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

Faculty of Engineering, Environment and Computing School of Computing, Electronics and Mathematics

MSc. Data Science

7150CEM

Data Science Project

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

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

Author: Kapil Srivastava

SID: 15105318

1st Supervisor: Dr. Stenford Ruvinga

2st Supervisor:

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

Academic Year: 2024/25

**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: 10/08/2025

|  |  |
| --- | --- |
| **First Name:** | **Kapil** |
| **Last Name:** | **Srivastava** |
| **Student ID Number:** | **15105318** |
| **Ethics Application Number:** | P187137 |
| **1st Supervisor Name:** | **Dr. Stenford Ruvinga** |
| **2nd Supervisor Name:** |  |

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 is one of the leading environmental health risks worldwide, contributing to millions of premature deaths annually and placing a substantial burden on healthcare systems. Short-term fluctuations in pollutant levels, particularly in densely populated urban environments, can trigger acute respiratory illnesses, exacerbating seasonal peaks in healthcare demand. Understanding these dynamics is essential for designing timely interventions, informing public health policy, and improving preparedness during high-risk periods.

This study explores the short-term effects of ambient air pollution on acute respiratory illness by integrating daily pollutant concentrations (PM₂.₅, PM₁₀, NO₂, NOₓ, O₃, black carbon) with syndromic health surveillance data over a one-year period. Seasonal-Trend Decomposition revealed pronounced winter peaks in both pollution and respiratory cases. Lag analysis identified strong delayed associations, particularly a 21–30 day cumulative lag of NO₂ showing the highest correlation with illness counts (r ≈ 0.59). 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. The findings demonstrate the value of combining statistical decomposition, lag structures, and machine learning to forecast respiratory health risks and support the development of pollution-informed early warning systems.

**Table of Contents:**

**Abstract**

**Chapter 1: Introduction**

**1.1 Background and Context**

**1.2 Problem Statement**

**1.3 Research Aim and Objectives**

**1.3.1 Aim**

**1.3.2 Objectives**

**1.4 Research Questions**

**1.5 Project Overview**

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

**2.1 Introduction**

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

**2.3 Geographical Variations in Air Pollution Impacts**

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

**2.5 Data Visualization in Environmental Health Research**

**2.6 Gaps and Opportunities for Further Research**

**2.7 Synthesis**

**Chapter 3: Methodology**

**3.1 Introduction**

**3.2 Dataset Overview and Collection**

**3.3 Data Preparation and Preprocessing**

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

**3.4.1 Descriptive Statistics**

**3.4.2 Time Series Visualization and Rolling Averages**

**3.4.3 Seasonal Decomposition**

**3.4.4 Correlation Exploration (Preliminary)**

**3.4.5 Lag and Advanced Lag Analysis**

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

**3.5 The Models**

**3.5.1 Random Forest Regressor**

**3.5.2 XGBoost**

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

**3.6 Evaluation Metrics**

**3.7 Implementation Details**

**3.7.1 Development Environment**

**3.7.2 Data Preprocessing and Feature Engineering**

**3.7.3 Train-Test Splitting Strategy**

**3.7.4 Model Implementations**

**3.7.5 Training and Optimization**

**3.7.6 Model Explainability**

**3.7.7 Reproducibility and Code Structure**

**Chapter 4: Results**

**4.1 Exploratory Data Analysis (EDA) Strategy**

**4.1.1 Descriptive Statistics**

**4.1.2 Time Series Visualization and Rolling Averages**

**4.1.3 Seasonal Decomposition**

**4.1.4 Correlation Analysis**

**4.1.5 Lag and Advanced Lag Analysis**

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

**4.2 Machine Learning Results**

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

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

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

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

**4.3 Results Summary and Comparative Evaluation**

**Chapter 5: Discussion and Critical Appraisal**

**5.1 Methodological Reflection on Analytical Approaches**

**5.1.1 Exploratory Data Analysis and Seasonal Decomposition**

**5.1.2 Lag Analysis of Pollution–Health Links**

**5.1.3 Machine Learning Predictive Modeling**

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

**5.3 Significance of Temporal Dependencies and Seasonality**

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

**5.4.1 Early Warning Systems**

**5.4.2 Seasonal Planning and Resource Allocation**

**5.4.3 Targeted Pollution Interventions**

**5.4.4 Integration with Syndromic Surveillance**

**5.4.5 Cost-Benefit and Policy Evaluation**

**5.5 Limitations of the Study and Challenges**

**5.6 Theoretical Implications and Future Perspectives**

**5.7 Conclusion**

**Chapter 6: Project Management**

**6.1 Introduction**

**6.2 Timeline and Milestones**

**6.3 Tools and Resources**

**6.4 Risk Management**

**6.5 Quality Assurance**

**6.6 Summary**

**Chapter 7: Conclusion and Future Work**

**7.1 Achievement of Objectives**

**7.2 Limitations**

**7.3 Future Work**

**Chapter 8: Student Reflections**

**8.1 Learning Outcomes**

**8.2 Challenges and Solutions**

**8.3 Skills Development**

**8.4 Future Applications**

**8.5 Summary**

**References**

**Appendices**

**List of figures:**

1. Architectural diagram for Random Forest.
2. Architectural diagram for XGBoost.
3. Architectural diagram for LSTM.
4. Raw Daily Time Series of Health Indicators.
5. Monthly Averages of Health Indicators.
6. Monthly Averages of Key Pollutants.
7. 7-Day Rolling Averages of Health Indicators.
8. 7-Day Rolling Averages of Pollutants.
9. STL Decomposition of Health Indicators (Observed, Trend, Residual).
10. STL Decomposition of Pollutants (Observed, Trend, Residual).
11. Pearson Correlation Matrix of Health Indicators and Air Pollutants.
12. Overlay Time Series Plot of NO₂ Concentration and Acute Respiratory Illness.
13. Bar Chart of NO₂-Based Feature Correlations with Acute Respiratory Illness.
14. Almon Lag Weight Profile for NO₂ Effect on Acute Respiratory Illness.
15. AQI Proxy Over Time with Seasonal Overlay.
16. SHAP summary plot for the Random Forest Regressor, showing the top 20 features by average SHAP value.
17. Feature importance plot from the untuned XGBoost model.
18. SHAP summary plot for the tuned XGBoost model.
19. Training and validation loss over 24 epochs for the initial LSTM model.
20. Refined LSTM model training and validation loss curves.
21. SHAP summary plot showing the top 10 most impactful features in the LSTM model.

**List of tables:**

1. Descriptive Statistics of Key Variables.
2. Top Positive Cross-Correlations between Lagged Pollutants and Health Indicators.
3. Top 10 most important features in the Random Forest model as determined by mean SHAP values.
4. Top 10 most important features in the untuned XGBoost model by raw importance score.
5. Top 10 most important features in the tuned XGBoost model by mean SHAP importance.
6. Comparative performance metrics across all models for ARI forecasting.

**Chapter 1: Introduction**

**1.1 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 originate from diverse sources, including vehicular emissions, industrial activities, domestic heating, and atmospheric photochemical reactions, and their distribution and health effects vary with meteorological conditions, topography, and seasonal cycles.

Urban environments are particularly vulnerable due to high population density and concentrated emission sources, while seasonal variations such as winter inversions and summer photochemical activity further influence pollutant behavior and population exposure. The burden on public health systems becomes especially pronounced during high-pollution periods that coincide with seasonal peaks in respiratory infections. While the links between pollution and health are well established, understanding the temporal structure and predictive potential of these relationships is increasingly important for public health preparedness.

Advances in real-time environmental monitoring and digital health surveillance have created new opportunities for data-driven forecasting and risk modeling. However, much of the existing literature continues to rely on coarse temporal granularity, a limited scope of pollutants, or statistical models that fail to capture the complex, non-linear, and lagged nature of pollution–health interactions. There is therefore a need for integrative approaches that combine robust statistical methods with interpretable machine learning to support both scientific understanding and practical policy implementation.

**1.2 Problem Statement**

Despite extensive research linking air pollution to respiratory morbidity, substantial gaps persist in year-long, high-resolution studies that combine traditional and emerging pollutant metrics with real-time syndromic health indicators. In particular, few studies leverage advanced machine learning to both forecast health risks and interpret pollutant-specific impacts over varying time lags. Additionally, the underutilization of modern visualization platforms hinders effective knowledge transfer to 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. This project aims to address these challenges by uniting diverse data streams and applying advanced analytical tools to enhance respiratory health surveillance and environmental health planning.

**1.3 Research Aim and Objectives**

**1.3.1 Aim**

The primary aim of this study is to analyze the temporal relationships between ambient air pollutants and acute respiratory health outcomes using a year-long, high-resolution dataset, and to develop predictive and interpretive tools that can support proactive public health interventions.

**1.3.2 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 through the use of interpretable machine learning methods such as SHAP and sensitivity analysis.

**1.4 Research Questions**

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

**Seasonality and Trends**

**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?**

**Pollution–Health Associations**

**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?**

**Predictive Modeling**

**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?**

**Pollutant Impact Interpretation**

**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?**

**1.5 Project Overview**

This dissertation investigates the temporal relationships between ambient air pollutants and acute respiratory health outcomes over a one-year period from June 2024 to June 2025 in an urban setting. It combines high-resolution environmental data—including conventional pollutants such as PM₁₀, PM₂.₅, NO₂, NOₓ, O₃, and black carbon—with spectrally resolved particulate matter data, alongside daily syndromic health surveillance indicators for acute respiratory infections, 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. In addition, the project incorporates visual outputs designed to support data-informed public health policy and the development of early-warning decision systems.

