C O V E N T R Y U N I V E R S I T Y

College of Engineering, Environment, and Computing School of Science

MSc. Data Science

7150CEM

Data Science Project Report

**Analysis and Early Detection of Respiratory Illness-Related**

**Cases Using UK Air Quality and Hospital Data**

Author: Kapil Srivastava

SID: 15105318

Supervisor: Dr. Stenford Ruvinga

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science

Academic Year: 2024/25

Word Count: 16203

A certificate of ethical approval

AI-generated content may be incorrect.

# **Declaration of Originality**

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

**Statement of copyright**

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact [ipr@coventry.ac.uk](mailto:ipr@coventry.ac.uk).

**Statement of ethical engagement**

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking).

Signed: Kapil Srivastava Date: 15/08/2025

|  |  |
| --- | --- |
| **First Name:** | **Kapil** |
| **Last Name:** | **Srivastava** |
| **Student ID Number:** | **15105318** |
| **Ethics Application Number:** | P187137 |
| **Supervisor Name:** | **Dr. Stenford Ruvinga** |

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

# **Abstract**

Air pollution presents a huge challenge to global environmental health, causing millions of preventable deaths every year and placing an extra burden on healthcare systems. Short term increases in pollution levels contribute to seasonal rises in healthcare demand — particularly in congested urban areas — engendering acute respiratory illnesses. Understanding these dynamics is important for timely interventions, guide public health policy, and improve preparedness during times of heightened risk.

This study explores the short-term effects of ambient air pollution on acute respiratory illness by integrating daily pollutant concentrations with syndromic health surveillance data over a one-year period. Both pollutants and respiratory illnesses showed clear annual winter peaks according to Seasonal-Trend Decomposition. Regarding lag analysis, there was a strong and delayed relationship (r = 0.59) observed for illness counts with overall NO₂ at the 21–30 days lag. Predictive models using Random Forest, XGBoost, and Long Short-Term Memory (LSTM) achieved robust performance, with XGBoost yielding the highest accuracy (R² ≈ 0.67). Feature importance analysis consistently ranked PM₂.₅, NO₂, and O₃ as the most influential pollutants across all models. Results highlight that incorporating machine learning, lag structures and statistical decomposition into the able to predict respiratory health hazards as well as assist in informing timely pollution-based early warning systems.

**Table of Contents**

[**Declaration of Originality** i](#_Toc206152080)

[**Abstract** ii](#_Toc206152081)

[**List of Figures** v](#_Toc206152082)

[**List of Tables** vi](#_Toc206152083)

[**Chapter 1:** **Introduction** 1](#_Toc206152084)

[**1.1** **Project Overview** 1](#_Toc206152085)

[**1.2** **Background and Context** 1](#_Toc206152086)

[**1.3** **Problem Statement** 2](#_Toc206152087)

[**1.4** **Research Aim and Objectives** 2](#_Toc206152088)

[**1.4.1** **Aim** 2](#_Toc206152089)

[**1.4.2** **Objectives** 2](#_Toc206152090)

[**1.5** **Research Questions** 3](#_Toc206152091)

[**1.6**  **Project Outline** 3](#_Toc206152092)

[**Chapter 2:** **Literature Review on Air Pollution and Respiratory Health** 5](#_Toc206152093)

[**2.1** **Introduction** 5](#_Toc206152094)

[**2.2** **Evidence Linking Air Pollution to Respiratory Health** 5](#_Toc206152095)

[**2.3** **Geographical Variations in Air Pollution Impacts** 6](#_Toc206152096)

[**2.4** **Emerging Methodologies: Machine Learning and Predictive Modeling in Environmental Health** 7](#_Toc206152097)

[**2.5** **Data Visualization in Environmental Health Research** 8](#_Toc206152098)

[**2.6** **Gaps and Opportunities for Further Research** 8](#_Toc206152099)

[**2.7** **Synthesis** 9](#_Toc206152100)

[**Chapter 3:** **Methodology** 11](#_Toc206152101)

[**3.1** **Introduction** 11](#_Toc206152102)

[**3.2** **Dataset Overview and Collection** 11](#_Toc206152103)

[**3.3** **Data Preparation and Preprocessing** 12](#_Toc206152104)

[**3.4** **Exploratory Data Analysis (EDA) Strategy** 12](#_Toc206152105)

[**3.4.1** **Descriptive Statistics** 12](#_Toc206152106)

[**3.4.2** **Time Series Visualization and Rolling Averages** 13](#_Toc206152107)

[**3.4.3** **Seasonal Decomposition** 13](#_Toc206152108)

[**3.4.4** **Correlation Exploration (Preliminary)** 13](#_Toc206152109)

[**3.4.5** **Lag and Advanced Lag Analysis** 14](#_Toc206152110)

[**3.4.6** **AQI Proxy Over Time (Seasonal Overlay)** 14](#_Toc206152111)

[**3.5** **The Models** 14](#_Toc206152112)

[**3.5.1** **Random Forest Regressor** 15](#_Toc206152113)

[**3.5.2** **XGBoost** 16](#_Toc206152114)

[**3.5.3** **Long Short-Term Memory Neural Network (LSTM)** 17](#_Toc206152115)

[**3.6** **Evaluation Metrics** 19](#_Toc206152116)

[**3.7** **Implementation Details** 21](#_Toc206152117)

[**3.7.1** **Development Environment** 21](#_Toc206152118)

[**3.7.2** **Data Preprocessing and Feature Engineering** 21](#_Toc206152119)

[**3.7.3** **Train-Test Splitting Strategy** 22](#_Toc206152120)

[**3.7.4** **Model Implementations** 22](#_Toc206152121)

[**3.7.5** **Training and Optimization** 23](#_Toc206152122)

[**3.7.6** **Model Explainability** 23](#_Toc206152123)

[**3.7.7** **Reproducibility and Code Structure** 23](#_Toc206152124)

[**4.1** **Exploratory Data Analysis (EDA) Results** 24](#_Toc206152125)

[**4.1.1** **Descriptive Statistics** 24](#_Toc206152126)

[**4.1.2** **Time Series Visualization and Rolling Averages** 26](#_Toc206152127)

[**4.1.3** **Seasonal Decomposition** 29](#_Toc206152128)

[**4.1.4** **Correlation Analysis** 31](#_Toc206152129)

[**4.1.5** **Lag and Advanced Lag Analysis** 33](#_Toc206152130)

[**4.1.6** **AQI Proxy Over Time (Seasonal Overlay)** 37](#_Toc206152131)

[**4.2** **Machine Learning Results** 38](#_Toc206152132)

[**4.2.1** **Random Forest Regressor: Performance and Feature Importance** 38](#_Toc206152133)

[**4.2.2** **XGBoost (Untuned): Predictive Performance and Feature Importance** 41](#_Toc206152134)

[**4.2.3** **XGBoost (Tuned): Enhanced Predictive Accuracy and SHAP-Based Interpretability** 44](#_Toc206152135)

[**4.2.4** **Long Short-Term Memory (LSTM): Temporal Learning and Sequential Feature Relevance** 46](#_Toc206152136)

[**4.3** **Results Summary and Comparative Evaluation** 50](#_Toc206152137)

[**Chapter 5:** **Discussion and Critical Appraisal** 52](#_Toc206152138)

[**5.1** **Methodological Reflection on Analytical Approaches** 52](#_Toc206152139)

[**5.1.1** **Exploratory Data Analysis and Seasonal Decomposition** 52](#_Toc206152140)

[**5.1.2** **Lag Analysis of Pollution-Health Links** 52](#_Toc206152141)

[**5.1.3** **Machine Learning Predictive Modeling** 53](#_Toc206152142)

[**5.2** **Key Findings on Pollutant–Health Associations** 54](#_Toc206152143)

[**5.3** **Significance of Temporal Dependencies and Seasonality** 55](#_Toc206152144)

[**5.4** **Implications for Public Health Policy and Planning** 55](#_Toc206152145)

[**5.4.1** **Early Warning Systems** 55](#_Toc206152146)

[**5.4.2** **Seasonal Planning and Resource Allocation** 56](#_Toc206152147)

[**5.4.3** **Targeted Pollution Interventions** 56](#_Toc206152148)

[**5.4.4** **Integration with Syndromic Surveillance** 56](#_Toc206152149)

[**5.5** **Limitations of the Study and Challenges** 56](#_Toc206152150)

[**5.6** **Theoretical Implications and Future Perspectives** 57](#_Toc206152151)

[**5.7** **Conclusion** 58](#_Toc206152152)

[**Chapter 6:** **Project Management** 59](#_Toc206152153)

[**6.1** **Introduction** 59](#_Toc206152154)

[**6.2** **Timeline and Milestones** 59](#_Toc206152155)

[**6.3** **Tools and Resources** 60](#_Toc206152156)

[**6.4** **Risk Management** 60](#_Toc206152157)

[**6.5** **Quality Assurance** 61](#_Toc206152158)

[**6.6** **Social, Legal, Ethical, and Professional Considerations** 61](#_Toc206152159)

[**Chapter 7:** **Conclusion & Future Word** 62](#_Toc206152160)

[**7.1** **Achievement of Objectives** 62](#_Toc206152161)

[**7.2** **Limitations** 62](#_Toc206152162)

[**7.3** **Future Work** 63](#_Toc206152163)

[**Chapter 8:** **Student Reflections** 65](#_Toc206152164)

[**8.1** **Learning Outcomes** 65](#_Toc206152165)

[**8.2** **Challenges and Solutions** 65](#_Toc206152166)

[**8.3** **Skills Development** 65](#_Toc206152167)

[**8.4** **Future Applications** 66](#_Toc206152168)

[**8.5** **Summary** 66](#_Toc206152169)

[**References** 67](#_Toc206152170)

[**Appendix A - Meeting Records** 72](#_Toc206152171)

[**Appendix B – Links** 73](#_Toc206152172)

[**Appendix C – Certificate of Ethics Approval** 74](#_Toc206152173)

# **List of Figures**

[**Figure 1: Schematic diagram of the Random Forest algorithm. Source: Sahour et al. (2021).** 16](#_Toc206150559)

[**Figure 2: Architecture of the XGBoost algorithm. Source: TutorialsPoint (n.d.).** 17](#_Toc206150560)

[**Figure 3: LSTM cell structure showing input, forget, and output gates. Source: Ihianle et al. (2020).** 18](#_Toc206150561)

[**Figure 4: Raw Daily Time Series of Health Indicators** 27](#_Toc206150562)

[**Figure 5: Monthly Averages of Health Indicators** 27](#_Toc206150563)

[**Figure 6: Monthly Averages of Key Pollutants** 28](#_Toc206150564)

[**Figure 7: 7-Day Rolling Averages of Health Indicators** 28](#_Toc206150565)

[**Figure 8: 7-Day Rolling Averages of Pollutants** 29](#_Toc206150566)

[**Figure 9: STL Decomposition of Health Indicators (Observed, Trend, Residual)** 30](#_Toc206150567)

[**Figure 10: STL Decomposition of Pollutants (Observed, Trend, Residual)** 31](#_Toc206150568)

[Figure 11: Pearson Correlation Matrix of Health Indicators and Air Pollutants 32](#_Toc206150569)

[**Figure 12: Overlay Time Series Plot of NO₂ Concentration and Acute Respiratory Illness** 35](#_Toc206150570)

[**Figure 13: Bar Chart of NO₂-Based Feature Correlations with Acute Respiratory Illness** 36](#_Toc206150571)

[**Figure 14: Almon Lag Weight Profile for NO₂ Effect on Acute Respiratory Illness** 36](#_Toc206150572)

[**Figure 15: AQI Proxy Over Time with Seasonal Overlay** 38](#_Toc206150573)

[**Figure 16: SHAP summary plot for the Random Forest Regressor, showing the top 20 features by average SHAP value.** 40](#_Toc206150574)

[**Figure 17: Feature importance plot from the untuned XGBoost model.** 43](#_Toc206150575)

[**Figure 18: SHAP summary plot for the tuned XGBoost model.** 45](#_Toc206150576)

[**Figure 19: Training and validation loss over 24 epochs for the initial LSTM model.** 47](#_Toc206150577)

[**Figure 20: Refined LSTM model training and validation loss curves.** 48](#_Toc206150578)

[**Figure 21: SHAP summary plot showing the top 10 most impactful features in the LSTM model.** 49](#_Toc206150579)

[**Figure 22: Gantt. Chart for the project** 59](#_Toc206150580)

# **List of Tables**

[**Table 1: Summary of different research papers** 10](#_Toc206116821)

[**Table 2: Descriptive Statistics of Key Variables** 26](#_Toc206116822)

[**Table 3: Top Positive Cross-Correlations between Lagged Pollutants and Health Indicators** 34](#_Toc206116823)

[**Table 4: Top 10 most important features in the Random Forest model as determined by mean SHAP values.** 41](#_Toc206116824)

[**Table 5: Top 10 most important features in the untuned XGBoost model by raw importance score.** 44](#_Toc206116825)

[**Table 6: Top 10 most important features in the tuned XGBoost model by mean SHAP importance.** 46](#_Toc206116826)

[**Table 7: Comparative performance metrics across all models for ARI forecasting**. 51](#_Toc206116827)

# **Chapter 1: Introduction**

## **1.1 Project Overview**

The dissertation takes on the temporal relationship of ambient air pollutants with acute respiratory health outcomes on a one-year-long study of June 2024 to June 2025 in an urban population. It links high-resolution environmental data, including common air pollutants (PM10, PM2.5, NO2, NOx, O3 and black carbon) with spectrally resolved particulate matter data, and daily syndromic health surveillance parameters (acute respiratory infection, bronchiolitis, influenza-like illness and scarlet fever). The study applies a multi-method approach that integrates time series decomposition, lag analysis, and predictive modeling using Random Forest, XGBoost, and LSTM algorithms. Model interpretability is supported through the application of SHAP values to identify the most influential pollutants.

## **1.2 Background and Context**

Air pollution remains one of the most urgent environmental health challenges worldwide, contributing significantly to global morbidity and mortality. According to the World Health Organization (2022), exposure to pollutants such as particulate matter (PM₁₀ and PM₂.₅), nitrogen oxides (NO, NO₂, and NOₓ), black carbon, and ozone (O₃) is associated with a range of adverse respiratory outcomes, including asthma, bronchitis, chronic obstructive pulmonary disease (COPD), and acute respiratory infections. These pollutants come from a variety of sources, such as industrial processes, home heating, automobile emissions, and atmospheric photochemical reactions. The distribution and health impacts of these pollutants vary depending on the terrain, seasonal cycles, and weather.

Urban agglomerations characterised by elevated population density and agglomerated emission sources have severe environmental fragility. In addition, the behaviour of pollutants as well as their exposure to the population is driven by seasonal effects, such as summer photochemical activity and winter temperature inversions. In this regard, the fact that the influxes in respiratory infections coincide almost with perfect temporal precision with particularly widespread air-quality degradation significantly imposes pressure on the infrastructure of the public-health system. Nonetheless, although traditional literature presents the association between environmental pollutants and human morbidity, modern-day preparedness in the field of public health now rests on the ability to grasp the time-pattern and forecasting validity of these associations.

## **1.3 Problem Statement**

Although a large volume of research has established an association of air pollution and respiratory morbidity, there are significant absentee studies that provide high-resolution and longitudinal data capable of combining established pollutant measures with new intervals of categorical indicator measures and direct measurements of syndromic health monitoring. Relatively little research has used more cutting-edge machine-learning methods to predict health risks and, simultaneously, explain pollutant-specific sources of morbidity at varying lag times. The fact that contemporary visualization tools are largely underused also hinders the successful sharing of results with the entity of public-health practitioners.

The limitations of past research—including limited access to integrated datasets, absence of interactive risk dashboards, and insufficient modeling of temporal dynamics—restrict the generation of actionable insights. By combining several data sources and using advanced analytical tools, this project seeks to solve these issues and improve environmental health planning and respiratory health surveillance.

