Spatio-Temporal Estimation of $PM_{2.5}$ With Transfer Learning

Jamie Welsh

Master of Science

Artificial Intelligence

School of Informatics

University of Edinburgh

2023

Abstract

The negative effects of air pollution are extensive, spanning from respiratory problems to harm to the environment and ecosystems. Consequently, the monitoring and forecasting of air pollution have become critically important. The Centre for Speckled Computing at the University of Edinburgh has developed *AirSpeck*, an air quality monitoring sensor that can either be worn by an individual or set up in a fixed location. This sensor is capable of recording the concentration of various pollutants in the air, including $PM_{2.5}$. In recent times, numerous models have been developed to estimate $PM_{2.5}$ levels spatially, temporally and spatio-temporally. In this study, a framework will be created to train spatial and spatio-temporal models for estimating $PM_{2.5}$ concentrations across three datasets, comparing various machine learning techniques. The best-performing spatial model was the feed-forward neural network trained on personal AirSpeck data from the INHALE study, which yielded a MAPE of 42.27%. The best performing spatio-temporal model was trained to predict INHALE stationary sensor AirSpeck data one hour into the future and combined spatial predictions of a feed-forward neural network with a long short term memory (LSTM) network. This model yielded a MAPE of 42.08%. Furthermore, a transfer learning approach to spatial $PM_{2.5}$ estimation was devised to improve predictive performance in scenarios with limited training data. This involved pre-training a feed-forward neural network architecture with data from another geographic location and then fine-tuning this model on the target dataset. For subsets of the INHALE and DAPHNE stationary datasets, this method results in reductions in MAPE from 53.72% to 48.49% and from 61.82% to 54.86%, respectively, relative to the second-best model. Additionally, the transfer learning method outperforms all other spatial estimation models on the Leon and Guadalajara personal sensor datasets, yielding MAPEs of 20.84% for Leon and 19.69% for Guadalajara.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jamie Welsh*)

Acknowledgements

I would like to thank Prof. Arvind for his guidance and feedback throughout the course of this project. I would also like to thank my family and friends for their unwavering encouragement and support.

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Objectives 2

1.3 Summary of Models and Results 2

1.4 Novelty 3

1.5 Report Structure 3

2 Background **4**

2.1 Particulate Matter 4

2.2 Data Sources 4

2.2.1 DAPHNE 5

2.2.2 INHALE 5

2.2.3 Leon 6

2.2.4 Guadalajara 6

2.2.5 Air Quality Reference Data 7

2.2.6 Meteorological Data 7

2.2.7 Road and Land Data 7

2.3 Machine Learning Algorithms 8

2.3.1 Random Forests 8

2.3.2 Extra Trees 8

2.3.3 Gradient Boosting Regression 8

2.3.4 Artificial Neural Networks 8

2.3.5 Recurrent Neural Networks 9

3 Related Work **10**

4 Methodology **13**

4.1 Data Preprocessing and Feature Selection 13

4.2 Stationary Sensor Model 18
4.2.1 Transfer Learning 19
4.3 Personal Sensor Model 20
4.3.1 Transfer Learning 21
4.4 Evaluation Metrics 21

5 Experiments and Results **23**
5.1 Stationary Sensor Models 23
5.1.1 INHALE 23
5.1.2 DAPHNE 27
5.2 Personal Sensor Models 30
5.2.1 INHALE 30
5.2.2 Transfer Learning 34
5.3 Summary of Results 36

6 Conclusions and Future Work **39**

Bibliography **41**

A Stationary Model Results - INHALE **47**
A.0.1 Transfer Learning 49

B Stationary Model Results - DAPHNE **50**
B.0.1 Transfer Learning 52

C Personal Model Results - INHALE **53**

D Personal Model Results - Guadalajara **55**

E Personal Model Results - Leon **56**

Chapter 1

Introduction

1.1 Motivation

Monitoring air pollution levels is of critical importance due to its negative health [1, 2, 3], environmental [4, 5], and socio-economic impacts [6, 7]. In 2019, 99% of people worldwide were living in places where World Health Organisation guidelines on air pollution were not met [8]. Moreover, in that same year, 4.2 million premature deaths were linked to ambient air pollution. Given the significance of studying air pollution levels, considerable efforts have been dedicated to developing efficient air quality monitoring sensors [9, 10, 11]. "The Centre for Speckled Computing" at the University of Edinburgh has developed a compact air quality monitoring sensor named AirSpeck that can either be worn on the body (AirSpeck personal) or deployed in a fixed location (AirSpeck stationary) [12]. The AirSpeck personal sensor captures spatio-temporal data at 30-second intervals, while the AirSpeck stationary sensor records time series data with time intervals ranging from five to 30 minutes depending on seasonal variations in the number of daylight hours for charging the solar cells. These sensors record the concentration of airborne particulate matter classified by diameter: PM_1, $PM_{2.5}$, and PM_{10}, as well as temperature and humidity. They also measure the concentration of nitrogen dioxide (NO_2) and ozone (O_3). These sensors have been utilised to collect data in various global locations, including London (United Kingdom), Delhi (India), Leon and Guadalajara (both Mexico). Previous studies utilising AirSpeck pollution data have sought to estimate the concentration of $PM_{2.5}$ spatially (i.e., predicting the concentration in one location based on the concentration in another location and other relevant factors) and temporally (i.e., predicting the concentration at a specific location based on time series data). Some studies have even estimated the $PM_{2.5}$ concentration

spatio-temporally based on the data collected by both the stationary and personal AirSpeck sensors. A model capable of predicting air quality for a specific location at a future time has numerous downstream benefits such as helping people plan walking and cycling routes that minimise pollution and informing environmental policy-making. In another MSc Dissertation from the School of Informatics, Estrada (2023) utilised the spatial INHALE personal sensor model from this study for route planning to minimize $PM_{2.5}$ exposure [13]. All previous contributions have produced models which are trained on data from a certain city (e.g. London) and predict exposure in the same city. This has led to strong predictive performance, but these models struggle to extend to new cities due to disparate $PM_{2.5}$ distributions [14]. It would be useful to produce a general method of spatial $PM_{2.5}$ estimation which can be applied to cities with little air quality data. In this work, a transfer learning approach will be devised to leverage information from larger datasets to aid predictive performance when there is a smaller quantity of air pollution data. This transfer learning framework will be used to train and evaluate spatial models on four different datasets.

1.2 Objectives

Robust estimation of particulate matter concentration is crucial. An accurate estimation model to predict personal $PM_{2.5}$ exposure in urban environments will generate many downstream benefits, including planning mitigation strategies that may have profound effects on public health and urban planning. One objective of this study is to create a robust framework for training spatio-temporal $PM_{2.5}$ estimation models. This framework will be used to train spatio-temporal models on various datasets. A further objective of this work is to develop a transfer learning approach to spatial $PM_{2.5}$ estimation whereby knowledge is leveraged from one model to another to improve spatial prediction when training estimation models with limited data.

1.3 Summary of Models and Results

Spatial estimation models have been trained on the INHALE stationary sensor, DAPHNE stationary sensor, and INHALE personal sensor datasets in order to predict the $PM_{2.5}$ concentration recorded by these sensors. For each of these datasets, the feed-forward neural network achieved the lowest MAPE, yielding values of 49.49%, 56.14%, and 42.27%, respectively. The spatio-temporal models for the INHALE and DAPHNE

stationary datasets integrate their spatial interpolations with a single-layer LSTM. However, the INHALE personal sensor model combines a two-layer stacked LSTM with its spatial interpolations. These optimised configurations result in MAPEs of 42.08%, 54.27%, and 43.04% for each dataset, respectively. For subsets of the INHALE and DAPHNE stationary sensor datasets, models trained using transfer learning techniques outperform all other algorithms, with MAPEs of 48.49% and 54.86%. Furthermore, for the Leon and Guadalajara personal sensor datasets, the transfer learning method yields MAPEs of 20.84% and 19.69%, respectively, surpassing the performance of the other machine learning techniques.

1.4 Novelty

This project introduces a deep learning methodology for enhanced spatio-temporal estimation of $PM_{2.5}$ as well as a transfer learning framework for training spatial estimation models in scenarios with limited data. The explicit novel contributions of this work are outlined below.

- Application of deep learning to both spatial and temporal modeling of $PM_{2.5}$ and amalgamating this into a unified spatio-temporal predictive model, trained on AirSpeck pollution data.
- Application of transfer learning to spatial $PM_{2.5}$ estimation in locations with limited data.
- Introduction of novel features based on proximity to various types of roads.

1.5 Report Structure

The data sources for this study along with the different machine learning algorithms and deep learning models are outlined in the ‘Background’ section. Previous approaches to $PM_{2.5}$ estimation are reviewed in the ‘Related Work’ section. In the ‘Methodology’ section, the data pre-processing steps as well as the implementation of the stationary and personal sensor models will be explained. Following this, the results of all the trained models will be presented, discussed and summarised in the ‘Experiments and Results’ section. In the ‘Conclusions and Future Work’ section, the findings of the research will be examined and the scope of future work discussed.

Chapter 2

Background

2.1 Particulate Matter

Particulate Matter (PM) is a complex mixture of extremely small solid and liquid particles suspended in the air. Some of these particles are emitted directly from sources such as vehicle exhausts and industrial processes while others are formed via chemical reactions in the air. $PM_{2.5}$ refers to particles that have an aerodynamic diameter of less than $2.5\mu m$ (which is approximately 3% of the diameter of a human hair). The composition of $PM_{2.5}$ can vary, a study into its constituents found that these often include sulfates, nitrates, ammonium and carbon [15]. Due to their minute size, these particles can remain suspended in the air for long periods and as a result they can pose significant health risks, including respiratory and cardiovascular diseases [16, 17].

2.2 Data Sources

The air quality data used in this study was obtained by the AirSpeck personal and AirSpeck stationary sensors from various studies conducted in different cities worldwide [12]. These sensors use an optical particle counter to separate particles into 16 bins. From these bins, the concentrations of PM_1, $PM_{2.5}$ and PM_{10} can be calculated. The sensors also record temperature and humidity. Moreover, every measurement comes with a corresponding timestamp and geographic coordinates. The AirSpeck personal sensor records data at intervals of 30 seconds and the AirSpeck stationary sensor records data at anywhere from five-minute to 30-minute intervals. In this study, $PM_{2.5}$ data obtained in various deployments of the AirSpeck sensors will be used as either the ground truth for training prediction models or as reference data to aid predictive

performance.

2.2.1 DAPHNE

The goal of this study was to examine the health impact of exposure to $PM_{2.5}$ in Delhi, India. There were two big cohorts of subjects, asthmatic adolescents and pregnant women. The asthmatic adolescents wore the AirSpeck personal sensor for 48 hours in three cycles, while the pregnant mothers wore the AirSpeck personal sensor for a period of 48 hours in each trimester of the pregnancy [18]. Stationary AirSpeck sensors were deployed in the home, community and schools of the subjects involved in the study. Data collection for this study began in July 2018 and concluded in March 2020. In this study, only the air pollution data obtained by the AirSpeck stationary sensors will be used.

2.2.2 INHALE

The goal of this study was to analyse the health effects of air pollution on subjects in London. Subjects wore the mobile AirSpeck personal sensor for a period of two weeks and there were also a number of stationary sensors deployed in various locations around London. This study is ongoing, the data used in this work was collected between February 2021 and June 2023 [19]. In this work, the data from both the stationary and personal sensors will be used to train spatial and spatio-temporal models. In Figure 2.1., the geographic locations of personal sensor readings for one of the subjects over a 24 hour period are plotted alongside the locations of all of the stationary AirSpeck sensor locations.

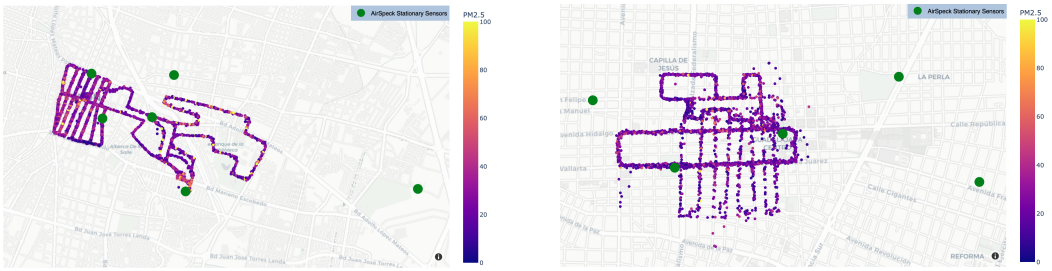
(a) AirSpeck personal sensor readings for subject INH002 over a 24-hour period

(b) INHALE stationary sensor locations

Figure 2.1: INHALE personal sensor data and stationary sensor locations

2.2.3 Leon

In this study, the AirSpeck personal sensor collected data for one hour a day, from 9am to 10am. The data collection period spanned 31 weekdays between the 21^{st} of October 2019 and the 2^{nd} of December 2019. Each day, the subjects followed the same route. The GPS locations of the sensors throughout the data collection are plotted in Figure 2.2a. Furthermore, six stationary AirSpeck sensors collected data in fixed locations in Leon between the 19^{th} of December 2019 and the 17^{th} of February 2020. The locations of these sensors are plotted in Figure 2.2a. The quantity of data collected in this deployment is far smaller than for the INHALE and DAPHNE deployments in London and Delhi, respectively.