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

**2.1 Introduction**

Air pollution remains a pressing global public health concern, with well-documented associations with adverse respiratory health outcomes. Pollutants such as particulate matter (PM2.5 and PM10), black carbon, nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3) contribute to a spectrum of respiratory conditions, including asthma, acute respiratory infections, and chronic obstructive pulmonary disease (COPD). These pollutants, stemming from sources like vehicle emissions, industrial activities, and natural phenomena, trigger inflammation and exacerbate chronic diseases, posing significant challenges to public health systems worldwide. Climate change further amplifies these effects through increased frequency of extreme weather events that elevate pollutant levels (Tran et al., 2023). This literature review synthesizes findings from key studies to provide a robust foundation for a project leveraging a year-long dataset of daily air pollutant measurements and syndromic health indicators to explore temporal dynamics and health impacts using advanced machine learning and visualization techniques.

**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, especially PM₂.₅ and black carbon, remains a major focus due to its ability to penetrate deep into the lungs. Black carbon, produced by incomplete fossil fuel and biomass combustion, is associated with 6–9% increases in daily asthma-related hospital visits per interquartile range increase. Its small particle size enables alveolar deposition, triggering inflammation and oxidative stress, and potentially acting as a carrier for pathogens. Chronic exposure is linked to both respiratory and cardiovascular disease. The State of Global Air (2024) report attributes 8.1 million global deaths in 2021 to air pollution, with PM₂.₅ responsible for 4.7 million, many from respiratory illnesses.

In the West Midlands, UK, Walters et al. (1995) reported that a 10 µg/m³ increase in NO₂ correlated with a 5–8% rise in asthma-related hospital admissions, especially among children. While socioeconomic factors influenced rates, they did not fully explain the association, highlighting the role of environmental determinants. Other pollutants, including ozone and nitrogen oxides, also contribute to respiratory morbidity. Ozone, formed via photochemical reactions, can irritate airways, and a 10 ppb rise in daily maximum ozone is associated with a 1–3% increase in respiratory emergency visits. Nitrogen oxides from vehicle exhausts promote airway inflammation and are linked to higher hospital admission rates. These findings underscore the need for multi-pollutant studies to capture potential interactive health risks.

**2.3 Geographical Variations in Air Pollution Impacts**

Regional studies provide valuable insight into the spatial distribution of air pollution and its health consequences. In Birmingham, UK, Hall et al. (2020) applied concentration–response functions recommended by the Committee on the Medical Effects of Air Pollutants to assess the impact of PM₂.₅. Their analysis estimated that air pollution contributes to approximately 720 early deaths, 7,500 lost life years, and 900 new asthma cases annually among children and adults. Geographical disparities were evident, with central wards such as Tyseley & Hay Mills, Holyhead, and Aston experiencing pollution-related mortality rates of up to 8.5%. These areas, characterised by high population density and proximity to industrial and traffic-related sources, bear a disproportionate health burden.

Globally, findings vary by region. In the United States, a nationwide analysis of air pollution from 1990 to 2021 reported an 80.5% decline in PM₂.₅-attributable mortality but heterogeneous trends in health outcomes, with states such as California showing increases in diabetes-related disability-adjusted life years (DALYs) (Frontiers Authors, 2025). In Southeast Asia, Tan et al. (2023) found that urban exposure to particulate matter in Malaysia was associated with respiratory symptoms such as coughing and wheezing in children, largely linked to diesel emissions and industrial activity. In Africa, Nkosi et al. (2024) highlighted the persistent health burden from household air pollution due to reliance on biomass fuels such as charcoal and kerosene, contributing to significant respiratory disease rates, with 1.1 million air pollution-related deaths in 2019. A multicity study by Zhang et al. (2024) across 2017–2022 revealed that PM₂.₅ and O₃ exposure were associated with varying hospitalisation and mortality rates, with older populations and socioeconomically disadvantaged areas being more vulnerable.

The State of Global Air report (Health Effects Institute, 2024) underscores the particularly high burden in countries such as India and China, with 2.1 million and 2.3 million air pollution-related deaths, respectively, in 2021—many linked to respiratory conditions.

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

The advent of machine learning has transformed environmental health research by enabling the modelling of complex, non-linear relationships between air pollutants, meteorological factors, and health outcomes. Shi et al. (2020) employed a Chained Random Forest Classifier (CRFC) to predict respiratory virus presence across 31 regions in China from 2016 to 2021, integrating air quality indices and meteorological variables, and achieving an average accuracy of 0.76 with an AUC of 0.9. Wang et al. (2023) tested eight machine learning algorithms to predict outpatient visits for acute respiratory infections from 2018 to 2021, finding the random forest model most accurate and identifying NO₂ as a significant predictor with a one-day lag. Similarly, Kim et al. (2022) applied gradient boosting and Gaussian process regression to forecast respiratory disease occurrence, achieving R² values of 0.67–0.68, with temperature, humidity, PM₂.₅, and SO₂ emerging as key predictors.

A longitudinal study by Su et al. (2024) monitored medication use among 3,386 asthma and COPD patients in California from 2012 to 2019 using digital health sensors. The results showed positive associations between NO₂, PM₂.₅, and O₃ exposure and increased rescue medication use, with random forest models confirming these links. Exposure was estimated to result in a 23.9% increase in daily rescue puffs, corresponding to substantial economic costs.

These findings underscore the potential of machine learning to improve predictive accuracy and support public health planning. The adaptation of early warning systems (EWS) from infectious disease surveillance to environmental health, as evaluated by Meckawy et al. (2020), provides a promising framework. Integrating real-time air quality monitoring with health surveillance could enable forecasting of pollution-related health events, although challenges such as data standardisation and accessibility must be addressed.

**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). The Atmotube blog (2024) highlights techniques such as heat maps, time series plots, and colour-coded indices to convey air quality information, emphasising intuitive interfaces and colourblind-friendly palettes to reach diverse audiences. 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 can also be a valuable research tool. Brown and Green (2022), for example, applied bibliometric analysis with cluster visualisation to identify research hotspots in air pollution and health studies. The current project’s use of Power BI follows these principles, aiming to create an accessible platform for exploring temporal trends and spatial variations in both air pollution and health outcomes. Nevertheless, further evaluation is required to determine how different audiences interpret and interact with such visual tools to maximise their impact on policy and public awareness.

**2.6 Gaps and Opportunities for Further Research**

Despite progress in the field, important gaps remain. Many studies examine specific pollutants or broad health outcomes, with limited attention to the temporal dynamics of exposure. By using daily measurements over a year, the current project can identify seasonal patterns and short-term fluctuations in pollutant levels and their health impacts. Including a wide range of pollutants, including spectral particulate matter, also addresses gaps in multi-pollutant research.

The use of machine learning to predict specific respiratory conditions from multi-pollutant exposures requires further exploration. While studies such as Wang et al. (2023) and Kim et al. (2022) show promise, broader validation is needed to ensure generalisability. Interpretable models, including those using SHAP analysis, are essential for translating predictions into actionable policy. Adapting early warning systems from infectious disease surveillance to environmental health offers potential for enhancing risk prediction during high-pollution episodes (Meckawy et al., 2020), though challenges such as data integration and accessibility in low-resource settings remain.

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; Atmotube, 2024). 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.

**Chapter 3: Methodology**

**3.1 Introduction**

This study adopts a data-driven methodology to investigate the temporal relationship between air quality pollutants and respiratory health outcomes in the West Midlands, UK. Through structured data acquisition, systematic preprocessing, and staged analytical workflows, the dataset is transformed into a robust foundation for time series analysis, exploratory evaluation, and predictive modeling. The machine learning component is framed as a regression task, where the objective is to predict continuous numerical outcomes—daily counts of respiratory health cases—based on environmental exposure variables. Regression models such as Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks are applied to quantify and forecast pollutant–health relationships over varying temporal lags. The methodology emphasizes reproducibility, analytical rigor, and the use of advanced tools for temporal and statistical insight.

**3.2 Dataset Overview and Collection**

The dataset used in this study was compiled by merging two publicly available government-sourced repositories, ensuring both reliability and comprehensive coverage of the West Midlands region. This integration offers a multidimensional perspective by combining syndromic healthcare data with atmospheric pollutant measurements.

**The health data, obtained from the UK Health Security Agency (UKHSA), consists of daily records detailing emergency department attendances associated with four respiratory syndromes—acute bronchiolitis, acute respiratory illness, influenza-like illness, and scarlet fever. These indicators act as proxies for real-time community-level respiratory health trends and provide a consistent basis for monitoring changes in public health across the study period.**

**The air quality data, sourced from the UK Department for Environment, Food & Rural Affairs (DEFRA) Air Quality Archive, contains daily measurements from monitoring stations across the West Midlands. The pollutants measured can be categorised into two groups. The first group comprises conventional pollutants, including particulate matter (PM₂.₅ and PM₁₀), nitrogen dioxide (NO₂), nitric oxide (NO), combined nitrogen oxides expressed as NO₂ equivalents, ozone (O₃), and black carbon. The second group includes spectral particulate matter, which captures optical responses across a range of wavelengths such as blue, green, red, yellow (590 nm), ultraviolet (370 nm), and infrared particulate matter.**

**Overall, the dataset spans a continuous 12-month period from June 2024 to June 2025, encompassing 365 daily entries and 18 variables. This comprehensive structure enables detailed temporal analysis of air pollution and its potential correlation with short-term fluctuations in respiratory-related healthcare utilisation, forming the basis for subsequent exploratory, statistical, and predictive analyses.**

**3.3 Data Preparation and Preprocessing**

Preprocessing was conducted in Python using structured Jupyter notebooks for both the health and air quality datasets. The primary objective was to ensure data quality, maintain temporal alignment, and prepare the datasets for analytical use.

