A Scalable Deep Learning System for Monitoring and Forecasting Pollutant Concentration Levels on UK Highways

Abstract

The construction of intercity highways by the government has resulted in a progressive increase in vehicle emissions and pollution from noise, dust, and vibrations despite its recognition of the air pollution menace. Efforts that have targeted roadside pollution still do not accurately monitor deadly pollutants such as nitrogen oxides and particulate matter. Reports on regional highways across the country are based on a limited number of fixed monitoring stations that are sometimes located far from the highway. These periodic and coarse-grained measurements cause inefficient highway air quality reporting, leading to inaccurate air quality forecasts. This paper, therefore, proposes and validates a scalable deep learning framework for efficiently capturing fine-grained highway data and forecasting future concentration levels. Highways in four different UK regions - Newport, Lewisham, Southwark, and Chepstow were used as case studies to develop a REVIS system and validate the proposed framework. REVIS examined the framework's ability to capture granular pollution data, scale up its storage facility to rapid data growth and translate high-level user queries to structured query language (SQL) required for exploratory data analysis. Finally, the framework's suitability for predictive analytics was tested using fastai's library for tabular data, and automated hyperparameter tuning was implemented using bayesian optimisation. The results of our experiments demonstrated the suitability of the proposed framework in building end-to-end systems for extensive monitoring and forecasting of pollutant concentration levels on highways. The study serves as a background for future related research looking to improve the overall performance of roadside and highway air quality forecasting models.

Keywords: Urban Air Pollution, Air Quality Prediction, Highway, Deep Learning, Big Data, Internet of Things

1. Introduction

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

Hence, there is a pressing and cogent need to find innovative and sustainable ways to monitor air pollutants and curb their devastating effects on health and human capital, as well as associated GDP losses (DEFRA 2020). According to (Alvanchi et al. 2020), monitoring particulate matter ($PM_{2.5}$, PM_{10}) and other highway pollutants like NO_2 is not a straightforward task because pollutants tend to decay and diffuse into the background concentration within 200m from the source. Furthermore, highway speed limits and traffic congestion complicate things further as they result in varying driving patterns such as sudden slow-downs and speedups, which elevate these pollution levels or limit their dispersion (Karner et al. 2010, Zhang & Batterman 2013). In response to the ever-increasing impacts of air pollution and its associated intricacies, the UK government has invested about £100 million to proactively tackle air quality (AQ) challenges to protect health and support clean growth (DEFRA 2019). However, despite these investments by the UK, the issue of how to proactively tackle and ultimately improve air quality across UK highways persists.

According to Barthwal & Acharya (2018), most countries monitor air pollution using stationary monitoring stations operated by government authorities. Figure 1 illustrates how the UK currently monitors highways to come up with its ultra low emission policies. Highways are monitored by Highways England (a government-owned company charged with operating, maintaining, and improving motorways in England) via its automatic urban and rural network (AURN), which collects sparse air pollutant data. However, evidence suggests that these air quality analysers are relatively heavy and expensive to install or maintain (Carullo et al. 2007, Barthwal & Acharya 2018). Therefore, it is impracticable for Highways Englands' monitoring stations to be deployed across the UK to capture pollutant concentration levels and improve air quality. On the other hand, low-cost/off-the-shelf IoT sensors that have been proposed in previous studies (e.g., Badura et al. (2018), Borghi et al. (2018), Budde et al. (2018)) for monitoring air quality are plagued with interference issues from weather, cross-sensitivities between pollutants and ageing effects of integrated circuit technology (Karagulian et al. 2019). These limitations are coupled with the fact that it would take a significant and environmentally unfriendly investment to install low-cost IoT sensors across every road in the UK.

Asides from inefficient highway air quality monitoring, another major challenge rests on the issue of how data disparity and isolated data sets affect the accurate prediction of pollutant concentration levels. Evidence suggests that data sources that could collectively predict pollutant concentration levels (e.g., historic pollution, GIS location data, traffic flow, weather and background pollution) exist in silos, thus resulting in numerous integration and data processing issues (Umadevi & Geraldine Bessie Amali 2020). In addition, existing forecasting methods are riddled with computational issues like scalability and memory demand which limits their optimal adoption (Zhang et al. 2012). For instance, Alléon et al. (2020) and Lee et al. (2020) attempted to develop large-scale air quality forecasting systems. However, the authors highlighted the inability to integrate additional granular data and insufficient computational power as the shortcomings of their research.

On the back of significant advancements in scalable machine learning (ML) approaches such as deep learning, which are known to thrive on huge data (Akinosho et al. 2020), this study, therefore, proposes a scalable deep learning framework for monitoring and forecasting pollutant concentration levels on UK highways. This framework leverages internet of things (IoT) sensors for real-time monitoring, graphics processing units (GPUs) for parallel computing, big data for scalable storage and deep learning for forecasting highway pollutant concentration. In the design of a system that implements the proposed framework, the following objectives were set for this study:

- Develop, calibrate, and deploy energy efficient hardware devices with practicable accuracy to capture real-time pollution data on four different UK highways.

- Integrate missing or inaccurate data from heterogeneous sources to enhance forecasting accuracy of the developed model.

- Develop and evaluate a baseline deep learning model to make hourly predictions of PM2.5, PM10 and NO2 concentration levels due to the deadly nature of these pollutants.

- Perform a system scalability test to determine response time and throughput as hardware device load increases.

This manuscript is structured as follows: Immediately after introduction follows a section that summarises the research methodology adopted to achieve the highlighted research objectives while section 3 highlights features of the proposed framework. Afterwards, the development process of a prototype system that implements the proposed framework is discussed in section 4. The discussion also includes the scalability test that was performed on the system. Section 5 correlates the findings of this study with existing research and highlights its relevance to practice. Finally, conclusions are drawn, and future research directions are indicated in Section 6.

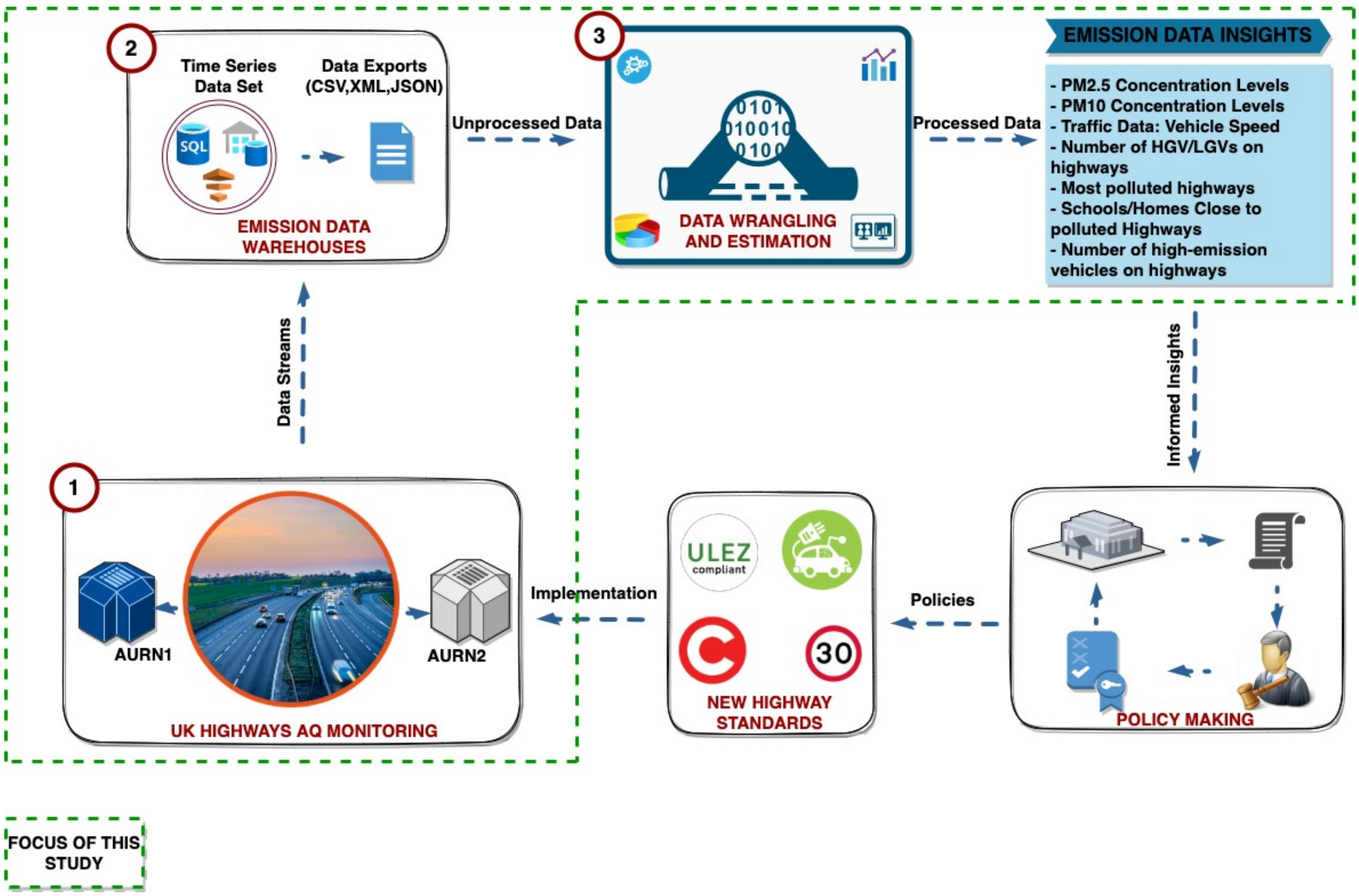


Figure 1: The current situation of highway AQ monitoring in the UK. This study seeks to address three main challenges which include: 1) expensive cost of deploying monitoring stations such as the AURN on highways 2) Data silos/segregated data operated by different data agencies 3) Inefficient air quality estimation methods