(a) Leon personal sensor data is plotted along with the locations of the stationary AirSpeck sensors

(b) Guadalajara personal sensor data is plotted along with the locations of the stationary AirSpeck sensors

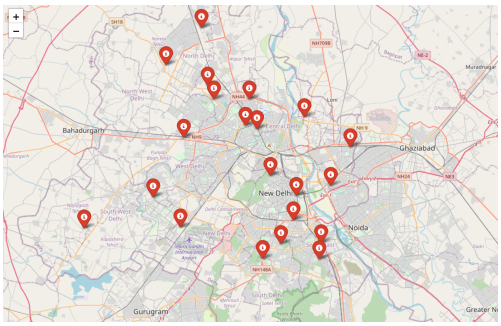
Figure 2.2: GPS locations of minute averaged readings as well as the locations of the stationary AirSpeck sensors are plotted for the Leon and Guadalajara datasets

2.2.4 Guadalajara

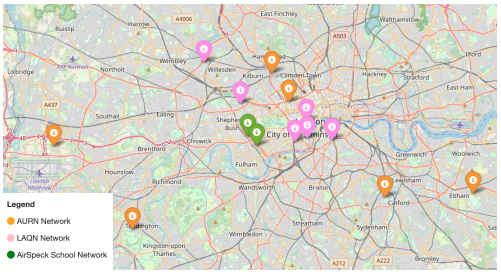
In Guadalajara, the AirSpeck personal sensors collected one hour of data per day, also between 9am and 10am. This was done between the 13^{th} of January 2020 and the 14^{th} of February 2020. The subjects followed a predetermined route while wearing the mobile AirSpeck sensor, and as a result all of the readings are taken outdoors. Furthermore, five AirSpeck stationary sensors were deployed throughout Guadalajara for a period of time between the 19^{th} of December 2019 and the 17^{th} of February 2020. The locations of these stationary sensors can be observed in Figure 2.2b.

2.2.5 Air Quality Reference Data

Reliable $PM_{2.5}$ data from air quality monitoring stations is crucial for a spatial estimation model. For the Leon and Guadalajara datasets, the AirSpeck stationary sensors which were deployed at the same time as the mobile sensors are used as reference data. The DAPHNE dataset utilises reference data collected from multiple monitoring stations, all of which are managed by the Central Pollution Control Board (CPCB) [20]. The locations of these CPCB monitoring sensors are plotted in Figure 2.3a. For the INHALE study, $PM_{2.5}$ reference data is obtained from three different sources: Automatic Urban and Rural Network [21], London Air Quality Network [22] and two AirSpeck stationary sensors which were deployed in and around the Kensington area of London.



(a) Locations of the CPCB monitoring stations in Delhi, India



(b) Locations of the London air quality reference sensors

Figure 2.3: Locations of air quality reference sensors in Delhi and London

2.2.6 Meteorological Data

Meteorological data for the relevant locations and times was obtained from the website *rp5.ru* [23]. This website contains data archives for several cities worldwide and this data includes various weather related fields which may be useful for estimation pollution levels such as air pressure, humidity, wind direction and wind speed.

2.2.7 Road and Land Data

Road and land usage data was extracted from the OpenStreetMap (OSM) platform for London, Delhi, Leon and Guadalajara [24]. This data is extracted in *geojson* format and contains details on roads, footways and land usage patterns within a specified region and will be used to incorporate several road type and land usage features into the spatial models.

2.3 Machine Learning Algorithms

2.3.1 Random Forests

Random Forests is a machine learning algorithm that involves the creation of multiple decision trees. Each decision tree is trained on a different sample of the data and for each tree, only a random subset of features is considered when deciding how to split the nodes. This randomness helps the model to be more robust and reduce the likelihood of overfitting. When making a prediction, the outputs of all the individual trees are averaged [25].

2.3.2 Extra Trees

Extremely Randomized Trees (also known as Extra Trees) is an ensemble learning method similar to Random Forests. When splitting attributes, a Random Forest algorithm will choose the optimal split point, while Extra Trees selects these split points completely at random. This additional randomness often leads to better model generalisation, making it a robust tool for complex machine learning tasks [26].

2.3.3 Gradient Boosting Regression

Gradient Boosting Decision Trees (GBDT) is an ensemble learning method that creates decision trees in a sequential fashion. Starting with a simple model it calculates the errors of this model and then builds a new decision tree specifically to predict these errors. This means that each subsequent tree in the sequence is learning from the mistakes of its predecessors. The final prediction is the sum of the predictions from all individual trees [27].

2.3.4 Artificial Neural Networks

Artificial Neural Networks (ANNs) often outperform tree-based models in capturing intricate, non-linear relationships within data sets due to the inherent ability of the architecture to model such complexity. ANNs are inspired by biological neural networks, incorporating layers of interconnected nodes that allow for the modeling of non-linear relationships in data that tree-based models may not be able to efficiently capture [28]. Feed-forward neural networks (FFNNs) are a type of ANN where the information flows in one direction from an input layer through one or more hidden layers to an output

layer. The nodes in each layer take as input the output from the previous layer and pass the information forward [29].

2.3.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of artificial neural network architecture designed to model patterns in sequential data [30]. Unlike feed-forward neural networks, RNNs take as input information from previous states in the sequence with this information being stored in a cell called the hidden state. However, conventional RNNs can struggle to learn long-term dependencies due to issues known as vanishing and exploding gradients. This limitation led to the development of more advanced types of RNNs, such as Long Short-Term Memory networks (LSTMs) [31] and Gated Recurrent Units (GRUs) [32] which are designed to better capture long-term dependencies by using gating mechanisms. These gates learn to make decisions on which information to retain and which information to forget thus ensuring that only the most relevant information is propagated through the network.

Chapter 3

Related Work

Khan (2021) utilises data obtained by the AirSpeck sensor in Leon, Guadalajara and Delhi in order to train spatio-temporal models for prediction in each location [14]. The best spatial model in all cases was an Extra Trees model, giving rise to mean absolute percentage errors of $31.69\%, 39.33\%$ and 11.02% for Leon, Guadalajara and Delhi, respectively. These spatial predictions are based on, amongst other factors, air pollution information from previous time steps. These spatial interpolations are used as input to an LSTM which is trained to predict the $PM_{2.5}$ level for the next time-step. This architecture leads to a MAPE of $27.01\%, 32.12\%$ and 5.17% for each dataset. Despite the effectiveness of these models, this approach fails to generalise well to new cities with a lack of input features cited as the main reason for this. This study includes meteorological features such as wind speed, wind direction and atmospheric pressure as well as some land use features.

Porchelvan (2021) also aims to estimate $PM_{2.5}$ spatio-temporally with AirSpeck sensor data [33]. Stationary and personal sensor data from the INHALE and DAPHNE studies is used to train separate predictive models for each dataset. In the stationary sensor spatial model, recorded pollution levels of the other AirSpeck sensor(s) are used as input features. This study includes air quality monitoring data for the city of London from the Automatic Urban and Rural Network (AURN) [21] as reference data to aid prediction. Our study will expand this approach by incorporating air quality monitoring data from three sources: AURN (Automatic Urban and Rural Network) [21] and London Air Quality Network (LAQN) [22] for London, and data from the Central Pollution Control Board (CPCB) [20] for Delhi.

Xhang (2023) [34] adopts a method similar to those described in both [14, 33] for spatio-temporal particulate matter estimation. This study includes weather, road type and

land use features as well as data from the three closest air quality monitoring stations. Spatial interpolation models are trained on both the mobile and stationary AirSpeck data from the INHALE study conducted in London. Among the techniques used, gradient boosting regression yielded the lowest mean absolute percentage error. The spatial predictions are fed through a temporal model to predict $PM_{2.5}$ spatio-temporally with a stacked LSTM model corresponding to the lowest MAPE in both cases. The stationary and personal sensor spatio-temporal models give rise to MAPEs of 38.09% and 73.48%, respectively. The road and land based features had minimal impact on the personal sensor model in this study, perhaps as the GPS latitude and longitude readings are averaged by hour thus capturing an average location of each subject every hour. In theory, a subject could walk along busy roads for the best part of an hour, yet the averaged GPS reading might pinpoint a spacious green area like a park. In light of this, spatial models will be trained on mobile AirSpeck data at the minute level granularity in this study.

Machine learning based approaches are common when it comes to spatial air pollution estimation. Baawain et al. (2014) train an artificial neural network to predict daily concentrations of several pollutants including PM_{10}, ozone and nitrogen dioxide [35]. They observed that factors such as temperature, wind direction and wind speed all play a significant role in determining PM_{10} concentration. Aditya et al. (2018) find that logistic regression and auto-regression can be used to detect air quality and predict future $PM_{2.5}$ concentration [36]. Alimissis et al. (2018) train a multiple linear regression model as well as a feed-forward ANN to estimate spatially the concentration of five different air pollutants including nitrogen dioxide and ozone [37]. This study finds that the neural network model significantly outperforms the linear interpolation model with regards to the mean absolute error and root mean squared error metrics. This improvement is attributed to the neural model's ability to comprehend the complex spatial variability of pollution levels. Mahalingam et al. (2019) use data from the CPCB to train a support vector machine and a neural network to predict the air quality index (AQI) [38]. It is found that the support vector machine predicts the AQI with a considerably higher accuracy than the neural network. Nevertheless, feed-forward neural networks stand out as a prominent and efficacious approach to spatial air quality estimation.

Huang et al. (2015) train a feed-forward neural network to forecast $PM_{2.5}$ concentration with input features including the forecast for weather on a given day, weather conditions the day before, the air quality index two days beforehand and the pollution levels in nearby cities the previous day [39]. Due to their innate ability to capture dependencies

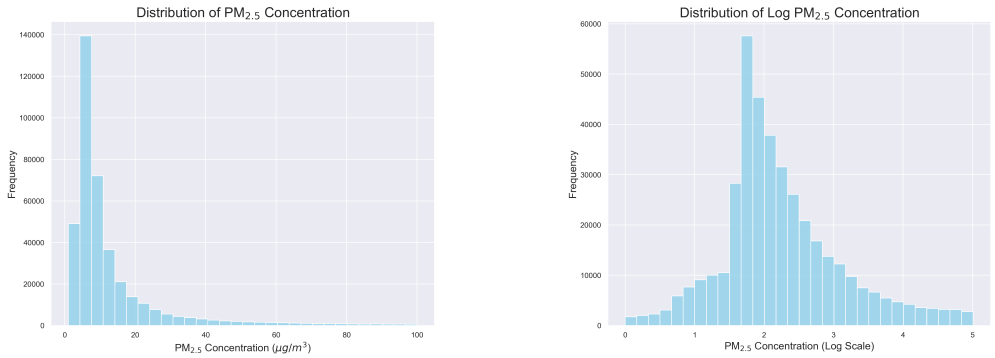
in sequential data, residual neural network architectures (which include LSTMs and GRUs) have been widely employed for temporal forecasting of air pollution. Tsai et al. (2018) train an LSTM network to forecast $PM_{2.5}$ concentration using air quality monitoring data from Taiwan [40]. Window sizes based on the last 3, 8, 24 and 72 hours $PM_{2.5}$ concentration are used to predict future values with a history size of 72 hours leading to the lowest mean absolute error. Intuitively, the mean absolute error is lower when predicting a shorter time into the future. Furthermore, the LSTM network is compared to a feed-forward ANN and the LSTM forecasts $PM_{2.5}$ concentration with a considerably lower mean absolute error. Tao et al. (2019) combine a one dimensional convolutional neural network with a bi-directional GRU to perform temporal forecasting [41]. This method exhibits improved temporal forecasting when compared to machine learning algorithms such as support vector regression and gradient boosting regression. Chang et al. (2020) presents an aggregated LSTM approach to temporal forecasting of air pollution up to eight hours into the future [42]. This LSTM model aggregates three separate LSTM models, each trained on different monitoring station data, into a single model. It evaluates with a considerably lower mean absolute error and root mean squared error than a support vector regression model as well as a gradient boosting decision trees model. It is evident that residual neural networks are a popular and powerful tool for temporal forecasting of air quality.

Transfer learning is a powerful technique in machine learning that leverages knowledge from one task and applies it to a different but related task. Models can benefit from pre-trained representations, which can significantly reduce the need for large labeled datasets. Transfer learning can help models to generalise better to unseen data as a result of the knowledge transfer. Numerous studies have shown the effectiveness of transfer learning in computer vision tasks [43, 44] and natural language processing tasks [45, 46]. The volume of training data is pivotal for regression tasks, and if there is insufficient data, the model can struggle to capture trends and relationships in the data [47]. Where there doesn't exist a sufficiently large dataset to train a neural network, fine-tuning, a form of transfer learning, is a potential solution [48]. This is when the weights of a neural model are pre-trained on some dataset for a similar task and then the pre-trained model is further trained, or fine-tuned, on the task specific dataset. The advantage of this approach is that it leverages the knowledge gained from the larger dataset to improve pattern recognition in the smaller dataset, thereby potentially improving the performance even with limited data.