For the health data, preprocessing began by filtering the dataset to retain only the four respiratory syndromes relevant to the study: acute bronchiolitis, acute respiratory illness, influenza-like illness, and scarlet fever. Dates were parsed and standardised using Python’s datetime module to ensure consistency across records. Missing values were addressed using a seven-day rolling mean, preserving the temporal structure of the data. Column names were then renamed for clarity, and data validation checks confirmed that all values were numeric and that date sequences were continuous.

The air quality data required more extensive cleaning. Metadata such as headers, footnotes, and non-numeric rows were removed from the DEFRA CSV files. Placeholder strings like “No data” were replaced with NaN values before converting all readings to numeric format. Daily pollutant concentrations were calculated by averaging values across monitoring stations, and unit consistency was verified to confirm that all measurements were expressed in micrograms per cubic metre (µg/m³). Missing values were imputed using a seven-day rolling average, and final pollutant values were rounded to three decimal places.

Once both datasets were prepared, they were merged using the date field as the key, covering the period from 9 June 2024 to 8 June 2025. Verification steps ensured complete daily coverage and the absence of null values. The resulting dataset comprises 365 daily records and 18 variables, including one date field, four health indicators, and thirteen pollutant measurements, with all pollutant concentrations aligned to national environmental monitoring standards.

**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**

Descriptive statistics were computed to characterise the distribution of each variable in the dataset. For all pollutant and health indicators, measures of central tendency, including the mean and median, were calculated alongside measures of dispersion such as the standard deviation, minimum, and maximum values. Indicators of distribution shape, including skewness, were also derived. These metrics were used to identify the presence of outliers, evaluate variability over time, and inform decisions regarding data normalisation and transformation.

**3.4.2 Time Series Visualization and Rolling Averages**

Time series plots were generated for all variables across the full date range to visually assess temporal continuity and identify periodic behaviours or disruptions. Raw daily values were first plotted to capture the natural variability present in the data. To reduce short-term noise and highlight mid-term patterns, seven-day rolling averages were then applied to each time series. Visual comparisons between pollutant concentrations and health indicators were also conducted, providing an initial basis for hypothesis generation. This stage established a foundational understanding of time-dependent behaviour prior to undertaking decomposition and lag analysis.

**3.4.3 Seasonal Decomposition**

To separate and analyse the different components of each time series, Seasonal-Trend-Loess (STL) decomposition was applied to both pollutant and health indicator variables. Each series was broken down into trend, seasonal, and residual components using Python’s statsmodels package in additive mode. This approach made it possible to isolate recurring temporal patterns and distinguish them from underlying trends or random noise. The resulting decompositions were subsequently used to support feature engineering and enhance interpretability in the modelling process.

**3.4.4 Correlation Exploration (Preliminary)**

A preliminary correlation analysis was carried out to identify potential linear relationships among the variables. Pearson correlation coefficients were calculated between all pollutant and health indicators, and the results were visualised through a full correlation matrix and accompanying heatmaps. Intra-group correlations, such as those among different pollutants, were also examined to detect redundancy or shared environmental patterns.

**3.4.5 Lag and Advanced Lag Analysis**

To investigate the potential delayed effects of air pollutant exposure on health outcomes, a lag analysis was conducted. This involved applying simple lag shifts of one to fourteen days to the pollutant variables and computing cross-correlation functions (CCFs) to assess temporal dependencies between pollutants and health indicators. Compound lag windows, such as three-day and seven-day moving averages, were also considered to capture cumulative exposure effects. In addition, polynomial lag models, including Almon lags, were explored as an optional approach for structuring and interpreting distributed lag effects. The findings from this analysis informed the selection of lagged features for use in time-aware modelling techniques.

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

A custom proxy for the Air Quality Index (AQI) was developed by aggregating the normalised concentrations of selected pollutants. This metric was visualised as a seasonal time series plot to examine its variation across the study period. Overlay plots were also produced to compare the AQI proxy with health indicator timelines, without making direct interpretative claims. These visualisations were intended to support further hypothesis development and provide a clearer understanding of the relationship between air quality and respiratory health trends.

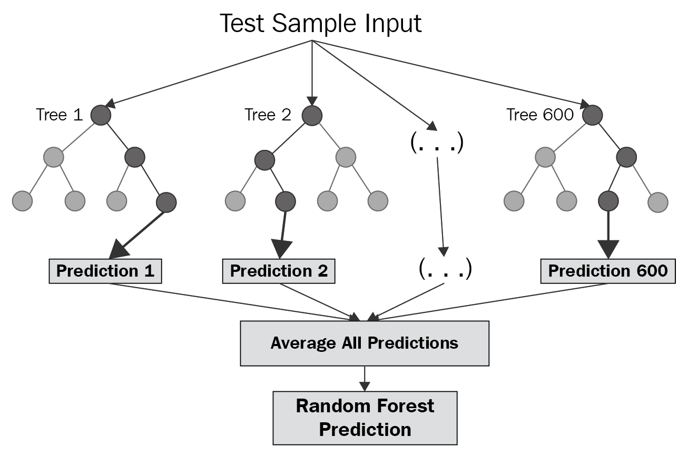
**3.5 The Models**

The aim of this modeling framework is to **predict acute respiratory illness (ARI)** using a variety of key air pollutants as predictive features. The pollutants considered include NO₂, NOx, PM₁₀, black carbon, ozone, and others—integrated in both current and lagged forms to capture their short- and medium-term effects. The selection of models spans across **ensemble-based machine learning models** and **recurrent deep learning architectures**, allowing for a comprehensive evaluation of both predictive performance and feature interpretability. Through these models, we assess not only the accuracy of ARI forecasting but also the **relative importance of each pollutant and time window** in contributing to health outcomes.

**3.5.1 Random Forest Regressor**

Random Forest is a bagging-based ensemble algorithm that constructs multiple decision trees during training and aggregates their predictions through averaging for regression tasks. In this study, a tuned RandomForestRegressor was implemented with parameters such as n\_estimators=100 and n\_jobs=-1 to balance model complexity and computational efficiency. Each pollutant-lag pair was treated as an independent feature, enabling the model to capture non-linear relationships and interactions. Although it does not inherently model temporal sequences, Random Forest performs reliably on tabular data with manually engineered lag variables.

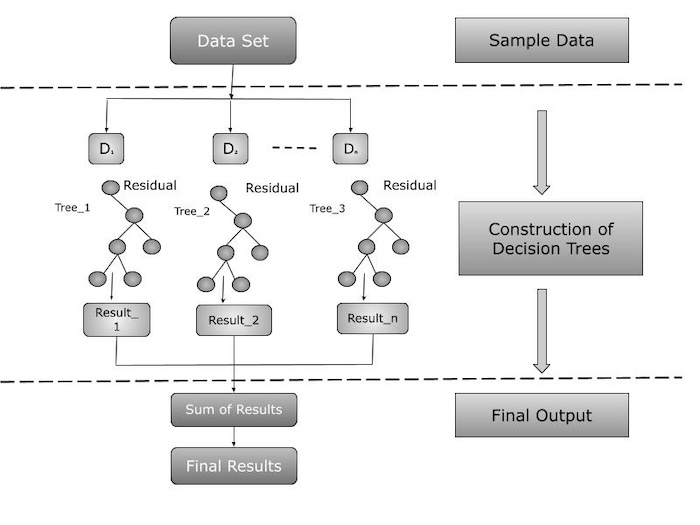
Its strengths include robustness to noise and overfitting, effective handling of mixed data types such as continuous variables and categorical encodings (e.g., seasons, weekdays), and the ability to provide interpretable feature importance rankings for pollutant-lag combinations. Limitations include the absence of native sequence modelling, which necessitates manual lag creation, and a potential curse of dimensionality when incorporating numerous lagged features across pollutants.



***Figure 1: Architectural diagram for Random Forest.***

**3.5.2 XGBoost**

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that builds additive models iteratively, with each stage correcting the errors of the previous one. Known for its high efficiency, scalability, and predictive power, it was applied in this study in two phases: first as an untuned version with default settings to establish a baseline, and then as a tuned model with optimised hyperparameters (e.g., max\_depth=6, eta=0.1, subsample=0.8) to improve performance. Operating on a tabular feature matrix containing current-day pollutant concentrations and multiple lagged values, XGBoost provides feature importance metrics that reveal which pollutants and lag periods are most predictive of acute respiratory illness (ARI).