2. Research Methodology

A combination of experimental design with case-study methodology was adopted to achieve the identified research objectives of this study. The rationale behind this approach is based on the need to fulfil two principal goals. One, system implementation is needed to demonstrate the practicality of the proposed framework using a developed system. This approach reflects an experimental design method of research. On the other hand, the adoption of case-study strategy is to test the results of the developed system in disparate real-life environments. This approach is quite effective and has been utilised in related research involving multiple case-studies (Chen et al. 2015, Zheng et al. 2015). The experimental design approach was also used to conceive the layered architecture of the proposed framework with each layer addressing at least one research objective. Layering remains a prevalent application design technique that allows the disintegration of a complex software system into modules. Layers within the proposed framework consist of libraries, programming languages, and services required for monitoring and forecasting.

A careful market analysis of "grey literature" and an extensive review of academic publications revealed several sensors suitable for designing monitoring units to address the first research objective. Google scholar and scientific databases such as Scopus and ScienceDirect were also used to search for academic publications, while Google's search engine revealed additional sensor manufacturers. We limited our search to the three pollutants of interest in this study - NO_2, $PM_{2.5}$ and PM_{10}. A similar search of relevant integration libraries and big data frameworks informed the framework's approach to solving integration and data storage challenges. From the array of available options, enterprise frameworks that allow the integration of data from legacy as well as newly built systems were selected. There are a lot of algorithms that are available for air quality forecasting. However, it was important to choose a scalable machine learning approach such as deep learning that has shown significant promise using distributed computing clusters (Sergeev & Del Balso 2018, Chen et al. 2019).

3. A Proposed Deep Learning Framework for Highway AQ Monitoring and Prediction

The proposed framework is a four-layered architecture composed of the hardware layer, data storage layer, integration layer and analytics layer as depicted in Figure 2. This section introduces these layers and their functionalities.

3.1. Hardware Layer

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

3.2. Data Storage Layer

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

3.3. Integration Layer

The data integration layer ingests data from third-party sources into a central repository. The layer handles this data ingestion using the extract, transform and load (ETL) process. External data can include pollution data captured by other monitoring stations, highway geographical data, meteorological data and traffic data. The essence of this layer is to ensure that data not captured in the hardware layer by the monitoring devices can be integrated into the system to improve the performance of developed estimation models. If the suggested

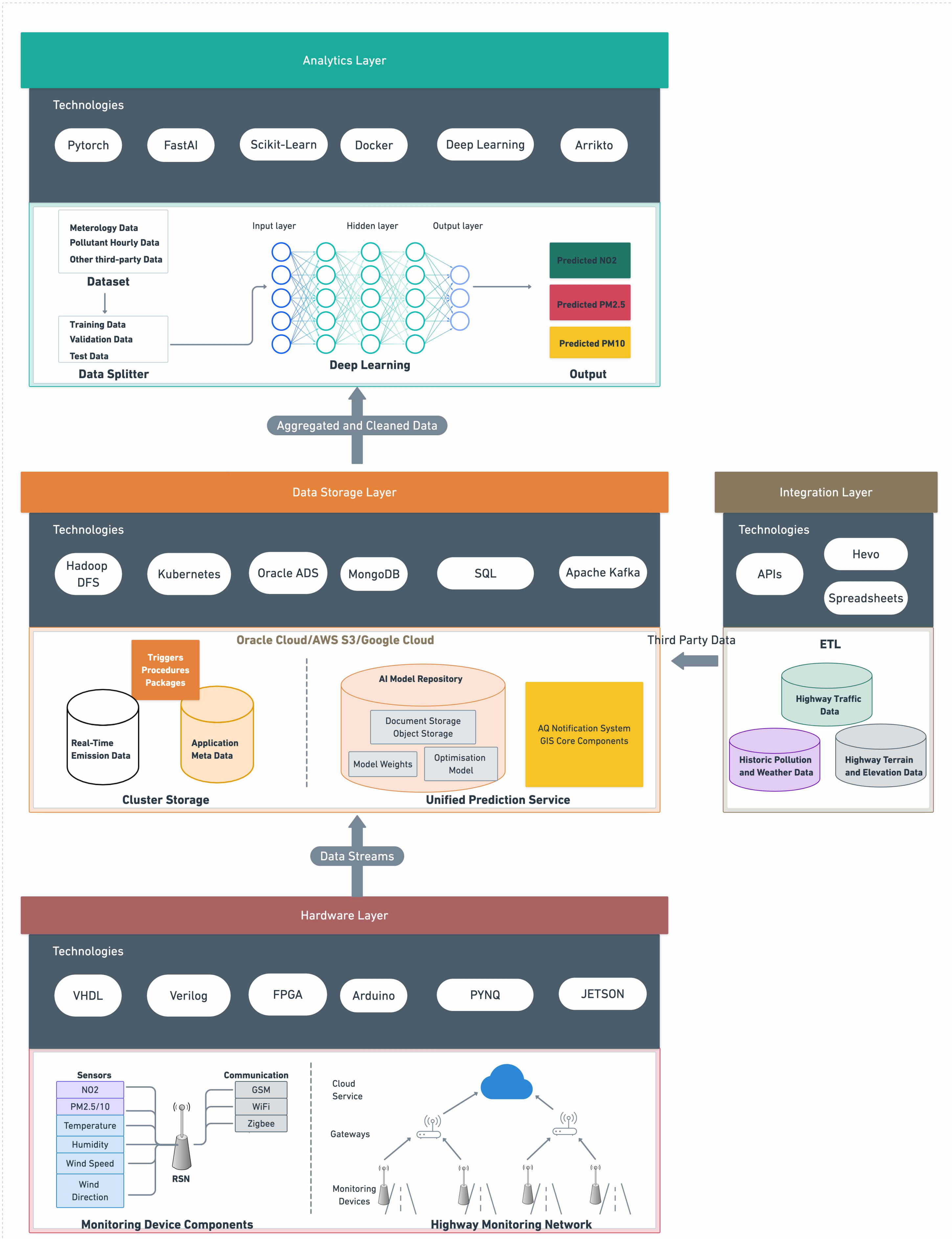


Figure 2: Deep Learning Framework for Highway Air Quality Monitoring and Prediction

functionalities of the hardware layer are too expensive to implement, this layer can grab open-source or paid data from available online sources. Data can be downloaded in different formats such as TXT, JSON, XML and CSV or exposed as external links. The data from this layer should be stored as separate tables in the data storage layer for unique identification and also to avoid mix-ups with existing data.

3.4. Analytics Layer

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

4. Development and Deployment of the REVIS System Prototype

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

4.1. REVIS Highway Monitoring Devices

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

4.1.1. REVIS Device Development and Evaluation

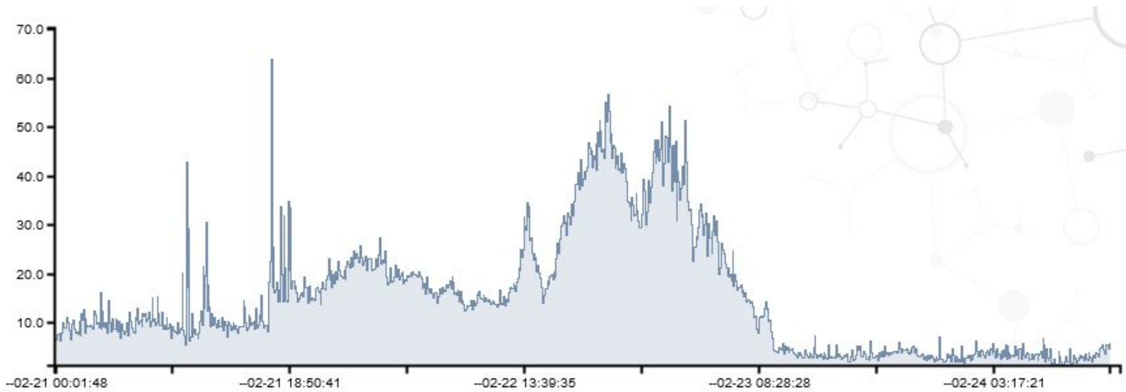
REVIS demonstrates the hardware layer through the development and calibration of devices with built-in sensors to measure the atmospheric composition of NO_2, $PM_{2.5}$ and PM_{10}, alongside weather parameters - pressure, temperature and relative humidity. Table 1 below summarises details of manufacturers of the chosen sensors and their accuracy figures. Each REVIS device required an excellent design of both analogue and digital circuitry around it and several stages of calibration. The Alphasense NO_2 sensor for example, showed during experimentation that it was best suited for fixed sensing installations and urban air monitoring since varying meteorological conditions had a significant influence on it's readings. The sensor's cross-interference with the $PM_{2.5}$ SPS30 sensor and detection range limits (DRL) were also evaluated using equation 1