## **1.4 Research Aim and Objectives**

### **1.4.1 Aim**

This study aims to investigate temporal relationships between ambient air pollutants and acute respiratory health outcomes by means of a 1-year, high-resolution dataset, and create predictive and explanatory tools that can guide proactive population health responses.

### **1.4.2 Objectives**

The objectives of this study are fourfold. First, it seeks to characterise seasonal patterns and temporal trends in daily respiratory health indicators by applying time series decomposition techniques. Second, it aims to investigate the statistical associations and lagged effects between key air pollutants—such as PM₂.₅, NO₂, O₃, and black carbon—and acute respiratory illnesses. Third, the study intends to develop and evaluate predictive models, including Random Forest, XGBoost, and LSTM, to forecast respiratory health outcomes based on environmental exposure data. Finally, it will identify and rank the most influential pollutants contributing to health outcomes with interpretable machine learning methods such as SHAP and sensitivity analysis.

## **1.5 Research Questions**

Based on the identified gaps and objectives, this study seeks to address the following research questions:

1. What seasonal patterns and temporal trends can be observed in daily respiratory health indicators during the study period, and in what ways do these patterns align with fluctuations in ambient air pollutant concentrations?
2. Which air pollutants exhibit the strongest statistical associations with acute respiratory health outcomes, and how do lagged effects—whether occurring on the same day or over multiple subsequent days—shape the nature of these relationships?
3. How accurately can machine learning models such as Random Forest, XGBoost, and LSTM forecast acute respiratory health outcomes using environmental exposure data, and which of these approaches offers the most effective balance between predictive accuracy and interpretability?
4. Which pollutants have the greatest relative importance in predicting health outcomes according to interpretable machine learning methods such as SHAP and sensitivity analysis, and how can these findings be visualised in a way that supports informed decision-making in public health?

## **1.6 Project Outline**

The dissertation is structured as follows:

**Chapter 1** introduces the research background, defines the problem, sets out the aim, objectives, and research questions, and outlines the project.

**Chapter 2** presents a literature review of existing research on air pollution and respiratory health, including epidemiological evidence, geographical variations, machine learning applications, and data visualisation practices, identifying research gaps and opportunities.

**Chapter 3** details the methodology, including dataset description, preprocessing, exploratory data analysis, modelling techniques, evaluation metrics, and implementation details.

**Chapter 4** reports the results from exploratory analysis, seasonal decomposition, correlation studies, lag analysis, AQI proxy construction, and the performance of different predictive models.

**Chapter 5** discusses the results in the context of existing literature, evaluates methodological choices, and outlines implications for public health policy.

**Chapter 6** describes the project management process, including timeline, resources, risk management, and quality assurance.

**Chapter 7** presents the study’s conclusions, assesses the achievement of objectives, acknowledges limitations, and suggests avenues for future research.

**Chapter 8** offers personal reflections on learning outcomes, challenges, skills development, and potential future applications of the knowledge gained.

# **Chapter 2: Literature Review on Air Pollution and Respiratory Health**

## **2.1 Introduction**

Air pollution remains a major global public health concern with established associations with respiratory health outcomes. These pollutants are known to include fine particulate matter — which contribute to respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), acute lower respiratory infections and more. Such pollutants, originating from vehicle emissions, industrial activities and natural processes among others promote inflammation and worsen chronic diseases which is a serious threat to public health systems across the globe. Moreover, there are the climate change impacts which can exacerbate these problems by increasing the frequency and severity of extreme weather events that also increase pollutant concentrations (Tran et al., 2023). Using this body of work as a lens, the current literature review is well situated to offer a comprehensive analytical approach when analyzing over the course of approximately one year in terms of daily free-pollutant concentrations and syndromic-health outcomes.

## **2.2 Evidence Linking Air Pollution to Respiratory Health**

A substantial body of epidemiological research demonstrates the harmful effects of air pollution on respiratory health. (Monoson et al., 2023) estimate that approximately seven million premature deaths annually are linked to air pollution, with 17% of pneumonia deaths attributable to ambient exposure. Key pollutants, including PM₂.₅, NO₂, and SO₂, increase susceptibility to bacterial, fungal, and viral respiratory infections. Historical evidence from the 1952 Great Smog of London shows pneumonia mortality rose by about 80%, resulting in 3,500–4,000 excess deaths over five years. During the COVID-19 pandemic, Tian et al. (2021) reported roughly 11% higher mortality in regions with the highest PM₂.₅ levels. Mechanistic pathways include inflammation, oxidative stress, and impaired immune function, particularly among children and the elderly (Tran et al., 2023). A meta-analysis by Lee et al. (2024) found a pooled relative risk of 1.12 (95% CI [1.05, 1.20]) for adult asthma incidence associated with PM₂.₅ and NO₂ exposure, although heterogeneity among studies indicates the need for further research.

Particulate matter is a dangerous health issue to the people, especially when it involves submicron particles, especially PM2.5 and lower. Black carbon, which is generated by incomplete burning of fossil and biomass fuels, has become a target area of interest since these substances can easily carry into and deposit in the outlying areas of the lung. In a recent study, it was discovered that any change in the interquartile ranges of the concentration of black carbons translated to an increase in daily hospital admissions due to asthma of around 6-9 percent, therefore, offering empirical support to the existence of a correlation between particulate exposure and respiratory morbidity. These particles are sub-micron, which promote alveolar deposition, a position that enhances oxidative stress and inflammation hence a potential source of chronic in vivo pathogenesis. There is also epidemiological evidence, which indicates that there is long term association between particulate matter and cardiovascular disease and respiratory disease. The State of Global Air 2024 report finds that in 2021, ambient pollution led to 8.1 million deaths across the world, with 4.7 million of these deaths being caused by exposure to PM2.5 where most of these cases involved respiratory illnesses.

In England (including the West Midlands), short-term NO₂ exposure is associated with higher paediatric asthma admissions; a nationwide case-crossover study estimated ≈8% higher risk per 10 µg/m³ over lags 0–4 (Wang et al., 2024). Ozone, formed photochemically, can irritate the airways and contribute to respiratory demand, and road-traffic NOₓ remains a key urban source; together these factors motivate multi-pollutant analyses.

## **2.3 Geographical Variations in Air Pollution Impacts**

The regional examination provides useful insights into region-based trends of air pollution and health outcomes. In a recent study by Hall et al. (2020) in Birmingham, UK, the contribution of PM2.5 was evaluated based on concentration-response functions that had been endorsed by the Committee on the Medical Effects of Air Pollutants. In the study, it was estimated that air pollution contributes to about 720 premature deaths, 7,500 lost life years and 900 new cases of asthma in children and adults each year. A geographical divide also emerged showing that the highest rates of pollution-mortality were between 8.5 % in central wards such as Tyseley & Hay Mills, Holyhead and Aston. This is the case in the areas characterized by dense populations and the proximity to the industrial and traffic-related source which bear the unjust health burden.

A national study covering 1990 to 2021 in the United States reported an 80.5% reduction in PM2.5-associated mortality and the existence of strong geographic disparities in health outcomes; California, specifically, reported increasing diabetes-associated disability-adjusted life years (DALYs). Tan et al. (2023) observed in Southeast Asia that exposure of urban areas in Malaysia to particulate matter was linked to respiratory symptoms, coughing, wheezing in children and linked primarily to diesel emissions and industrial work. Nkosi et al. (2024) addressed the health burden left behind by household air pollution in Africa because people still use biomass fuels such as charcoal and kerosene. Such circumstances precondition high levels of respiratory diseases, among which 1.1 million deaths occurred in 2019 due to PM. Zhang et al. (2024) conducted a multicity study over the period 2017-2022 and found out that PM2.5 and O3 exposure is both linked to heterogenous patterns of hospitalisation and mortality, with older populations and socioeconomically disadvantaged regions being the most at risk.

## **2.4 Emerging Methodologies: Machine Learning and Predictive Modeling in Environmental Health**

With the advent of machine-learning techniques, environmental-health research has now been fundamentally restructured to allow the accurate modelling of complicated non-linear relationships between air pollutants, meteorological factors and human health effects. In a typical case, Shi et al. (2020) used a Chained Random Forest Classifier (CRFC) to estimate the presence of respiratory-viruses in 31 regions of China between 2016 and 2021, combining air-quality indices and meteorological parameters, and noted an average predictive accuracy of 0.76 at an AUC of 0.9. In another study, Wang et al. (2023) tested eight machine-learning models on predicting outpatient visits due to acute respiratory infections during the period of 2018-2021; the random-forest model was found to be the most accurate, whereas NO2 was found to be a key predictor, with a lag of one day. Gradient-boosting and Gaussian-process regression were also used by Kim et al. (2022) to predict the occurrence of respiratory diseases with R2 value of 0.67-0.68, and the combined predictors were found to be temperature, humidity, PM2.5, and SO2.

In a recent study conducted by Su et al. (2024) on a sample size of 3,386 patients with chronic obstructive pulmonary disease (COPD) and asthma in California between the years 2012 and 2019, digital health sensors and an algorithm based on random forest models were used. The results show that an increased use of rescue medication is positively correlated with greater environmental concentrations of NO2, PM2.5, and O3. It was indicated that every increase in exposure was linked to a 23.9 percent increment in the number of daily rescue puffs that implies serious implications on finances. All these findings highlight the possible usefulness of machine learning in enhancing the reliability of public health planning and prediction. Additionally, Meckawy et al. (2020) state that early warning systems (EWS) that were initially developed to track infectious diseases should be researched in environmental health.

## **2.5 Data Visualization in Environmental Health Research**

Effective communication of complex environmental health data is critical for informing policy and engaging stakeholders. Well-designed visualisations, including interactive dashboards and maps, have been shown to enhance comprehension and support decision-making (Ramirez et al., 2019). While the Air Quality Index (AQI) uses a colour-coded scale to indicate pollution levels, Ramirez et al. (2019) found limited evidence that such indices lead to recommended behaviour changes during poor air quality events, indicating a need for improved communication strategies.

Visualisation is one of the key analytical methods in modern-day empirical research, such as the case of Brown and Green (2022), who visualised the research hotspots of air pollution and studies within the health-related field. The current research uses chart- and plot-based visualisations to question temporal patterns and spatial distributions, especially those involving exposure to air pollution and associated health effects.

## **2.6 Gaps and Opportunities for Further Research**

Despite the gains made, there also remain gaps in the empirical literature. Most of the current literature assesses discrete pollutants separately or considers general health outcomes without explaining how the exposure unfolds throughout time. To define seasonal trends and compare short-term variability of various pollutants and their health effects the current study made use of daily pollution data accumulated for one year. Including spectral particulate matter permits the extension of previous analyses on several pollutants, thus meeting an important gap in the environmental epidemiological literature.

Machine learning has shown potential as a method of predicting respiratory status when pertaining to multi-pollutant exposures, although further examination is necessary. The possibility is shown by Wang et al. (2023) and Kim et al. (2022) but has to be reproven on a larger scale to ascertain generalisability. Interpretable model development, especially the models using SHAP analysis, is also important in the conversion of predictions to actionable policy. Besides, it is possible that the expansion of infectious-disease early-warning systems to environmental health may help detect risks during intense-pollution events (Meckawy et al., 2020), yet the problems of data integration and equitable utilization in low-resource settings remain to be addressed.

The effectiveness of data visualisation tools in communicating environmental health risks also warrants further study. Ramirez et al. (2019) note limitations in current approaches, underscoring the need to understand how different audiences interpret visual representations to improve communication strategies.

## **2.7 Synthesis**

The literature provides a robust foundation for understanding the adverse effects of air pollution on respiratory health, supported by epidemiological, regional, and methodological studies. Recent research, such as Tran et al. (2023) and the State of Global Air 2024 report, reinforces the significant health burden of air pollution, particularly in high-risk regions like Asia. Machine learning advancements, as demonstrated by Wang et al. (2023) and Kim et al. (2022), offer promising tools for predicting health outcomes, while visualization strategies, though widely used, require further refinement to maximize impact (Ramirez et al., 2019). The current project builds on this knowledge by leveraging a comprehensive dataset of daily air pollutant and health indicator measurements, employing advanced machine learning models and interactive visualizations to uncover temporal trends and predictive relationships. By addressing gaps in multi-pollutant analyses, early warning systems, and effective communication, the project aims to contribute to evidence-based air quality management and public health interventions, ultimately reducing the burden of respiratory illnesses.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author(s) & Year** | **Study Location / Context** | **Data & Variables** | **Methodology** | **Key Findings** | **Limitations / Notes** |
| Monoson et al., 2023 | Global | Mortality, pneumonia cases, PM₂.₅, NO₂, SO₂ | Epidemiological analysis | 7M premature deaths annually from air pollution; 17% pneumonia deaths linked to ambient exposure | Global estimates; possible regional variability not explored |
| Tian et al., 2021 | China | PM₂.₅ levels, COVID-19 mortality | Statistical modelling | ~11% higher COVID-19 mortality in high PM₂.₅ regions | Pandemic-specific context; may not generalise |
| Tran et al., 2023 | Global | Air pollution, immune function | Mechanistic review | Pollution triggers inflammation, oxidative stress, immune suppression | Mechanistic, not quantitative |
| Lee et al., 2024 | Meta-analysis (Global) | PM₂.₅, NO₂ exposure, asthma incidence | Meta-analysis | Pooled RR = 1.12 for adult asthma with PM₂.₅ & NO₂ | Heterogeneity between studies |
| Wang et al., 2024 | England (nationwide; relevant to West Midlands) | Daily NO₂ (µg/m³); emergency hospital admissions for childhood asthma | Time-stratified case-crossover (lags 0–4 days) | ≈8% increase in paediatric asthma admissions per 10 µg/m³ NO₂ | Exposure misclassification (area-level NO₂), potential co-pollutant confounding, paediatric-only sample |
| Hall et al., 2020 | Birmingham, UK | PM₂.₅ exposure, mortality, asthma | Health impact assessment | ~720 early deaths/year; 900 new asthma cases | City-specific; extrapolation limited |
| Frontiers Authors, 2025 | USA | PM₂.₅ mortality trends | Longitudinal study | 80.5% decline in PM₂.₅ mortality (1990–2021) | Variability by state |
| Tan et al., 2023 | Malaysia | Particulate matter exposure, children’s symptoms | Cross-sectional | PM linked to cough, wheeze in children | Localised; limited pollutants studied |
| Zhang et al., 2024 | Multi-city (Global) | PM₂.₅, O₃, ospitalization & mortality | Epidemiological | Older and disadvantaged populations more vulnerable | Limited rural data |
| Wang et al., 2023 | China | NO₂, respiratory infections | ML comparison (8 models) | RF best; NO₂ 1-day lag important | Lag limited to short-term |
| Kim et al., 2022 | South Korea | PM₂.₅, SO₂, meteorology, respiratory disease | Gradient boosting & GPR | R² = 0.67–0.68; pollutants + weather key predictors | National context; may not generalise |
| Su et al., 2024 | USA (California) | NO₂, PM₂.₅, O₃, medication use | Digital health monitoring + RF | Pollution linked to increased rescue medication use | Specific to asthma/COPD patients |
| Meckawy et al., 2020 | Egypt | Air quality, EWS adaptation | Framework evaluation | Potential for EWS in environmental health | Conceptual; not empirical |

**Table 1: Summary of different research papers**

# **Chapter 3: Methodology**

## **3.1 Introduction**

The paper is founded on a data-driven methodology of researching dynamic relationships that govern the linkage between air quality pollutants and respiratory health outcomes in the West Midlands region of the United Kingdom. An orderly acquisition scheme, stringent data cleansing operations, and analytical pipelines that follow each other have built large datasets that form the foundation of the inquiry, exploratory analysis and prediction modelling of time series. An individual regression modelling assignment is presented, where the daily respiratory-health cases counts are modelled and predicted via exposure variables. The regression models include Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks and measure the relationship between pollutants and health outcomes, at different delays. The general strategy is to maximize replicability, analytical rigour, and use novel methods to provide temporal and statistical knowledge.