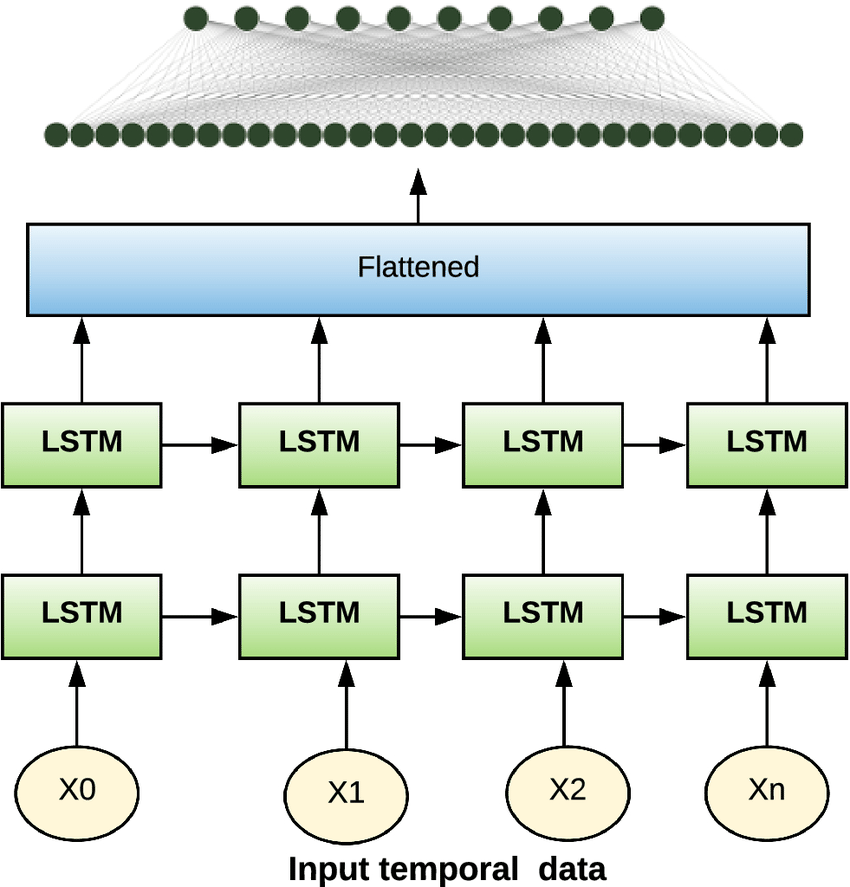


***Figure 2: Architectural diagram for XGBoost.***

Its strengths include superior predictive accuracy in structured data, effective regularisation through L1/L2 penalties to improve generalisation, and valuable insight into lag effects—such as identifying NO₂ lag 19 as highly relevant to ARI through SHAP analysis. Limitations include tuning complexity, as hyperparameters require careful optimisation to avoid under- or overfitting, and the lack of native sequential modelling, which necessitates manual lag feature engineering.

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

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to model long-term temporal dependencies in sequential data. In this study, the LSTM was used to process sequences of pollutant and syndromic health data over a 14-day input window, enabling the model to learn temporal patterns directly rather than relying solely on manually engineered lag features. The architecture included sequential input of 14-day pollutant and health indicator vectors, a single-output prediction for next-day acute respiratory illness (ARI) occurrence, and optimisation using mean squared error as the loss function. Post hoc interpretability was provided through SHAP analysis, identifying the most influential time-lagged features within the sequence. This design allowed the model to capture both short-term and medium-range temporal dependencies by integrating autoregressive health information with pollutant exposure patterns.



***Figure 3: Architectural Model for LSTM.***

The LSTM’s strengths include its ability to model sequential dependencies, eliminating the need for manual lag engineering, and its suitability for capturing cumulative exposure effects. While SHAP-based interpretation provides insights into influential lags and pollutant–health relationships, the model is computationally demanding, requires large, high-quality datasets for optimal performance, and remains less transparent than tree-based models.

**3.6 Evaluation Metrics**

To rigorously assess the performance and reliability of the models developed for air quality prediction, a set of standard regression evaluation metrics were employed. These metrics quantitatively capture the degree of accuracy, error, and explanatory power of the models. Given the continuous nature of the target variable (e.g., AQI, PM2.5), regression-based metrics were most appropriate. In addition, feature importance analysis was included to enhance interpretability and extract actionable insights from the models. In this report the following are used to evaluate the performance of the models the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Absolute Percentage Error (MAPE) and the Coefficient of Determination (R² Score). The metrics are briefly described below.

**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 quantitative performance metrics, feature importance analysis was used to improve model interpretability and identify the input variables that most strongly influenced predictions. In tree-based models such as Random Forest and Gradient Boosting, importance scores were derived from the reduction in impurity (e.g., Gini importance), reflecting each feature’s relative contribution to the model. While error metrics indicate how well a model performs, feature importance explains why it makes certain predictions—an essential consideration in environmental modelling, where understanding key pollution drivers can guide policy and intervention strategies.

**3.7 Implementation Details**

This study was implemented using Python, employing a suite of machine learning and deep learning techniques to investigate the relationship between air pollutant exposure and acute respiratory illness. The workflow was designed to maintain temporal integrity, support interpretability, and facilitate model comparison. All models were developed in a time-series-aware setting, with preprocessing, training, and evaluation conducted in a reproducible and modular fashion.

**3.7.1 Development Environment**

All experiments were conducted in Python 3.x. Core libraries included pandas and numpy for data manipulation, scikit-learn for classical machine learning, xgboost for gradient boosting, and tensorflow/keras for deep learning. Visualization was performed using matplotlib, while shap was employed for model explainability. Computations were run on a standard workstation with optional GPU acceleration for neural network training.

**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**

To prevent data leakage and uphold causality, a **time-based split** was used. The first 80% of the data was allocated for model training and validation, while the remaining 20% was held out for testing. This ensured that models only had access to historical data when making future predictions.

**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**

All models were trained on the same data splits to ensure comparability. For LSTM, a 10% validation split from the training set was used for early stopping. Tree-based models were optimized either manually or via grid search, using negative mean squared error as the scoring metric.

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 support interpretability, SHAP (SHapley Additive exPlanations) was integrated across all models. For tree-based algorithms, TreeExplainer was used to compute feature attributions. For LSTM models, KernelExplainer was applied using a custom prediction function that reshaped flattened input sequences into their original 3D 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**

Randomization was controlled using fixed seeds (random\_state=42), and code was modularized for clarity and reuse. Model artifacts, including SHAP values, scalers, and intermediate datasets, were saved for traceability and reproducibility. Visualization outputs were saved in high-resolution formats to support reporting and dissemination.

**Chapter 4: Results**

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

**4.1.1 Descriptive Statistics**

Descriptive analysis of the 365-day dataset revealed key characteristics of both pollutant concentrations and health indicators. Measures of central tendency, dispersion, and distribution asymmetry were used to assess variability, detect potential outliers, and determine the need for normalisation before modelling.

**Health indicators.** Acute respiratory illness showed the highest daily mean count (μ = 214.44, σ = 73.72) with values ranging from 105 to 470, indicating substantial day-to-day variability. Acute bronchiolitis syndromic cases had a lower mean (μ = 14.45, σ = 12.26) but a highly skewed distribution, with a maximum of 65 against a median of 11. Influenza-like illness was similarly skewed, with a mean of 10.86, a median of 3, and a range from 0 to 83, suggesting episodic outbreaks. Scarlet fever syndromic cases were rare, with a mean of 0.50 and median near 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.** Ozone (O₃) had a relatively high mean concentration (50.59 ppb) and moderate variability (σ = 16.99), ranging from 4.27 to 96.36 ppb. Nitrogen dioxide (NO₂) averaged 17.36 ppb, while nitric oxide (NO) averaged 8.16 ppb but peaked sharply at 88.25 ppb, suggesting short-term traffic or combustion spikes. Total nitrogen oxides (NOₓ as NO₂) ranged from 7.94 to 192.31 ppb, indicating cumulative nitrogen-based pollution episodes.

**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.

Overall, the statistics highlight substantial heterogeneity in both pollutant and health variables, with several showing heavy-tailed distributions. These patterns justify the application of normalisation or transformation techniques and the use of robust modelling approaches capable of handling 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 1: Descriptive Statistics of Key Variables***

**4.1.2 Time Series Visualization and Rolling Averages**

Temporal patterns in health indicators and pollutant concentrations were examined using daily line plots and 7-day rolling averages. This visualisation approach (Figures 4–8) revealed seasonal fluctuations, episodic peaks, and potential co-evolution between health outcomes and environmental exposures.

**Health indicators. Daily time series (Figure 4) showed that acute respiratory illness consistently had the highest incidence, peaking sharply between November 2024 and January 2025 with daily counts approaching 450. This increase coincided with marked rises in acute bronchiolitis and influenza-like illness, which peaked in near synchrony, suggesting a common seasonal or environmental driver. Scarlet fever cases remained sparse and largely stable. Monthly averages (Figure 5) confirmed these trends: acute respiratory illness reached over 370 cases in December 2024, influenza-like illness peaked in December, and bronchiolitis in November before declining—patterns indicative of strong winter seasonality.**

**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.** The application of 7-day smoothing (Figures 7–8) reduced high-frequency noise and made mid-term trends more visible. For health indicators, rolling averages emphasised the co-peaking of influenza-like illness and bronchiolitis in late 2024 before tapering. For pollutants, smoothing highlighted episodic surges, particularly a pronounced NOₓ spike in December 2024, likely caused by seasonal emission increases and poor dispersion, coinciding with morbidity peaks. Ozone maintained a semi-regular oscillation throughout, reinforcing its seasonal modulation.