Chapter 4

Methodology

4.1 Data Preprocessing and Feature Selection



(a) Histogram of $PM_{2.5}$ concentration for the INHALE personal sensor data

(b) Histogram of log $PM_{2.5}$ concentration for the INHALE personal sensor data

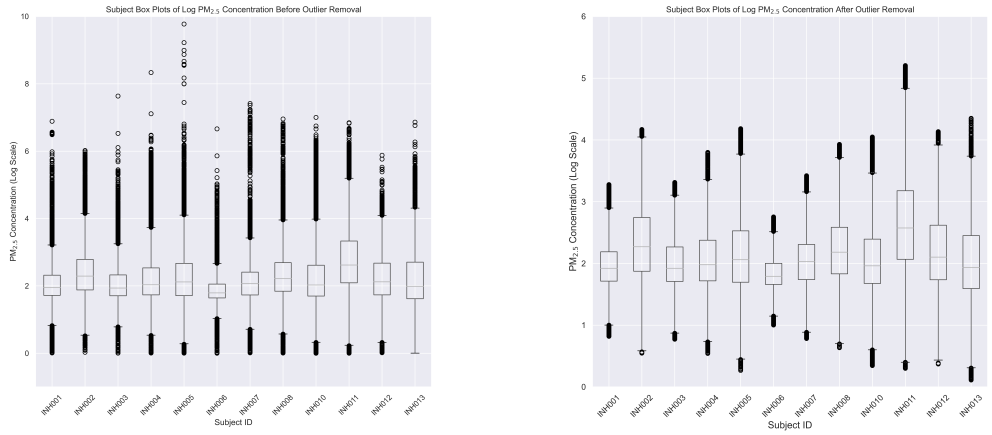
Figure 4.1: Contrasting the distribution of $PM_{2.5}$ concentration and its logarithmic transformation for the INHALE personal sensor dataset

As is common in data processing, outliers are removed from the stationary and personal AirSpeck data. Observe in Figure 4.1 that the distribution of $PM_{2.5}$ readings for the INHALE personal sensor dataset is positively skewed and has a rather long tail while taking the natural logarithm of the $PM_{2.5}$ concentrations leads to a far more symmetrical distribution. Similar distributions are observed for all of the datasets considered in this study. Hence, removing outliers on the logarithmic scale is reasonable as it ensures a more balanced removal of extreme values from both ends of the scale instead of mostly omitting high $PM_{2.5}$ values which are very important to be able to predict. Outlier

removal is done on a subject-by-subject basis using Tukey's fences method of outlier detection [49]. Data points where the following identity is not satisfied are dropped from the dataset,

$$Q1 - k.IQR < x < Q3 + k.IQR. \tag{4.1}$$

In this equation, Q1 and Q3 are the lower and upper quartiles, IQR is the inter-quartile range and k is some positive constant. In this study, $k = 1.5$ is chosen as this is a popular option for statistical analysis tasks [50]. Furthermore, any readings with GPS locations outside the specific city under investigation are excluded, as these readings are likely to be the result of reading errors. The logarithmic distributions of $PM_{2.5}$ concentration before and after outlier removal are plotted for 12 subjects of the INHALE personal sensor study in Figure 4.2. As desired, several outliers below the lower fence and above the upper fence are removed for each of the subjects in the plots. Removing outliers on the log scale leads to a more even balance of small and large outliers being removed. Data points with obviously erroneous latitude or longitude readings are also removed from the datasets.



(a) Box plots of log $PM_{2.5}$ concentration prior to outlier removal

(b) Box plots of log $PM_{2.5}$ concentration after outlier removal

Figure 4.2: Logarithmic Distribution of $PM_{2.5}$ concentration before and after removing outliers for mobile sensor data of 12 INHALE subjects

The $PM_{2.5}$ readings are averaged by the minute for the mobile sensor data and by the hour for the stationary sensor data. For the INHALE personal sensor data, measurements are recorded at all times and it is important to attempt to isolate the readings taken

outdoors. There is no perfectly accurate method of classifying data points as indoors or outdoors however, readings taken between 12 am and 6 am are excluded as these will likely correspond to instances where the subjects are indoors. Furthermore, consecutive readings where the GPS location is near identical and measured $PM_{2.5}$ concentration is below $2\mu g/m^3$ are eliminated from the dataset. In this case, it is a reasonable assumption that a subject is stationary indoors. This classification method is somewhat crude and certainly doesn't exclude all of the indoor readings from the dataset but will reduce their influence. For each dataset, reference sensor, meteorological, road type and land use features are created. The GPS latitude and longitude of a given sensor reading are used to calculate the closest three reference sensors (either from air quality monitoring stations e.g. AURN, LAQN, CPCB or stationary AirSpeck sensors) as well as the distances to these sensors. Due to the spherical nature of the earth, the distance calculations are carried out using the Haversine formula [51] which is defined as

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{4.2}$$

In this formula, d is the distance between the two points (along the surface of the earth), r is the radius of the earth (≈ 6371km), ϕ_1, ϕ_2 are the latitudes of the two points, and λ_1, λ_2 are the longitudes of the two points. The distances to the three closest reference sensors, as well as the $PM_{2.5}$ readings at these stations, are averaged to create two features: Average Reference $PM_{2.5}$ and Average Reference Distance. In Figure 4.3, the locations of the three closest reference sensors are plotted for two randomly selected static AirSpeck sensors in Delhi, with the connecting lines corresponding to the shortest distance between the AirSpeck sensor and the monitoring station in question, as calculated by the Haversine formula. For the left-most AirSpeck sensor, the third closest reference sensor is located quite far away, in fact, this distance is over 10 kilometres. Even though data from 20 separate reference stations is utilised in this study, it is often the case that even the closest monitoring stations can be located quite far away from the AirSpeck sensor due to the vast size of the metropolitan area of Delhi.

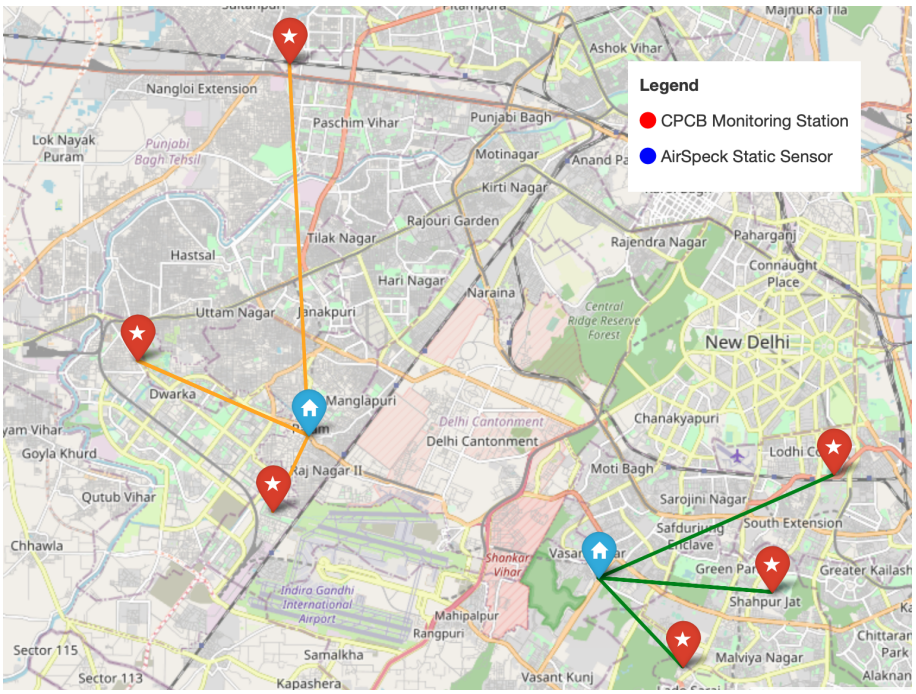
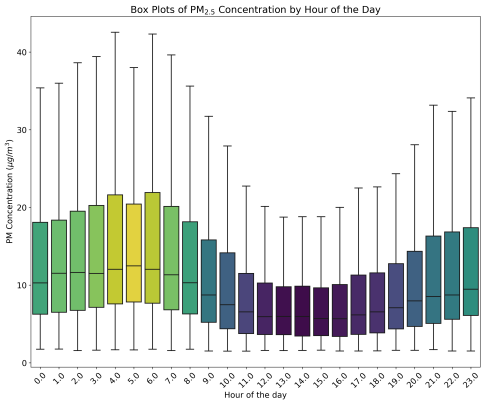
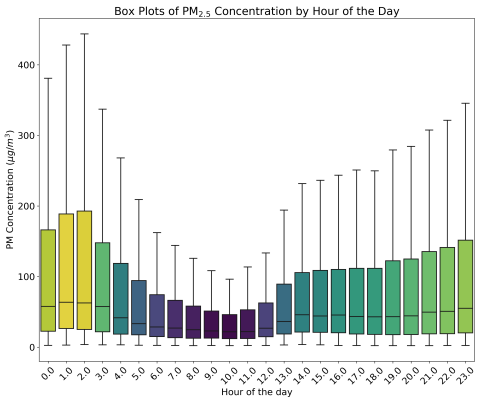


Figure 4.3: Illustration of the three closest reference sensor locations for two stationary AirSpeck sensors in Delhi



(a) Box plots of $PM_{2.5}$ concentration by hour of the day for the INHALE stationary sensors



(b) Box plots of $PM_{2.5}$ concentration by hour of the day for the DAPHNE stationary sensors

Figure 4.4: Box plots illustrating the distribution of $PM_{2.5}$ concentration by hour of the day in London and Delhi

In both London and Delhi, the distribution of $PM_{2.5}$ can vary significantly depending on the hour of the day, see Figure 4.4. There seems to be a cyclic component between the hour of the day and $PM_{2.5}$. As is common practice for cyclical features, the cosine

and sine transformations of the hour will be integrated into the baseline feature set [34, 52]. Moreover, the cosine and sine of the day of the week will be incorporated, considering the potential influence of weekly pollution patterns. For each city in the study, meteorological data is sourced from the website *www.rp5.ru* at 30-minute intervals [23]. For clarity, this data is averaged hourly. Using this data, the following features were added: air temperature, humidity, atmospheric pressure (raw and sea-level adjusted), wind speed, dew point temperature, horizontal visibility, and sine and cosine of wind direction (as it's cyclical). OpenStreetMap data is used to determine the nearest road type and represents it using one-hot encoding — a method where categorical variables are converted into a set of binary columns — and also calculates the distance to it. Land use is categorised into three separate designations: *green space*, *commercial*, and *industrial*. Distances to the nearest of each of these designations are included as features. OSM designates a *primary road* as a key highway linking towns, and a *secondary road* as a secondary highway [24]. Distances to the nearest motorway, primary road, and secondary road are calculated for each data point, as proximity may suggest increased pollution. The features are divided into three groups. The baseline feature set comprises grid average $PM_{2.5}$ concentration, average distance, and the sine and cosine of hour of the day and day of the week. The OSM and the weather features comprise the other two sets. To understand feature importance, five machine learning algorithms are applied to the INHALE mobile sensor data. The top performer was the Random Forest model. Observe in Figure 4.5 that weather features tend to elevate the MAPE, while OSM features produce varied outcomes. Even though the weather can greatly influence pollution [53], these features seem to confuse the models, possibly due to intricate inter-feature relationships. The optimal model is the Random Forest with only baseline features, though its MAPE is marginally better than the Gradient Boosting Decision Tree with the full complement of features. Later, deep neural network architectures will model the relationship between features and $PM_{2.5}$. Therefore, it might be useful to include all features, as these networks can capture intricate, non-linear relationships in data. So, despite the somewhat unsatisfactory impact of extra feature sets, both the OSM and the weather features will be included from now on in this study.

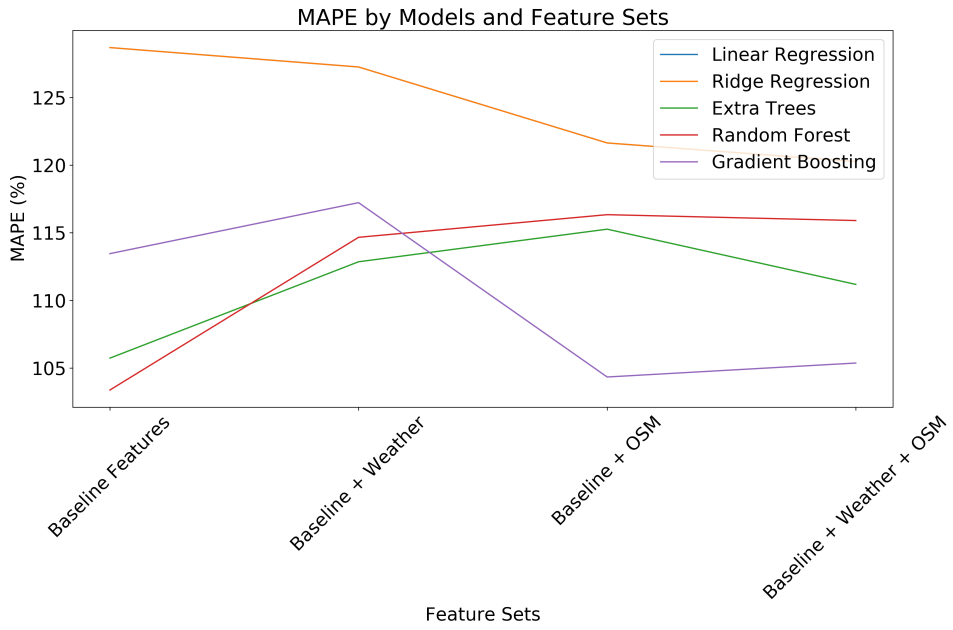


Figure 4.5: Evaluation of the mean absolute percentage error (MAPE) across different models and feature sets

4.2 Stationary Sensor Model

The AirSpeck stationary sensor data is augmented with the various features outlined in Section 4.1 and is split into training and validation datasets using an 80/20 split — a typical division in machine learning tasks [54]. The hour-averaged $PM_{2.5}$ concentration recorded by the static AirSpeck sensors is used as the ground truth to train and evaluate the models. Essentially, the models aim to predict this based on the provided input features. Land Use Regression (LUR) is a commonly used method for the spatial prediction of pollutant concentration [14, 33, 34, 55]. LUR spatially models pollutant concentration using land use and other geographical indicators as explanatory variables. The features discussed in Section 4.1 serve as the explanatory variables in these models, with the stationary $PM_{2.5}$ concentration being the target to predict. For both the DAPHNE and INHALE stationary sensor datasets, a simple five-layer fully connected neural network is trained and evaluated in comparison with several machine learning algorithms, namely: linear regression, ridge regression, extra trees, random forests, and gradient boosting regression. For each dataset, the hyperparameters of the best-performing spatial model will be tuned. This model will then be used to create a time series of spatial interpolations to train the spatio-temporal model. This method proved

more effective than first predicting temporally and then interpolating spatially as found in [14]. The temporal model adopts the leave-one-subject-out method, where it is trained on data from all subjects except one and then evaluated using the data of the excluded subject. Three temporal models will be trained and evaluated: a single-layer Long Short Term Memory network, a Gated Recurrent Unit network, and a stacked LSTM, which consists of two LSTM layers. These models are trained for 100 epochs with a mean absolute percentage error loss function. The learning rate and the number of units in the feed-forward layer will be tuned for the model that achieves the lowest MAPE.

