlighted the significance of optimising IoT hardware for comprehensive and accurate air quality assessments.

Shakhov & Sokolova (2019) brought a unique perspective by introducing the concept of mobile sensors mounted on vehicles for air pollution monitoring. By distinguishing between deterministic and Poisson traffic flows, the study focused on detecting elevated pollution levels when vehicles were present. This approach introduced dynamism to air pollution data collection, improving the understanding of pollution sources and variations. Kumaresan et al. (2021) underscored the grave health concerns related to air pollution. The introduction of an IoT roadside air pollution monitoring system, coupled with a mobile app, provided an effective solution for real-time monitoring and data sharing. This approach could make a significant impact in high-traffic areas and signal zones prone to pollution spikes, enhancing public health protection and city planning. Al-Dweik et al. (2017) introduced a modular Road Side Unit (RSU) that harnessed IoT technology to collect data from various sensors and cameras mounted on vehicles. This data-driven approach opened new avenues for real-time traffic management, such as speed limit adjustments and traffic congestion reduction. The addition of weather advisory warnings aligned with the broader goal of reducing emissions and improving road safety. Pal et al. (2018) concentrated on real-time pollution monitoring and

personalised air quality information for individual vehicles. Their IoT-based solution was valuable for addressing vehicle-generated pollution in urban areas. The approach ensured that drivers and relevant authorities had access to real-time data, contributing to a better understanding of pollution sources.

Goyal et al. (2018) introduced a prototype for vehicle-mounted sensing of $PM_{2.5}$ and PM_{10} . This innovative technology addressed real-world challenges, enabling the creation of dense air pollution data. The system provided insights into the root causes of air pollution and offered potential solutions for managing and mitigating pollution in urban areas. Mateichyk et al. (2020) explored the impact of modern transport infrastructure and the use of IoT for roadside pollution monitoring. The information and analytical system they introduced combined various advanced technologies to assess and forecast the influence of different factors on roadside pollution. This technology held the potential to provide essential insights for optimising transport infrastructure and ensuring sustainable urban development. Rushikesh & Sivappagari (2015) introduced an IoT-based roadside air pollution monitoring system with a mobile app. Their emphasis on real-time monitoring and data sharing facilitated public awareness and engagement. By focusing on high-traffic zones and signal areas, this technology ensured that pollution spikes were promptly detected and addressed. These studies were able to highlight the potential of real-time data collection and analysis in creating healthier urban environments.

2.2.3 Barriers To The Adoption of IoT For TRAP Monitoring

The adoption of IoT for TRAP monitoring faces several barriers, despite its potential to revolutionise environmental data collection and management. One of the primary challenges is the high initial cost of implementing IoT infrastructure. Deploying a network of sensors and devices across urban areas, along with the necessary data storage and processing infrastructure, can be financially demanding for municipalities and local authorities (Idrees &

Zheng 2020). The cost of maintenance and system upgrades over time further compounds the financial burden. Data privacy and security concerns present another significant barrier. IoT systems collect vast amounts of sensitive data, including location information and potentially personally identifiable information. Protecting this data from unauthorised access and cyber threats is a crucial consideration. Ensuring robust data encryption, secure data storage, and adherence to privacy regulations becomes paramount, which can be complex and resource-intensive.

Interoperability and standardisation challenges are also notable impediments to IoT adoption in TRAP monitoring. There is a lack of uniform standards for IoT devices and platforms, making it difficult to integrate various sensor types and data sources seamlessly (Toma et al. 2019). Without standardised protocols, data compatibility and system interoperability become challenging, limiting the effectiveness of IoT systems in addressing TRAP concerns. Scalability and network coverage are additional barriers. The deployment of IoT sensors must cover extensive urban areas to provide comprehensive TRAP data. However, expanding and maintaining network coverage can be a complex undertaking, especially in densely populated and geographically diverse cities. Ensuring uninterrupted network connectivity and sensor functionality is a logistical challenge.

Data management and analytics pose another barrier. IoT systems generate vast amounts of data. Efficiently storing, processing, and analysing this data require advanced data management infrastructure and expertise. Without the capability to manage and extract meaningful insights from the data deluge, the potential benefits of IoT for TRAP monitoring remain unrealised. Moreover, community acceptance and participation are essential for IoT adoption in TRAP monitoring. Ensuring public trust and involvement is a hurdle that requires community outreach and education efforts. In addition, regulatory and compliance issues present challenges for IoT adoption. Governments and municipalities must navigate a

complex web of regulations and compliance requirements, particularly in the realm of data protection and environmental monitoring. Ensuring that IoT systems adhere to relevant regulations while still maintaining their effectiveness is a delicate balance.

2.2.4 Big Data And TRAP Monitoring

The relevance of big data in the context of TRAP monitoring is of paramount importance, especially in today's rapidly urbanising world. As cities continue to expand, the volume of traffic and the associated pollution intensifies. Big data, characterised by its immense volume, velocity, and variety, plays a pivotal role in both comprehending and addressing the impact of highway pollution, specifically TRAP. First and foremost, the vast amount of data generated from various sources, including traffic sensors, satellite imagery, weather stations, and air quality monitoring stations, provides a comprehensive and real-time view of urban environments. This data serves as a powerful tool in understanding the intricate dynamics of TRAP. By amalgamating diverse data streams, researchers and policymakers can identify trends and patterns that would be impossible to discern from traditional monitoring methods alone. Furthermore, big data facilitates spatial analysis, allowing for a granular understanding of TRAP hotspots and vulnerable areas. This insight is invaluable in designing targeted mitigation strategies, such as the placement of green infrastructure or the rerouting of traffic away from sensitive locations. By harnessing the geospatial aspects of big data, cities can optimise resource allocation and focus their efforts where they are most needed.

In addition, big data offers an opportunity for citizen engagement and crowd-sourced data collection. Mobile apps and online platforms enable residents to report TRAP incidents, contributing real-time data for monitoring. This citizen science approach not only improves data accuracy but also fosters community awareness and participation in TRAP management. The interconnectedness of big data systems allows for collaborative efforts between cities, regions, and even countries. Shared data and experiences can lead to more effective regional

TRAP control strategies and the standardisation of monitoring and reporting practices. This global perspective is vital in addressing trans-boundary TRAP issues and striving for more sustainable, cleaner urban environments. Also, big data contributes to transparency and accountability in TRAP management. By making TRAP data accessible to the public, it empowers citizens to hold authorities and industries accountable for their pollution levels. This transparency can stimulate action and encourage the adoption of cleaner technologies and practices, ultimately driving positive change in urban TRAP dynamics.

2.2.5 Existing Applications of Big Data For TRAP Monitoring

Reddy et al. (2021) presented a comprehensive system for monitoring ambient air quality on roads, with a focus on identifying vehicles emitting pollution beyond predefined limits. Their approach integrates IoT, electrochemical toxic gas sensors, RFID, and big data techniques including stream processing and data visualisation to monitor air contamination patterns. Their proposed framework was designed to also aid in intelligent traffic light control to reduce emissions, demonstrating the potential of Big Data in traffic pollution management. The study of Mateichyk et al. (2020) recognised the need for digitisation in traffic management and utilised artificial intelligence, Big Data, and predictive analytics to develop an information and analytical system for monitoring roadside pollution due to traffic flows. This innovative approach leverages various technologies to assess and forecast roadside pollution, thereby contributing to efficient transport infrastructure development. In another investigation, Wang & Huang (2017) sought to tackle air pollution issues around metropolitan areas in China. They designed an Environmental Monitoring Vehicle to collect data on air pollutants, utilising GPS, IMU, and a range of sensors. Big Data analysis, including spatial correlation and spatial clustering, allowed for the creation of periodic reports on roadside air quality. This approach demonstrates the promise of Big Data in enhancing air quality monitoring, especially in densely populated urban areas.

El Fazziki et al. (2017) proposed an agent-based system for modelling urban road network infrastructure, real-time air pollution indexes, and dynamic traffic regulation. By integrating agent technology with machine learning and Hadoop-based frameworks - HBase and MapReduce, they aimed to optimise traffic management and reduce vehicle emissions. Their system holds the potential to significantly improve air quality in urban environments through real-time monitoring and regulation. In yet another study, Cecilia et al. (2018) introduced a high-throughput hardware-software infrastructure for gathering information from vehicles to provide novel Intelligent Transportation Systems (ITS) services. By parallelising fuzzy clustering techniques on CPUs and GPUs, they efficiently processed vehicle data to identify highly polluting traffic areas and drivers. This approach showcases the applicability of Big Data in creating smart ITS services for better traffic and air quality management. Apte et al. (2017) recognised the limitations of conventional fixed-site pollution monitoring methods and employed a novel approach. Equipping Google Street View vehicles with pollution measurement platforms allowed for highly detailed urban air quality mapping. By utilising Big Data, they achieved greater spatial precision and revealed pollution patterns that significantly impact public health and environmental equity. Chang (2019) used Big Data-oriented Social Network Analysis to analyse event co-occurrence and spatial correlation characteristics of pollution scenarios at monitoring stations. This data-driven approach improved understanding of regional high pollution characteristics, providing valuable insights for real-time air quality management and pollution precaution.

In another context, Tarek et al. (2018) focused on efficiently analysing air quality data using large-scale data mining techniques. By applying clustering methods and time-series analysis to air pollution data, they successfully identified pollution hotspots and temporal pollution trends. This approach has the potential to provide more dynamic and efficient insights into air pollution patterns. Sridhar et al. (2022) proposed an IoT-based air quality monitoring system. They utilised air sensors and data processing to detect harmful gases and

provide real-time air quality information. The system, which can display air quality based on standard criteria, offers essential information for assessing air pollution and taking preventive measures. N. Genikomsakis et al. (2018) harnessed low-cost sensors to develop a portable air pollution monitoring system focused on $PM_{2.5}$. Their on-field testing demonstrated the system's accuracy and potential for collecting spatio-temporal $PM_{2.5}$ profiles. This innovation contributes to fine-detailed air quality monitoring in support of intelligent transportation systems.

In this study, implementing best practices from recent research, such as the utilisation of advanced IoT systems and big data processing, as shown in studies like those by Reddy et al. (2021) and El Fazziki et al. (2017), is essential. A successful approach to consider would be the application of integrated systems combining IoT sensors with AI-driven analytics for traffic pollution management. In addition, adopting technologies for real-time data processing and visualisation, similar to the systems developed by Sridhar et al. (2022), can further enhance the effectiveness of pollution monitoring. These technologies can facilitate more accurate and timely responses to fluctuating pollution levels, thereby supporting the development of a robust, scalable forecasting model as well as demonstrating the complexities of data integration in environmental monitoring.

2.2.6 Barriers To The Adoption of Big Data For TRAP Monitoring

The implementation of big data for TRAP monitoring is not without its challenges. These challenges must be addressed to fully leverage the potential of big data in understanding and mitigating TRAP in urban environments. Data privacy and security concerns are significant. Handling sensitive air quality and location data raises worries about data breaches and misuse, potentially hindering data sharing and collaboration. Protecting this data through robust encryption, access controls, and adherence to data protection regulations is imperative. Data quality and standardisation issues also pose challenges. Inconsistencies in data

quality and format across various sources can make data integration and analysis difficult (Krogstie 2015). Establishing standards and protocols for data collection and sharing is essential to ensure the reliability of data.

Access to high-speed internet and reliable connectivity is crucial for real-time TRAP monitoring. However, inadequate infrastructure, particularly in developing regions, can limit the effectiveness of big data solutions. Expanding infrastructure and ensuring connectivity in remote areas is a pressing concern. The cost and resource constraints associated with implementing big data solutions can be a barrier, especially for smaller municipalities and organisations with limited budgets. Funding is needed to invest in sensor networks, data centres, and maintenance. A shortage of expertise in data science and analytics presents another challenge. Specialised skills are crucial for effective big data analysis, but a lack of data experts and limited training opportunities can hinder progress. Integrating data from diverse sources, including traffic sensors, meteorological stations, and air quality monitoring stations, can be complex due to data silos and incompatible formats. Creating a unified dataset for analysis is a challenge that needs to be addressed.

As data volumes continue to grow, scalability becomes a pressing issue. Expanding big data infrastructure to accommodate increasing data loads is a significant undertaking that requires careful planning and investment. The complex regulatory landscape surrounding data sharing, privacy, and compliance can create obstacles for cross-border data exchange. Navigating these regulatory hurdles is essential for effective TRAP monitoring. Public awareness and acceptance of data collection and monitoring initiatives can significantly impact their success (Conrad & Hilchey 2011). Transparency, public education, and gaining public trust are essential to ensure the cooperation of residents. Ethical and legal considerations related to data ownership, consent, and ethical use also need to be addressed. Establishing clear ethical guidelines is essential to ensure the responsible use of data in TRAP monitoring.

2.3 Traffic-Related Air Pollution Forecasting - Enabling Technologies and Methods

This section highlights the applications of chemical transport models and deep learning in TRAP forecasting, examining how these technologies model the complex interactions between various pollutants and environmental factors. It also discusses the barriers to their full adoption, including computational demands, the need for extensive training data.

2.3.1 Chemical Transport Models And TRAP Forecasting

Chemical Transport Models (CTMs) are complex and comprehensive tools used in the field of atmospheric science and environmental research. These models are designed to simulate the dispersion and behaviour of air pollutants in the atmosphere. CTMs serve a critical role in understanding how pollutants are transported, diffused, and transformed in the air, making them indispensable for assessing air quality and addressing environmental and public health concerns (Ward 2019). A key element of Chemical Transport Models (CTMs) is the inclusion of emission sources. These models account for emissions from both natural and anthropogenic sources. To do this, emission inventories compile comprehensive data on the types and amounts of pollutants emitted into the atmosphere, serving as crucial inputs for CTMs. In addition, CTMs simulate the movement and dispersal of these pollutants in the atmosphere, incorporating meteorological data such as wind speed, wind direction, temperature, and atmospheric stability to enhance the accuracy of their predictions. These meteorological factors have a significant impact on how pollutants move through the air. By integrating high-resolution meteorological data from sources like weather stations and numerical weather prediction models, CTMs can provide accurate simulations of how pollutants spread over time and space (Gariazzo et al. 2020).

A crucial aspect of CTMs is their ability to account for chemical reactions in the at-

mosphere. Pollutants are not static; they can undergo transformations through chemical processes, resulting in the formation of secondary pollutants. CTMs capture these chemical reactions, providing insights into the changing composition of the atmosphere (Cui & Wang 2021). This is particularly important for understanding the complexities of air quality and the formation of various air pollutants, including ozone, particulate matter, and secondary aerosols. To facilitate their modelling work, CTMs adopt a grid-based approach (Karamchandani et al. 2011). This involves dividing the study area into a grid of cells, both horizontally and vertically. Each grid cell represents a specific location within the domain being studied. The use of grids allows for the spatial and temporal discretisation of the atmosphere, which is critical for modelling air quality at different scales. CTMs can, therefore, be applied to assess air quality in small urban areas, larger regional settings, and even global domains.

2.3.2 Existing Applications Of Chemical Transport Models For TRAP Forecasting

The study of Pohjola et al. (2007) conducted a field measurement campaign near a major road in Helsinki, focusing on aerosol measurements. It compared measured concentration data with predictions from the road network dispersion model CAR-FMI in combination with the MONO32 aerosol process model. The study evaluated vehicular exhaust emissions, atmospheric dispersion, and particle transformation within a distance of 200 meters. The most critical process affecting particle concentrations was atmospheric dilution, with coagulation and condensation playing a minor role. Condensation was found to affect particle diameter in the smallest modes. Modelling with 109–1010 molecules of condensable organic vapor yielded the closest match to measured values. In a similar study, Masood et al. (2017) modelled CO emissions from Mathura road in New Delhi, India using the CALINE4 model, employing emission factors and vehicle classification methodology. Predicted and monitored CO concentrations at the receptor location showed fair agreement, and the model’s performance was assessed, with a root mean square error (RMSE) of 302 and a correlation coefficient (r) of

0.87. Predicted CO emissions were integrated with ArcGIS to generate digital maps, identifying pollution hot spots for policy formulation and environmental impact assessment. Lin & Ge (2006) introduced an alternative approach, combining a cell-transmission traffic propagation model with a Gaussian dispersion model to estimate traffic emissions and roadside air quality. This model captured time-dependent vehicular traffic characteristics and highlighted high carbon monoxide concentrations at intersections during off-peak traffic hours.

The study of Beevers et al. (2012) investigated coupling CMAQ and ADMS air quality models for predicting NO_X , NO_2 , and O_3 concentrations in London. The model showed reasonable agreement with monitoring data from 80 sites. The study pointed out the need for improved prediction of road transport-related NO_X emissions, especially hourly scaling. The combination of regional and local scale models demonstrated promise as a tool for policy development and epidemiological studies. Mishra et al. (2016)'s study investigated gaseous emission dispersion from urban roadside sites in Delhi. It used the General Finite Line Source Model (GFLSM) to predict CO, NO_2 , and SO_2 concentrations, showing a high level of agreement between observed and predicted values. The model exhibited reasonable prediction capabilities for gaseous pollutant dispersion from on-road vehicles in an urban environment. In another study, Heist et al. (2013) developed a GIS-based air pollution model, STEMS-Air, for PM_{10} , offering daily and annual predictions. The model performed well in predicting concentrations of PM_{10} , making it a valuable tool for air pollution mapping in urban areas. Gulliver & Briggs (2011) focused on air pollution mapping for short-term exposure studies using the STEMS-Air model for PM_{10} . The model achieved good agreement with observed concentrations, making it suitable for air quality planning, health risk assessment, and epidemiological studies. In related research, Stocker et al. (2019) conducted a model inter-comparison with four dispersion models, evaluating their abilities to capture near-road pollutant dispersion using experimental datasets. The results highlighted the challenges and uncertainties associated with near-road air dispersion modelling. This study shed

light on the complexities of modelling near-road exposures to traffic-generated pollutants, emphasising the need for validation and uncertainty reduction in dispersion models.

The referenced studies highlight the importance of integrating detailed modelling techniques to accurately assess urban air quality, which presents a good practice worth adopting in this research. Particularly, the combination of traffic data with air quality models, as demonstrated in the works of Lin & Ge (2006) and Beevers et al. (2012), offers a robust method for understanding the fluctuations in pollutant levels. However, these studies also expose a significant gap in the fine-tuning of emission factor calculations and their temporal resolution. Enhancing the granularity of emission factors, especially on an hourly basis, could substantially improve the predictive accuracy of the models. Addressing this gap in the current study could lead to more precise strategies for air pollution control and provide actionable insights for urban planning and public health policy.

2.3.3 Barriers To The Adoption of CTM Tools For TRAP Forecasting

ADMS-Roads and ADMS-Urban, integral components of the Atmospheric Dispersion modelling System (ADMS) suite, developed by Cambridge Environmental Research Consultants (CERC), represent advanced CTMs with their unique advantages and constraints. One prominent limitation inherent in these models is the simplification of complex physical and chemical processes (Liang et al. 2023). While these models offer precision in various scenarios, they rely on simplifications that may not encompass the full complexity of real-world atmospheric interactions (Cha et al. 2023). Consequently, these simplifications introduce a degree of approximation into the model's predictions. The reliance on assumptions and generalisations is another limitation (Suleiman et al. 2019). ADMS-Roads may assume uniform emissions from vehicles, which may not always align with the diverse and dynamic nature of real-world traffic conditions. These simplifications can impact the precision of the model's output, particularly in scenarios with varying traffic patterns. ADMS-Roads is designed for

a specific spatial scope, primarily focusing on the air quality near roads and highways. This means that its application is not well-suited for assessing air quality over larger areas, such as regional or global scales (Lv et al. 2020). Consequently, the model's utility is constrained to the immediate vicinity of roadways. The accuracy of ADMS-Roads is highly contingent on the quality and precision of input data. Errors or uncertainties in input data, including emission inventories and meteorological information, can significantly affect the reliability of the model's predictions. Ensuring accurate and up-to-date input data is critical for obtaining dependable results. The model may also lack a high level of vertical resolution, primarily focusing on the horizontal dispersion of pollutants. This limitation can be particularly relevant in areas with tall buildings or complex terrain where understanding vertical dispersion is crucial (CERC 2020).

Modelling air quality in areas with complex terrain, such as hilly or mountainous regions, can present challenges for ADMS-Roads. The model may not fully capture the intricate effects of terrain on pollutant dispersion, potentially leading to less accurate results in such areas (Horizon Nuclear Power 2018). Accurate calibration and validation of the model are vital, and the lack of precise validation data can compromise its reliability. Users must carefully validate the model with observed data to ensure the accuracy of its predictions. ADMS-Roads may not account for rapid changes in traffic conditions, such as temporary road closures, traffic congestion, or shifts in vehicle types. These variations in traffic patterns can influence emissions and dispersion patterns, and the model's inability to adapt quickly to such changes is a limitation. When it comes to modelling non-exhaust emissions, such as tire and brake wear, ADMS-Roads may face challenges. These emissions are not as well-documented as exhaust emissions, and modelling them accurately can be complex due to variations in wear rates and limited data availability (CERC 2020). In addition, real-time data assimilation can be difficult to implement. It requires continuous access to accurate real-time data, which may not always be readily available in all regions where the model is

applied.

2.3.4 Deep Learning And TRAP Forecasting

Deep learning has emerged as a transformative approach in the field of TRAP forecasting, providing significant advancements over traditional modelling techniques. The relevance of deep learning in TRAP forecasting is underscored by its ability to handle complex, high-dimensional data, capture non-linear relationships, and provide accurate, scalable solutions for real-time air quality management. Air pollution dynamics are inherently non-linear, driven by a multitude of interacting factors such as traffic volume, weather conditions, and human activities. Traditional linear models may fail to capture these intricate relationships, leading to less accurate predictions. Deep learning models, however, are designed to model non-linear relationships through their complex network structures. Deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can identify and learn from patterns in spatial and temporal data, respectively. This enables deep learning models to capture the underlying processes driving air pollution, resulting in more reliable forecasts. Accurate TRAP forecasts generated by deep learning models provide essential information for policymakers, urban planners, and public health officials. By understanding the predicted patterns of air pollution, these stakeholders can make informed decisions about traffic management, urban development, and public health interventions. For example, forecasts can inform the implementation of traffic restrictions, the placement of green spaces, or the timing of public health advisories. The ability of deep learning models to provide detailed, location-specific predictions enhances their utility in supporting targeted and effective decision-making.

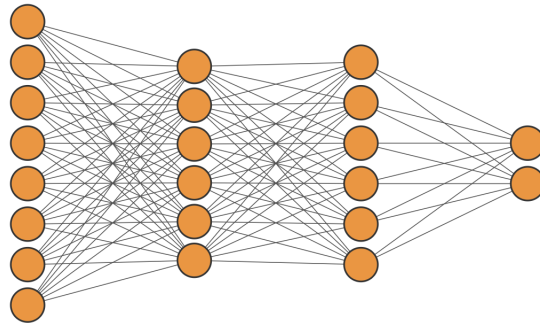
2.3.5 Deep Learning Architectures

Recent theoretical results inspired by the brain and cognition show that to be able to learn some high-level representation of complex functions involved in complicated forecasting tasks,

one may have to implement *deep* architectures (Bengio 2009). This section describes some of the common Deep Learning architectures relevant to TRAP forecasting.

2.3.5.1 Deep Neural Network (DNN)

A Deep Neural Network is typically a standard neural network with “depth”. The depth of a neural network is determined by the number of hidden layers (second and third layers in Figure 2.1) between the input and output layers. Even though no threshold determines when a neural network can be identified as “Deep”, most researchers have agreed that a CAP (Credit Assignment Path) depth > 2 can be considered “Deep” while Schmidhuber (2015) considers CAP > 10 to be very deep learning. DNNs are trained to model complex non-linear relationships by extracting uniquely abstract features that help improve its performance. Each layer of its multi-layered composition is dedicated to a particular feature identification (Carreira-Perpiñán & Hinton 2005).



Input Layer $\in \mathbb{R}^8$ Hidden Layer $\in \mathbb{R}^6$ Hidden Layer $\in \mathbb{R}^6$ Output Layer $\in \mathbb{R}^2$

Figure 2.1: Deep Neural Networks(DNNs) Architecture

2.3.5.2 Convolutional Neural Network (CNN)

CNNs are widely used for image processing applications (Krizhevsky et al. 2012). The architecture came into limelight after the results of AlexNet (A deep learning network used for image classification) at the ImageNet competition (Krizhevsky et al. 2012). Unlike con-

ventional MLPs, CNN neurons are arranged in a way that matches the width, height and depth of images. In addition to input layers, output layers and activation functions, CNNs particularly have two additional layers, the convolution and pooling layers (depicted as the second to fourth hidden layers in Figure 2.2). The convolution layer convolves the image by using different convolutional filters and shifting the receptive fields gradually. It is common practice to insert a pooling layer between successive convolutional layers. The pooling layer, on the other hand, reduces the size of the output from the convolution layer by calculating the mean, max, median or other statistical features of the image at different pixels.

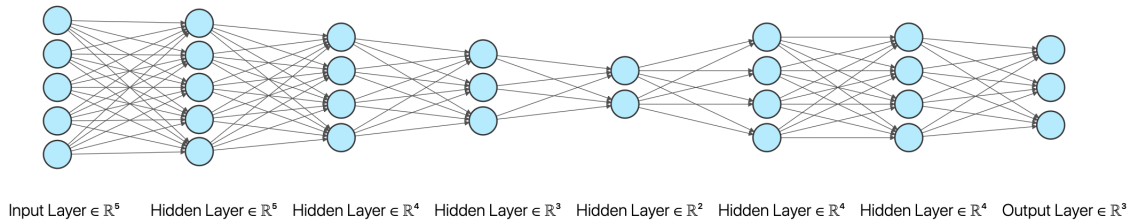


Figure 2.2: Convolutional Neural Networks(CNNs) Architecture

2.3.5.3 Recurrent Neural Network (RNN)

RNNs are best suited for handling sequential data. They outshine other forms of deep learning when processing time-dependent information (Mikolov et al. 2010). Parameters across different time steps are shared based on sequential data properties. RNNs are mostly applied in video and speech processing since they can keep information on a previously processed audio chunk or video frame in order to make predictions of successive data. A RNN's output y_t at any time t is dependent not only on input x_t but also on x_{t-i} at times $t-i$. Like other deep learning architectures, RNNs can also be trained using the backpropagation algorithm. More specifically, a backpropagation variant – Back Propagation Through Time (BPTT), is the standard training algorithm for RNNs (Werbos 1988, Schmidhuber 2015). A sample of RNN architecture is shown in Figure 2.3.

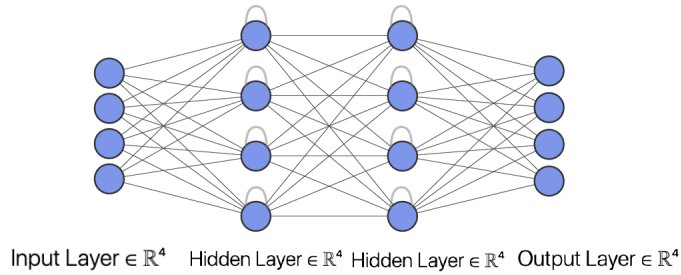


Figure 2.3: Recurrent Neural Networks (RNNs) Architecture

2.3.5.4 Auto-Encoder (AE)

Auto-Encoders(AEs) are mainly used for data denoising and dimensionality reduction (Alain & Bengio 2014, Wang et al. 2016). Unlike other MLPs, AEs extract features from the input layer with the aim of replicating the same input data in the output layer. AEs involve an encoding and decoding process which forces the network to ignore the noisy part of the input and instead focus on encoding/representation of the more informative segments. The output layer in AEs has the same dimension(number of nodes) as the input layer(illustrated in figure 2.4) aimed at replicating the input data rather than having to predict Y given X like in most MLPs. The hidden layer plays a vital role by ensuring that the network actually learns the features of the input data and not just output the same version of the input data.

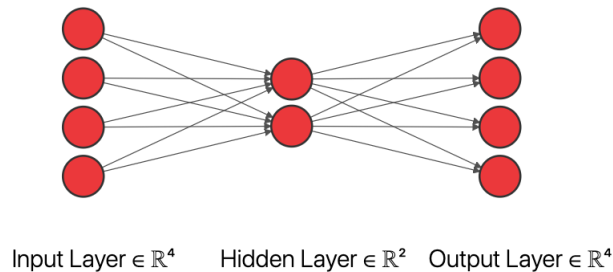


Figure 2.4: Autoencoders (AEs) Architecture

2.3.6 Existing Applications of Machine Learning Methods For TRAP Forecasting

Catalano & Galatioto (2017)'s research proposed an innovative Machine Learning approach to enhance urban air pollution forecasting. The study introduced the concept of a self-adaptive model, capable of selecting the most suitable prediction model from a range of alternatives, ultimately leading to more precise predictions, especially for extreme pollution events. The research's real-world application was tested in the Greater Manchester Area, UK, focusing on nitrogen dioxide concentration prediction. The primary discovery was that the self-adaptive approach outperformed standard statistical methods and artificial neural networks by up to 27% and 113%, respectively, for extreme air pollution event predictions. Meng et al. (2021) aimed to address the health concerns related to traffic pollution near residential buildings. To achieve this, the research delved into the spatial distribution of PM_{2.5} concentration emanating from road traffic emissions. A spatial distribution model for estimating PM_{2.5} concentration (SDC) was developed using Machine Learning techniques. Subsequently, the impact on residents' life expectancy was evaluated using this model. In another study, Wu et al. (2022) investigated the environmental impacts of elevated roads, particularly their effects on air quality in Shanghai, offering novel insights. Notably, the study employed a Long Short-Term Memory (LSTM) model to predict air quality, considering the daily periodicity of pollutants.

The study of Lozhkin et al. (2016) introduced a pioneering differential neural network model for estimating CO emissions' dispersion near highways, providing practical applications. The approach optimised the model based on simulated and experimental measurements, enabling the prediction of emergencies when parameters like wind speed, direction, and fire conditions change. In similar research, Pandey et al. (2013) addressed submicron particle prediction in Hangzhou using Machine Learning. The standout feature of the research was the comprehensive use of Machine Learning techniques to predict PM_{1.0} and ultrafine

particle (UFP) levels, improving our understanding of their relationships with meteorological and traffic variables. The study's key discovery was that tree-based classification models, such as Alternating Decision Trees and Random Forests, were highly effective in predicting PM1.0 and UFP levels, suggesting their potential for air quality forecasting. It underlined the importance of systematically collecting and analysing datasets using Machine Learning in predicting submicron-sized ambient air pollutants. Wang et al. (2020) used two years observation data from Shanghai's roadside air quality monitoring stations to develop an air quality forecasting model. The study uncovered distinct daily and weekly variations in fine particulate matter (PM2.5) and carbon monoxide (CO) concentrations, closely linked with traffic patterns. The study of Suleiman et al. (2016) focused on air quality modelling through Machine Learning, specifically predicting particle mass concentration and particle number counts. The research highlighted the effectiveness of Boosted Regression Trees (BRT) and Artificial Neural Networks (ANN) for air quality prediction, providing valuable tools for air quality management. The primary contribution was the emphasis on model interpretation, with BRT offering an advantage in this aspect. It introduced an approach that balances prediction accuracy and model interpretability, aiding in the decision-making process for air quality management.

Fong et al. (2020) utilised Long Short-Term Memory (LSTM) Recurrent Neural Networks for air pollution prediction in Macau. The distinctive element was the application of transfer learning to enhance prediction accuracy, making the models more efficient. The LSTM models proved highly accurate and demonstrated potential in improving air quality forecasting. This research provided a valuable tool for urban planning, public health management, and environmental quality monitoring, contributing to better air quality. In another study, Hashad et al. (2021) study explored machine learning (ML) methods to predict particle concentrations downwind of various vegetation barrier designs. Data from 83 computational fluid dynamics simulations were employed to train and test ML models. Notably, random

forest (RF), neural networks (NN), and XGBoost (XGB) models excelled, achieving low normalised root mean square errors and high R2 values, outperforming support vector machine (SVM) and linear regression (LR) models. The study emphasised the importance of vegetation dimensions and particle size in predicting pollutant concentrations.

While the individual contributions of these studies are impressive, there remains an underutilised potential in synthesising these varied approaches for a holistic prediction model. In particular, integrating models like the LSTM used in Fong et al. (2020) for capturing temporal patterns in pollution data, alongside spatial distribution techniques as applied in Meng et al. (2021), could offer a more comprehensive understanding of urban air quality dynamics. By adopting these advanced techniques, this study aims to not only predict but also actively manage air quality through informed, data-driven decisions. Thus, the current research initiative is focused on integrating these innovative Machine Learning strategies to develop a robust, adaptable air quality forecasting framework tailored to complex urban environments.

2.3.7 Barriers To The Adoption of Machine Learning Methods For TRAP Forecasting

The implementation of machine learning and deep learning techniques for predicting traffic-related air pollution encounters multiple significant obstacles. A primary issue is the quality and volume of the required data. Predictive models depend on large, high-quality datasets that include variables such as air pollution levels, and traffic flow. Yet, the collection of such detailed and precise data frequently faces challenges due to inadequate monitoring infrastructure and the limited availability of comprehensive historical data in various regions. Merging data from varied sources like traffic sensors, air quality monitors, and meteorological stations also introduces significant challenges (Zheng et al. 2013). Issues such as data formatting, standardisation, and alignment must be tackled to facilitate accurate model development, a process that can be notably time-intensive. Additionally, the computational requirements of

deep learning models present a further obstacle. These models demand substantial computing power and typically require advanced hardware like GPUs for efficient operation. This high resource dependency means that smaller organisations or areas with restricted computational resources might find it difficult to adopt and sustain the necessary infrastructure for these advanced predictive models.

Model complexity and opacity are also significant hurdles. Deep learning models, often described as “black boxes”, can be challenging to interpret, making it difficult to understand their decision-making processes. This lack of transparency may hinder their adoption, particularly by regulatory bodies and policymakers who require clear, interpretable models for decision support. Training deep learning models is a specialised skill, demanding expertise in hyperparameter tuning, network architecture design, and addressing issues like overfitting. Securing individuals with this expertise, such as data scientists or machine learning engineers, can be challenging in some areas. Privacy and security concerns associated with sensitive data, such as personal information from traffic cameras or air quality monitoring stations, must also be addressed. Striking a balance between data privacy and security while enabling its use for machine learning applications is a complex task (Wachter 2018). Regulatory compliance is paramount, as implementing machine learning and deep learning models for air pollution prediction may require adherence to environmental regulations and standards. Ensuring that these models comply with the relevant rules can be a substantial barrier, especially in regions with strict environmental regulations. Resource constraints, particularly in terms of funding and skilled personnel, can hinder the development, implementation, and maintenance of machine learning models. Smaller municipalities or regions may struggle to overcome these limitations. Scalability is another pressing concern as regions grow and urbanise. Ensuring that machine learning models can adapt to changing traffic patterns, urban expansion, and evolving pollution sources is crucial for their long-term adoption. Ensuring the accuracy and reliability of predictions is also paramount. While machine learning models

have the potential to improve prediction accuracy, they are not infallible. Ongoing calibration and validation are necessary to ensure that predictions meet the necessary standards for decision-making.

2.4 Chapter Summary

The chapter extensively explores the role of digital technologies in managing traffic-related air pollution, emphasising IoT, Big Data, and Machine Learning as crucial components. These technologies have been instrumental in enhancing the monitoring and forecasting of traffic pollution through various applications and innovations highlighted throughout the studies. Key findings include the transformative potential of IoT in creating interconnected networks of sensors that provide real-time, detailed environmental data, thus enabling precise pollution monitoring and management strategies. However, significant barriers such as the integration of diverse data sources and ensuring the quality and granularity of data have been identified. These challenges underscore the gaps in current practices, particularly in data standardisation and the application of machine learning to interpret complex environmental datasets. The hurdles related to data privacy, security, and the computational demands of processing large datasets also present limitations that could impede the adoption of these technologies on a wider scale.

Moving forward in this study, the plan is to leverage the insights gained from these technologies to develop a more integrated approach to traffic pollution management. This will involve enhancing data collection methods to improve the quality and accuracy of the data used for pollution forecasting. In addition, advancing the application of machine learning algorithms will be crucial to effectively analyse and interpret the integrated data, thereby improving the accuracy of pollution forecasts. These steps will aim to bridge the identified knowledge gaps, particularly in data integration and machine learning application, to foster a more effective and comprehensive traffic pollution management system.

Chapter 3

Research Methodology

3.1 Chapter Overview

This chapter embarks on an exploration of the research theory, design principles, and philosophy that form the foundation of this study. It commences with a general overview of these fundamental concepts before delving into the rationale and motivations guiding their selection within the specific context of this study.

3.2 Research Theory and Design Principles

3.2.1 Research Theory

Research involves gathering knowledge and collecting facts about unknown occurrences around us. Understanding what knowledge is and adopting a philosophical perspective on these occurrences is essential before embarking on a research project. Everyone has a conscious or unconscious philosophical perception of their daily life, although some argue against having theoretical standpoints, believing that new theories often override old ones, causing confusion. However, this view is flawed, as philosophical stances enhance research arguments and make them explicit. Philosophical viewpoints help recognize fundamental assumptions in research, allowing for accurate analysis of methods and the authenticity of findings.

Empiricists and rationalists represent two foundational philosophical perspectives that influence research methodologies and knowledge interpretation. Empiricists argue that knowledge is derived from sensory experience, emphasizing evidence and observation. This approach is central to scientific methods, which rely on quantitative measurements and systematic observation. Rationalists, in contrast, believe that reason and intellect are the primary sources of knowledge, independent of sensory experience. They argue for innate ideas embedded in the human mind at birth and emphasize logical reasoning as the basis of all knowledge. These perspectives highlight the relationship between research and theory, which is further discussed in subsequent sections.

3.2.1.1 Inductive Research Method

The word “theory” has several meanings and is often used in diverse ways. The widely accepted definition, however, depicts it as a way of explaining observations. Researchers often find it challenging to relate the abstractness of a theory to the actual world. It is crucial to understand that scientific theory development is unachievable if observed information cannot be explained (Chibucos et al. 2005). As suggested by Boss et al. (1993), independent empirical information must be used to test conceptual ideas. Inductive research starts with observations that are fed into a series of theories within a domain to determine which theory supports the findings. Induction involves drawing conclusions from observations. After theoretical deliberation on collected data, a researcher may collect more data to further investigate whether a theory holds. This strategy of additional data collection to confirm a theory is known as *iterative* and is supported by the *grounded theory* (Glaser & Strauss 1967).

3.2.1.2 Deductive Research Method

Deductive research method, also known as the “top-down” approach is often based on a theoretical foundation from which several alternative hypotheses can be drawn (Johnson-Laird 1999). It is one of the shared perspectives on the interrelation between research and

theory. A research hypothesis is deduced and scrutinised based on known ideology about a particular field of interest (Eysenck & Keane 2015). Research starts with general assumptions which have to be proven through logical arguments (Walliman 2017). An example of such supposition is “*Every living thing must grow*”. So if an animal or plant is considered to be a living thing, it must grow at some point. The conclusion drawn is as a result of the theory about living things as well as several observations on the growth of living things. This is a typical example of a deductive argument. Although theories can be truthfully proven through observations, they can also be refuted through observations which do not conform with the hypothesis.

3.2.1.3 Abductive Research Method

Abductive research, sometimes known as *reduction* (Rotaru et al. 2014) is another approach that investigates the kind of inferences or best explanation that can be drawn after a set of observations (Sober 2001). In his first introduction of the idea of abduction, Peirce (1931) presented abduction as a reasoning process which yields explanations with respect to the effect and cause of a phenomenon. Abduction is seen as closely related to induction with the latter considered as a hypothetical argument resulting from the former (Fann 2012).

3.2.2 Research Design

A research work can be aimed at **Grouping** – categorising similar entities together, **Examining** – describing possible outcomes from observed situations, **Explaining** – making sense of findings from descriptive research, **Evaluating** – providing comparative judgements on the quality of happenings or objects, **Correlating** – investigating relationships between different phenomena or entities, **Comparing** – investigating contrasting differences between different or **Predicting** – suggesting possible future occurrences based on already known correlation between past events (Walliman 2017). These are often regarded as research objectives, and particular research can either have one or multiple objectives. Some of these objectives are

reliant on the successful completion of one another; for example, one should have properly evaluated an event before making predictions. Research design can be viewed as a framework that guides the creation of evidence that best suites set criteria and methods by a researcher (Kerlinger 1986). Several research designs are applicable based on the objectives of the research of interest.

3.2.2.1 Experimental Research Design

Experimental research design investigates the implications of manipulating the conditions of an event. For real experiments to be conducted, variations need to be made to an independent variable while its effects are observed on a dependent variable. This variation is known as the *manipulation* phase of an experiment. However, in most social experiments, non-numeric experimental variables are quite difficult to manipulate. For example, gender cannot be manipulated by replacing male with a female; neither can people be allocated to different social groups than the one they usually belong. A high level of social engineering is required to effect this kind of manipulation. Every experiment needs to fulfil an internal validity constraint which requires that an experimental result provides evidence to support the causal relationship between variables. There should be enough proof that a variable x actually causes the changes in a dependent variable y . An experiment which matches specific attributes of experimental design but does not meet all the internal validity prerequisites is known as a *quasi-experiment* design. The experimental design is considered to be a standard for judging quantitative research.

3.2.2.2 Comparative Research Design

Comparative design examines contrasting differences between two or more observations. Walliman (2017) highlights that this design type compares present and past happenings, especially when a researcher has little or no control over occurrences. Comparative design thrives in studies involving concept-building, theory-building and recognition of causal

processes (Bloemraad 2013). Schenker & Rumrill Jr (2004) in their study, specify that causal-comparative design mostly encompasses the use of already existing groups to identify differences within the group based on dependent variables. The research further points out that these dependent variables are usually not manipulable in ethical or practicable scenarios.

3.2.2.3 Case-study Research Design

Case study research design involves conducting a detailed and comprehensive investigation of a single entity or a small group of closely related entities. It is particularly useful when researchers seek to explore and understand complex real-world situations in depth. One of the defining characteristics of case studies is the focus on in-depth investigation. This means that it allow researchers to delve deeply into the specifics and intricacies of a particular case, which can be especially valuable when dealing with complex or unique phenomena. Another key feature of case study research is its emphasis on contextual understanding. Researchers not only examine the case itself but also consider the broader context in which it operates. This may include historical factors, cultural influences, and environmental elements that can provide a more comprehensive perspective.

3.2.2.4 Cross-sectional Research Design

Cross-sectional Design, also commonly known as survey design, is usually applied to research carried out to predict the outcome of interest for a particular population, mainly in issues relating to public health planning (Levin 2006). Cross-sectional research is characterised with attributes such as having multiple cases, quantitative method of data collection, non-manipulable variables and focuses on a single point in time.

3.3 Research Philosophy

At every step of research, assumptions are always made; either consciously or unconsciously (Burrell & Morgan 1979). These assumptions about the nature of reality or the development

of knowledge are known as research philosophies (Pietrobon & Dai 2012). The way research questions are formed, the methodology used to address these questions and the interpretation of discoveries are influenced by philosophical assumptions (Crotty 1998). In this section, epistemological, ontological and axiological types of research philosophies are discussed.

3.3.1 Epistemology

Coined from greek words *episteme* and *epistania*, meaning *knowledge* and *to understand or know* respectively, epistemology is the study of what counts as knowledge in the world (Cooksey & McDonald 2011). Epistemology focuses on the very foundation of knowledge – that is, its form, nature, mode of acquisition and how it can be transferred to others (Burrell & Morgan 1979). Or as posited by Schwandt (1997), epistemology is the study and justification of the nature of knowledge. A researcher can look to answer questions like *What counts as knowledge?*, *What relationship exists between me, as a researcher and acquired knowledge?*, *Can knowledge be personally experienced or needs to be acquired from one point of view?* *What relationship exists between the knower and the unknown?*. Answers to these sort of questions help researchers discover what is new with respect to what is already known (Kivunja & Kuyini 2017). It is essential to have an in-depth understanding of available epistemological assumptions in order to examine the weakness and cogency of subsequent research findings. Gray (2013) has identified three epistemological positions on knowledge gathering; namely *Objectivism*, *Subjectivism* and *Constructivism*

3.3.1.1 Objectivism

Objectivists believe that scientific methods require observable facts which are only accessible through unconcealed behaviours (Diesing 1966). From their point of view, objectivists maintain that social and physical phenomena exist independently and therefore seek to understand the social world through quantifiable and observable facts that lead to theories about social reality Thornhill et al. (2009). Objective researchers try as much as possible to

detach themselves from their beliefs throughout the research process by not involving their own sentiments and values. Nevertheless, Bunge (1993) highlight that objectivism is not entirely against subjectivism and sometimes involves studying people's subjective attitudes and values through objective lenses. Positivism is a research paradigm linked with objectivism (Gray 2013).

3.3.1.2 Constructivism

Constructivists argue that truth and meaning are not discovered but constructed based on the participant's association with the external world (Crotty 1998). Each participant formulates their own meaning to the same phenomenon based on personal experience (Hendry et al. 1999). As a result, this leads to a series of opposing but valid interpretations of the subject matter. According to Raskin (2002) constructivism is divided into three main categories: social, radical and psychological constructivism, with all three assuming the same epistemological stance that knowledge is not discovered but constructed by the human mind. Interpretivism is a research paradigm linked with constructivism (Gray 2013).

3.3.1.3 Subjectivism

Unlike constructivists, subjectivists believe that interpretations are not a result of the subject's interaction with the outside world but rather a result of the meaning obtruded on the object by the subject (Crotty 1998). While an objectivist researcher attempts to uncover facts and general laws regarding social happening, a subjectivist researcher is more concerned about discovering knowledge that can help constitute diverse social realities for different subjects (Thornhill et al. 2009). As opined by Rand (1990), subjectivists argue that conceptions do not match referents in the world and are just definitions.