$$DRL = 3.3\sigma/S \tag{1}$$

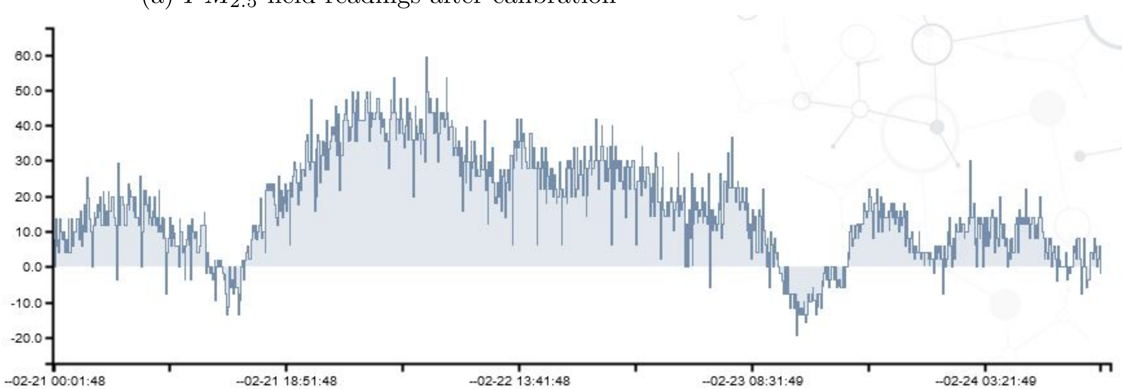
where S denotes the calibration curve's slope, and σ denotes the standard deviation of the sensor response in the absence of air (Shrivastava et al. 2011). The nearest AURN stations to the monitoring devices were identified for field evaluation. The selected stations were deemed suitable for calibration since they were close to deployed sensors and mainly provided missing weather data and also average hourly measurement of the pollutants of interest. Data from the REVIS devices were averaged over an hour for appropriate comparison with the reference data. Figure 3 shows $PM_{2.5}$ and NO_2 readings on one of the REVIS devices after calibration. Aside from the occasional underestimated measurement of the NO2 sensors, other sensors such as $PM_{2.5}$ and PM_{10} showed close estimates to the reference measurements with correlation coefficient r $>$ 0.8.

Table 1: Sensor Specifications and Accuracy

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



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



(b) NO_2 field readings after calibration

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

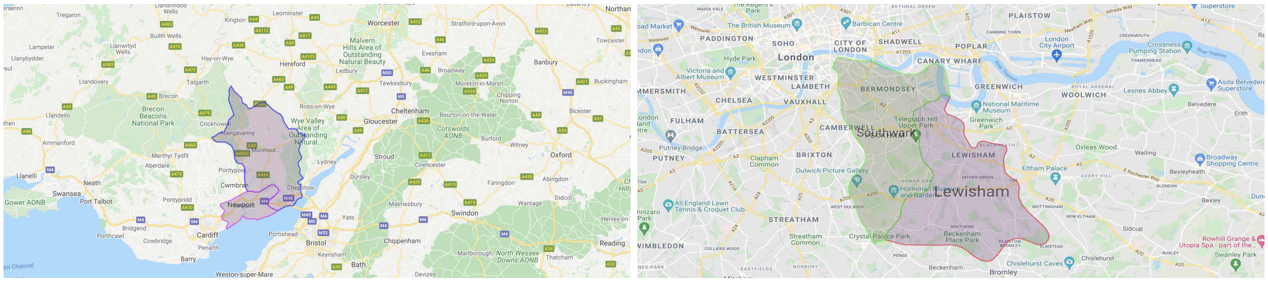
4.1.2. Device Deployment in Case Study Regions

Major highways in London, Newport and Chepstow were chosen as case studies for this research. The city of London is made up of 9 million inhabitants which includes 4.49 million males and 4.51 million males (ONS 2020). With 74.9% of this population belonging to the working age 16-64 years (ONS 2021), the government faces the challenge of addressing traffic congestion hurdles across the city. A survey reported in TFL (2019) showed that 59% of Londoners tend to use the bus at least once a week, with car passenger commuting more popular among the younger populace. Newport and Chepstow are located in the south eastern region of Wales. According to ONS (2018), 1.53 million residents live in the region with a population density of 546 person/km_2. The region also has the highest and lowest life expectancy figures across wales with Monmouthshire boasting the highest life expectancy. 74.3% of the employed populace prefer to journey by motorcycle,van or car while 8.8% choose to travel by bus or train (Statswales 2020). Major highways in the region such as the M4, A48 and A466 highways connect neighbouring cities.

Figure 4 depicts the distribution of REVIS devices in these cities. In London, twelve devices were distributed on sections of the A302, A2209 and A1203 highways. One device was placed 92.79m from Junction 25 of the M4 highway in Newport and another device was positioned close to The A48 motorway in Chepstow. The deployment approach that was adopted during the distribution of these devices ensured three key requirements: (1) sufficient highway length (2) cellular data connectivity and (3) electrical/solar power availability. It was also necessary that device installation required minimum technical skills and data was captured for a minimum of 6-8 months.

4.2. Exploratory Analysis of Pollution and Weather Data

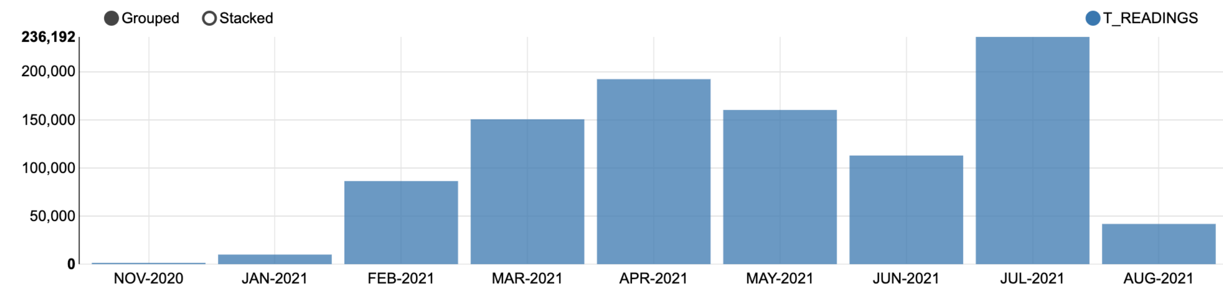
It is important to verify data consistency before commencing model training in ML regression tasks such as the one being considered. The minimum recommendation is to confirm the total number of rows and columns within the data, as this may have been compromised during data transfer (Bilal & Oyedele 2020). This section analyses the impact of weather parameters and the case-study region on pollutant levels. Although data was captured between November 2020 and August 2021, missing data in the early stages of deployment (shown in Figure 5 below) influenced the decision to analyse data between February 2021 and August 2021 when missing data was minimal.



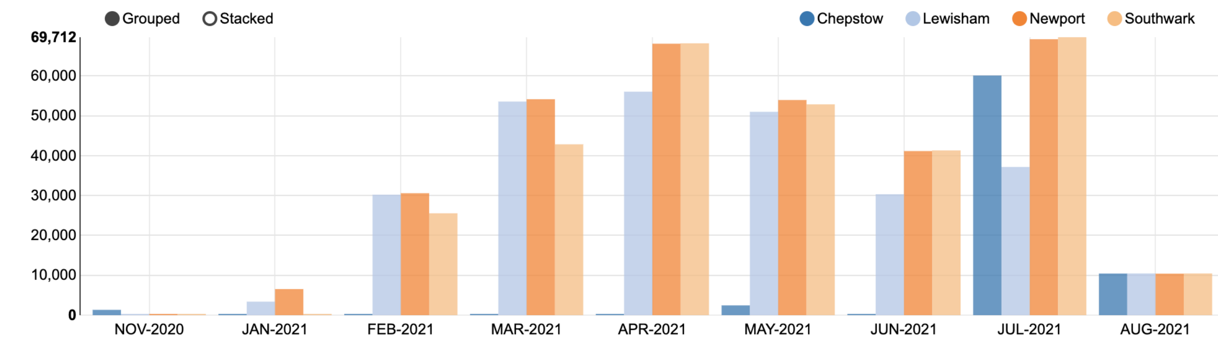
(a) Map of regions where case-study highways are located.