4.2.1 Transfer Learning

A significant amount of data has been collected by the AirSpeck stationary sensors for both the INHALE and DAPHNE studies, comprising 11483 and 19339 hour-averaged data points respectively. To mimic a scenario where there exists far less data for one of the locations, a small subset of the DAPHNE data will be extracted on a subject-by-subject basis. Data from 10% of the subjects was included in this as this ensured that the pre-training dataset (INHALE stationary sensor data) had over five times as many data points as the target dataset. Several machine learning methods will be trained on this small dataset as well as a simple fully connected neural network. These methods will be compared to a neural network that is pre-trained on the INHALE data and then fine-tuned on the subset of the DAPHNE dataset. It is hoped that the model can learn patterns from the larger INHALE dataset which can improve the validation performance on the target dataset.

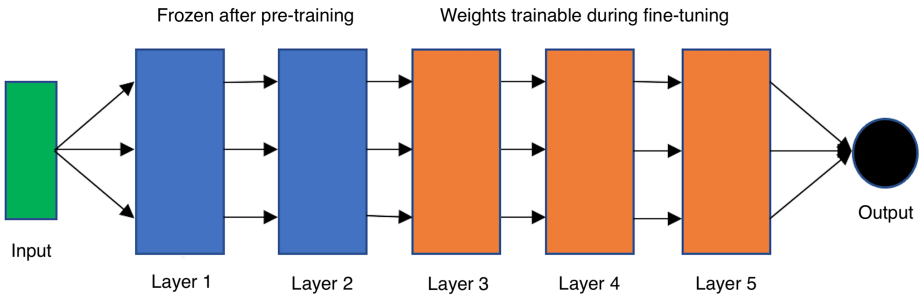


Figure 4.6: A simple visualisation of the neural network architecture for the transfer learning model

To force the network to retain information from the pre-training dataset, the first two

layers of the network are frozen after pre-training. These layers are shaded blue in Figure 4.6, with the orange layers representing those with trainable weights during the fine-tuning process. If no layers are frozen and the network is trained for an extended period on the target dataset, there is a risk of the model experiencing *catastrophic forgetting*. This term refers to the phenomenon wherein the model loses the knowledge encapsulated in its initial weights (learned during pre-training) as it overfits to the new data. Despite the importance of weight initialization in deep learning models, the useful generalisations acquired during pre-training can be unintentionally overridden during fine-tuning if none of the layers are frozen [56]. The network will be fine-tuned for the number of epochs that minimises the validation MAPE. The same process will be followed in the reverse direction: a 25% the INHALE subjects will be taken (which ensured that the pre-training dataset was at least five times as large as the target dataset), and several machine learning prediction models will be trained on this dataset. These will be evaluated alongside a neural network which is pre-trained on the entire DAPHNE stationary dataset and fine-tuned on the small INHALE set. As discussed above, the first two layers of this model will also be frozen during fine-tuning.

4.3 Personal Sensor Model

The personal sensor models use minute-averaged personal exposure data from the mobile AirSpeck sensors as the ground truth for $PM_{2.5}$ concentration. The mobile sensor data is augmented with the various features outlined in Section 4.1 and split into a training and a validation dataset using an 80/20 split. Several spatial models, including a feed-forward neural network, are trained to predict the minute level $PM_{2.5}$ concentration recorded by the INHALE personal sensors. The model that achieves the lowest MAPE is then fine-tuned used to create spatial interpolations. Both the true and interpolated $PM_{2.5}$ concentrations are resampled, taking the mean on an hourly basis, to prepare for temporal modeling. Due to the somewhat volatile nature of the minute-by-minute readings and the evident cyclical nature of $PM_{2.5}$ at the hourly level, the spatio-temporal model is trained on hourly data points. Spatial interpolations from the last three hours are employed to train a temporal model that predicts one hour into the future. As discussed in Section 4.2, three different temporal models are trained and evaluated: a single-layer LSTM, a Gated Recurrent Unit, and a two-layer stacked LSTM. The model producing the lowest MAPE will be fine-tuned to establish the final spatio-temporal model.

4.3.1 Transfer Learning

Table 4.1: Comparison of AirSpeck mobile sensor datasets

Dataset	No. of Data Points	Mean $PM_{2.5}$ Concentration $(\mu g/m^3)$
INHALE	352,638	12.74
Leon	4,947	27.39
Guadalajara	2,993	23.51

The Leon and Guadalajara personal sensor datasets are dwarfed in size by the INHALE personal sensor data, see Table 4.1. In a manner analogous to the stationary sensor model, a neural network will be pre-trained on the entire INHALE personal sensor dataset and then fine-tuned on each of the smaller datasets in order to predict personal $PM_{2.5}$ exposure. This approach aims to capture some of the insights gained by the model from the much larger dataset. Even though the model is pre-trained on data from a different city, which will inherently have a different distribution of $PM_{2.5}$ concentration, the hope is that by initialising the model with extensive training data from a similar task, the model can discern patterns otherwise obscured due to the restricted data from the target dataset. All personal sensor models are trained on minute-averaged data points, with predictions made at this granularity. Given that some of the true values can be notably volatile, the estimated exposure values are averaged on an hourly basis and juxtaposed with the hourly averages of the true exposure values for evaluation. This approach assesses the ability of the models to approximate $PM_{2.5}$ concentration over an hour, even while being trained on minute resolution data. By training on minute-resolution data, the spatial model can swiftly and accurately adjust to abrupt transitions. For instance, one data point might represent a subject walking alongside a busy road, and the subsequent data point might reflect them settling in a park for a brief interval. These environmental shifts are more accurately captured at a finer time resolution.

4.4 Evaluation Metrics

In this study, two main metrics will be used to compare and analyse the performance of the prediction models: mean absolute error (MAE) and mean absolute percentage error (MAPE). Given a set of predictions $y_1, y_2, ..., y_N$ with corresponding true values $d_1, d_2, ..., d_N$, then the formula for mean absolute error is given in Equation 4.3.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - d_j| \tag{4.3}$$

Mean absolute error is expressed in the same units as the data being measured (in this case $\mu g/m^3$) and indicates how close the predictions are to the true values. Mean absolute percentage error normalises the absolute error of each prediction by the true value. The formula for mean absolute percentage error is defined in Equation 4.4.

$$MAPE = \frac{100}{N}\sum_{j=1}^{N}|\frac{y_j - d_j}{d_j}| \tag{4.4}$$

The dichotomy between MAE and MAPE is evident in their sensitivity to the magnitude of true values. While MAE provides a direct measure in the same units as the data, MAPE offers a normalised assessment that prevents larger true values from disproportionately influencing the error. Hence, while both metrics will be used to analyse and evaluate predictive models in this study, greater emphasis will be placed on MAPE.

Chapter 5

Experiments and Results

5.1 Stationary Sensor Models

The stationary sensor models use all of the features outlined in Section 4.1 to predict the $PM_{2.5}$ concentration recorded by the stationary AirSpeck sensors for both the INHALE and DAPHNE studies. For each dataset, different spatial prediction models are compared by their MAEs and MAPEs with the best-performing model being combined with the best performing of three temporal models to construct a spatio-temporal model to predict $PM_{2.5}$ concentration for the stationary AirSpeck sensors one hour into the future.

5.1.1 INHALE

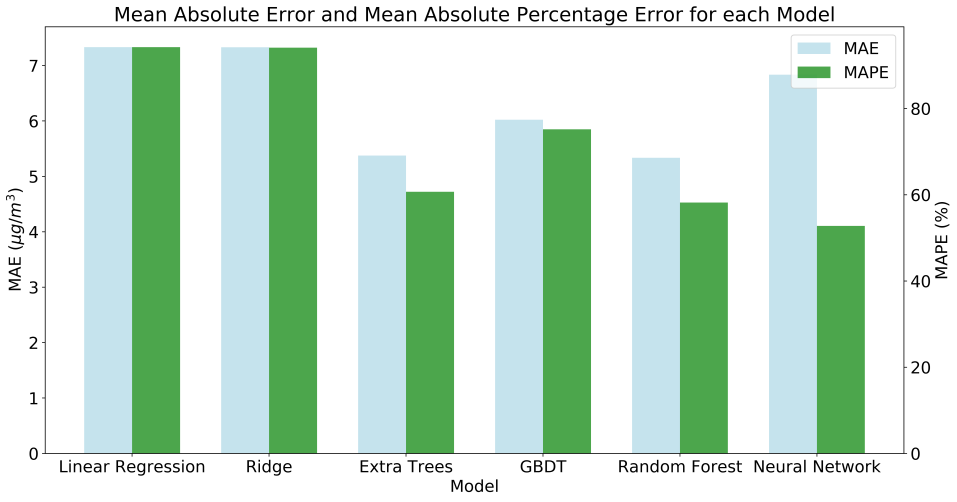


Figure 5.1: Evaluation of the spatial estimation models on the held-out data

As observed in Figure 5.1, the artificial neural network model registers the smallest MAPE with 52.77%. However, the MAE for the neural network is higher than that of several other models. One possible explanation is that the neural network is specifically trained to minimise MAPE, which isn't the case for the other models. Alternatively, the inherent nature of the neural network might enable it to discern intricate relationships between the various input features that other machine learning models cannot grasp. The learning rate, number of hidden units, and dropout probability of the neural network model were adjusted, with the optimal hyperparameters yielding a reduction in MAPE from 52.77% to 49.49%. The most significant contributor to this reduction was the integration of dropout layers after each hidden layer with a dropout probability of $p = 0.5$. Dropout often enhances the ability of neural networks to generalise by diminishing overfitting [57]. Comprehensive results from the hyperparameter tuning can be found in Appendix A.

Table 5.1: Evaluation of the spatio-temporal models for varying window sizes on the held-out data

Model	History Size (Hours)	MAE $(\mu g/m^3)$	MAPE (%)
LSTM	3	**5.902**	**42.08**
	5	6.057	42.99
	8	6.160	43.29
GRU	3	5.935	42.37
	5	5.964	42.58
	8	5.952	42.70
Stacked LSTM	3	6.061	42.75
	5	6.474	45.43
	8	6.592	46.16

Of the models trained, the single-layer LSTM model trained on a window size of three hours resulted in the lowest MAE and MAPE, as shown in Table 5.1. Perhaps counter-intuitively, increasing the window size of previous spatial interpolations that the model is trained on does not enhance performance. The shortest window size corresponds to the lowest MAPE for all three temporal model architectures. This suggests that forecasting one hour into the future doesn't necessitate an extensive history of spatial interpolations. For the rest of this study, all spatio-temporal models will be trained on a three-hour history of spatial interpolations. After hyperparameter tuning, the optimal

spatio-temporal model has a significantly lower MAPE (42.08%) than the spatial model used for the interpolations (49.49%). The spatio-temporal prediction of the LSTM combined with the feed-forward spatial model is illustrated for one of the subjects in Figure 5.2. For this subject, the spatio-temporal model captures the cyclical nature of the $PM_{2.5}$ concentration very well.

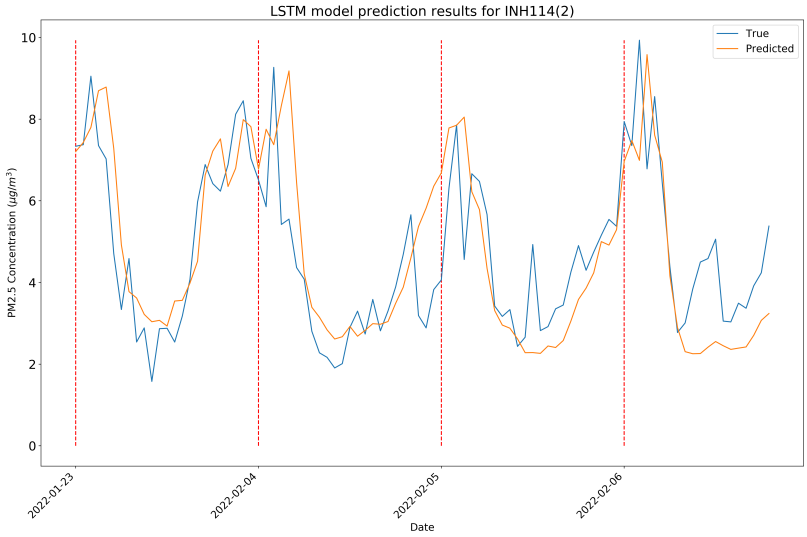


Figure 5.2: Spatio-temporal predictions for the optimal model are plotted against the true $PM_{2.5}$ concentration for one of the stationary sensors associated with subject INH114 of the INHALE study

5.1.1.1 Transfer Learning

A feed-forward neural network which was pre-trained on the DAPHNE dataset and then fine-tuned on the target dataset was trialled for the entire INHALE dataset as well as a 50% subset of it, however in both cases, it yielded a higher MAPE than the neural network trained solely on the target dataset, see Appendix A. However, for the 25% subset of the INHALE stationary dataset, the fine-tuned feed-forward neural network attains a lower MAPE (48.61%) than all of the other machine learning techniques, as shown in Figure 5.3.