3.3.2 Ontology

Ontology is the philosophical study of what constitutes reality (Scotland 2012). It focuses on the assumptions made to confirm the existence or nature of a social phenomenon being investigated. As opined by Scott & Usher (2010), ontology provides an interpretation of entities that constitutes the world, making it essential to a research paradigm. Researchers need to take a stance with regards their awareness of the reality of things and how they work (Scotland 2012). According to Rowland (1995), certain ontological assumptions lead to certain epistemological beliefs and vice versa. As argued by Bunge (1993), *realism* is an ontological assumption of the world as an independent entity which exists irrespective of our volition or formation of ideas about it, and that it can be known. Thornhill et al. (2009) identified two variants of realism: *critical realism* and *direct realism*. While critical realists assert that whatever we experience are not actual representations of the world and we may be deceived by our senses, direct realist argue that what we experience is the reality. *Relativism*, another ontological supposition, on the other hand, believes that reality is subjective and is effectuated by our senses (Guba et al. 1994). Idealism argues that reality is mentally constructed or immaterial. Idealists believe that the foundation of reality lies in the mind's perceptions and ideas. For idealists, our consciousness or subjective experiences don't just interpret a pre-existing world; they actually constitute it. Essentially, the very existence or essence of objects in the world depends on being perceived or conceptualised (Guyer & Horstmann 2015).

3.3.3 Axiology

Axiology addresses ethical issues that should be considered the process of research. According to Finnis (1980), axiology takes into consideration the philosophical approach to right decision making which is of value. Heron (1996) argues that human values guide all his actions. Furthermore, he explains that researchers with axiological skills are able to effectively attribute their values to research judgements and how research is being carried out. Axiology

tackles questions such as *How can participants' rights be protected?*, *How can psychological, physical or legal risks be prevented or minimised?*, *What moral concerns should be considered?* (NHMRC et al. 2007). Guevara-López et al. (2015) highlights that the axiological foundation is based on values such as beneficence, compassion, autonomy, non-maleficence and justice.

3.4 Research Paradigms

The word *paradigm* is derived from the Greek/Latin word '*paradigma*' meaning *a model or pattern* and has been used in educational research to mean a researcher's perspective of interpreting research data (Mackenzie & Knipe 2006). Guba et al. (1994) further describes the word to mean a set of guides or perspectives to govern research investigations. A research paradigm indicates a researcher's philosophical inclination and describes the study interest, process of study and presentation of research results. As highlighted by Lincoln & Guba (1985), every research paradigm has four critical components: methodology, epistemology, ontology and axiology, which constitutes the values and beliefs each paradigm presents. In this section, three research paradigms, namely positivist, interpretivist and pragmatism, will be briefly introduced with the aim of justifying the paradigm choice. Adequate comparisons will be made between the three paradigms and the stance for each paradigm on the aforementioned critical components will be highlighted.

3.4.1 Positivism

This form of research paradigm was first proposed by Auguste Comte when he sought to apply the scientific method of investigation to a study involving the natural and social world (Cohen et al. 2002). Comte (1856) posited that experimentation and observation should be the standard for better comprehension of human behaviour. Research established on this paradigm draws conclusions through theory formulation and testing as well as the derivation of mathematical equations. A Positivist's methodology seeks to identify causal relationships in research outcomes and is considered value-neutral (Creswell 2009). However, this

value-neutral ideology has been argued to be false as positivists tend to make value-laden conclusions in their choice of variables, observed actions and final interpretations (Salomon 1991). Considering the four critical elements of a research paradigm, the Epistemological position for positivism is *objectivism*, its ontological position is *realism*, its methodology is *experimental* (Kivunja & Kuyini 2017) and finally, its axiology is *beneficence* – meaning conducted research should be aimed at optimising research outcomes for participants and the general public (Mertens 2014)

3.4.2 Interpretivism

Interpretivism, unlike positivism, is centralised around gaining insights into the subjective perspective of humans (Guba & Lincoln 1989). A major effort is placed in understanding the point of view of the individual being observed instead of that of the researcher. Therefore, the key principle of this paradigm is that reality is formulated based on social interactions (Bogdan & Biklen 1998). Research-based on this paradigm involves data gathering in an approach congruent with the grounded theory (Straus & Corbin 1990). The *subjectivist* epistemological position of interpretivism indicates that the researcher’s understanding of accrued data is subjective and is based on interactions with observed subjects (Grix 2004). Interpretivism’s relativist ontological position suggests that the researcher considers the investigated scenario to be composed of several realities and that these realities can be better understood through interactions with research participants (Chalmers et al. 2009). A *naturalist* methodology and *balanced* axiology is assumed for an interpretivism paradigm.

3.4.3 Pragmatism

Philosophers who did not agree with either the positivist idea of accessing the truth about the actual world through scientific methods or the determination of social reality as proposed by interpretivism postulated the pragmatic paradigm (Peirce 1997). Teddlie & Tashakkori (2003), Biesta (2010) contend that a perspective that would allow research methods that are

most appropriate for investigating a phenomenon of interest should be considered as against the mono-directional approach of interpretivists and positivists. Hence, a more practical approach that uses a pragmatic approach to combine multiple methods and help comprehend the behaviour and beliefs of participants was postulated. The pragmatic approach follows a *relational* epistemology which allows the researcher to decide what relationships are suitable for his study as he deems fit (Bird-David 1999). Also, its *non-singular or relative* ontology means that a single reality does not exist and every participant can have his own understanding of reality (McCaslin 2008). A mixed-methods methodology is adopted by pragmatists, and it allows for a combination of quantitative and qualitative methods of data collection (Morgan 2014). Lastly, the pragmatic axiology is *value-laden* and allows the researcher to adopt an objective or subjective point of view (Thornhill et al. 2009)

3.5 Data Collection Methods

It is common for studies to categorise the terms *quantitative* and *qualitative* as data collection methods or research methods due to the misconception of researchers being informally referenced sometimes as qualitative or quantitative researchers (Mackenzie & Knipe 2006). However, as suggested by O’leary (2004), cited in the study of Mackenzie & Knipe (2006), one way of defining these terms is to associate each term as adjectives to different kinds of data and modes of analysis. For example, qualitative data should be associated with data representing words and pictures with thematic analysis and quantitative data should be associated with numeric data with a statistical form of data analysis. This definition describes both terms as modes of data collection as well as for analytic and reporting methods. Quantitative research derives conclusions through objective realisations from collected data while qualitative research involves subjective analysis of collected data (Creswel 2009).

3.5.1 Qualitative Research

Majority of the data for qualitative research are as a result of fieldwork (Patton 2005). As Barnes (1992) elucidates, there are three reasons why researchers opt for qualitative research rather than quantitative. Firstly, these researchers admit that they are unable to set aside their knowledge about the social world with the goal of being objective. Secondly, they presume that statistical and experimental methods are not sufficient in explaining and studying our everyday life. Thirdly, they claim the views of objective researchers are not analytically valid because researchers are occupied with the different perceptions of the world. Qualitative research focuses more on beliefs, motives and interpretations which relates more occurrences or happenings, which cannot be implemented through variables (Maxwell 2012). This research type deals with unquantifiable data. (Winter 2000). Leedy & Ormrod (2005) cited in the study of Williams (2007) has identified five methods of qualitative research namely: Content analysis, case studies, ethnography, phenomenology and grounded theory. All of these methods are applicable in different qualitative research cases (Creswel 2009).

3.5.2 Quantitative Research

Yanow & Schwartz-Shea (2015) in their study emphasised that quantitative data instigates the derivation of quantifiable information through statistical analysis with the aim of supporting or disproving “knowledge claims”. In its raw form, quantitative data convey little or no meaning to a researcher until analytic techniques are used to provide more information. Newman et al. (1998) highlights that quantitative research approach is used when reality is being observed and interpreted with intentions of developing a theory. According to Leedy & Ormrod (2005), quantitative research uses specific experimentation methods to improve existing theories. Quantitative researchers are inclined towards using mathematical models for the analysis of data collected using a predetermined apparatus. The conclusions drawn from quantitative research can either be justifying, descriptive or predictive (Williams 2007).



Figure 3.1: Philosophical Perspective For This Study

3.6 Research Choices for This Study

Figure 3.1 outlines the research choices for this study, with detailed justifications for each choice provided in the following subsections.

3.6.1 Deductive Research Method

This research does not propose a new theory; instead, it clearly utilises a deductive research method, evident from several hallmark characteristics of this approach. The study begins with established theories about traffic pollution, which guide the formulation of specific hypotheses. These theories recognise that traffic pollution is mainly caused by vehicle emissions,

including exhaust, brake, and tire wear, and that meteorological conditions and other highway parameters such as traffic flow significantly influence the concentration and dispersion of pollutants. Based on these theories, hypotheses are developed that predict the relationships between these highway parameters and pollutant levels, and how these relationships can be monitored and forecasted using modern technology.

A key feature of deductive research is testing these hypotheses through empirical data collection and analysis (Casula et al. 2021). In this context, the deployment of IoT sensors to gather real-time data on traffic and pollution levels serves as a practical application of empirical methods to test theoretical predictions. The study employs quantitative data analysis, utilising Big Data analytics to process extensive datasets and identify patterns or relationships as dictated by the initial hypotheses. This analytical approach facilitates statistical testing to either confirm these theories or suggest modifications. Furthermore, the research adheres to a structured and objective methodology typical of deductive approaches, ensuring that the results are reliable and replicable. This methodological rigour is in line with the scientific realism ontology, which underscores that objective and systematic scientific methods are essential for effectively measuring and analysing real-world phenomena (Mukumbang 2023). This structured approach not only strengthens the validity of the research findings but also enhances their applicability in practical settings, particularly in the development of traffic pollution management strategies.

3.6.2 Experimental and Case Study Research Design

The choice of research design is typically driven by the research objectives and the specific challenges they present. In the context of this study, a mix of the experimental and simulation research design has been identified as the suitable choice of research design. After careful consideration of available research design options, this choice of design approach is considered fitting for the experimental nature of this research. Poor air quality mitigation

studies are commonly based on experimental research since it is usually impossible to manipulate the variables of interest. The dependent variable y in this study is the highway air quality forecasts and the matrix of independent variables X are pollutant concentration for gases as well as meteorological data on temperature, humidity, pressure, amongst others. Similarly, the case study research design is essential as it allows the thorough exploration and understanding of specific and closely related cases identified in this study. This depth of analysis is often not possible with other research methods. Cross-sectional research design is unsuitable because it will fail to account for the varying pollutant concentration levels over time. Similarly, other research design approaches, such as comparative and descriptive methods, do not directly align with the objectives of our research.

3.6.3 Epistemological Stance

In the context of this study, the epistemological stance can be identified as objectivism. This stance reflects a belief in the possibility of obtaining objective knowledge about the world through empirical observation and scientific methodology. Objectivism in epistemology holds that reality exists independently of human thoughts and beliefs, and that it is possible to understand this reality through observation, measurement, and logical analysis. The objectivist stance in this research supports the use of quantitative methods for data analysis. This is consistent with the belief that numbers and statistical outcomes represent objective truths about the world. Through these methods, the study aims to produce reliable, reproducible, and universal knowledge about how traffic-related pollutants behave and can be controlled. This knowledge is expected to be valid across different contexts and settings, adhering to the objectivist view that scientific knowledge is universally applicable.

Furthermore, the objectivist epistemology aligns with the study's goal of developing predictive models. These models are based on the premise that the patterns and relationships identified through data analysis can be used to make accurate predictions about future con-

ditions. In essence, objectivism here supports the development of tools that can objectively forecast pollution levels based on empirical data, reinforcing the view that through rigorous scientific methods, we can gain a true understanding of natural phenomena and use this knowledge to make informed decisions and interventions.

3.6.4 Ontological Stance

The ontological position for this study is that of a critical realist. This approach assumes that pollutants such as nitrogen oxides and particulate matter exist independently of human perceptions and that their properties and effects are objective realities that can be empirically measured and analysed. Critical realism supports the idea that these pollutants and their interactions with environmental factors, such as meteorological conditions and traffic volumes, are discoverable through scientific methods. In adopting critical realism, the study inherently trusts that empirical data collection, whether through IoT sensors or other monitoring technologies, accurately captures real-world phenomena. It also presupposes that the causal relationships identified through data analysis—such as the impact of traffic flow on pollutant levels—are real and consistent, regardless of human observation. This stance justifies the use of advanced computational models and statistical techniques to forecast future pollutant levels based on observed data, under the belief that these models can reliably simulate real-world conditions.

Moreover, a realist ontology in this context underlines the development of predictive models and supports their application in crafting effective traffic management and pollution control strategies. It implies that interventions based on model forecasts can lead to tangible improvements in air quality. Therefore, the study's alignment with scientific realism not only enhances its methodological rigour but also reinforces its relevance and applicability to policy-making and practical environmental management

3.6.5 Positivist Research Paradigm

The research paradigm guiding this study is grounded in positivism. This paradigm believes that the world can be objectively observed and measured. In this study, this manifests through the deployment of IoT sensors to collect quantifiable data on air pollutants — data that are seen as objective truths about the state of the environment. These sensors measure specific, observable phenomena, adhering to the positivist principle that reality exists independently and can be empirically verified. The study’s approach also strongly emphasises hypothesis testing. It begins with theoretical assumptions based on existing knowledge about air pollution, which are then tested through empirical data collected via sensors. This data is processed and analysed using Big Data analytics, allowing the research to confirm or refute the initial hypotheses based on statistical evidence. This methodological rigour is characteristic of the positivist approach, which values systematic, scientific inquiry to achieve certainty and predict future conditions.

Moreover, the positivist paradigm in this research is aimed at establishing causal relationships and generalising findings beyond the specific conditions under study. By identifying how variables such as traffic volume and weather conditions affect pollutant levels, the study seeks to apply these findings to similar urban environments, thereby broadening the applicability of the results. The ability to generalise findings and the emphasis on replicability across different contexts highlight the study’s commitment to providing actionable insights that are not only scientifically valid but also practically applicable in real-world settings. This structured and objective approach underscores the study’s alignment with the positivist paradigm, which prioritises observable, measurable outcomes over subjective interpretations. Such a framework is particularly effective in environmental studies where precise data and clear causal links are crucial for developing effective policy and management strategies to mitigate pollution.

3.6.6 Quantitative Data Collection Method

According to Hulin et al. (2012), exposure evaluation based on quantitative-measurement is more practical in estimating the health effects of human exposure to pollutants. It is also useful for executing preventive policies to help mitigate the hazards involved. Also, Piedrahita et al. (2014) elucidates the importance of quantitative measurements of pollutants to apply procedures to ensure no intrusion from other pollutants. This requirement can only be met through quantitative research. Although this study involves a lot of fieldwork data collection; most of the data collected are quantitative, and studies have shown that fieldwork data can be largely quantitative (Brannen 2005). Also, real-time measurements of pollutants are quantitative and qualitative data gathered through surveys and questionnaires cannot serve the purpose of this study. For example, human participants are not able to give precise details on their daily car emissions on the highway. This sort of analysis is seemingly tricky to be carried out qualitatively due to the abstractness of air pollutants.

3.7 Sampling Method

Case study research strategy has been continuously criticised as being one of the social research methods that lack rigour and objectivity when compared to other methods (Rowley 2002). Nevertheless, studies have been conducted in fields such as nursing, sociology, technology and education that have all adopted the case study method of evaluation (Zucker 2001, Grassel & Schirmer 2006, Sadik et al. 2006, Hamilton & Corbett-Whittier 2012). Case study utilises quantitative and qualitative data obtained through the observation and investigation of different cases to demonstrate the outcome of a phenomenon (Gerring 2006). According to Yin (1984), the case study research method is defined as an empirical approach of inquiry aimed at investigating an event within a bounded context while employing several sources of evidence. Furthermore, Yin highlights that case study can be categorised into three, namely: exploratory, explanatory and descriptive. While exploratory case study

attempts to explore hidden phenomenon within data, descriptive case study describes this phenomenon and explanatory case studies provides a more in-depth explanation for the phenomenon. Also, McDonough & McDonough (2014) in their study provided two additional categories of case study research. The interpretive case study interprets data by supporting or criticising predefined assumptions while evaluative case study provides conclusions on the phenomenon within the data. Tellis (1997) identified two case study designs which are the *single* and *multiple* case study designs.

In this study, the evaluative type of case study is adopted as it enables the testing and evaluation of the proposed intervention in a real-life scenario. A multiple case study rather than a single case study design is employed since it involves studying several cases, often with variations in context, characteristics, or conditions. This approach permits the identification of commonalities and differences across cases, leading to more robust and generalised findings that can be applied to a wider range of contexts. The findings from multiple case studies are typically more applicable to real-world situations due to the diversity of cases studied. This enhances the external validity of the research, making the results more relevant and meaningful to a broader audience. By analysing multiple cases, the study can identify patterns, trends, and variations that might not be apparent in a single case study.

The choice of highways for this study rather than local roads is based on two reasons - 1) Highways typically experience much higher traffic volumes compared to local roads. This increased traffic will provide this study with a larger and more diverse sample of vehicles, drivers, and behaviors to observe and analyse. This diversity can enhance the representativeness of the data collected, leading to findings that are more applicable to a wider range of driving populations. 2) Highways often involve longer travel distances compared to local roads. This extended duration of travel will provide the study with continuous data streams that cover various driving conditions, such as merging, lane changes, and steady-state cruis-

ing. Continuous data collection allows for a more comprehensive understanding of pollution patterns and will lead to more nuanced insights and accurate analysis of these patterns and trends. However, It is crucial to acknowledge that selecting highways as case study subjects for this research entails specific challenges, including addressing safety concerns, securing the required permissions for data collection, and managing the complexities of high-speed environments. Consequently, to ensure an effective approach, a set of criteria was initially established based on factors like accessibility to the selected highways and ethical considerations. Among the numerous highways in the UK, four key highways were identified that best aligned with the predetermined criteria, aligning seamlessly with the study's objectives. Following a careful analysis of existing systems through literature review, the following criteria were used to select suitable highways for this study are the identified criteria for a suitable case study for this research:

- **Traffic variation:** An optimal case study scenario necessitates consistent traffic flow throughout both peak and off-peak hours, facilitating significant fluctuations in pollution levels. To illustrate, an exemplary road would experience substantial congestion during rush hours and notably lighter traffic during other periods.
- **Power source:** The chosen trial site must offer viable resources for powering the IoT device, either through solar energy or electrical means.
- **Device Installation and Cellular data connectivity:** IoT sensors should be mounted on lamp posts at a height of 1.5 to 2 meters above the ground, strategically positioned on both sides of the highway and, if possible, near the center to monitor emissions from all traffic lanes. The selected trial location should have numerous lamp posts to ensure extensive coverage and be characterised by reliable cellular coverage for seamless connectivity to the cloud infrastructure.
- **Highway Length:** The strategic placement of sensors mandates a reasonable spacing between them. Consequently, the length of the road should permit the deployment of

numerous sensors at suitable intervals.

- **Success Metric For Deployment:** The deployment of sensors for the study depends on the resources available for the project. Nevertheless, it is necessary to install multiple sensors along the case-study highways to collect pollution data effectively. The data collection phase is expected to span 8-12 months, and this duration will be clearly specified in the requests for site access.

Based on the above criteria, necessary site access was sought from Costain Plc, a UK highway contractor and collaborator providing case studies for this research. A non-probability sampling method, driven primarily by site availability and ease of access, was employed to select the final case-study highways. When collecting data from the entire population is impractical, sampling becomes essential. Non-probability sampling was chosen because it allows for case study selection based on the researcher's subjective decisions, necessary when factors like accessibility and collaborative agreements influence site selection. This approach is well-documented in research methodology literature, such as in Thornhill et al. (2009), which discusses the implications and utility of non-probabilistic case study selection in practical research scenarios.

Unlike probability sampling, determining the sample size in non-probability sampling can be ambiguous and often relies on the research aims and objectives. Various non-probability sampling techniques, such as quota sampling, purposive sampling, snowball sampling, self-selection sampling, and convenience sampling, exist (See Figure 3.2). After careful consideration, convenience sampling was deemed most appropriate for this study. This conclusion was based on the location choice being driven by the collaborator's availability, providing easily obtainable and accessible highways considering bureaucratic procedures. This method of selection is a feature of convenience sampling (Etikan et al. 2016).

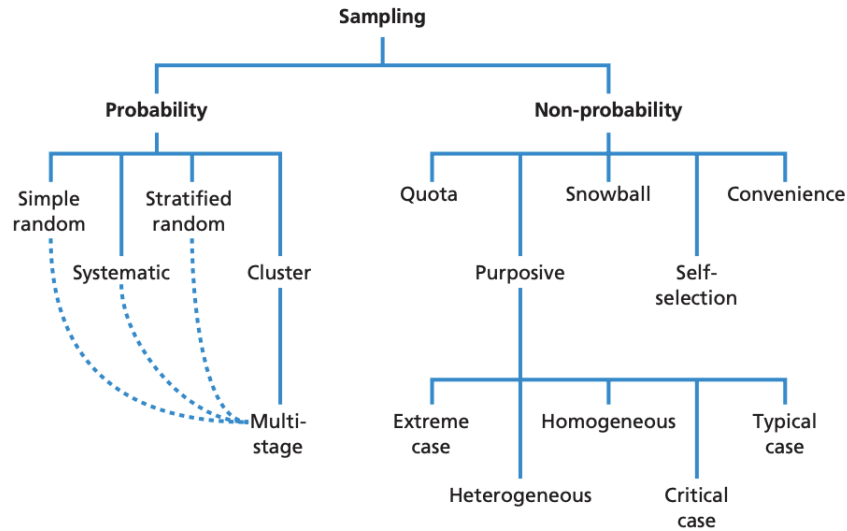


Figure 3.2: Sampling Techniques, Source: (Thornhill et al. 2009)

3.8 Chapter Summary

This chapter provides a detailed insight into the research methodology adopted in this study. The research methodology incorporates both experimental and case study research designs, catering to the study’s experimental nature. The positivist research paradigm ensures an objective and unbiased analysis of data, while quantitative data collection is chosen for its precision in air quality monitoring. The case study research design, particularly the evaluative and multiple case study types, offers a comprehensive assessment of IoT-based air quality monitoring systems. Convenience sampling facilitates the selection of suitable highways for case studies, aligned with the availability of highway locations offered by the collaborator on this study, Costain Plc. This systematic methodology aims to investigate the effectiveness of IoT sensors in monitoring and forecasting highway air quality, with a focus on emissions and pollution, ensuring the collection of reliable and applicable results. The selection of highways for the case studies adheres to predefined criteria, further enhancing the diversity and viability of the chosen sites. Site access requirements have been duly requested to support the study’s robust data collection and analysis.

Chapter 4

A Scalable Framework for TRAP Monitoring and Forecasting

4.1 Chapter Overview

In this chapter, a cost-effective framework designed for monitoring and forecasting pollutant levels along UK highways is presented. This framework serves as a pivotal step in achieving the first research objective and lays the foundation for the fourth objective through the deployment of the REVIS system. This comprehensive framework encompasses hardware, data storage solutions, and predictive tools, all of which are applied in practical scenarios involving selected highways as case studies. Additionally, the chapter explores the framework's scalability and its capacity for real-time monitoring in scenarios involving more sensors and highways than those considered in this study. Figure 4.1 shows an illustration of the workflow for the entire process of framework design and validation

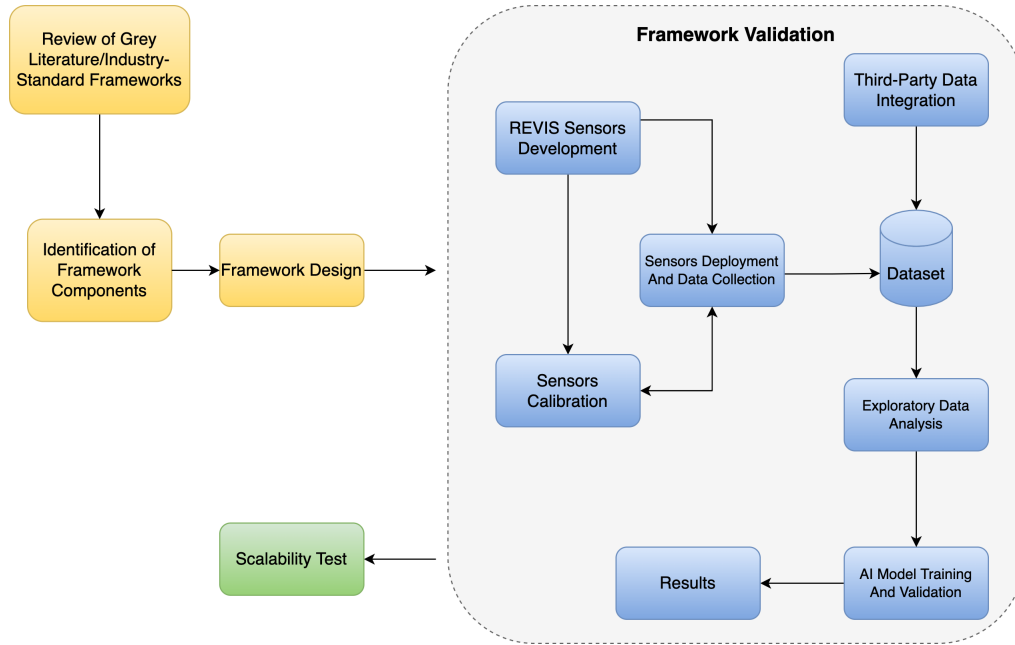


Figure 4.1: Pipeline diagram for framework design and validation

4.2 Framework Design Methodology

Layering is an established application design strategy that simplifies the management of complex software systems by dividing them into distinct, manageable modules. This modular approach not only facilitates easier development and maintenance but also enhances the scalability and flexibility of system design. In the proposed framework, layering is strategically used to organise critical components that ensure the system’s functionality. These components are distributed across various layers, including libraries that provide specific functionalities, programming languages that dictate the implementation of these functionalities, and services that are crucial for performing monitoring and forecasting of environmental data. By organising the framework in this way, each layer can be developed and refined independently while maintaining a cohesive overall system architecture, essential for effectively managing the complex processes involved in traffic pollution monitoring.

The design of the framework was significantly influenced by a comprehensive market analysis of “grey literature” and an in-depth review of academic publications. This research was

instrumental in selecting several sensor components and programming languages that were integrated into the framework following literature recommendations. Searches conducted through Google Scholar and scientific databases such as Scopus and ScienceDirect were crucial in identifying relevant academic publications, while Google’s search engine uncovered additional sensor development approaches. This process ensured a robust selection of the latest and most effective technologies.

Further investigation into relevant integration libraries and big data frameworks provided insights into addressing the challenges of data integration and storage within the framework. A variety of enterprise frameworks were evaluated, with selections made to facilitate the integration of data from both existing legacy systems and new constructions. Additionally, a range of algorithms known for their effectiveness in air quality forecasting was reviewed, with both academic and industry use considered. For this study, a scalable machine learning approach, specifically deep learning, was selected due to its demonstrated success in distributed computing environments, as highlighted by recent research cited in studies such as those by Sergeev & Del Balso (2018) and Chen et al. (2019). This approach ensures that the framework is not only robust in handling complex data but also adaptable and effective in providing accurate and timely environmental forecasts.

4.3 Framework Components

The framework design is a four-layered architecture composed of the hardware layer, data storage layer, integration layer and analytics layer as depicted in figure 4.2. This section introduces these layers and their functionalities.

4.3.1 Hardware Layer

This layer serves as the entry point for the entire framework. It initiates the monitoring and analytics process by ensuring that real-time data are captured and subsequently transferred

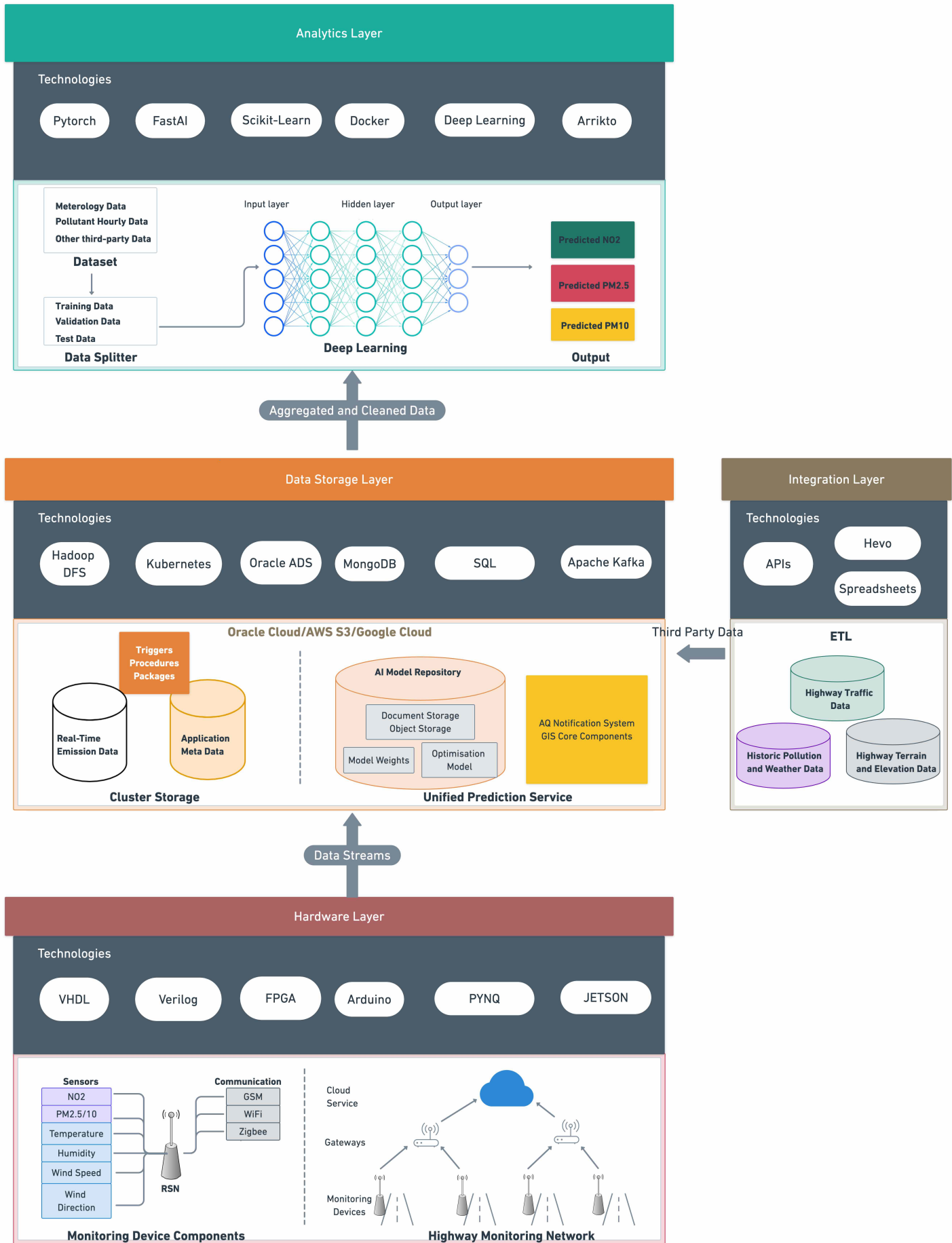


Figure 4.2: Scalable Framework for Highway Air Quality Monitoring and Prediction

to a cloud platform for data aggregation. A typical real-time sensing device in this layer would push data at an interval of 30secs-1min and be able to sense multiple pollutants and capture weather data. Other device functionalities such as self-powering capability, edge computing and on-board intelligence are desirable but not entirely mandatory for monitoring. Multiple gateways and a cloud platform are essential for this layer to function as required. The cloud platform will store captured data, but on-device storage will also be helpful to avoid data loss when data transfer fails. Additional data on vehicle categories and traffic flow in this layer will provide more insights into the 'culprit' vehicle that contributes the most to highway pollution. Advanced computer vision and edge computing technologies can enable this functionality in monitoring devices through embedded ML models. Development technologies relevant to this layer include VHDL, Verilog, FPGA, and Arduino.

4.3.2 Data Storage Layer

This layer stores pollution data and model weights. Readings captured from deployed sensing devices are either sent immediately to this layer or stored temporarily and pushed later through HTTP post requests. The data storage layer is responsible for ensuring data consistency, security and integrity. According to Ahmed et al. (2017), it is best practice to have the unified prediction service (UPS) reside close to the historic pollution data to reduce latency. Hence, this layer also houses weights and parameters from training pollutant concentration forecasting models. Data stored in this layer are bound to increase exponentially, and necessary technologies to configure big data storage must be put in place. Relevant technologies such as hadoop, spark and hive are possible open-source options to consider in this configuration. Data streaming frameworks like Apache Kafka or ActiveMQ are also available for real-time sensing of changes in this layer and to send alerts in the event of data transfer failures. Triggers, procedures and packages are useful to automate most of the required database tasks such as populating tables, generating logs or automatically generating SQL for data aggregation.

4.3.3 Integration Layer

The data integration layer ingests data from third-party sources into a central repository. The layer handles this data ingestion using the extract, transform and load (ETL) process. External data can include pollution data captured by other monitoring stations, highway geographical data, meteorological data and traffic data. The essence of this layer is to ensure that data not captured in the hardware layer by the monitoring devices can be integrated into the system to improve the performance of developed estimation models. If the suggested functionalities of the hardware layer are too expensive to implement, this layer can grab open-source or paid data from available online sources. Data can be downloaded in different formats such as TXT, JSON, XML and CSV or exposed as external links. The data from this layer should be stored as separate tables in the data storage layer for unique identification and also to avoid mix-ups with existing data.

4.3.4 Analytics Layer

The analytics layer handles exploratory and inferential analysis of historic highway pollution data to estimate future air quality. The layer extracts data from the data storage layer for model training and validation. Essential data pre-processing steps such as data consistency verification, target attribute transformation, feature extraction, data encoding and data imputation are carried out in this layer as part of the first stages of training. A machine learning approach suitable for tabular or time-series data such as the historic pollution data is required for estimation. Deep learning is one of many machine learning approaches that has stood the test of time (Akinosho et al. 2020). Frameworks and libraries such as fastai, scikit-learn, PyTorch and TensorFlow make it relatively easy to train a baseline model. Additional functionalities that are beginning to gain traction and could be included in implementing this layer is MLOps - model maintenance in the production environment. MLOps encompasses automation and monitoring steps such as continuous integration, deployment and training on data collected in production.

4.4 Development and Deployment of the REVIS System Prototype

In this section, the proposed framework is validated for practicality through the implementation of a Real-Time Highways Emission Visualisation (REVIS) platform use case. The framework was tested for scalability and performance through different stages of data collection, exploratory data analysis and predictive model development.

4.4.1 REVIS Highway Monitoring Devices

The development and evaluation steps of the monitoring devices and the deployment strategy adopted are highlighted in this section.

Table 4.1: Sensor Specifications and Accuracy

Measured Quantity	Units	Sensor used	Accuracy	Comments
Temperature	$^{\circ}C$	Texas: HDC2010	± 40	Could be affected by direct sunlight, depending on how well airflow works within the unit - may require additional physical shading.
Relative Humidity	%	Texas: HDC2010	± 3 start of life $\pm 0.25/\text{yr}$ drift	As above
Pressure	hPa	ST: LPS22HB	± 1	
$PM_{2.5}$ and PM_{10}	$\mu g/m^3$	Sensirion: SPS30	$\pm 10 \mu g/m^3$ $\pm 10\%$	Over 0-100 $\mu g/m^3$ range Over 100-1000 $\mu g/m^3$ range
NO_2	ppb	Alphasense: B43F	NO2- Approx. ± 20	Careful design and several stages of calibration are required when measuring tiny gas concentrations

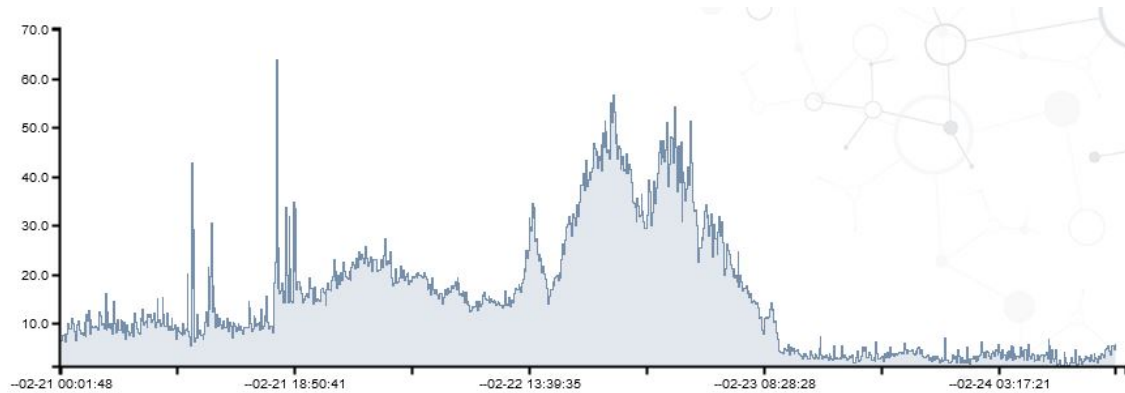
4.4.1.1 REVIS Device Development and Evaluation

REVIS demonstrates the hardware layer through the development and calibration of devices with built-in sensors to measure the atmospheric composition of NO_2 , $PM_{2.5}$ and PM_{10} ,

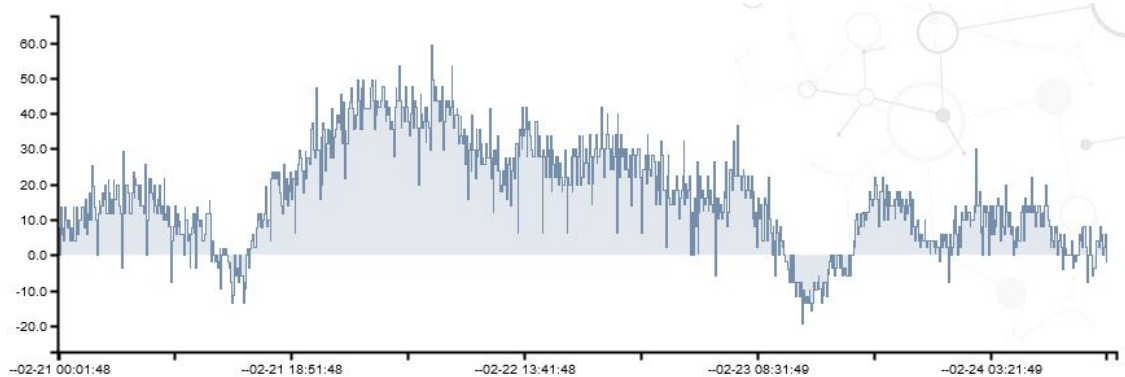
alongside weather parameters - pressure, temperature and relative humidity. Table 4.1 below summarises details of manufacturers of the chosen sensors and their accuracy figures. Each REVIS device required an excellent design of both analogue and digital circuitry around it and several stages of calibration. The Alphasense NO_2 sensor for example, showed during experimentation that it was best suited for fixed sensing installations and urban air monitoring since varying meteorological conditions had a significant influence on its readings. The sensor's cross-interference with the $PM_{2.5}$ SPS30 sensor and detection range limits (DRL) were also evaluated using equation 4.1

$$DRL = 3.3\sigma/S \quad (4.1)$$

where S denotes the calibration curve's slope, and σ denotes the standard deviation of the sensor response in the absence of air (Shrivastava et al. 2011). The nearest AURN stations to the monitoring devices were identified for field evaluation. The selected stations were deemed suitable for calibration since they were close to deployed sensors and mainly provided missing weather data and also hourly measurement of the pollutants of interest. Data from the REVIS devices were averaged over an hour for appropriate comparison with the reference data. Figure 4.3a shows $PM_{2.5}$ and NO_2 readings on one of the REVIS devices after calibration. Aside from the occasional underestimated measurement of the NO_2 sensors, other sensors that measured $PM_{2.5}$ and PM_{10} showed close estimates to the reference measurements with correlation coefficient $r > 0.8$.



(a) $PM_{2.5}$ field readings after calibration



(b) NO_2 field readings after calibration

Figure 4.3: Calibrated NO_2 and $PM_{2.5}$ readings from field. Vertical units are in $\mu g/m^3$ for $PM_{2.5}$ and ppb for NO_2 . Even with the calibration, NO_2 readings sometimes record negative readings.

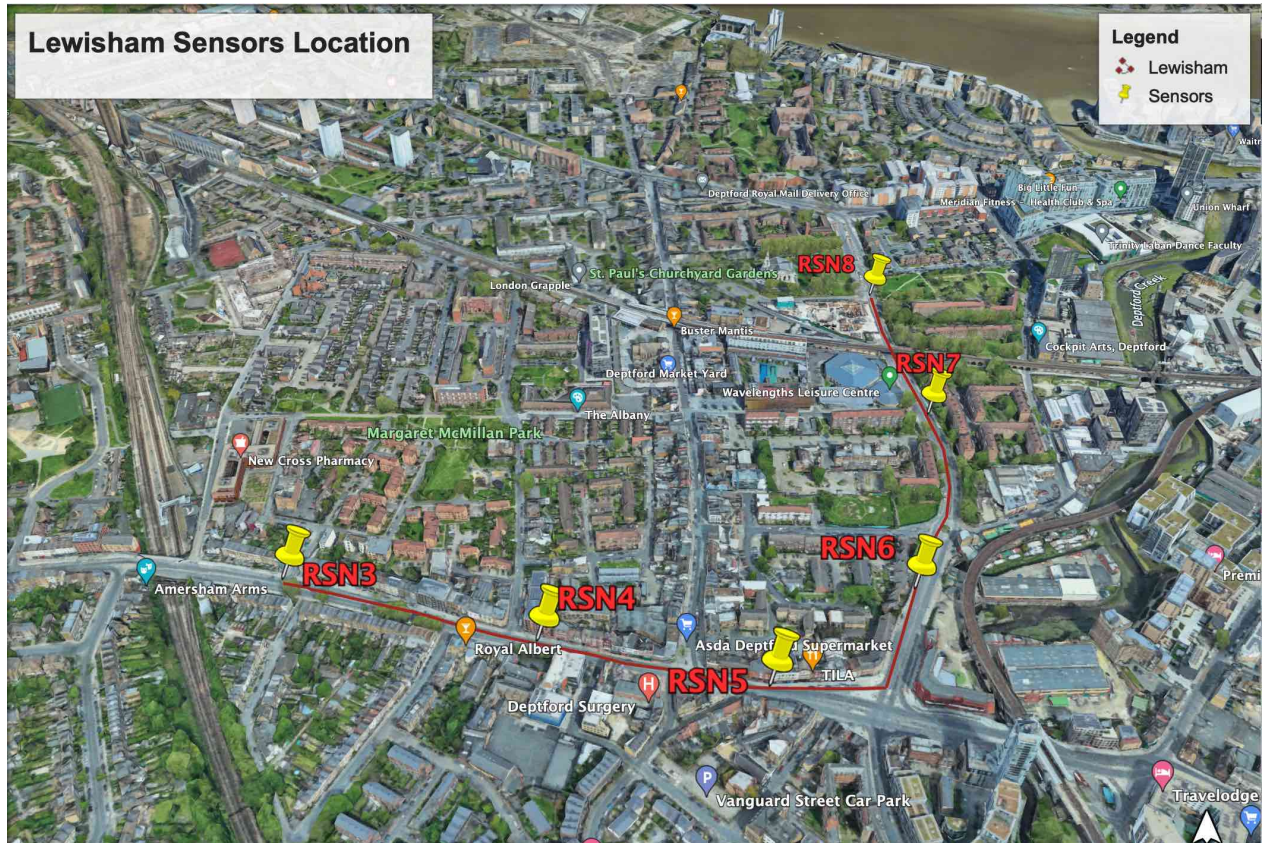
4.4.1.2 Device Deployment in Case Study Regions

The research focuses on major highways in London (A2209, A302), Newport (M4), and Chepstow (A48). London, with a population of 9 million and a density of 5,598 persons per square kilometre, poses significant traffic congestion challenges, especially with 74.9% of its inhabitants falling within the 16-64 age group (ONS 2021). A study by TFL (2019) revealed that 59% of Londoners use buses weekly, while car commuting remains popular among the younger demographic. Newport and Chepstow, situated in southeastern Wales, accommodate 1.53 million residents with a density of 546 persons per square kilometre ONS (2018). Monmouthshire, within this region, exhibits the highest life expectancy in Wales. Statistically, 74.3% of the employed population in the area prefer motorcycles, vans, or cars for commuting, whereas 8.8% opt for buses or trains (Statswales 2020). The M4, A48, and A466 highways serve as vital connectors between neighbouring cities in this region.