(b) Sensor distributions across the highways

Figure 4: Maps showing the distribution of 14 REVIS devices in four regions across the UK - Newport(1), Chepstow(1), Lewisham (6), Southwark(6). The devices in London were deployed to capture readings from the A302, A2209 and A1203 highways, while the devices in Newport and Chepstow were deployed close to the M4 and A48 highways, respectively



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



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

Figure 5: Total monthly readings captured by deployed sensors between November 2020 and August 2021. These plots illustrate the amount of missing data in the first two months when some devices were offline. Chepstow had the lowest monitored readings overall.

4.2.1. **The Impact of Weather on $PM_{2.5}$, PM_{10} and NO_2**

Weather parameters influence the dispersion rates of pollutants (Barrera-Animas et al. 2022). It is worthwhile to first check the correlation between the weather parameters before investigating the impact of weather on highway pollution. Figure 6 illustrates a correlation matrix constructed to identify the hierarchical similarities between these parameters which revealed a strong correlation between temperature, wind speed and wind direction. To understand the effects of temperature on the four pollutants, the seasonal trends were plotted as shown in Figure 7. The average temperature for all four regions ranged between 8.6 and 12.56$°C$ in Winter, 9.73 and 19.76$°C$ in spring and 19.41 and 21.78$°C$ in summer. A regression analysis of temperature against each pollutant as presented in Table 2a depicts a positive correlation between $PM_{2.5}$ and PM_{10} and temperature in Newport, Southwark and Lewisham for spring and summer seasons. Chepstow had no correlation calculated for winter when there was no temperature reading recorded and a negative correlation in spring and summer. NO2 had a negative correlation with temperature in all of these regions in winter and spring but had a positive correlation in Southwark and Lewisham in Summer. These findings corroborate studies that suggest that concentration levels are highest when the temperature is high (Pearce et al. 2011, Analitis et al. 2014).

Figure 6: Distance matrix of weather parameters using Pearson's correlation. A strong correlation can be noticed between "temp","temp_min","temp_max","wind_speed", "wind_degree" and "feels_like". There is also a discernible correlation between "clouds_all" and "humidity"/"windspeed".

Table 2: Regression analysis of weather parameters vs pollutant concentration

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

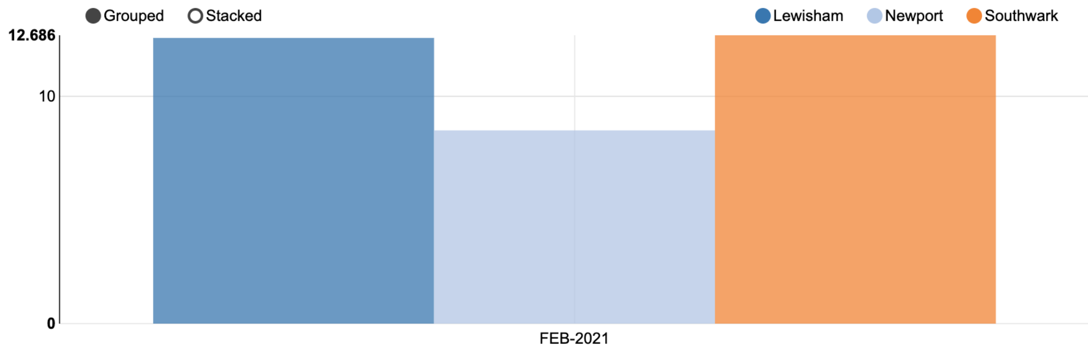
(a) Correlation between regional temperature and pollutants in spring, winter and summer

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

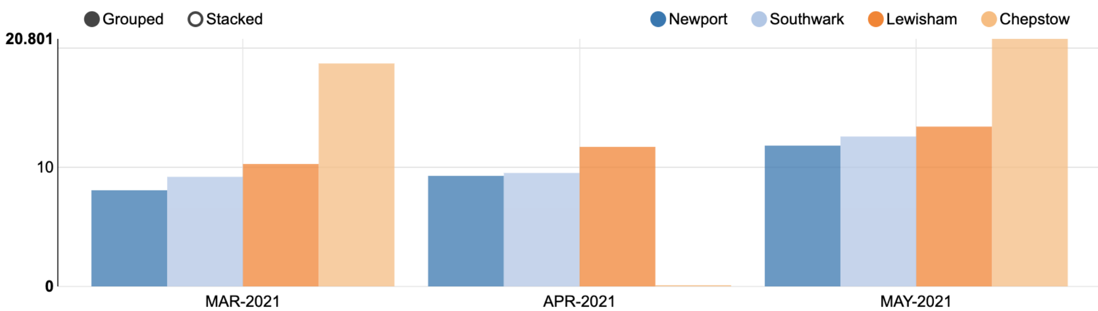
(b) Correlation between regional pressure and pollutants in spring, winter and summer

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

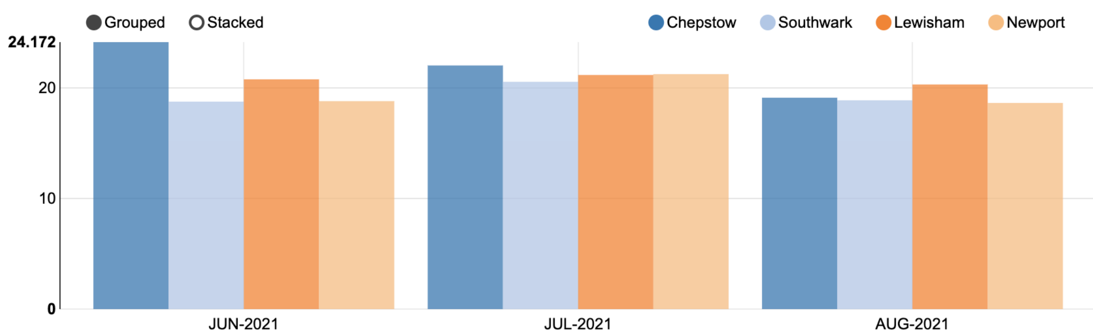
(c) Correlation between regional humidity and pollutants in spring, winter and summer



(a) Average winter temperature for all four regions

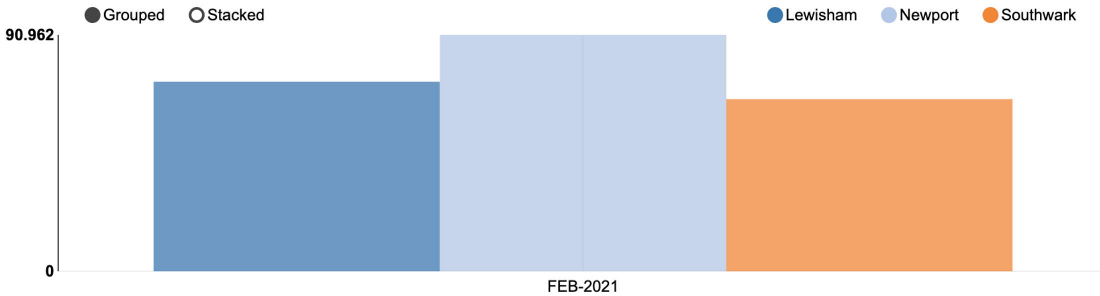


(b) Average spring temperature for all four regions

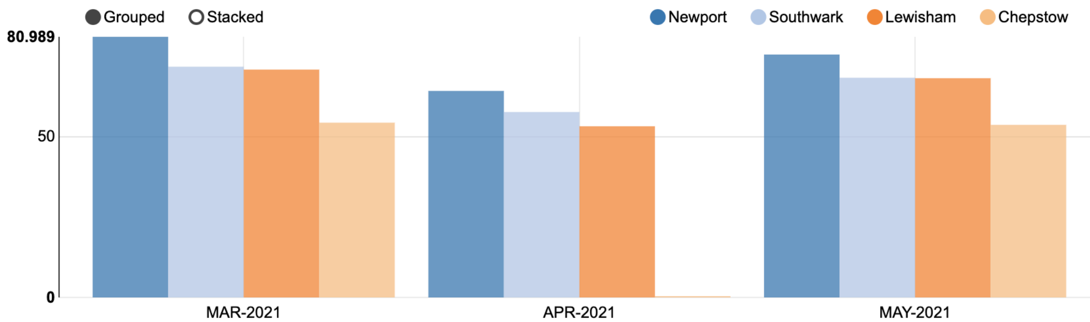


(c) Average summer temperature for all four regions

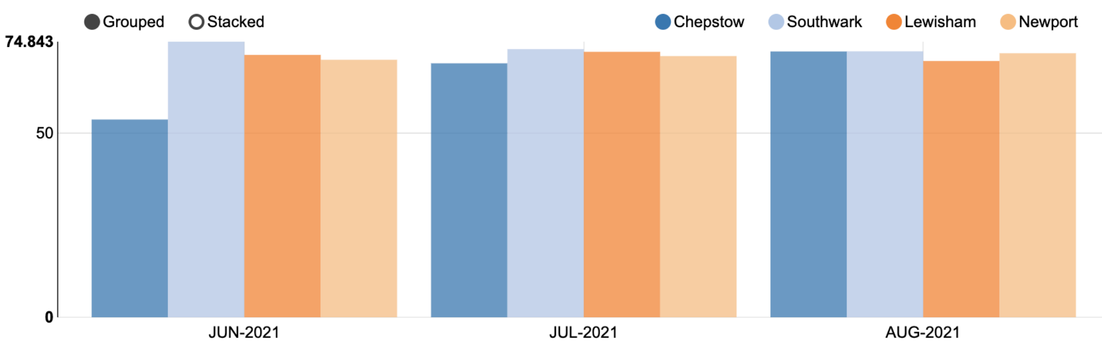
Figure 7: The seasonal trends for temperature in Newport, Southwark, Lewisham and Chepstow. Newport has the lowest temperature of $8.6°C$ in winter as there was also no reading recorded for Chepstow, as illustrated in plot (a). Chepstow had the highest average temperature of $19.76°C$ in spring and $21.78°C$ in summer, as shown in plots (b) and (c)