**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. The figure-based diagnostics collectively highlight the dynamic and seasonal nature of both pollutant and health indicator patterns. The close temporal proximity of pollutant and morbidity peaks supports the hypothesis of delayed or cumulative effects, providing the basis for the decomposition and lag modelling described in later sections.**

**4.1.3 Seasonal Decomposition**

To disentangle temporal complexity, each time series was decomposed into trend, seasonal, and residual components using Seasonal-Trend-Loess (STL) under the additive assumption. This method enabled the identification of persistent trends, periodic patterns, and short-term anomalies not visible in raw or smoothed series alone.

**Health Indicators: Component Analysis**

STL decompositions of the health indicators (Figure 9) present three panels for each variable: observed series, long-term trend, and residual fluctuations. Acute respiratory illness showed a clear seasonal peak between days 180–200, aligning with winter months, with its trend rising steadily until early January before declining. Acute bronchiolitis followed a similar pattern, peaking slightly earlier at around day 170. Influenza-like illness displayed the sharpest seasonal effect, with a short, intense peak suggestive of epidemic dynamics. Scarlet fever lacked a defined seasonal trend, its variation driven mainly by sporadic residual spikes, reflecting low prevalence and unpredictability.

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**

The pollutant decompositions (Figure 10) reveal distinct seasonal profiles. Nitrogen-based pollutants (NO, NO₂, NOₓ) trended upward from early autumn, peaking near day 200, before declining into spring, reflecting increased heating and traffic emissions during colder months. Black carbon mirrored this pattern, consistent with its combustion origins. PM₂.₅ and PM₁₀ also peaked in winter but with flatter seasonal profiles, suggesting contributions from resuspension or non-combustion sources. In contrast, ozone exhibited an inverse cycle, with a winter trough and gradual increase toward spring and summer (around day 250), consistent with 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 confirms that both pollutant and health indicator series exhibit interpretable seasonality. Health indicators—particularly acute respiratory illness and influenza-like illness—coincide temporally with winter peaks in nitrogen-based and particulate pollutants. These findings support the inclusion of seasonally lagged pollutant variables in time-aware predictive models. Additionally, trend components from the decomposition offer a means of feature engineering that captures long-term dynamics while filtering short-term noise.

**4.1.4 Correlation Analysis**

To assess linear associations between environmental and health variables, Pearson correlation coefficients were calculated for all pairwise combinations in the dataset. These were visualised as a full correlation matrix (Figure 11), enabling both inter-group (pollutant–health) and intra-group (within pollutants or within health indicators) comparisons.

**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**

Correlations between pollutants and respiratory health indicators were generally weak to moderate, with the clearest positive relationships observed between PM₁₀/PM₂.₅ and acute respiratory illness. Acute bronchiolitis showed weak links to some optical PM measures but negligible association with PM₁₀/PM₂.₅, along with a moderate negative correlation with ozone. Within pollutants, very strong intercorrelations among spectral PM and black carbon, and among the NO, NO₂, and NOₓ group, highlight substantial multicollinearity. These results underscore the need for dimensionality reduction, regularisation, or careful feature selection in subsequent modelling to ensure stability and interpretability.

**4.1.5 Lag and Advanced Lag Analysis**

To examine temporal dependencies between pollutant exposure and respiratory health outcomes, both exploratory and parametric lag modelling techniques were applied. Analyses focused on NO₂ as a leading predictor of acute respiratory illness (ARI) and incorporated methods ranging from brute-force cross-correlation scans to structured polynomial lag regression and cumulative exposure metrics.

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

Pearson correlations between lagged pollutant values (1–30 days) and each health indicator were computed. The strongest lagged association was observed between NO₂ and ARI, peaking at a 24-day lag (r = 0.37). Moderate correlations were also observed across multiple lags for other pollutants.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 2: Top Positive Cross-Correlations between Lagged Pollutants and Health Indicators***

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

Visual inspection of NO₂ levels (7-day and 30-day rolling averages) alongside ARI cases revealed consistent temporal alignment, with illness peaks trailing NO₂ surges by approximately 3–4 weeks. This visual evidence supports the statistical lag relationships identified.

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 validate the temporal effect structure, an Almon regression with 30 lags was fitted using two polynomial basis terms (Z₁ and Z₂). This reduced dimensionality model achieved an R² of 0.356, close to the full 30-lag regression (R² = 0.383). Only Z₁ was statistically significant (p < 0.001), indicating a dominant concave lag shape centred in the 24–28 day range. These results confirm that pollutant impacts are concentrated in specific temporal windows rather than being evenly distributed, offering both interpretability and modelling efficiency. This specification was later adopted for lagged feature engineering in forecasting models.

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

A custom AQI proxy was developed by aggregating normalised concentrations of seven key pollutants—black carbon, NO, NO₂, NOₓ, ozone, PM₁₀, and PM₂.₅—providing a standardised measure of overall pollution load independent of official AQI scales. The proxy was analysed across temporal dimensions, including seasonal groupings and weekday–weekend patterns, to support exploratory interpretation and subsequent modelling.

The seasonal overlay (Figure 15) revealed distinct annual fluctuations. Winter months (highlighted in blue) exhibited the highest variability and peak values, with the absolute maximum occurring near day 210. These peaks align with meteorological factors such as atmospheric inversion and reduced dispersion, as well as increased combustion-related emissions. Summer and spring displayed lower and more stable AQI proxy levels, with summer recording the lowest baseline. Autumn showed a gradual rise, marking the transition toward winter’s high-pollution period.

A graph showing the time and the time

AI-generated content may be incorrect.

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

Weekday–weekend comparisons indicated higher concentrations of traffic-associated pollutants (e.g., NO₂, NO) on weekdays, yet slightly higher respiratory morbidity rates on weekends. Although weekend AQI proxy averages were marginally lower (0.19 vs. 0.22), illnesses such as influenza-like illness and bronchiolitis remained elevated. This divergence between exposure and health outcomes may reflect lagged effects, behavioural differences, or healthcare system factors. Overall, the AQI proxy offers a valuable composite measure for identifying temporal pollution patterns and serves as a useful input for downstream health-outcome modelling.

**4.2 Machine Learning Results**

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

The Random Forest Regressor, configured as a tuned baseline model (n\_estimators=100, n\_jobs=-1), was employed to forecast daily acute respiratory illness (ARI) counts using lagged air pollutant concentrations and prior illness indicators. 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 metrics offer a nuanced view of the model's capabilities. The RMSE of 19.86 indicates a moderate level of dispersion between predicted and observed ARI values. The MAE of 16.31 further confirms this average prediction error magnitude, while the **median** absolute error (13.87) implies that the model tends to make smaller errors more frequently, but is occasionally affected by larger deviations — a pattern typical of ensemble regressors on noisy health datasets.

The **R² value of 0.5504** demonstrates that approximately 55% of the variance in ARI outcomes is explained by the model, suggesting **moderate explanatory power**. This is non-trivial given the complex, multifactorial etiology of respiratory illnesses and the inherently noisy nature of environmental health data. Finally, a **MAPE of 10.25%** indicates that the average prediction error was slightly above 10% of actual ARI values — a reasonable level of precision for public health forecasting applications, though not sufficient for high-stakes clinical decision-making without further refinement.

#### **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.***

The most influential feature, **acute\_respiratory\_illness\_lag1**, accounted for approximately **77.85%** of the total feature importance. This underscores the autoregressive nature of ARI progression: short-term historical illness levels serve as the most immediate predictors of current-day incidence. Other moderately important features included lag2, lag4, and lag3, contributing incrementally to model performance.

Interestingly, environmental variables — such as **PM2.5 (lag11)** and **ozone (lag4–6)** — appeared among the top 20 features but held **very limited importance**, each contributing less than 1% to the model’s output. This suggests that the **Random Forest Regressor is far more responsive to internal dynamics of ARI time series** than to the exogenous air pollution features in this configuration.

|  |  |  |
| --- | --- | --- |
| 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 3: Top 10 most important features in the Random Forest model as determined by mean SHAP values.***

The steep drop-off in importance beyond lag1 suggests that the model maintains a **short effective memory**, performing well in detecting near-term illness trends but underutilizing longer-lag pollutant effects. This limitation reflects the model’s **inherent lack of sequential awareness**, where time-ordered dependencies must be engineered manually rather than learned dynamically.

In summary, the Random Forest Regressor serves as a **robust, interpretable baseline** with respectable predictive accuracy. However, its strong reliance on autoregressive illness signals, and limited exploitation of environmental drivers, highlights the potential value of models capable of learning richer temporal and causal structures — such as XGBoost with advanced tuning, or LSTM networks with sequence-aware architectures.

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

The untuned XGBoost model was deployed as a rapid baseline to evaluate the predictive capacity of gradient-boosted trees on the ARI forecasting task. Unlike Random Forest, XGBoost builds additive decision trees sequentially to minimize loss functions, which can enhance 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%

These results closely mirror those of the Random Forest model, with a **slightly lower RMSE and higher R²**, indicating a **marginal improvement in predictive accuracy and explained variance**. The drop in MAPE (from 10.25% to 9.43%) suggests **better proportional prediction fidelity**, a key consideration in epidemiological forecasting where both magnitude and directionality are critical.