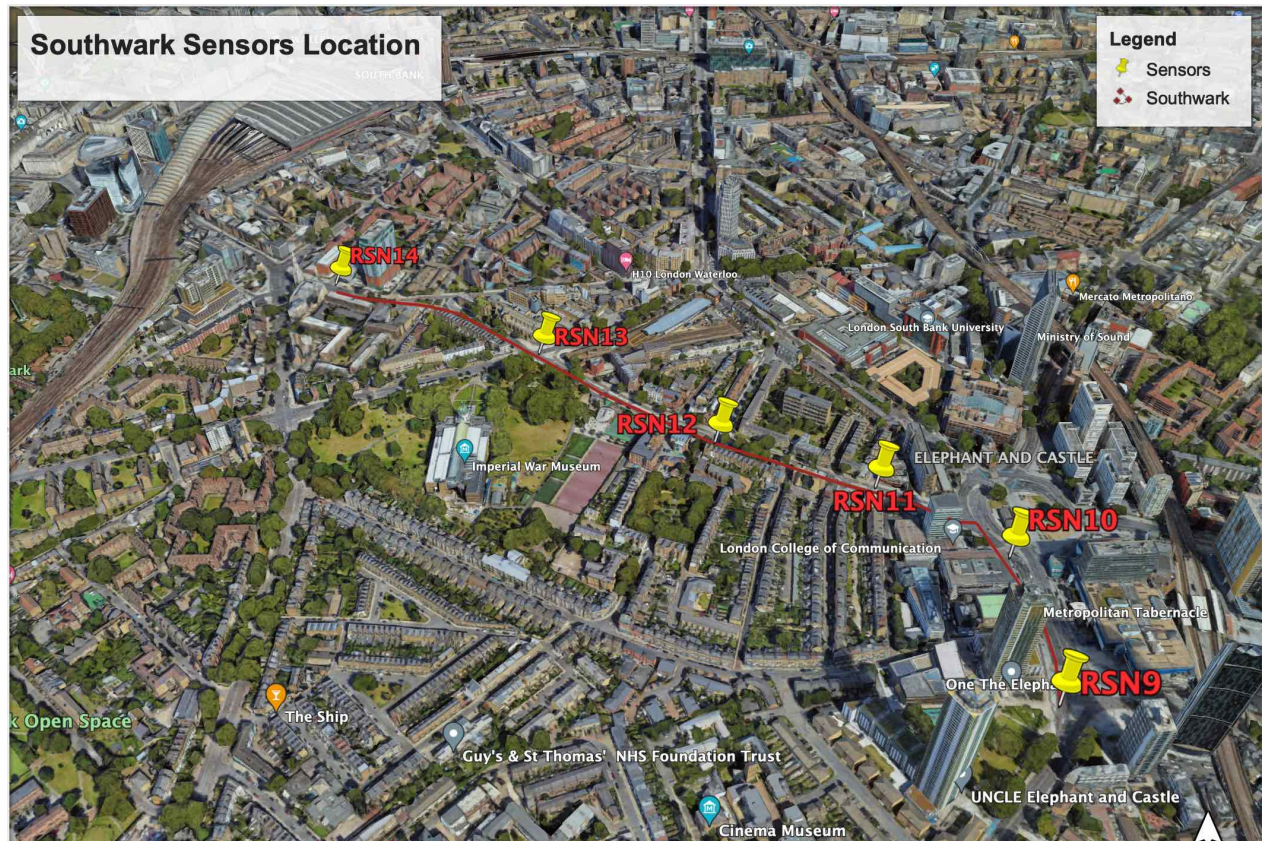
Figure 4.4 depicts the distribution of REVIS devices in these cities. In London, twelve devices were distributed on sections of the A302 and A2209 highways. One device was placed 92.79m from Junction 25 of the M4 highway in Newport and another device was positioned close to The A48 motorway in Chepstow. The deployment approach that was adopted during the distribution of these devices ensured three key requirements: (1) sufficient highway length (2) cellular data connectivity and (3) electrical/solar power availability. It was also necessary that device installation required minimum technical skills and data was captured for a minimum of 6-8 months.

4.4.2 Exploratory Analysis of Pollution and Weather Data

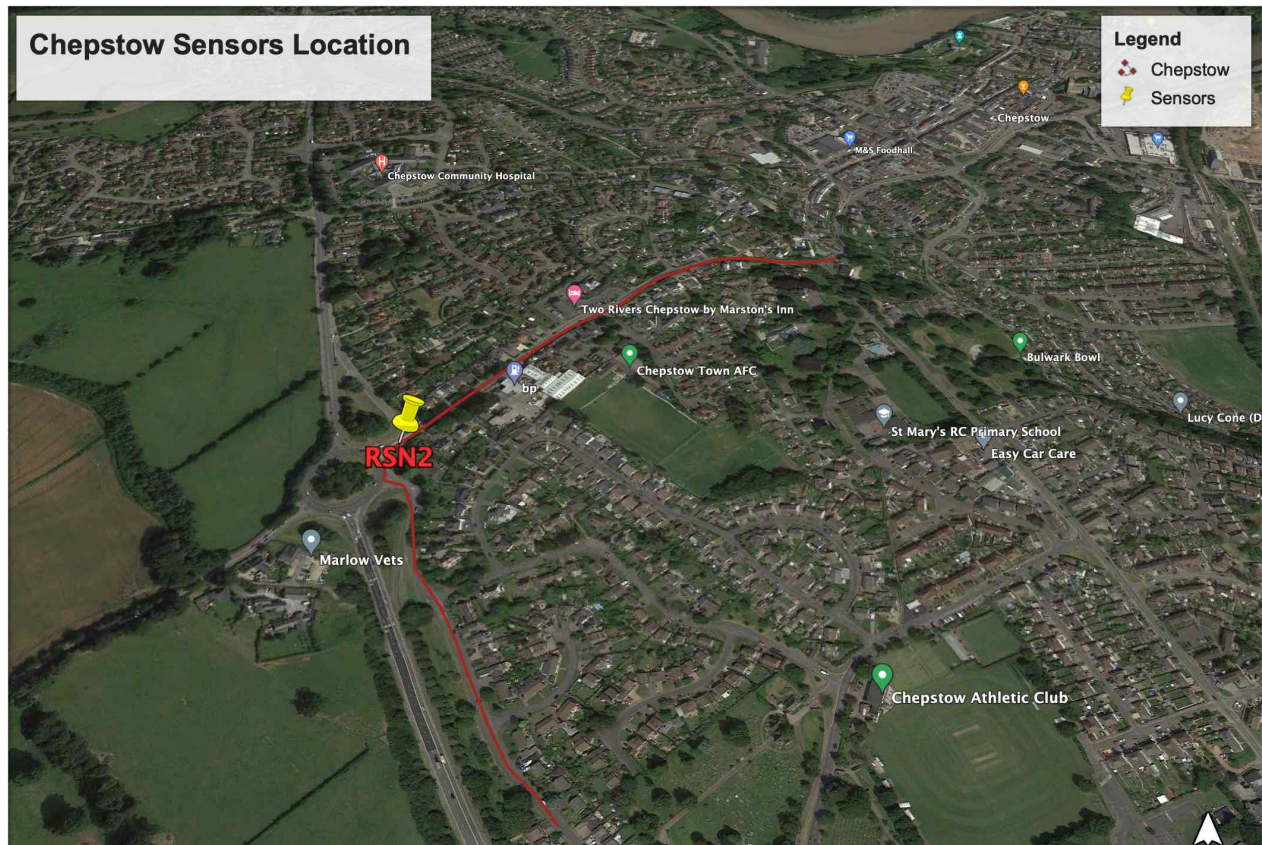
It is important to verify data consistency before commencing model training in ML regression tasks such as the one being considered. The minimum recommendation is to confirm the total number of rows and columns within the data, as this may have been compromised during data transfer (Bilal & Oyedele 2020). This section analyses the impact of weather parameters and



(a) Sensors distribution is Lewisham A2209 highway.



(b) Sensors distribution in Southwark A302 highway.



(c) Sensors distribution in Chepstow A48 highway.



(d) Sensors distribution in Newport M4 highway.

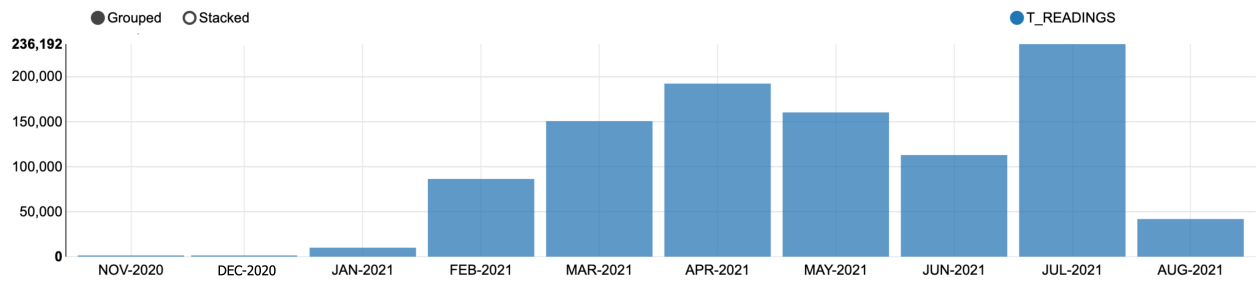
Figure 4.4: Maps showing the distribution of 14 REVIS devices across four UK regions: Newport (1), Chepstow (1), Lewisham (6), and Southwark (6). For the Southwark and Lewisham locations in London, devices captured readings from the A302 and A2209 highways, while those in Newport and Chepstow were deployed near the M4 and A48 highways, respectively.

the case-study region on pollutant levels. Although data was captured between November 2020 and August 2021, missing data in the early stages of deployment (shown in Figure 14 below) influenced the decision to analyse data between February 2021 and August 2021 when missing data was minimal.

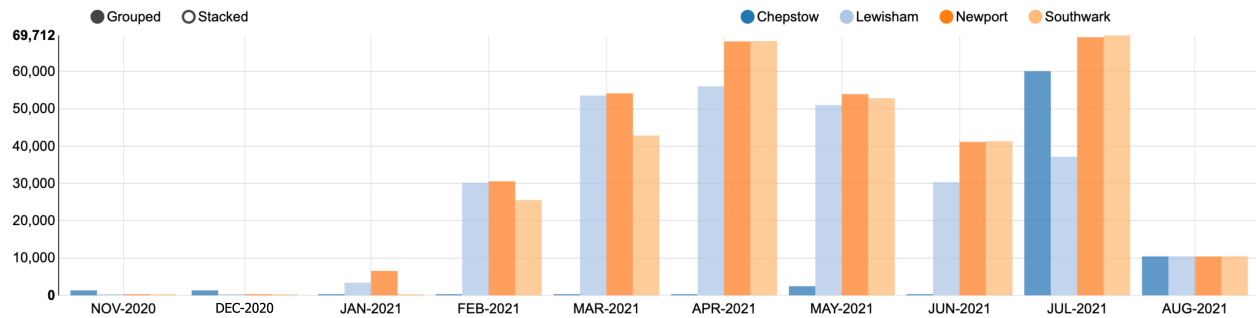
4.4.2.1 The Impact of Weather on $PM_{2.5}$, PM_{10} and NO_2

Weather parameters significantly impact the dispersion rates of pollutants (Barrera-Animas et al. 2022). Given the seemingly similar attributes of some weather data parameters, such as “feels_like,” “temp_min,” and “temp_max,” it was crucial to examine the correlations between these parameters to be able to identify whether certain parameters provide redundant information. Figure 4.6 presents a correlation matrix used to identify the hierarchical similarities among these parameters, revealing a strong correlation between temp, temp_min, temp_max, and feels_like.

To understand the effects of temperature on four pollutants, seasonal trends were plotted, as shown in figure 4.7. The average temperature for all four regions ranged between 8.6 and 12.56°C in winter, 9.73 and 19.76°C in spring, and 19.41 and 21.78°C in summer. Regression analysis of temperature against each pollutant, presented in Table 4.2, indicates a positive correlation between $PM_{2.5}$ and PM_{10} and temperature in Newport, Southwark, and Lewisham during the spring and summer seasons. Chepstow showed no correlation in winter due to the lack of temperature readings and a negative correlation in spring and summer. NO_2 exhibited a negative correlation with temperature in all regions during winter and spring, but a positive correlation in Southwark and Lewisham in summer. These findings support studies suggesting that concentration levels are highest when the temperature is elevated (Pearce et al. 2011, Analitis et al. 2014).



(a) Average monthly readings captured by deployed REVIS devices.



(b) This plot illustrates the number of readings captured per region.

Figure 4.5: Total monthly readings captured by deployed sensors between November 2020 and August 2021. These plots illustrate the amount of missing data in the first two months when some devices were offline. Chepstow had the lowest monitored readings overall.

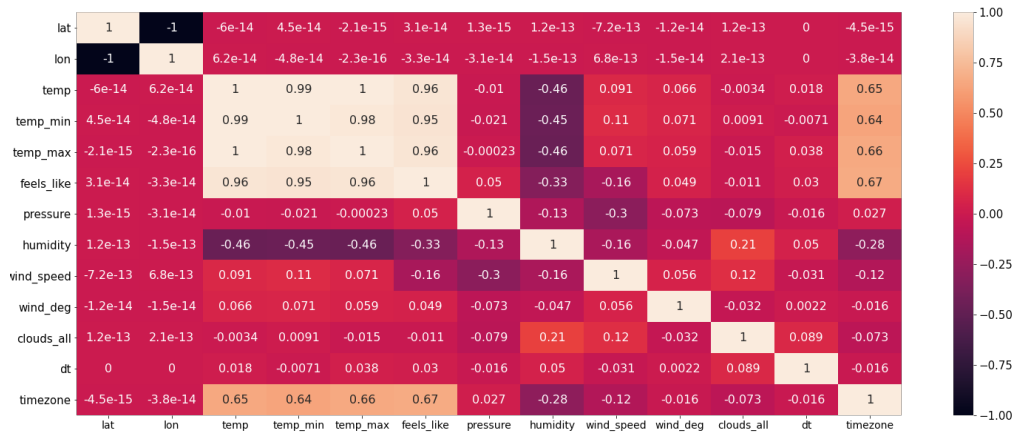


Figure 4.6: Distance matrix of weather parameters using Pearson's correlation. A strong correlation can be noticed between "temp", "temp_min", "temp_max", "wind_speed", "wind_degree" and "feels_like". There is also a discernible correlation between "clouds_all" and "humidity"/"windspeed".

Table 4.2: Regression analysis of weather parameters vs pollutant concentration

Regions	Winter				Spring				Summer			
	temp($^{\circ}C$)	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$	temp($^{\circ}C$)	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$	temp($^{\circ}C$)	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$
Newport	8.60	-0.53	-0.03	-0.46	9.73	-0.56	0.59	0.48	19.58	-0.32	0.61	0.51
Southwark	12.68	-0.40	-0.10	-0.32	10.44	-0.33	0.23	0.18	19.41	0.11	0.24	0.26
Lewisham	12.56	-0.46	-0.13	0	11.80	-0.41	0.38	0.10	20.77	0.33	0.35	0.09
Chepstow	-	-	-	-	19.76	-0.44	-0.38	-0.33	21.78	-0.19	-0.20	-0.17

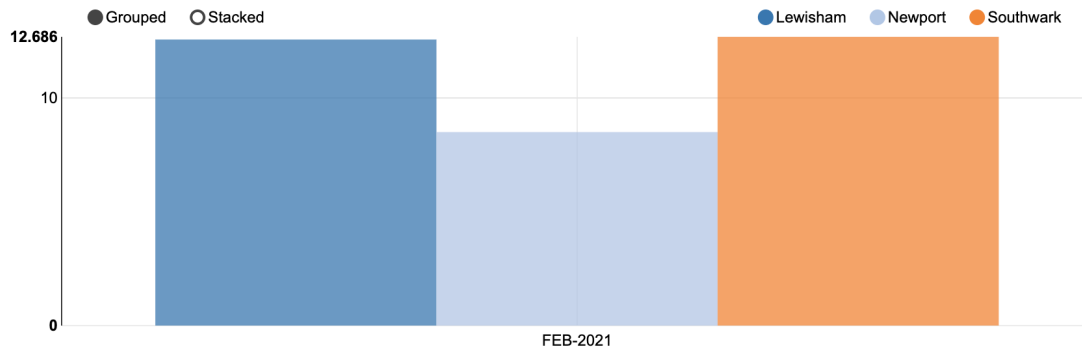
(a) Correlation between regional temperature and pollutants in spring, winter and summer

Regions	Winter				Spring				Summer			
	pressure	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$	pressure	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$	pressure	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$
Newport	1014	-0.10	0.42	0.38	1022.50	-0.06	0.12	0.31	1012.90	-0.13	0.33	0.55
Southwark	1018.50	0.22	0.10	0.13	1026.10	-0.15	-0.17	-0.22	1015.30	0.26	0.08	0.03
Lewisham	1018.80	0.07	0.03	0.11	1026.30	-0.01	-0.10	-0.18	1014.20	0.19	0.16	0.15
Chepstow	-	-	-	-	1009.50	0.44	0.22	0.15	1007.20	0.19	0.10	0.09

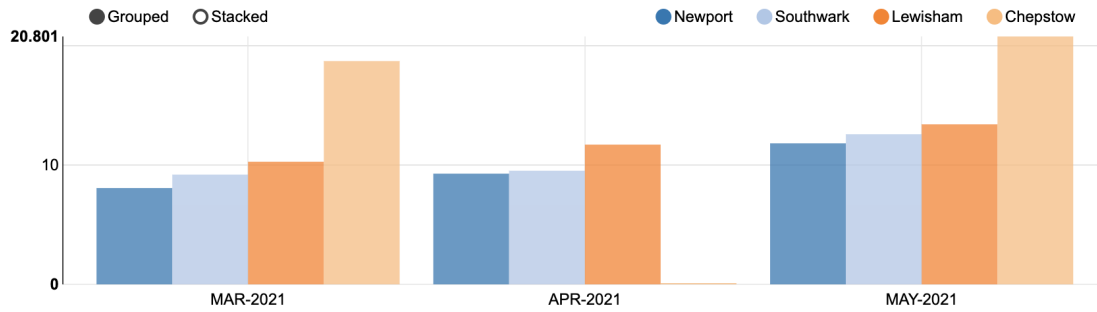
(b) Correlation between regional pressure and pollutants in spring, winter and summer

Regions	Winter				Spring				Summer			
	humidity(%)	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$	humidity(%)	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$	humidity(%)	$NO_2(r^2)$	$PM_{2.5}(r^2)$	$PM_{10}(r^2)$
Newport	90.96	0	-21	-18	73.54	2	-11	-3.40	70.85	1.30	-13.70	-4.80
Southwark	66.27	7	-1	-15	65.85	13	-6.50	-8.90	73.30	6.80	-3	-2.20
Lewisham	72.93	3	-8	-5	64.04	11	-15.20	-4.20	70.98	17.6	-17	-5.60
Chepstow	-	-	-	-	53.95	8	-1.70	-6	64.95	13.30	-3.4	-11.20

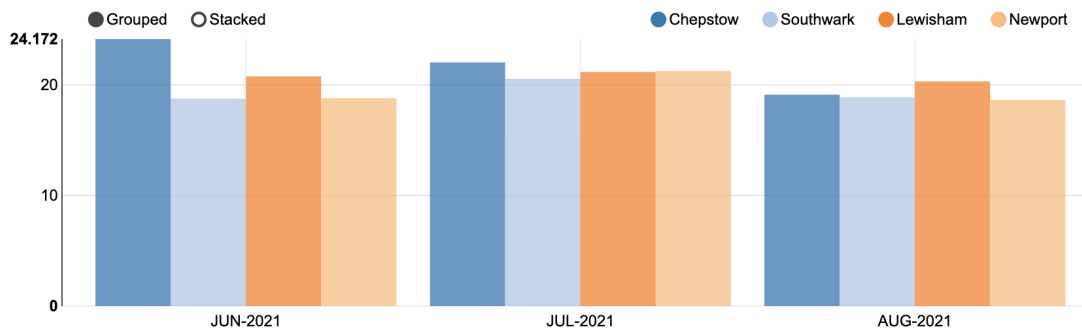
(c) Correlation between regional humidity and pollutants in spring, winter and summer



(a) Average winter temperature for all four regions



(b) Average spring temperature for all four regions



(c) Average summer temperature for all four regions

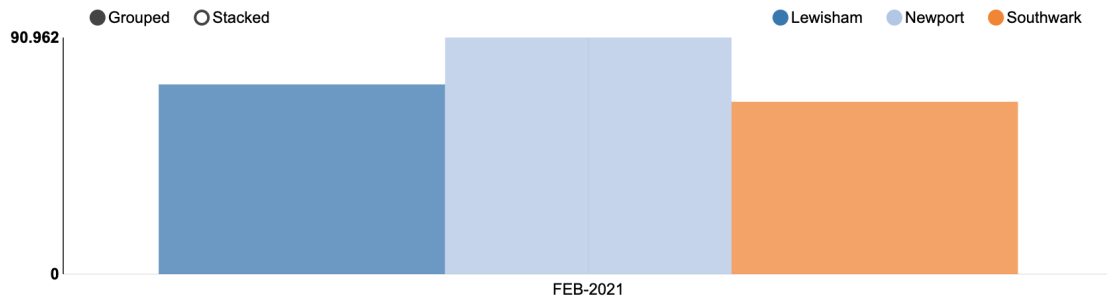
Figure 4.7: The seasonal trends for temperature in Newport, Southwark, Lewisham and Chepstow. Newport has the lowest temperature of 8.6°C in winter as there was also no reading recorded for Chepstow, as illustrated in plot (a). Chepstow had the highest average temperature of 19.76°C in spring and 21.78°C in summer, as shown in plots (b) and (c)

For pressure, the lowest readings were recorded in Chepstow during Spring and Summer seasons while Lewisham and Southwark recorded the highest pressures in spring. Table 4.2b summarises the pressure readings during these seasons and the correlation figures with the pollutants. The $PM_{2.5}$ and PM_{10} concentrations in Newport and Chepstow were positively correlated with pressure, indicating that an increase in atmospheric pressure will increase the concentration levels of these highway pollutants. All three pollutants negatively correlate with pressure in Southwark and Lewisham in spring but positive in winter and summer. The conclusion drawn from this result is a strong correlation between pressure and $PM_{2.5}$ and PM_{10} but a significant negative correlation with NO_2 . Figure 4.8 illustrates the average seasonal humidity across the regions with the lowest humidity value was recorded in Chepstow during summer and the highest in Newport during winter. It can be deduced from Table 4.2c that $PM_{2.5}$ and PM_{10} were negatively correlated with humidity for winter, spring and summer seasons. In particular, both pollutants are prone to be absorbed in the atmosphere as humidity increases. Naturally, rain results in higher relative humidity and soaks up these particles, resulting in a lower level of particulate in winter. (Odat 2009).

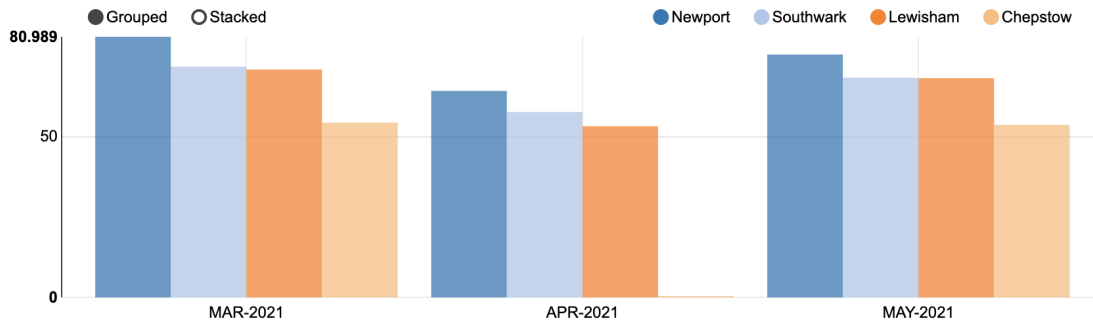
4.4.2.2 The Impact of Region on $PM_{2.5}$, PM_{10} and NO_2

Each region has its unique attributes which can influence the concentration level of pollutants measured over the experimentation period. Aside from the weather, other attributes such as the highway gradient, region terrain, residential development, background coefficient and traffic flow can also contribute to the concentration levels across regions (Sayegh et al. 2016, Pasquier & André 2017). Although some of these attributes were not captured in this research, their effects on the captured concentration levels remain to be seen. This section presents some primary insights across the four regions in the data set.

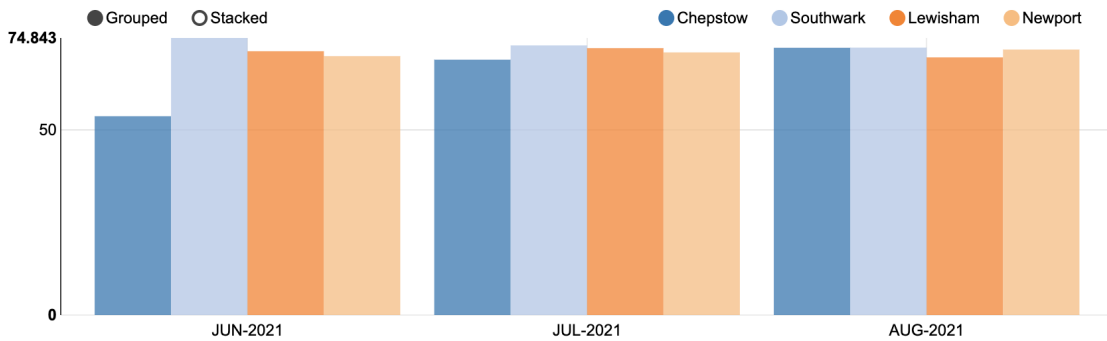
Table 4.3 shows that the average concentration levels across regions vary significantly. Chepstow and Southwark seemed to have the most practical NO_2 averages, with Southwark



(a) Average winter humidity for all four regions

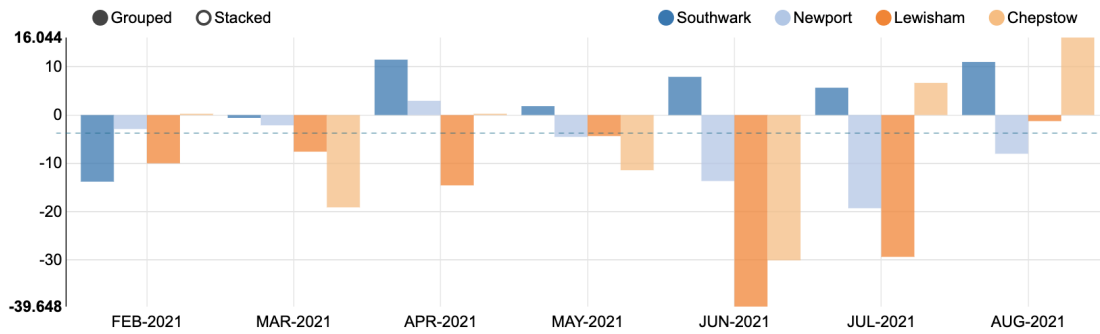


(b) Average spring humidity for all four regions

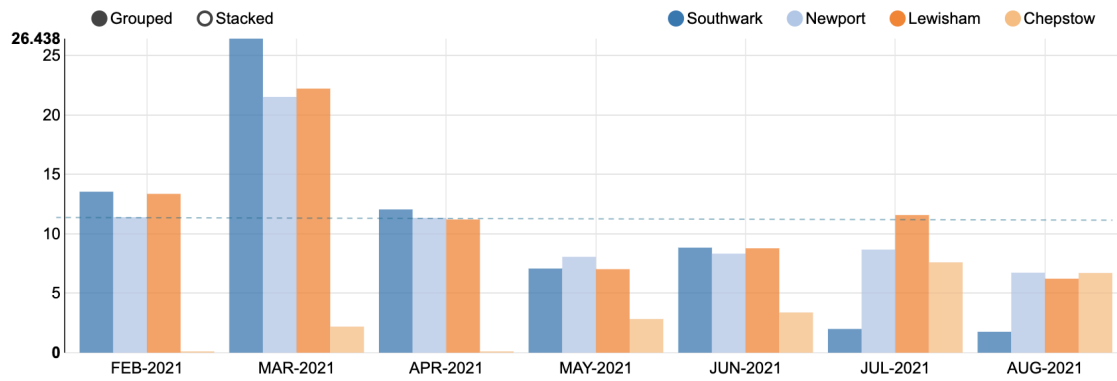


(c) Average summer humidity for all four regions

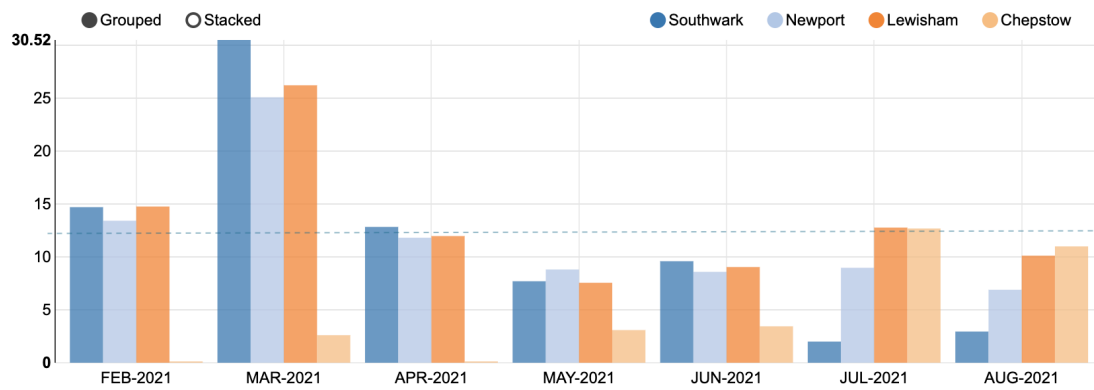
Figure 4.8: The seasonal trends for humidity in Newport, Southwark, Lewisham and Chepstow. Similar to temperature and pressure, no reading was captured for Chepstow in winter. However, the region recorded the least humidity of 53.95% in spring, as illustrated in plot (b). Newport had the highest average humidity of 90.96% in winter and 73.54% in spring, as shown in plots (a) and (b)



(a) Monthly NO_2 average for all four regions



(b) Monthly $PM_{2.5}$ average for all four regions



(c) Monthly PM_{10} average for all four regions

Figure 4.9: Plots highlighting the varying monthly averages for the three monitored pollutants. These averages varied significantly and are an indication that some influential factors may have affected the concentration levels

Table 4.3: Pollutant summary statistics based on region

Regions	NO_2				$PM_{2.5}$				PM_{10}			
	count	mean	min	max	count	mean	min	max	count	mean	min	max
Newport	40326	-6.85	-602.37	111.99	40326	11.41	0.28	745.45	40326	12.49	0.28	746.04
Southwark	38757	4.35	-714.97	1094.81	38757	10.27	0.55	4384.20	38757	11.35	0.60	6888.54
Lewisham	32986	-15.168	-1406.17	93.06	32986	12.42	0.60	277.02	32986	13.98	0.60	424.42
Chepstow	9138	7.33	-190.18	180.30	9138	7.31	0.43	127.45	9138	12.11	0.431	179.02

having the highest. Lewisham has the highest $PM_{2.5}$ and PM_{10} average of $12.42\mu g/m^3$ and $13.98\mu g/m^3$, respectively. This analysis and the plot in Figure 4.9 reveal some prevalent calibration issues within the recorded values, which were sometimes exaggerated, as in the case of the maximum values for $PM_{2.5}$ and PM_{10} . Nevertheless, a one-way ANOVA variance test carried out to check the variance in NO_2 , $PM_{2.5}$ and PM_{10} by region resulted in p values of $2.36e^{-4}$, $1.45e^{-3}$ and $1.68e^{-4}$, respectively. This result indicates that the impact of regions on the concentration levels of these three pollutants is notable.

4.4.3 Forecasting Model Training and Evaluation

Fastai was used for data pre-processing and model training. The library is built on the PyTorch framework and allows quick analysis using its readily encoded best practices. The aim was to develop a model capable of efficiently making hourly predictions of the pollutant of interest. This section introduces the data processing procedure, the network’s architecture used for training and the validation method.

4.4.3.1 Data Description

The dataset provides an extensive collection of weather parameters essential for understanding and modelling traffic-related air pollution. Key features include geographical coordinates, timestamp, temperature metrics (temp, temp_min, temp_max, feels_like), atmospheric conditions (pressure, humidity), wind metrics (wind_speed, wind_dir), cloud cover, and precipitation metrics (Rain_1h, Rain_3h, Snow_1h, Snow_3h). Table 4.4 shows a summary statistics

of the dataset with temperature averaging around 51° F with moderate variability, ranging from 21.33° F to 87.15° F. Perceived temperature, considering factors like wind chill and humidity, averages at 44.17° F. Rainfall and snowfall metrics indicate occasional precipitation events, with high variability in 1-hour rainfall. Atmospheric pressure is relatively stable, averaging 1014.02 hPa, while humidity levels are generally high, averaging 81.47%. Wind speed shows significant variability, with an average of 10.47 m/s, and wind direction covers a wide range with a mean of 181.82 degrees. Cloud cover averages 50.47%, indicating varied sky conditions.

Table 4.4: Descriptive statistics for the dataset

	lat	lon	date	Rain_1h(mm)	Rain_3h(mm)	Snow_1h(cm)	Snow_3h(cm)	temp(F)	temp_min(F)	temp_max(F)	feels_like(F)
count	991662	991662	991662	18000	650	200	400	991662	991662	991662	991662
mean	51.454	-2.587	15511	1.057	0.841	2.030	0.425	50.966	48.301	53.566	44.174
std	1.789	0.2935	2.100	2.318	0.554	0.000	0.510	10.120	10.227	10.164	12.214
min	51.454	-2.587	15142	0.110	0.130	2.030	0.130	21.330	18.000	23.000	5.310
25%	51.454	-2.587	15329	0.250	0.310	2.030	0.175	43.920	43.900	46.400	35.185
50%	51.454	-2.587	15511	0.510	1	2.030	0.190	49.690	46.900	52.000	42.480
75%	51.454	-2.587	15690	1.150	1	2.030	0.440	58.190	50.100	60.800	53.445
max	51.454	-2.587	15870	42.930	2.690	2.030	1.190	87.150	84.550	90.000	86.540

	pressure(hPa)	humidity	wind_speed(knots)	wind_dir(degrees)	clouds_all(%)
count	991662	991662	991662	991662	991662
mean	1014.024	81.471	10.471	181.816	50.469
std	11.664	14.981	5.472	93.663	36.235
min	968	13	0.360	0	0
25%	1007	73	6.930	100	8
50%	1015	86	9.170	210	68
75%	1022	93	13.870	260	90
max	1049	100	38.030	360	100

4.4.3.2 Meteorology Data Integration and Data Set Pre-Processing

Weather data such as wind speed and direction, precipitation, visibility, pressure, cloud cover, dew point, and wind gust which were not captured by the REVIS devices, were integrated

from OpenWeather (sample data available at ¹). Also, ozone data from the AURN stations were integrated into the data set to be analysed and used for training. These integration exemplify the integration capabilities of the framework while enriching the data needed to train an estimation model. Appendix A presents a complete list of the columns, their description and data types before processing. An SQL procedure for automatically generating SQL codes such as the one illustrated in Figure 4.10 was implemented to summarise the pollution data. This generated hourly, 3-hourly and 6-hourly average of the pollutant concentration levels with the aim of capturing periodicity within the training data (sample of integrated data shown in Figure 4.11).

```

SELECT city_name, lat, lon, TO_CHAR(edate, 'yyyy-mm-dd hh24:mi:ss') edate, Rain_desc, Rain_1h, Rain_3h, Snow_1h, Snow_3h,
Drizzle_desc, Fog_desc, Clouds_desc, Haze_desc, Mist_desc, Clear_desc, Snow_desc, Thunderstorm_desc, temp, temp_min,
temp_max, feels_like, pressure, humidity, wind_speed, wind_deg, clouds_all,
ROUND(Ozone, 4) Ozone, ROUND(AVG(Ozone) OVER (ORDER BY edate ROWS BETWEEN 6 PRECEDING AND 1 PRECEDING), 4) Ozone_avg6h, Ozone_factor,
ROUND(no, 4) no, ROUND(AVG(no) OVER (ORDER BY edate ROWS BETWEEN 6 PRECEDING AND 1 PRECEDING), 4) no_avg6h, no_factor,
ROUND(no2, 4) no2, ROUND(AVG(no2) OVER (ORDER BY edate ROWS BETWEEN 6 PRECEDING AND 1 PRECEDING), 4) no2_avg6h, no2_factor,
ROUND(nono2, 4) nono2, ROUND(AVG(nono2) OVER (ORDER BY edate ROWS BETWEEN 6 PRECEDING AND 1 PRECEDING), 4) nono2_avg6h, nono2_factor,
ROUND(pm10, 4) pm10, ROUND(AVG(pm10) OVER (ORDER BY edate ROWS BETWEEN 6 PRECEDING AND 1 PRECEDING), 4) pm10_avg6h, pm10_factor,
ROUND(pm25, 4) pm25, ROUND(AVG(pm25) OVER (ORDER BY edate ROWS BETWEEN 6 PRECEDING AND 1 PRECEDING), 4) pm25_avg6h, pm25_factor
FROM (SELECT city_name, lat, lon, edate, TO_CHAR(edate, 'mm-dd hh24') edate_hr, Rain_desc, Rain_1h, Rain_3h,
Snow_1h, Snow_3h, Drizzle_desc, Fog_desc, Clouds_desc, Haze_desc,
Mist_desc, Clear_desc, Snow_desc, Thunderstorm_desc, temp, temp_min, temp_max, feels_like,
pressure, humidity, wind_speed, wind_deg, clouds_all,
nvl(last_value(nullif((CASE WHEN Ozone<0 THEN null ELSE Ozone END), 0)) IGNORE NULLS OVER (ORDER BY edate), 0) Ozone,
nvl(last_value(nullif((CASE WHEN no<0 THEN null ELSE no END), 0)) IGNORE NULLS OVER (ORDER BY edate), 0) no,
nvl(last_value(nullif((CASE WHEN no2<0 THEN null ELSE no2 END), 0)) IGNORE NULLS OVER (ORDER BY edate), 0) no2,
nvl(last_value(nullif((CASE WHEN nono2<0 THEN null ELSE nono2 END), 0)) IGNORE NULLS OVER (ORDER BY edate), 0) nono2,
nvl(last_value(nullif((CASE WHEN pm10<0 THEN null ELSE pm10 END), 0)) IGNORE NULLS OVER (ORDER BY edate), 0) pm10,
nvl(last_value(nullif((CASE WHEN pm25<0 THEN null ELSE pm25 END), 0)) IGNORE NULLS OVER (ORDER BY edate), 0) pm25
FROM emissions_main) JOIN emission_factors USING (city_name, edate_hr);

```

Figure 4.10: Auto-SQL generation to pre-process the data set. An SQL command which generates 3-hour and 6-hour pollutant averages from the preceding readings is depicted.

FEELS_LIKE	PRESSURE	HUMIDITY	WIND_SPEED	WIND_DEG	OZONE_PREV1H	OZONE_PREV2H	OZONE_PREV3H	OZONE_PREV4H	OZONE_PREV5H
72.989999	1024.000000	13.000003	10.290000	39.999998	118.867102	116.905410	116.057104	110.808298	110.490195
38.970001	1017.000000	76.000000	16.110001	260.000002	76.970596	76.970597	80.746400	77.240303	75.568198
40.099999	1026.000000	100.000001	6.930000	240.000001	34.425801	49.293800	51.688599	53.085602	33.178500
53.759998	1000.000000	100.000001	8.050000	220.000000	23.998301	23.998301	23.998301	23.998301	23.998300
45.279999	1004.000000	61.000000	11.410000	279.999997	78.618401	74.202905	55.625199	58.963799	69.248802
36.369999	1031.000001	86.000000	8.050000	169.999999	27.790099	27.241301	40.263302	40.013802	41.660198
15.400000	1006.000000	59.000000	19.459999	60.000005	76.849099	69.947800	70.881797	71.452598	70.051499
29.620001	1010.000000	71.000000	8.050000	310.000001	85.109703	88.281503	85.920302	76.492996	83.178901
53.709999	1016.000000	93.000000	8.050000	220.000000	34.425801	41.460701	44.603901	46.898998	41.061500
53.959999	1023.000000	93.000000	9.170000	229.999998	34.575500	36.321701	34.791698	23.050300	34.226300

Figure 4.11: Sample of integrated weather dataset after pre-processing.

¹<https://doi.org/10.17632/b8dw3w868h.1>

Three key data pre-processors: *categorify*, *fillMissing* and *normalize* from *fastai* were adopted for additional data pre-processing. These pre-processors map categorical columns to distinct categories, replaces null values with column median values and normalises continuous columns by subtracting the mean and dividing by the standard deviation. The “*add_datepart*” helper function of the library allows the specification of the date column which generates additional predictors such as “*Year*”, “*DayofWeek*”, “*DayOfYear*”, “*Is_Month_End*” and so on. Appendix B provides a detailed list of independent/dependent and categorical/continuous variables in the data set after processing.

4.4.3.3 Validation Set Creation and Training Architecture

Model training typically begins by splitting the dataset into training, validation, and test sets. The training data is used to train the model, while the validation data helps in selecting the best-performing model, which is then verified using the test data. When dealing with target imbalance, it is customary to randomise the dataset before splitting, known as stratification. However, since this problem resembles a time-series problem where the chronological order is crucial, the validation and test sets cannot be randomly selected. Instead, the common practice is to use the most recent weeks or months of data for validation and testing Duan et al. (2023). In this case, the dataset, comprising approximately 991,662 rows and 34 columns, showed no significant target imbalance for the three pollutants, making stratification unnecessary. Therefore, the last 45 days of the dataset, covering July and August, were chosen for validation (15 days) and testing (30 days), representing 10% of the dataset. *Fastai*’s *TrainTestSplitter* class was employed to implement this division.

Suitable optimisers, loss functions and activation functions had to be selected from an array of available options. Series of experimentation were carried out on popular optimisation functions such as *SGD*, *RMSProp*, *LAMB*, *LARS* and *Adam* and regression loss functions like *BCELossFlat*, *MSELossFlat* and *L1LossFlat* before deciding the most suitable. Eventually,

Adam optimiser and *MSELossFlat* were chosen for model training. *Bayesian-optimization* library was used to test and optimise the number of architecture layers, the size of each layer and dropout rates for the network. The final architecture used to train the model was made up of 14 embedding layers, 3 dropout layers, 3 batchnorm1d layers, 3 linear layers and 2 ReLU activation functions. The embedding layer was adopted for improved performance as inspired by the architecture proposed in Guo & Berkhahn (2016). Finally, the learning rate finder (*lr_finder*) function of *TabularLearner* class was used to determine the best learning rate to be used for training. This resulted in a minimum value of $2.5e^{-4}$, and steep value of $1.3e^{-4}$. Figure 4.12 below shows the plot of the learning rate against the loss. Experts recommend selecting the learning rate at the point where the plot starts to dip. (i.e., 10^{-4}).

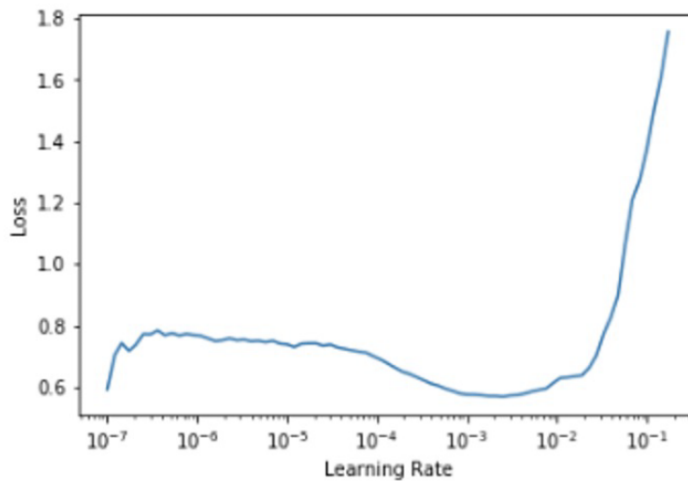


Figure 4.12: The model’s training loss against the learning rate to determine the appropriate learning rate. The learning rate was fixed at the point where the plot started dipping (i.e., 10^{-4})

4.4.3.4 Model Evaluation

In this section, the results of the deep learning model developed are presented. The model was trained to make day-ahead predictions of the three pollutants, but first, an appropriate evaluation metric had to be selected. The top metrics for regression problems are mean squared error/root mean squared error(MSE/RMSE), mean absolute error(MAE) and R

Square. The fastai library has two variants of RMSE: *rmse* and *exp_rmse*. The mean absolute error and root mean squared error (*exp_rmse* variant), defined as shown in equations 4.2 and 4.3 below, were selected as the metrics for evaluating the developed model.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|^2} \quad (4.3)$$

Figure 4.13 illustrates the model training and validation losses after 20000 epochs. It is noteworthy that the training loss gradually as the number of epoch increased. The validation loss took a slightly different pattern and dropped significantly after 2500 epochs but became steady for the remaining training epochs. The final MAE and exponential RMSE after training were 0.350 and 1.591 respectively. Figure 4.14 captures the actual NO_2 concentration levels (highlighted in blue) and the model's day ahead prediction(highlighted in red). The difference in the model's predicted NO_2 and actual values is slight, and the predicted values were close to the actual.