(a) Average winter humidity for all four regions



(b) Average spring humidity for all four regions



(c) Average summer humidity for all four regions

Figure 8: The seasonal trends for humidity in Newport, Southwark, Lewisham and Chepstow. Similar to temperature and pressure, no reading was captured for Chepstow in winter. However, the region recorded the least humidity of 53.95% in spring, as illustrated in plot (b). Newport had the highest average humidity of 90.96% in winter and 73.54% in spring, as shown in plots (a) and (b)

For pressure, the lowest readings were recorded in Chepstow during Spring and Summer seasons while Lewisham and Southwark recorded the highest pressures in spring. Table 2b summarises the pressure readings during these seasons and the correlation figures with the pollutants. The $PM_{2.5}$ and PM_{10} concentrations in Newport and Chepstow were positively correlated with pressure, indicating that an increase in atmospheric pressure will increase the concentration levels of these highway pollutants. All three pollutants negatively correlate with pressure in Southwark and Lewisham in spring but positive in winter and summer. The conclusion drawn from this result is a strong correlation between pressure and $PM_{2.5}$ and PM_{10} but a significant negative correlation with NO_2. Figure 8 illustrates the average seasonal humidity across the regions with the lowest humidity value was recorded in Chepstow during summer and the highest in Newport during winter. It can be deduced from Table 2c that the three pollutants were negatively correlated with humidity for winter, spring and summer seasons. In particular, particulate matter ($PM_{2.5}$ and PM_{10}) are prone to be absorbed in the atmosphere as humidity increases. Naturally, rain results in higher relative humidity and soaks up these particles, resulting in a lower level of particulate in winter. (Odat 2009).

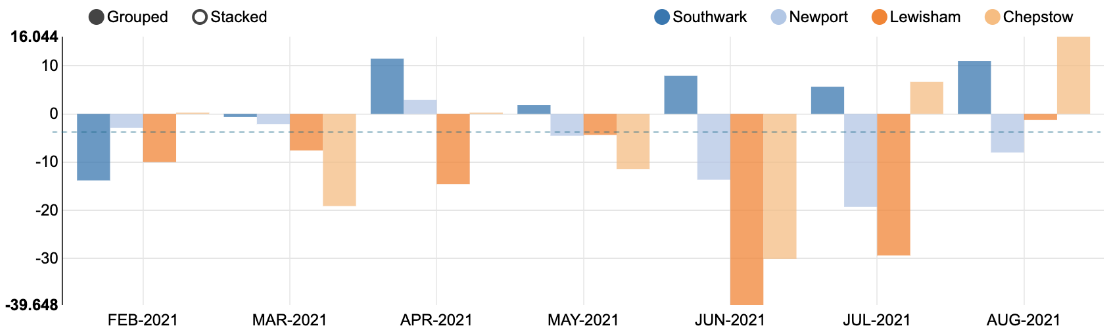
4.2.2. The Impact of Region on $PM_{2.5}$, PM_{10} and NO_2

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

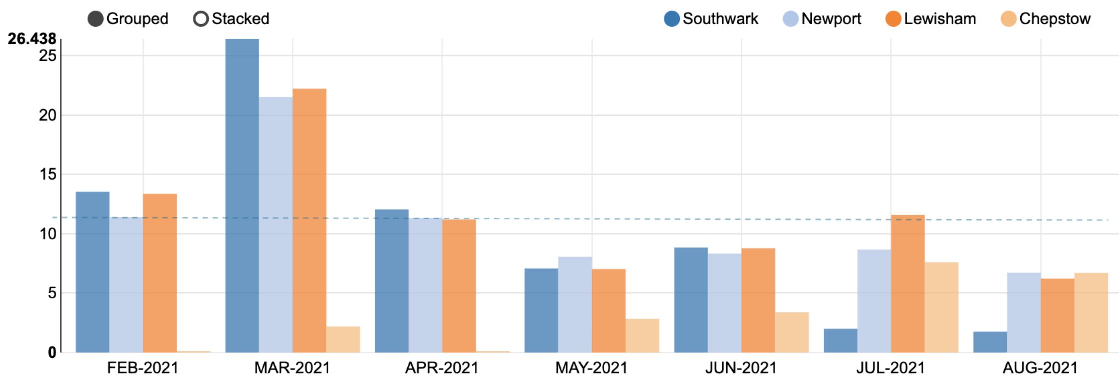
Table 3: Pollutant summary statistics based on region

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

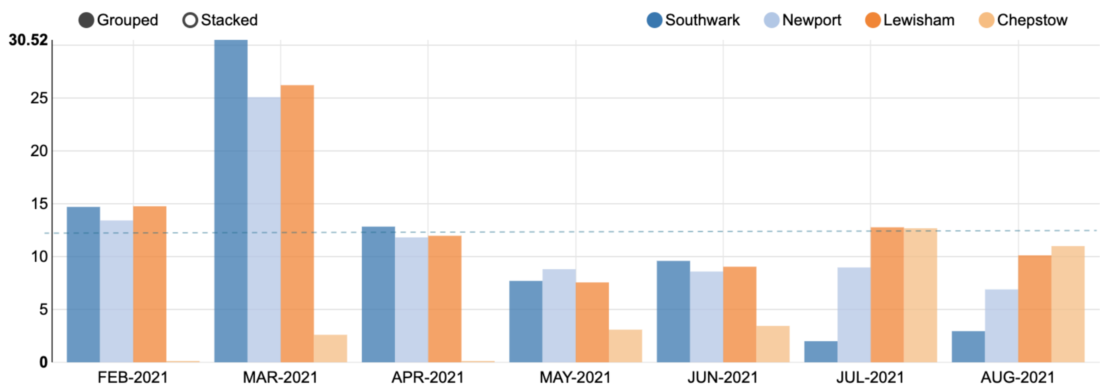
Table 3 shows that the average concentration levels across regions vary significantly, and this can be linked to the calibration accuracy of the sensor devices. Chepstow and Southwark seemed to have the most practical NO_2 averages, with Southwark having the highest. Lewisham has the highest $PM_{2.5}$ and PM_{10} average of $12.42 \mu g/m^3$ and $13.98 \mu g/m^3$, respectively. This analysis and the plot in Figure 9 reveal some prevalent calibration issues within the recorded values, which were sometimes exaggerated, as in the case of the maximum values for $PM_{2.5}$ and PM_{10}. Nevertheless, a one-way ANOVA variance test carried out to check the variance in NO_2, $PM_{2.5}$ and PM_{10} by region resulted in p values of 2.36e-4, 1.45e-3



(a) Monthly NO_2 average for all four regions



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



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

Figure 9: Plots highlighting the varying monthly averages for the three monitored pollutants. These averages varied significantly and are an indication that some influential factors may have affected the concentration levels

and 1.68e-4, respectively. This result indicates that the impact of regions on the concentration levels of these three pollutants is notable.

4.3. Forecasting Model Training and Evaluation

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

4.3.1. Meteorology Data Integration and Dataset Pre-Processing

Weather data such as wind speed and direction, precipitation, visibility, pressure, cloud cover, dew point, and wind gust which were not captured by the REVIS devices, were integrated from OpenWeather. Also, ozone data from the AURN stations were integrated into the dataset to be analysed and used for training. These integration exemplify the integration capabilities of the framework while enriching the data needed to train an estimation model. Appendix A presents a complete list of the columns, their description and data types before processing. An SQL procedure for automatically generating SQL codes such as the one illustrated in Figure 10 was implemented to summarise the pollution data. This generated hourly, 3-hourly and 6-hourly summaries of the pollutant concentration levels with the aim of capturing periodicity within the training data.

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

4.3.2. Validation Set Creation and Training Architecture

Model training is typically initiated by splitting the dataset into training, test and validation datasets. As the name implies, training data is used for training, while validation data is used for selecting the model that works best after verification using the test data. It is customary to randomise the dataset before splitting when there is a class imbalance - stratification; but since this problem is similar to a time-series problem where the date order is important, the validation and test sets can not be randomly selected. The common practice is to select the last few weeks or months of the dataset for validation and testing. Our dataset of 991662 rows and 34 columns had no class imbalance for the three pollutants which meant that stratification was not necessary. An experiment with different numbers of last days was carried out to determine the best validation approach, and the last 45 days of the dataset from July and August were eventually chosen for validation(15 days) and test(30 days). Fastai's "TrainTestSplitter" class was used to implement this division.