The training curves depicted in Figure 5.4 illustrate the progression of MAE and MAPE as the neural network adapts to the target data. Both the training and validation MAE and MAPE are very high before the model is fine-tuned, since the model originally trained on the DAPHNE data is being used to predict the $PM_{2.5}$ concentration for the INHALE dataset. It takes only six epochs of fine-tuning for the validation MAPE to

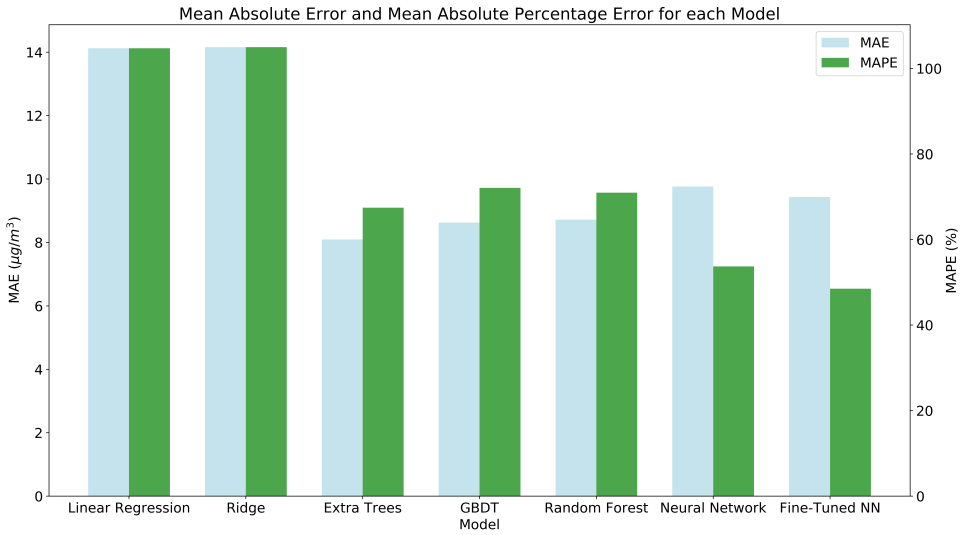


Figure 5.3: Evaluation of spatial estimation models on the held-out data for a 25% subset of the INHALE stationary sensor data

reach its minimum value. Both training and validation MAPE reduce rapidly, indicating the model's adeptness at fitting to the target data. The use of transfer learning evidently induces a noticeable improvement in generalisation performance as it pertains to the validation MAPE. However, the MAE achieved by the fine-tuned ANN exceeds that of a few of the machine learning algorithms.

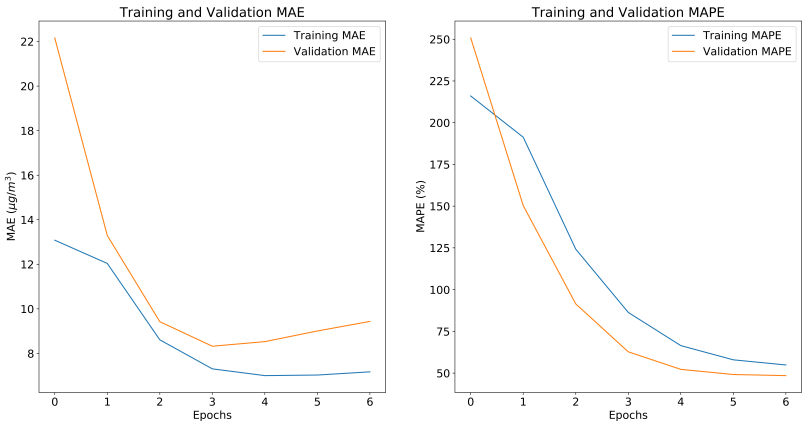


Figure 5.4: Training curves of the feed-forward neural network as it is fine-tuned to a 25% subset of the INHALE stationary data

5.1.2 DAPHNE

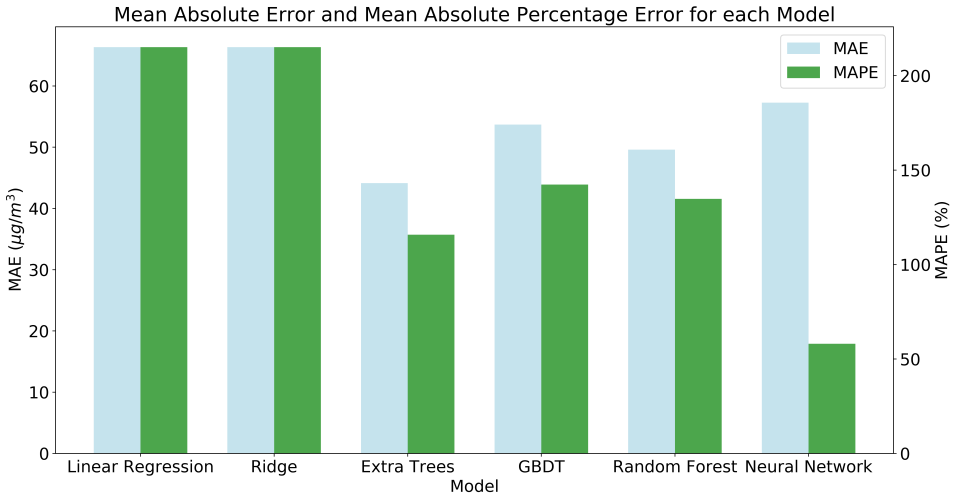


Figure 5.5: Evaluation of spatial estimation models on the held-out data

Refer to Figure 5.5 to observe that the neural network estimation model yields the smallest MAPE (58.06%). While this model has a higher MAE compared to a few other models, its MAPE is substantially lower than all of the other models (the next lowest is the Extra Trees model with a MAPE of 115.85%). After hyperparameter tuning of the neural network, MAPE falls to 56.14% and the spatial interpolations from the optimal model were applied to the DAPHNE dataset. This data was used to train three different temporal models. The architecture that resulted in the lowest MAE and MAPE was the single-layer LSTM network, as highlighted in Table 5.2. The MAPE of this spatio-temporal model is marginally lower than that of the purely spatial model. This indicates that introducing the task of predicting one data point (one hour) into the future, based on the three most recent spatial interpolations, doesn't lead to increased error. After hyperparameter tuning, the MAPE is reduced to 54.27%, as detailed in Appendix B.

Table 5.2: Performance comparison of spatio-temporal estimation models

Model	**MAE** $(\mu g/m^3)$	**MAPE (%)**
Long Short Term Memory (LSTM)	**67.08**	**54.65**
Gated Recurrent Unit (GRU)	67.37	54.66
Stacked LSTM	69.00	55.05

Spatio-temporal predictions for one of the subjects are plotted in Figure 5.6. Over the

course of 24 hours, the predictions effectively capture the peaks and dips in $PM_{2.5}$ concentration. However, the highest peak is under-predicted, and the lowest trough is over-predicted. Furthermore, it's the very high true values for some of the other subjects that this model struggles to estimate accurately, contributing to the relatively high MAPE for this model.

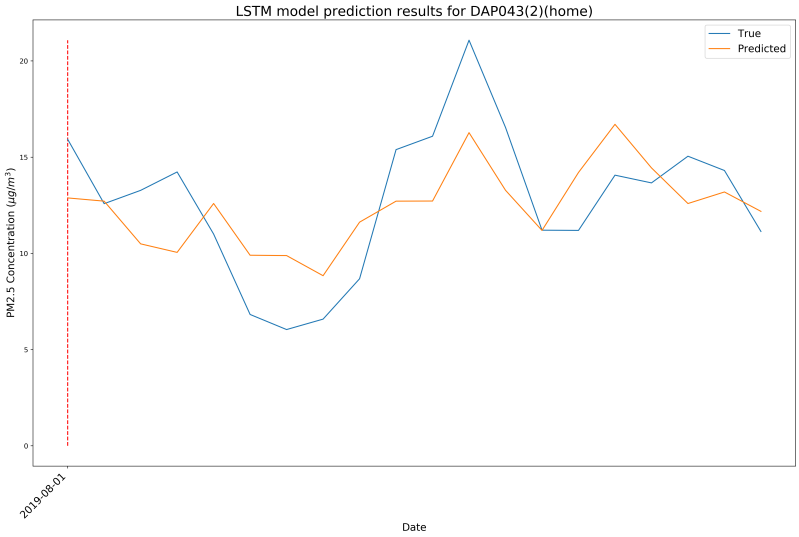


Figure 5.6: Spatio-temporal predictions are plotted against true $PM_{2.5}$ concentration for one of the stationary sensors associated with one of the subjects of the DAPHNE study

5.1.2.1 Transfer Learning

Observe in Figure 5.7 that for the 10% subset of the DAPHNE stationary sensor dataset, the fine-tuned neural network exhibits a lower MAPE than all the other predictive models, with a slight improvement $(54.86\% \text{ vs } 61.82\%)$ on the neural network trained on solely this dataset. Furthermore, there is a slight increase in MAE as well (from $29.29\mu g/m^3$ to $33.40\mu g/m^3$). These two deep learning models vastly outperform the other machine learning algorithms. It would appear that for this dataset, building upon existing knowledge by incorporating knowledge from a larger related dataset improves the ability of the estimation model on the task-specific dataset.

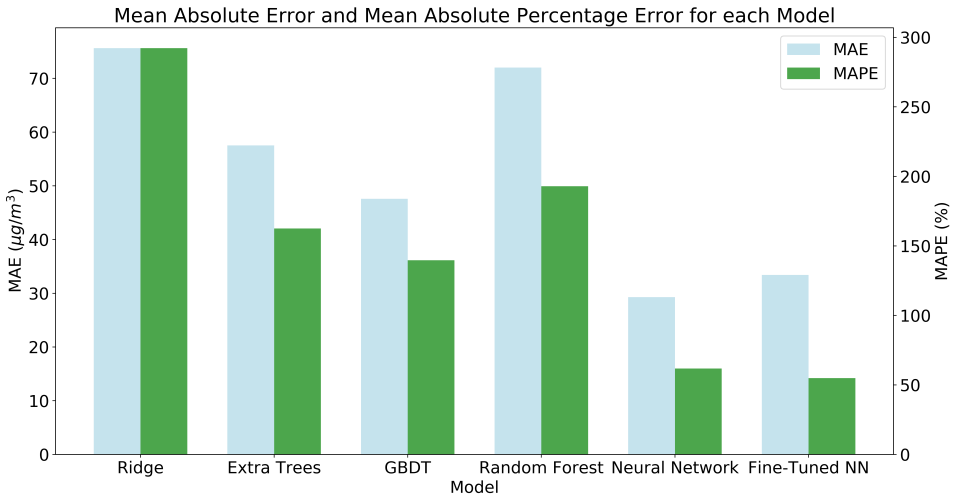


Figure 5.7: Evaluation of spatial estimation models on the held-out data for the 10% subset of the DAPHNE dataset

Furthermore, it can be seen in Figure 5.8. that before any fine-tuning has been carried out on the DAPHNE data, the MAPE is quite high and comes down significantly after only one epoch of fine-tuning, with the validation MAPE reaching a minimum after two epochs of fine-tuning. This backs up the intuition that the pre-trained network would exhibit poor initial predictive performance since it was trained on data from a different location. For both a 50% subset and the entire DAPHNE dataset, the transfer learning approach doesn't reduce the MAPE. This could be because the datasets are adequately sized, rendering the pre-training unnecessary.

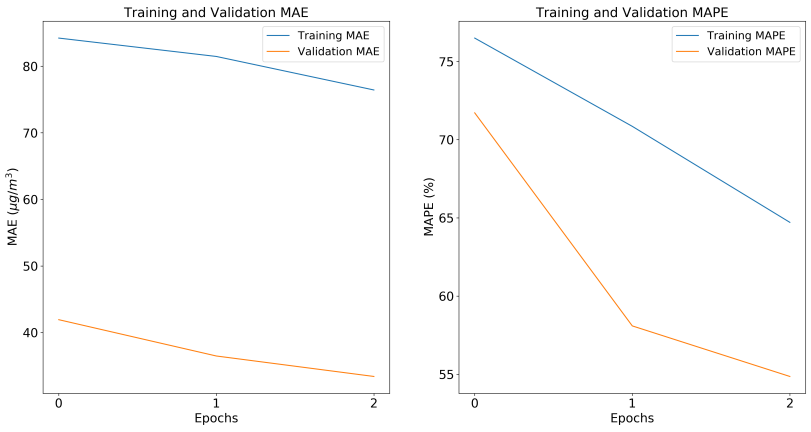


Figure 5.8: Training curves of the neural network as it is fine-tuned on the 10% subset of the DAPHNE dataset

5.2 Personal Sensor Models

The personal sensor models are trained to predict the $PM_{2.5}$ exposure, as measured by the mobile AirSpeck sensors, for the INHALE, Leon and Guadalajara datasets.