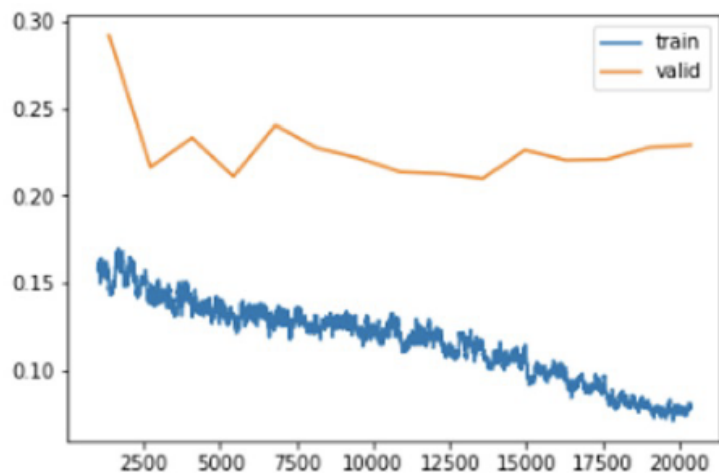


Figure 4.13: A plot showing the model’s training and validation losses against the number of epochs. It is worth noting that there was a gradual decrease in both losses as the training epochs increased which indicates that the model was learning. Further training beyond 20000 epochs would have either resulted in overfitting or no further drop in both losses

```
learn.show_results()
```

R	NO2_AVG6H	NO2_FACTOR	NONO2_AVG6H	NONO2_FACTOR	PM10_AVG6H	PM10_FACTOR	PM25_AVG6H	PM25_FACTOR	Elapsed	NO2	NO2_pred
4	-0.885981	-0.332012	-0.567841	0.662468	-0.162312	-0.699817	-0.219881	-0.814794	1.733253	2.034706	2.129269
6	-0.748497	0.903192	-0.447542	0.058060	-0.162312	0.790438	-0.219881	1.451507	1.800624	2.850389	3.122491
8	-1.236534	-0.112674	-0.615339	-0.334880	-0.162312	0.790438	-0.219881	0.883862	1.835902	1.898369	2.188342
7	-0.354666	2.035958	-0.354947	0.910231	-0.162312	1.176936	-0.219881	1.997850	1.832319	2.943470	3.006840
1	0.368544	0.594365	-0.066416	0.100828	-0.162312	1.176936	-0.219881	1.284325	1.764231	2.849961	3.192918
4	-1.124172	-0.726089	-0.597230	-0.589171	-0.162312	0.790438	-0.219881	0.728293	1.843507	2.016155	2.143995
9	0.467780	0.327762	1.666261	0.880060	-0.162312	-0.274668	-0.219881	-0.643668	1.757342	3.549833	3.641593
5	-0.125720	0.678973	-0.298762	0.533345	-0.162312	0.790438	-0.219881	0.549499	1.785652	2.364968	2.505113
9	0.341336	2.563796	-0.033054	0.957226	-0.162312	1.541260	-0.219881	1.607077	1.766261	3.858867	3.541561

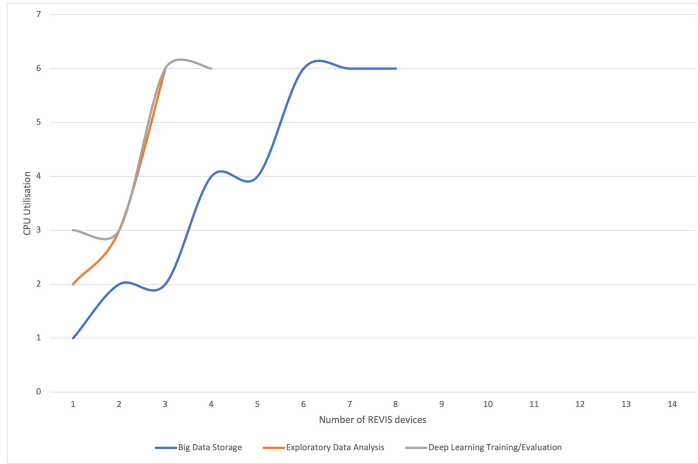
Figure 4.14: An illustration of captured NO_2 pollutant readings (blue highlight) and the deep learning model predictions (red highlight). These results were derived from an evaluation using the validation data set. It should be pointed out that the model’s predictions are not too far off the actual readings.

Table 4.5: Hardware specifications of the two oracle cloud instances used to test scalability

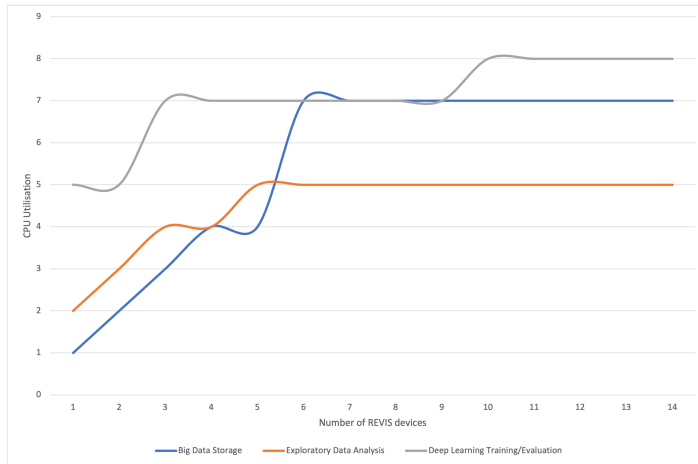
Name	Instance Type	Processor	GPU type	CPU cores	CPU memory	GPU memory
Compute A1 – OCPU	Ampere Bare Metal	OCPU	-	6	32GB	-
VM.GPU2.1	GPU	Pascal	1 NVIDIA P100	12	72GB	16GB

4.4.4 Evaluating the Scalability Performance of the REVIS System

The REVIS system was tested for scalability using the IoT asset monitoring tool and database performance hub of two different oracle cloud instances. The fourteen REVIS devices were deployed sequentially to capture both system’s response time and throughput. The first experiment was run on a bare metal cloud instance with specifications as shown in Table 4.5. Figure 4.15a shows the performance of this cloud instance as it could not scale past 8 devices and exploded at 3 and 4 devices for EDA and deep learning analysis. However, the GPU cloud instance performed better due to its auto-scale feature. Figure 4.15b shows a plot of the CPU cores utilised for exploratory data analysis, data storage and deep learning analysis as the number of deployed devices increased. It can be observed that the number of CPU cores increased gradually for each task and then stabilised at some point. The system was able to scale up its resources according to the computation/storage requirements. For the database performance, the test was run between November 2020 and Jan 2021 on the GPU instance and evaluated for utilisation, execution count, number of running statements and number of sessions metrics as shown in Figure 4.16. The maximum GPU utilisation was under 20% even with over 1.5 million execution queries.



(a) System performance of the bare metal instance as the number of REVIS devices increased.



(b) System performance of the GPU instance as the number of REVIS devices increased.

Figure 4.15: Plots of bare metal vs GPU instance as number of devices increased

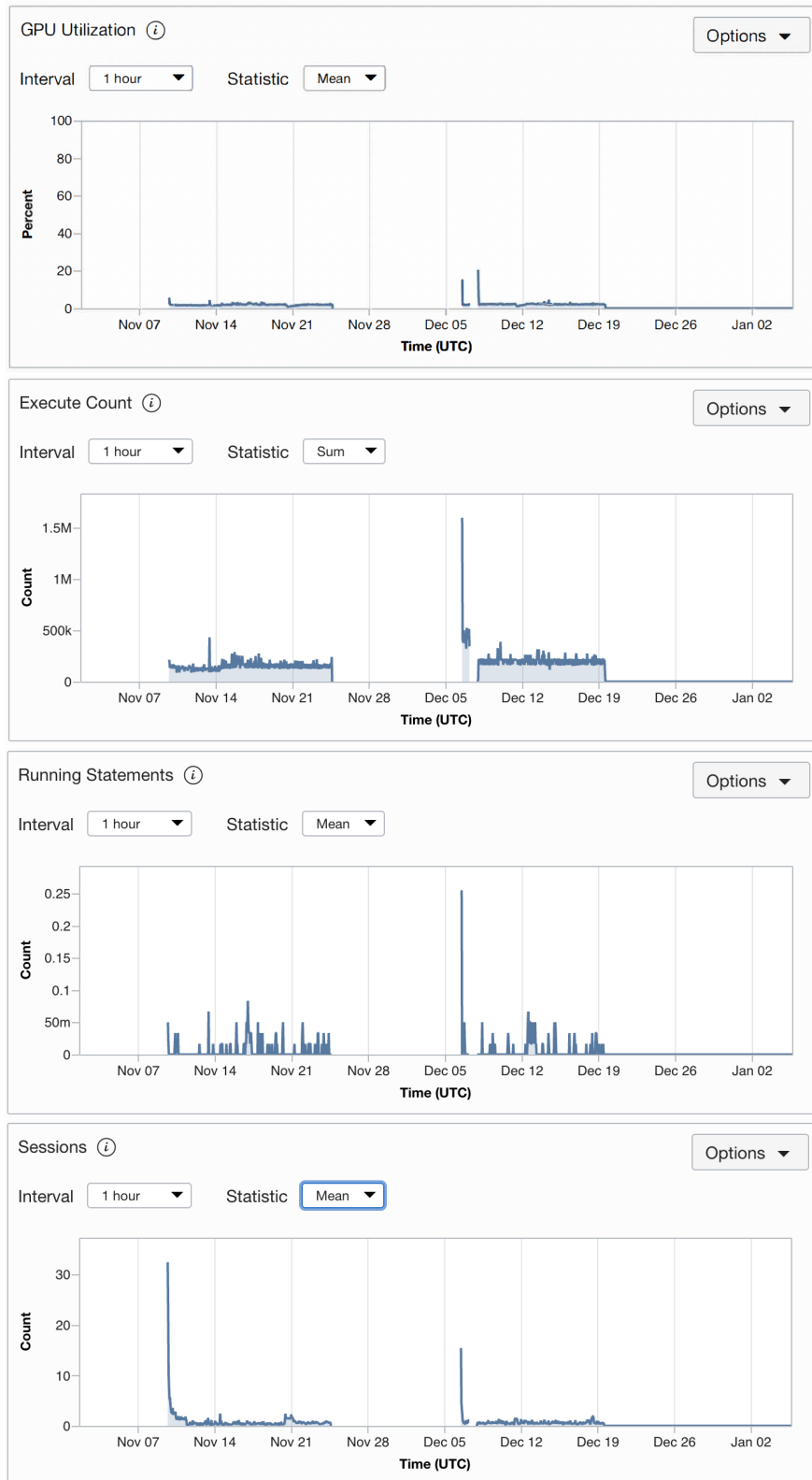


Figure 4.16: Plots of scalability metrics showing database performance as the number of devices increased

4.5 Chapter Summary

This chapter introduced a cost-effective framework for ubiquitous monitoring and predicting pollutant concentration levels on highways in the UK. The practical implementation of this framework is also demonstrated through the utilisation of the REVIS (Real-time Emission Visualisation System) system. This demonstration showcases comprehensive insights into the development of IoT hardware for data collection, the configuration of advanced big data tools for efficient data storage, and the presentation of results obtained from trained deep learning forecasting models. In addition, the chapter also highlights the scalability aspect of the framework, as demonstrated by its deployment on two distinct cloud instances equipped with varying computational resources. This chapter effectively illustrates that the realisation of real-time monitoring and forecasting capabilities is within reach when supported by adequate computational resources.

Chapter 5

Multi-target Regression for TRAP forecasting

5.1 Chapter Overview

This chapter delves into an innovative expansion of the deep learning-based air quality forecasting method adopted in Chapter Four by integrating crucial highway information with traditional meteorological and pollution data. This integration provides an improved understanding of the multifaceted factors influencing air quality, marking a significant contribution to the field. A novel approach is proposed in which a single Multi-Target Regression (MTR) model trained on the integrated data set is used to predict multiple traffic-related pollutants simultaneously, streamlining forecasting processes and potentially performing better. The chapter also explores the use of categorical embeddings within tabular data models and compares their performance against established time series and regression algorithms using cutting-edge libraries. This comparison provides valuable insights into the most effective modelling techniques for air quality prediction. To comprehensively interpret these models, the feature importance within the data set is evaluated, identifying key contributors to air quality, which not only aids in model interpretation but also guides future data collection endeavours.

5.2 Monitoring Site and Integrated Third-Party Data

The data collection sites remains consistent with those described in the preceding chapter, but it now incorporates additional highway data. Supplementary weather data, which was not captured by the REVIS devices were integrated from the AURN stations nearest to the highways of interest. Publicly accessible background mapping data was used to augment this study's analysis and obtained from the Department for Environment, Food and Rural Affairs (DEFRA) website ¹. Similarly, their emissions factor toolkit was employed to estimate traffic exhaust emissions across various vehicle categories. Highways England's webtris application ² provided essential data on traffic congestion, average vehicle speed, and traffic volume. Lastly, terrain information for the case study sites was extracted using the Google Earth application. Detailed descriptions of these data sets are provided in the subsequent sections.

5.3 Data Description

The approach used to collect data in this study was to imagine the highways as consisting of multiple segments. Deployed devices were mapped to different segments of the highway and data captured for each device represented that highway segment. This way, it was easier to match device measurements with other data set such as background concentration that are represented by 1x1km grids. This section describes the data set specification which is also summarised in Appendix C.

5.3.1 Pollution Data

NO_2 , PM_{10} , and $PM_{2.5}$ data captured every five minutes by the REVIS devices were included in the data set. After collocating the NO_2 readings of the devices with the nearest AURN

¹<https://uk-air.defra.gov.uk/data/laqm-background-home>

²<https://webtris.highwaysengland.co.uk/>

stations in Chepstow ³, Newport ⁴ and London (Lewisham ⁵ and Southwark ⁶) it was clear that the NO_2 readings were inaccurate with the average correlation of 0.07. This inaccuracy was linked to the analogue NO_2 sensors used on the REVIS devices, which responded strongly to changes in temperature and relative humidity, to get negative readings sometimes. As a result, NO_2 measurements from AURN stations were used in place of the REVIS NO_2 data. The REVIS data for PM_{10} , and $PM_{2.5}$ were retained since there was a good correlation of 0.73 and 0.8 with the AURN data. To ensure efficient data mapping, the REVIS data had to be summarised into hourly aggregates to match the hourly readings in the integrated AURN data (see Figure 5.1).

	ROAD_NAME	REGION_NAME	SEGMENT_NAME	PM2.5 ($\mu\text{g}/\text{m}^3$)	PM10 ($\mu\text{g}/\text{m}^3$)	NO2 (ppb)
Rownum						
0	A48 Road	Chepstow	A48 Road - S02	NaN	NaN	39.10924
1	A48 Road	Chepstow	A48 Road - S02	NaN	NaN	48.46526
2	A48 Road	Chepstow	A48 Road - S02	NaN	NaN	48.88828
3	A48 Road	Chepstow	A48 Road - S02	NaN	NaN	63.21779
4	A48 Road	Chepstow	A48 Road - S02	NaN	NaN	48.33805

Figure 5.1: Snapshot of pollution data.

5.3.2 Traffic Data

Traffic information was integrated from Highways Englands' traffic monitoring unit (TMU) sites. The data which can be downloaded through an API or a web interface includes counts for vehicles less than 5.2m or greater than 11.6m in length, counts for each vehicle type, total traffic volume and average traffic speed. The measurements only included descriptions of vehicle lengths so it was necessary to map different vehicle types to the appropriate lengths for easy comprehension. Cars were mapped to 0-520cm, buses to 521-660cm, light goods vehicle (LGV) to 661-1160cm and heavy goods vehicle (HGV) to 1160cm+ (Bálint et al.

³https://uk-air.defra.gov.uk/networks/site-info?site_id=CHP

⁴https://uk-air.defra.gov.uk/networks/site-info?site_id=NPT3

⁵https://uk-air.defra.gov.uk/networks/site-info?site_id=LW1

⁶https://uk-air.defra.gov.uk/networks/site-info?site_id=SK5

2014). TMU data are captured every minute so just like the historic pollution data, this data was also summarised into hourly aggregates(see Figure 5.2).

Rownum	Bus Avg Speed(mph)	Bus Count	Car Avg Speed(mph)	Car Count	HGV Avg Speed(mph)	HGV Count	LGV Avg Speed(mph)	LGV Count
0	54.00	40.0	54.00	380.0	54.00	31.0	54.00	32.0
1	53.50	43.0	53.50	506.0	53.50	25.0	53.50	32.0
2	53.25	33.0	53.25	295.0	53.25	24.0	53.25	32.0
3	53.50	17.0	53.50	448.0	53.50	22.0	53.50	9.0
4	52.50	58.0	52.50	731.0	52.50	32.0	52.50	24.0

Figure 5.2: Snapshot of traffic data.

5.3.3 Weather Data

The temperature, humidity and pressure for the four highways of interest were measured in real-time along with pollution data. However, previous studies have shown the impact of other meteorological parameters such as wind speed and wind direction in aiding pollutant dispersion (Chen & Ye 2019). The modelled wind speed and direction data were therefore integrated from same AURN stations used for NO_2 while data from REVIS devices were aggregated to match. Wind direction across the four regions ranged between 16° and 360° and the wind speed was between 0 and 16 knots (see Figure 5.3).

Rownum	Humidity(phi)	Wind Direction(degrees)	Wind Speed(knots)	Temperature(celcius)	Pressure(hPa)
11985	79.669189	329.8	10.0	7.233315	1000.168030
11986	76.290894	311.6	6.7	9.282722	1006.908813
11987	76.371329	336.0	8.8	8.219354	1001.844552
11988	78.375244	325.1	6.1	8.699036	1008.303304
11989	70.254517	336.5	8.1	11.230164	1002.729004

Figure 5.3: Snapshot of weather data.

5.3.4 Elevation Data

Research into emission modelling in recent years has shown that vehicle exhaust outputs varies in uphill and downhill situations (Zhai et al. 2020, Xu et al. 2020). The vehicle’s engine is under more pressure as it goes uphill and under less pressure downhill. It is unknown whether capturing this sort of highway information would result in an improved estimation accuracy. More importantly, highway terrain data such as elevation and gradient data are required to compute the vehicle emissions factor for different vehicle types. Google Earth’s desktop application was used to capture this information. The elevation data for a designated road trajectory was obtained by drawing the path on the application and selecting “Show Elevation Profile” from the right-click menu. This generated a graph displaying the elevation along the path. Specific elevation points were then manually recorded from the profile, as direct data export is not supported (see Figure 5.4).

Highway Elevation(m)	
Rownum	
11986	68.932396
11987	3.902021
11988	68.932396
11989	3.902021
11990	68.932396

Figure 5.4: Snapshot of elevation data.

5.3.5 Emissions Factor Data

Version 11.0 of DEFRA’s emission factor toolkit (EFT) was used to compute the source apportionment of particulate matter and NO_2 for the different vehicle categories. EFT allows the specification of parameters such as the year of interest, road type, vehicle speed and vehicle type from the onset and automatically computes the required output based on COPERT 5 specifications (COPERT is the standard EU vehicle emissions calculator). The traffic was selected as ‘Detailed Option 2’ since the traffic data that was collected did not

include information on vehicle types as either petrol, diesel or hybrid. This option allows non-detailed vehicle counts for cars, buses, LGVs and HGVs to be used as traffic flow input for EFT. The highway gradient information from Google earth was also fed into the tool while the ‘*flow direction*’ was determined from the elevation chart in the application. As a result, the Newport, Lewisham and Southwark highways were specified as ‘Up Hill’ while Chepstow was specified as ‘Down Hill’ flow direction due to the single direction by which vehicles travelled. Finally, the below equations were used to verify the estimations from the toolkit and the values were close.

$$\text{For Uphill: } EF_2 = EF_1(1 + G \times [C_1 \times V + C_2]) \quad (5.1)$$

$$\begin{aligned} \text{For Downhill: } EF_2 &= EF_1(1 - G \times [C_1 \times V + C_2]) \text{ if } G \leq 2.5\% \\ EF_2 &= EF_1(1 - 0.025 \times [C_1 \times V + C_2]) \text{ if } G > 2.5\% \end{aligned} \quad (5.2)$$

where EF_1 and EF_2 denote emission factor for vehicles travelling at speed V on a level and uphill/downhill road respectively, G is the highway gradient and C_1 and C_2 are the gradient coefficients based on vehicle type and pollutant of concern (CERC 2019)(see Figure 5.5).

	No2 Emission Factor	PM Emission Factor
Rownum		
0	1735.0	2999.93920
1	2242.0	3836.07528
2	1367.0	2364.85630
3	1883.0	3193.08780
4	3178.0	5374.81560

Figure 5.5: Snapshot of emission factor data.

5.3.6 Background Air Pollution Concentration Data

Background concentration maps for a particular pollutant provide data on contributions from various sources, including natural, local sources such as household coal burning, industries, and other means of transportation, mixed with contributions from the source of interest—in this case, road transport. It is crucial to consider and eliminate these other sources to avoid double counting, where the pollutant concentration is unknowingly repeated. Therefore, in this study, the highway contribution was subtracted from the background concentration to address this issue. Publicly available background pollution maps from the DEFRA UK AIR resource website (UKAIR 2018) were used to capture this information for the four case study locations. It is important to note that these were 2018 background maps covering the years 2020 and 2021, and they do not account for the long-term or short-term impacts of COVID-19 lockdowns on local sources. The data provides grid-based modeled background concentrations for $PM_{2.5}$, PM_{10} , NO_x , and NO_2 from 2018 to 2030. The background concentrations for 2020 and 2021, as indicated in Table 5.1, include only rail, domestic, industrial, and point sources. Minor road and motorway background concentrations were excluded to prevent double counting. This approach is consistent with the method proposed in the study by Arunachalam et al. (2014) (see Figure 5.6).

Table 5.1: Pollutant background concentration for the four regions of interest in the year 2020 and 2021

Regions	Grid_ref.x	Grid_ref.y	$NO_2(ppb)$		$PM_{2.5} (\mu g/m^3)$		$PM_{10} (\mu g/m^3)$	
			2020	2021	2020	2021	2020	2021
Newport	332500	189500	17.711	16.761	10.386	10.278	15.785	15.648
Chepstow	353500	193500	8.409	8.067	7.986	7.883	12.069	11.941
Lewisham	537500	177500	24.698	23.827	12.090	11.941	18.560	18.347
Southwark	531500	178500	28.954	27.997	12.706	12.555	19.768	19.552

Rownum	Background NO2 (ppb)	Background PM2.5 ($\mu\text{g}/\text{m}^3$)	Background PM10 ($\mu\text{g}/\text{m}^3$)
0	8.409935	7.986203	12.06984
1	8.409935	7.986203	12.06984
2	8.409935	7.986203	12.06984
3	8.409935	7.986203	12.06984
4	8.409935	7.986203	12.06984

Figure 5.6: Snapshot of background concentration data.

5.4 Machine learning approach

This section describes the approach taken in this study to address the multi-target prediction problem. The pseudo-code for the proposed approach is highlighted below while the entire workflow is summarised in Figure 5.7.

Algorithm 1: Multi-target algorithm for predicting NO_2, PM_{10} and $PM_{2.5}$.

Input: data set $\mathcal{D}(X, Y)$, Fastai tabular model \mathcal{F} , Prophet model \mathcal{P} , Multioutputregressor model

\mathcal{M} , epochs ϵ , learning rate η , batch size β , estimators n , max depth d

Output: $(\hat{y}_1, \hat{y}_2, \hat{y}_3)$

Initialize: ϵ, η, β

Categorify(\mathcal{D})

FillMissing(\mathcal{D})

Normalize(\mathcal{D})

Split \mathcal{D} into trainSet, testSet and validationSet

for $e = 1, \dots, \epsilon$ **do**

 train \mathcal{F} using trainSet, η and β

 validate(\mathcal{F} , validationSet)

end

Return: Trained tabular model $\mathcal{F}_{trained}$

Initialize: \mathcal{P}

for x_i, \dots, x_n **do**

\mathcal{P} .addRegressor(x)

end

train \mathcal{P} using trainset

validate(\mathcal{P} , validationSet)

Return: Trained model $\mathcal{P}_{trained}$

Initialize: n, d, \mathcal{M}

train \mathcal{M} using trainSet, n and d

validate(\mathcal{M} , validationSet)

Return: Trained model $\mathcal{M}_{trained}$

for $model \in (\mathcal{F}_{trained}, \mathcal{P}_{trained}, \mathcal{M}_{trained})$ **do**

for $t = 1, \dots, 24$ **do**

Get: x_t

if $t \neq 1$ **then**

Predict: $(\hat{y}_1, \hat{y}_2, \hat{y}_3)_t$ using (model, $(\hat{y}_1, \hat{y}_2, \hat{y}_3)_{t-1}, x_t$)

else if $t = 1$ **then**

Predict: $(\hat{y}_1, \hat{y}_2, \hat{y}_3)_t$ using (model, x_t)

Return: \mathcal{F} : $(\hat{y}_1, \hat{y}_2, \hat{y}_3)_t, \mathcal{P}$: $(\hat{y}_1, \hat{y}_2, \hat{y}_3)_t, \mathcal{M}$: $(\hat{y}_1, \hat{y}_2, \hat{y}_3)_t$

end

end

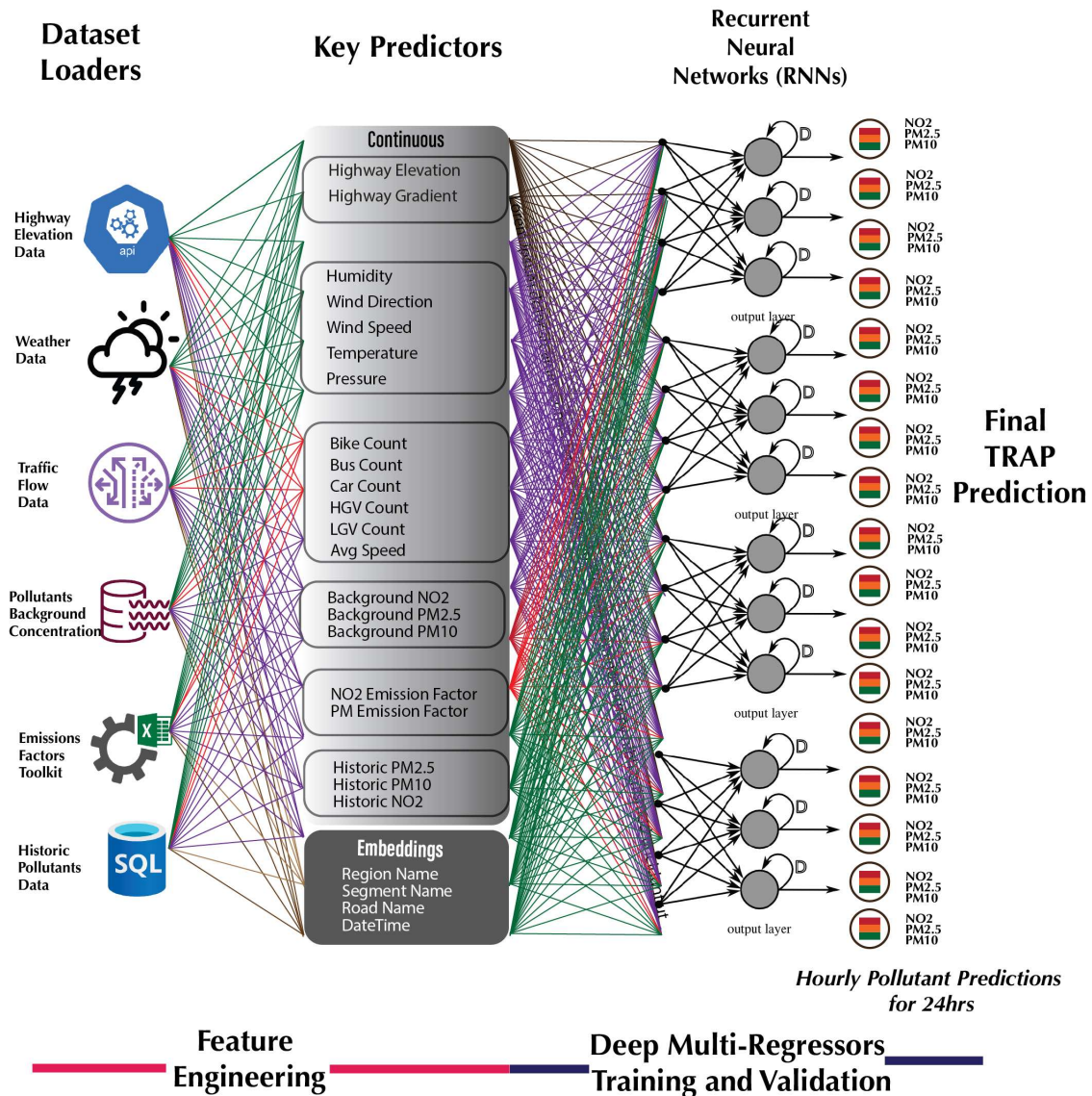


Figure 5.7: Multi-target model training architecture using the newly curated data set. Feature engineering steps including normalisation and log transformation were carried out before training on three different algorithms used for experimentation.

5.4.1 Multi-target regression and RNNs

Neural Networks have become a familiar term among the artificial intelligence (AI) and machine learning research community. The ML approach which became more popular in 2012 as a result of its performance at the imagenet classification competition, has since grown into a widely adopted method for not just classification but also regression problems. Multi-target models in general refers to models that are able to automatically detect relationships between target variables, thereby resulting in better predictions (Korneva & Blockeel 2020). A multi-target regression neural network differs from its single-target counterpart by the number of predicted outputs. As illustrated in figures 5.8a and 5.8b, single-target predicts just one output using the set of features characterising the data set while multi-target can predict multiple outputs simultaneously. In terms of performance, multi-target outputs are simpler and faster to train than an ensemble of single-target models (Kocev et al. 2009). Multi-target models are more widely adopted for classification problems such as object classification, face recognition and sporadically used for regression problems (Spyromitros-Xioufis et al. 2012).

Recurrent neural networks are mainly associated with research involving time-series, sequence labelling and classification using visual, audio or text data. This class of neural networks and its variants - Gated Feedback Recurrent Neural Network (GRU) and Long-Short term memory (LSTM) are suitable for time-series problems since they are capable of keeping track of the temporal information within input data. Other neural network architectures like CNN and GANs struggle with these kind of data (Yu et al. 2019). Despite the competitiveness of RNNs over other architectures, its application to domains such as air quality forecasting is limited due to the inadequate understanding of its internal mechanisms (Shen et al. 2020). Fortunately, several libraries and frameworks have been introduced in recent times to take away the intricacies of the RNN implementation.

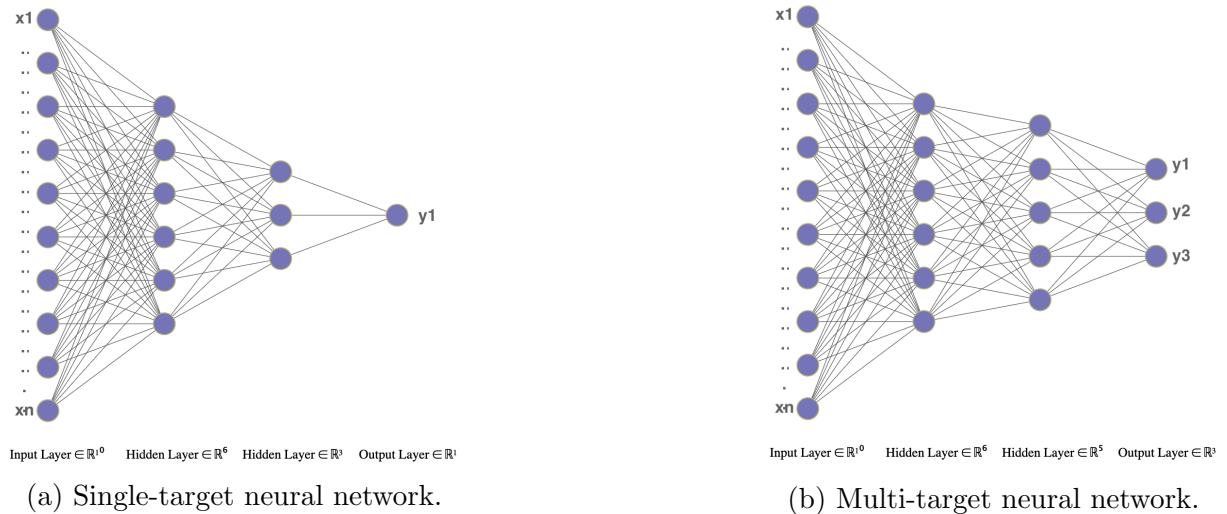


Figure 5.8: Multi-target vs single-target neural networks.

5.4.2 Fastai, prophet and multioutputregressor methods

Fastai was first introduced in 2016 as a library built with a high level of abstraction to help AI enthusiasts with limited maths background to quickly develop deep learning models. With as little as 10 lines of codes, the complexities of developing such models are handled by fastai’s customisable low, mid and high level APIs (Howard & Gugger 2020). The library is put forward as being capable of achieving state-of-the-art results in computer vision, natural language processing, collaborative filtering, and time-series problems. Another key attribute of the library which has caught the eye of researchers is the library’s implementation of entity embeddings for encoding categorical features to achieve state-of-the-art results.

Prophet, on the other hand, is a library developed by Facebook to strategically introduce some modifications to traditional time-series algorithms. The library uses the idea of “changepoints” to generate additive regression models capable of automatically detecting and adapting to sudden changes in time-series trajectories (Taylor & Letham 2018). This implies a reduction in the efforts required to manually specify data shifts before training a model. The library is designed to be robust against missing data and is originally built for univariate daily, weekly and yearly time-series forecasting. However, with a few modifications to the

library, such as the use of multiple regressors, multivariate prediction is possible. The default configuration in prophet is known to produce estimates similar to professional forecasters and therefore encourages quick experimentation. The library is famously used for sales as well as weather forecasting. The easiest way to install prophet is through its python or R package on PyPI and CRAN repositories.

Scikit-learn (Sklearn) is one of the most useful python library that houses different regression, classification and time-series algorithms. One of the wrapper regressor classes in sklearn is the MultiOutputRegressor class which permits the definition of one regressor from any of the available regression algorithms and then creates an instance for each output. One key advantage of the class is that it can be used to identify outputs that are independent of each other and also used to evaluate the performance of other multioutput models.

The adoption of these libraries and frameworks in this study is based on their distinct advantages tailored to the needs of traffic-related air pollution forecasting. Fastai was chosen for its ability to simplify the integration of complex neural networks and efficiently handle large datasets, which are typical in traffic and air quality monitoring. Prophet was selected for its capability to facilitate quick experimentation while still producing reliable results, particularly beneficial for managing the irregularities and sudden changes inherent in traffic patterns. Lastly, Sklearn was employed for its versatility in quickly implementing and comparing different modelling approaches, ensuring that the most effective model is selected to match the specific characteristics of the dataset. These tools collectively enhance the robustness and accuracy of the forecasting models used in this study.

5.4.3 Data preprocessing

All the available data were first pulled together and merged into a single csv file using Oracle SQL procedures before preprocessing was initiated. It was important that these procedures

were used to extract the data into separate database tables since they were generated as JSON strings directly from the IoT devices. The tables were joined using matching columns such as region or highway id and then loaded into a jupyter notebook for pre-processing and data cleansing. This data fusion technique is known as the early multi-view integration approach where the data sets are first joined together into a vector using a matching feature before training on a machine learning algorithm (Noble et al. 2004, Li et al. 2018, Guarino et al. 2022). The matching feature in this case is the region/highway id. Two versions of the data were created to adapt to the needs of the algorithms that were explored. The feature engineering steps that were taken are as follows:

- Data straight from the database had 232,553 rows and 10 columns. Each row represented a single reading for particular pollutant or weather data at 5 min intervals. One of the columns captured the *trend_type_id*, an integer which indicates the type of measurement (weather, pollutant, emission factor etc) that was measured. A dictionary was then created to convert these ids into meaningful and more descriptive strings. Pandas library was used for data manipulation and its pivot function was used to turn rows with matching dates into one single row while retaining the measurement type as columns. Missing measurements for a particular time point was represented with 'Nan'. The shape of the data set after this preprocessing step was 11,990 rows x 44 columns
- Next was to create the first version of the data set which includes extracted date information. Additional date attributes such as *day*, *month*, *year*, *dayofweek*, *ismonthend* etc were added to this data set. This step makes it easier for the algorithm to extract the date information from the datetime object. The second version of the data had just the date and pollutants data like a typical time series data set.
- Inspecting the data for missing values revealed 1111 missing data for the REVIS $PM_{2.5}$ and PM_{10} while the integrated AURN NO_2 had none. The missing values were replaced

with data from the previous day using the last observation carried forward (LOCF) method which is one of the famous imputation methods for time series data (Hadeed et al. 2020). The same approach was used to fill missing values in other weather and traffic attributes.

- It was difficult to identify the underlying distribution of the pollutants since their min and max has a smaller scale of values as shown in Table 5.2. Hence, the log transform of all three pollutants was taken to make the distributions less skewed. The resulting plot of the distribution is shown in figure 5.9.
- Finally, the features were split into categorical and continuous features based on the type of values they hold as shown in Appendix C. This step facilitates the use of tabular models.

Table 5.2: Descriptive statistics of the pollutants data

Variable	count	mean	std	min	25%	50%	75%	max
NO_2 (ppb)	11990	21.954	16.405	0.631	9.753	16.910	30.379	132.370
$PM_{2.5}$ ($\mu g/m^3$)	10879	9.711	14.922	0.699	3.717	5.932	10.205	401.012
PM_{10} ($\mu g/m^3$)	10879	11.801	17.882	0.778	4.828	8.042	12.587	617.351

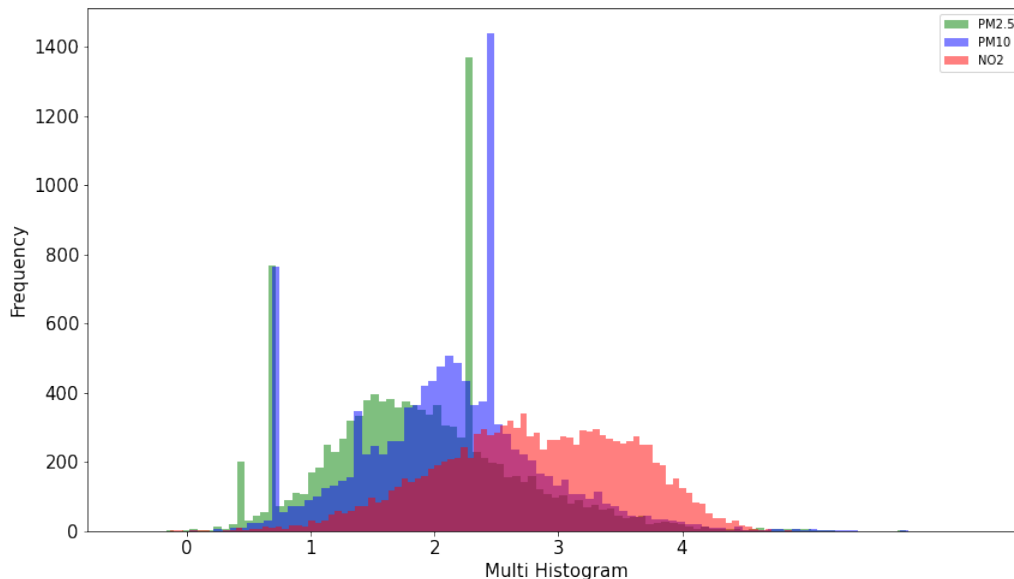


Figure 5.9: Data distribution for all three pollutants.

5.5 Experimentation and Model Training

This section highlights the experiments and optimisation techniques carried out in this study while results of each experiment are presented in subsequent sections. Figure 5.10 shows the difference between two sets of experiments carried out using fastai, prophet and multioutputregressor algorithms. Each experiment was carried out using separate jupyter notebooks and a dedicated high performance computer with 64gb RAM and Nvidia RTX 3080 GPU.

5.5.1 Experiment 1 - Comparing Fastai, Prophet and MultiOutputRegressor defaults

The first experiment involved training models with different combinations of data sets and methods. The aim was to initially try out the default configurations of the choice libraries and see how they perform with hourly, 3-hourly and 6-hourly MTR predictions before attempting any hyperparameter tuning. Out of the box, fastai permits the customisation of the number of features to predict and this can be set to as many as possible if a custom loss function is

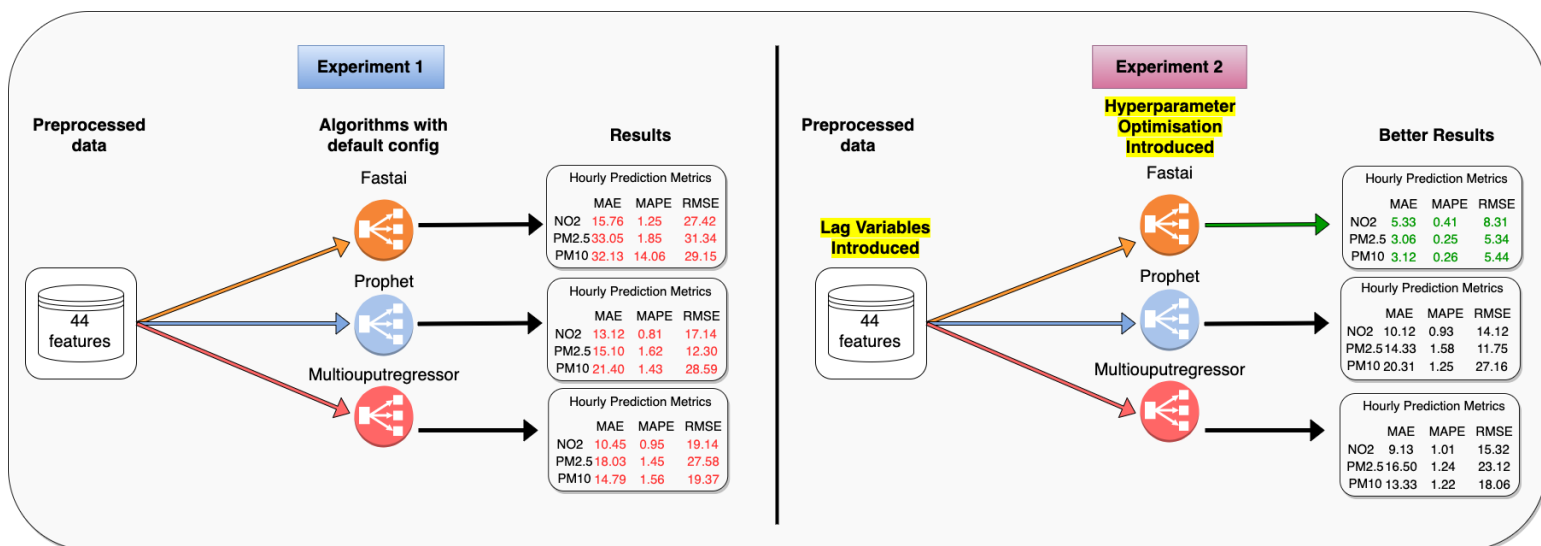


Figure 5.10: Summary of experiments carried out in this study.

Table 5.3: Hyperparameters used for experiment 1 - default configurations

Algorithm	Hyperparameter name	Hyperparameter value
Fastai	Number of layers	2
	First layer neurons	200
	Second layer neurons	100
	Dropout probability	0.04
	Learning rate	$1e^{-1}$
Prophet	Period	365
	Changepoint prior scale	0.001
MultiOutputRegressor	Number of estimators	100
	Learning rate	0.1
	Max depth	3
	Minimum samples split	2
	Minimum samples leaf	1
	Alpha	0.9

configured alongside. The default design of fastai’s tabular learner (a class within its mid-level API) is a two-layered neural network with 200 neurons in the first layer and 100 in the second layer. Other fastai default parameters and values are shown on Table 5.3.

Prophet uses a conventional time-series method of forecasting and requires just the date column and one dependent variable (y). However, for this experiment, the library’s *add_regressor* function was used to include the other features but this required that the historic and future values of these additional regressors be included during training. Since prophet does not support multi-output forecasting by default, another package called *multi-prophet* was used which allowed the prediction of all three pollutants simultaneously. Also, UK holiday effects were captured using the built-in country holidays feature.

Randomforestregressor, gradientboostingregressor and kneighboursregressor were explored with the multioutputregressor estimator classes to see which performed better. The best performing regressor with the default configurations was to then be used for subsequent experiments. Gradientboostingregressor produced the best result when compared in terms of the mean absolute error (MAE). The default configuration used is shown in Table 5.3. The result of experiment 1 is reported in section 6.6 but overall, it showed that most of the models did not perform too well and more experimentation or parameter optimisation was required.

5.5.2 Hyperparameter tuning with optuna and gridsearchcv

Following the not-so-impressive results of experiment 1, it was essential that the training parameters were optimised. Optuna is a mildly famous parameter optimisation framework for deep learning models. It was chosen for the purpose of this study due to its ease of use and also its recently introduced integration module for fastai. Optuna requires the definition of an objective function to be optimised, and in this case was defined as the model’s prediction of the three pollutants. Table 5.4 shows the search space for each of the

Table 5.4: Details of Hyperparameters optimised using Optuna and GridSearchCV

Optimiser	Hyperparameter	Search space	Result
Optuna	Number of layers	(1,7)	3
	Neurons per layer	(50,200)	200,162,134
	Weight decay	(0.01,0.1)	0.01
	Learning rate	($1e^{-5}$, $1e^{-1}$)	$1e^{-3}$
	Dropout probability	($1e^{-3}$, $1e^{-1}$)	0.2
GridSearchCV	Number of estimators	(10,300)	250
	Learning rate	($1e^{-5}$, $1e^{-1}$)	$1e^{-1}$
	Max depth	(1,40)	12
	Minimum split	(0.01,1)	0.6
	Alpha	(0.1,2)	1.3

optimised hyperparameter and the associated value after 50 optuna trials. GridSearchCV is an estimator within the sklearn library used to carry out brute force parameter search on regression algorithms such as the one being explored in this study. The technique uses cross-validation for this purpose while fitting and scoring each fold independently. GridSearchCV was used to optimise the number of estimators, learning rate, max depth, minimum sample split and alpha values for the gradientboostingregressor algorithm. Table 5.4 also shows the selected hyperparameter values after optimisation.

5.5.3 Experiment 2 - exploring lagged dependent variables (LDVs)

This experiment sought better model performance through the introduction of lagged variables. Introducing lagged variables in regression analysis is not new as discussed in the study of Wilkins (2018). The concept has been explored in several studies including air quality research with some scholars arguing that it may introduce bias in the data set if not defined properly (Grubb & Symons 1987). In this study the concept was implemented by carefully creating a structured data set which contained actual readings from previous time points leading to the current time point to be predicted. Each of these time points were depicted as

separate columns and fed into each model to be trained. The effect of this experiment was that information of the previous time points needed to be provided for any future time point. This was the sensitive bit that could easily lead to data leakage. A function was therefore written to implement this idea while sequentially predicting all the timing points leading to the current one. Results of experiment 2 are also reported in section 6.6 and it shows an improvement from the previous experiment.

5.6 Model Validation and Results

This section highlights results of the experiments carried out in this study. Details of the choice evaluation metrics and the methods used to select a suitable validation data are also highlighted.