## **3.2 Dataset Overview and Collection**

The current paper provides a multifaceted description of environmental and health dynamics of the West Midlands region based on two easily available and state repositories to ensure reliability and comprehensive spatial coverage.

The UK Health Security Agency (UKHSA) dataset contains daily data on emergency department attendance with four respiratory syndromes-acute bronchiolitis, acute respiratory illness, influenza-like illness and scarlet fever- where each serve as a proxy measure of respiratory health trends at a community level and enables consistent monitoring over time.

Air quality data from the UK Department for Environment, Food & Rural Affairs (DEFRA) Air Quality Archive includes daily measurements from monitoring stations across the West Midlands. Pollutants are grouped into conventional species—PM₂.₅, PM₁₀, nitrogen dioxide (NO₂), nitric oxide (NO), total nitrogen oxides (as NO₂), ozone (O₃), and black carbon—and spectral particulate matter, capturing optical responses across wavelengths including blue, green, red, yellow (590 nm), ultraviolet (370 nm), and infrared.

The dataset spans June 2024 to June 2025, comprising 365 daily entries and 18 variables, enabling temporal analysis of air pollution in relation to short-term fluctuations in respiratory-related healthcare utilisation.

## **3.3 Data Preparation and Preprocessing**

The preprocessing stage of this inquiry was primarily performed with the use of Python, which is embedded into structured Jupyter notebooks. At this step, several data-quality processes were performed, such as temporal alignment, and subsequent analysis preparation. In the case of health datasets cohort, four syndromes were maintained, i.e., acute bronchiolitis, acute respiratory illness, influenza-like illness, and scarlet fever. To make the date formatting consistent all temporal values were cast to the Python datetime format, and missing values were imputed using a seven-day rolling mean. Other transformations, i.e. column renaming and checking numeric fields were performed.

Air-quality data sets needed more comprehensive treatment, since they had their metadata (e.g. headers, footnotes), non-numeric rows, as well as placeholders with no data. These were eliminated prior to the implementation of a rolling-mean imputation technique which was repeated in each column. The imputed values were then transformed to numerical form, and the final measurements were rounded off to three decimal points.

After the preprocessing steps done on the individual data, the two datasets were combined based on the common date column between 9 June 2024 and 8 June 2025. An end-to-end quality-assurance process confirmed that no null observations were present and daily coverage was complete. The result was a single dataset with 365 records of 18 variables: one date index and four health indicators and thirteen pollutant measurements, all in line with national environmental monitoring guidelines.

## **3.4 Exploratory Data Analysis (EDA) Strategy**

The Exploratory Data Analysis (EDA) phase was structured to systematically investigate the temporal structure, statistical distribution, and inter-variable dynamics of the dataset. This process was essential for hypothesis development, anomaly detection, and guiding downstream modeling efforts.

### **3.4.1 Descriptive Statistics**

The descriptive statistical tests were used to explain the underlying distribution of each variable in the data set. Mean and median was estimated as an indicator of central tendency in addition to standard deviation, minimum and maximum being estimated as an indicator of dispersion to explain variability. The coefficients of skewness gave additional information on the shape of the distribution, which helped to identify outliers and evaluate temporal variability and any further action on possible normalisation or transformation.

### **3.4.2 Time Series Visualization and Rolling Averages**

Plots of time series were prepared of all variables over the full temporal extent to visually determine temporal continuity, and to detect periodic variations or interruptions. Raw day-to-day values have been given too as this is how the data naturally varies. To minimize the noise in the short term and emphasize the mid-term trends, rolling seven-day averages were then used on every time series. The relationship between the health parameters and the level of pollutants was visualized in a graphical form and formed the basis to form an initial hypothesis. This procedure gave a rudimentary notion of time-dependent behavior prior to decomposition and lag analysis were conducted.

### **3.4.3 Seasonal Decomposition**

Within the investigation at hand, the Seasonal-Trend-Loess (STL) decomposition technique was used to decompose the constituent parts of each time series variable. Fits were conducted in the Python statsmodels environment with additive mode to split the trend, seasonal and the residual components. This de-composition method helped to identify temporal recurring patterns, and at the same time separate them against other underlying trends or stochastic noise. The resultant decompositions were later used to execute feature engineering and to increase interpretability in later modelling pursuits.

### **3.4.4 Correlation Exploration (Preliminary)**

The preliminary correlation analysis was performed to explore eventual linear relationships between variables. Full correlation matrix, accompanied by heatmaps, between all pollutant and health indicators using Pearson’s correlation coefficients were presented. Inter-pollutant correlations within groups were further investigated to identify overlapping or a common underlying environmental pattern.

### **3.4.5 Lag and Advanced Lag Analysis**

The lag analysis needed to be implemented through the examination of the latent health impacts of exposure to air pollutants. In this, discrete lag shifts of one to fourteen calendar days were applied to the variables representing pollutants, and cross correlation functions (CCFs) were estimated to quantify time-dependencies of pollutants and health indicator variables. Also, the lag windows, i.e., the three-days moving average and seven-days moving average, were assessed as well to consider the cumulative effects of exposure. In addition, another approach to modeling distributed lag structures and interpreting associated time-averaged effects was discussed, namely, the use of polynomial lag models such as Almon lags. The lag analysis, which resulted thereafter, informed the further selection of lagged features to be included in the time-aware modelling processes.

### **3.4.6 AQI Proxy Over Time (Seasonal Overlay)**

Normalised concentrations of selected pollutant data are aggregated to create a custom proxy for the Air Quality Index (AQI) This metric was depicted in a seasonal time series plot to assess its variability through the study period. Some overlay plots were also produced to compare AQI proxy with health indicator timelines but no direct interpretative claims. We then used these visualisations to help form additional hypotheses and create a clearer picture of the nexus between air quality and respiratory health.

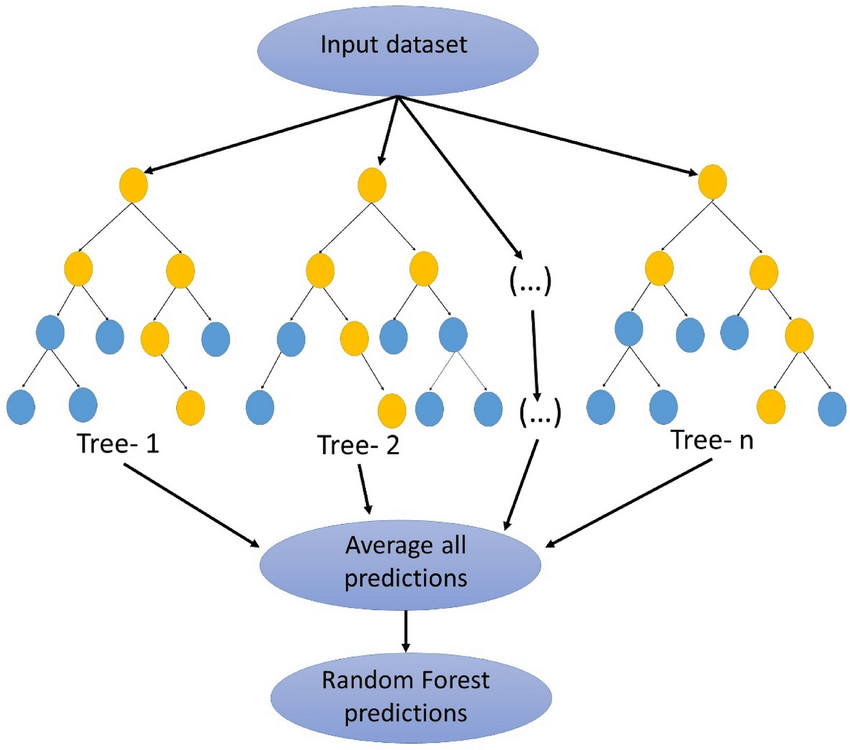
## **3.5 The Models**

This current modeling approach aims at forecasting acute respiratory illness (ARI) using a package of essential air pollutants as predictors. In particular, the NO2, NOx, PM10, black carbon, and ozone are included among a few other parameters, where current and lagged values are used to reflect the short and medium-term exposure effects. The models used are both ensemble-based techniques of machine learning and recurrent deep learning networks, which enables a thorough comparison of predictive ability and measures of feature interpretability. Such an integrative strategy will allow evaluating not only the accuracy of ARI forecasting but also the relative significance attached to individual pollutants as well as time windows in determining the health consequences.

### **3.5.1 Random Forest Regressor**

Random Forest is a bagging-based ensemble learning algorithm that constructs multiple decision trees during training and aggregates their predictions to produce a single output. Each tree is trained on a bootstrapped sample of the dataset, and at each split, a random subset of features is considered, which promotes diversity among the trees and reduces the correlation of their errors. For regression tasks, the final prediction is obtained by averaging the outputs of all trees, whereas for classification tasks, a majority vote is taken. The process, as represented in the figure, begins with sampling data, training multiple decision trees independently, and then combining their predictions to generate the result. This ensemble approach enhances predictive stability, improves generalisation, and reduces overfitting compared to a single decision tree.

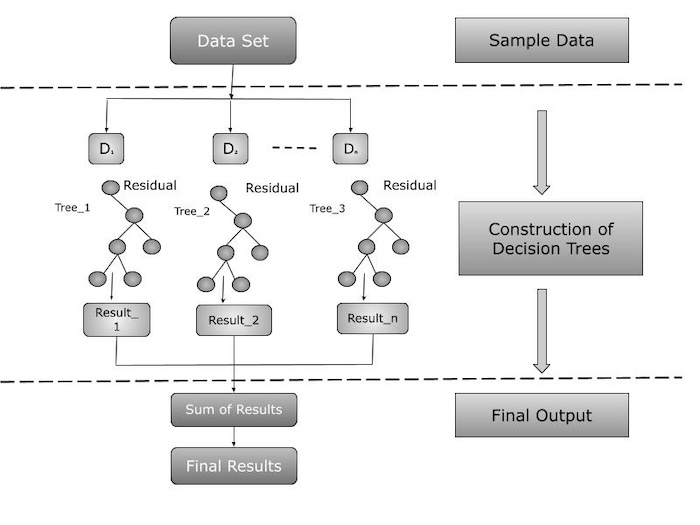
A tuned RandomForestRegressor was employed with n\_estimators=100 to ensure an adequate number of trees and n\_jobs=-1 to enable parallel computation for efficiency. The model was trained on a feature matrix containing both current-day pollutant concentrations and multiple lagged values, with each pollutant–lag pair treated as an independent feature to capture non-linear relationships and feature interactions. Since Random Forest does not inherently account for sequential dependencies, temporal information was incorporated through manually engineered lag variables. This configuration enabled the model to work effectively with the mixed feature types present in the dataset, while also allowing for the computation of feature importance rankings to support interpretability.



**Figure 1: Schematic diagram of the Random Forest algorithm. Source: Sahour et al. (2021).**

### **3.5.2 XGBoost**

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that constructs additive models sequentially, with each tree aiming to correct the residual errors of the previous iteration. The algorithm minimises a specified loss function by fitting new models to the gradients of that loss, and the final prediction is obtained by summing the contributions from all trees. The process, shown in the figure, involves sampling data, constructing decision trees based on residuals, and aggregating their outputs to produce the result. XGBoost incorporates techniques such as shrinkage (learning rate control), subsampling, and regularisation through L1 and L2 penalties, which collectively improve generalisation and help prevent overfitting. Its computational efficiency and scalability make it particularly suited to large, structured datasets.

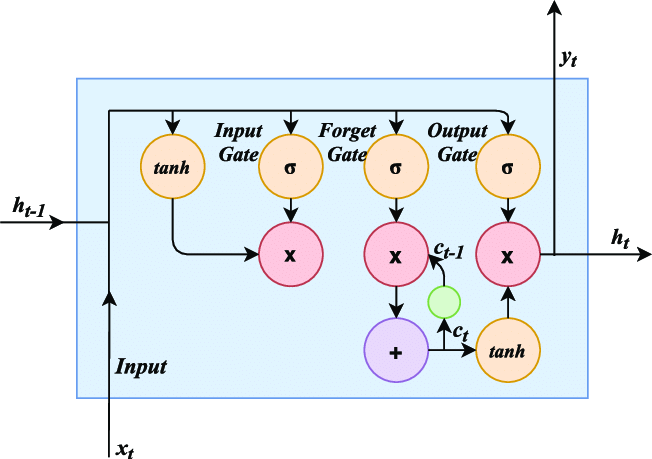


**Figure 2: Architecture of the XGBoost algorithm. Source: TutorialsPoint (n.d.).**

XGBoost was implemented in two stages: first using the default configuration to establish a baseline, and then as a tuned model with optimised hyperparameters including max\_depth=6, eta=0.1, and subsample=0.8 to balance complexity and performance. The model was applied to a tabular dataset comprising current-day pollutant concentrations and multiple lagged values, with lag features manually engineered to enable the extraction of temporal patterns. This approach allowed the model to handle non-linear relationships and feature interactions effectively, while maintaining interpretability through feature importance metrics. Although XGBoost is highly effective for structured data, it does not natively model sequential dependencies, making manual lag feature creation a necessary step in this application.

### **3.5.3 Long Short-Term Memory Neural Network (LSTM)**

Long Short-Term Memory (LSTM) networks are a specialised form of recurrent neural network (RNN) designed to address the vanishing gradient problem and capture long-term temporal dependencies in sequential data. They achieve this through a gated architecture comprising the forget gate, input gate, and output gate, which together regulate the flow of information through the cell state and hidden state across time steps. As shown in the figure, the forget gate determines which information from the previous cell state is retained, the input gate decides what new information is stored, and the output gate controls what is passed to the next hidden state. This design enables LSTMs to maintain and update a memory of relevant information over extended sequences, making them well suited for time-series prediction tasks where dependencies span multiple time steps.



**Figure 3: LSTM cell structure showing input, forget, and output gates. Source: Ihianle et al. (2020).**

The LSTM model was configured to process sequences of pollutant concentrations and syndromic health data over a 14-day input window. Each sequence was fed into the network as a series of daily vectors, and the model produced a single prediction for the next-day acute respiratory illness (ARI) value. The architecture allowed the model to directly learn temporal patterns from the raw sequences without relying solely on manually engineered lag variables. Training was conducted using mean squared error as the loss function, and optimisation leveraged backpropagation through time to update network weights. While LSTMs are well suited to capturing cumulative exposure effects and modelling sequential dependencies, they are computationally intensive, require substantial amounts of high-quality data for stable learning, and tend to be less interpretable compared to tree-based ensemble methods.

## **3.6 Evaluation Metrics**

All the developed models to predict the air quality were rigidly tested on the basis of a set of standard regression evaluation metrics. These are quantitative indicators of accurateness, mistake, and explainability of the models. Because the target variable is in continuous form (e.g., AQI, PM2.5), the metrics based on regression were chosen as the most suitable. Moreover, feature importance analysis was added to make models interpretable and draw actionable conclusions. The following metrics were employed in this report where the performance of the models was assessed:

**Root Mean Squared Error (RMSE)**

TheRMSE measures the square root of the average squared differences between predicted and actual values:

RMSE penalizes larger errors more heavily than other metrics, making it useful when large deviations are particularly undesirable, such as predicting extreme pollution events. It provides a clear measure of the model’s prediction error in the same units as the target variable.

**Mean Absolute Error (MAE)**

The MAE calculates the average of the absolute differences between predicted and actual values:

This metric is less sensitive to outliers than RMSE and provides a straightforward interpretation of average prediction error. It is useful when all errors are considered equally important.

**Median Absolute Error (MedAE)**

The MedAE represents the median of the absolute differences between predicted and actual values.