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

Figure 10: Auto-SQL generation to pre-process the dataset. An SQL command which generates 3-hour and 6-hour pollutant averages from the preceding readings is depicted.

Suitable optimisers, loss functions and activation functions had to be selected from an array of available options. Series of experimentation were carried out on popular optimisation functions such as *SGD, RMSProp, LAMB, LARS and Adam* and regression loss functions like *BCELossflat, MSELossFlat and L1LossFlat* before deciding the most suitable. Eventually, *Adam* optimiser and *MSELossFlat* were chosen for model training. *Bayesian-optimization* library was used to test and optimise the number of architecture layers, the size of each layer and dropout rates for the network. The final architecture used to train the model was made up of 14 embedding layers, 3 dropout layers, 3 batchnorm1d layers, 3 linear layers and 2 ReLU activation functions. The embedding layer was adopted for improved performance as inspired by the architecture proposed in Guo & Berkhahn (2016). Finally, the learning rate finder (*lr_find*) function of *TabularLearner* class was used to determine the best learning rate to be used for training. This resulted in a minimum value of $2.5e^{-4}$, and steep value of $1.3e^{-4}$. Figure 11 below shows the plot of the learning rate against the loss. Experts recommend selecting the learning rate at the point where the plot starts to dip. (i.e., $10^{-}4$).

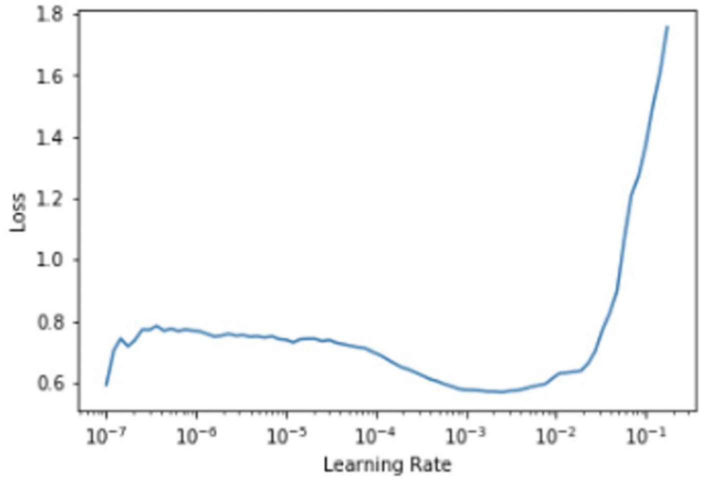


Figure 11: The model's training loss against the learning rate to determine the appropriate learning rate. The learning rate was fixed at the point where the plot started dipping (i.e., $10^{-}4$)

329 *4.3.3.* **Model Evaluation**

In this section, the results of the deep learning model developed are presented. The model was trained to make day-ahead predictions of the three pollutants, but first, an appropriate evaluation metric had to be selected. The top metrics for regression problems are mean squared error/root mean squared error(MSE/RMSE), mean absolute error(MAE) and R Square. The fastai library has two variants of RMSE: *rmse* and *exp_rmse*. The mean absolute error and root mean squared error (exp_rmse variant), defined as shown in equations 2 and 3 below, were selected as the metrics for evaluating the developed model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2} \tag{3}$$

Figure 12 illustrates the model training and validation losses after 20000 epochs. It is noteworthy that the training loss gradually as the number of epoch increased. The validation loss took a slightly different pattern and dropped significantly after 2500 epochs but became steady for the remaining training epochs. The final MAE and exponential RMSE after training were 0.350 and 1.591 respectively. Figure 13 captures the actual NO_2 concentration levels (highlighted in blue) and the model's day ahead prediction(highlighted in red). The difference in the model's predicted NO_2 and actual values is slight, and the predicted values were close to the actual.

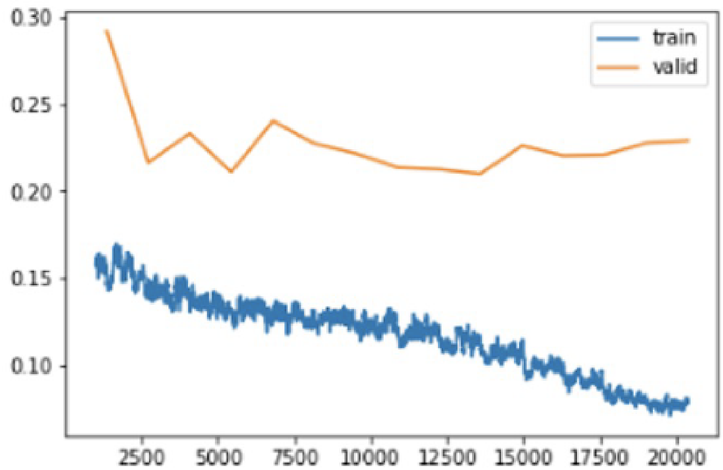


Figure 12: A plot showing the model's training and validation losses against the number of epochs. It is worth noting that there was a gradual decrease in both losses as the training epochs increased which indicates that the model was learning. Further training beyond 20000 epochs would have either resulted in overfitting or no further drop in both losses

```
learn.show_results()
```

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

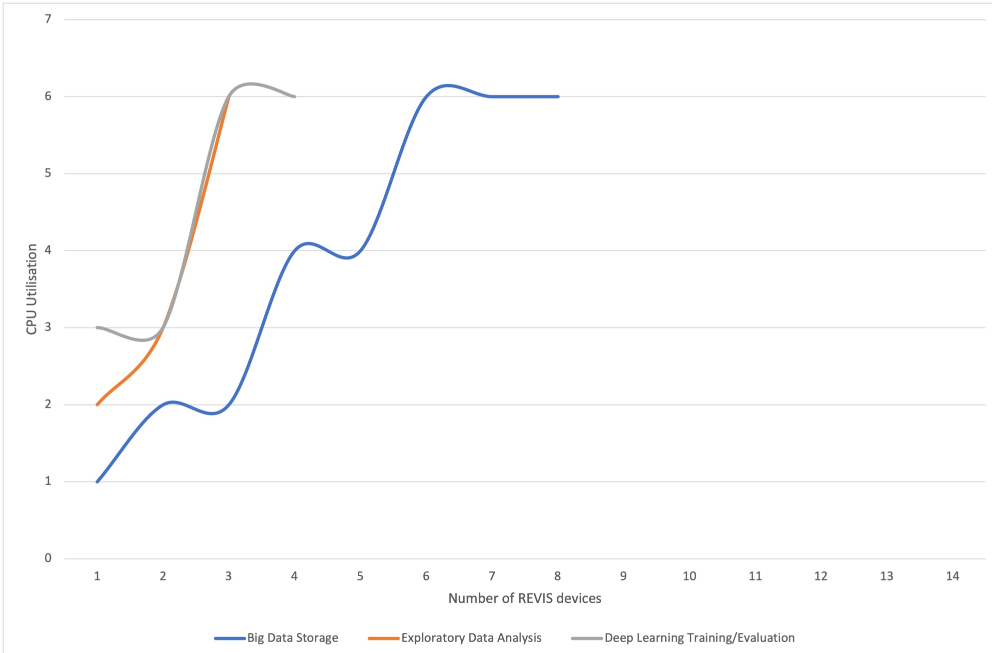
Figure 13: An illustration of captured NO_2 pollutant readings (blue highlight) and the deep learning model predictions (red highlight). These results were derived from an evaluation using the validation dataset. It should be pointed out that the model's predictions are not too far off the actual readings.