5.6.1 Performance Metrics

Evaluation metrics are used to check the performance of models during and after training. Hence, it was necessary that suitable metrics for MTRs were first chosen even before training was started. More importantly, the metrics were also used to validate the developed models to make sure they were actually learning. Existing regression studies adopt metrics such as mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE) and mean square error (MSE) for model evaluation. Equations 5.3 to 5.5 illustrate the MAE, RMSE and MAPE metrics that were chosen as performance measures where y is the actual value and \hat{y}_i is the predicted value. For fastai, a custom loss function that could compute the model's performance for each pollutant, average it and then update the model's weights accordingly was implemented. This was an important step to force the model to learn appropriately and not perform exceptionally on one pollutant and poorly on another.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5.3)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (5.4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|^2} \quad (5.5)$$

5.6.2 Test and Validation Data

Seventy percent of the entire dataset was used for training, while the remaining 30% was split into validation (20%) and test (10%) sets. To preserve the seasonality within the data, the dataset was first sorted by date and then split by index, ensuring no randomisation occurred. Consequently, 8,953 rows were used for training, 2,398 rows for validation, and 1,199 rows for testing. In terms of days, this translated to 39 days for validation and 27 days for testing. Each data point represents hourly readings for all 44 features, with some data, such as highway elevation and emission factors, being constant for each region. The validation set was used to optimise the models' parameters after each training loop, while the test set was employed to evaluate the performance of the final model. Cross-validation, specifically K-fold cross-validation, is a widely adopted validation method in regression analysis (Morin & Davis 2017). K-fold cross-validation works by partitioning the dataset into 'k' equal-sized subsets or folds, then iteratively training and validating the model 'k' times, using a different fold as the validation set each time. This ensures that every data point is used for both training and validation, providing a comprehensive assessment of the model's performance. This method was chosen to validate and test the accuracy of the trained models.

The implementation of K-fold cross-validation in this study involved combining the training and validation data sets ($8,953 + 2,398 = 11,351$) into 5 chunks of approximately 2,270 rows each. In the first step, the first chunk was used for validation while the other chunks were used for training. In the second step, the second chunk was used for validation and

the other chunks for training. This process was repeated until all chunks had been used for cross-validation. Sklearn’s *cross_val_score* helper function was used to facilitate this cross validation process for the fastai and multioutputregressor models. For the Prophet model, this chunk is referred to as the *period*, while the number of days to be predicted is referred to as the *horizon*. The horizon was set at 1 hour, 8 hours, 16 hours, and 24 hours based on the model that was being trained and validated. This method ensured a robust evaluation of the models’ performance across different time frames and data segments.

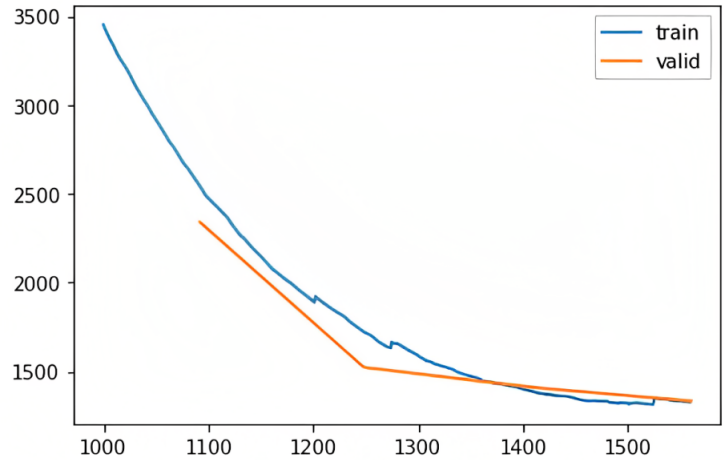
Table 5.5: Experiment 1 results of MTR models prediction for different timesteps

Pollutant & Timestep		Fastai			Multioutputregressor			Prophet		
		MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
NO_2 (ppb)	1hr	15.760	1.256	27.420	10.452	0.952	19.145	13.128	0.811	17.142
	8hr	16.321	1.076	31.329	17.334	1.772	21.768	14.372	0.816	20.099
	16hr	18.167	1.321	34.771	21.982	2.306	24.911	14.714	0.852	23.146
	24hr	21.159	1.442	35.682	23.057	2.512	21.156	15.591	0.994	26.044
$PM_{2.5}$ ($\mu g/m^3$)	1hr	33.051	1.858	31.341	18.036	1.452	27.588	15.103	1.623	12.304
	8hr	34.111	2.328	33.142	23.911	1.641	33.612	19.145	1.815	18.142
	16hr	38.440	2.416	36.189	27.105	1.952	35.145	10.232	2.012	22.356
	24hr	40.099	2.512	38.146	26.830	1.835	36.875	15.344	2.458	23.198
PM_{10} ($\mu g/m^3$)	1hr	32.130	14.063	29.156	14.798	1.568	19.376	21.403	1.434	28.599
	8hr	37.156	7.342	31.002	18.233	1.734	22.157	20.123	2.583	32.048
	16hr	38.360	10.222	35.158	21.156	1.912	28.523	22.041	5.168	37.145
	24hr	33.127	8.066	36.360	24.076	1.820	32.142	23.487	3.443	33.640

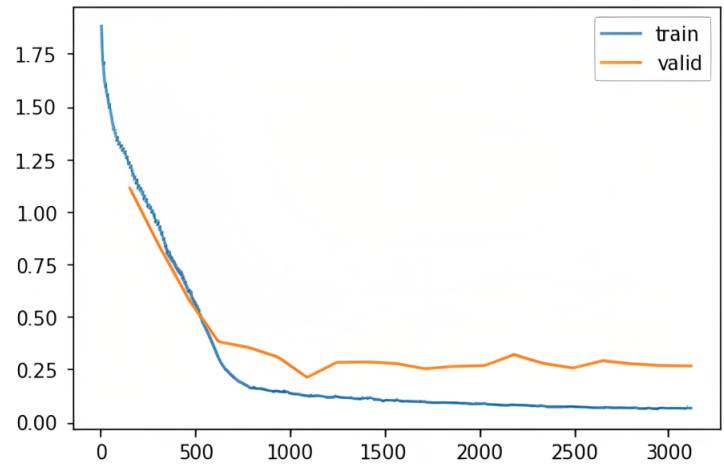
5.6.3 Experiment 1 Results

Models trained in the first experiment were evaluated over hourly, 8-hourly, 16-hourly, and 24-hourly timesteps, chosen for their practical relevance to stakeholders’ needs and consistency with previous air quality studies (Bui et al. 2018, ?). These intervals allow for immediate operational responses, alignment with work shifts and daily cycles, and support day-ahead planning. Evaluating at multiple intervals captures various temporal patterns, ensuring robust and effective models across different forecasting horizons. This comprehensive

approach provides a thorough assessment of model performance, highlighting their strengths and weaknesses, and ensuring versatility and reliability in various scenarios. Figure 5.11a shows the training and validation loss for fastai after 1,500 epochs. From the plot, it can be seen that the training loss reduced progressively but this was not indicative of the final evaluation results shown in Table 5.6. The table shows the scores recorded for each algorithm in each timestep. It is evident that all the models struggled with the 24hr and 16hr predictions and performed slightly better with the hourly and 8hr predictions. The overall minimum MAE, MAPE and RMSE 1hr scores for NO_2 in this experiment was 10.452, 0.952, 19.145 respectively with the multioutputregressor model. Likewise, the best performance for $PM_{2.5}$ was on the prophet model with 15.103, 1.623 and 12.304 scores. For the most part, fastai recorded the worst performance in this experiment with scores as high as 40.099, 2.512 and 38.146. To further strengthen the assumptions that the scores recorded on these models were too high, a graphical plot of the actual readings and models' predictions were made as illustrated in figures 5.12-5.14. None of the models were able to perform well on all three pollutants simultaneously. An ensemble of predictions from the two better performing models - multioutputregressor and prophet was also explored but there was no improvement with the achieved scores.

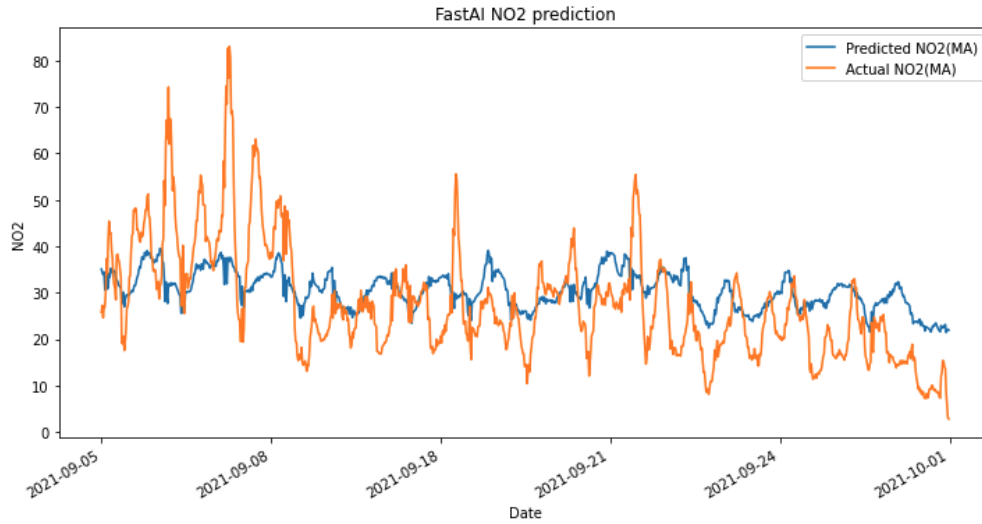


(a) Experiment 1 - Fastai's training and validation loss after 1500 epochs.

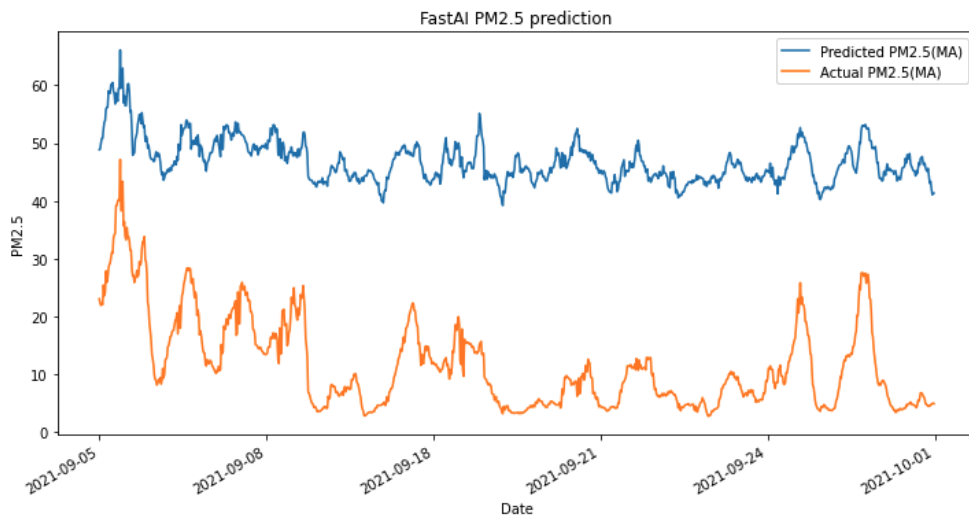


(b) Experiment 2 - Fastai's training and validation loss after 3000 epochs.

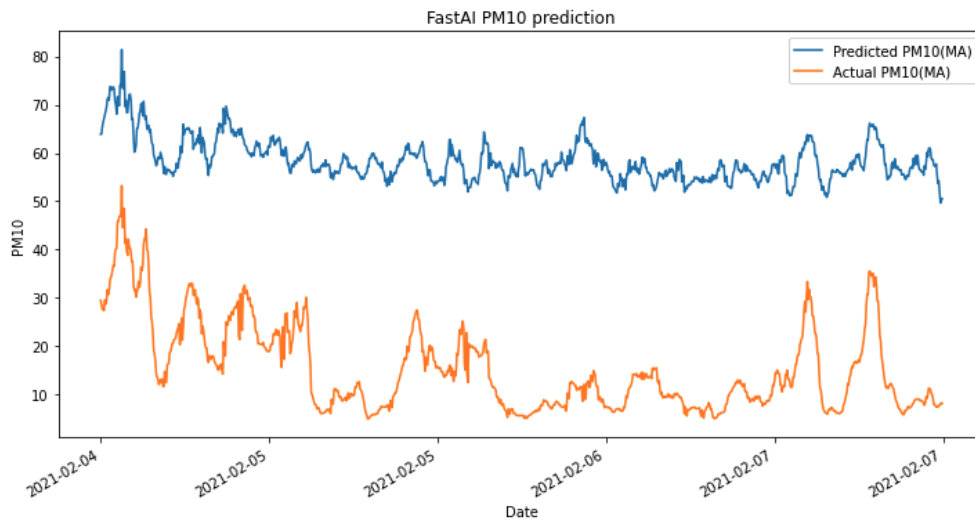
Figure 5.11: Training and validation losses on Fastai after 1500 and 3000 epochs for experiments 1 and 2 respectively.



(a) Predicted vs Actual hourly NO_2 readings.

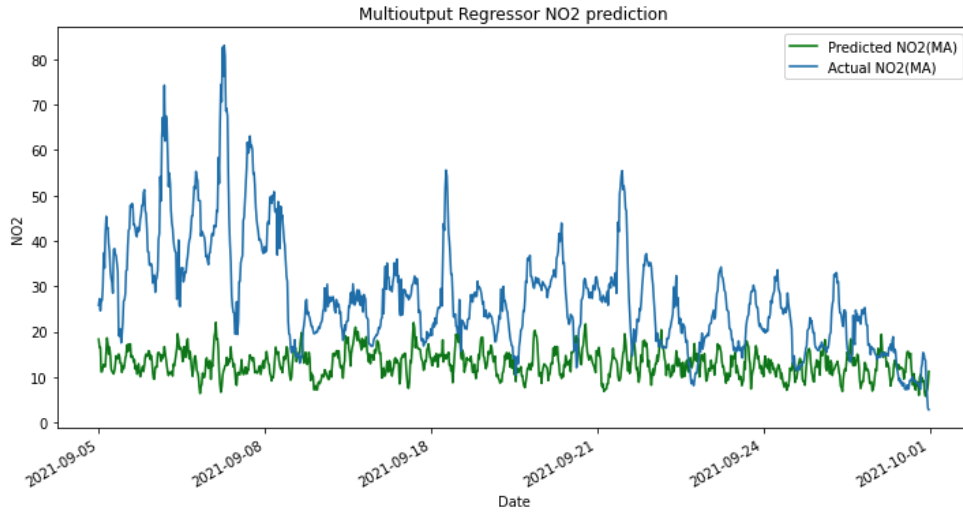


(b) Predicted vs Actual hourly $PM_{2.5}$ readings.

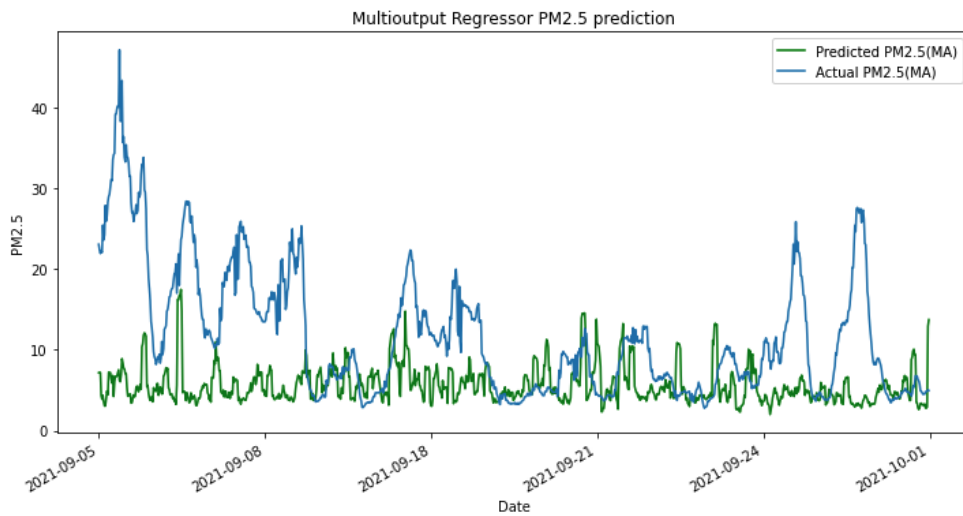


(c) Predicted vs Actual hourly PM_{10} readings.

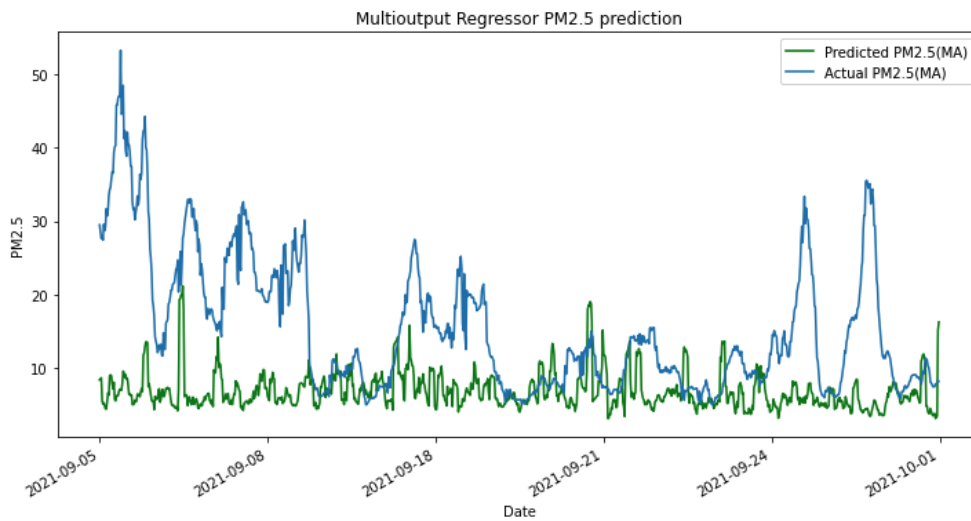
Figure 5.12: Experiment 1 - Fastai's model predictions.



(a) Predicted vs Actual hourly NO_2 readings.

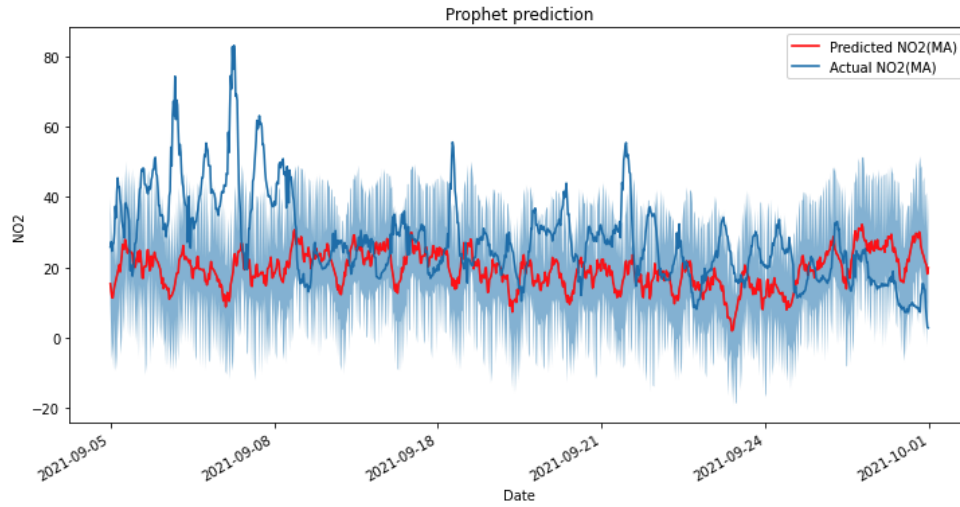


(b) Predicted vs Actual hourly $PM_{2.5}$ readings.

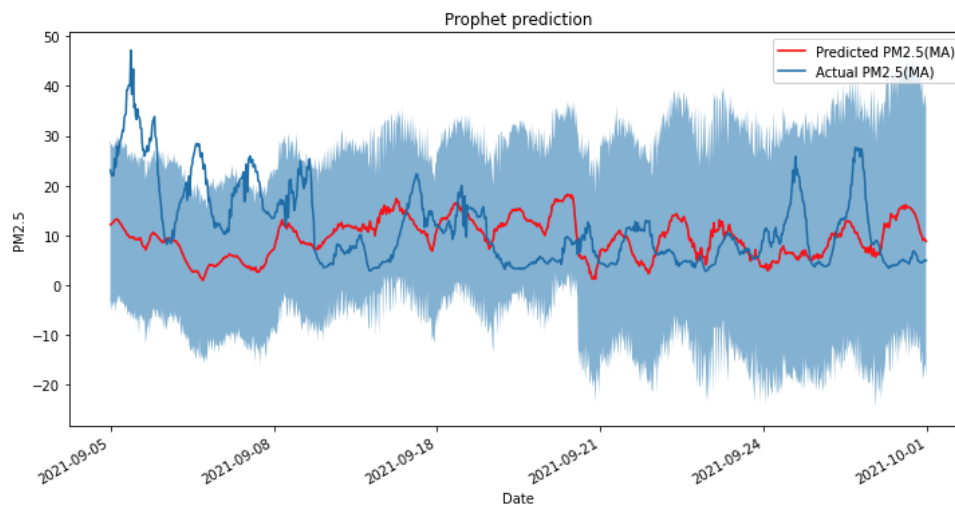


(c) Predicted vs Actual hourly PM_{10} readings.

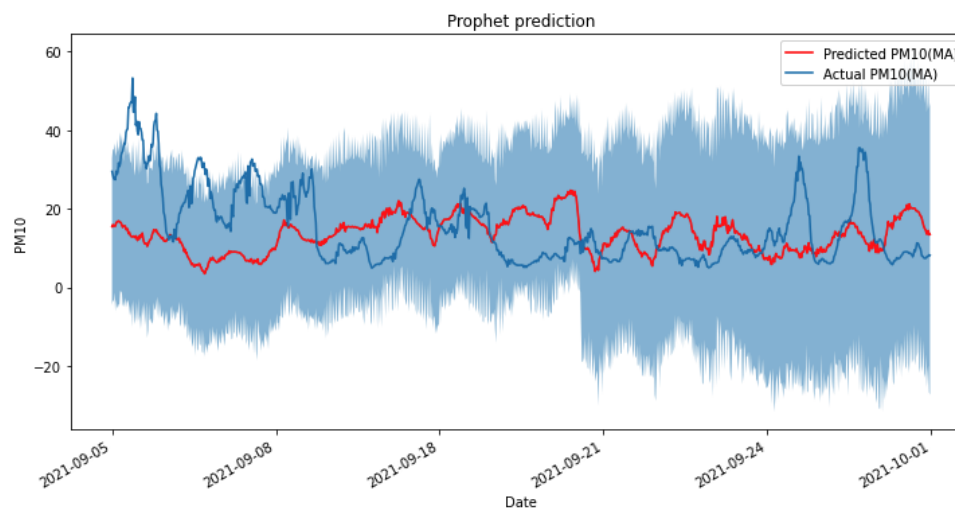
Figure 5.13: Experiment 1 - MultiOutputRegressor's model predictions.



(a) Predicted vs Actual hourly NO_2 concentration levels.



(b) Predicted vs Actual hourly $PM_{2.5}$ concentration levels.



(c) Predicted vs Actual hourly PM_{10} concentration levels.

Figure 5.14: Experiment 1 - Prophet's model predictions.

Table 5.6: Experiment 2 results of MTR models prediction for different timesteps

Pollutant & Timestep		Fastai			Multioutputregressor			Prophet		
		MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
NO_2	1hr	5.333	0.412	8.312	9.132	1.012	15.325	10.122	0.931	14.122
	8hr	7.182	0.676	9.042	13.562	1.622	19.328	13.306	0.826	19.059
	16hr	6.325	0.521	8.763	20.152	2.133	22.541	14.334	0.782	22.326
	24hr	8.058	0.731	10.324	22.034	2.262	20.331	15.591	0.924	24.134
$PM_{2.5}$	1hr	3.062	0.258	5.341	16.506	1.243	23.124	14.332	1.589	11.752
	8hr	4.251	0.328	4.142	21.121	1.476	33.612	18.032	1.629	16.302
	16hr	4.430	0.399	5.189	23.105	1.432	35.145	9.112	1.892	20.126
	24hr	5.639	0.435	6.146	22.498	1.835	36.875	13.763	2.298	21.156
PM_{10}	1hr	3.124	0.267	5.443	13.332	1.228	18.069	20.313	1.254	27.169
	8hr	4.022	0.354	4.783	18.023	1.734	21.100	19.523	2.383	30.124
	16hr	4.129	0.378	5.034	19.326	1.912	26.613	20.376	4.198	32.225
	24hr	5.123	0.462	6.343	21.312	1.820	31.298	21.809	3.213	31.004

5.6.4 Experiment 2 Results

There was an immediately noticeable improvement in the results obtained in experiment 2. The metrics scores dropped considerably for the fastai model while the multioutputregressor and prophet models also saw some improvements. The best scores were recorded by fastai in this round of experiment for all three pollutants simultaneously. Although the model in this experiment was run for 1,500 more epochs than experiment 1, this was not the reason for the improved scores. The first experiment was only run for shorter epochs to avoid overfitting since the validation and training losses were not reducing as the epochs increased. A plot of the validation loss illustrated in figure 5.11b shows that the loss from this experiment was lower from the beginning and reduced in a stable manner as compared to experiment 1. The model's worst performance was on NO_2 24hr predictions with MAE as high as 8.058. However, this result still outperforms the previous NO_2 results for all the models in experiment 1. From Table 5.6, it is hard to determine the model's best prediction performance since the results for $PM_{2.5}$ and PM_{10} were quite similar on 1hr

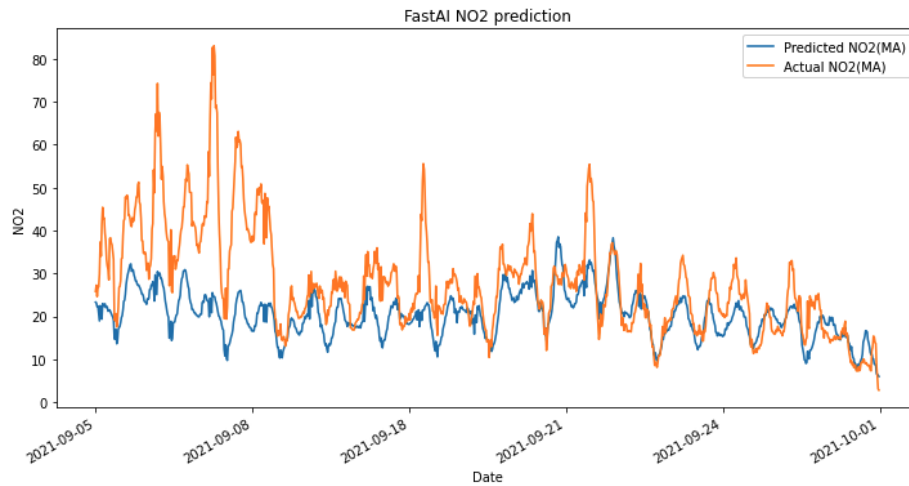
timestep predictions. The best average MAE, MAPE and RMSE scores was recorded as 3.062, 0.258 and 5.341 respectively. This improvement in the performance of the fastai model can be associated with the introduction of lagged variables as well as the hyperparameter tuning in this round of experiment. As illustrated in figures 5.16 and 5.17 and also Table 5.6, the prophet and multioutputregressor models also performed slightly better in this as a result of these changes but the improvement was not as significant as fastai’s (see Figure 5.15).

5.6.5 Statistical significance of results

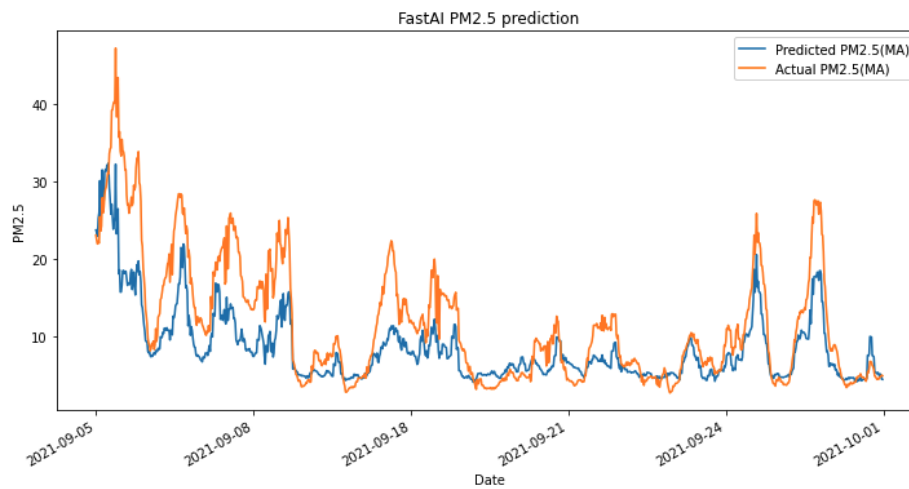
To further strengthen the confidence in the results achieved with fastai, it was necessary that statistical hypothesis tests were carried out to weigh its performance against the two other models. The non-parametric Friedman and the Wilcoxon signed-rank test were selected with a null hypothesis (H_o) that there is no statistical difference between the predictions from the three models. This hypothesis would be rejected if the chi-square was > 3.84 for the Friedman test and p-value was below 0.05 for both tests. Both tests were performed on 20 MAE, MAPE and RMSE error readings from cross-validation in experiment 2. The Friedman test for the 3 models resulted in a chi-square score of 6.45 and p-value of 0.03. Table 5.7 shows the result of the Wilcoxon test for pair-wise comparisons of the models. Just like the Friedman test, all the p-value scores were less than 0.05. The result of both statistical tests indicates that the hypothesis can be rejected and the predictions from fastai are statistically different from the multiouputregressor and prophet models.

Table 5.7: Statistical significance and model evaluation using Wilcoxon signed rank test

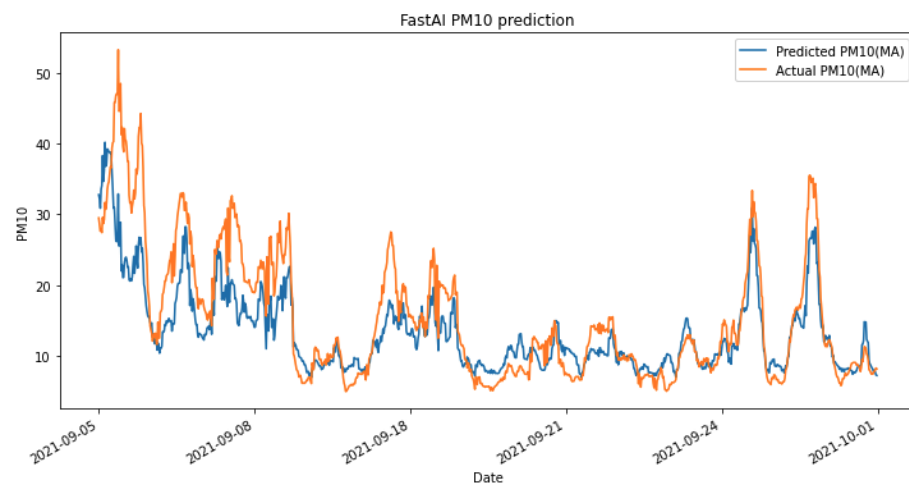
Pair-wise comparison	P-value	Significance
Fastai and Prophet	0.02	Yes
Multioutputregressor and Prophet	0.03	Yes
Fastai and Multioutputregressor	0.02	Yes



(a) Predicted vs Actual hourly NO_2 concentration levels.

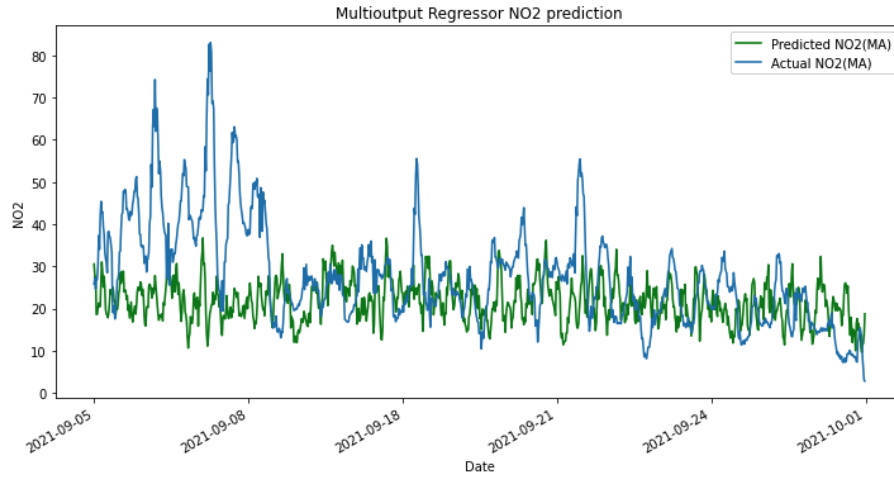


(b) Predicted vs Actual hourly $PM_{2.5}$ concentration levels.

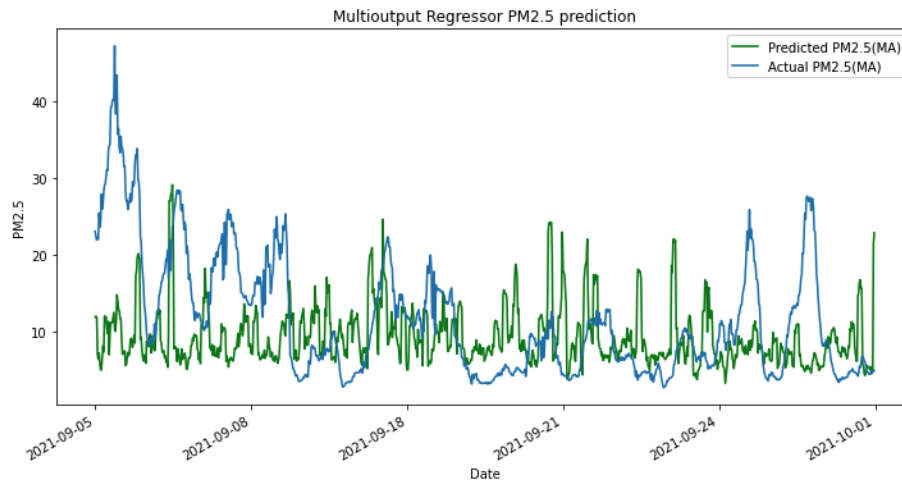


(c) Predicted vs Actual hourly PM_{10} concentration levels.

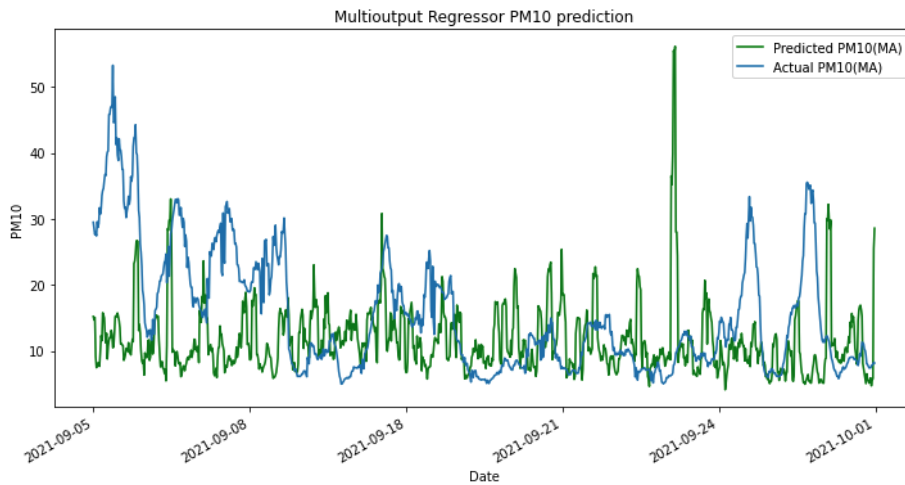
Figure 5.15: Experiment 2 - Fastai MTR predictions for NO_2 , $PM_{2.5}$ and PM_{10} .



(a) Predicted vs Actual hourly NO_2 concentration levels.

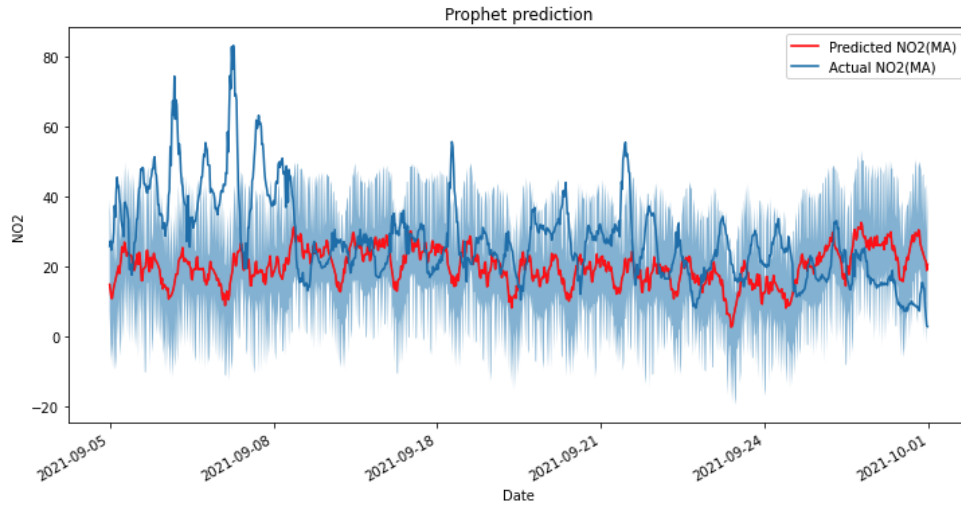


(b) Predicted vs Actual hourly $PM_{2.5}$ concentration levels.

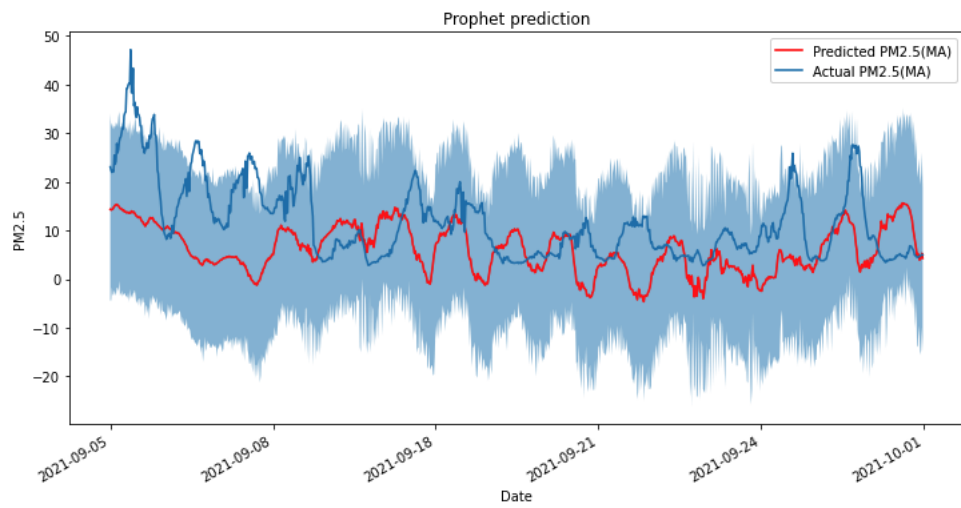


(c) Predicted vs Actual hourly PM_{10} concentration levels.

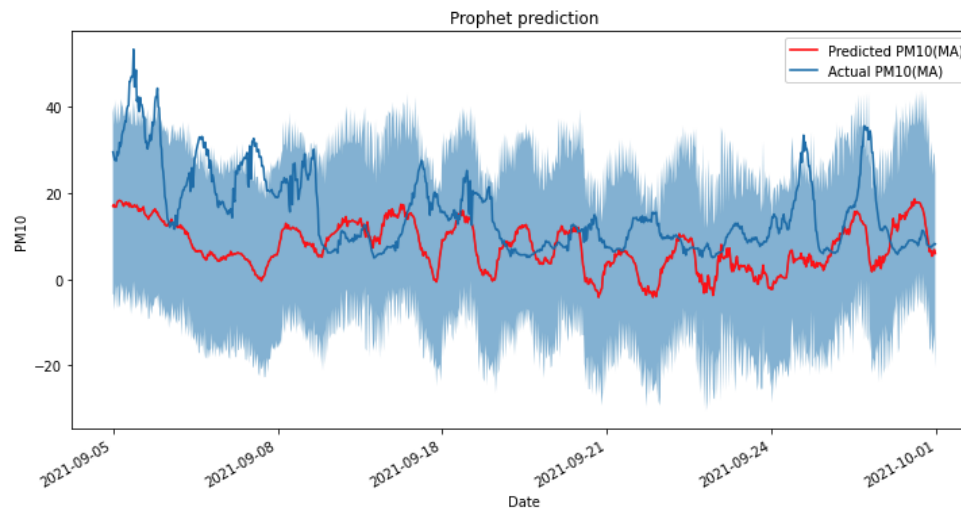
Figure 5.16: Experiment 2 - MultiOutputRegressor's MTR predictions for NO_2 , $PM_{2.5}$ and PM_{10} .



(a) Predicted vs Actual hourly NO_2 concentration levels.



(b) Predicted vs Actual hourly $PM_{2.5}$ concentration levels.



(c) Predicted vs Actual hourly PM_{10} concentration levels.

Figure 5.17: Experiment 2 - Prophet's model predictions.

5.6.6 Results comparison with related work

Although numerous studies have focused on traffic-related air pollution prediction, very few have explored multi-target prediction of pollutants or utilised the specific combination of datasets employed in this study. To validate the performance of the algorithm proposed in this study, a number of closely related studies conducted in locations similar to those used for experiments in this research were selected. These studies were chosen not only for their geographical relevance but also for their adoption of artificial neural network variants similar to the deep learning algorithms applied in this research. For instance, the studies by Suleiman et al. (2019) and Cabaneros et al. (2017) were conducted in London and focused on forecasting roadside concentrations of NO_2 , $PM_{2.5}$, and PM_{10} using artificial neural networks. We replicated the data cleaning processes reported in these studies and used their datasets to train and evaluate models using our proposed algorithm. This approach ensured a fair comparison by maintaining consistency in data preprocessing and modelling techniques.

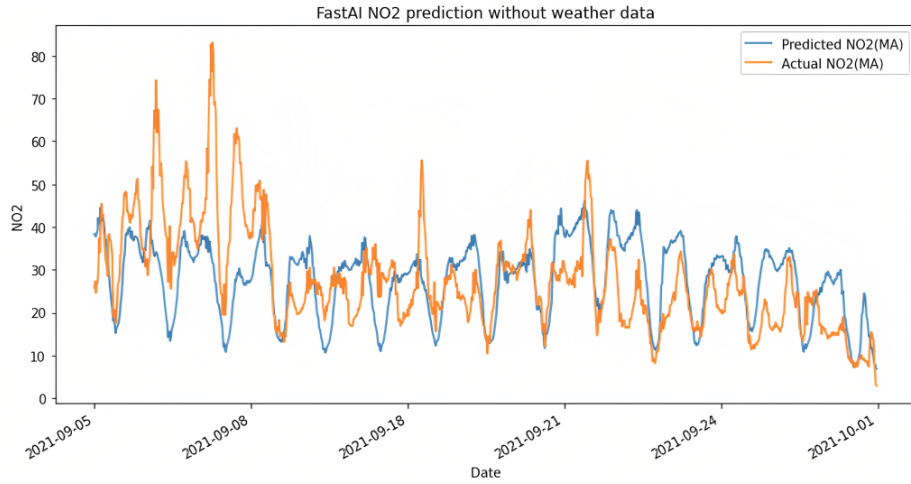
The results of the proposed method in this research, compared with those reported in the selected studies, are presented in Table 5.8. The resulting models from each retraining process consistently outperformed the benchmarks, achieving a minimum improvement of 5% in RMSE scores. This enhanced performance can be attributed to several factors. Firstly, the use of additional training data enriched our model’s learning process, enabling it to capture more complex patterns and variations in the data. Secondly, the adoption of categorical embeddings allowed the developed models to handle categorical variables more effectively, leading to more accurate predictions. This validation underscores the robustness and effectiveness of the proposed approach, making it a valuable tool for urban air quality management. The consistent performance gains across different locations further validate the generalisability and reliability of the approach.