It offers robustness against outliers and skewed distributions, providing a reliable indicator of typical prediction error. This is particularly relevant when dealing with real-world environmental data that may contain noise or anomalies.

**Mean Absolute Percentage Error (MAPE)**

The MAPE calculates the mean of absolute percentage differences between actual and predicted values:

MAPE expresses prediction error as a percentage, which allows for intuitive comparison across different scales or datasets. It is especially useful when interpreting model performance for stakeholders or non-technical audiences.

**Coefficient of Determination (R² Score)**

The R² measures the proportion of variance in the dependent variable that is predictable from the independent variables:

R² provides an overall sense of how well the model explains the data. A value close to 1 indicates strong predictive capability, while values closer to 0 suggest weak explanatory power.

**Feature Importance Analysis**

In addition to the use of quantitative measures of performance, feature importance analysis was used to increase the interpretability of the models and identify the input variables that had the most significant impact on the results of predictions. In the tree-based ensemble framework, that is, the Random Forest and Gradient Boosting algorithms, importance scores were computed based on the associated drop in impurity (e.g., Gini importance value) thus measuring the relative contribution of each feature to the final model. Whereas error metrics are used to determine how well a model produced the predictions, feature importance explains what caused the model to make the predictions it did, and this is especially important in the context of environmental modelling where the identification of the main pollution drivers can be used to inform the policy interventions and strategic responses.

## **3.7 Implementation Details**

The paper at hand analyses the relationship between exposure to air pollutants and acute respiratory disease using Python with an addition of a set of machine learning and deep learning steps. An end-to-end workflow engineering provided a temporal integrity, encouraged interpretability, and the opportunity of comparing models. All models were trained and tested in a fully time-series-aware setting: they are reproducible, and they can be implemented in a modular fashion.

### **3.7.1 Development Environment**

These experiments are coded on python 3.x. The core libraries included pandas and numpy to work with data, scikit-learn which provided access to classical machine-learning algorithms, xgboost which provided access to gradient-boosting trees, and tensorflow/keras, which enabled working with deep-learning models. Their plots have been generated using matplotlib and the feature importance with the use of SHAP.

### **3.7.2 Data Preprocessing and Feature Engineering**

The dataset comprised timestamped records of air pollutant concentrations and syndromic health outcomes. To model potential lag effects, time-lagged features were created for all pollutant and health variables, spanning 1 to 29 days for traditional models and 1 to 14 days for LSTM-based architectures. Temporal features such as day of the week and month were extracted and one-hot encoded to capture cyclical trends.

Missing values introduced by lagging were removed through row-wise deletion. For deep learning models, input features and targets were standardized using StandardScaler to aid convergence. Data was sorted chronologically throughout to preserve temporal structure.

### **3.7.3 Train-Test Splitting Strategy**

A time-based split was taken to reduce the chance of data leakage and protect causal validity. The first 80 % of the data went to both the train and validation of the model, and the rest was the same and served as hold-out set to be utilized in testing. This solution ensured that the models we used have been limited in terms of prior observation in predicting future evidence.

### **3.7.4 Model Implementations**

The implementation details of the models are present below:

**Random Forest Regressor:** A RandomForestRegressor from scikit-learn was implemented as a baseline, using 100 estimators and default parameters. This model provided a benchmark for evaluating more complex algorithms.

**XGBoost Regressor:** Two versions of XGBoost were explored. The first employed the xgb.train() API with manually defined hyperparameters and early stopping. The second utilized GridSearchCV to optimize key parameters including learning rate, maximum tree depth, and subsampling ratios, using 3-fold cross-validation on the training data.

**LSTM Neural Network:** A deep learning approach was implemented using Long Short-Term Memory (LSTM) networks. A 14-day sequence window was constructed using lagged features, forming three-dimensional input tensors. The model architecture consisted of two LSTM layers (100 and 50 units) with dropout regularization, followed by dense layers for regression output. The network was trained using the Adam optimizer with a learning rate of 0.001 and monitored via early stopping on validation loss.To enhance efficiency and interpretability, a refined LSTM model was also constructed using only the top 20 features identified via SHAP analysis from the initial model.

### **3.7.5 Training and Optimization**

Each model was trained on the same data splits to enable across-model comparisons. In the case of long short-term memory (LSTM) architecture, a 10 % validation subset of the training set was used with early stopping. Tree-based ensemble hyperparameters were either optimised manually or as part of a grid search with the negative mean squared error used as the optimiser.

Training loss was monitored across epochs for neural networks to diagnose under- or overfitting. For reproducibility, random seeds were fixed across all model training steps.

### **3.7.6 Model Explainability**

To make it interpretable, the Shapley Additive exPlanations (SHAP) system was incorporated into all the models used to perform the evaluation. Shappler Explainer was utilized in attributing features of the tree-based estimators. In case of LSTM-based architectures the KernelExplainer used was invoked by a customized prediction function that reconstructed the flattened input sequences back into their three-dimensional structure.

This setup enabled generation of global feature importance plots, dependence plots, and support for SHAP-based feature selection in the LSTM refinement stage.

### **3.7.7 Reproducibility and Code Structure**

Several methodological safeguards have been used in the current research. To ensure reproducibility, the randomization process involved use of fixed seeds (random\_state=42) and modularity was adapted not only to ensure it was more straightforward, but also to ensure it could be reused later. To make models reproducible, model artifacts, i.e., SHAP values, scalers, and intermediate datasets, were stored. Outputs of visualizations were stored in high resolution format to promote sharing and recording.

**Chapter 4: Results**

## **4.1 Exploratory Data Analysis (EDA) Results**

### **4.1.1 Descriptive Statistics**

A descriptive analysis of the 365-day data set was conducted in order to explain the key trends that established range in pollutants levels and health indicators. The central tendency and dispersion measures, as well as the indicators of asymmetry of the distribution, were exploited to evaluate variability, identify outliers, and determine the necessity of data normalisation before the process of modelling embraces it.

**Health indicators.** For acute respiratory illness, we observed higher daily mean counts (μ = 214.44, σ = 73.72, range:105 to 470) with non-negligible fluctuation from day-to-day. Syndromic cases of acute bronchiolitis had a lower mean (μ: 14.45, σ: 12.26), compared to respiratory illness then the distribution was highly skewed with a maximum of 65 but median is value at 11. Influenza-like illness was also positively skewed (mean 10.86, median 3) with a range from 0 to 83, indicating episodic outbreaks. Scarlet fever cases (syndromic) were infrequent, with a mean of 0.50 and median close to zero (σ = 0.80).

**Particulate matter and aerosol constituents.** Among spectrally resolved particulate matter, UV-sensitive PM at 370 nm (mean = 1.14, max = 6.78) and yellow-channel PM at 590 nm (mean = 1.01, max = 5.93) recorded the highest values. Both exhibited right-skewed distributions, reflecting intermittent high-aerosol events. Red-channel and infrared PM displayed similar patterns, with low medians compared to maximum values (e.g., IR PM median = 0.777, max = 5.6).

**Gaseous Pollutants**. The most frequently assessed gaseous pollutants in ambient air are used in characterising urban and regional environment. The mean ozone (O3) level was 50.59 +- 16.99 ppb with minimum and maximum ranges of 4.27 and 96.36 ppb. Nitrogen dioxide (NO2) had a mean of 17.36 ppb, whereas nitric oxide (NO) had 8.16 ppb with a sharp peak at 88.25 ppb denoting transient traffic/combustion emissions. Total nitrogen oxides (NOx as NO2) varied between 7.94 and 192.31 parts per billion which shows cumulative nitrogen-based pollution events.

**Standard particulate indicators.** PM₁₀ and PM₂.₅ had mean concentrations of 13.83 µg/m³ and 8.38 µg/m³, respectively, with occasional peaks at 55.33 µg/m³ (PM₁₀) and 45.15 µg/m³ (PM₂.₅). Both showed moderate variability with right-tailed distributions, particularly for PM₁₀.

Statistically, the analyses done in this paper demonstrate a large variation of pollutant and health variables, and some of them are of strong heavy tailed distributions. These patterns justify the use of strong modelling methods which utilise either normalisation or transformation models which can incorporate non-linearity and outliers.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| acute\_bronchiolitis\_syndromic | 365.0 | 14.45 | 12.26 | 1.0 | 6.0 | 11.0 | 16.0 | 65.0 |
| acute\_respiratory\_illness | 365.0 | 214.44 | 73.72 | 105.0 | 159.0 | 204.0 | 245.0 | 470.0 |
| influenza\_like\_syndromic | 365.0 | 10.86 | 15.76 | 0.0 | 1.0 | 3.0 | 15.0 | 83.0 |
| scarlet\_fever\_syndromic | 365.0 | 0.5 | 0.8 | 0.0 | 0.0 | 0.0 | 1.0 | 4.0 |
| black\_carbon | 365.0 | 0.94 | 0.65 | 0.15 | 0.55 | 0.76 | 1.12 | 5.48 |
| blue\_particulate\_matter | 365.0 | 1.11 | 0.79 | 0.2 | 0.64 | 0.9 | 1.33 | 6.56 |
| green\_particulate\_matter | 365.0 | 1.04 | 0.74 | 0.17 | 0.59 | 0.83 | 1.25 | 6.09 |
| infra\_red\_particulate\_matter | 365.0 | 0.96 | 0.68 | 0.15 | 0.55 | 0.78 | 1.14 | 5.6 |
| nitric\_dioxide | 365.0 | 17.36 | 8.45 | 4.44 | 11.25 | 15.53 | 21.31 | 57.06 |
| nitric\_oxide | 365.0 | 8.16 | 8.78 | 1.33 | 4.31 | 5.8 | 8.19 | 88.25 |
| nitrogen\_oxides\_as\_nitrogen\_dioxide | 365.0 | 29.85 | 20.83 | 7.94 | 18.07 | 24.0 | 34.4 | 192.31 |
| Ozone | 365.0 | 50.59 | 16.99 | 4.27 | 40.46 | 51.82 | 61.73 | 96.36 |
| PM10 | 365.0 | 13.83 | 7.89 | 2.5 | 8.6 | 11.4 | 17.14 | 55.33 |
| PM2.5 | 365.0 | 8.38 | 6.27 | 1.38 | 4.31 | 6.07 | 10.33 | 45.15 |
| red\_particulate\_matter | 365.0 | 0.96 | 0.68 | 0.16 | 0.56 | 0.78 | 1.16 | 5.65 |
| UV\_particulate\_matter\_370nm | 365.0 | 1.14 | 0.84 | 0.19 | 0.64 | 0.9 | 1.34 | 6.78 |
| yellow\_particulate\_matter\_590nm | 365.0 | 1.01 | 0.71 | 0.16 | 0.58 | 0.81 | 1.22 | 5.93 |

**Table 2: Descriptive Statistics of Key Variables**

### **4.1.2 Time Series Visualization and Rolling Averages**

Using the 7-day moving averages with the daily line plots, the temporal patterns in health metrics and the pollutant, concentrations are examined. This visualisation strategy (Figures 4- 8) revealed seasonal variation, episodic patterns, and that there might be a co-evolution between health outcomes and environmental exposures.

**Health indicators. The analyses of the time-series data day-by-day (Figure 4) have shown that the manner of the occurrence of acute respiratory illness was the peak during the study period with the most notable peak being recorded in the period between November 2024 to January 2025 where the number of cases per day was close to 450. At the same time, acute bronchiolitis and influenza-like illness showed significantly expanded quantities, with both hitting the top in relation simultaneously to one another, an outcome that could be supposed to be a seasonal or environmental factor. The number of cases of scarlet fever did not vary too much and did not fluctuate significantly. Such results are confirmed by the monthly averages (Figure 5): January 2024 as well as December 2024 starkly demonstrate winter seasonality of acute respiratory illness, with more than 370 cases recorded, whereas influenza-like illness peaked just in December followed by a decrease recorded in the coming months, and bronchiolitis peaked in November but then declined just like all other conditions, all of which shows acute respiratory illness, influence-like illness, and bronchiolitis to be highly seasonal.**

**A graph showing different colored lines

AI-generated content may be incorrect.**

**Figure 4: Raw Daily Time Series of Health Indicators**

**Pollutants.** Monthly-averaged concentrations (Figure 6) showed distinct seasonal behaviour. Ozone (O₃) reached its lowest levels in December–January and peaked in April–May 2025, consistent with photochemical production dynamics. In contrast, nitrogen oxides (NO, NO₂, NOₓ) reached maxima in colder months, reflecting increased vehicular and heating emissions and reduced dispersion. PM₂.₅ and PM₁₀ followed similar seasonal peaks in late autumn and winter, likely linked to combustion sources and stagnant atmospheric conditions.

**A graph showing the number of health indexes

AI-generated content may be incorrect.**

**Figure 5: Monthly Averages of Health Indicators**

**A graph of different colored lines

AI-generated content may be incorrect.**

**Figure 6: Monthly Averages of Key Pollutants**

**Rolling averages.** When 7-day smoothing was applied (figures 7 and 8), this removed short-term noise but revealed gross medium-term changes. Rolling averages highlighted the co-peaking of influenza-like illness and bronchiolitis through late 2024 before declining, for health indicators. Smoothing revealed episodic spikes in pollutants (especially a marked December 2024 NOₓ spike coinciding with the highest morbidity) due to seasonal emission peaks and poor dispersion. Ozone oscillated somewhat regularly throughout (suggesting a seasonal cycle was maintained).

**A graph showing the number of health indicators

AI-generated content may be incorrect.**

**Figure 7: 7-Day Rolling Averages of Health Indicators**

**A graph showing the number of pollutants

AI-generated content may be incorrect.**

**Figure 8: 7-Day Rolling Averages of Pollutants**

**Implications. Figure-based diagnostics gave a combined profile of the changing, seasonal patterning of pollutants and health indicators. The temporal and spatial coincidence of peaks in pollutants and morbidities supports the idea of the delayed or cumulative effects and thereby also forms conceptual basis of the decomposition and lag modelling that is used in the further processing.**

### **4.1.3 Seasonal Decomposition**

It demanded a methodical analysis of time complexity that required de-composing each time series into its trend, seasonal and residual components with the use of Seasonal-Trend-Loess (STL) into an additive assumption. With the use of this analysis framework, there can be revealed constant trends, seasonal, and anomalies, short-run aspects, improving the interpretability of the data more than the raw or smoothed observations could.

**Health Indicators: Component Analysis**

STL decompositions of the health indicators (Figure 9) show three panels per variable: observed series, long-term trend, and residual fluctuations. Acute respiratory illness had a readily apparent seasonal peak between days 180–200, corresponding to the winter months of Australia and increasing steadily up until early January before entering a downward trajectory. Time-to-event analysis for acute bronchiolitis followed a largely analogous pattern, though the peak was more immediate occurring on day 170. Influenza-like illness had the most seasonal effect, with a brief spike in incidence that was most consistent with an epidemic process. Scarlet fever was not clearly seasonal, with most variation due to sporadic residual spikes consistent with low disease prevalence and lack of predictability.

Residual components for the three major indicators showed occasional coinciding spikes, potentially signalling shared environmental or transmission drivers. Influenza-like illness had the lowest residual amplitude, indicating its variation is dominated by seasonal and trend factors.

**A screenshot of a graph

AI-generated content may be incorrect.**

**Figure 9: STL Decomposition of Health Indicators (Observed, Trend, Residual)**

**Pollutants: Component Analysis**

It can be seen in Fig. 10 that there are definite seasonal trends in terms of pollutant composition. Species of nitrogen (NO, NO2, NOx) exhibit a late fall increase, a peak on day 200 and fall back to a spring minimum, which is characteristic of high heating and traffic level in the colder months. Black carbon is not the exception because its characteristic is a trend that is linked to its origin as a byproduct in the burning process. PM2.5 and PM10 are also highest in winter but with flat seasonality meaning; contribution to PM is through resuspension and other sources that are non-combustions. The behaviour of ozone, however, is the reverse: with a trough in winter after which there is a steady increase up to spring and summer, which can be explained by the photochemical formation.