4.4. Evaluating the Scalability Performance of the REVIS System

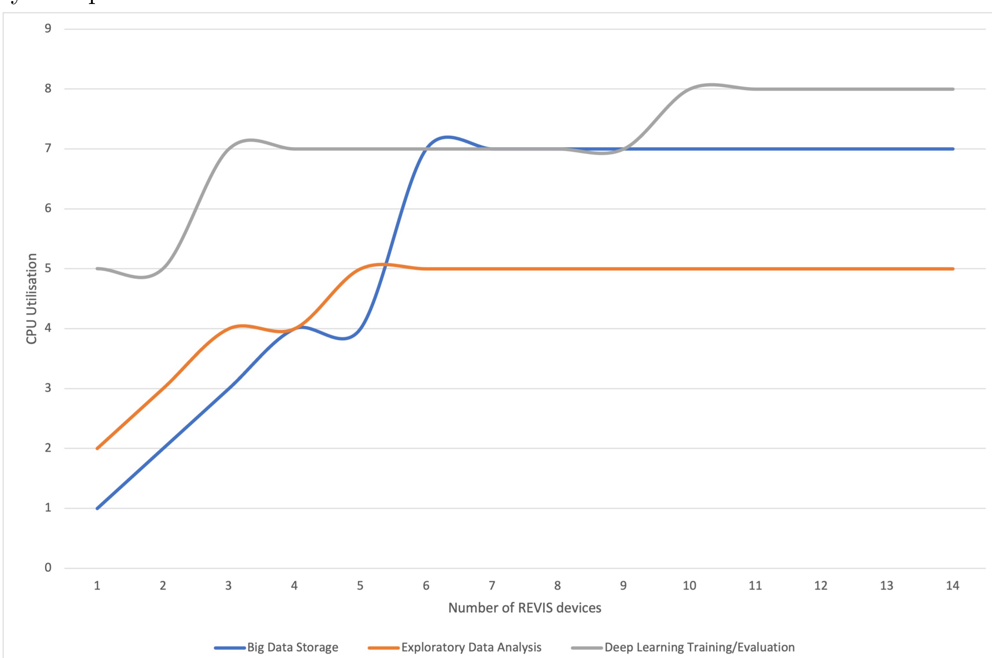
The REVIS system was tested for scalability using the IoT asset monitoring tool and database performance hub of two different oracle cloud instances. The fourteen REVIS devices were deployed sequentially to capture both system's response time and throughput. The first experiment was run on a bare metal cloud instance with specifications as shown in Table 4. Figure 14a shows the performance of this cloud instance as it could not scale past 8 devices and exploded at 3 and 4 devices for EDA and deep learning analysis. However, the GPU cloud instance performed better due to its auto-scale feature. Figure 14b shows a plot of the CPU cores utilised for exploratory data analysis, data storage and deep learning analysis as the number of deployed devices increased. It can be observed that the number of CPU cores increased gradually for each task and then stabilised at some point. The system was able to scale up its resources according to the computation/storage requirements. For the database performance, the test was run between November 2020 and Jan 2021 on the GPU instance and evaluated for utilisation, execution count, number of running statements and number of sessions metrics as shown in Figure 15. The maximum GPU utilisation was under 20% even with over 1.5 million execution queries.

Table 4: Hardware specifications of the two oracle cloud instances used to test scalablity

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



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



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

Figure 14: Plots of bare metal vs GPU instance as number of devices increased

(a) Metrics showing system's performance after the deployment of 14 sensors

Figure 15: Plots of scalability metrics showing database performance as the number of devices increased

5. Discussion and Implication of Study

The REVIS system was used to demonstrate the possibility of optimising the cost, efficiency and environmental impact of hardware IoT devices through the development, calibration and deployment of monitoring units to capture real-time pollution data on highways. The devices were developed through an excellent design of both analogue and digital circuitry around it and an iterative approach of calibration and performance optimisation. Although we were able to address the energy interference and cross-sensitivity issues of existing sensing devices, the developed units still had some NO_2 data inconsistencies which were directly linked to the chosen pollutant sensor. A probable solution is the adoption of machine learning techniques for sensor data calibration. This technique is increasingly becoming popular and has been explored in studies such as that of Zimmerman et al. (2018) and Si et al. (2020). Nevertheless, our implementation still demonstrates the hardware layer of the proposed framework and also the effectiveness of carefully designed low-cost and environmentally-friendly sensors in capturing and processing accurate data on highway air quality.

It is important to note that sensors data alone are not sufficient for ensuring accuracy in air quality forecasting models. There are a number of air quality data sources, which exist separately but can provide better insights about air quality if well explored and integrated. An important aspect of this study is to integrate missing or inaccurate data from heterogeneous sources to enhance forecasting accuracy of the developed deep learning model. The essence of this layer is to ensure that data not captured in the hardware layer by the monitoring devices can be integrated into the system to improve the performance. Similarly, an exploratory analysis on the captured and integrated data was conducted to evaluate the impact of different parameters on pollutant concentration. It is well established in literature that weather parameters such as rainfall and temperature influence the dispersion rates of pollutants (Barrera-Animas et al. 2022). Hence, there is need for a more coordinated approach such as the one proposed in this study to manage multiple data sources, which are relevant for accurately forecasting air quality on highways through common data environment and data integration.

Finally, this study set out to develop and evaluate a baseline deep learning model to make hourly predictions of $PM_{2.5}$, PM_{10} and NO_2 concentration levels. The problem is modelled as a structured data with additional features added to extend a typical time-series problem. This method allows us to explore entity embeddings for the categorical features. The performance of the baseline model on individual pollutant concentration is good, thereby suggesting that this approach of modelling is practicable. An improved forecasting model is directly applicable to highways where air quality sensing devices are not available but data on other features such as traffic flow and weather are captured. These models can then be used as a substitute to estimate the air quality on these highways. The scalability test performed on the REVIS system indicated it was able to scale up its resources according to the computation/storage requirements. This study has addressed an important aspect of air quality management on highways, provided a scalable solution for academics and industry practitioners; and a pathway for policy makers and highway regulators to make more informed decisions.

6. Conclusion

A cost-effective deep learning framework for ubiquitous monitoring and predicting pollutant concentration levels on UK highways was proposed in this study. An implementation of the framework was demonstrated using the REVIS system. Details of the development of the REVIS IoT hardware for data collection, the configuration of big data tools for data storage and the and results of trained deep learning forecasting models were reported. The scalability feature of the framework was also highlighted using two cloud instances with different computational resources. This study showed that real-time monitoring and forecasting could be achieved with the right computational resources. Although the scope of the research was limited to NO_2 $PM_{2.5}$ and PM_{10} pollutants and evaluation using deep learning, future research can focus on investigating other pollutants such as CO_2, SO_2 and Ozone as well as other machine learning approaches for estimation. This study is a part of a series of publications highlighting research findings on the REVIS project. This is just a scratch on the surface of our research outputs, as other articles will elaborate on developing more advanced deep learning models. Future research will integrate other relevant highway attributes such as traffic flow, highway terrain, background concentration, and pollutant characteristics such as washout coefficient, dispersion rate, and emission, critical to developing highway air quality estimation models.

Acknowledgement

The authors would like to express their sincere gratitude to InnovateUK (Grant Application No 10137 and File No 104367) and EPSRC(Grant Ref No EP/N509012/1) for providing the financial support for this study.

Appendix A: Data summary for pollutant estimation before processing

S/No	Column	Column Description	Non-Null Count	Data type
1	city_name	The name of the city of interest	991662 non-null	object
2	lat	The geographic coordinate of the city of interest (Latitude)	991662 non-null	float64
3	lon	The geographic coordinate of the city of interest (Latitude)	991662 non-null	float64
4	date	The observation time to include date, time, hour and second	991662 non-null	datetime64[ns]
5	rain_desc	Description of measured precipitation	5975 non-null	object
6	rain_1h	Integrated average hourly precipitation measurement (mm)	5658 non-null	float64
7	rain_3h	Integrated precipitation measurement averaged over 3 hrs preceding the observation time (mm)	65 non-null	float64
8	snow_1h	Integrated average hourly snow depth measurement (cm)	77 non-null	float64
9	snow_3h	Integrated snow depth measurement averaged over 3 hrs preceding the observation time (cm)	4 non-null	float64
10	drizzle_desc	Description of measured drizzle	244 non-null	object
11	fog_desc	Description of measured fog	193 non-null	object
12	clouds_desc	Description of measured clouds	72395 non-null	object
13	haze_desc	Description of measured haze	46 non-null	object
14	mist_desc	Description of measured mist	312 non-null	object
15	clear_desc	Description of measured clear	11342 non-null	object
16	snow_desc	Description of measured snow	103 non-null	object
17	storm_desc	Description of measured thunderstorm	1 non-null	object
18	temp	Captured average hourly temperature (°C)	991662 non-null	float64
19	temp_min	Captured minimum temperature over a 24-hr period (°C)	991662 non-null	float64
20	temp_max	Captured maximum temperature over a 24-hr period (°C)	991662 non-null	float64

21	feels_like	Integrated measurement of human impression of weather (K)	991662 non-null	float64
22	pressure	Captured average hourly pressure (hPa)	991662 non-null	int64
23	humidity	Captured average hourly relative humidity (ϕ)	991662 non-null	int64
24	wind_speed	Integrated average hourly wind speed (knots)	991662 non-null	float64
25	wind_direction	Integrated average hourly wind direction (true degrees)	991662 non-null	int64
26	clouds_all	Integrated hourly measurement of cloudiness (%)	991662 non-null	float64
27	ozone	Integrated average hourly ozone ($\mu g/m^3$)	181233 non-null	float64
28	ozone_avg6h	Integrated ozone readings averaged over 6 hrs preceding the observation time ($\mu g/m^3$)	181233 non-null	float64
29	NO_2	Captured average hourly NO_2 (ppb)	121207 non-null	float64
30	NO_2_avg6h	Captured NO_2 readings averaged over 6 hrs preceding the observation time (ppb)	121207 non-null	float64
31	PM_{10}	Captured average hourly PM_{10} ($\mu g/m^3$)	121207 non-null	float64
32	PM_{10}_avg6h	Captured PM_{10} readings averaged over 6 hrs preceding the observation time ($\mu g/m^3$)	121207 non-null	float64
33	$PM_{2.5}$	Captured average hourly $PM_{2.5}$ ($\mu g/m^3$)	121207 non-null	float64
34	$PM_{2.5}$_avg6h	Captured $PM_{2.5}$ readings averaged over 6 hrs preceding the observation time ($\mu g/m^3$)	121207 non-null	float64