Table 5.8: Comparison of prediction results with existing studies based on RMSE score

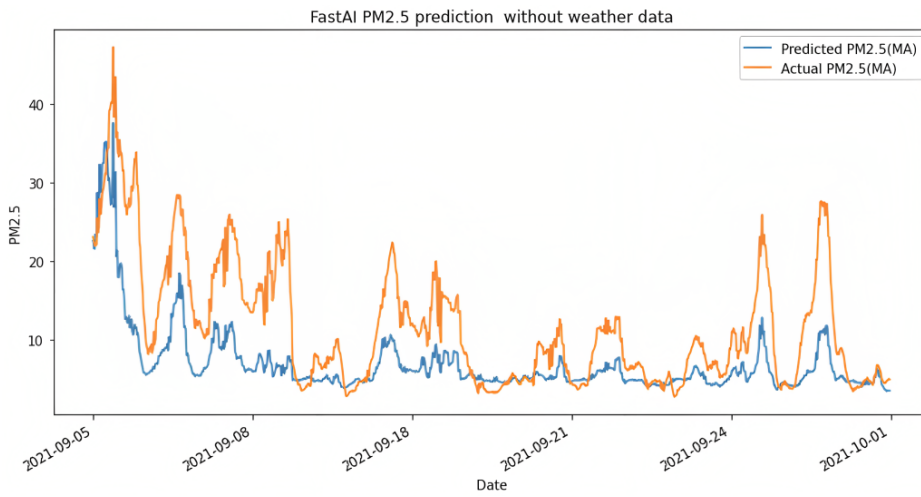
Reference	Data Source	Method	Pollutant	RMSE (lowest)	MTR RMSE (This study)
Cabaneros et al. (2017)	Marlyeborne Road Monitoring sites	Hybrid Artificial Neural Networks	NO_2	22.05	17.03
Suleiman et al. (2019)	Monitoring sites	Artificial Neural Network, SVM, BRT	$PM_{2.5}$	4.67	2.13
			PM_{10}	10.05	8.23
Li et al. (2020)	Hong Kong Roadside station	SVM, GAM, XGBoost, RF, BRT	$PM_{2.5}$	7.90	6.10
			NO_x	30	28
Jida et al. (2021)	Aeroqual AQ sensor	Artificial Neural Network	$PM_{2.5}$	8.45	7.11
			PM_{10}	12.42	11.09
Wu et al. (2022)	Shanghai Roadside stations	Neural Networks - LSTM	NO_2	9.61	8.54
Mengara Mengara et al. (2022)	South Korea Roadside stations	LSTM, Auto Encoder, Convolutional Neural Networks	$PM_{2.5}$	7.40	6.12
			PM_{10}	9.81	8.33

5.6.7 Experiment 3 - verify model's performance on missing data

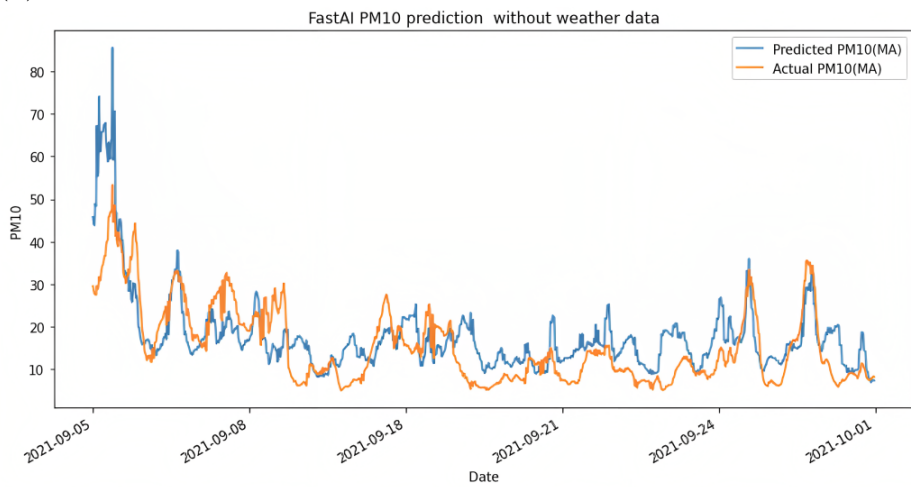
An additional test was carried out to evaluate the performance of the fastai model from experiment 2 in a real-life scenario where some of the integrated data might be missing. It is suggested that as much data as possible is sourced to get optimum performance, but this may not always be the case. To replicate this scenario, the values for the intended missing data were replaced with zeros in the test data before model inferencing. It was important to not drop the columns entirely since the model was originally trained on 44 features and dropping them would result in errors. Similarly, replacing with Nan instead of zeros results in errors too. The model's predictive performance when traffic, weather, emissions factor, background concentration or elevation data are missing can be seen on figures 5.18-5.22. The illustrations indicate varying predictive accuracy depending on the missing data. The model's performance is worse when weather data is missing and poor when elevation or background concentration data are missing. NO_2 prediction is the most affected in these missing data scenarios. This performance variation with certain missing data begs the question - *What are the most important features that must be captured for a reasonable prediction accuracy?*



(a) NO_2 hourly predictions missing traffic data.

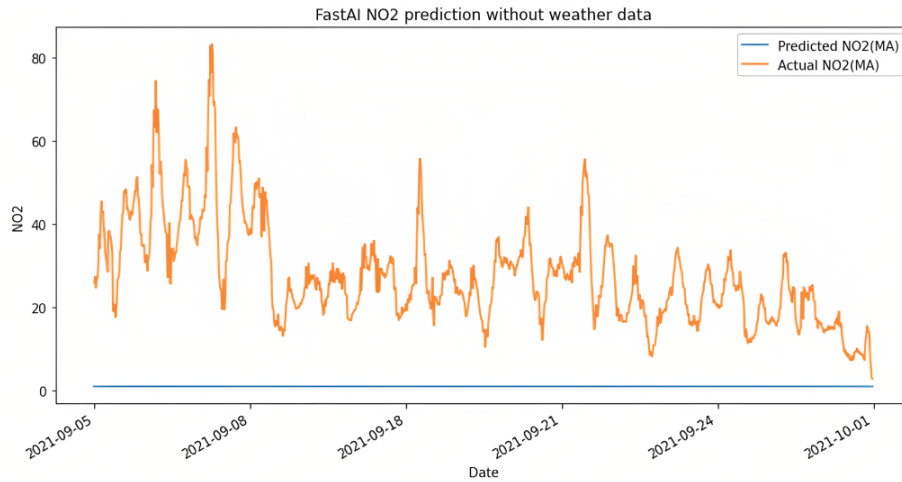


(b) $PM_{2.5}$ hourly predictions missing traffic data.

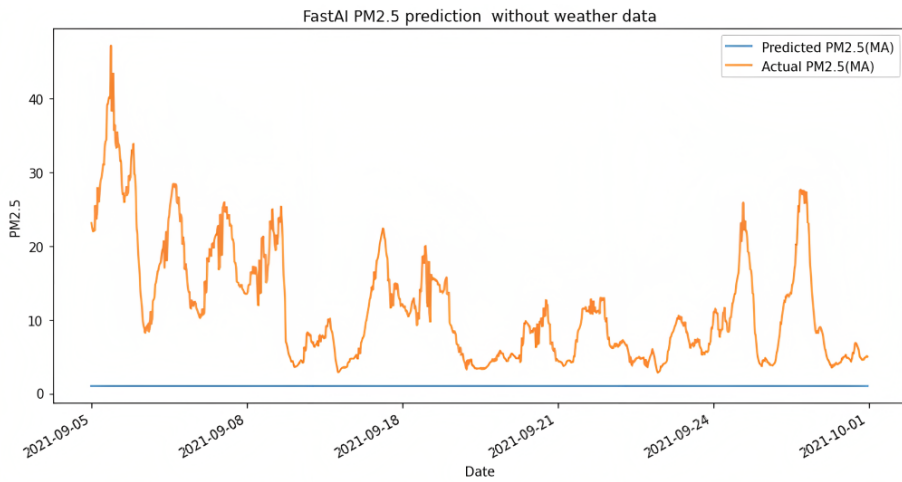


(c) PM_{10} hourly predictions missing traffic data.

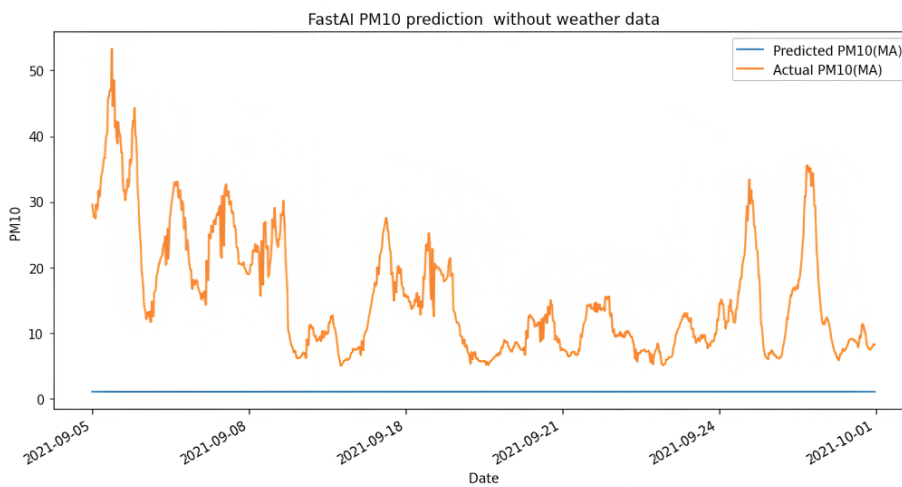
Figure 5.18: Fastai model's performance when missing traffic data.



(a) NO_2 hourly predictions missing weather data.

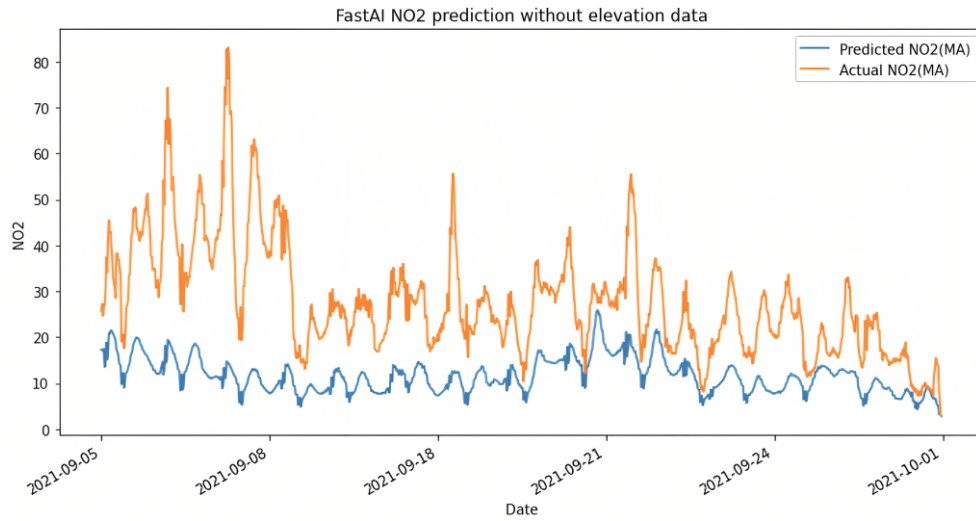


(b) $PM_{2.5}$ hourly predictions missing weather data.

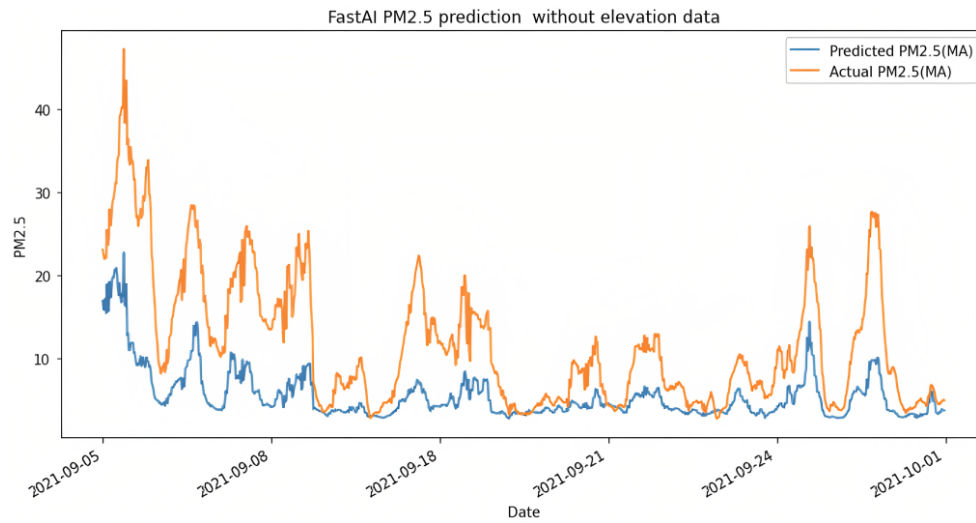


(c) PM_{10} hourly predictions missing weather data.

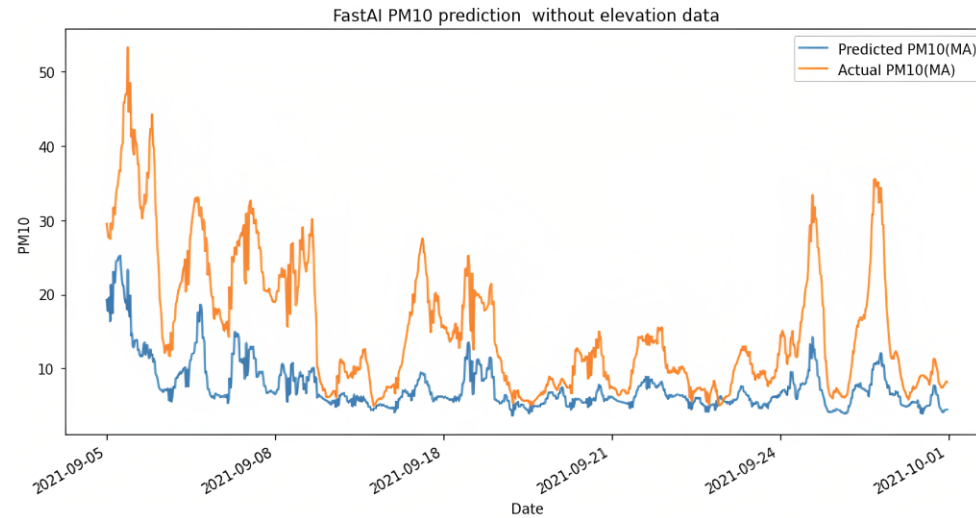
Figure 5.19: Fastai model's performance when missing weather data



(a) NO_2 hourly predictions missing elevation data.

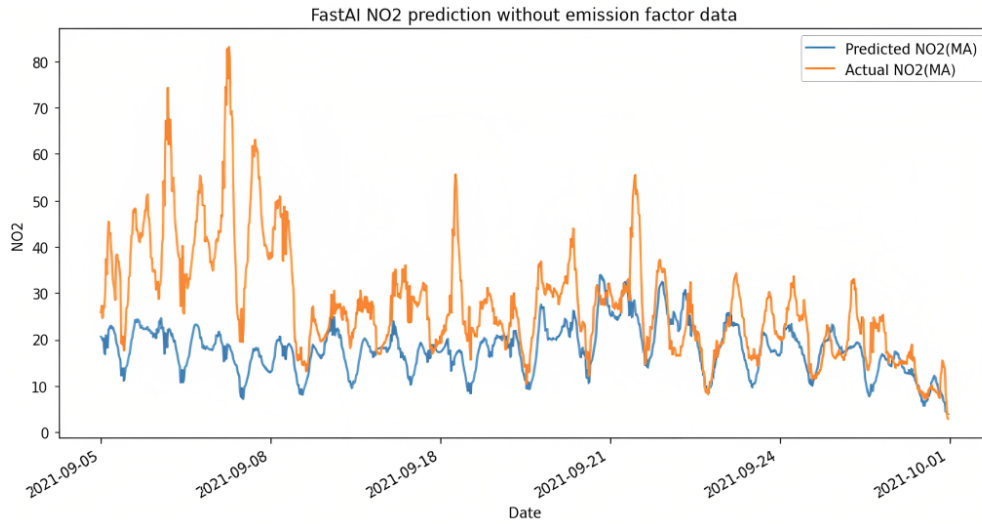


(b) $PM_{2.5}$ hourly predictions missing elevation data.

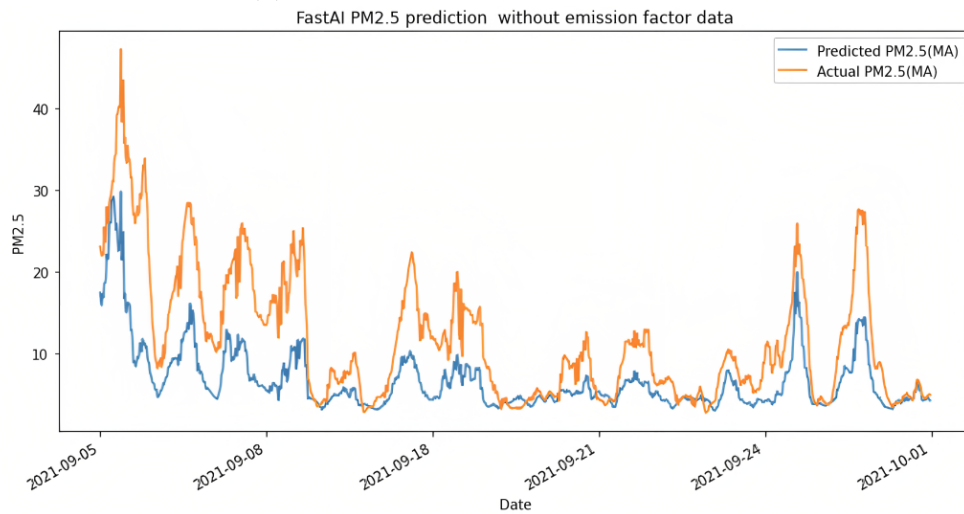


(c) PM_{10} hourly predictions missing elevation data.

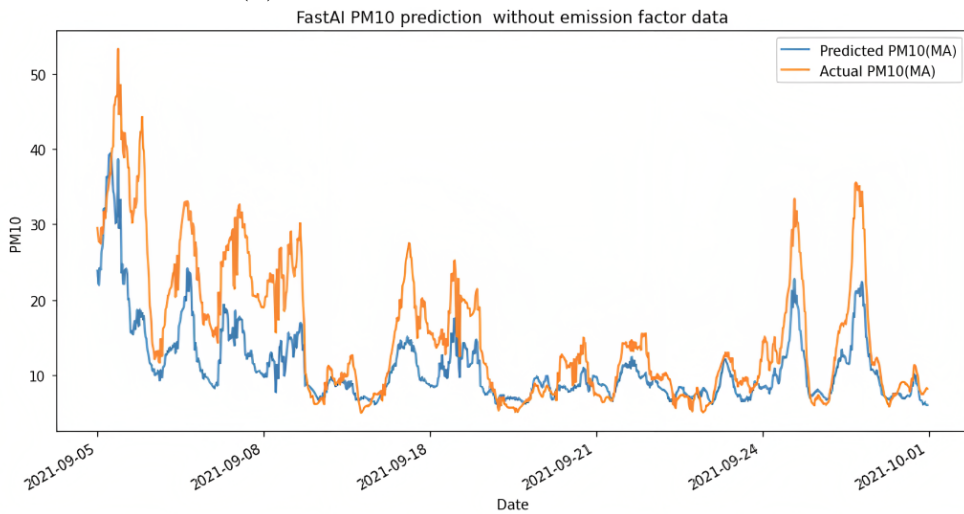
Figure 5.20: Fastai model's performance when missing elevation data



(a) NO_2 hourly predictions missing emissions factor data.

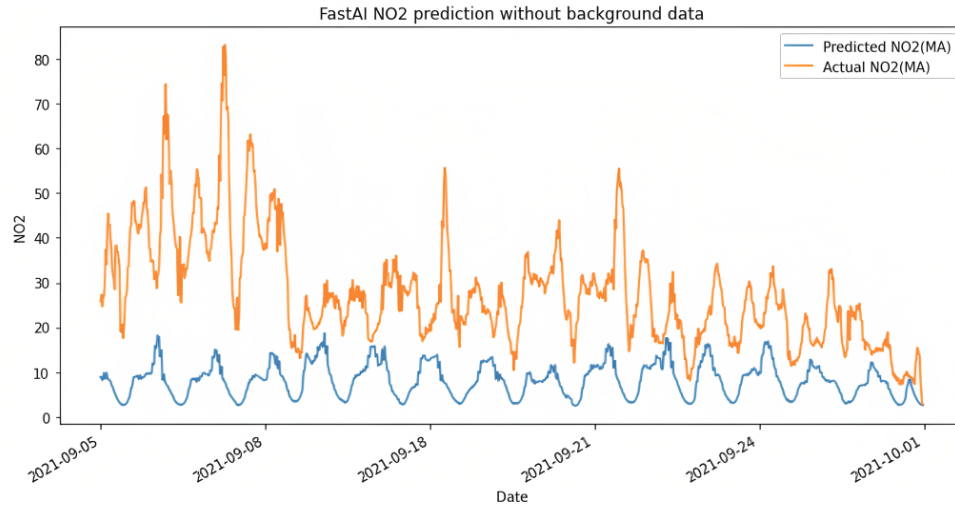


(b) $PM_{2.5}$ hourly predictions missing emissions factor data.

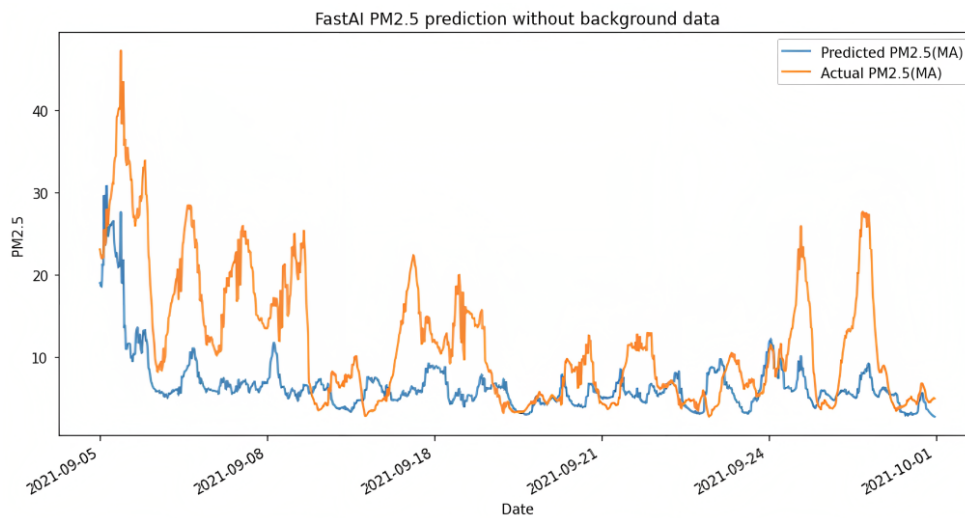


(c) PM_{10} hourly predictions missing emissions factor data.

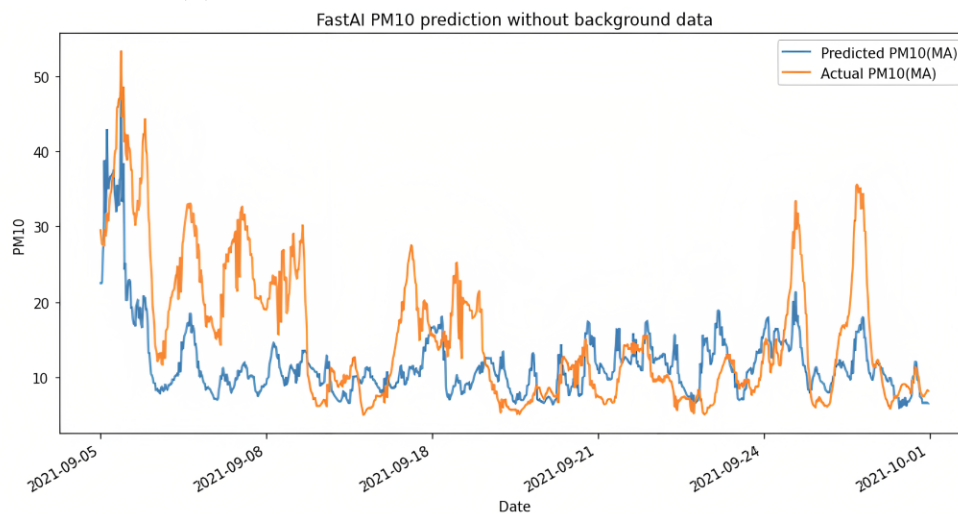
Figure 5.21: Fastai model's performance when missing emissions factor data



(a) NO_2 hourly predictions missing background concentration data.



(b) $PM_{2.5}$ hourly predictions missing background concentration data.



(c) PM_{10} hourly predictions missing background concentration data.

Figure 5.22: Fastai model's performance when missing background concentration data

5.7 Feature importance on best model results

Following the improvement of fastai model’s performance in experiment 2, further investigation was carried out to understand which of the input parameters were the most influential in the model’s predictions. This section highlights the outcome of this analysis.

5.7.1 Experiment 4 - fewer features, same accuracy

Machine learning models developed with advanced algorithms such as deep learning are considered black box models (Akinosho et al. 2020). This is as a result of the complexities involved in understanding what happens behind the scenes for most of these models. It is particularly important in the air quality domain to highlight the main contributors to pollution through this kind of understanding. Thankfully, various tools are now available to make models explainable and fastai’s *Interpretation* classes further facilitate this task. A feature importance plot as shown in figure 5.23 was plotted using one of these tools and this gave many insights into which of the 44 input parameters were the least and most contributing. From the plot it is observable that ‘LGV Count’, ‘Other Avg speed’, ‘Bus Count’, ‘Wind Direction’, ‘Car Count’, ‘HGV Count’, ‘ NO_2 emission factor’ and ‘DATETimeHour’ were the most influential features. These are mainly traffic parameters except the ‘Wind Direction’ and ‘DATETimeHour’ features. All the additional date variables that were added to the data set had none to little impact with some even recording negative importance. Similarly, ‘highway elevation’, ‘background NO_2 ’ and other weather parameters were not important for the model’s predictions. The fastai model was retrained while dropping these low and negative influencing parameters to see if its performance would be any different and if the feature importance will be reshuffled.

Figure 5.24 shows the feature importance after retraining on just the top 12 features from experiment 2. The model’s accuracy remained similar to what was achieved in experiment 2 but the feature importance was reorganised. It can be noticed that most of the traffic

parameters maintained the top spot with only *car count* dropping behind. The date parameter were also influential with the hour of the day having the highest influence. The wind direction and NO_2 emission factor features dropped to the bottom of the list in this round. However, it is worth reiterating that these least influential features are only not so important for this minimised data set but had significant impact in the overall data set

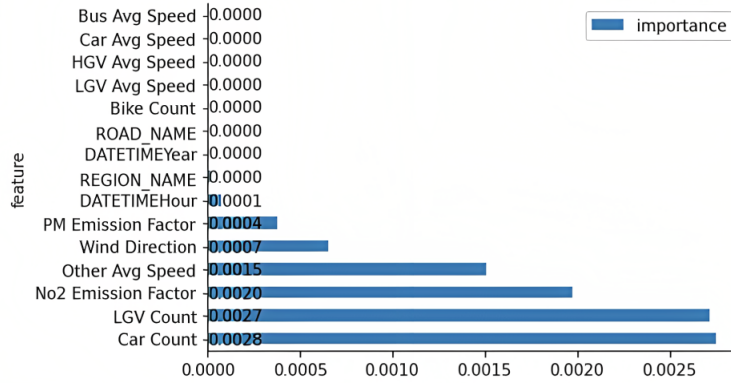


Figure 5.23: Feature importance from experiment 2. Traffic features including ‘LGV count’ and ‘car count’, ‘average speed’ were in the top list with the hour of the day, ‘wind direction’, ‘PM emission factor’ and ‘ NO_2 emission factor’ also part of this list. Some of the least influential parameters were ‘bike count’, minute of the day and similar date parameters.

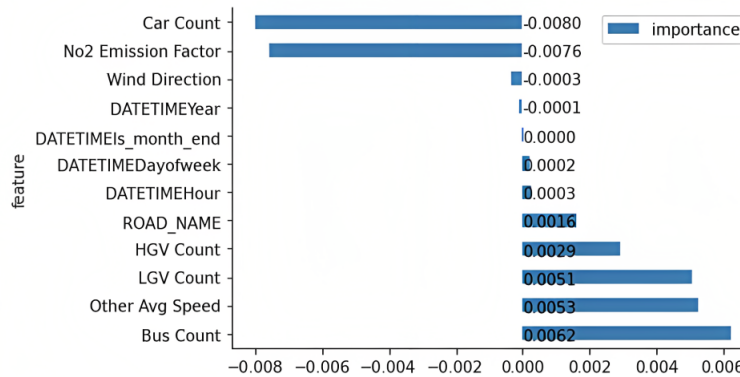


Figure 5.24: Feature importance after retraining on the top twelve features from experiment 2. All the traffic features except ‘car count’ maintained the top spot while ‘wind direction’ and ‘ NO_2 emission factor’ dropped further down the importance list.

5.7.2 Experiment 5 - features ablation test

The result of running an ablation test on the fastai model to further corroborate the importance of the training features is illustrated in figure 5.25. The test was carried out by dropping each feature one at a time and then retraining the model on the remaining features to predict all three pollutants. The RMSE score on the test data for each pollutant was recorded once the model retraining process was complete and the model was cross validated. This score was then compared to the RMSE score when all the features were used. The x-axis on the figure represents each feature that was dropped while the y-axis represents the recorded RMSE score. It can be observed that the impact of dropping most of the additional date parameters was almost non-significant except for the hour parameter. Similarly, dropping the weather parameters, background pollution data and traffic parameters all resulted in a significant increase in the RMSE score to a level that is almost similar to experiment 1. Removing the other features had less impact on the model's performance. The result of this ablation test corresponds with the feature importance from the previous section where traffic and weather parameters were highlighted as important.

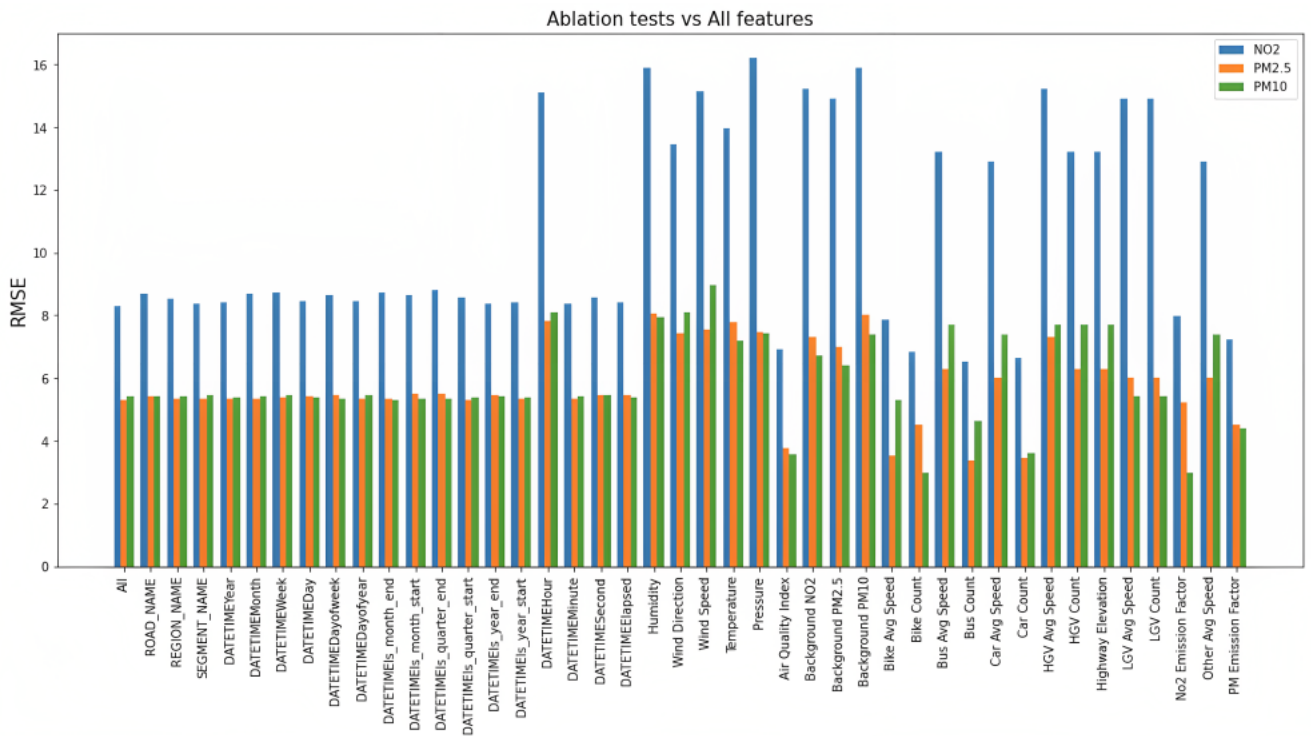


Figure 5.25: Feature ablation test to reveal features with the most impact on fastai model's predictions. The x-axis contains the feature list with each tick representing the feature that was removed when the model was retrained and RMSE score recalculated. The RMSE scores are represented on the y-axis. This chart indicates the importance of traffic and weather data as the RMSE scores increased when these features were removed from the data set.

5.8 Chapter Summary

This chapter builds on the forecasting model developed in the previous chapter. Unconventional training data, including terrain information, background pollutant concentrations, and emissions factors, was integrated with conventional data such as traffic flow, weather conditions, and historical pollution levels. Multi-target prediction models were trained for NO_2 , $PM_{2.5}$, and PM_{10} , with the results indicating the effectiveness of the models, albeit requiring extensive hyperparameter tuning. The best performance was achieved with fastai for simultaneous hourly predictions of all three pollutants, excelling with $PM_{2.5}$ and PM_{10} but encountering challenges in accurately predicting NO_2 spikes, particularly those caused by localised pollution. Key contributors to model performance were identified as traffic, weather, time of day, and emission factors. This chapter underscores the potential improvement in prediction accuracy by incorporating additional highway features, while acknowledging the persistent challenge of unusual spikes, prompting further investigation into such scenarios.

Chapter 6

Algorithmic Audit and Model

Integration In Journey Planner

6.1 Chapter Overview

In this chapter, a comprehensive exploration of the constraints and shortcomings of the developed MTR model is conducted, particularly focusing on its challenges in handling peak scenarios for pollutant levels. The primary objective is to discern the underlying factors contributing to these limitations while seeking avenues for enhancing the model's performance. Additionally, this chapter sheds light on the practical aspects of integrating and deploying the model within a mobile journey planning application, highlighting the steps involved in this process.

6.2 Baseline Model Evaluation

6.2.1 Multi-target regression deep learning model

This study builds upon the MTR-1 model developed in the previous chapter. MTR-1 was designed to simultaneously predict concentration levels of $PM_{2.5}$, PM_{10} , and NO_2 at time-points - 8hrs, 16hrs, and 24hrs. To develop this model, an innovative approach was employed,

leveraging the capabilities of fastai, a robust deep learning framework. The model’s architectural design features a 3-layered Recurrent Neural Network (RNN) with 200 neurons in the initial layer, 162 neurons in the second layer, and 134 neurons in the third layer.

The process of gathering training data occurred between November 2020 and November 2021, employing REVIS sensors strategically placed along dedicated sections of the four highways. While historical pollution and meteorological data were acquired through these sensors, additional data such as pollution levels, geographical factors, emission indicators, and traffic flow statistics were integrated from external sources using an integration approach proposed in an earlier study (Akinosho et al. 2022). The incorporation of novel highway parameters, previously unexplored in traditional Traffic-Related Air Pollution (TRAP) prediction, yielded promising outcomes, enabling the model to unveil the intricacies of pollution dynamics along the studied highways. However, the model encountered challenges during specific peak events, prompting this study’s exploration into the underlying causes and potential avenues for enhancing efficiency, particularly concerning infrequent occurrences.

Table 6.1: Boundary values for outlier detection in the target pollutants

Pollutant	Boundary Value (Lower Threshold)	Boundary Value (Upper Threshold)
NO_2	-27.26	71.17
$PM_{2.5}$	-35.05	54.47
PM_{10}	-41.84	65.44

6.2.2 Identifying outliers from original data set

The MTR-1 model was audited using the framework proposed by the study of Raji et al. (2020) to spot and describe inaccuracies. According to their study, these inaccuracies refer to any artificial intelligence system results that don’t match accuracy expectations and with the potential of derailing outputs if not detected. To achieve this, an exploratory error

analysis was carried out to understand the reasons behind the model’s failures and to identify potential outliers. For outlier detection, the z-score method was adopted, chosen based on the observed data distribution, which closely adheres to a normal distribution. The z-score technique allowed the effective identification and isolation of data points that exhibit substantial deviations from the data set’s central tendency. The z-score method gauges the extent to which each data point deviates from the data set’s mean in terms of standard deviations. Those data points exceeding a specified threshold, typically set at around ± 2 to ± 3 standard deviations, are categorised as outliers. This approach proves to be particularly suitable for the data set, as it enables us to pinpoint those values that may signal uncommon events or reflect measurement errors. In the process, a total of 447 outliers were identified within the data set. Figure 6.1 shows these outliers consisting of 168 associated with NO_2 , 150 $PM_{2.5}$, and 129 PM_{10} outliers. These findings provide valuable insights into the data set’s unique characteristics and help in understanding which pollutants or variables exhibit the most significant deviations from the norm. To offer a comprehensive overview of these findings, Table 6.1 summarises the boundary values for the three pollutants which were computed using equation 6.1, and indicates the thresholds that determined outlier status for each pollutant in the analysis.

$$z = \frac{x - \mu}{\sigma}, \quad z \in \{-3, +3\} \tag{6.1}$$

6.2.3 Model’s performance on outliers

A comprehensive assessment of the MTR-1 model’s performance was conducted in the presence of outliers. To achieve this, cross-validation techniques were employed while the model’s results were compared on both outlier-prone and outlier-free subsets of the data set. This approach provided valuable insights into the model’s generalisation capabilities, shedding light on its robustness in handling challenging data points. For performance evaluation, traditional regression metrics such as Mean Squared Error (MSE) and Root Mean Squared Error

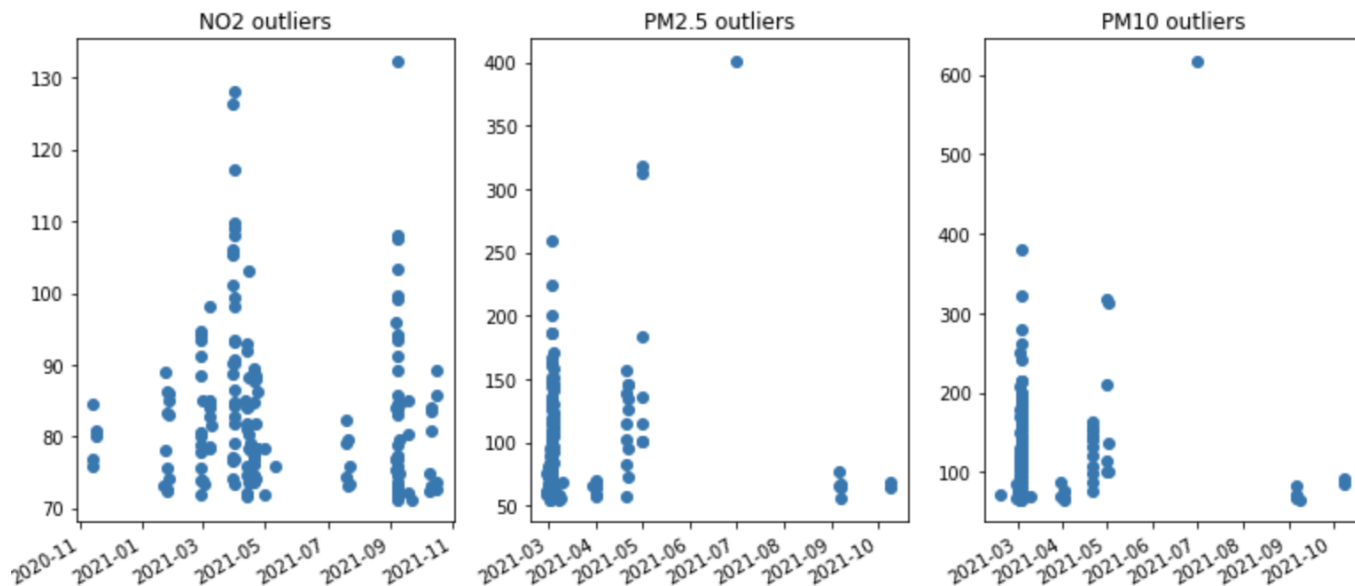
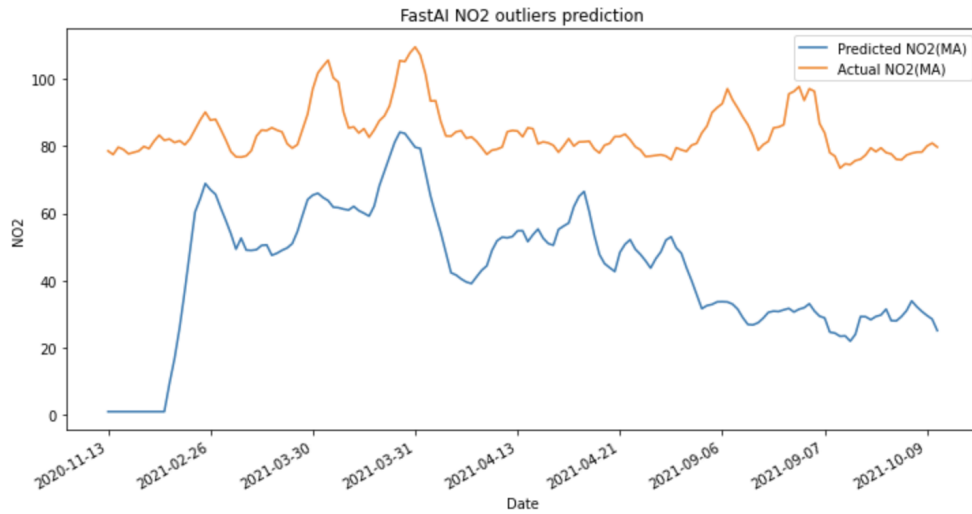


Figure 6.1: Outliers detected within target pollutants: A total of 447 outliers detected, comprising 168 NO_2 outliers, 150 $PM_{2.5}$ outliers, and 129 PM_{10} outliers.

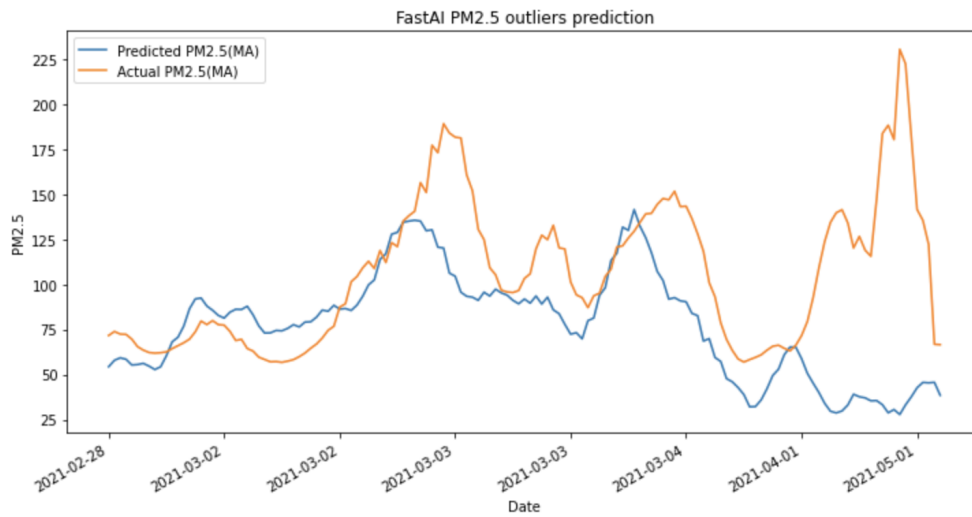
(RMSE) were adopted to gauge overall predictive accuracy. Special attention was given to the potential impact of outliers on these metrics, as extreme values can inflate error measurements. To mitigate this, other alternative metrics like Mean Absolute Error (MAE) were considered and the use of robust regression techniques known for their reduced sensitivity to extreme values was also explored. In addition, visualisation plots were used to offer a more in-depth examination of the MTR-1 model’s performance. These plots illustrated the model’s predictive accuracy by comparing predicted values against actual outlier values, providing a clear visual representation of its performance, as depicted in Figure 6.2. This multi-faceted approach allowed us to thoroughly assess the model’s ability to handle outliers and provided a comprehensive evaluation of its predictive capabilities.

6.2.4 Synthetic data generation from outliers

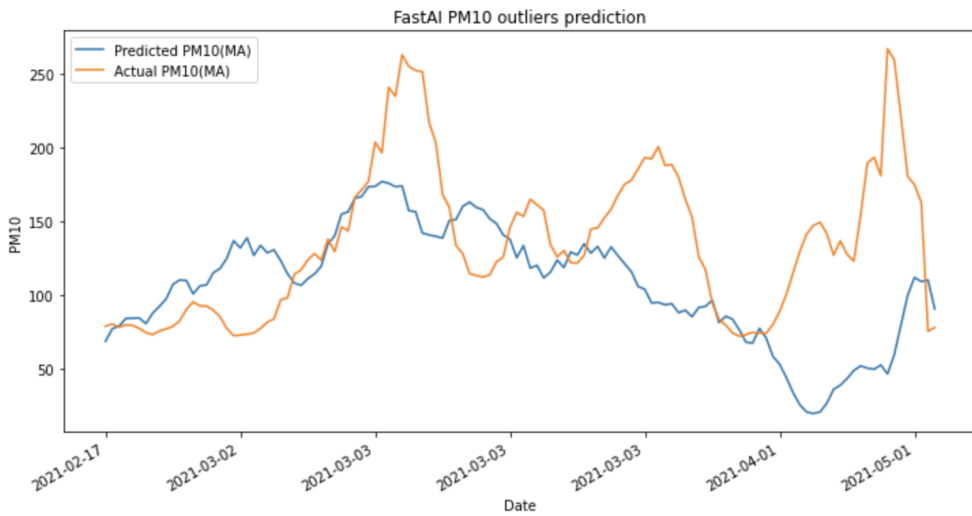
Following the identification of outliers, the next step was to generate synthetic data sets using the GaussianCopula method from the *sdv.tabular* library. This step was important for fine-tuning the MTR model to get a better performance on unique scenarios where it struggled.



(a) Actual vs predicted NO_2 outliers using the MTR-1 model.