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.***

The top-ranked feature, **black\_carbon**, achieved an importance score of **230**, far exceeding all others, suggesting a strong association between current black carbon levels and acute respiratory illness. Similarly, nitric\_dioxide, PM10, and ozone were among the most impactful variables. The inclusion of **nitric\_dioxide\_lag19** and **black\_carbon\_lag1/lag19** supports the hypothesis that **medium-range lag effects play a detectable role in ARI dynamics**, a finding less pronounced in the Random Forest model.

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

These findings suggest that even in its untuned form, **XGBoost is more sensitive to pollutant-derived signals** than Random Forest, offering potential for deeper insight once hyperparameters and regularization are optimized. Moreover, its implicit handling of interaction terms may help capture synergistic effects (e.g., between black carbon and ozone) that influence respiratory morbidity.

In summary, the untuned XGBoost model delivered comparable predictive performance to Random Forest while **shifting the interpretability landscape toward exogenous features** — a promising indicator for its suitability in pollutant-health impact modeling.

### **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%

Compared to the untuned XGBoost model (RMSE = 19.79, R² = 0.5533), the tuned version achieved a **15% reduction in RMSE** and an **absolute gain of over 12 percentage points in R²**, indicating a markedly improved capacity to explain variation in ARI incidence. Additionally, a **MAPE under 9%** reinforces the model’s robustness in maintaining low relative prediction error, making it especially promising for practical public health forecasting.

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

XGBoost’s internal feature importance rankings and SHAP (SHapley Additive exPlanations) analysis revealed a strong autoregressive pattern similar to 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 5: Top 10 most important features in the tuned XGBoost model by mean SHAP importance.***

Despite its improved accuracy, the tuned XGBoost model still leaned heavily on autoregressive illness signals, with environmental pollutant variables largely absent from the top 10. This suggests that, while tuning enhances quantitative performance, **predictive insights remain largely grounded in recent illness dynamics rather than pollution-based causal drivers**, at least under the current feature engineering regime.

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

The Long Short-Term Memory (LSTM) model was developed as a sequence-aware deep learning approach tailored to forecast acute respiratory illness (ARI) using temporally structured environmental and syndromic data. By operating over a 14-day input sequence for each prediction point, the model aimed to uncover not only short-term dependencies but also subtle medium-range pollutant-health interactions that non-sequential models might overlook.

To explore the learning capacity and generalization behavior of the LSTM model, two configurations were examined: an **initial baseline model** and a **refined architecture** with optimized training behavior. Performance was evaluated using both traditional error metrics and SHAP-based interpretability.

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

The initial LSTM configuration was trained over 24 epochs on a dataset of 338 sequences (each with 14 time steps and 98 features). Model performance was evaluated on a hold-out test set comprising 68 sequences. Results are as follows:

* **Root Mean Squared Error (RMSE):** 48.5050
* **Mean Absolute Error (MAE):** 35.9780
* **R-squared (R²):** 0.6573
* **Median Absolute Error (MedAE):** 28.7052
* **Mean Absolute Percentage Error (MAPE):** 38.3948%

These metrics indicate that the model explained approximately 66% of the variance in ARI outcomes, a moderately strong result considering the complexity and noisiness of environmental health data. However, the **absolute error values (MAE and RMSE)** were noticeably higher than those of tuned XGBoost, and the **high MAPE (~38%)** revealed that predictions often deviated significantly from the true scale—particularly when actual values were low.

**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.***

Despite the overall downward trend in training and validation loss, the volatility spikes around epochs 10–15 suggest some instability in optimization, potentially linked to sensitivity to input variance, lack of dropout regularization, and other hyperparameter factors such as learning rate or optimizer configuration. While these aspects were not explicitly tuned in this initial model, they represent potential avenues for improvement in future iterations.

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

To address the fluctuations observed in the initial training, a refined model was constructed with more conservative regularization and adjusted learning dynamics. The training process was shortened to 15 epochs, based on early stopping criteria and visual inspection of validation loss trajectory.

The refined model achieved the following test-set metrics:

* **Root Mean Squared Error (RMSE):** 48.1043
* **Mean Absolute Error (MAE):** 35.8182
* **R-squared (R²):** 0.6629
* **Median Absolute Error (MedAE):** 26.6723
* **Mean Absolute Percentage Error (MAPE):** 42.7393%

Compared to the initial model, the refined version achieved **slightly lower RMSE and MAE**, and a better MedAE—indicating reduced susceptibility to large outliers. However, MAPE slightly increased, pointing to **higher relative errors when true values were small**. This behavior is not uncommon in health outcome prediction where daily ARI counts may vary across a wide scale, and minor absolute errors can yield disproportionately large percentage deviations.

**A graph of a line graph

AI-generated content may be incorrect.**

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

Importantly, the refined model exhibited **markedly smoother convergence**, indicating more robust learning and reduced overfitting risk. These results affirm that **careful model tuning—even within deep learning architectures—can enhance stability and interpretability** without necessarily requiring drastic changes to architecture depth or sequence length.

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

Due to the inherent "black-box" nature of deep learning architectures, particularly recurrent models like LSTM, post hoc interpretability is critical for understanding and validating model behavior. To this end, we employed SHAP (SHapley Additive Explanations) to evaluate the **contribution of each feature across the full 14-day input sequences** used by the LSTM model.

Using GradientExplainer, SHAP values were computed for 68 test samples. The **maximum absolute SHAP values** were then used to rank feature importance, producing a bar plot that revealed a broad set of relevant inputs. Interestingly, the analysis yielded **11 dominant predictors**, instead of the commonly reported top-10, due to a tie in SHAP value magnitudes among lower-ranked features.

**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 most influential feature in the LSTM model was black\_carbon\_lag4, with a maximum SHAP value of 0.5554, indicating strong model sensitivity to short-term black carbon spikes—consistent with literature on its acute respiratory effects. PM₂.₅-related features (lag12, lag7, lag1) followed closely, with SHAP values between 0.425 and 0.455, underscoring the importance of sustained fine particulate exposure in respiratory distress.

Lagged ARI indicators (acute\_respiratory\_illness\_lag5, lag10, lag11, lag12) also appeared in the top 10, but with less dominance than in Random Forest or XGBoost models, reflecting LSTM’s ability to incorporate autoregressive memory without over-reliance on any single lag. Nitric oxide (NO) lags (lag1, lag12, lag13) were also influential, aligning with delayed inflammatory responses to gaseous pollutants.

The feature importance profile demonstrates that the LSTM leveraged both exogenous pollutant exposures and endogenous illness progression signals, capturing complex nonlinear and delayed relationships. Unlike tree-based models, which tended to focus heavily on autoregressive lags, the LSTM distributed attention more evenly between environmental and historical health features.

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

While the LSTM did not outperform the tuned XGBoost model in error-based metrics, it demonstrated several conceptual strengths. It achieved **temporal generalisation** by learning lagged relationships directly from raw sequences without the need for explicit feature engineering. **Pollutant integration** was evident in SHAP analysis, which confirmed the influence of black carbon and PM₂.₅—signals less prominent in Random Forest outputs. The model also exhibited **reduced autoregressive bias**, distributing predictive weight more evenly between environmental and endogenous health indicators compared to tree-based models. However, limitations included higher prediction error variance (particularly MAPE) due to sensitivity to sequence quality and limited training data, longer training and tuning times, and interpretability challenges—although SHAP analysis partially mitigated the latter.

Overall, the LSTM model offers valuable complementary insights to tree-based approaches, particularly in capturing nonlinear and delayed pollutant effects on respiratory illness. Future research could enhance its utility by incorporating attention mechanisms or hybrid ensemble architectures, combining the deep learning capabilities of LSTM with the interpretability and efficiency of gradient boosting methods.

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

The results from the predictive modeling of acute respiratory illness (ARI) across four model configurations—Random Forest, XGBoost (untuned), XGBoost (tuned), and LSTM—are summarized in Table 6. These models differ not only in architecture but also in their ability to capture temporal structure, variable interactions, and feature relevance. The comparison encapsulates their performance on multiple evaluation metrics and provides insight into their interpretive utility.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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, 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 6: Comparative performance metrics across all models for ARI forecasting.***

The tuned XGBoost model outperformed all others in terms of **predictive accuracy**, attaining the lowest RMSE and highest R². It also achieved the **lowest MAPE**, making it more consistent in relative error performance. Notably, the LSTM model, while less precise in magnitude-based metrics, offered **comparable variance explanation (R² ≈ 0.66)** and uncovered a **broader range of influential features**, especially pollutant-based lags not emphasized 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**

The study employed a multi-faceted analytical approach to explore the relationship between air pollutants and acute respiratory illness counts, integrating exploratory data analysis (EDA), seasonal-trend decomposition using Loess (STL), lag analysis, and machine learning models. Each method provided unique insights but also presented limitations that require careful consideration.

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

EDA offered initial insights into the distributional properties and correlations between air pollutants and respiratory illness counts. However, observational patterns from EDA can be confounded by seasonal effects, necessitating robust decomposition methods. STL was used to partition time series data into trend, seasonal, and residual components, effectively isolating systematic variations such as annual seasonality and weekly cycles (Cleveland et al., 1990). This decomposition clarified seasonal peaks in respiratory cases, likely tied to winter viral outbreaks, by removing predictable fluctuations (Schwartz et al., 1996). STL’s additive separability assumption, however, may miss nonlinear interactions between seasonal factors and pollution, such as extreme weather events influencing both emissions and health susceptibility. Despite this, STL provided a solid foundation for subsequent analyses by reducing spurious correlations.