Appendix B: List of attributes after processing and classification as categorical or continuous

S/No	Attribute Name	Attribute Type
1	city_name	Categorical
2	lat	Categorical
3	lon	Categorical
4	year	Categorical
5	month	Categorical

6	week	Categorical
7	day	Categorical
8	dayofweek	Categorical
9	dayofyear	Categorical
10	is_month_end	Categorical
11	is_month_start	Categorical
12	is_quarter_end	Categorical
13	is_quarter_start	Categorical
14	is_year_end	Categorical
15	is_year_start	Categorical
16	rain_1h	Continuous
17	snow_1h	Continuous
18	temp	Continuous
19	temp_min	Continuous
20	temp_max	Continuous
21	feels_like	Continuous
22	pressure	Continuous
23	humidity	Continuous
24	wind_speed	Continuous
25	wind_direction	Continuous
26	clouds_all	Continuous
27	ozone	Continuous
28	ozone_avg6h	Continuous
29	no_2	Continuous
30	no_2_avg6h	Continuous
31	$pm_{2.5}$	Continuous
32	$pm_{2.5}$_avg6h	Continuous

34	PM_{10}	Continuous
34	PM_{10}_avg6h	Continuous

References

Ahmed, E., Ahmed, A., Yaqoob, I., Shuja, J., Gani, A., Imran, M. & Shoaib, M. (2017), 'Bringing computation closer towards user network: Is edge computing the solution?', *IEEE Communications Magazine* **55**, 138 – 144.

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O. & Ahmed, A. A. (2020), 'Deep learning in the construction industry: A review of present status and future innovations', *Journal of Building Engineering* p. 101827.

Alléon, A., Jauvion, G., Quennehen, B. & Lissmyr, D. (2020), 'Plumenet: Large-scale air quality forecasting using a convolutional lstm network', *arXiv preprint arXiv:2006.09204* .

Alvanchi, A., Rahimi, M., Mousavi, M. & Alikhani, H. (2020), 'Construction schedule, an influential factor on air pollution in urban infrastructure projects', *Journal of Cleaner Production* **255**, 120222.

Analitis, A., Michelozzi, P., D'Ippoliti, D., De'Donato, F., Menne, B., Matthies, F., Atkinson, R. W., Iñiguez, C., Basagaña, X., Schneider, A. et al. (2014), 'Effects of heat waves on mortality: effect modification and confounding by air pollutants', *Epidemiology* pp. 15–22.

Badura, M., Batog, P., Drzeniecka-Osiadacz, A. & Modzel, P. (2018), Optical particulate matter sensors in pm2. 5 measurements in atmospheric air, *in* 'E3S Web of Conferences', Vol. 44, EDP Sciences, p. 00006.

Barikayeva, N., Nikolenko, D. & Ivanova, J. (2018), About forecasting air pollution in the construction of highways, *in* 'IOP Conference Series: Materials Science and Engineering', Vol. 463, IOP Publishing, p. 042016.

Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D. & Akanbi, L. A. (2022), 'Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting', *Machine Learning with Applications* **7**, 100204.

Barthwal, A. & Acharya, D. (2018), An internet of things system for sensing, analysis & forecasting urban air quality, *in* '2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)', IEEE, pp. 1–6.

Bilal, M. & Oyedele, L. O. (2020), 'Guidelines for applied machine learning in construction industry—a case of profit margins estimation', *Advanced Engineering Informatics* **43**, 101013.

Borghi, F., Spinazzè, A., Campagnolo, D., Rovelli, S., Cattaneo, A. & Cavallo, D. M. (2018), 'Precision and accuracy of a direct-reading miniaturized monitor in pm2. 5 exposure assessment', *Sensors* **18**(9), 3089.

Budde, M., Müller, T., Laquai, B., Streibl, N., Schwarz, A., Schindler, G., Riedel, T., Beigl, M. & Dittler, A. (2018), Suitability of the low-cost sds011 particle sensor for urban pm-monitoring, *in* '3rd International Conference on Atmospheric Dust'.

Carullo, A., Corbellini, S. & Grassini, S. (2007), 'A remotely controlled calibrator for chemical pollutant measuring-units', *IEEE Transactions on Instrumentation and Measurement* **56**(4), 1212–1218.

Chen, J., Li, K., Deng, Q., Li, K. & Philip, S. Y. (2019), 'Distributed deep learning model for intelligent video surveillance systems with edge computing', *IEEE Transactions on Industrial Informatics* .

Chen, M., Wang, S. & Xu, Q. (2015), 'Multiobjective optimization for air-quality monitoring network design', *Industrial & Engineering Chemistry Research* **54**(31), 7743–7750.

DEFRA (2019), 'Clean air strategy'.
URL: *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment$_d$ata/fil*
air − strategy − 2019.pdf

DEFRA (2020), 'Air quality appraisal: impact pathways approach'.
URL: *https://www.gov.uk/government/publications/assess-the-impact-of-air-quality/air-quality-appraisal-impactpathways-approach*

Guo, C. & Berkhahn, F. (2016), 'Entity embeddings of categorical variables', *arXiv preprint arXiv:1604.06737* .

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S. & Borowiak, A. (2019), 'Review of the performance of low-cost sensors for air quality monitoring', *Atmosphere* **10**(9), 506.

Karner, A. A., Eisinger, D. S. & Niemeier, D. A. (2010), 'Near-roadway air quality: synthesizing the findings from real-world data', *Environmental science & technology* **44**(14), 5334–5344.

Lee, M., Lin, L., Chen, C.-Y., Tsao, Y., Yao, T.-H., Fei, M.-H. & Fang, S.-H. (2020), 'Forecasting air quality in taiwan by using machine learning', *Scientific reports* **10**(1), 1–13.

Mabahwi, N. A. B., Leh, O. L. H. & Omar, D. (2014), 'Human health and wellbeing: Human health effect of air pollution', *Procedia-Social and Behavioral Sciences* **153**, 221–229.

Odat, S. (2009), 'Diurnal and seasonal variation of air pollution at al-hashimeya town, jordan', *Earth Environ Sci* **2**, 1–6.

ONS (2018), '2017 uk greenhouse gas emissions, provisional figures', *Statistical Release: National Statistics* .

ONS (2020), 'Population estimates for regions in england and wales by sex and age', Available from: https://www.statista.com/statistics/1064772/population-of-london-by-gender/ [Accessed: 15-12-2021].

ONS (2021), 'Labour market in the regions of the uk: October 2021', Available from: https://www.gov.uk/government/statistics/labour-market-in-the-regions-of-the-uk-october-2021 [Accessed: 15-12-2021].

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J. & Tapper, N. J. (2011), 'Quantifying the influence of local meteorology on air quality using generalized additive models', *Atmospheric Environment* **45**(6), 1328–1336.

Public Health England (2019), 'Review of interventions to improve outdoor air quality and public health'.
URL: *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment$_d$ata/fil* 2019 − 2018572*.pdf*

Sergeev, A. & Del Balso, M. (2018), 'Horovod: fast and easy distributed deep learning in tensorflow', *arXiv preprint arXiv:1802.05799* .

Shrivastava, A., Gupta, V. B. et al. (2011), 'Methods for the determination of limit of detection and limit of quantitation of the analytical methods', *Chronicles of young scientists* **2**(1), 21.

Si, M., Xiong, Y., Du, S. & Du, K. (2020), 'Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods', *Atmospheric Measurement Techniques* **13**(4), 1693–1707.

Statswales (2020), 'Summary statistics for wales, by region: 2020', Available from: https://gov.wales/sites/default/files/statistics-and-research/2020-05/summary-statistics-regions-wales-2020-629.pdf [Accessed: 15-12-2021].

TFL (2019), 'Travel in london: Understanding our diverse communities 2019'.

Umadevi, K. & Geraldine Bessie Amali, D. (2020), Data visualization and analysis for air quality monitoring using ibm watson iot platform, *in* 'Data Visualization', Springer, pp. 15–32.

Vohra, K., Marais, E. A., Suckra, S., Kramer, L., Bloss, W. J., Sahu, R., Gaur, A., Tripathi, S. N., Van Damme, M., Clarisse, L. et al. (2021), 'Long-term trends in air quality in major cities in the uk and india: A view from space', *Atmospheric Chemistry and Physics* **21**(8), 6275–6296.

World Bank (2022), 'The global health cost of pm2.5 air pollution: A case for action beyond 2021'.
URL: *https://openknowledge.worldbank.org/handle/10986/36501*

Zhang, K. & Batterman, S. (2013), 'Air pollution and health risks due to vehicle traffic', *Science of the total Environment* **450**, 307–316.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. & Baklanov, A. (2012), 'Real-time air quality forecasting, part i: History, techniques, and current status', *Atmospheric Environment* **60**, 632–655.

Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E. & Li, T. (2015), Forecasting fine-grained air quality based on big data, *in* 'Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 2267–2276.

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L. & Subramanian, R. (2018), 'A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring.', *Atmospheric Measurement Techniques* **11**(1).