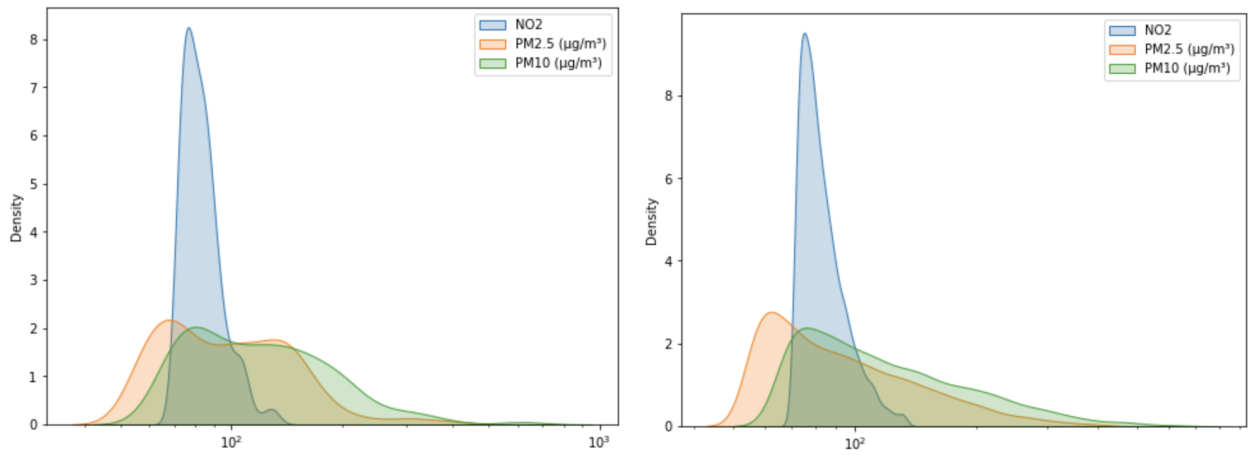
(b) Actual vs predicted $PM_{2.5}$ outliers using the MTR-1 model.



(c) Actual vs predicted PM_{10} outliers using the MTR-1 model.

Figure 6.2: Plots of MTR-1 model's performance on predicting NO_2 , $PM_{2.5}$ and PM_{10} outliers

The GaussianCopula method is a useful approach for creating realistic and representative data for various data science tasks. It uses a copula-based generative model that captures the dependencies between variables in the data, making it suitable for preserving complex multivariate relationships. To generate a synthetic data set, the first step is to define the metadata and constraints of the target data set, such as column data types, primary keys, and unique constraints. Once the metadata is set up, the GaussianCopula model can then be used to sample data. The GaussianCopula method works by transforming the data into a multivariate Gaussian distribution with specified correlations, and then sampling from this distribution. This ensures that the generated data preserves the statistical characteristics and dependencies present in the original data set. Figure 6.3 illustrates the gaussian distribution of the outliers that was used to generate 12000 (4000 per pollutant) additional training data and the distribution of the generated data which is quite similar.



(a) Distribution of original outlier data set

(b) Distribution of generated outlier data set.

Figure 6.3: Distribution of outlier data sets (a) and generated training data set using the GaussianCopula method (b). Both plots depict normal distributions for both data sets, a key characteristic and prerequisite for employing the GaussianCopula data generation technique. Notably, a similarity is also evident in the distributions of both data sets

6.3 Experimentation

The MTR-1 model underwent fine-tuning on the newly generated data set, resulting in the creation of the MTR-1E(Enhanced) model. While retaining the original fastai architecture for model training, a hyperparameter search was carried out to optimise its performance. This section provides a detailed overview of the model training and experimentation steps that were undertaken in this process.

6.3.1 MTR-1E Model Training and Validation

The initial phase of experiments involved standard data cleaning procedures aimed at ensuring the reliability and accuracy of the data set for subsequent analyses and machine learning tasks. Duplicate records were identified and subsequently removed to mitigate any potential data distortions. In addition, data points with missing target data were addressed through careful data imputation techniques. Inconsistent data formats, such as date formats or categorical variables, were looked for and standardised for uniformity. The data set was segregated into categorical and continuous variables as required by the subsequent modelling process. Rather than conducting a complete retraining, the initial model underwent fine-tuning when exposed to the new data set. This approach was chosen to leverage MTR-1's existing knowledge about the target pollutants. To facilitate this, slight modifications were made to the training architecture. Specifically, the last layer of the pre-trained model was frozen to prevent it from being updated during fine-tuning, while keeping the rest of the architecture largely unchanged. The architecture comprised three fully connected layers with 200, 162, and 134 nodes, a weight decay of $1e^{-2}$, a learning rate of $1e^{-3}$, and a dropout probability of 0.2.

The data set, consisting of 12,000 data points, was split into training (70%, 8,400), validation (20%, 2,400), and test (10%, 1,200) sets. The resulting model, now referred to as MTR-1E was trained for 1300 epochs. Figure 6.4 illustrates the validation and training loss

trends following the training process. For model evaluation, two separate test sets were used: Test set A was derived from the synthetic dataset created in section 5.4.2 and was used to evaluate the model’s performance under controlled conditions. This test set is designed to mimic realistic but systematically varied data, ensuring that the model’s ability to generalise to typical scenarios is rigorously assessed. Test set B, on the other hand, comprised outliers from the original dataset. The purpose of using Test set B is to evaluate the model’s robustness and resilience when faced with anomalous or unexpected conditions. By testing the model on these outliers, we can assess its ability to maintain accuracy and reliability in real-world scenarios where data may not always conform to typical patterns. This dual testing approach ensures a comprehensive evaluation of the model’s performance, covering both standard and extreme conditions. one from the synthetic dataset (Test set A) and another containing outliers from the original dataset (Test set B).

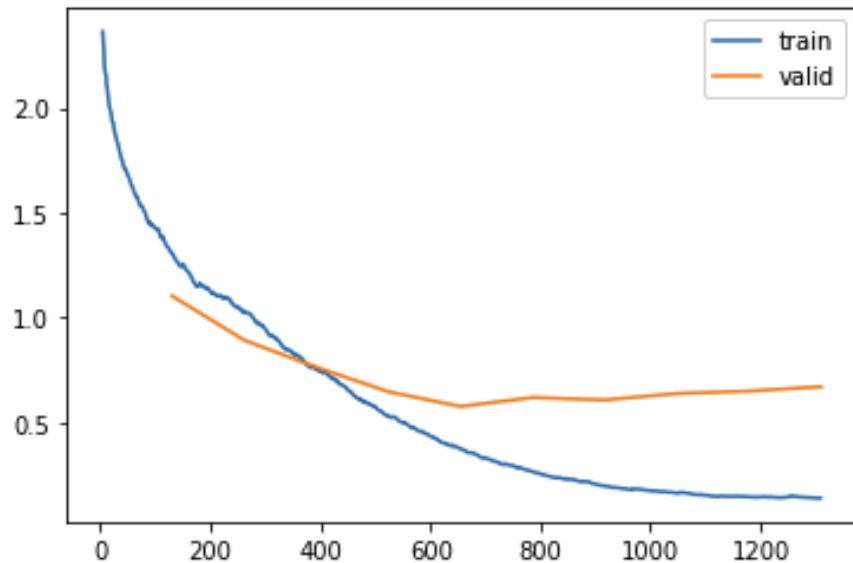


Figure 6.4: Training and validation loss after 1300 epochs training a model with the synthetic data set.

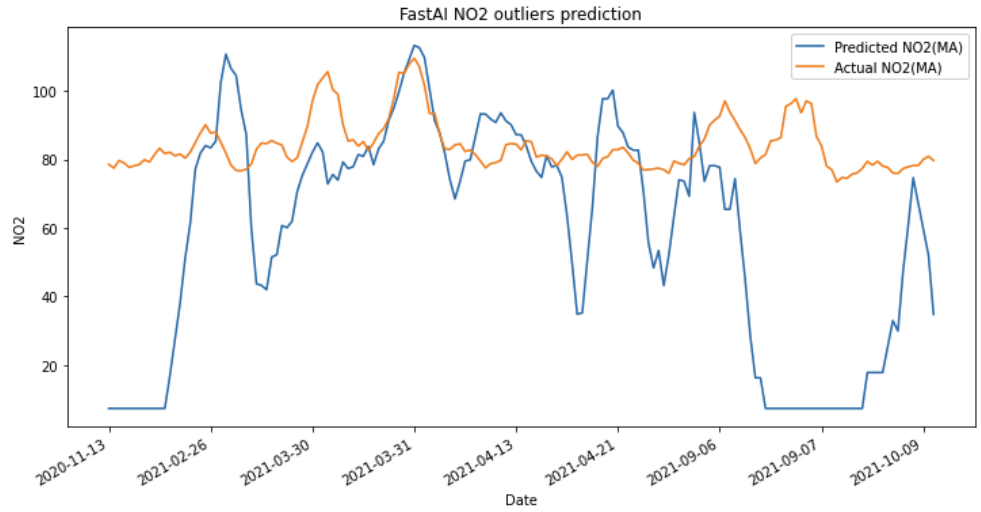
6.3.2 Model Performance Evaluation and Results

Model evaluation and validation are essential steps in assessing the performance of machine learning models. Two commonly used metrics for this purpose are Root Mean Squared

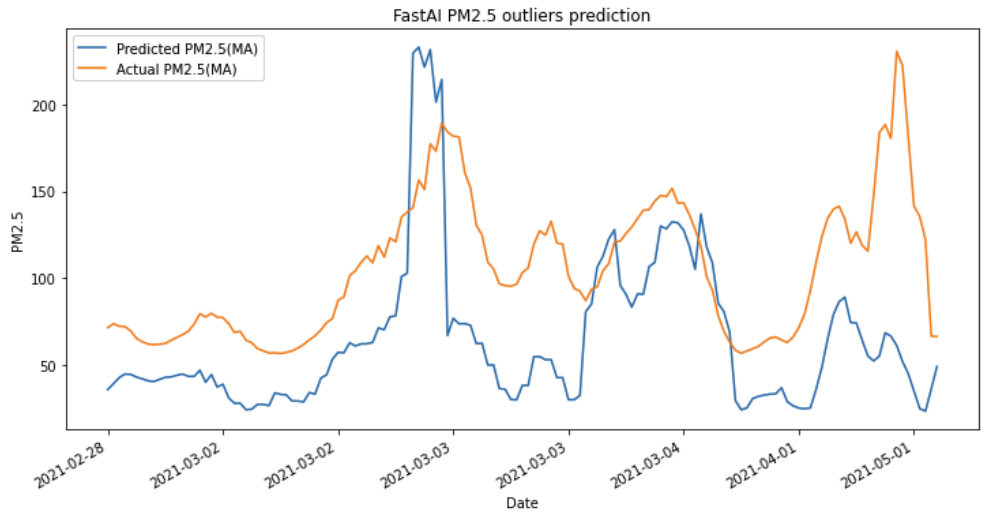
Table 6.2: Comparing results of MTR-1 vs MTR-1E’s performance on Test set A and B

Pollutant	MTR-1				MTR-1E			
	Test set A		Test set B		Test set A		Test set B	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
NO_2	8.74	10.57	7.32	9.83	7.77	9.19	6.22	8.84
$PM_{2.5}$	6.42	8.14	5.13	7.11	5.58	7.08	4.36	6.32
PM_{10}	7.83	9.77	6.98	8.15	6.81	8.30	6.29	7.09

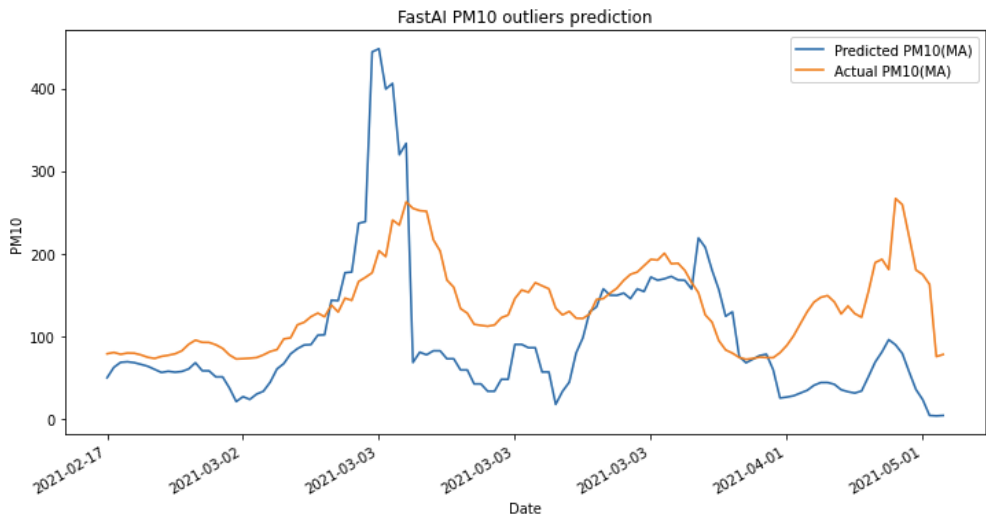
Error (RMSE) and Mean Absolute Error (MAE). RMSE provides a measure of the average magnitude of prediction errors, taking into account both the size and direction of deviations between predicted and actual values. Lower RMSE values indicate better model accuracy, and it is particularly useful when large errors should be penalized more heavily. On the other hand, MAE measures the average absolute magnitude of prediction errors, disregarding their direction. MAE is more robust to outliers and provides a straightforward interpretation as it represents the average prediction error in the same units as the target variable (Morley et al. 2018). Combining these metrics in model evaluation offers a comprehensive view of predictive performance, with RMSE providing insight into error magnitude and MAE offering a more interpretable measure of average error. Both MAE and RMSE were used as evaluation metrics for the MTR-1E model. In comparison to MTR-1, which the experiment sought to enhance, Table 6.2 presents the evaluation scores when the model was tested on two distinct test sets. It is noteworthy that MTR-1E exhibited significant improvement, recording approximately 10-15% enhancement over MTR-1. Additionally, a visual plot of predicted versus actual values, as shown in Figure 6.5, further illustrates a closer alignment than what was initially depicted in Figure 6.2.



(a) Actual vs predicted NO_2 outliers using the MTR-1E model.



(b) Actual vs predicted $PM_{2.5}$ outliers using the MTR-1E model.



(c) Actual vs predicted PM_{10} outliers using the MTR-1E model.

Figure 6.5: Plots of MTR-1E model's performance on predicting NO_2 , $PM_{2.5}$ and PM_{10} outliers

6.4 Model Deployment and Journey Planner Integration

The improved MTR-1E model was deployed on Oracle Cloud to predict pollution levels along road segments using weather data, traffic data, and other relevant highway parameters. This section explains the deployment process of the model and its integration into a route planning mobile application.

6.4.1 Oracle Cloud Deployment

Deploying a fastai model on Oracle cloud involves a series of steps to ensure a seamless transition from model development to real-world usage. Initially, the MTR-1E model needed to be prepared for deployment. This typically involved exporting the model in a format compatible with the Oracle cloud environment. Common choices include exporting it as a PyTorch '.pth' file or converting it to 'ONNX' format, which ensures compatibility with different deployment platforms. With the model ready, the next step was to set up the deployment environment on Oracle cloud. This involved creating a compute instance that includes the necessary dependencies, such as Python, PyTorch, fastai, and any additional libraries used in the model. Oracle cloud provides flexible options for setting up this environment, whether through virtual machines or containerisation using services like Oracle container engine for Kubernetes.

After configuring the environment, the model was then deployed as a serverless function orchestrated through Oracle Functions to serve predictions. Oracle Functions is primarily used for building event-driven applications. One can write functions that respond to various types of events, such as HTTP requests, messages from messaging systems, changes in data storage, and more. These functions execute in response to specific triggers, reducing the need for manual intervention. In this instance, the function loads the model, handles incoming data, and returns predictions via http responses. Oracle cloud's infrastructure provided a scalable and reliable hosting options to accommodate varying levels of traffic and demand.

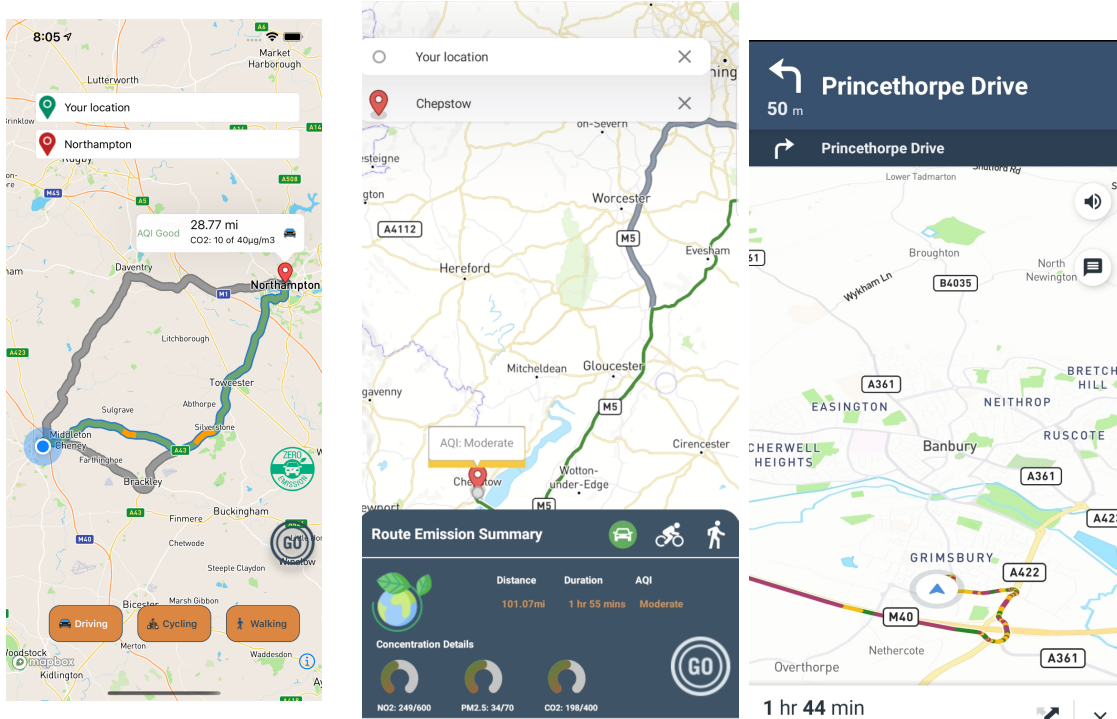
Before deploying the model in production environment, rigorous testing and quality assurance were carried out. This includes comprehensive testing to ensure that the model performs as expected, handling input data correctly, and meeting performance benchmarks. Monitoring and logging mechanisms were also put in place to keep track of the model's behaviour in the production environment and facilitate troubleshooting when issues arise.

Finally, continuous monitoring, maintenance, and updates were integral to the successful deployment of the prediction model. Regular monitoring of the model's performance, applying necessary updates to its dependencies, and being prepared to retrain and redeploy the model when new data becomes available was necessary. Continuous integration and continuous deployment (CI/CD) pipelines were implemented to streamline this process, making it easier to manage and update the deployed fastai model on Oracle Cloud.

6.4.2 Development of REVIS Travel Planner (RTP)

REVIS Travel Planner (RTP) is a multifunctional travel planning application tailored to assist users in making well-informed and efficient travel decisions. It simplifies the process of planning journeys by offering users various essential features and capabilities. Android and iOS versions of RTP were developed as part of this study to enhance availability to users. The Android version of the app was developed using Java programming language and Android SDK, while the iOS version was developed using Swift programming language and Xcode. Figure 6.6 shows screenshots of both versions. One main function of RTP is traffic and emission-based route planning which allows users to decide how to chart out their journeys from a starting point to their desired destination. The app uses a location search feature that simplifies the task of finding start and end locations using postcodes or street addresses. Users of the app can select between cycling, walking and driving travel modes and also choose pollution or traffic-based route suggestions. RTP also provides comprehensive journey details including the total distance to be covered and the estimated duration of their

journey.



(a) RTP iOS automatic user location detection and emission-based route suggestion.

(b) RTP Android automatic user location detection and emission-based route suggestion.

(c) RTP Android colour-coded map display.

Figure 6.6: Screenshots demonstrating key features of the RTP iOS and Android app: user location detection, travel modes, emission-based route planning, and colour-coded route display on maps

6.4.3 Integration of Mapbox Software Development Kit

Mapbox was first introduced in 2010. It was founded by Eric Gundersen and provides a platform for customizing maps and geospatial data visualization, making it a valuable tool for developers and businesses to integrate location-based services into their applications (Zastrow 2015). Since its inception, Mapbox has gained popularity and is widely used for various mapping and location-based applications, including navigation, gaming, and data visualisation. Integrating the Mapbox SDK for emission-based route planning involves a systematic process. Firstly, it was important to ensure the correct configuration was in place including

the integration of API keys. Next was to harness the power of the Mapbox Directions API. This API allows for dynamic and traffic-aware route planning by making requests that fetch real-time traffic information, road closures, and estimated travel times. The functionality of this API was extended by integrating emission data from the MTR-1E prediction model. The aim was to tailor route planning options to meet specific user preferences, such as selecting the quickest or shortest route, optimising routes for various modes of transportation or selecting least polluted routes. Mapbox provided the possibility of creating visually appealing and user-friendly maps that not only display routes but also provide additional context, such as nearby points of interest, landmarks, and terrain information.

6.4.4 Integrating Model Forecasts into RTP

The model was integrated into RTP through web service invocations. The process took two iterations to determine the best integration approach. In the first iteration, the route between two points, A and B, was split into segments, and a prediction was made by the model for each segment. For example, if the route had 50 road segments, the model had to be invoked 50 times. This approach was time-consuming and took too long for the user to get the results. In the second approach, the app utilised historical pollution predictions for common routes. The prediction for frequently queried routes and road segments were stored and readily available in a cache on the user's device, making it less computationally demanding. This approach offered a faster and more resource-efficient solution for the users of RTP. For app visualisation, each road segment was assigned a colour code based on the standard UKAIR daily air quality index (DAQI). Each pollutant had its associated DAQI band on each segment, and the overall DAQI band (computed by averaging these bands) for the route was used to determine the air quality situation on the route.

6.5 Chapter Summary

In this chapter, an audit of the multi-target regression model designed for predicting NO_2 , $PM_{2.5}$, and PM_{10} concentrations on highways is carried out. To accomplish this, an exploratory error analysis was conducted, uncovering previously undetected outliers that had an impact on the model's predictive accuracy. These identified outliers were subsequently used to generate additional training data with the aim of enhancing the model's performance. In addition, the chapter also demonstrated the deployment process of the model and its practical utilisation in real-world scenarios by integrating it into a mobile journey planning application. This application allows users to plan their routes based on emission levels from their starting point to their destination.

Chapter 7

Conclusion and Recommendations

7.1 Chapter Overview

This concluding chapter encapsulates the findings of this study, emphasising its practical implications in technology, society, and economics. It also addresses the challenges encountered during the research and the corresponding mitigation strategies. Lastly, recommendations for future research directions are provided to guide upcoming scholarly pursuits in this field.

7.2 Summary Of The Study

Air Quality has been an age-long issue in the UK and around the world since the industrial revolution of the mid-19th century with its effects significantly felt in the mid-20th century, such as the great smog of London in 1956. Subsequently, there has been tremendous efforts in the UK to eliminate or reduce the impact of air pollution, especially on highways through highway standards and air quality policies. A major problem with developing these standards and policies is the absence of informed decision-making through the acquisition of accurate data and derivation of relevant insight about air quality challenges and opportunities. This study aimed to propose a scalable deep learning framework for monitoring and forecasting pollutant concentration levels on UK highways. A mixed method approach was adopted to understand the necessary aspects of the proposed framework, develop and evaluate a

prototype system (REVIS) in a bid to implement the framework and address the identified gaps in air quality management on highways.

The REVIS system was used to demonstrate the possibility of optimising the cost, efficiency and environmental impact of hardware IoT devices through the development, calibration and deployment of monitoring units to capture real-time pollution data on highways. The devices were developed through an excellent design of both analogue and digital circuitry around it and an iterative approach of calibration and performance optimisation. For data modelling and air quality forecasting, it is important to note that sensors data alone are not sufficient for ensuring accuracy in these models. There are a number of air quality data sources, which exist separately but can provide better insights about air quality if well explored and integrated. An important aspect of this study is to integrate missing or inaccurate data from heterogeneous sources to enhance forecasting accuracy of the developed deep learning model. The essence of this layer is to ensure that data not captured in the hardware layer by the monitoring devices can be integrated into the system to improve the performance. Similarly, an exploratory analysis on the captured and integrated data was conducted to evaluate the impact of different parameters on pollutant concentration. It is well established in literature that weather parameters such as rainfall and temperature influence the dispersion rates of pollutants (Barrera-Animas et al. 2022). Hence, there is need for a more coordinated approach such as the one proposed in this study to manage multiple data sources, which are relevant for accurately forecasting air quality on highways through common data environment and data integration.

Finally, this study contributes to existing body of air quality monitoring knowledge by investigating how additional data which are rarely integrated in TRAP forecasting could help improve accuracy. Unconventional training data for AI models such as terrain data, pollutants background concentration and emissions factor were integrated with the traditional

traffic flow, weather and historic pollution data and used to train multi-target prediction models for NO_2 , $PM_{2.5}$ and PM_{10} . The results of the experiments demonstrate the efficacy of the MTR models albeit with a lot of hyperparameter tuning required. The best performance was achieved with fastai on simultaneous hourly predictions for all three pollutants. The model performed well with $PM_{2.5}$ and PM_{10} and was able to capture peak episodes but struggled with similar spikes for NO_2 . In addition, evaluating the feature importance revealed key contributors to the model’s performance, with traffic, weather, hour of the day and emission factor being the most significant. The limitations of this baseline model prompted further investigation in form of an algorithmic audit into why the model’s performance was not too good with unique peak events. An exploratory error analysis revealed previously undetected outliers that negatively affected the model’s predictive accuracy. These identified outliers were subsequently used to generate additional training data with the objective of enhancing the model’s performance. Standard data processing techniques, including data imputation, aggregation, and transformation, were employed to preprocess the data before model training. The resultant model showed an improvement of approximately 10-15% when tested against both the outlier data set and a subset of the newly generated data set. This outcome highlights the potential for improving underperforming deep learning models through algorithmic audits.

7.3 Reflections on the Quantitative Results

The quantitative results presented in this study demonstrate the effectiveness and limitations of various models in TRAP forecasting across different pollutants and time-frames. By employing multiple performance metrics — MAE, MAPE, and RMSE — the study ensures a comprehensive evaluation of model accuracy and robustness. In the first experiment reported in Chapter Five, models were evaluated over hourly, 8-hourly, 16-hourly, and 24-hourly timesteps. The results revealed that all models performed better on shorter timeframes, with the MultiOutputRegressor model achieving the lowest errors for NO_2 at 1-hour intervals

and the Prophet model excelling in $PM_{2.5}$ predictions. The Fastai model, however, recorded the highest errors across most metrics, indicating it struggled with capturing the intricate dynamics of TRAP. The significant disparity in performance across different pollutants and timeframes highlights the challenge of developing a universally effective model for TRAP forecasting.

For the second experiment, after increasing the number of epochs and introducing lagged variables and hyperparameter tuning, there was a noticeable improvement in the Fastai model's performance. It outperformed the other models across all pollutants, especially in shorter timeframes. This improvement underscores the importance of model tuning and the inclusion of relevant temporal features in enhancing predictive accuracy. The reduction in validation loss over extended epochs indicates a better generalization capability of the Fastai model in this setup. Statistical significance tests, including the Friedman and Wilcoxon signed-rank tests, confirmed that the performance improvements of the Fastai model were statistically significant compared to the MultiOutputRegressor and Prophet models. These tests reinforce the reliability of the improved results obtained in the second experiment.

The comparison with existing studies shows that the proposed method, particularly with the refined Fastai model, achieves lower RMSE scores for NO_2 , $PM_{2.5}$, and PM_{10} , outperforming most of the reviewed approaches. This superior performance can be attributed to the use of categorical embeddings and a comprehensive dataset integrating traffic, weather, and environmental factors. Experiment 3, which tested the model's robustness to missing data, revealed that the Fastai model's accuracy significantly drops when critical features like weather data and background concentrations are missing. This finding emphasises the necessity of comprehensive and continuous data collection to maintain high forecasting accuracy. Results of the feature importance analysis and ablation testing also validated the critical role of traffic and weather parameters in the model's predictions. The insights gained from

these analyses highlight the importance of specific features in TRAP forecasting and suggest directions for optimising data collection efforts to improve model performance.

In summary, the experiments demonstrate that while deep learning models, particularly Fastai, can significantly enhance TRAP forecasting accuracy, their effectiveness is highly dependent on comprehensive data and appropriate feature engineering. The study's findings provide valuable insights into the optimisation of model parameters and the critical importance of continuous, high-quality data collection for accurate air pollution forecasting.

7.4 Challenges

While this study effectively tackled energy interference and cross-sensitivity issues in the developed sensing devices, it encountered certain inconsistencies in the NO₂ data, which could be directly attributed to the chosen pollutant sensor. Further investigation revealed that the sensor's performance was influenced by various factors, including its sensitivity to temperature fluctuations and potential interference from other gases in the atmosphere. To resolve these NO₂ data inconsistencies, an exploration of potential solutions was initiated. Key strategies included the selection of sensors specifically designed to resist environmental cross-sensitivities and implementing hardware filters to shield and protect the sensors from unwanted frequencies and noise. Additionally, environmental compensation algorithms were planned to adjust readings based on known impacts such as temperature or humidity, aiming to enhance the reliability and accuracy of the NO₂ measurements.

In addition to these sensor-related challenges, the study involved the integration of data from various sources, which presented its own set of difficulties. While some data, including historic pollution and certain weather data, were publicly available, access to the remaining data required additional research authorisation requests. Traffic flow data, in particular,

posed accessibility challenges. Data integration also had to contend with the disparity in data formats from different sources, which was resolved through the creation of data integration maps. These integrated data sets were subsequently used to train models with three prominent algorithms: deep learning, time-series, and linear regression. The goal was to demonstrate how well AI models performed with the newly curated data compared to conventional air quality modelling tools.

The results from this study show that just like any other machine learning task, sufficient hyperparameter tuning is required when training these models irrespective of the quality or type of data being used. Despite `fastai`'s default incorporation of new deep learning techniques such as 'entity embeddings for categorical variables', the library's training parameters still needed to be tweaked for better results. The trained model was able to capture general pollution levels including rise in pollution and drop off but was not able to capture unpredictable peak events that could have been caused by specific occurrences such as an extra congestion. This is an indication that more features or peak events data can still be captured in the data set in order to model the specific causes of these peaks. Another approach is to tackle the prediction as a classification problem rather than a regression one. This will enable the use of advanced loss functions like *focal loss* which are designed to force an algorithm to learn rare trends in the data.

Some challenges were encountered during model integration and development of the RTP app. First was the issue of latency which impacted the app's responsiveness until a solution was found. Platform compatibility was also an issue which led to the development of Android and iOS applications separately. Ideally, this study should have adopted newer technologies such as Flutter or Ionic which allows the development of Programmable Web Apps (PWAs) that are able to run on multiple platforms. Similarly, ensuring the security and privacy of user data was another constraint. Handling sensitive location and pollution-related information

required stringent security measures and compliance with data protection regulations, adding complexity to the development process. Real-time data and connectivity issues occasionally affected the app’s performance. Dependence on up-to-the-minute data, including traffic and weather conditions, meant that limited connectivity or data delays could impact the app’s real-time functionality. Finally, User engagement and trust were important but often difficult to establish. Convincing users to adopt the app and place confidence in its pollution predictions required clear explanations of the AI model’s workings and its limitations.

7.5 Implication For Practice

The study’s timing aligns perfectly with the growing global mandate for regular air quality assessment in major cities (Zeng et al. 2019). From a social perspective, the approach proposed here holds the promise of mitigating traffic-related pollution risks faced by citizens worldwide. It addresses the environmental justice concerns, especially in developed countries, where vulnerable communities often lack adequate resources and are disproportionately affected by traffic pollution (Barnes et al. 2019). The improved air quality management system, supported by accurate forecasting, empowers governmental agencies to implement targeted traffic restrictions, provide early warnings of peak pollution episodes, and allocate resources more effectively to the most affected areas. Economically, the cost of air pollution in terms of healthcare expenses and reduced agricultural yields has led to substantial economic losses in many nations (Pandya et al. 2022). While the prediction system proposed in this study cannot single-handedly resolve these economic issues, it plays a significant role when integrated into existing air quality systems, aiding informed decision-making.

From a technological perspective, this study offers a path for streamlining the production of air quality models for practical applications. The multi-target regression models developed in this study provide a solution to the challenge of deploying separate models for each pollutant of interest. Tools such as AWS Lambda, Oracle ADS, and MLflow can automate

this process, offering opportunities for real-time predictions. However, it's crucial to remain vigilant about potential "model drift", where the model's performance can deteriorate as the environment deviates from the training scenarios. A possible remedy involves implementing automatic drift detection and model retraining using updated data, followed by performance comparison with the deployed model.

Conducting an audit of an air pollution prediction algorithm, as demonstrated here, bears profound social and technological consequences. On a social level, it enhances transparency and trust in algorithmic systems, assuaging fears of bias or unaccountability. These audits promote social equity by mitigating biases, ensuring pollution predictions treat diverse demographic groups equally, and contributing to better-informed decisions that may reduce health risks associated with air pollution. This is particularly vital for vulnerable communities. From a technological standpoint, these audits drive advancements in algorithm development, encouraging the creation of more accurate and reliable prediction models. They also ensure compliance with legal regulations, guaranteeing algorithm adherence to environmental standards and data quality requirements. Furthermore, algorithm audits raise awareness about the importance of accurate pollution predictions and educate the public about their potential health and environmental impacts. They can result in substantial cost savings by optimizing pollution control measures and strategies. Lastly, these audits foster international collaboration in addressing global environmental challenges, promoting cooperation among nations to effectively combat air pollution and its consequences.

The study also addresses ethical concerns in AI. Publicised scandals related to biased outcomes, lack of transparency, and data misuse have eroded trust in AI systems and led to calls for mandatory algorithmic ethical assessments (Alon-Barkat & Busuioc 2023). This research bridges the gap between high-level ethical principles and technical fairness and transparency guidelines. It introduces a practical audit approach that promotes transparency

and trust.

7.6 Directions For Future Research

The findings and insights gained from this study open up promising avenues for future research in the realm of air quality monitoring and prediction. One noteworthy direction for further investigation lies in the realm of sensor technology. Given the significance of sensors in capturing air quality data, continued research can focus on the development of more robust and versatile sensors, capable of addressing issues such as cross-sensitivity and temperature variations. Researchers can delve into novel sensor materials and designs, exploring innovative approaches to enhance accuracy and reliability. Expanding on the theme of sensor technology, future research may also explore the integration of advanced sensor networks, including distributed sensor systems and remote sensing technologies. These networks can provide comprehensive, real-time data coverage over broader geographical areas, contributing to more accurate and detailed air quality predictions. Developing sensor networks that can adapt to dynamic environmental conditions and account for various pollutant sources would be a challenging yet rewarding pursuit.

Additionally, ethical considerations in air quality prediction represent a critical research domain. As AI and machine learning models continue to influence environmental and health-related decisions, research can delve into the development of ethical frameworks and audit mechanisms for these algorithms. The aim is to ensure fair, transparent, and unbiased predictions while safeguarding individuals' privacy and rights. From a global perspective, international collaboration and data sharing can foster the development of a unified air quality monitoring system. Future research can explore ways to establish international standards and protocols for data sharing, ensuring consistent and accurate air quality information across borders. This collaborative approach can lead to more effective strategies for combating air pollution, which often transcends geographical boundaries. Finally, further research can

concentrate on the social and economic impacts of improved air quality prediction. Analysing the financial benefits, healthcare savings, and the socio-economic implications of enhanced air quality prediction systems can provide governments and stakeholders with tangible incentives for investing in advanced monitoring and prediction technologies. Addressing environmental justice issues and enhancing the equitable distribution of air quality monitoring resources could also be a significant area of exploration.

In summary, the research conducted in this study acts as a stepping stone towards a comprehensive understanding of air quality dynamics and prediction. Future research endeavours can build upon these foundations to develop innovative sensor technologies, advanced prediction models, ethical frameworks, and decision support systems, ultimately contributing to a cleaner, healthier, and more sustainable environment.

Appendix A: Data summary for pollutant estimation before processing

S/No	Column	Column Description	Non-Null Count	Data type
1	city_name	The name of the city of interest	991662 non-null	object
2	lat	The geographic coordinate of the city of interest (Latitude)	991662 non-null	float64
3	lon	The geographic coordinate of the city of interest (Longitude)	991662 non-null	float64
4	date	The observation time to include date, time, hour and second	991662 non-null	datetime64[ns]
5	rain_desc	Description of measured precipitation	5975 non-null	object
6	rain_1h	Integrated average hourly precipitation measurement (mm)	5658 non-null	float64
7	rain_3h	Integrated precipitation measurement averaged over 3 hrs preceding the observation time (mm)	65 non-null	float64
8	snow_1h	Integrated average hourly snow depth measurement (cm)	77 non-null	float64
9	snow_3h	Integrated snow depth measurement averaged over 3 hrs preceding the observation time (cm)	4 non-null	float64
10	drizzle_desc	Description of measured drizzle	244 non-null	object

11	fog_desc	Description of measured fog	193 non-null	object
12	clouds_desc	Description of measured clouds	72395 non-null	object
13	haze_desc	Description of measured haze	46 non-null	object
14	mist_desc	Description of measured mist	312 non-null	object
15	clear_desc	Description of measured clear	11342 non-null	object
16	snow_desc	Description of measured snow	103 non-null	object
17	storm_desc	Description of measured thunderstorm	1 non-null	object
18	temp	Captured average hourly temperature (°C)	991662 non-null	float64
19	temp_min	Captured minimum temperature over a 24-hr period (°C)	991662 non-null	float64
20	temp_max	Captured maximum temperature over a 24-hr period (°C)	991662 non-null	float64
21	feels_like	Integrated measurement of human impression of weather (K)	991662 non-null	float64
22	pressure	Captured average hourly pressure (hPa)	991662 non-null	int64
23	humidity	Captured average hourly relative humidity (ϕ)	991662 non-null	int64
24	wind_speed	Integrated average hourly wind speed (knots)	991662 non-null	float64

25	wind_direction	Integrated average hourly wind direction (true degrees)	991662 non-null	int64
26	clouds_all	Integrated hourly measurement of cloudiness (%)	991662 non-null	float64
27	ozone	Integrated average hourly ozone ($\mu g/m^3$)	181233 non-null	float64
28	ozone_avg6h	Integrated ozone readings averaged over 6 hrs preceding the observation time ($\mu g/m^3$)	181233 non-null	float64
29	NO_2	Captured average hourly NO_2 (ppb)	121207 non-null	float64
30	NO_2 _avg6h	Captured NO_2 readings averaged over 6 hrs preceding the observation time (ppb)	121207 non-null	float64
31	PM_{10}	Captured average hourly PM_{10} ($\mu g/m^3$)	121207 non-null	float64
32	PM_{10} _avg6h	Captured PM_{10} readings averaged over 6 hrs preceding the observation time ($\mu g/m^3$)	121207 non-null	float64
33	$PM_{2.5}$	Captured average hourly $PM_{2.5}$ ($\mu g/m^3$)	121207 non-null	float64
34	$PM_{2.5}$ _avg6h	Captured $PM_{2.5}$ readings averaged over 6 hrs preceding the observation time ($\mu g/m^3$)	121207 non-null	float64

Appendix B: List of Attributes After Processing, Including Classification as Categorical, Continuous, Independent, and Dependent Variables

S/No	Attribute Name	Attribute Type
1	city_name	Categorical, Independent
2	lat	Categorical, Independent
3	lon	Categorical, Independent
4	year	Categorical, Independent
5	month	Categorical, Independent
6	week	Categorical, Independent
7	day	Categorical, Independent
8	dayofweek	Categorical, Independent
9	dayofyear	Categorical, Independent
10	is_month_end	Categorical, Independent
11	is_month_start	Categorical, Independent
12	is_quarter_end	Categorical, Independent
13	is_quarter_start	Categorical, Independent
14	is_year_end	Categorical, Independent
15	is_year_start	Categorical, Independent

16	rain_1h	Continuous, Independent
17	snow_1h	Continuous, Independent
18	temp	Continuous, Independent
19	temp_min	Continuous, Independent
20	temp_max	Continuous, Independent
21	feels_like	Continuous, Independent
22	pressure	Continuous, Independent
23	humidity	Continuous, Independent
24	wind_speed	Continuous, Independent
25	wind_direction	Continuous, Independent
26	clouds_all	Continuous, Independent
27	ozone	Continuous, Independent
28	ozone_avg6h	Continuous, Independent
29	no ₂	Continuous, Dependent
30	no ₂ _avg6h	Continuous, Independent
31	pm _{2.5}	Dependent

32	$pm_{2.5}$ -avg6h	Continuous, Independent
34	PM_{10}	Continuous, Dependent
34	PM_{10} -avg6h	Continuous, Independent

Appendix C: List of attributes captured for MTR pollutant concentration forecasting

S/No	Column	Column Description	Range	Non-Null Count	Variable type
1	datetimehour	Hour variable extracted after preprocessing of datetime column	0-23	11990 non-null	Categorical
2	datetimeminute	Minute variable extracted after preprocessing of datetime column	0-59	11990 non-null	Categorical
3	datetimesecond	Second variable extracted after preprocessing of datetime column	0-59	11990 non-null	Categorical
4	datetimeelapsed	Time elapsed variable extracted after preprocessing of datetime column	1.60e+9-1.63e+9	11990 non-null	Continuous
5	datetimeyear	Year variable extracted after preprocessing of datetime column	2020-2021	11990 non-null	Categorical
6	datetimemonth	Month variable extracted after preprocessing of datetime column	1-11	11990 non-null	Categorical
7	datetimeweek	Week variable extracted after preprocessing of datetime column	1-47	11990 non-null	Categorical
8	datetimeday	Day variable extracted after preprocessing of datetime column	1-31	11990 non-null	Categorical

9	datetimedayofweek	Day of week variable extracted after pre-processing of datetime column	0-6	11990 non-null	Categorical
10	datetimedayofyear	Day of year variable extracted after pre-processing of datetime column	8-322	11990 non-null	Categorical
11	datetimeis_month_end	Boolean variable to indicate if the day is month end	0/1	11990 non-null	Categorical
12	datetimeis_month_start	Boolean variable to indicate if the day is start of the month	0/1	11990 non-null	Categorical
13	datetimeis_quarter_end	Boolean variable to indicate if the day is the end of a quarter	0/1	11990 non-null	Categorical
14	datetimeis_quarter_start	Boolean variable to indicate if the day is the start of a quarter	0/1	11990 non-null	Categorical
15	datetimeis_year_end	Boolean variable to indicate if the day is the start of the year	0/1	11990 non-null	Categorical
16	datetimeis_year_start	Boolean variable to indicate if the day is the end of the year	0/1	11990 non-null	Categorical
17	road_name	The name of the highway of interest	-	11990 non-null	Categorical
18	region_name	The name of the region where the highway is located	-	11990 non-null	Categorical
19	segment_name	The name of the highway segment where the IoT device is located	-	11990 non-null	Categorical
20	NO_2	Integrated average hourly NO_2 (ppb) reading from AURN station	0.63-132.37	11990 non-null	Continuous
21	$PM_{2.5}$	Captured $PM_{2.5}$ ($\mu g/m^3$) reading from REVIS IoT devices	0.69-401.01	10879 non-null	Continuous
22	PM_{10}	Captured PM_{10} ($\mu g/m^3$) reading from REVIS IoT devices	0.77-617.35	10879 non-null	Continuous

23	air_quality_index	The AQI for the highway segment of interest computed from the pollutant concentration readings	0-6.5	11990 non-null	Continuous
24	background_NO ₂	The background NO ₂ concentration for the highway segment of interest	8.06-27.99	11990 non-null	Continuous
25	background_PM _{2.5}	The background PM _{2.5} concentration for the highway segment of interest	7.88-12.55	11990 non-null	Continuous
26	background_PM ₁₀	The background PM ₁₀ concentration for the highway segment of interest	11.94-19.55	11990 non-null	Continuous
27	NO ₂ _emission_factor	Calculated NO ₂ emission factor based on different vehicle types on the highway at that time point	0-14823	11990 non-null	Continuous
28	PM_emission_factor	Calculated PM ₁₀ emission factor based on different vehicle types on the highway at that time point	0-19982	11990 non-null	Continuous
29	bike_count	Captured bike count from REVIS IoT devices	-	6 non-null	Continuous
30	bike_avg_speed	Captured bike avg speed	-	6 non-null	Continuous
31	car_count	Integrated car count from TMU sites	0-3515	10949 non-null	Continuous
32	car_avg_speed	Captured car avg speed from REVIS IoT devices	-	6 non-null	Continuous
33	bus_count	Integrated bus count from TMU sites	0-412	10949 non-null	Continuous
34	bus_avg_speed	Integrated bus avg speed	-	6 non-null	Continuous
35	lgv_count	Integrated LGV count from TMU sites	0-245	10949 non-null	Continuous
36	lgv_avg_speed	Captured LGV avg speed from REVIS IoT devices	-	6 non-null	Continuous
37	hgv_count	Integrated HGV count from TMU sites	0-383	10949 non-null	Continuous

38	hgv_avg_speed	Captured HGV avg speed from REVIS IoT devices	-	6 non-null	Continuous
39	other_avg_speed	Integrated average travelling speed from TMU sites	0-76.25	10949 non-null	Continuous
40	humidity	Captured average hourly relative humidity from REVIS IoT devices (ϕ)	23.65-99.99	11990 non-null	Continuous
41	wind_speed	Integrated hourly modelled wind speed (knots) from AURN station	0-16.2	11990 non-null	Continuous
42	wind_direction	Integrated hourly modelled wind direction (true degrees) from AURN station	0-360	11990 non-null	Continuous
43	temperature	Captured average hourly temperature ($^{\circ}\text{C}$) from REVIS IoT devices	-2.95-44.07	10879 non-null	Continuous
44	pressure	Captured average hourly pressure (hPa) from REVIS IoT devices	979.31-1042.72	10879 non-null	Continuous

Bibliography

- Ahlers, D., Driscoll, P. A., Kraemer, F. A., Anthonisen, F. V. & Krogstie, J. (2016), A measurement-driven approach to understand urban greenhouse gas emissions in nordic cities, NIK.
- Ahmed, E., Ahmed, A., Yaqoob, I., Shuja, J., Gani, A., Imran, M. & Shoaib, M. (2017), ‘Bringing computation closer towards user network: Is edge computing the solution?’, *IEEE Communications Magazine* **55**, 138 – 144.
- Akinosho, T. D., Bilal, M., Hayes, E. T., Ajayi, A., Ahmed, A. & Khan, Z. (2023), ‘Deep learning-based multi-target regression for traffic-related air pollution forecasting’, *Machine Learning with Applications* p. 100474.
- Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O. & Ahmed, A. A. (2020), ‘Deep learning in the construction industry: A review of present status and future innovations’, *Journal of Building Engineering* p. 101827.
- Akinosho, T. D., Oyedele, L. O., Bilal, M., Barrera-Animas, A. Y., Gbadamosi, A.-Q. & Olawale, O. A. (2022), ‘A scalable deep learning system for monitoring and forecasting pollutant concentration levels on uk highways’, *Ecological Informatics* **69**, 101609.
- Al-Dweik, A., Muresan, R., Mayhew, M. & Lieberman, M. (2017), Iot-based multifunctional scalable real-time enhanced road side unit for intelligent transportation systems, *in* ‘2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE)’, IEEE, pp. 1–6.