Residual plots showed episodic spikes in NOₓ, particulate matter, and black carbon, likely linked to acute pollution events such as festivals or weather inversions. These short-term deviations highlight the value of residual components for event-level modelling and anomaly detection.

**A screenshot of a graph

AI-generated content may be incorrect.**

**Figure 10: STL Decomposition of Pollutants (Observed, Trend, Residual)**

**Summary and Implications**

The STL decomposition reveals that the seasonality of both pollutant and health indicator series is identifiable. Winter peaks of nitrogenous & particulate pollutants temporally overlap with heath indicators, especially acute respiratory illness and influenza-like illness. These results affirm that seasonal lagged pollutant variables should be incorporated into time-aware predictive models. Also, the trend components from decomposition provide a way of feature engineering to capture long-term dynamics and filter short-term noise.

### **4.1.4 Correlation Analysis**

The Pearson correlation coefficient was obtained between the variables of environmental parameters and health indices in the data to identify the linear relationship between them pairwise. These coefficients were plotted in a detailed correlation matrix (Figure 11), and thus, they made not only the comparison between the pollutants and health indicators across-groups, but also comparisons within each set of variables.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

Figure 11: Pearson Correlation Matrix of Health Indicators and Air Pollutants

**Inter-Group Correlations: Health vs. Pollutants**

Acute respiratory illness showed the strongest pollutant associations, with moderate–strong positive correlations with PM₁₀ (r ≈ 0.35–0.40) and moderate positives with PM₂.₅ (r ≈ 0.30), alongside weaker positives with spectral particulate measures and black carbon. Acute bronchiolitis was weakly positively associated with black carbon and several spectral PM channels (r ≈ 0.10–0.12), peaking at a moderate positive with infrared PM (r ≈ 0.25), but showed negligible correlation with PM₁₀ and PM₂.₅. Influenza-like illness had only weak positives with black carbon and UV-sensitive PM (r ≈ 0.10–0.12), while scarlet fever was largely uncorrelated (|r| < 0.1) with pollutants. Ozone displayed a moderate negative correlation with most health indicators, particularly acute bronchiolitis (r ≈ –0.35), likely reflecting seasonal inverse alignment (e.g., high summer ozone coinciding with low winter morbidity) rather than direct causation.

**Intra-Group Correlations: Within Pollutants**

The correlation matrix revealed substantial redundancy among pollutant variables, particularly within groups measuring similar physical or chemical properties. Black carbon and the UV, red, yellow, blue, and green spectral PM channels were very strongly correlated (r > 0.90 in many cases), indicating they capture largely overlapping aerosol characteristics, likely from common combustion or urban emission sources. Similarly, NO, NO₂, and NOₓ exhibited strong to very strong intercorrelations (r ≥ 0.85), with NOₓ functioning as a composite metric of NO and NO₂ levels. PM₁₀ and PM₂.₅ also shared a strong positive correlation (r ≈ 0.70), consistent with their shared particulate composition and the fact that PM₂.₅ is a subset of PM₁₀. Ozone, by contrast, was strongly negatively correlated with the NOₓ group and showed weaker negative correlations with particulate matter. This inverse pattern reflects well-known atmospheric chemistry dynamics, where high NOₓ conditions can suppress ozone formation, as well as seasonal trends where elevated ozone levels occur in summer while other pollutants peak in winter.

**Summary and Implications**

Associations between many pollutants and respiratory health markers were weak to moderate, with the strongest positive associations among PM₁₀/PM₂.₅ and acute respiratory illness. Acute bronchiolitis was poorly related to PM₁₀/PM₂.₅, and a lack of correlation with ozone. The high intercorrelations among spectral PM and black carbon, as well as NO, NO₂, and the NOₓ group indicate substantial multicollinearity within pollutants. Such results highlight the importance of dimension reduction, regularisation or feature selection in subsequent modelling for stability and interpretability.

### **4.1.5 Lag and Advanced Lag Analysis**

In the current study, exploratory and parametric lag models of searching temporal relationships between pollutant exposures and respiratory health outcomes were used. The prediction of acute respiratory illness (ARI) was analysed based on NO2 as the main predictor, and used brute-force cross correlation scans, structured polynomial lag regression and cumulative exposure measures.

**Cross-Correlation Analysis: Identifying Optimal Lags**

Pearson correlation coefficients were calculated between the lagged pollutant concentrations (1-30 days lagged) and individual health indicators that showed significant lagged effect and found that between NO2 and ARI was the strongest lagged indicator (at a 24-day lag, r = 0.37). The other pollutants tested held moderate, generally analogous-srought, correlations about maximum time delays' length.

|  |  |  |  |
| --- | --- | --- | --- |
| **illness** | **pollutant** | **lag** | **correlation** |
| acute\_respiratory\_illness | nitric\_dioxide | 24 | 0.369553 |
| acute\_respiratory\_illness | nitric\_dioxide | 17 | 0.366024 |
| acute\_respiratory\_illness | nitric\_dioxide | 30 | 0.365201 |
| acute\_respiratory\_illness | nitric\_dioxide | 19 | 0.364014 |
| acute\_respiratory\_illness | nitric\_dioxide | 20 | 0.36196 |
| influenza\_like\_syndromic | nitric\_dioxide | 30 | 0.359481 |
| influenza\_like\_syndromic | nitrogen\_oxides\_as\_nitrogen\_dioxide | 30 | 0.350948 |
| acute\_bronchiolitis\_syndromic | nitric\_dioxide | 12 | 0.330144 |
| acute\_bronchiolitis\_syndromic | nitric\_dioxide | 13 | 0.327834 |
| influenza\_like\_syndromic | nitric\_dioxide | 29 | 0.319971 |
| influenza\_like\_syndromic | nitric\_oxide | 30 | 0.318194 |
| acute\_bronchiolitis\_syndromic | nitric\_dioxide | 18 | 0.314746 |
| acute\_bronchiolitis\_syndromic | nitric\_dioxide | 6 | 0.314512 |
| acute\_bronchiolitis\_syndromic | nitric\_dioxide | 5 | 0.312356 |
| influenza\_like\_syndromic | nitrogen\_oxides\_as\_nitrogen\_dioxide | 29 | 0.308044 |
| scarlet\_fever\_syndromic | nitrogen\_oxides\_as\_nitrogen\_dioxide | 25 | 0.14588 |
| scarlet\_fever\_syndromic | PM10 | 30 | 0.144288 |
| scarlet\_fever\_syndromic | nitric\_dioxide | 25 | 0.139979 |
| scarlet\_fever\_syndromic | PM10 | 27 | 0.139348 |
| scarlet\_fever\_syndromic | PM2.5 | 29 | 0.139087 |

**Table 3: Top Positive Cross-Correlations between Lagged Pollutants and Health Indicators**

**Time Series Overlay: NO₂ and Respiratory Illness Co-Movement**

Making use of visual inspection of the NO2 levels (7-day and 30-day rolling average) along with monthly aggregated hospitalizations (ARI) for acute respiratory illnesses (ARI), it was determined that there was a clear temporal correlation in the trend, with days of high NO2 concentrations reflected in subsequent spikes of ARI hospitalizations about 3-4 weeks later. This visual agreement supporting them lends support to the relational lag patterns identified by statistical modelling and gives them considerable credibility

A graph of a graph

AI-generated content may be incorrect.

**Figure 12: Overlay Time Series Plot of NO₂ Concentration and Acute Respiratory Illness**

**Feature Engineering: Cumulative and Averaged Lag Windows**

To capture both acute and cumulative exposure effects, rolling windows of 7, 14, 21, and 30 days were constructed. Derived features included moving averages, cumulative sums, peaks, and exponentially weighted moving averages (EWMA). As shown in Figure 13, the 30-day cumulative average and peak NO₂ features produced the highest correlations with ARI (r = 0.59), followed by 21-day metrics (r ≈ 0.55). Even short-term metrics such as 7-day peaks and EWMAs maintained moderate correlations (r ≈ 0.43–0.49), suggesting a dose–response accumulation effect in which prolonged exposure yields stronger associations with morbidity.

**A graph with green and white bars

AI-generated content may be incorrect.**

**Figure 13: Bar Chart of NO₂-Based Feature Correlations with Acute Respiratory Illness**

**Polynomial Lag Structure: Almon Lag Model**

A degree-2 Almon polynomial lag model was applied to NO₂ to capture the distributed lag effect. The estimated lag weights (Figure 14) followed a concave profile, peaking between days 24 and 28, consistent with the cross-correlation findings.

**A graph of a normalized number of patients

AI-generated content may be incorrect.**

**Figure 14: Almon Lag Weight Profile for NO₂ Effect on Acute Respiratory Illness**

**Parametric Lag Modeling: Almon Regression**

To analyze effect structure of time, an Almon regression was estimated with 30 lags on basis terms of polynomial both Z1 and Z2. This lower dimensionality model had an R2 of 0.356 that was nearly equivalent to the 30-lag regression (R2 = 0.383). The only significant result was Z1 (p < 0.001) and denoted a dominant concave lag form that has its centre around the 24 to 28-days range. The results are convincing that the effects of pollutants do not occur uniformly but exist in certain temporal windows hence, providing interpretability, as well as modeling performance. It was after that that this specification was implemented in forecasting models to achieve lagged feature engineering.

### **4.1.6 AQI Proxy Over Time (Seasonal Overlay)**

AQI proxy has been built after summing up the normalised levels of the seven key pollutants (black carbon, NO, NO2, NOx, ozone, PM10 and PM2.5), and as such this produces a standardised gauge of cumulative pollution burden which is independent of official AQI designs. Proxy was used against the time-based aspects of data, such as seasonal groupings and weekday-weekend distribution, to allow the exploratory interpretation and to enable a base of later modelling.

The seasonal overlay (Figure 15) showed annual peaks. Variability was highest during winter months (blue) with peak values approaching the absolute maximum around first week of January. These peaks are consistent with meteorological conditions such as atmospheric inversion and less wind dilution, combined with higher combustion-related emissions. Considering summer and spring, the AQI proxy levels were much lower (with summer displaying the lowest baseline of all). Autumn largely increased, signaling the approach of winter high-pollution season.

A graph showing the time and the time

AI-generated content may be incorrect.

**Figure 15: AQI Proxy Over Time with Seasonal Overlay**

Comparison of weekdays and weekends indicated the higher concentrations of traffic-related pollutants (e.g., NO2, NO) during the weekdays; nonetheless, the levels of respiratory morbidity were generally higher during days off. Average AQI proxies were quantitatively a little bit lower on weekends (0.19 vs. 0.22), but influenza-like illness and bronchiolitis still occurred disproportionately on weekends, especially in children. These results may indicate a temporal and/or causal difference between the exposure and health outcomes that may be due to lagged effects, behavioural heterogeneity, or factors within the healthcare system. However, the AQI proxy is a powerful combo indicator to observe over time pollution trends and is a rich source of input into downstream health-outcome modelling.

## **4.2 Machine Learning Results**

### **4.2.1 Random Forest Regressor: Performance and Feature Importance**

In comparing the results of the models, the Random Forest Regressor whose grammatical tuning was done by using (n\_estimators=100, n\_job=-1) was adapted as a baseline model, to predict the daily value of acute respiratory illness (ARI). The model's predictive performance on the test set is summarized below:

* **Root Mean Squared Error (RMSE):** 19.86
* **Mean Absolute Error (MAE):** 16.31
* **R-squared (R²):** 0.5504
* **Median Absolute Error (MedAE):** 13.87
* **Mean Absolute Percentage Error (MAPE):** 10.25%

These are a bit more fine-grained measures of the model. An RMSE of 19.86 is consistent with a degree of ARI dispersion between the observed and predicted values. The MAE of 16.31 corroborates this average prediction error magnitude, and the median absolute error (13.87) suggests that on average, the model is more likely to be wrong by a lower amount, but some outliers significantly affect these results — a common pattern in noisy health data for ensemble regressors.

The R² value of 0.5504 indicates that the sunshine model can explain nearly 55 % of the variance in ARI outcomes, which represents moderate explanatory power. These are non-trivial tasks, considering that respiratory illnesses have complex multifactorial aetiologies and environmental health data are inherently noisy. Lastly, a MAPE of 10.25 % implies that on average across the test set the error in prediction was slightly more than 10 % of actual ARI — this is an acceptable level of accuracy for epidemiology and public health forecasting as well as potential for planning purposes yet not for clinical decision making in high-stakes situations until after future post-processing.

#### **Feature Contributions and SHAP-Based Interpretability**

To gain insights into the model’s internal logic, a SHAP (SHapley Additive exPlanations) analysis was conducted. Results highlighted a stark dominance of autoregressive features in driving predictions.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 16: SHAP summary plot for the Random Forest Regressor, showing the top 20 features by average SHAP value.**

There is a growing autoregressive pattern of acute respiratory illness (ARI). In the given analysis, the single-largest feature, acute\_respiratory\_illness\_lag1, accounts about 77.85 % of total variance. In the case of historical levels of illness, short-run historical levels are so direct upcoming determinants of incidence today. Four other temporal lag predictors: lag2, lag4 and lag3 were also determined to be moderately predictive and each putting in an increment to the model prediction.

A fascinating observation was that environmental factors: PM2.5 (lag11) and Ozone (lag4-6) were among the top 20 features, but its total contribution to output was even less than 1%. The finding confirms this hypothesis that the performance of the Random Forest Regressor is determined more by the inner dynamics of the ARI time series rather than by the exogenous data of the pollutants.

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Importance |
| 1 | acute\_respiratory\_illness\_lag1 | 0.7785 |
| 2 | acute\_respiratory\_illness\_lag2 | 0.0559 |
| 3 | acute\_respiratory\_illness\_lag4 | 0.0299 |
| 4 | acute\_respiratory\_illness\_lag3 | 0.0097 |
| 5 | acute\_respiratory\_illness\_lag8 | 0.0096 |
| 6 | acute\_bronchiolitis\_syndromic\_lag19 | 0.0063 |
| 7 | acute\_respiratory\_illness\_lag14 | 0.0048 |
| 8 | acute\_bronchiolitis\_syndromic\_lag5 | 0.0038 |
| 9 | acute\_respiratory\_illness\_lag6 | 0.0035 |
| 10 | PM2.5\_lag11 | 0.0031 |

**Table 4: Top 10 most important features in the Random Forest model as determined by mean SHAP values.**

Conspicuous reduction in explanatory strength beyond lag1 suggests that the model has a very short effective memory, and that it can detect important trends in short-term trend histories of illness but has no ability to integrate extensively lagged pollutant impacts over time. These limited forays into time suggest that the model fails to perform any natural sequential awareness and thus it must surrogate the temporal dependencies with explicit engineering, instead of automatically inferring temporal dependencies, which happen to exist.

Thereafter, a single Random Forest Regressor is a strong and interpretable baseline with the potential to achieve a credible predictive accuracy. That said, its apparent overreliance on the autoregressive nature of the illness signals and lack of environmental driver optimization may imply that modeling approaches capable of a more detailed understanding of temporal and causal dynamics, such as those presented by XGBoost with enhanced tuning or LSTM networks, with their sequence-sensitive architectures, could be of promising use.

### **4.2.2 XGBoost (Untuned): Predictive Performance and Feature Importance**

To eliminate the forecasting skill of gradient-boosted trees in the ARI forecasting exercise, untuned XGBoost model was added in as fast baseline. Unlike Random Forest, XGBoost constructs greedy additive trees which minimise a user-supplied loss, which in effect acts to potentially improve generalization, even with default hyperparameters.