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

To capture delayed health effects of pollution, lag analysis was conducted, examining pollutant levels 1–7 days prior to respiratory illness counts. This approach aligns with epidemiological evidence that short-term pollution effects can persist for up to a week (COMEAP, 2018). The analysis revealed stronger associations for certain pollutants at specific lags, suggesting delayed respiratory responses. However, testing multiple lags increases the risk of false positives, and high inter-correlation among daily pollutant levels complicates attribution to a specific day or pollutant (Dominici et al., 2002). Multi-pollutant models and literature comparisons were used to mitigate these issues, confirming lagged effects consistent with prior studies (COMEAP, 2018). Unmeasured confounders, such as temperature or viral epidemics, remain 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 tree-based models excelled at capturing non-linear relationships and interactions. RF achieved low prediction errors, comparable to similar studies (Chen et al., 2019), with SHAP (Shapley Additive Explanations) analysis identifying a small subset of key predictors, such as PM₂.₅ and NO₂ (Lundberg & Lee, 2017). SHAP enhanced interpretability, confirming the models’ reliance on epidemiologically plausible features. However, these models treat data as independent observations, potentially missing long-range temporal dependencies, and struggled to predict extreme case surges (Liu et al., 2020). XGBoost showed slightly better accuracy but risked overfitting, mitigated through cross-validation and early stopping.

**LSTM Sequence Model**: The LSTM modeled temporal sequences explicitly, capturing long-term dependencies and autoregressive patterns. It performed comparably to tree-based models, with slight advantages in certain forecast horizons, echoing findings that deep learning may not always outperform well-tuned classical models for short-term predictions (Lee et al., 2021). The LSTM highlighted the importance of recent respiratory case counts, aligning with studies showing strong autocorrelation in health outcomes (Kim et al., 2020). However, its data requirements and sensitivity to hyperparameters posed challenges, and its complex dynamics reduced interpretability compared to tree-based models.

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 study identified PM₂.₅, NO₂, and ozone (O₃) as the primary pollutants associated with acute respiratory illness, with PM₂.₅ and NO₂ showing the strongest correlations and feature importance across models. These findings align with evidence linking traffic-related pollutants to respiratory morbidity (COMEAP, 2018; Atkinson et al., 2016). SHAP analysis confirmed that higher PM₂.₅ and NO₂ levels consistently increased predicted illness counts, while low particulate levels had a protective effect. Ozone’s influence was season-dependent, significant in warmer months, consistent with its role in acute respiratory irritation (Zhang et al., 2019).

Other pollutants, like black carbon and particulate matters showed weaker associations, likely due to low ambient levels or collinearity with PM₂.₅ and NO₂. SHAP analysis helped disentangle these effects, suggesting NO₂ as a stronger proxy for traffic-related toxicity than CO (Lundberg & Lee, 2017). Meteorological factors, while correlated with illness, were less predictive than pollutants, indicating that pollution and seasonality capture much of the weather-related effect indirectly.

The LSTM reinforced these findings, with PM₂.₅ exclusion causing significant drops in predictive performance, and recent illness counts emerging as a critical autoregressive predictor (Kim et al., 2020). These results suggest that pollutants trigger acute responses within a context of temporal health inertia, where prior cases influence future ones.

While these associations are epidemiologically plausible, unmeasured confounders (e.g., pollen, viral activity) and observational design limit causal claims. The consistency with prior literature strengthens confidence in targeting PM₂.₅, NO₂, and O₃ for public health interventions.

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

Temporal dynamics were central to this study. STL decomposition revealed strong seasonal cycles in respiratory illness, with winter peaks aligning with influenza seasons (Schwartz et al., 1996). This ensured that pollutant effects were not confounded by seasonal trends. A modest day-of-week effect was also detected, possibly reflecting healthcare-seeking behavior, and was adjusted for in models (Cleveland et al., 1990).

Autocorrelation in illness counts, where prior days’ cases predicted future ones, was a significant finding, particularly in the LSTM, which leveraged this inertia effectively (Kim et al., 2020). This suggests short-term persistence in health outcomes, possibly due to ongoing outbreaks or lagged pollution effects. Pollutant time series also exhibited autocorrelation, with STL capturing seasonal trends (e.g., higher NO₂ in winter, O₃ in summer) and lag features allowing models to account for multi-day exposure effects.

Time-sensitive validation ensured realistic forecasting performance, with residual diagnostics confirming that major temporal dependencies were modeled. However, potential regime changes (e.g., policy shifts or COVID-19 impacts) could disrupt stationarity assumptions, requiring ongoing model adaptation (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 demonstrated the potential to forecast respiratory illness surges based on pollutant levels and recent case trends. Such models could underpin real-time early warning systems, enabling hospitals to prepare resources and issue public advisories during high-risk periods (Chen et al., 2019). For instance, elevated PM₂.₅ and NO₂ could trigger alerts for vulnerable populations to stay indoors, extending current air quality indices like the UK’s DAQI (COMEAP, 2018). The multi-day lag effects suggest prolonged monitoring post-pollution events.

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

The pronounced winter peak in respiratory illness, exacerbated by pollution, underscores the need for seasonal healthcare planning. Strategies include pre-winter vaccination campaigns, increased staffing during high-risk months, and targeted pollution controls (e.g., reducing wood-burning emissions) to mitigate hospital burdens (Atkinson et al., 2016). Day-of-week patterns could further optimize resource allocation, such as scheduling specialists for busier weekdays.

### **5.4.3 Targeted Pollution Interventions**

The prominence of PM₂.₅, NO₂, and O₃ supports policies like Low Emission Zones and stricter vehicular emissions standards, which have shown health benefits (Mudway et al., 2020). Temporary measures, such as restricting industrial emissions during forecasted high-pollution days, could reduce acute respiratory impacts. Public advisories tailored to specific pollutants (e.g., limiting outdoor activity during high O₃ afternoons) could further protect at-risk groups.

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

Integrating environmental data with syndromic surveillance could enhance outbreak detection by distinguishing pollution-driven spikes from infectious causes (Schwartz et al., 1996). This could sharpen public health responses, directing resources to environmental or infectious interventions as needed.

### **5.4.5 Cost-Benefit and Policy Evaluation**

The models enable quantification of health benefits from pollution reductions, supporting cost-benefit analyses for interventions like cleaner transport systems (COMEAP, 2018). Monitoring post-intervention health outcomes could validate policy effectiveness, ensuring alignment with public health goals.

## **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.

The study period, while capturing multiple seasons, may miss longer-term trends or rare events. Limited sample size constrained deep learning models like LSTM, and unmeasured confounders (e.g., meteorological factors, viral activity) could bias results (COMEAP, 2018). The observational design precludes definitive causal claims, and multicollinearity among pollutants complicates attribution (Atkinson et al., 2016).

Model limitations include potential overfitting in RF/XGBoost and sensitivity to hyperparameters in LSTM. Prediction errors, particularly for extreme events, highlight challenges in forecasting rare surges (Liu et al., 2020). Data quality issues, such as syndromic indicator noise or policy-driven changes in healthcare utilization, further complicate interpretations (Zhang et al., 2019).

Broader challenges include the complexity of isolating small pollution effects against dominant factors like infections, nonlinear relationships, and varying population susceptibility. Ambient pollution measurements may underestimate true exposure, biasing effect estimates downward.

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

This study underscores the complexity of environmental health relationships, advocating for integrated modeling approaches. Combining STL with machine learning leverages structured temporal patterns and complex interactions, suggesting a modular framework for future research (Cleveland et al., 1990). SHAP values bridge predictive accuracy and interpretability, supporting explainable AI in health studies (Lundberg & Lee, 2017).

The prominence of specific pollutants prompts toxicological inquiries into mechanisms, such as NO₂’s inflammatory effects. Multi-day lag structures align with distributed lag theories, refining exposure timing models (COMEAP, 2018). Integrating syndromic surveillance with environmental data challenges siloed approaches, advocating for interdisciplinary systems (Schwartz et al., 1996).

Future research should incorporate comprehensive data (e.g., meteorological, viral), explore causal inference techniques, and develop uncertainty quantification for ML models, such as quantile regression forests. Hybrid models combining LSTM and tree ensembles could enhance forecasting accuracy.

## **5.7 Conclusion**

This study robustly identified PM₂.₅, NO₂, and O₃ as key drivers of acute respiratory illness, using a comprehensive methodology integrating EDA, STL, lag analysis, and machine learning. Temporal dependencies, particularly seasonality and autocorrelation, were critical in isolating genuine pollutant effects. The findings support practical applications like early warning systems and targeted pollution controls, while acknowledging limitations in generalizability, confounding, and model uncertainties. By blending epidemiology and data science, this work contributes to understanding environmental health linkages, offering a foundation for policies to improve air quality and protect public health.