5.2.1 INHALE

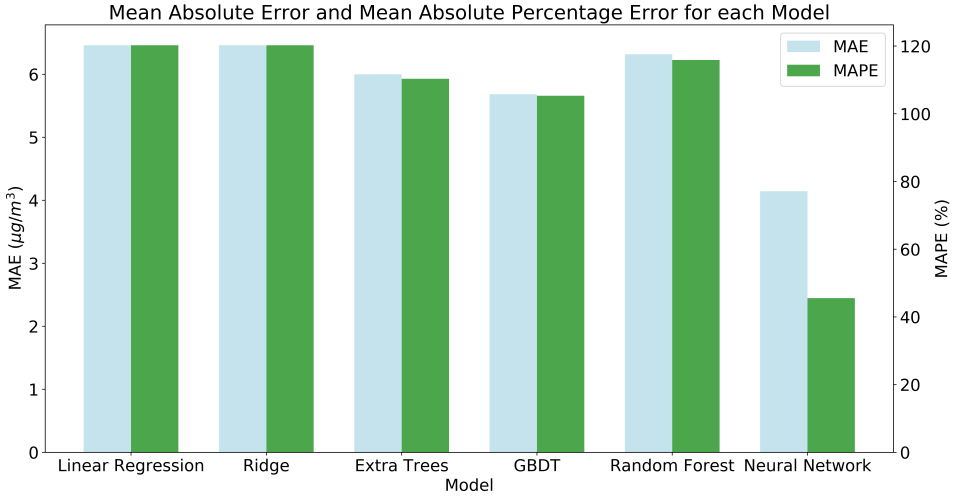


Figure 5.9: Evaluation of spatial estimation models on the held-out data

Firstly, several models are trained on the INHALE personal data. The MAE and MAPE achieved by each model are depicted in Figure 5.9. It is evident that the artificial neural network (ANN) outperforms all other models; it attains a significantly lower MAE and MAPE on the unseen validation data than all of the machine learning models. The learning rate, number of hidden units, and dropout probability were tuned. The lowest MAPE was achieved with a learning rate of 0.001, 32 hidden units, and a dropout probability of 0.5. With these optimal parameters, the MAPE decreases from 45.54% to 42.27%. The comprehensive results of the hyperparameter tuning are detailed in Appendix C. Furthermore, the neural network aligns well with the objective of this study: transferring knowledge from one model to another. The ANN trained on the INHALE mobile sensor data serves as a foundation for the fine-tuned neural networks that will estimate the spatial data for Leon and Guadalajara.

Figure 5.10: Contour plot of estimated $PM_{2.5}$ concentration $(\mu g/m^3)$ on the 20th of July 2023 at 7am [13]

Estrada (2023) uses this spatial predictive model to generate walking and cycling routes which minimise airborne particulate matter pollution [13]. For a specific date and time, the spatial estimates generated by the model exhibit minimal variance. Figure 5.10 offers a visualisation of this in the form of a contour plot, taken from [13]. Across an area of around 50 square kilometers, the estimates consistently fall within the range of $4-8\mu g/m^3$. The estimates for the plot were calculated for a grid, with each grid point representing the centroid of a 100m x 100m square. Only the OSM features, and in certain cases the Grid Average $PM_{2.5}$ and the Grid Average Distance features, differ for each GPS locations that $PM_{2.5}$ concentration was estimated for. As a result, it's plausible for the model to produce relatively consistent estimates across this area. Moreover, there's a trend of higher estimated $PM_{2.5}$ concentrations in more built-up areas. For instance, estimates in Hyde Park are generally lower than those in the surrounding built-up regions. This appears to be influenced by the road type and land usage features incorporated in the spatial model. In Figure 5.11, a journey for a subject from the held out set in plotted. Colours toward to left hand side of the colour bar correspond to the spatial model under-predicting $PM_{2.5}$ concentration while colours to the right of the bar correspond to over-prediction. The model is predicting most of the exposure values

with minimal error however, a few are being considerably under-predicted. The subject seems to be walking alongside roads throughout the journey, and the model might not be capturing some peak $PM_{2.5}$ exposures possibly due to dense traffic at specific times. For minute resolution data, there will be some amount of randomness in exposure that the spatial model isn't able to capture due to factors such as traffic and wind gusts. However, the model may be capturing more general traffic trends based on the hour of the day as well as proximity to busier roads.

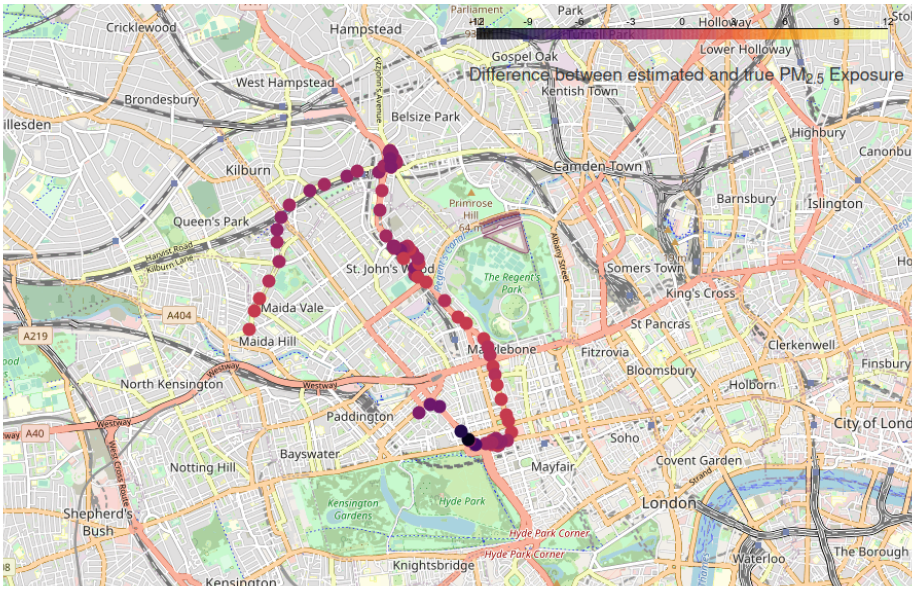


Figure 5.11: Illustration of the difference between real and $PM_{2.5}$ exposure and the spatial predictions for a journey taken by subject INH006

After hyperparameter tuning of the spatial model, the spatial model with the lowest MAPE is used to create spatial predictions which are in turn used to train three different temporal models. The results reported in Table 5.3. indicate that the single layer LSTM achieves a lower MAPE and MAE than both of the other RNN variants. After tuning the learning rate and number of hidden units in the dense layers of the LSTM there is no decrease in MAPE. Unlike for the stationary sensor data, the MAPE of the spatio-temporal model is higher than that of the spatial model. It is reasonable that the spatio-temporal model's MAPE slightly exceeds that of the spatial model since, in addition to generating spatial predictions, these are being used to predict one hour into the future.

The predictions for the optimal spatio-temporal model have been plotted against true $PM_{2.5}$ concentration for subject INH126 in Figure 5.12. There is an obvious cyclic nature to the true $PM_{2.5}$ concentration which is picked up on to a certain extent by

Table 5.3: Performance comparison of spatio-temporal estimation models

Model	**MAE** $(\mu g/m^3)$	**MAPE** (%)
Single Layer LSTM	**5.402**	**43.04**
Gated Recurrent Unit (GRU)	5.410	43.07
Stacked LSTM	5.423	43.26

the spatio-temporal predictions. The peaks in $PM_{2.5}$ are consistently under-predicted and the model misses a few of these peaks completely. This is likely due to the spatial interpolations under-predicting or completely missing these spikes in $PM_{2.5}$. Furthermore, since the LSTM is trained with minimising MAPE as the objective, it may be more desirable for the model to err on the side of under-prediction heightened $PM_{2.5}$ concentration if the model is finding it difficult to predict the occurrence of these spikes accurately.

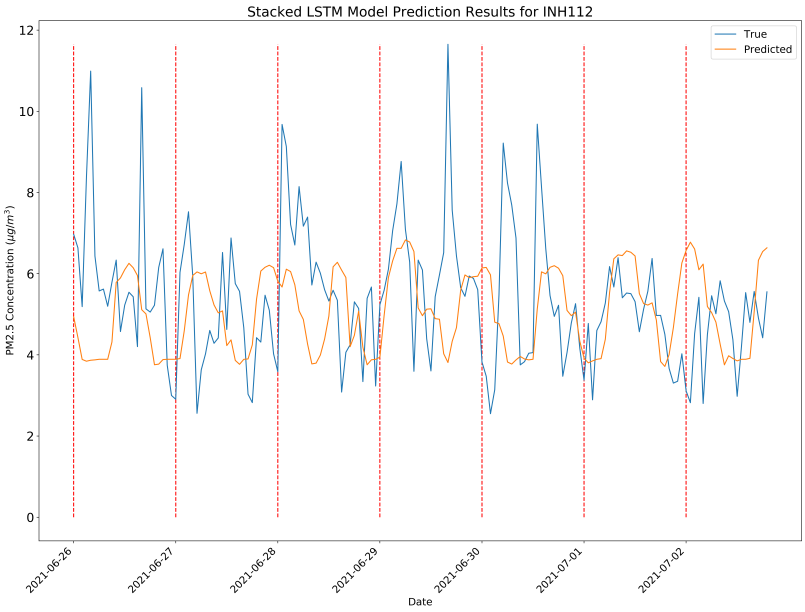


Figure 5.12: Spatio-temporal predictions of the hybrid feed-forward neural network and LSTM network are plotted against true $PM_{2.5}$ exposure measured by the personal AirSpeck sensor for INHALE subject INH112

5.2.2 Transfer Learning

5.2.2.1 Leon

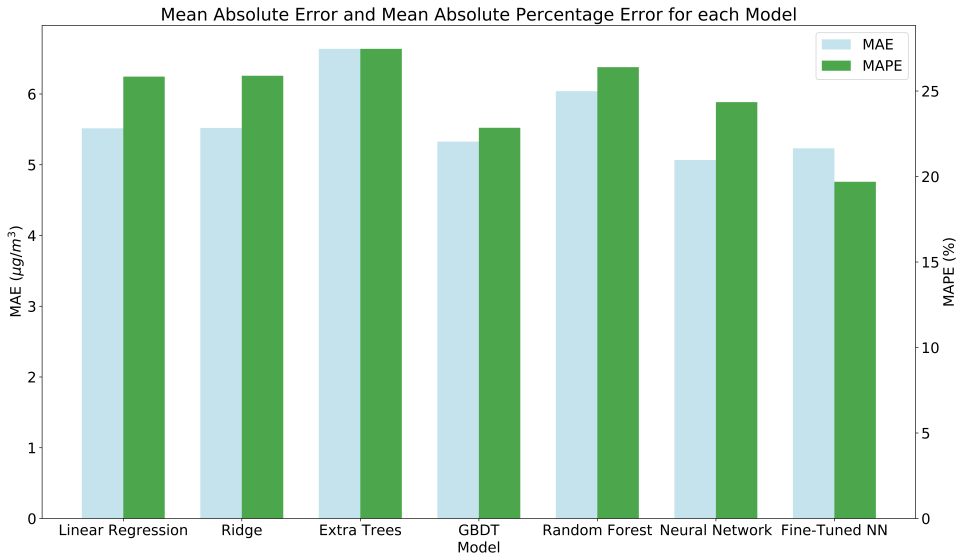


Figure 5.13: Evaluation of spatial estimation models on the held-out Leon data

In Figure 5.13, it's evident that for the Leon data, the fine-tuned neural network showcases a noticeably lower MAPE than the standard neural network trained exclusively on the target dataset. These two networks have nearly identical model architectures since the number of hidden layers and units are consistent. However, the learning rate is altered for the fine-tuning process since there are fewer weights to adjust. This network also surpasses all of the baseline regression models in terms of both MAPE and MAE. The reduction in MAPE is relatively modest - the extra trees spatial model presents a MAPE of 24.28% and the fine-tuned neural network yields a MAPE of 20.84%, - reflecting a MAPE reduction of approximately 14%. The training curves for the neural network's fine-tuning can be observed in Figure 5.14. Predictably, the initial training and validation losses are lofty, but after just one epoch, both the loss and MAPE plummet. Subsequently, the training loss tapers off as the network adapts to the Leon training data. The MAPE for the unseen data also diminishes incrementally as the model becomes more attuned to the target data, bottoming out after 29 epochs. Some fluctuation in the validation MAPE is evident throughout the fine-tuning, possibly a result of the relatively limited sample size.