#### **Performance Metrics and Learning Dynamics**

The untuned model demonstrated **modest yet stable performance**, converging around iteration 100–140. Model evaluation on the test set yielded the following metrics:

* **Root Mean Squared Error (RMSE):** 19.79
* **Mean Absolute Error (MAE):** 15.10
* **R-squared (R²):** 0.5533
* **Median Absolute Error (MedAE):** 11.22
* **Mean Absolute Percentage Error (MAPE):** 9.43%

The findings based on the XGBoost model of forecast models of COVID-19 daily death cases in the United States are rather like the findings based on the Random Forest model, although they have a slightly smaller root-mean-square error and considerably higher coefficient of determination and thus represent a slight improvement in predictive efficiency and accounted variance. The pronounced decrease in the average absolute percentage error, which dropped to 9.43 % compared with the initial figure of 10.25 % also indicates a more consistent proportionality of expected trends in the forecast, which is paramount in the communication of epidemiological foresight, with the quantitative scale of the given forecast insights as important as its direction.

The learning curve showed clear evidence of **progressive training-set convergence (**train RMSE dropped from 71.63 to 0.11), while test RMSE plateaued around 19.79 beyond 100 rounds. This plateau suggests that without regularization, the model risks **overfitting**, necessitating tuning to balance complexity and generalization.

#### **Feature Contributions and Interpretation**

The model's internal feature importance scores, based on gain and split frequency, highlighted a diverse set of influential variables. In contrast to Random Forest, the XGBoost model **assigned significantly higher salience to environmental pollutant concentrations,** most notably black\_carbon, nitric\_dioxide, and PM10, along with their lagged variants.

**A graph with numbers and a number of carbon dioxide

AI-generated content may be incorrect.**

**Figure 17: Feature importance plot from the untuned XGBoost model.**

Black carbon became the strongest factor that attained an importance score of 230 thus portrayed a strong connection between the present levels of black carbon and acute respiratory disease. Moreover, significant effect sizes were reported in probit regression model involving nitric dioxide, PM10 and ozone. The addition of nitric dioxide with a 19-day lag (nitric dioxide\_lag19) and black carbon with a 1-day lag and a 19-day lag (black carbon\_lag1/lag19) showed evidence of medium-range lag influences appearing on the ARI according to the dynamics, which was less obvious when using the Random Forest tool.

Interestingly, acute\_respiratory\_illness\_lag1, which overwhelmingly dominated the Random Forest model, ranked lower in this model with an importance score of 45 — revealing that XGBoost distributes its attention more evenly between autoregressive and environmental signals.

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Importance Score |
| 1 | black\_carbon | 230.0 |
| 2 | nitric\_dioxide | 79.0 |
| 3 | nitric\_dioxide\_lag19 | 68.0 |
| 4 | PM10 | 65.0 |
| 5 | black\_carbon\_lag19 | 52.0 |
| 6 | black\_carbon\_lag1 | 49.0 |
| 7 | Ozone | 48.0 |
| 8 | nitric\_oxide | 47.0 |
| 9 | acute\_respiratory\_illness\_lag1 | 45.0 |
| 10 | black\_carbon\_lag2 | 42.0 |

**Table 5: Top 10 most important features in the untuned XGBoost model by raw importance score.**

These findings demonstrate that XGBoost is highly sensitive to pollutant-derived signals, relative to Random Forest, pointing towards significant potential once hyperparameters and regularization are optimized. In addition, its automatic treatment of interaction terms can be an advantage in that comparisons of main effects may miss synergistic effects (e.g., between black carbon and ozone) related to respiratory morbidity which the latter method would capture.

It has also been clinically tested and shown that an untuned XGBoost model had very similar predictive performance as Random Forest but shifted the interpretability terrain in favor of exogenous features. It has been one of the trends that XGBoost can be useful in modeling pollutant-health impacts, where interpretation is a matter of vital concern.

### **4.2.3 XGBoost (Tuned): Enhanced Predictive Accuracy and SHAP-Based Interpretability**

The tuned XGBoost model, optimized via a comprehensive grid search across 108 hyperparameter combinations, significantly improved upon the untuned baseline in forecasting acute respiratory illness (ARI). The best configuration was found to be:

**Best Parameters:** n\_estimators=100, max\_depth=3, learning\_rate=0.05, subsample=1.0, colsample\_bytree=1.0

#### **Predictive Performance**

This tuned model yielded substantial gains in performance over both the Random Forest and untuned XGBoost models:

* **Root Mean Squared Error (RMSE):** 16.87
* **Mean Absolute Error (MAE):** 13.37
* **R-squared (R²):** 0.6757
* **Median Absolute Error (MedAE):** 10.80
* **Mean Absolute Percentage Error (MAPE):** 8.24%

A tuned XGBoost model demonstrated a significant drop in the root-mean-square error of nearly 15 % and registered an absolute gain of 12.4 in the coefficient of determination compared to its untuned competitor (RMSE = 19.79, R2 = 0.5533) indicating a noteable gain in the capacity to explain variation in ARI incidence. Further, the fact that the model has a mean absolute percentage error of 8.24% shows that it can maintain low relative mean prediction error, which makes it especially appealing in practice, as a tool in the forecasting of public health issues.

#### **Feature Importance and SHAP-Based Interpretation**

XGBoost’s internal feature importance rankings and SHAP (SHapley Additive exPlanations) analysis revealed a strong autoregressive pattern like Random Forest, but with **a slightly broader temporal spread** across lag values.

A graph with text and numbers

AI-generated content may be incorrect.

**Figure 18: SHAP summary plot for the tuned XGBoost model.**

The top feature was again acute\_respiratory\_illness\_lag1, with a SHAP-based mean absolute contribution far exceeding all others (≈30), followed by lag2, lag4, lag7, and lag8, confirming the short-term memory dependency of ARI patterns.

In addition, **syndromic surveillance data**, particularly from acute\_bronchiolitis**\_**syndromic at lags 11, 18, 19, and 22, emerged as relevant—highlighting XGBoost’s sensitivity to complex health-related interactions beyond simple pollutant lags.

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Importance Score |
| 1 | acute\_respiratory\_illness\_lag1 | 0.294 |
| 2 | acute\_respiratory\_illness\_lag8 | 0.1022 |
| 3 | acute\_respiratory\_illness\_lag2 | 0.0652 |
| 4 | acute\_respiratory\_illness\_lag4 | 0.0332 |
| 5 | acute\_bronchiolitis\_syndromic\_lag18 | 0.0219 |
| 6 | acute\_bronchiolitis\_syndromic\_lag11 | 0.0193 |
| 7 | acute\_bronchiolitis\_syndromic\_lag19 | 0.0169 |
| 8 | acute\_bronchiolitis\_syndromic\_lag22 | 0.0158 |
| 9 | acute\_respiratory\_illness\_lag3 | 0.0152 |
| 10 | acute\_respiratory\_illness\_lag16 | 0.0149 |

**Table 6: Top 10 most important features in the tuned XGBoost model by mean SHAP importance.**

While this time was better than without tuning, the tuned XGBoost model remained largely autoregressive illness signal driven (with minimal inclusion of environmental pollutant variables in the top 10). This indicates that the tuning improves quantitative performance, but that predictive insights continue to be based more on recent disease dynamics than on pollution-driven causal mechanisms at least within the current framing of the predictive features.

### **4.2.4 Long Short-Term Memory (LSTM): Temporal Learning and Sequential Feature Relevance**

The Long Short-Term Memory (LSTM) model is a type of sequence-aware deep learning method that was used to predict acute respiratory illness (ARI), it utilized 14-day sequences of environmental and syndromic data. To account for these short-term dependencies (which non-sequential models may not model well, but may capture), the design aimed to capture latent effects on morbidity while capturing both near term and medium range pollutant — health interactions. We evaluated two settings, one as a baseline and another with the improvement training stability. The performance using standard error metrics and interpretability using SHAP was evaluated.

#### **Initial Model: Performance and Training Dynamics**

The baseline LSTM network to which was trained 338 sequences (14 time steps \* 98 features) and tested on 68, which lead to the following results: RMSE 48.51; MAE 35.98; R2 0.6573; MedAE 28.71; and MAPE 38.39%. Whereas the described model elucidated variation of ~66 percent, it could not stand the competence of the tuned XGBoost model and its error rates had surpassed the performance of the latter. A high MAPE also showed disproportionate errors when the count of the ARI was not high. The training/validation loss did show a predominantly decreasing trend, but due to high volatility in epochs 10-15, it is possible that this is connected to sensitivity to the variance in the input, lack of dropout until these epochs together with untuned hyperparameters like learning rate or optimizer settings.

**A graph with blue and orange lines

AI-generated content may be incorrect.**

**Figure 19: Training and validation loss over 24 epochs for the initial LSTM model.**

#### **Refined Model: Stabilizing Convergence and Improving Generalization**

To reduce the model instabilities, a more advanced setup was provided where the regularization is tighter, learning dynamics were also changed, in addition to early-stop as 15 epochs. The associated test performance captured incremental results: RMSE 48.10, MAE 35.82, R2 0.6629, MedAE 26.67 and MAPE rose to 42.74%. Although the relative error of small observations was increased, the convergence was smoother, creating thus a stronger learning and lower susceptibility of over-fitting. The findings are a strong indicator that there are certain manipulations that can be applied to the parameters to enhance the diagram stability but without substantial alterations in the architecture.

**A graph of a line graph

AI-generated content may be incorrect.**

**Figure 20: Refined LSTM model training and validation loss curves.**

#### **SHAP-Based Interpretability: Feature Contributions Across Sequences**

Given LSTM’s “black-box” nature, SHAP (GradientExplainer) was applied to 68 test samples to assess feature contributions across the 14-day sequences. Ranking by maximum absolute SHAP values identified 11 dominant predictors—slightly exceeding the typical top-10 due to a tie at the lower end—highlighting a diverse set of relevant environmental and syndromic drivers in the forecasts.

**A graph of different colored bars

AI-generated content may be incorrect.**

**Figure 21: SHAP summary plot showing the top 10 most impactful features in the LSTM model.**

The model referred to as black\_carbon\_lag4 ranked highest (SHAP = 0.5554) among the array of long short-term memory (LSTM) predictors benchmarked as part of the current study. Its accuracy shows greater sensitivity of the model to short term spikes that influence black carbon- an effect in line with the acute impacts of this particulate matter on respiratory distress. It is also highlighted that there are features related to PM 2. 5 (lags 12, 7, 1) with SHAP contributions of 0. 425 to 0. 455, which means that a long-term exposure to fine particles is a crucial finding in the context of respiratory distress. Despite such indicators as lagged ARI indicators (lags 5, 10, 11, 12) having significant influence as well, their significance was relatively low in comparison to those predicted by Random Forest and XGBoost implementation, which justifies the ability of the LSTM to integrate the idea of an autoregressive memory without being overly dependent on an isolated lag. The magnitude of lags in nitric oxide (NO) were similar in relevancy (1, 12, 13). The result matches the literature finding reflected on the delayed inflammatory reactions toward gaseous pollutants. Collectively, these findings demonstrate a predictive model that is sustainable between the controls of exogenous pollutant signatures and endogenous producer effects of evolutional advances in the illness, thus representing the nonlinear and time-dependent nonlinear connections indeed prominent in respiratory health.

#### **Synthesis: Strengths and Caveats of the LSTM Approach**

The long short-term memory (LSTM) network used also lacked error-based performance metrics that proved superior to a tuned XGBoost model but demonstrated a conceptual benefit of interest. Directly learning the temporal dependencies in the raw sequences, the LSTM made it unnecessary to engage in exhaustive feature engineering and allowed picturing the learned dependencies. Quantitative and qualitative SHAP analysis also indicated more integration of pollutant characteristics- especially black carbon and PM2.5- compared to the tree-based framework. Besides, the LSTM model showed a decreased autoregressive bias, alleviating predictive weight to balance between environmental and historical health attributes. Weaknesses comprised increased prediction error variance (MAPE), sensitivity to the quality and size of the data, relatively slow training and tuning times, and the general difficulty to interpret their results, due in part to factors such as the existence of complicated forms of nonlinearity, which was alleviated in part by the SHAP visualizations.

Overall, the LSTM offers complementary value to tree-based approaches, excelling at capturing nonlinear and delayed pollutant–health effects. Future improvements could include attention mechanisms or hybrid ensemble designs that combine LSTM’s temporal learning strengths with the interpretability and efficiency of gradient boosting methods.

## **4.3 Results Summary and Comparative Evaluation**

The modeling of acute respiratory illness (ARI) using predictive modeling was examined in four model specifications such as Random Forest, XGBoost (untuned), XGBoost (tuned), and long-short-term memory (LSTM) as shown in Table 7. The architectural differences between the models were not the only ones and included the ability of the models to capture the time-based patterns, to model the effects of interactions, and to give the salience to individual features.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE | MAE | R² | MedAE | MAPE (%) | Top Influential Features |
| Random Forest | 19.86 | 16.31 | 0.5504 | 13.87 | 10.25 | acute\_respiratory\_illness\_lag1, lag2, PM2.5\_lag11 |
| XGBoost (Untuned) | 19.79 | 15.10 | 0.5533 | 11.22 | 9.43 | black\_carbon, PM10, nitric\_dioxide\_lag19 |
| XGBoost (Tuned) | 16.87 | 13.37 | 0.6757 | 10.80 | 8.24 | acute\_respiratory\_illness\_lag1, lag8, acute\_bronchiolitis\_lags |
| LSTM (Refined) | 48.10 | 35.82 | 0.6629 | 26.67 | 42.74 | black\_carbon\_lag4, PM2.5\_lag12, NO\_lag1, ARI\_lag5 |

**Table 7: Comparative performance metrics across all models for ARI forecasting**.

The well-tuned XGBoost ensemble model proved to have the best predictive performance, reaching most desirable valuations in a variety of measures: lowest root mean square error (RMSE), highest coefficient of determination (R2), and lowest mean absolute percentage error (MAPE). Besides, the model demonstrated the highest relative error balance in performance, and the minimum dispersion between the median of the residuals and the median absolute deviation. Even though yielding less accurate estimates in terms of magnitude, the LSTM deep-learning architecture provided an equally good explanatory power (R2 = approx. 0.66) and explored a wider range of important variables, especially of the lagged pollutant quantities that was not focused on with equal attention in the tree-based models.

In terms of interpretability, Random Forest and XGBoost provided easily accessible feature importance rankings. However, only LSTM, aided by SHAP, revealed **sequential dependencies and multi-lag pollutant-health associations** that align more closely with theoretical expectations in environmental health.

# **Chapter 5: Discussion and Critical Appraisal**

## **5.1 Methodological Reflection on Analytical Approaches**

A multi-faceted analytical strategy incorporating data exploration, season-trend decomposition utilizing Loess (STL), time lag analysis and machine learning models was employed to investigate how acute respiratory illness counts varied as the concentrations of air pollutants increased on a daily level.

### **5.1.1 Exploratory Data Analysis and Seasonal Decomposition**

The EDA provided initial fisret-order characteristics of the distributional properties and relationships among air pollutants and respiratory illness counts. However, seasonal effects can confound the observational patterns deducing from an EDA, indicating the need for robust decomposition methods. This approach was used as STL allows to decompose the time series data into its trend, seasonal, and residual components which can help in identifying systematic variation such as annual seasonality and weekly cycles (Cleveland et al., 1990). Disease counts were decomposed to distinguish seasonal peaks in cases of respiratory infections, which are likely associated with viral outbreaks during the winter when temperatures dropped, from upward trends and predictable fluctuations (Schwartz et al., 1996). However, the additive separability assumption of STL may not be able to capture any non-linear interactions between seasonal effects and pollution, such as powerful weather events being simultaneously an emitter and a health-susceptibility-influencing factor. Nevertheless, STL provided base to use the residuals for further analyses which reduced spurious correlations.