- Alain, G. & Bengio, Y. (2014), ‘What regularized auto-encoders learn from the data-generating distribution’, *The Journal of Machine Learning Research* **15**(1), 3563–3593.
- Alon-Barkat, S. & Busuioc, M. (2023), ‘Human–ai interactions in public sector decision making: “automation bias” and “selective adherence” to algorithmic advice’, *Journal of Public Administration Research and Theory* **33**(1), 153–169.
- Alvanchi, A., Rahimi, M., Mousavi, M. & Alikhani, H. (2020), ‘Construction schedule, an influential factor on air pollution in urban infrastructure projects’, *Journal of Cleaner Production* **255**, 120222.
- Amato, F., Alastuey, A., De La Rosa, J., Gonzalez Castanedo, Y., Sánchez de la Campa, A., Pandolfi, M., Lozano, A., Contreras González, J. & Querol, X. (2014), ‘Trends of road dust emissions contributions on ambient air particulate levels at rural, urban and industrial sites in southern Spain’, *Atmospheric Chemistry and Physics* **14**(7), 3533–3544.
- Analitis, A., Michelozzi, P., D’Ippoliti, D., De’Donato, F., Menne, B., Matthies, F., Atkinson, R. W., Iñiguez, C., Basagaña, X., Schneider, A. et al. (2014), ‘Effects of heat waves on mortality: effect modification and confounding by air pollutants’, *Epidemiology* pp. 15–22.
- Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. & Hamburg, S. P. (2017), ‘High-resolution air pollution mapping with Google Street View cars: exploiting big data’, *Environmental Science & Technology* **51**(12), 6999–7008.
- Arthurs, P., Gillam, L., Krause, P., Wang, N., Halder, K. & Mouzakitis, A. (2021), ‘A taxonomy and survey of edge cloud computing for intelligent transportation systems and connected vehicles’, *IEEE Transactions on Intelligent Transportation Systems* **23**(7), 6206–6221.
- Arunachalam, S., Valencia, A., Akita, Y., Serre, M. L., Omary, M., Garcia, V. & Isakov, V. (2014), ‘A method for estimating urban background concentrations in support of hybrid

- air pollution modeling for environmental health studies’, *International journal of environmental research and public health* **11**(10), 10518–10536.
- Bálint, A., Fagerlind, H., Martinsson, J. & Holmqvist, K. (2014), ‘Accident analysis for traffic safety aspects of high capacity transports’.
- Barikayeva, N., Nikolenko, D. & Ivanova, J. (2018), About forecasting air pollution in the construction of highways, in ‘IOP Conference Series: Materials Science and Engineering’, Vol. 463, IOP Publishing, p. 042016.
- Barnes, C. (1992), ‘Qualitative research: valuable or irrelevant?’, *Disability, handicap & society* **7**(2), 115–124.
- Barnes, J. H., Chatterton, T. J. & Longhurst, J. W. (2019), ‘Emissions vs exposure: Increasing injustice from road traffic-related air pollution in the united kingdom’, *Transportation research part D: transport and environment* **73**, 56–66.
- Baron, R. & Saffell, J. (2017), ‘Amperometric gas sensors as a low cost emerging technology platform for air quality monitoring applications: A review’, *ACS sensors* **2**(11), 1553–1566.
- Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D. & Akanbi, L. A. (2022), ‘Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting’, *Machine Learning with Applications* **7**, 100204.
- Barthwal, A. & Acharya, D. (2018), An internet of things system for sensing, analysis & forecasting urban air quality, in ‘2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)’, IEEE, pp. 1–6.
- Beevers, S. D., Kitwiroon, N., Williams, M. L. & Carslaw, D. C. (2012), ‘One way coupling of cmaq and a road source dispersion model for fine scale air pollution predictions’, *Atmospheric environment* **59**, 47–58.

- Bengio, Y. (2009), ‘Learning deep architectures for ai’, *Foundations and trends in Machine Learning*, **2**, 1–55.
- Biesta, G. (2010), ‘Pragmatism and the philosophical foundations of mixed methods research’, *Sage handbook of mixed methods in social and behavioral research* **2**, 95–118.
- Bilal, M. & Oyedele, L. O. (2020), ‘Guidelines for applied machine learning in construction industry—a case of profit margins estimation’, *Advanced Engineering Informatics* **43**, 101013.
- Bird-David, N. (1999), ‘“animism” revisited: personhood, environment, and relational epistemology’, *Current anthropology* **40**(S1), S67–S91.
- Bloemraad, I. (2013), ‘The promise and pitfalls of comparative research design in the study of migration’, *Migration Studies* **1**(1), 27–46.
- Bogdan, R. C. & Biklen, S. K. (1998), ‘Foundations of qualitative research in education’, *Qualitative research in education: An introduction to theory and methods* pp. 1–48.
- Borchani, H., Varando, G., Bielza, C. & Larranaga, P. (2015), ‘A survey on multi-output regression’, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **5**(5), 216–233.
- Borrego, C., Coutinho, M., Costa, A. M., Ginja, J., Ribeiro, C., Monteiro, A., Ribeiro, I., Valente, J., Amorim, J., Martins, H. et al. (2015), ‘Challenges for a new air quality directive: The role of monitoring and modelling techniques’, *Urban Climate* **14**, 328–341.
- Boss, P., Doherty, W. J., LaRossa, R., Schumm, W. R. & Steinmetz, S. K. (1993), *Sourcebook of family theories and methods: A contextual approach*, Springer Science & Business Media.
- Brannen, J. (2005), ‘Mixing methods: The entry of qualitative and quantitative approaches into the research process’, *International journal of social research methodology* **8**(3), 173–184.

- Bui, T.-C., Le, V.-D. & Cha, S.-K. (2018), ‘A deep learning approach for forecasting air pollution in south korea using lstm’, *arXiv preprint arXiv:1804.07891* .
- Bunge, M. (1993), ‘Realism and antirealism in social science’, *Theory and Decision* **35**(3), 207–235.
- Burrell, G. & Morgan, G. (1979), “sociological paradigms and organisational analysis” elements of the sociology of corporate life’.
- Cabaneros, S. M. S., Calautit, J. K. S. & Hughes, B. R. (2017), ‘Hybrid artificial neural network models for effective prediction and mitigation of urban roadside no2 pollution’, *Energy Procedia* **142**, 3524–3530.
- Carabetta, J. L. M. (2019), Mining jams into pollution: how Waze data helps estimating air pollution in large cities, PhD thesis.
- Carreira-Perpiñán, M. Á. & Hinton, G. E. (2005), On contrastive divergence learning.
- Carullo, A., Corbellini, S. & Grassini, S. (2007), ‘A remotely controlled calibrator for chemical pollutant measuring-units’, *IEEE Transactions on Instrumentation and Measurement* **56**(4), 1212–1218.
- Casula, M., Rangarajan, N. & Shields, P. (2021), ‘The potential of working hypotheses for deductive exploratory research’, *Quality & Quantity* **55**(5), 1703–1725.
- Catalano, M. & Galatioto, F. (2017), ‘Enhanced transport-related air pollution prediction through a novel metamodel approach’, *Transportation Research Part D: Transport and Environment* **55**, 262–276.
- CCC (2017), ‘Carbon budgets: how we monitor emissions targets’, Available from: <https://www.theccc.org.uk/tackling-climate-change/reducing-carbon-emissions/carbon-budgets-and-targets/> [Accessed: Accessed: 2019-01-30].

- Cecilia, J. M., Timón, I., Soto, J., Santa, J., Pereñíguez, F. & Muñoz, A. (2018), ‘High-throughput infrastructure for advanced its services: A case study on air pollution monitoring’, *IEEE Transactions on Intelligent Transportation Systems* **19**(7), 2246–2257.
- CERC (2019), ‘Cambridge environmental research consultants - detailed modelling of no2 at market hill aqma, maldon’.
- CERC (2020), ‘Adms-roads extra user guide’, https://www.cerc.co.uk/environmental-software/assets/data/doc_userguides/CERC_ADMS-RoadsExtra5.0_User_Guide.pdf. Accessed: 2023-04-30.
- Cha, Y., Song, C.-K., Jeon, K.-h. & Yi, S.-M. (2023), ‘Factors affecting recent pm2. 5 concentrations in china and south korea from 2016 to 2020’, *Science of The Total Environment* **881**, 163524.
- Chalmers, D., Manley, D. & Wasserman, R. (2009), *Metametaphysics: New essays on the foundations of ontology*, Oxford University Press.
- Chang, I.-C. (2019), ‘Identifying leading nodes of pm2. 5 monitoring network in taiwan with big data-oriented social network analysis’, *Aerosol and Air Quality Research* **19**(12), 2844–2864.
- Chauhan, R., Kaur, H. & Alankar, B. (2021), ‘Air quality forecast using convolutional neural network for sustainable development in urban environments’, *Sustainable Cities and Society* **75**, 103239.
- Chen, J., Li, K., Deng, Q., Li, K. & Philip, S. Y. (2019), ‘Distributed deep learning model for intelligent video surveillance systems with edge computing’, *IEEE Transactions on Industrial Informatics* .
- Chen, X. & Ye, J. (2019), ‘When the wind blows: spatial spillover effects of urban air pollution in china’, *Journal of Environmental Planning and Management* **62**(8), 1359–1376.

Chibucos, T. R., Leite, R. W. & Weis, D. L. (2005), *Readings in family theory*, Sage.

Cohen, L., Manion, L. & Morrison, K. (2002), *Research methods in education*, routledge.

Comte, A. (1856), ‘A general view of positivism [discours sur l’esprit positif 1844]’.

Conrad, C. C. & Hilchey, K. G. (2011), ‘A review of citizen science and community-based environmental monitoring: issues and opportunities’, *Environmental monitoring and assessment* **176**, 273–291.

Cooksey, R. W. & McDonald, G. M. (2011), *Surviving and thriving in postgraduate research*, Springer.

Creswel, J. W. (2009), ‘Research design: Qualitative, quantitative, and mixed methods approaches’, *Los angeles: University of Nebraska–Lincoln* .

Creswell, J. W. (2009), ‘Research design: Qualitative and mixed methods approaches’, *London and Thousand Oaks: Sage Publications* .

Crotty, M. (1998), *The foundations of social research: Meaning and perspective in the research process*, Sage.

Cui, L. & Wang, S. (2021), ‘Mapping the daily nitrous acid (hono) concentrations across china during 2006–2017 through ensemble machine-learning algorithm’, *Science of The Total Environment* **785**, 147325.

DEFRA (2019), ‘Clean air strategy’.

URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/81787/air-strategy-2019.pdf

DEFRA (2020), ‘Air quality appraisal: impact pathways approach’.

URL: <https://www.gov.uk/government/publications/assess-the-impact-of-air-quality/air-quality-appraisal-impactpathways-approach>

- Devarakonda, S., Sevusu, P., Liu, H., Liu, R., Iftode, L. & Nath, B. (2013), Real-time air quality monitoring through mobile sensing in metropolitan areas, *in* ‘Proceedings of the 2nd ACM SIGKDD international workshop on urban computing’, ACM, p. 15.
- Diesing, P. (1966), ‘Objectivism vs. subjectivism in the social sciences’, *Philosophy of Science* **33**(1/2), 124–133.
- Duan, J., Gong, Y., Luo, J. & Zhao, Z. (2023), ‘Air-quality prediction based on the arima-cnn-lstm combination model optimized by dung beetle optimizer’, *Scientific Reports* **13**(1), 12127.
- EEA (2016), ‘Premature deaths attributable to air pollution’, *Air quality in Europe — 2015 report* .
- El Fazziki, A., Benslimane, D., Sadiq, A., Ouarzazi, J. & Sadgal, M. (2017), ‘An agent based traffic regulation system for the roadside air quality control’, *IEEE Access* **5**, 13192–13201.
- Etikan, I., Musa, S. A. & Alkassim, R. S. (2016), ‘Comparison of convenience sampling and purposive sampling’, *American journal of theoretical and applied statistics* **5**(1), 1–4.
- Eysenck, M. W. & Keane, M. T. (2015), *Cognitive psychology: A student’s handbook*, Psychology press.
- Fann, K. T. (2012), *Peirce’s theory of abduction*, Springer Science & Business Media.
- Fayaz, S. A., Zaman, M., Kaul, S. & Butt, M. A. (2022), ‘Is deep learning on tabular data enough? an assessment’, *International Journal of Advanced Computer Science and Applications* **13**(4).
- Finnis, J. (1980), ‘Natural rights and natural law’.
- Fong, I. H., Li, T., Fong, S., Wong, R. K. & Tallon-Ballesteros, A. J. (2020), ‘Predicting concentration levels of air pollutants by transfer learning and recurrent neural network’, *Knowledge-Based Systems* **192**, 105622.

- Gariazzo, C., Carlino, G., Silibello, C., Renzi, M., Finardi, S., Pepe, N., Radice, P., Forastiere, F., Michelozzi, P., Viegi, G. et al. (2020), 'A multi-city air pollution population exposure study: Combined use of chemical-transport and random-forest models with dynamic population data', *Science of The Total Environment* **724**, 138102.
- Gerring, J. (2006), *Case study research: Principles and practices*, Cambridge university press.
- Glaser, B. & Strauss, A. (1967), 'Grounded theory: The discovery of grounded theory', *Sociology the journal of the British sociological association* **12**(1), 27–49.
- Goyal, T., Singh, A., Chhaya, S., Vikas, A., Garg, P., Malik, R. & Sen, R. (2018), Low cost platform design for pollution measurement in delhi-ncr using vehicle-mounted sensors, *in* 'Proceedings of the 24th Annual International Conference on Mobile Computing and Networking', pp. 759–761.
- Grassel, E. & Schirmer, B. (2006), 'The use of volunteers to support family caregivers of dementia patients: results of a prospective longitudinal study investigating expectations towards and experience with training and professional support', *Zeitschrift Fur Gerontologie Und Geriatrie* **39**(3), 217–226.
- Gray, D. E. (2013), *Doing research in the real world*, Sage.
- Grix, J. (2004), *The foundations of research: a student's guide*, Macmillan International Higher Education.
- Grubb, D. & Symons, J. (1987), 'Bias in regressions with a lagged dependent variable', *Econometric Theory* **3**(3), 371–386.
- Guarino, A., Lettieri, N., Malandrino, D., Zaccagnino, R. & Capo, C. (2022), 'Adam or eve? automatic users' gender classification via gestures analysis on touch devices', *Neural Computing and Applications* **34**(21), 18473–18495.

- Guba, E. G., Lincoln, Y. S. et al. (1994), ‘Competing paradigms in qualitative research’, *Handbook of qualitative research* **2**(163-194), 105.
- Guba, E. & Lincoln, Y. (1989), ‘What is this constructivist paradigm anyway’, *Fourth Generation Evaluation*. London: Sage .
- Guevara-López, U., Altamirano-Bustamante, M. M. & Viesca-Treviño, C. (2015), ‘New frontiers in the future of palliative care: real-world bioethical dilemmas and axiology of clinical practice’, *BMC medical ethics* **16**(1), 11.
- Gulliver, J. & Briggs, D. (2011), ‘Stems-air: A simple gis-based air pollution dispersion model for city-wide exposure assessment’, *Science of the total environment* **409**(12), 2419–2429.
- Guo, C. & Berkhahn, F. (2016), ‘Entity embeddings of categorical variables’, *arXiv preprint arXiv:1604.06737* .
- Guyer, P. & Horstmann, R.-P. (2015), ‘Idealism’.
- Hadeed, S. J., O’Rourke, M. K., Burgess, J. L., Harris, R. B. & Canales, R. A. (2020), ‘Imputation methods for addressing missing data in short-term monitoring of air pollutants’, *Science of The Total Environment* **730**, 139140.
- Hamilton, L. & Corbett-Whittier, C. (2012), *Using case study in education research*, Sage.
- Harrison, H., Birks, M., Franklin, R., Mills, J. et al. (2017), Case study research: Foundations and methodological orientations, in ‘Forum qualitative Sozialforschung/Forum: qualitative social research’, Vol. 18.
- Hashad, K., Gu, J., Yang, B., Rong, M., Chen, E., Ma, X. & Zhang, K. M. (2021), ‘Designing roadside green infrastructure to mitigate traffic-related air pollution using machine learning’, *Science of The Total Environment* **773**, 144760.
- Hatcher, W. G. & Yu, W. (2018), ‘A survey of deep learning: Platforms, applications and emerging research trends’, *IEEE access* **6**, 24411–24432.

- Heist, D., Isakov, V., Perry, S., Snyder, M., Venkatram, A., Hood, C., Stocker, J., Carruthers, D., Arunachalam, S. & Owen, R. C. (2013), 'Estimating near-road pollutant dispersion: A model inter-comparison', *Transportation Research Part D: Transport and Environment* **25**, 93–105.
- Hendry, G. D., Frommer, M. & Walker, R. A. (1999), 'Constructivism and problem-based learning', *Journal of further and higher education* **23**(3), 369–371.
- Heron, J. (1996), *Co-operative inquiry: Research into the human condition*, Sage.
- Horizon Nuclear Power (2018), 'Wylfa newydd project 6.3.27 es volume c - road traffic-related effects (project-wide) app c4-1 - project-wide modelling of road traffic emissions'. Accessed: 2023-04-30.
- House of Commons (2024), 'Air quality: policies, proposals and concerns', Available from: <https://researchbriefings.files.parliament.uk/documents/CBP-9600/CBP-9600.pdf> [Accessed: 2024-03-01].
- Howard, J. & Gugger, S. (2020), 'Fastai: A layered api for deep learning', *Information* **11**(2), 108.
- Hulin, M., Simoni, M., Viegi, G. & Annesi-Maesano, I. (2012), 'Respiratory health and indoor air pollutants based on quantitative exposure assessments'.
- Idrees, Z. & Zheng, L. (2020), 'Low cost air pollution monitoring systems: A review of protocols and enabling technologies', *Journal of Industrial Information Integration* **17**, 100123.
- Jackson, K. R., Ramakrishnan, L., Muriki, K., Canon, S., Cholia, S., Shalf, J., Wasserman, H. J. & Wright, N. J. (2010), Performance analysis of high performance computing applications on the amazon web services cloud, in '2010 IEEE second international conference on cloud computing technology and science', IEEE, pp. 159–168.

- Jida, S. N., Hetet, J.-F., Chesse, P. & Guadie, A. (2021), ‘Roadside vehicle particulate matter concentration estimation using artificial neural network model in addis ababa, ethiopia’, *journal of environmental sciences* **101**, 428–439.
- Johnson, J., Kennedy, R. & Khoshgoftaar, T. (2023), ‘Learning from highly imbalanced big data with label noise’, *International Journal on Artificial Intelligence Tools* .
- Johnson-Laird, P. N. (1999), ‘Deductive reasoning’, *Annual review of psychology* **50**(1), 109–135.
- Kadri, A., Yaacoub, E., Mushtaha, M. & Abu-Dayya, A. (2013), Wireless sensor network for real-time air pollution monitoring, *in* ‘2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)’, IEEE, pp. 1–5.
- Kaginalkar, A., Kumar, S., Gargava, P. & Niyogi, D. (2021), ‘Review of urban computing in air quality management as smart city service: An integrated iot, ai, and cloud technology perspective’, *Urban Climate* **39**, 100972.
- Karamchandani, P., Vijayaraghavan, K. & Yarwood, G. (2011), ‘Sub-grid scale plume modeling’, *Atmosphere* **2**(3), 389–406.
- Karner, A. A., Eisinger, D. S. & Niemeier, D. A. (2010), ‘Near-roadway air quality: synthesizing the findings from real-world data’, *Environmental science & technology* **44**(14), 5334–5344.
- Kerlinger, F. N. F. N. (1986), *Foundations of behavioral research*, 3rd ed. edn, Holt, Rinehart and Winston, New York.
- URL:** <https://www.worldcat.org/title/foundations-of-behavioral-research/oclc/12135319>
- Kivunja, C. & Kuyini, A. B. (2017), ‘Understanding and applying research paradigms in educational contexts.’, *International Journal of higher education* **6**(5), 26–41.

- Kocev, D., Džeroski, S., White, M. D., Newell, G. R. & Griffioen, P. (2009), ‘Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition’, *Ecological Modelling* **220**(8), 1159–1168.
- Korneva, E. & Blockeel, H. (2020), Towards better evaluation of multi-target regression models, in ‘Joint European Conference on Machine Learning and Knowledge Discovery in Databases’, Springer, pp. 353–362.
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, NIPS’12, Curran Associates Inc., USA, pp. 1097–1105.
URL: <http://dl.acm.org/citation.cfm?id=2999134.2999257>
- Krogstie, J. (2015), ‘Capturing enterprise data integration challenges using a semiotic data quality framework’, *Business & information systems engineering* **57**, 27–36.
- Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L. & Britter, R. (2015), ‘The rise of low-cost sensing for managing air pollution in cities’, *Environment international* **75**, 199–205.
- Kumaresan, S. J., Abhilash, A., Prasad, S. R., Ekambaranath, T. & Kamaraj, P. (2021), ‘Air pollution monitoring system using internet of things.’, *Turkish Online Journal of Qualitative Inquiry* **12**(9).
- Kuss, P. & Nicholas, K. A. (2022), ‘A dozen effective interventions to reduce car use in european cities: Lessons learned from a meta-analysis and transition management’, *Case studies on transport policy* **10**(3), 1494–1513.
- Leedy, P. D. & Ormrod, J. E. (2005), *Practical research*, Pearson Custom.
- Levin, K. A. (2006), ‘Study design iii: Cross-sectional studies’, *Evidence-based dentistry* **7**(1), 24.

- Lewis, A. C., Lee, J. D., Edwards, P. M., Shaw, M. D., Evans, M. J., Moller, S. J., Smith, K. R., Buckley, J. W., Ellis, M., Gillot, S. R. et al. (2016), 'Evaluating the performance of low cost chemical sensors for air pollution research', *Faraday discussions* **189**, 85–103.
- Li, Y., Wu, F.-X. & Ngom, A. (2018), 'A review on machine learning principles for multi-view biological data integration', *Briefings in bioinformatics* **19**(2), 325–340.
- Li, Z., Yim, S. H.-L. & Ho, K.-F. (2020), 'High temporal resolution prediction of street-level pm_{2.5} and nox concentrations using machine learning approach', *Journal of Cleaner Production* **268**, 121975.
- Liang, M., Chao, Y., Tu, Y. & Xu, T. (2023), 'Vehicle pollutant dispersion in the urban atmospheric environment: A review of mechanism, modeling, and application', *Atmosphere* **14**(2), 279.
- Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X. & Wang, Z. (2020), 'Deep learning for air quality forecasts: a review', *Current Pollution Reports* **6**, 399–409.
- Lin, J. & Ge, Y. (2006), 'Impacts of traffic heterogeneity on roadside air pollution concentration', *Transportation Research Part D: Transport and Environment* **11**(2), 166–170.
- Lincoln, Y. & Guba, E. (1985), 'Thousand oaks', *Naturalistic Inquiry* pp. 290–296.
- Lisboa, P., Saralajew, S., Vellido, A., Fernández-Domenech, R. & Villmann, T. (2023), 'The coming of age of interpretable and explainable machine learning models', *Neurocomputing* **535**, 25–39.
- Lozhkin, V., Tarkhov, D., Timofeev, V., Lozhkina, O. & Vasilyev, A. (2016), Differential neural network approach in information process for prediction of roadside air pollution by peat fire, in 'IOP conference series: materials science and engineering', Vol. 158, IOP Publishing, p. 012063.

- Lv, Y., Chen, X., Wei, S., Zhu, R., Wang, B., Chen, B., Kong, M. & Zhang, J. J. (2020), 'Sources, concentrations, and transport models of ultrafine particles near highways: a literature review', *Building and Environment* p. 107325.
- Mabahwi, N. A. B., Leh, O. L. H. & Omar, D. (2014), 'Human health and wellbeing: Human health effect of air pollution', *Procedia-Social and Behavioral Sciences* **153**, 221–229.
- Mackenzie, N. & Knipe, S. (2006), 'Research dilemmas: Paradigms, methods and methodology', *Issues in educational research* **16**(2), 193–205.
- Manna, S., Bhunia, S. S. & Mukherjee, N. (2014), Vehicular pollution monitoring using iot, *in* 'International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)', IEEE, pp. 1–5.
- Martín-Baos, J. Á., Rodríguez-Benitez, L., García-Ródenas, R. & Liu, J. (2022), 'Iot based monitoring of air quality and traffic using regression analysis', *Applied Soft Computing* **115**, 108282.
- Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O. & Kallel, A. (2020), 'A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection', *Science of the Total Environment* **715**, 136991.
- Masood, A., Kafeel, A. & Shamshad, A. (2017), 'Urban roadside monitoring, modeling and mapping of air pollution', *Applied Journal of Environmental Engineering Science* **3**(2), 3–2.
- Mateichyk, V., Tsuman, M., Weigang, G., Kozodoy, D., Sansyzbajeva, Z. & Grytsuk, Y. (2020), The information and analytical system for monitoring roadside pollution by traffic flows, *in* 'ICTE in Transportation and Logistics 2019', Springer, pp. 352–359.
- Maxwell, J. A. (2012), *Qualitative research design: An interactive approach*, Vol. 41, Sage publications.

- McCaslin, M. L. (2008), 'Pragmatism', *The Sage encyclopedia of qualitative research methods* **1**, 671–675.
- McDonough, J. & McDonough, S. (2014), *Research methods for English language teachers*, Routledge.
- Meng, M.-R., Cao, S.-J., Kumar, P., Tang, X. & Feng, Z. (2021), 'Spatial distribution characteristics of pm2. 5 concentration around residential buildings in urban traffic-intensive areas: From the perspectives of health and safety', *Safety Science* **141**, 105318.
- Mengara Mengara, A. G., Park, E., Jang, J. & Yoo, Y. (2022), 'Attention-based distributed deep learning model for air quality forecasting', *Sustainability* **14**(6), 3269.
- Mertens, D. M. (2014), *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*, Sage publications.
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. & Khudanpur, S. (2010), 'Recurrent neural network based language model', *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010* **2**, 1045–1048.
- Mishra, R. K., Shukla, A., Parida, M. & Pandey, G. (2016), 'Urban roadside monitoring and prediction of co, no2 and so2 dispersion from on-road vehicles in megacity delhi', *Transportation Research Part D: Transport and Environment* **46**, 157–165.
- Morgan, D. (2014), 'Pragmatism as a paradigm for social research', *Qualitative Inquiry* **20**, 1045– 1053.
- Morin, K. & Davis, J. L. (2017), 'Cross-validation: What is it and how is it used in regression?', *Communications in Statistics-Theory and Methods* **46**(11), 5238–5251.
- Morley, S. K., Brito, T. V. & Welling, D. T. (2018), 'Measures of model performance based on the log accuracy ratio', *Space Weather* **16**(1), 69–88.

- Mukumbang, F. C. (2023), 'Retroductive theorizing: a contribution of critical realism to mixed methods research', *Journal of Mixed Methods Research* **17**(1), 93–114.
- N. Genikomsakis, K., Galatoulas, N.-F., I. Dallas, P., Candanedo Ibarra, L. M., Margaritis, D. & S. Ioakimidis, C. (2018), 'Development and on-field testing of low-cost portable system for monitoring pm_{2.5} concentrations', *Sensors* **18**(4), 1056.
- Newman, I., Benz, C. R. & Ridenour, C. S. (1998), *Qualitative-quantitative research methodology: Exploring the interactive continuum*, SIU Press.
- NHMRC, A. et al. (2007), 'National statement on ethical conduct in human research', *Canberra, NHMRC*.
- Noble, W. S. et al. (2004), 'Support vector machine applications in computational biology', *Kernel methods in computational biology* **71**, 92.
- Odat, S. (2009), 'Diurnal and seasonal variation of air pollution at al-hashimeya town, Jordan', *Earth Environ Sci* **2**, 1–6.
- O'leary, Z. (2004), *The essential guide to doing research*, Sage.
- ONS (2018), '2017 uk greenhouse gas emissions, provisional figures', *Statistical Release: National Statistics*.
- ONS (2021), 'Labour market in the regions of the uk: October 2021', Available from: <https://www.gov.uk/government/statistics/labour-market-in-the-regions-of-the-uk-october-2021> [Accessed: 15-12-2021].
- Pal, S., Ghosh, A. & Sethi, V. (2018), Vehicle air pollution monitoring using iots, in 'Proceedings of the 16th ACM conference on embedded networked sensor systems', pp. 400–401.
- Pandey, G., Zhang, B. & Jian, L. (2013), 'Predicting submicron air pollution indicators: a machine learning approach', *Environmental Science: Processes & Impacts* **15**(5), 996–1005.

- Pandya, S., Gadekallu, T. R., Maddikunta, P. K. R. & Sharma, R. (2022), 'A study of the impacts of air pollution on the agricultural community and yield crops (indian context)', *Sustainability* **14**(20), 13098.
- Pascal, M., Corso, M., Chanel, O., Declercq, C., Badaloni, C., Cesaroni, G., Henschel, S., Meister, K., Haluza, D., Martin-Olmedo, P. et al. (2013), 'Assessing the public health impacts of urban air pollution in 25 european cities: results of the aphekom project', *Science of the Total Environment* **449**, 390–400.
- Pasquier, A. & André, M. (2017), 'Considering criteria related to spatial variabilities for the assessment of air pollution from traffic', *Transportation research procedia* **25**, 3354–3369.
- Patton, M. Q. (2005), 'Qualitative research', *Encyclopedia of statistics in behavioral science* .
- Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J. & Tapper, N. J. (2011), 'Quantifying the influence of local meteorology on air quality using generalized additive models', *Atmospheric Environment* **45**(6), 1328–1336.
- Peirce, C. S. (1931), 'Collected papers of cs peirce.(eds.) c. hartshorne, et al'.
- Peirce, C. S. (1997), *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*, SUNY Press.
- Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., Li, K., Dick, R. P., Lv, Q., Hannigan, M. et al. (2014), 'The next generation of low-cost personal air quality sensors for quantitative exposure monitoring', *Atmospheric Measurement Techniques* **7**(10), 3325–3336.
- Pietrobon, A. & Dai, Y. (2012), 'Branding for start-ups: A case study of spotify', p. 5.
URL: <http://www.diva-portal.org/smash/get/diva2:536929/FULLTEXT01.pdf>

Pohjola, M., Pirjola, L., Karppinen, A., Härkönen, J., Korhonen, H., Hussein, T., Ketzel, M. & Kukkonen, J. (2007), ‘Evaluation and modelling of the size fractionated aerosol particle number concentration measurements nearby a major road in helsinki—part i: Modelling results within the lipika project’, *Atmospheric Chemistry and Physics* **7**(15), 4065–4080.

Public Health England (2019), ‘Review of interventions to improve outdoor air quality and public health’.

URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/2019-2018572.pdf

Rai, A. C. & Kumar, P. (2017), ‘Summary of air quality sensors and recommendations for application’, *iSCAPE - Improving the Smart Control of Air Pollution in Europe Ref. Ares* (689954), 33–54.

URL: https://www.iscapeproject.eu/wp-content/uploads/2017/09/iSCAPE_D1.5_Summary-of-air-quality-sensors-and-recommendations-for-application.pdf

Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Di Sabatino, S., Ratti, C., Yasar, A. & Rickerby, D. (2017), ‘End-user perspective of low-cost sensors for outdoor air pollution monitoring’, *Science of The Total Environment* **607**, 691–705.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. & Barnes, P. (2020), Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, *in* ‘Proceedings of the 2020 conference on fairness, accountability, and transparency’, pp. 33–44.

Ramírez-Moreno, M. A., Keshtkar, S., Padilla-Reyes, D. A., Ramos-López, E., García-Martínez, M., Hernández-Luna, M. C., Mogro, A. E., Mählknecht, J., Huertas, J. I., Peimbert-García, R. E. et al. (2021), ‘Sensors for sustainable smart cities: A review’, *Applied Sciences* **11**(17), 8198.

Rana, A., Rawat, A. S., Afifi, A., Singh, R., Rashid, M., Gehlot, A., Akram, S. V. & Alsham-

- rani, S. S. (2022), ‘A long-range internet of things-based advanced vehicle pollution monitoring system with node authentication and blockchain’, *Applied Sciences* **12**(15), 7547.
- Rand, A. (1990), *Introduction to Objectivist Epistemology: Expanded Second Edition*, Penguin.
- Raskin, J. D. (2002), ‘Constructivism in psychology: Personal construct psychology, radical constructivism, and social constructionism’, *American communication journal* **5**(3), 1–25.
- RCP (2016), ‘Every breath we take: the lifelong impact of air pollution’, *Royal college of paediatrics and child health* .
- Reddy, Y. P., Parameswaran, T. & Sathiyaraj, R. (2021), A smart environment monitoring framework using big data and iot, in ‘2021 IEEE Mysore Sub Section International Conference (MysuruCon)’, IEEE, pp. 399–404.
- Rotaru, K., Churilov, L. & Flitman, A. (2014), ‘Can critical realism enable a journey from description to understanding in operations and supply chain management?’, *Supply Chain Management: An International Journal* **19**(2), 117–125.
- Rowland, G. (1995), ‘Archetypes of systems design’, *Systems practice* **8**(3), 277–288.
- Rowley, J. (2002), ‘Using case studies in research’, *Management research news* **25**(1), 16–27.
- Rushikesh, R. & Sivappagari, C. M. R. (2015), Development of iot based vehicular pollution monitoring system, in ‘2015 International Conference on Green Computing and Internet of Things (ICGCIoT)’, IEEE, pp. 779–783.
- Sadik, S., Ali, A., Ahmad, F. & Suguri, H. (2006), Using honey bee teamwork strategy in software agents, in ‘2006 10th International Conference on Computer Supported Cooperative Work in Design’, IEEE, pp. 1–6.

- Saide, P. E., Mena-Carrasco, M., Tolvett, S., Hernandez, P. & Carmichael, G. R. (2016), 'Air quality forecasting for winter-time pm2. 5 episodes occurring in multiple cities in central and southern chile', *Journal of Geophysical Research: Atmospheres* **121**(1), 558–575.
- Salomon, G. (1991), 'Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research', *Educational researcher* **20**(6), 10–18.
- Sayegh, A., Tate, J. E. & Ropkins, K. (2016), 'Understanding how roadside concentrations of nox are influenced by the background levels, traffic density, and meteorological conditions using boosted regression trees', *Atmospheric Environment* **127**, 163–175.
- Schenker, J. D. & Rumrill Jr, P. D. (2004), 'Causal-comparative research designs', *Journal of vocational rehabilitation* **21**(3), 117–121.
- Schmidhuber, J. (2015), 'Deep learning in neural networks: An overview', *Neural networks* **61**, 85–117.
- Schwandt, T. A. (1997), *Qualitative inquiry: A dictionary of terms.*, Sage Publications, Inc.
- Scotland, J. (2012), 'Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms.', *English language teaching* **5**(9), 9–16.
- Scott, D. & Usher, R. (2010), *Researching education: Data, methods and theory in educational enquiry*, Bloomsbury Publishing.
- Sergeev, A. & Del Balso, M. (2018), 'Horovod: fast and easy distributed deep learning in tensorflow', *arXiv preprint arXiv:1802.05799* .
- Shakhov, V. & Sokolova, O. (2019), Towards air pollution detection with internet of vehicles, in '2019 15th International Asian School-Seminar Optimization Problems of Complex Systems (OPCS)', IEEE, pp. 183–186.

- Shen, Q., Wu, Y., Jiang, Y., Zeng, W., Alexis, K., Vianova, A. & Qu, H. (2020), Visual interpretation of recurrent neural network on multi-dimensional time-series forecast, *in* ‘2020 IEEE Pacific Visualization Symposium (PacificVis)’, IEEE, pp. 61–70.
- Shrestha, A. & Mahmood, A. (2019), ‘Review of deep learning algorithms and architectures’, *IEEE access* **7**, 53040–53065.
- Shrivastava, A., Gupta, V. B. et al. (2011), ‘Methods for the determination of limit of detection and limit of quantitation of the analytical methods’, *Chronicles of young scientists* **2**(1), 21.
- Sloman, L., Hopkinson, L. & Taylor, I. (2017), ‘The impact of road projects in england’, *Campaign to Protect Rural England, London* .
- Sober, E. (2001), *Core questions in philosophy: A text with readings*, Prentice Hall Upper Saddle River, NJ.
- Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W. & Vlahavas, I. (2012), ‘Multi-label classification methods for multi-target regression’, *arXiv preprint arXiv:1211.6581* pp. 1159–1168.
- Sridhar, B., Mounika, R., Babu, P. N., Kumar, Y. P. & Raja, S. R. (2022), Design a low-cost air pollution monitoring iot system, *in* ‘Advances in Micro-Electronics, Embedded Systems and IoT: Proceedings of Sixth International Conference on Microelectronics, Electromagnetics and Telecommunications (ICMEET 2021), Volume 1’, Springer, pp. 191–202.
- Statswales (2020), ‘Summary statistics for wales, by region: 2020’, Available from: <https://gov.wales/sites/default/files/statistics-and-research/2020-05/summary-statistics-regions-wales-2020-629.pdf> [Accessed: 15-12-2021].
- Stocker, J., Heist, D., Hood, C., Isakov, V., Carruthers, D., Perry, S., Snyder, M., Venkatram, A. & Arunachalam, S. (2019), Road source model intercomparison study using new and

- existing datasets, *in* ‘15th International Conference on Harmonisation, Madrid, Spain. http://www.harmo.org/Conferences/Proceedings/_Madrid/publishedSections/H15-78.pdf. Accessed’, Vol. 3.
- Straus, A. & Corbin, J. (1990), ‘Basics of qualitative research: Grounded theory procedures and techniques’.
- Suleiman, A., Tight, M. & Quinn, A. (2016), ‘Hybrid neural networks and boosted regression tree models for predicting roadside particulate matter’, *Environmental Modeling & Assessment* **21**, 731–750.
- Suleiman, A., Tight, M. & Quinn, A. (2019), ‘Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (pm10 and pm2.5)’, *Atmospheric Pollution Research* **10**(1), 134–144.
- Sun, H., Fung, J. C., Chen, Y., Chen, W., Li, Z., Huang, Y., Lin, C., Hu, M. & Lu, X. (2021), ‘Improvement of pm2.5 and o3 forecasting by integration of 3d numerical simulation with deep learning techniques’, *Sustainable Cities and Society* p. 103372.
- Tarek, M. F. B., Asaduzzaman, M. & Patwary, M. (2018), Spatio-temporal analysis of large air pollution data, *in* ‘2018 10th International Conference on Electrical and Computer Engineering (ICECE)’, IEEE, pp. 221–224.
- Taylor, S. J. & Letham, B. (2018), ‘Forecasting at scale’, *The American Statistician* **72**(1), 37–45.
- Teddlie, C. & Tashakkori, A. (2003), *Handbook of mixed methods in social & behavioral research*, Sage.
- Tellis, W. M. (1997), ‘Introduction to case study’, *The qualitative report* **3**(2), 1–14.
- TFL (2019), ‘Travel in london: Understanding our diverse communities 2019’.

- Thornhill, A., Saunders, M. & Lewis, P. (2009), *Research methods for business students*, Prentice Hall: London.
- Toma, C., Alexandru, A., Popa, M. & Zamfiroiu, A. (2019), ‘Iot solution for smart cities’ pollution monitoring and the security challenges’, *Sensors* **19**(15), 3401.
- UKAIR (2018), ‘Background mapping data for local authorities’.
URL: <https://uk-air.defra.gov.uk/data/laqm-background-home>
- Vohra, K., Marais, E. A., Suckra, S., Kramer, L., Bloss, W. J., Sahu, R., Gaur, A., Tripathi, S. N., Van Damme, M., Clarisse, L. et al. (2021), ‘Long-term trends in air quality in major cities in the uk and india: A view from space’, *Atmospheric Chemistry and Physics* **21**(8), 6275–6296.
- Wachter, S. (2018), ‘Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the gdpr’, *Computer law & security review* **34**(3), 436–449.
- Walliman, N. (2017), *Research methods: The basics*, Routledge.
- Wang, D., Wang, H.-W., Li, C., Lu, K.-F., Peng, Z.-R., Zhao, J., Fu, Q. & Pan, J. (2020), ‘Roadside air quality forecasting in shanghai with a novel sequence-to-sequence model’, *International Journal of Environmental Research and Public Health* **17**(24), 9471.
- Wang, L. & Huang, Y. (2017), ‘Mobile atmospheric sensing’, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-3/W2**, 217–221.
URL: <https://isprs-archives.copernicus.org/articles/XLII-3-W2/217/2017/>
- Wang, S., Feng, X., Zeng, X., Ma, Y. & Shang, K. (2009), ‘A study on variations of concentrations of particulate matter with different sizes in lanzhou, china’, *Atmospheric Environment* **43**(17), 2823–2828.

- Wang, Y., Yao, H. & Zhao, S. (2016), 'Auto-encoder based dimensionality reduction', *Neurocomputing* **184**, 232–242.
- Ward, E. (2019), *Use of remote sensing and in situ observations of the atmosphere in chemical transport models*, Chalmers Tekniska Hogskola (Sweden).
- Werbos, P. (1988), 'Generalization of backpropagation with application to a recurrent gas market model', *Neural Networks* **1**, 339–356.
- Wilkins, A. S. (2018), 'To lag or not to lag?: Re-evaluating the use of lagged dependent variables in regression analysis', *Political Science Research and Methods* **6**(2), 393–411.
- Williams, C. (2007), 'Research methods', *Journal of Business & Economics Research (JBER)* **5**(3).
- Winter, G. (2000), 'A comparative discussion of the notion of 'validity' in qualitative and quantitative research', *The qualitative report* **4**(3), 1–14.
- World Bank (2022), 'The global health cost of pm2.5 air pollution: A case for action beyond 2021'.
URL: <https://openknowledge.worldbank.org/handle/10986/36501>
- Wu, C.-l., Song, R.-f., Peng, Z.-r. et al. (2022), 'Prediction of air pollutants on roadside of the elevated roads with combination of pollutants periodicity and deep learning method', *Building and Environment* **207**, 108436.
- Wu, Shaowei, Ni, Yang, Li, Hongyu, Pan, Lu, Yang, Di, Baccarelli, Andrea A, Deng, F., Chen, Y., Shima, M. & Guo, X. (2016), 'Short-term exposure to high ambient air pollution increases airway inflammation and respiratory symptoms in chronic obstructive pulmonary disease patients in beijing, china', *Environment international* **94**, 76–82.
- Xu, J., Dong, Y. & Yan, M. (2020), 'A model for estimating passenger-car carbon emissions that accounts for uphill, downhill and flat roads', *Sustainability* **12**(5), 2028.

- Yang, Y., Zheng, Z., Bian, K., Song, L. & Han, Z. (2018), Sensor deployment recommendation for 3d fine-grained air quality monitoring using semi-supervised learning, *in* ‘2018 IEEE International Conference on Communications (ICC)’, IEEE, pp. 1–6.
- Yanow, D. & Schwartz-Shea, P. (2015), *Interpretation and method: Empirical research methods and the interpretive turn*, Routledge.
- Yin, R. (1984), ‘Case study research: design and methods (beverly hills, ca, sage)’.
- Yli-Tuomi, T., Aarnio, P., Pirjola, L., Mäkelä, T., Hillamo, R. & Jantunen, M. (2005), ‘Emissions of fine particles, nox, and co from on-road vehicles in finland’, *Atmospheric Environment* **39**(35), 6696–6706.
- Yu, Y., Si, X., Hu, C. & Zhang, J. (2019), ‘A review of recurrent neural networks: Lstm cells and network architectures’, *Neural computation* **31**(7), 1235–1270.
- Yuehong, Y., Zeng, Y., Chen, X. & Fan, Y. (2016), ‘The internet of things in healthcare: An overview’, *Journal of Industrial Information Integration* **1**, 3–13.
- Zastrow, M. (2015), ‘Data visualization: Science on the map’, *Nature* **519**(7541), 119–120.
- Zeng, Y., Cao, Y., Qiao, X., Seyler, B. C. & Tang, Y. (2019), ‘Air pollution reduction in china: Recent success but great challenge for the future’, *Science of the Total Environment* **663**, 329–337.
- Zhai, Z., Tu, R., Xu, J., Wang, A. & Hatzopoulou, M. (2020), ‘Capturing the variability in instantaneous vehicle emissions based on field test data’, *Atmosphere* **11**(7), 765.
- Zhang, K. & Batterman, S. (2013), ‘Air pollution and health risks due to vehicle traffic’, *Science of the total Environment* **450**, 307–316.
- Zheng, Y., Liu, F. & Hsieh, H.-P. (2013), U-air: When urban air quality inference meets big data, *in* ‘Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining’, pp. 1436–1444.

Zucker, D. M. (2001), 'Using case study methodology in nursing research', *The Qualitative Report* **6**(2), 1–13.