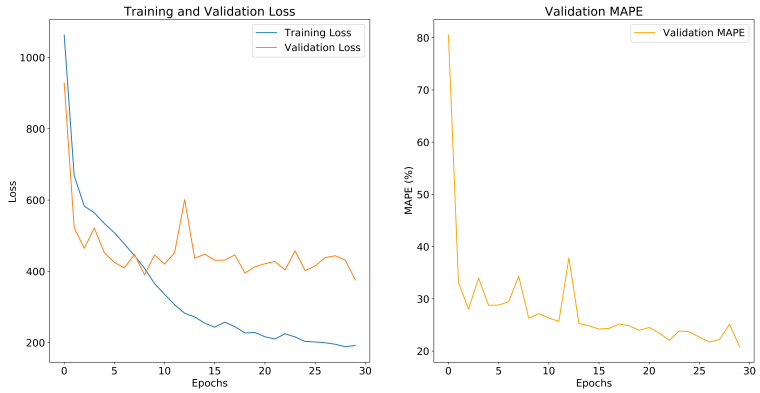


Figure 5.14: Loss and MAPE plots illustrating the fine-tuning process for the Leon model

5.2.2.2 Guadalajara

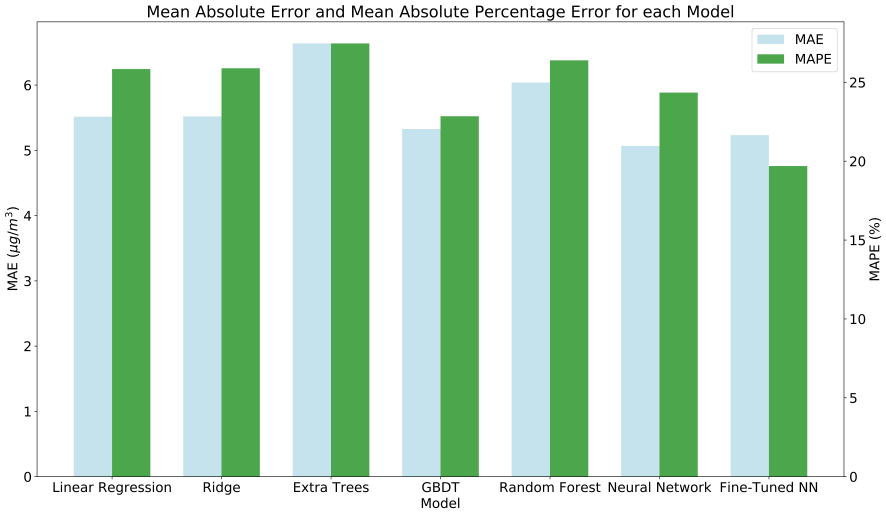


Figure 5.15: Evaluation of spatial estimation models for the held-out Guadalajara data

For the Guadalajara dataset, the fine-tuned neural network exhibits a lower mean absolute percentage error than all of the other models, including the neural network trained solely on the target dataset. This indicates that the patterns and trends learned on the INHALE data are beneficial to the feed-forward neural network as it learns to predict the Guadalajara data. For both the Guadalajara and Leon datasets, using a mean squared error loss function resulted in a lower MAPE than a MAPE loss function did. Training and validation loss are plotted alongside validation MAPE in Figure 5.16. Both losses start very high and plummet rapidly as the model begins to fit to the Guadalajara

data. After 12 epochs of fine-tuning, the validation MAPE reaches its minimum of 19.69%. This is a relative improvement of more than 10% compared to the next best model, which is the gradient boosting decision trees model with a MAPE of 22.85%.

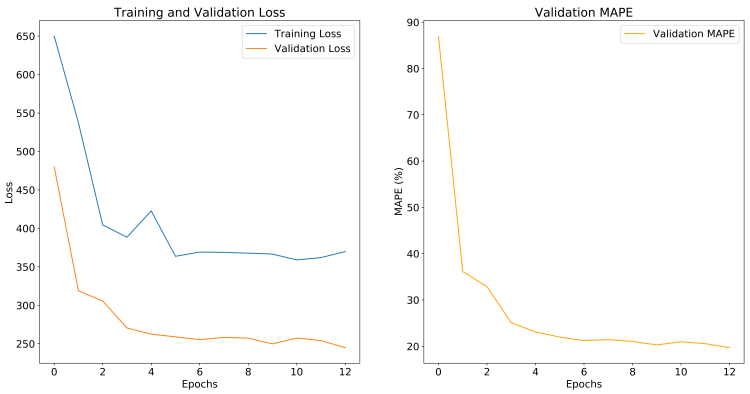


Figure 5.16: Loss and MAPE plots illustrating the fine-tuning process for the Guadalajara model

5.3 Summary of Results

The mean absolute errors are much higher for the DAPHNE models because the $PM_{2.5}$ concentration is on average much higher in Delhi than it is in London. Therefore, comparing the MAPEs of these models would be a more judicious approach. The spatial and spatio-temporal models for the INHALE dataset achieve lower MAPEs than the equivalent models for the DAPHNE dataset. Many factors could account for this, including data quality and differences in atmospheric conditions between London and Delhi. Additionally, there are two reference sensors in a similar area of London where most of the stationary AirSpeck sensors in the INHALE study are located, while only CPCB monitoring stations scattered across the vast Delhi metropolitan area are used as reference data for the DAPHNE stationary sensors. The mean value of the Grid Average Distance feature, defined as the average distance to the three closest reference sensors, is 3.66km for INHALE and 6.88km for DAPHNE. This might account for the superior performance of the INHALE spatial model. Surprisingly, the INHALE spatio-temporal model achieves a slightly higher MAPE than the equivalent INHALE static sensor model in [34], which records a MAPE of 38.09%, even though the model in this study was trained on a larger dataset with added input features. Such variance could be attributed to minor differences in methodology like outlier removal or hyperparameter

optimisation strategy.

Model	**DAPHNE**		**INHALE**	
	MAE $(\mu g/m^3)$	**MAPE**$(\%)$	**MAE** $(\mu g/m^3)$	**MAPE**$(\%)$
Spatial	58.02	56.14	6.789	49.49
Spatio-temporal	66.48	54.27	5.902	42.08
	DAPHNE-small		**INHALE-small**	
	MAE $(\mu g/m^3)$	**MAPE**$(\%)$	**MAE** $(\mu g/m^3)$	**MAPE**$(\%)$
Spatial	33.40	54.86	9.435	48.49

Table 5.4: Comparison of the optimal spatial and spatio-temporal models for INHALE and DAPHNE stationary sensor datasets

The fine-tuned spatial model for the smaller INHALE dataset delivers performance closely aligned with the best spatial model for the entire INHALE dataset. Thus, the transfer learning approach is compensating effectively for the limited training data in this scenario. The fine-tuned spatial model for the DAPHNE dataset registers a notably lower MAE and MAPE than the top model for the entire dataset. This discrepancy might be attributed to the small size of the validation data, as it's counter-intuitive to think this model would outperform the model trained on the full DAPHNE dataset in terms of generalisation ability.

Model	**INHALE**		**Leon**		**Guadalajara**	
	MAE	**MAPE**$(\%)$	**MAE**	**MAPE**$(\%)$	**MAE**	**MAPE**$(\%)$
Spatial	4.169	42.27	7.149	20.84	5.232	19.69
Spatio-temporal	5.402	43.04	N/A	N/A	N/A	N/A

Table 5.5: Comparison of the optimal spatial and spatio-temporal models for INHALE, Leon and Guadalajara personal sensor datasets

The spatial model trained on the INHALE personal sensor data achieves a markedly lower MAPE than its stationary sensor counterpart. The AirSpeck personal sensors, worn by individuals during daily activities, record data with a significantly broader spatial variance than the AirSpeck stationary sensors, possibly enhancing the generalisation capability of the personal sensor model over the stationary one. The spatio-temporal personal sensor model yields a comparable MAPE to the equivalent stationary sensor

model and significantly surpasses the spatio-temporal model trained on INHALE personal sensor data in [34], which registered a MAPE of 73.48%. The optimal spatial models for the Leon and Guadalajara datasets adopt the INHALE spatial model as a foundational blueprint. While the MAPE for spatial predictions is reasonably similar between Leon and Guadalajara, the MAE is appreciably elevated for Leon due to a higher average $PM_{2.5}$ concentration than in Guadalajara. Both of these fine-tuned spatial models outperform the spatial models trained specifically on these datasets in [14]. Furthermore, they both significantly undercut the MAPE of the INHALE spatial model. This divergence might stem from the different data gathering approaches across studies. The INHALE personal sensors are worn (for the most part) at all hours by participants, whereas in Leon and Guadalajara, the personal sensors were worn for an hour at a time while subjects adhered to a uniform outdoor route. The consistency of this collection method likely induces reduced data variability, yielding more accurate predictions as a result. Another vital factor is the reference $PM_{2.5}$ data. In both Leon and Guadalajara, six fixed AirSpeck sensors serve as the $PM_{2.5}$ reference sensors. These stationary sensors are situated in proximity to the data collection zones of the mobile AirSpeck sensors, as depicted in Figure 2.2. The INHALE models incorporate data from seven AURN sensors, six LAQN sensors, and two stationary AirSpeck sensors. These AURN and LAQN monitoring stations are dispersed throughout London, as illustrated in Figure 2.3b. Consequently, the average Grid Average Distance value across the dataset is considerably greater in London (4.68km) than in either Leon (0.94km) or Guadalajara (0.69km). This discrepancy likely factors into the superior predictive accuracy observed in Leon and Guadalajara relative to London.

Chapter 6

Conclusions and Future Work

Spatial estimation models were trained on static AirSpeck data in both London and Delhi in order to predict the $PM_{2.5}$ concentrations as measured by these sensors. The INHALE model (London) achieved a MAPE of 49.49% and the DAPHNE model (Delhi) achieved a MAPE of 56.14%. The spatio-temporal models for these datasets predict $PM_{2.5}$ concentration one hour into the future using the previous three spatial predictions and yield lower MAPEs than their spatial counterparts in both cases. The INHALE model achieves a MAPE of 42.08% and the DAPHNE model has a MAPE of 54.27%. A fully connected feed-forward neural network was pre-trained on the DAPHNE data and then fine-tuned on a subset of 25% of the INHALE dataset. This model achieved a MAPE of 48.49% which was lower than for any of the other models trained on the same subset of the INHALE data. Similarly, a feed-forward neural network pre-trained on the INHALE data and fine-tuned on a subset of 10% of the DAPHNE data gave rise to a MAPE of 54.86% which made it the best performing model with regards to this metric on this subset of the DAPHNE data. This transfer learning method was found to be ineffective for larger subsets of each dataset.

For the INHALE data, the neural network spatial model outperformed other architectures with a MAPE of 42.27%. For the INHALE mobile AirSpeck data the feed-forward neural network combined with a stacked LSTM was the optimal spatio-temporal model with a MAPE of 43.04%. The spatial neural network model was fine-tuned on mobile AirSpeck data from both Leon and Guadalajara. The Leon spatial model yielded a MAPE of 20.84% which was considerably lower than any of the other models trained on this dataset. Similarly, the Guadalajara model exhibits strong performance on the held-out validation set with a MAPE of 19.69%. For both of these datasets, pre-training the fully connected neural network on the INHALE mobile AirSpeck data, freezing

the first two layers of the network and then fine-tuning on the target dataset leads to a lower MAPE than training the same neural network architecture solely on the target dataset. In other words, the transfer learning approach has resulted in improved spatial predictions for both the Leon and Guadalajara datasets.

The primary limitation of the spatio-temporal models is their spatial predictions, as the addition of temporal forecasting doesn't significantly affect MAPE. There is always going to be some unpredictability when estimating air pollution levels, however it is highly probable that the mean absolute percentage errors of the both INHALE spatial models as well as the DAPHNE spatial model can be reduced significantly. The feature set in this study did not include any explicit traffic features as it was hoped that the features representing the hour of the day as well as proximity to different types of busy roads would implicitly model this to some degree. It is well established that vehicle exhaust fumes contribute significantly to $PM_{2.5}$. Therefore, introducing features into the analysis that explicitly represent traffic levels could help to extract more performance out of the predictive models. Furthermore, it was found that a dataset having a lower average distance to the closest three reference sensors tended to correspond to a smaller MAPE. Intuitively, the closer a reference sensor is located to a given location, the more likely it will be a good predictor of $PM_{2.5}$ concentration at that location. It is likely that the MAPE of all spatial models in this study could be improved if there was data from a wider variety reference sensors to draw upon.

It was demonstrated that transfer learning can be a powerful tool when training spatial $PM_{2.5}$ estimation models for locations with limited data. In this transfer learning framework, the feed-forward neural network architecture is pre-trained on one relatively large dataset in comparison to the target dataset. Combining several large datasets from a wide variety of geographical locations may improve the generalisability of the pre-trained model as it would be learning a more general representation of the input features which can then be fine-tuned on the target dataset. Furthermore, the transfer learning approach devised in this study focused on leveraging knowledge from one spatial model to another. Future studies may wish to investigate if this is a profitable approach to temporal modeling of $PM_{2.5}$. A fine-tuning approach to forecasting $PM_{2.5}$ concentration with limited data availability could be devised with one of the RNN variants defined in this work. This may be a valuable contribution due to the importance of forecasting pollution levels.

Bibliography

[1] Bert Brunekreef and Stephen T Holgate. Air pollution and health. *The lancet*, 360(9341):1233–1242, 2002.

[2] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.

[3] Philip J Landrigan. Air pollution and health. *The Lancet Public Health*, 2(1):e4–e5, 2017.

[4] Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: a review. *Frontiers in public health*, page 14, 2020.

[5] Bhola R Gurjar, Luisa T Molina, and C Shekhar P Ojha. *Air pollution: health and environmental impacts*. CRC press, 2010.

[6] Deepak Gautam and Nomesh B. Bolia. Air pollution: impact and interventions. *Air Quality, Atmosphere & Health*, 13:209–223, 2020.

[7] Jackson G Lu. Air pollution: A systematic review of its psychological, economic, and social effects. *Current opinion in psychology*, 32:52–65, 2020.

[8] World Health Organization. Ambient (outdoor) air quality and health. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, 2022. (Accessed on 11/07/2023).

[9] Kan Zheng, Shaohang Zhao, Zhe Yang, Xiong Xiong, and Wei Xiang. Design and implementation of lpwa-based air quality monitoring system. *IEEE Access*, 4:3238–3245, 2016.

[10] Federico Karagulian, Maurizio Barbiere, Alexander Kotsev, Laurent Spinelle, Michel Gerboles, Friedrich Lagler, Nathalie Redon, Sabine Crunaire, and Annette Borowiak. Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere*, 10(9):506, 2019.

[11] Xiaoting Liu, Rohan Jayaratne, Phong Thai, Tara Kuhn, Isak Zing, Bryce Christensen, Riki Lamont, Matthew Dunbabin, Sicong Zhu, Jian Gao, et al. Low-cost sensors as an alternative for long-term air quality monitoring. *Environmental research*, 185:109438, 2020.