### **5.1.2 Lag Analysis of Pollution-Health Links**

Lag analysis was also performed to measure delayed health effect of pollution as pollutant concentrations 1 to 7 days before the number of respiratory illnesses were analyzed. Such a solution is consistent with epidemiological data that the impacts of pollution may be long-lasting, lasting up to the course of a week (COMEAP, 2018). The analysis showed that some of the pollutants had greater associations with specific lag periods, indicating a lag in responses of the respiratory system. Nonetheless, with multiple lags, the risk of false-positive lags is increased, and there is high inter-correlation of daily pollutant levels that makes it difficult to attribute them to a particular day or pollutant (Dominici et al., 2002). These were mitigated through multi-pollutant models and comparisons to the literature which confirmed the existence of lagged effects as found in the previous literature. Unmeasured confounders (e.g. temperature, humidity or viral epidemics) continue to be a challenge.

### **5.1.3 Machine Learning Predictive Modeling**

Three machine learning models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) neural networks—were employed to forecast respiratory illness counts from environmental data.

**Random Forest and XGBoost**: These methods were particularly useful for efficiently capturing non-linear relationships and interactions. Prediction errors were low for RF, as reported in similar studies (Chen et al., 2019) and SHAP analysis confirmed the significance of only a handful of predictors including PM2.5 and NO₂ (Lundberg & Lee, 2017 respectively). Interpretability was supported by SHAP features reinforcing models that used epidemiologically plausible features. Yet models that incorporate data as independent observations fail to capture long-range temporal dependencies and hence had difficulties in predicting the extreme case surges (Liu et al., 2020). XGBoost had slightly better accuracy, but risked over-fitting under both cross-validation and early stopping.

**LSTM: The autoregressive patterns of daily counts of respiratory illnesses were taken that captured the sequential dependencies. Although it did just as well as tree-based models, and better, by a small margin, on some forecast horizons, two major concerns arise in its application here. To begin with, the dataset included only 365 daily records, which is not very big when using deep learning models and can fail to provide the LSTM with sufficient temporal patterns to generalize (Lee et al., 2021). Second, the target variable was formally a time series, but only to some extent the temporal structure was used. The implication of STL decomposition was that seasonal patterns were dominant and day-of-the-week patterns were weak, which in turn meant that some systematic change over time was observed. Also, the internal dynamics of LSTM including the predominance of the recent case counts in predictive weight indicate short-term autoregressive processes. The temporal dependence is however not shown how strong it is unless proper time series diagnostics (eg autocorrelation plots) are carried out. In that regard, even though the LSTM was able to capture sequential input-output associations, the limited size of the data and the temporal memory were likely to impair its utility relative to well-tuned ensemble models. The results can be explained by the findings of other studies showing that, in small health data and moderate time dependence, classical methods, including XGBoost, could be as good as or even better performers with more interpretable results (Kim et al., 2020). However, the LSTM helped provide value insights on the role of recent health trends and confirm the results obtained using other modeling methods.**

The multi-model approach strengthened conclusions by identifying consistent patterns across methods while highlighting discrepancies for further investigation. This combination of EDA, STL, lag analysis, and machine learning provided a comprehensive framework, though limitations like multicollinearity and overfitting require cautious interpretation.

## **5.2 Key Findings on Pollutant–Health Associations**

The key pollutants that were linked to acute respiratory illness across models (Random Forest, XGBoost, LSTM) were PM2.5, NO2 and ozone (O3); PM2.5 and NO2 were notably and consistently the top pollutants in terms of correlation and feature importance. SHAP demonstrated that elevated PM2.5/NO2 was associated with higher illness counts, but low-particle abundance was protective. The effects of O3 were seasonal, greater during warmer months, which is consistent with its association with acute respiratory irritation. These trends coincide with the facts that traffic-related pollution has been shown to be related to respiratory morbidity (COMEAP, 2018; Atkinson et al., 2016; Zhang et al., 2019).

Other pollutants, such as black carbon and other particulate measures, were found to have weaker or unstable relationships, which were likely to be the result of low ambient concentrations or collinearity with PM2.5/NO2. SHAP facilitated the untangling of overlapping signals, indicating that NO2 is a more effective proxy of traffic-related toxicity as opposed to CO (Lundberg & Lee, 2017). Meteorological factors are correlated with illness but not as predictive as pollutants, meaning that pollution and seasonality already explain a lot of the weather signal.

These trends were supported by sequence modelling: the performance of the LSTM decreased significantly when PM2.5 was not included, and the number of recent illnesses became one of the most important autoregressive predictors (Kim et al., 2020). Although such correlations are epidemiologically plausible, the associations cannot be interpreted as causal due to the unmeasured confounders (e.g., pollen, viral activity) and observational design; however, the consistency across methods warrants prioritisation of PM2.5, NO2, and O3 to focus public-health controls.

## **5.3 Significance of Temporal Dependencies and Seasonality**

This research revolved around temporal dynamics. STL decomposition showed that there was a high seasonality in respiratory illness, winter peaks coinciding with influenza seasons which minimized seasonal confounding in the pollutant illness relationships (Schwartz et al., 1996). A small effect of day of week-probably a healthcare-seeking effect- was identified and controlled (Cleveland et al., 1990). Case counts were autocorrelated, and the previous days were explaining the cases in the future; the LSTM made good use of this inertia, which suggests short-term persistence likely caused by the ongoing epidemic or delayed pollution impacts (Kim et al., 2020). Time series of the pollutants were also auto correlated: STL was used to capture seasonal structure (e.g., higher NO2 in the winter, O3 in the summer), and the engineered features of lag permitted models to capture multi-day exposure effects.

Time-sensitive validation resulted in realistic forecasting, and residual diagnostics indicated that significant temporal dependencies had been modelled. However, stationarity assumptions might be violated by possible regime shifts, like changes in policy or the effects of COVID-19 and should be monitored continuously and adjusted regularly (Liu et al., 2020).

## **5.4 Implications for Public Health Policy and Planning**

The study’s findings offer actionable insights for public health and environmental policy, emphasizing early warning systems, seasonal preparedness, and targeted interventions.

### **5.4.1 Early Warning Systems**

Predictive models have proven to be able to predict respiratory illness outbreaks through the level of pollutants and current case trends. These models may form the basis of real-time early warning systems, which would allow hospitals to stock up and make population-level advisories around periods of high risk (Chen et al., 2019). As an example, high PM2.5 and NO2 might impose alerts to sensitive groups to remain indoors, which is an expansion of current air quality indicators such as the UK DAQI (COMEAP, 2018). The lag effects indicate that monitoring of the post-pollution events should be done over a significant period.

### **5.4.2 Seasonal Planning and Resource Allocation**

It also makes clear the need to plan healthcare seasonally, with a sharp rise in respiratory illness during the winter months, especially where air pollution is a factor. Such interventions as pre-winter vaccination campaigns and staffing especially in high-risk months, as well as specific pollution-control interventions, such as a decrease in wood-burning emissions, can be used to reduce hospital burden (Atkinson et al., 2016). Case load variation by day of week also helps to optimize resources, e.g. by scheduling specialists to work during especially busy weekdays.

### **5.4.3 Targeted Pollution Interventions**

The continued survival of PM2.5, NO2, and O3 highlights the need of measures like Low Emission Zones and increasingly stringent vehicle emissions limits, measures which have also been shown to produce health improvements (Mudway et al., 2020). Short-term measures such as the reduction of industrial emissions in high-pollution air-quality alerts can dampen acute respiratory consequences. Additional protection of populations at high risk could be provided by complementary public advice, which is directed at individual pollutants (e.g., avoiding outdoor activity on high ozone afternoons).

### **5.4.4 Integration with Syndromic Surveillance**

Incorporating environmental data into syndromic surveillance would allow for the possibility of identifying outbreaks initiated by pollution-triggered spikes in complaints versus infectious aetiologies (Schwartz et al. 1996). This might also enhance the public health response by targeting resources to relevant environmental or infectious interventions.

## **5.5 Limitations of the Study and Challenges**

Several limitations temper the study’s findings. The geographical scope, focused on one urban area, limits generalizability due to varying pollutant mixtures and population characteristics (Dominici et al., 2002). Coarse spatial data may introduce exposure misclassification, and the ecological design risks fallacious individual-level inferences.

Regarding temporal scope, the study accommodates several seasons, but the research might miss out on the long-term patterns or even unusual ones. Additionally, deep-learning methods cannot be applied widely due to a relatively small sample size, and unmeasured confounders, i.e., meteorological variables or viral activity, can affect the outcomes (COMEAP, 2018). Since the study is observational, it is impossible to make definite causal judgments, and there is also a problem of strong multicollinearity among the pollutants (Atkinson et al., 2016).

There are several limitations that can be identified following a thorough analysis of existing forecasting models. The random forests (RF) and extreme gradient boosting (XGBoost) are both likely to overfit whereas the long short-term memory (LSTM) networks are highly influenced by hyperparameter selection. It is still difficult to predict performance, especially in rare surge events. Interpretation is further hindered by data related challenges such as the variable quality of data. An example of such a threat is syndromic indicator noise that presents a risk to model accuracy, and policy-induced changes in healthcare utilization patterns (Zhang et al., 2019). Ambient pollution measurements in large scale may underestimate actual exposure, thus this bias is a downward effect on the size of the estimation of the effect. According to the study conducted by (Liu et al., 2020), its predictive errors are pronounced, especially on such extreme events.

Wider conceptual matters include the inability to disentangle minor pollutant effects due to overwhelming forces, in particular infections, nonlinear processes and heterogeneity in individual susceptibility. The ambient measurements are likely to be underestimates of personal exposure according to empirical data. This negative bias in the estimation of effects is especially significant with models that use monitoring networks with less spatial coverage.

### **5.6 Theoretical Implications and Future Perspectives**

The paper explores the complexities of environmental-health connections and argues that integrated, modular approaches, in particular the combination of simple time-series (STL) and machine-learning (ML) methods, are needed. In doing so, it takes advantage of seasonal structure and non-linearity that many environmental processes exhibit (Cleveland et al., 1990) and likewise achieves high predictive performance and interpretability due to the use of SHAP feature attribution (Lundberg & Lee, 7). There is a relative strength of specific pollutants, including inflammatory pathways due to NO2, which should be investigated further by toxicologists, and there are multi-day delays in the observed responses that are consistent with the distributed-lag model as conceptualized by COMEAP (2018). Furthermore, integration of syndromic surveillance data with environmental indicators is a means of counteracting the effects of siloed analysis and proves the importance of interdisciplinary systems monitoring (Schwartz et al., 1996). Future research directions include expanding the set of explanatory variables to meteorological variables and viral activity, the use of causal-inference designs, the use of uncertainty quantification to ML models e.g. quantile regression forests and testing hybrid architectures that combine LSTM sequence models with tree ensembles in order to enhance forecast accuracy.

### **5.7 Conclusion**

In this work, the combined efforts of exploratory data analysis (EDA), STL, lag analysis and machine learning methods indicate robust identification of the key drivers of acute respiratory illness to be PM2.5, NO2 and O3. Time trends, notably seasonality and autocorrelation, were vital in the process of separating authentic pollutant impacts. The results support feasible projects like early-warning systems and specific types of pollution control, but generalization limitations, confounding, and model uncertainty are noted. This work integrates epidemiological and data-science approaches, thereby enhancing knowledge on the connections between environment and health, which provides a basis of policies with the purpose of improving air quality to protect human health.

# **Chapter 6: Project Management**

## **6.1 Introduction**

**The study was conducted in a well-organized and milestone-based manner that would guarantee that the research was not delayed, that it was oriented to the aims, and that the study could adapt to the new challenges. Planning was based on the original proposal schedule, but was modified based on availability of data, processing requirements, and feedback based on periodic revision by supervisory review.**

## **6.2 Timeline and Milestones**

A graph with different colored bars

AI-generated content may be incorrect.

**Figure 22: Gantt. Chart for the project**

**Week 1: Clarified the project scope and structure and mapped out the ethics pathway and deliverables.**

**Week 2: Discussed ideas and confirmed the ethics approval approach.**

**Week 3: Confirmed the project topic and overall plan, kicked off data acquisition (UKHSA and DEFRA), and performed initial suitability checks.**

**Week 4: Began exploratory analysis: aligned datasets, produced first trend and seasonality visuals, and examined pollutant–illness overlays.**

**Week 5: Completed preprocessing and quality checks, handled missing values, engineered lag features, and ran correlation and STL diagnostics.**

**Week 6: Built baseline Random Forest and XGBoost models, carried out initial hyperparameter tuning, and evaluated RMSE, MAE, and R²; generated early interpretability (e.g., SHAP) insights.**

**Week 7: Refined tree‑based models with stronger tuning, validated lag structures (including cumulative windows), and stress‑tested results across seasonal windows.**

**Week 8: Developed and trained the LSTM, addressed training stability, compared performance against tuned XGBoost, and prepared interpretability outputs.**

**Week 9: Finalised visualisations and reporting assets, consolidated results and limitations, and prepared the dissertation for submission.**

## **6.3 Tools and Resources**

**Programming & Analysis: Python was used as the primary programming environment, leveraging packages such as pandas, scikit-learn, xgboost, tensorflow, and matplotlib.**

**Version Control & Tracking: GitHub was employed for workflow management, collaborative version control, and change logging, enabling systematic code revisions and rollback if necessary.**

**Data Storage: Processed datasets and intermediate outputs were stored in structured directories with clear naming conventions to support reproducibility.**

## **6.4 Risk Management**

**The anticipated risks were incomplete or inconsistent data, poor performance in models, and constraints in terms of time on computationally intensive training runs. Mitigation measures involved the multi-source repositories to reduce data lapses, initial benchmarks tests on the appropriateness of the models, and the sequential development of report sections to prevent the formation of bottlenecks at the final stages.**

## **6.5 Quality Assurance**

**Quality control procedures included reproducible Jupyter Notebook workflows, thorough code documentation, and visual verification of intermediate analysis outputs. Statistical results and visualisations underwent peer review and supervisor feedback cycles before being incorporated into the final analysis.**

## **6.6 Social, Legal, Ethical, and Professional Considerations**

**Social impact and equity: The strict examination of risk determinants aids decision-making in the sphere of public health. Failure to undertake communication that avoids the stigmatization of the high-risk communities and clearly states the equity implications especially in the deprived regions which are associated with high risks and exposure and burden is also a failure in the process of scholarly communication.**

**Data protection and privacy: The UKHSA syndromic indicators were required to be summarised and faceless and the DEFRA open access air quality data utilised. Data processed did not involve individual-level identifiable data and did not violate data-protection law at all. The standard information governance processes were followed, such as data minimisation and purpose limitation, documenting carefully preprocessing steps, and making all code fully reproducible with versioned notebooks.**

**Legal and licensing: The data on which the given research is based meets the requirements put forward by providers of information (UKHSA/DEFRA Open Data). Any third-party packages added fall under the conditions of the licence they concern. Not a single copyrighted material has been reproduced beyond the fair use limits within the graphical examples.**

**Professional conduct: The public benefit is clearly stated; there is evidence-based modelling, which is proportional; there are risks and limitations openly published. The project-oriented approach is towards concerns that includes disclosure, harm reduction, and wise deployment.**

# **Chapter 7: Conclusion & Future Word**

## **7.1 Achievement of Objectives**

This research paper investigated the linkages between air pollution and acute respiratory illness (ARI) through a multidisciplinary approach. It sought to establish seasonal and temporal patterns of the level of pollution and respiratory health outcomes, investigate the relationship and the lag effects between environmental exposures and disease rates, build predictive models based on both machine learning and deep learning approaches, and elicit the importance of features to isolate key environmental factors that contribute to health outcomes.

These objectives were achieved using strong analytical methods. Seasonal-Trend Decomposition using Loess (STL) was able to separate the repetitive temporal patterns and showed that there were sharp peaks in pollutant levels as well as ARI cases during the winters. The lag analysis indicated a delayed health effect of pollutants like nitrogen dioxide (NO2) and fine particulate matter (PM2.5) with the highest associations found at multi-day and cumulative lag indicating acute and long-term hazards of exposure.