**Chapter 6: Project Management**

**6.1 Introduction**

**The research was managed using a structured, milestone-driven approach to ensure timely completion, alignment with objectives, and adaptability to emerging challenges. Planning followed the initial proposal timeline but was refined in response to data availability, computational demands, and iterative feedback from supervisory reviews. Regular progress assessments were conducted to track deliverables and allow for contingency measures where delays or technical obstacles arose.**

**6.2 Timeline and Milestones**

**Weeks 1–2: Data acquisition and preprocessing — downloading UKHSA health data and DEFRA pollutant datasets, cleaning, aligning by date and geography, and engineering lag features.**

**Weeks 3–4: Exploratory Data Analysis (EDA) to identify seasonality, pollution–illness correlations, and spatial/temporal patterns, alongside the generation of initial visualisations for hypothesis refinement.**

**Week 5: Predictive modeling and validation — Random Forest, XGBoost, and LSTM models evaluated using RMSE, MAE, and R², with SHAP analysis for interpretability.**

**Week 6: Visualization and reporting — production of illness–pollution overlays, static figures, and the final dissertation draft.**

**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**

**Anticipated risks included incomplete or inconsistent datasets, model underperformance, and time constraints related to computationally intensive training runs. Mitigation strategies involved sourcing from multiple repositories to reduce data gaps, conducting early benchmark tests for model suitability, and incrementally drafting report sections to prevent end-stage bottlenecks. Hardware resource limitations were addressed by using optimised batch processing and selective feature subsets during preliminary model runs.**

**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. Additionally, results from different modelling approaches were cross-compared to ensure consistency and robustness.**

**6.6 Summary**

**Effective project management enabled the study to progress in line with deadlines while maintaining analytical rigour. The combination of structured planning, proactive risk mitigation, and quality assurance processes ensured the delivery of reproducible, interpretable, and policy-relevant outputs for environmental health research.**

**Chapter 7: Conclusion & Future Word**

## **7.1 Achievement of Objectives**

This study aimed to examine the relationship between air pollution and acute respiratory illness through a multidisciplinary analytical lens. The key objectives were to (1) identify seasonal and temporal trends in pollutant levels and respiratory health outcomes; (2) explore associations and lag effects between environmental exposures and illness rates; (3) develop predictive models using machine learning and deep learning methods; and (4) interpret feature importance to isolate key environmental contributors to health outcomes.

Each objective was met with methodologically sound techniques. Seasonal-Trend Decomposition using Loess (STL) successfully disentangled regular temporal structures, enabling the detection of distinct winter peaks in respiratory illness and pollutant levels. Lag analysis confirmed that pollutants such as nitrogen dioxide (NO₂) and particulate matter (PM₂.₅) exhibit delayed health effects, with correlations peaking at multi-day and cumulative lags, suggesting both acute and sustained exposure risks.

Predictive modeling with ensemble tree methods (Random Forest and XGBoost) achieved strong accuracy, particularly for forecasting acute respiratory illness (ARI), with the tuned XGBoost model achieving an R² of approximately 0.67. These models consistently identified NO₂, PM₂.₅, and ozone (O₃) as the dominant environmental predictors. Deep learning via Long Short-Term Memory (LSTM) networks provided complementary insights, particularly in modeling temporal dependencies and autoregressive health trends.

Collectively, the methodologies applied enabled a comprehensive and interpretable framework that met the study’s original goals, while producing findings with both scientific and policy relevance.

## **7.2 Limitations**

This research, while rigorous, carries several limitations. First, the study used data from a single geographic area and a limited temporal window of one year, which restricts generalizability and may miss long-term or inter-annual variation. Additionally, the ecological nature of the analysis, relying on aggregated population-level data, precludes definitive individual-level inferences.

Second, not all potential confounding factors were incorporated. While seasonal and weekly trends were modeled, meteorological variables (e.g., temperature, humidity), circulating respiratory viruses, and allergens were not explicitly controlled for. These are known contributors to respiratory disease and may interact with pollution in complex ways (World Health Organization [WHO], 2022).

Third, while syndromic surveillance data enables timely analysis, it lacks diagnostic precision. Variability in healthcare-seeking behavior and coding practices can introduce noise into the outcome variable. Similarly, using ambient pollutant measurements as proxies for personal exposure introduces potential misclassification, particularly in spatially diverse urban areas.

Modeling limitations also exist. Deep learning models like LSTM require large datasets for optimal performance and are less interpretable. While methods like SHAP values improved transparency, multicollinearity among pollutants remains a challenge for isolating independent effects. Furthermore, although the study used cross-validation and held-out data, formal uncertainty quantification (e.g., confidence intervals) was not included, which limits inferential robustness.

Lastly, the findings represent associations rather than proven causal relationships. Though they align with established biological mechanisms (Monoson et al., 2023), and lag structures support temporal plausibility, further causal inference techniques would strengthen 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**: Expanding the study period would allow detection of multi-year trends and reduce sensitivity to year-specific events. This is especially valuable for training deep learning models, which benefit from larger datasets.

**Inclusion of Confounders and Interacting Variables**: Integrating meteorological data (e.g., temperature, humidity) and virological surveillance (e.g., flu or COVID-19 prevalence) would help isolate pollution-specific effects and account for multi-causal dynamics (UK Health Security Agency [UKHSA], 2023).

**Spatially Resolved and Individual-Level Exposure**: Incorporating high-resolution geographic data or wearable sensors would reduce exposure misclassification and permit finer-scale analyses, especially across demographic groups or neighborhoods.

**Real-Time Surveillance Integration**: Developing operational tools that merge predictive models with live environmental and health data could support early warning systems. Such platforms would enhance seasonal preparedness and allow for targeted public health responses.

**Causal Modeling Techniques**: Future work should apply methods like distributed lag nonlinear models or causal machine learning to estimate pollutant effects more precisely and formally account for uncertainty.

**Model Refinement and Optimization**: Future iterations of the LSTM and other deep learning models could explore hyperparameter tuning (e.g., learning rate schedules, optimizer choice such as AdamW or RMSprop, batch size adjustments), along with regularization techniques like dropout, early stopping, and gradient clipping. These measures could help stabilize training, reduce validation loss volatility, and improve generalization performance.

**Cross-City and Comparative Studies**: Testing this modeling framework in different cities or countries would assess its transferability and reveal context-specific risk factors.

**Health Impact Assessment and Policy Simulation**: Linking pollution reductions to specific health benefits—such as avoided hospital visits—would support cost-benefit analyses of environmental policies (COMEAP, 2022).

Together, these directions would help move from descriptive and predictive insights toward actionable, causal understanding of pollution’s health impact. Ultimately, translating such work into responsive public health interventions could reduce preventable respiratory illness and promote environmental equity.

**Chapter 8: Student Reflections**

**8.1 Learning Outcomes**

This project significantly enhanced my ability to integrate heterogeneous environmental and health datasets, manage their preprocessing, and prepare them for advanced analysis. I developed proficiency in applying machine learning and deep learning models—such as Random Forest, XGBoost, and LSTM—to address real-world public health problems. Beyond technical implementation, I learned to interpret and contextualise model outputs for diverse audiences, including those without technical backgrounds, ensuring results were both scientifically robust and communicable for policy and health decision-making.

**8.2 Challenges and Solutions**

**Data Gaps:** Missing values in pollutant and health datasets posed a risk to model integrity. This was addressed using rolling mean imputation, which preserved temporal continuity while mitigating bias from interpolation.

**Model Interpretability:** The complexity of ensemble and neural network models made it challenging to explain their decisions. This was mitigated through SHAP analysis, which provided pollutant-level impact estimates that could be visualised and clearly communicated.

**Computational Load:** LSTM training and extensive lag feature creation increased computational demand. To address this, code was optimised, redundant features were pruned, and more efficient batch processing was implemented.

**8.3 Skills Development**

Throughout the project, I advanced my expertise in time-series modelling, lag feature engineering, and cumulative exposure analysis. My proficiency in Python-based data science workflows improved, particularly in leveraging packages such as pandas, scikit-learn, xgboost, and tensorflow. I also developed stronger technical communication skills, learning to produce clear, insightful figures and concise academic writing that balanced statistical detail with accessibility for multidisciplinary audiences.

**8.4 Future Applications**

The skills and methodologies developed through this project are directly transferable to professional roles in public health analytics, environmental monitoring, and policy-driven data science. In particular, the integration of real-time environmental data with predictive models could contribute to the design of early-warning systems for respiratory illness surges, enabling more proactive health interventions.

**8.5 Summary**

This project provided a valuable academic and professional development experience, blending technical problem-solving with the real-world urgency of environmental health challenges. By bridging advanced analytical techniques with clear, policy-relevant interpretation, I have strengthened both my technical expertise and my capacity to contribute meaningfully to data-driven public health initiatives.

**References**

* Atmotube. (2024). Air Quality Data Visualization for Environmental Research. https://atmotube.com/blog/air-quality-data-visualization-for-environmental-research
* 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. [Journal Not Specified].
* Monoson, A., Schott, E., Ard, K., Kilburg-Basnyat, B., Tighe, R. M., Pannu, S., & Gowdy, K. M. (2020). 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
* Shi, S., Lin, H., Jiang, L., Zeng, Z., Lin, C., Li, P., Li, Y., & Yang, Z. (2020). Development of a respiratory virus risk model with environmental data based on interpretable machine learning methods. [Journal Not Specified].
* 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
* Walters, S., Phupinyokul, M., & Ayres, J. (1995). Hospital admission rates for asthma and respiratory disease in the West Midlands: their relationship to air pollution levels. [Journal Not Specified].
* 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>

**Appendices**