[12] DK Arvind, Janek Mann, Andrew Bates, and Konstantin Kotsev. The airspeck family of static and mobile wireless air quality monitors. In *2016 Euromicro Conference on Digital System Design (DSD)*, pages 207–214. IEEE, 2016.

[13] Madeleine Santiago Estrada. Route generation for minimal airborne particulate exposure for walking and cycling in london. *MSc Dissertation, School of Informatics, University of Edinburgh*, 2023.

[14] Azam Khan. Sensing spaces: Spatiotemporal modelling of airborne particulate concentrations (pm2.5) using stationary and personal exposure monitors. *4th Year Project Report, Artificial Intelligence and Computer Science, School of Informatics, University of Edinburgh*, 2021.

[15] Junji Cao, Hongmei Xu, Qun Xu, Bingheng Chen, and Haidong Kan. Fine particulate matter constituents and cardiopulmonary mortality in a heavily polluted chinese city. *Environmental health perspectives*, 120(3):373–378, 2012.

[16] C. Arden Pope III and Douglas W Dockery. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air & Waste Management Association*, 56(6):709–742, 2006.

[17] World Health Organization. Health effects of particulate matter: Policy implications for countries in eastern europe, caucasus and central asia. 2013.

[18] Delhi Air Pollution: Health aNd Effects (DAPHNE). https://gtr.ukri.org/projects?ref=NE. (Accessed on 25/07/2023).

[19] Health assessment across biological length scales for personal pollution exposure and its mitigation (INHALE). https://gtr.ukri.org/projects?ref=EP. (Accessed on 25/07/2023).

[20] Central Pollution Control Board (CPCB). https://cpcb.nic.in. (Accessed on 17/07/2023).

[21] Automatic Urban and Rural Network (AURN). https://uk-air.defra.gov.uk/networks/network-info?view=aurn. (Accessed on 07/06/2023).

[22] London Air Quality Network (LAQN). https://www.londonair.org.uk/. (Accessed on 06/07/2023).

[23] RP5. https://rp5.ru. (Accessed on 05/07/2023).

[24] Open Street Map. https://www.openstreetmap.org. . (Accessed on 15/04/2023).

[25] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[26] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[27] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[29] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62, 1997.

[30] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[33] Ahalya Porchelvan. Sensing spaces: A framework for predicting pm2.5. *MSc Project Report, School of Informatics, University of Edinburgh*, 2021.

[34] Xudong Zhang. Machine learning based prediction of airborne particulates concentrations and composition in central london. *MInf Project (Part 2)*, 2023.

[35] Mahad S Baawain, Aisha S Al-Serihi, et al. Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network. *Aerosol and air quality research*, 14(1):124–134, 2014.

[36] CR Aditya, Chandana R Deshmukh, DK Nayana, and Praveen Gandhi Vidyavastu. Detection and prediction of air pollution using machine learning models. *International Journal of Engineering Trends and Technology (IJETT)*, 59(4):204–207, 2018.

[37] Anastasios Alimissis, Kostas Philippopoulos, CG Tzanis, and Despina Deligiorgi. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmospheric environment*, 191:205–213, 2018.

[38] Usha Mahalingam, Kirthiga Elangovan, Himanshu Dobhal, Chocko Valliappa, Sindhu Shrestha, and Giriprasad Kedam. A machine learning model for air quality prediction for smart cities. In *2019 International conference on wireless communications signal processing and networking (WiSPNET)*, pages 452–457. IEEE, 2019.

[39] Min Huang, Tao Zhang, Jingyang Wang, and Likun Zhu. A new air quality forecasting model using data mining and artificial neural network. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 259–262. IEEE, 2015.

[40] Yi-Ting Tsai, Yu-Ren Zeng, and Yue-Shan Chang. Air pollution forecasting using rnn with lstm. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1074–1079. IEEE, 2018.

[41] Qing Tao, Fang Liu, Yong Li, and Denis Sidorov. Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru. *IEEE access*, 7:76690–76698, 2019.

[42] Yue-Shan Chang, Hsin-Ta Chiao, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, and Kuan-Ming Lin. An lstm-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, 11(8):1451–1463, 2020.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[46] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 1(8), 2018.

[47] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[48] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.

[49] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[50] Helge Erik Solberg and Ari Lahti. Detection of outliers in reference distributions: performance of horn's algorithm. *Clinical chemistry*, 51(12):2326–2332, 2005.

[51] Glen Van Brummelen. *Heavenly mathematics: The forgotten art of spherical trigonometry*. Princeton University Press, 2012.

[52] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[53] S Pal Arya et al. *Air pollution meteorology and dispersion*, volume 310. Oxford University Press New York, 1999.

[54] Afshin Gholamy, Vladik Kreinovich, and Olga Kosheleva. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. 2018.

[55] Mateus Habermann, Monica Billger, and Marie Haeger-Eugensson. Land use regression as method to model air pollution. previous results for gothenburg/sweden. *Procedia Engineering*, 115:21–28, 2015.

[56] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[57] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Appendix A

Stationary Model Results - INHALE

Table A.1: Evaluation of spatial estimation models on the validation data

Model	**MAE** $(\mu g/m^3)$	**MAPE (%)**
Linear Regression	7.333	94.26
Ridge Regression	7.330	94.13
Extra Trees	5.375	60.70
Gradient Boosting Decision Trees	6.021	75.18
Random Forests	**5.335**	58.20
Feed-forward Neural Network	6.833	**52.77**
Fine-tuned Neural Network	7.279	53.14

Table A.2: Spatial ANN model hyperparameter tuning. The MAE and MAPE columns correspond to the evaluation results on the held-out validation set

Hidden Units	**Learning Rate**	**Dropout Prob.**	**MAE** $(\mu g/m^3)$	**MAPE (%)**
32	0.0001	0	6.915	50.50
32	0.0005	0	7.106	50.58
32	0.001	0	6.761	50.93
64	0.0001	0	6.833	52.77
64	0.0005	0	6.978	52.69
64	0.001	0	6.592	52.82
32	0.0001	0.5	7.140	49.62
32	0.0005	0.5	6.758	49.56
32	0.001	0.5	6.789	**49.49**
64	0.0001	0.5	7.364	50.75
64	0.0005	0.5	**6.585**	50.81
64	0.001	0.5	6.719	50.62

Table A.3: Hyperparameter tuning results for the LSTM spatio-temporal model. The MAE and MAPE columns correspond to the evaluation results on the held-out validation set

Dense Units	**Learning Rate**	**MAE** $(\mu g/m^3)$	**MAPE (%)**
32	0.005	6.347	43.04
32	0.01	6.157	42.74
32	0.05	6.021	42.72
64	0.005	6.018	42.43
64	0.01	5.902	**42.08**
64	0.05	5.849	42.47
128	0.005	5.917	42.46
128	0.01	5.873	42.17
128	0.05	**5.826**	42.45

A.0.1 Transfer Learning

Table A.4: Evaluation of spatial estimation models on the validation data for a 25% subset of INHALE dataset

Model	**MAE** $(\mu g/m^3)$	**MAPE (%)**
Linear Regression	14.12	104.70
Ridge Regression	14.16	104.97
Extra Trees	**8.093**	67.44
Gradient Boosting Decision Trees	8.625	72.08
Random Forests	8.719	70.95
Feed-forward Neural Network	9.763	53.720
Fine-tuned Neural Network	9.435	**48.49**

Table A.5: Evaluation of spatial estimation models on the validation data for a 50% subset of the INHALE dataset

Model	**MAE** $(\mu g/m^3)$	**MAPE (%)**
Linear Regression	1834040611631	37082938990993
Ridge Regression	6.817	71.96
Extra Trees	6.222	50.61
Gradient Boosting Decision Trees	**5.931**	46.91
Random Forests	5.956	49.22
Feed-forward Neural Network	6.629	**43.03**
Fine-tuned Neural Network	6.871	44.56

Appendix B

Stationary Model Results - DAPHNE

Table B.1: Evaluation of spatial estimation models on the validation data

Model	**MAE** $(\mu g/m^3)$	**MAPE (%)**
Linear Regression	66.34	215.11
Ridge Regression	66.33	215.11
Extra Trees	**44.14**	115.85
Gradient Boosting Decision Trees	53.70	142.37
Random Forests	49.61	134.78
Feed-forward Neural Network	57.27	**58.06**
Fine-tuned Neural Network	70.01	59.90

Table B.2: Spatial ANN model hyperparameter tuning. The MAE and MAPE columns correspond to the evaluation results on the held-out validation set

Hidden Units	**Learning Rate**	**Dropout Prob.**	**MAE** $(\mu g/m^3)$	**MAPE (%)**
32	0.0001	0	58.40	58.12
32	0.0005	0	56.82	57.74
32	0.001	0	59.19	58.14
64	0.0001	0	57.27	58.06
64	0.0005	0	**56.18**	57.49
64	0.001	0	56.43	57.30
32	0.0001	0.5	59.77	57.48
32	0.0005	0.5	58.28	56.69
32	0.001	0.5	58.30	56.86
64	0.0001	0.5	59.74	57.40
64	0.0005	0.5	57.85	56.48
64	0.001	0.5	58.02	**56.14**

Table B.3: Hyperparameter tuning results for the stacked LSTM temporal model. The MAE and MAPE columns correspond to the evaluation results on the held-out validation set

Dense Units	**Learning Rate**	**MAE** $(\mu g/m^3)$	**MAPE (%)**
32	0.005	70.09	55.16
32	0.01	68.74	54.68
32	0.05	67.76	55.21
64	0.005	68.10	54.58
64	0.01	67.15	54.58
64	0.05	66.48	**54.27**
128	0.005	66.50	54.41
128	0.01	**65.83**	54.70
128	0.01	66.44	54.47

B.0.1 Transfer Learning

Table B.4: Evaluation of spatial estimation models on the validation data for a 10% subset of the DAPHNE dataset

Model	**MAE** $(\mu\mathbf{g}/\mathbf{m}^3)$	**MAPE (%)**
Linear Regression	76.65	295.01
Ridge Regression	75.64	292.27
Extra Trees	57.51	162.49
Gradient Boosting Decision Trees	47.58	139.65
Random Forests	72.00	192.94
Feed-forward Neural Network	**29.29**	61.82
Fine-tuned Neural Network	33.40	**54.86**

Table B.5: Evaluation of spatial estimation models on the validation data for a 50% subset of the DAPHNE dataset

Model	**MAE** $(\mu\mathbf{g}/\mathbf{m}^3)$	**MAPE (%)**
Linear Regression	77.37	196.73
Ridge Regression	77.37	196.70
Extra Trees	**63.10**	117.58
Gradient Boosting Decision Trees	67.55	143.47
Random Forests	64.50	130.76
Feed-forward Neural Network	79.29	**60.93**
Fine-tuned Neural Network	82.79	63.59

Appendix C

Personal Model Results - INHALE

Table C.1: Evaluation of spatial estimation models on the validation data

Model	**MAE** $(\mu g/m^3)$	**MAPE** (%)
Linear Regression	6.462	120.27
Ridge Regression	6.462	120.27
Extra Trees	6.000	110.33
Gradient Boosting Decision Trees	5.685	105.35
Random Forests	6.312	115.92
Feed-forward Neural Network	**4.145**	**45.54**

Table C.2: Spatial ANN model hyperparameter tuning. The MAE and MAPE columns correspond to the evaluation results on the held-out validation set

Hidden Units	**Learning Rate**	**Dropout Prob.**	**MAE** $(\mu g/m^3)$	**MAPE (%)**
32	0.0001	0	**4.122**	44.97
32	0.0005	0	4.126	45.48
32	0.001	0	4.134	44.72
64	0.0001	0	4.139	45.47
64	0.0005	0	4.145	45.54
64	0.001	0	4.126	46.35
32	0.0001	0.5	4.255	42.83
32	0.0005	0.5	4.141	42.84
32	0.001	0.5	4.169	**42.27**
64	0.0001	0.5	4.233	42.52
64	0.0005	0.5	4.247	42.36
64	0.001	0.5	4.262	42.44

Table C.3: Hyperparameter tuning results for the stacked LSTM temporal model. The MAE and MAPE columns correspond to the evaluation results on the held-out validation set

Dense Units	**Learning Rate**	**MAE** $(\mu g/m^3)$	**MAPE (%)**
32	0.005	5.481	43.75
32	0.01	5.411	43.13
32	0.05	5.452	43.38
64	0.005	5.439	43.63
64	0.01	5.402	**43.04**
64	0.05	5.402	43.20
128	0.005	5.424	43.14
128	0.01	**5.401**	43.07

Appendix D

Personal Model Results - Guadalajara

Table D.1: Evaluation of spatial estimation models on the validation data

Model	MAE ($\mu g/m^3$)	MAPE (%)
Linear Regression	5.514	25.84
Ridge Regression	5.520	25.90
Extra Trees	6.637	27.47
Gradient Boosting Decision Trees	5.327	22.85
Random Forests	6.039	26.39
Feed-forward Neural Network	**5.067**	24.35
Fine-tuned Neural Network	5.232	**19.69**

Appendix E

Personal Model Results - Leon

Table E.1: Evaluation of the spatial estimation models on the validation data

Model	**MAE** $(\mu g/m^3)$	**MAPE (%)**
Linear Regression	9.985	31.56
Ridge Regression	9.973	31.54
Extra Trees	8.079	24.28
Gradient Boosting Decision Trees	8.795	33.29
Random Forests	10.24	39.78
Feed-forward Neural Network	8.593	29.00
Fine-tuned Neural Network	**7.321**	**20.84**