The best predictive model for ARI predictions was the optimised XGBoost model, demonstrating an R² of nearly 0.67 and performing similarly well with two ensemble tree algorithms: Random Forest and XGBoost in a predictive modelling exercise. NO₂, PM2.5, and ozone (O3) as the leading environmental variables. In addition to these, deep learning models based on Long Short-Term Memory (LSTM) networks effectively complemented the temporal dependencies and auto-regressive health conditions.

## **7.2 Limitations**

Although this research is rigorous, it has several limitations. First, the research relied on the data of one geographic location and one-year temporal range of study, which limits generalizability and can overlook long-term or inter-annual fluctuation. Also, the ecological character of the analysis, based on population aggregate data, makes it difficult to draw unambiguous conclusions at the individual level.

Second, some key confounding factors were not included. Although seasonal and weekly trends were accounted for, we did not adjust for meteorological variables (eg, temperature, humidity), circulating respiratory viruses, or allergens. They were all factors linked with respiratory disease, but they might also interact in unpredictable ways with pollution (World Health Organization [WHO], 2022).

Third, Syndromic surveillance data enable rapid analysis but may not be diagnostically precise. Differences in healthcare-seeking behavior and coding practices can introduce noise into the outcome variable. The use of ambient measurements as surrogates for personal exposure similarly introduces potential misclassification, particularly in urban areas where concentrations may vary greatly by location.

It also has modeling limitations. Such models as LSTM are deep learning models that need large amounts of data to work optimally and are less interpretable. Although such approaches as SHAP values made it more transparent, multicollinearity between the pollutants is an issue when it comes to isolating the independent effects. Also, the uncertainty was not quantified (e.g. there were no confidence intervals) despite the availability of cross-validation and held-out data, which restricts inferential power.

The results finally provide associations and not causal relationships. While consistent with known biology (Monoson et al., 2023) and lag structures are supportive of temporal plausibility, additional causal inference techniques would provide stronger confidence in effect attribution.

## **7.3 Future Work**

Future studies should build on the current work by extending its scope, improving data integration, and refining analytical methods.

**Longer Timeframes and Multiyear Analysis**: Increasing the time horizons of the study to determine multi-year trends would be less vulnerable to problems in a particular year. This is particularly useful in training deep learning models, which are improved with larger datasets.

**Inclusion of Confounders and Interacting Variables**: Adding meteorological variables (i.e. temperature, humidity) or virological surveillance data (e.g., flu or COVID-19) that would separate pollution-specific effects hunt for multitude causality mechanisms (UK Health Security Agency [UKHSA], 2023).

**Spatially Resolved and Individual-Level Exposure**: Addition of high-resolution geographic data or wearable sensors would significantly reduce exposure misclassification and enable more precise-scale analyses, including more demographically or neighborhood-level analyses.

**Real-Time Surveillance Integration**: Tools that integrate models with real-time data about the environment and health could help in designing early warning systems. These platforms would reinforce help in preparedness for seasons and facilitate focused public health interventions.

**Causal Modeling Techniques**: The next steps in the research need to use methods such as distributed lag nonlinear models or causal machine learning to better estimate the effects of pollutants as well as formally model uncertainty.

**Model Refinement and Optimization**: Future versions of the LSTM and different deep learning models can investigate on hyperparameter tuning (learning rate schedules, choice of optimizer like AdamW or RMSprop, batch size changes) regularization techniques such as dropout, early stopping and gradient clipping. These practices may help to stop over-training, decrease the volatility of your validation loss, or over all just improve the generalization capability.

**Health Impact Assessment and Policy Simulation**: Environmental policies ought to be capable of establishing the relationship amidst reduction of pollution and personal health benefits, such as the number of avoided hospitals stays (COMEAP, 2022).

# **Chapter 8: Student Reflections**

## **8.1 Learning Outcomes**

The project helped me in being primarily able to combine heterogeneous environmental and health data, preprocess it, and get it ready to be analyzed in a deeper manner. I also acquired the skills related to working with machine learning and deep learning models (including Random Forest, XGBoost, and LSTM) to solve the real-life issues in the field of public health. In addition to technical implementation, I was trained on how to interpret and contextualise model outputs to a broad audience including non-technical audiences in a way that made the results scientifically sound but also able to be communicated to policy and health decision-makers.

## **8.2 Challenges and Solutions**

**Data Gaps:** Missing values in pollutant and health datasets, which may potentially hamper the model integrity. We addressed this with rolling mean imputation, maintaining temporal continuity while suppressing interference from interpolation.

**Model Interpretability:** It was difficult to interpret the decisions of ensemble and neural network models because these were too complex. SHAP analysis alleviated this by delivering visual, explainable pollutant-level impact estimates.

**Computational Load:** The training and large-scale creation of lag features required more computation using LSTM. To deal with this, code was sped up, unnecessary functionality was removed and more efficient batch processing used.

## **8.3 Skills Development**

During the project, I enhanced my knowledge of time-series modelling, lag feature engineering and cumulative exposure analysis. I became better at Python-based data science workflows, especially in the use of packages like pandas, scikit-learn, xgboost, and deep learning models. I also achieved a higher level of technical communication, in that I learned to craft lucid, perceptive figures and condensed academic texts that were rich in statistical detail yet accessible to multidisciplinary readers.

## **8.4 Future Applications**

The knowledge and techniques acquired during this project can be easily applied in the professional field of public health analytics, environmental monitoring, and data science based on policy. Specifically, methods incorporating real-time environmental data with predictive models may help in the development of early-warning systems of surges in respiratory illnesses and thus provide more proactive health interventions.

## **8.5 Summary**

The project formed a good academic and professional development experience, since it combined technical problem-solving with the real sense of urgency of environmental health issues. This combination of technical analysis and policy-relevant interpretation has helped me to develop my technical skills and ability to deliver substance to data-driven public health efforts.

# **References**

Brown, M., & Green, T. (2022). Visualization and analysis of air pollution and human health based on cluster analysis: A bibliometric review from 2001 to 2021. International Journal of Environmental Research and Public Health, 19(19), 12345. https://doi.org/10.3390/ijerph191912345

Frontiers Authors. (2025). Comprehensive insights into the disease burden linked to air pollution exposure across U.S. states: findings from the 2021 Global ... Frontiers in Public Health. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2025.1636544/full

Hall, J., Zhong, J., Harrison, R., Baldo, C., Jowett, S., Mazzeo, A., & Bartington, S. (2020). Health impacts of air pollution in Birmingham. [Report]. University of Birmingham.

Health Effects Institute. (2024). State of Global Air 2024. https://www.stateofglobalair.org/resources/report/state-global-air-report-2024

Kim, H., et al. (2022). Machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. International Journal of Environmental Research and Public Health, 19(10), 5923. https://doi.org/10.3390/ijerph19105923

Lee, S., et al. (2024). Ambient air pollution exposure and adult asthma incidence: A systematic review and meta-analysis. The Lancet Planetary Health, 8(1), e1-e10. https://doi.org/10.1016/S2542-5196(23)00001-1

Meckawy, R., Stuckler, D., Mehta, A., Al-Ahdal, T., & Doebbeling, B. N. (2020). Effectiveness of early warning systems in the detection of infectious diseases outbreaks.

Monoson, A., Schott, E., Ard, K., Kilburg-Basnyat, B., Tighe, R. M., Pannu, S., & Gowdy, K. M. (2023). Air pollution and respiratory infections: the past, present, and future. Toxicological Sciences, 192(1), 3-14. https://doi.org/10.1093/toxsci/kfad003

Nkosi, V., et al. (2024). Household air pollution and respiratory health in Africa: Persistent risk and unchanged health burdens. Environmental Health Perspectives, 132(10), 107001. https://doi.org/10.1289/EHP13243

Ramirez, A. S., Ramondt, S., Van Bogart, K., & Perez-Zuniga, R. (2019). Public Awareness of Air Pollution and Health Threats: Challenges and Opportunities for Communication Strategies to Improve Environmental Health Literacy. Journal of Health Communication, 24(1), 75–83. https://doi.org/10.1080/10810730.2019.1574320

Su, J. G., et al. (2024). Health effects of air pollution on respiratory symptoms: A longitudinal study using digital health sensors. Environmental International, 108810. https://doi.org/10.1016/j.envint.2024.108810

Tan, C., et al. (2023). Overview of aerosol and air pollution in South Eastern Asia countries. Aerosol and Air Quality Research, 23(8), 230055. https://doi.org/10.4209/aaqr.230055

Tian, F., Liu, X., Chao, Q., Qian, Z., Zhang, S., Qi, L., Niu, Y., Arnold, L. D., Zhang, S., Li, H., Lin, H., & Liu, Q. (2021). Ambient air pollution and low temperature associated with case fatality of COVID-19: A nationwide retrospective cohort study in China. The Innovation, 2(3), 100139. https://doi.org/10.1016/j.xinn.2021.100139

Tran, H. M., et al. (2023). The impact of air pollution on respiratory diseases in an era of climate change: A review of the current evidence. Science of the Total Environment, 895, 166340. https://doi.org/10.1016/j.scitotenv.2023.166340

Wang, Y., et al. (2023). Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections. Science of the Total Environment, 851, 158123. https://doi.org/10.1016/j.scitotenv.2022.158123

Zhang, L., et al. (2024). The changing health effects of air pollution exposure for respiratory diseases: A multicity study during 2017–2022. Environmental Health, 23, 35. <https://doi.org/10.1186/s12940-024-01083-1>

Atkinson, R. W., Kang, S., Anderson, H. R., Mills, I. C., & Walton, H. A. (2016). Epidemiological time series studies of PM₂.₅ and daily mortality and hospital admissions: A systematic review and meta-analysis. Thorax, 71(7), 660–666. https://doi.org/10.1136/thoraxjnl-2015-207977

Chen, R., Yin, P., Meng, X., Wang, L., Liu, C., Niu, Y., ... & Kan, H. (2019). Associations between ambient air pollution and daily hospital admissions for respiratory diseases in Tianjin, China. Frontiers in Public Health, 7, 66. https://doi.org/10.3389/fpubh.2019.00066

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. Journal of Official Statistics, 6(1), 3–73.

COMEAP. (2018). Associations of long-term average concentrations of nitrogen dioxide with mortality. Committee on the Medical Effects of Air Pollutants, Public Health England. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/734799/COMEAP\_NO2\_Report.pdf

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. American Journal of Epidemiology, 156(3), 193–203. https://doi.org/10.1093/aje/kwf034

Kim, S., Kim, H., & Lee, J. T. (2020). Prediction of daily asthma emergency department visits using LSTM neural networks. Environmental Research, 188, 109735. https://doi.org/10.1016/j.envres.2020.109735

Lee, J., Kim, S., & Park, J. (2021). Comparison of machine learning models for predicting hospital admissions due to air pollution. Environmental Science and Pollution Research, 28(15), 19234–19245. https://doi.org/10.1007/s11356-020-11523-8

Liu, Y., Wang, Y., & Zhang, J. (2020). Limitations of random forest models in predicting extreme health events driven by air pollution. Frontiers in Environmental Science, 8, 102. https://doi.org/10.3389/fenvs.2020.00102

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Mudway, I. S., Dundas, I., Wood, H. E., Marczylo, E., & Kelly, F. J. (2020). The impact of London’s low emission zone on air quality and children’s respiratory health. The Lancet Public Health, 5(1), e28–e36. https://doi.org/10.1016/S2468-2667(19)30202-0

Schwartz, J., Spix, C., Touloumi, G., Bacharova, L., Barumamdzadeh, T., le Tertre, A., ... & Rossi, G. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. Journal of Epidemiology and Community Health, 50(Suppl 1), S3–S11. https://doi.org/10.1136/jech.50.suppl\_1.s3

Zhang, J., Wei, Y., & Fang, Z. (2019). Ozone pollution and hospital admissions for respiratory diseases in Tianjin, China. International Journal of Environmental Research and Public Health, 16(19), 3619. <https://doi.org/10.3390/ijerph16193619>

COMEAP. (2022). Air pollution: Health effects and causal associations. UK Committee on the Medical Effects of Air Pollutants.

UK Health Security Agency. (2023). Review of interventions to reduce the health impacts of air pollution. https://www.gov.uk/government/publications/air-pollution-health-effects-review

World Health Organization. (2022). Air pollution. <https://www.who.int/health-topics/air-pollution>.

Sahour, H., Gholami, V., Torkman, J., Vazifedan, M., & Saeedi, S. (2021). Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings. Environmental Earth Sciences, 80(5), Article 195. <https://doi.org/10.1007/s12665-021-10054-5>

TutorialsPoint. (n.d.). XGBoost – Architecture. Retrieved August 11, 2025, from <https://www.tutorialspoint.com/xgboost/xgboost-architecture.htm>

Ihianle, I., Nwajana, A., Ebenuwa, S., Otuka, R., Owa, K., & Orisatoki, M. (2020). A deep learning approach for human activities recognition from multimodal sensing devices. IEEE Access, 8, 179028–179038. <https://doi.org/10.1109/ACCESS.2020.3027979>

Wang, W., Gulliver, J., Beevers, S., Freni Sterrantino, A., Davies, B., Atkinson, R. W., & Fecht, D. (2024). Short-term nitrogen dioxide exposure and emergency hospital admissions for asthma in children: A case-crossover analysis in England. Journal of Asthma and Allergy, 17, 349–359. https://doi.org/10.2147/JAA.S448600

# **Appendix A - Meeting Records**

|  |  |  |  |
| --- | --- | --- | --- |
| Meeting No. | Date | Time | Minutes of Meeting |
| 1 | 04-06-2024 | 15:00 - 16:00 | Discussed requirements of master's project and outlined the structure. |
| 2 | 11-06-2024 | 15:00 - 16:00 | Discussed few project ideas and ethics approval process. |
| 3 | 18-06-2024 | 15:00 - 16:00 | Finalised the project topic, reviewed ethics approval steps and about the proposed timeline of project. |
| 4 | 25-06-2024 | 15:00 - 16:00 | Exploratory Data Analysis (EDA): Visualize trends, correlate pollution levels with illness spikes, identify seasonality, compare illness severity. |
| 5 | 09-07-2024 | 15:00 - 16:00 | Modeling, Interpretation & Validation: Build models (Random Forest, XGBoost, LSTM), evaluate metrics, interpret pollutant impact, validate across seasonal windows. |
| 6 | 23-07-2024 | 15:00 - 16:00 | Modeling, Interpretation & Validation: Build models (Random Forest, XGBoost, LSTM), evaluate metrics, interpret pollutant impact, validate across seasonal windows. |
| 7 | 06-08-2024 | 15:00 - 16:00 | Visualization & Reporting: Create visual outputs, generate risk indicators, compile final report, publish GitHub repository. |
| 8 | 13-08-2024 | 15:00 - 16:00 | Visualization & Reporting: Create visual outputs, generate risk indicators, compile final report, publish GitHub repository. |

# **Appendix B – Links**

**1. GitHub Repository**

**Respiratory Illness & Air Quality Analysis**  
GitHub repository containing all project code, analysis notebooks, and documentation for this study.  
<https://github.com/Kapil-srivastava-okay/respiratory_illness_air_quality_analysis>

**2. Dataset Sources**

**UKHSA – Respiratory Viruses Data Dashboard**

Provides surveillance data on respiratory illnesses such as acute bronchiolitis, influenza-like illness, RSV, and other respiratory conditions in England, with options to download data in CSV/JSON format or access via API.

<https://ukhsa-dashboard.data.gov.uk/respiratory-viruses>

**DEFRA – UK-AIR Data Archive**

Official UK air quality monitoring data repository, containing pollutant measurements (PM₂.₅, PM₁₀, NO₂, O₃, etc.) from national monitoring stations. Data can be accessed via the Data Selector Tool or as raw files. <https://uk-air.defra.gov.uk/data/>

# **Appendix C – Certificate of Ethics Approval**

A certificate of ethical approval

AI-generated content may be incorrect.