**What is left to be done :**

**Table of Contents**

**List of figures**

**List of tables**

**Appendix**

**Abstract**

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 study demonstrates the value of combining statistical decomposition, lag structures, and machine learning for understanding and forecasting respiratory health risks. Despite limitations such as the use of aggregated data and omission of meteorological or viral covariates, the results align with established environmental health literature. These findings support the development of pollution-informed early warning systems and targeted interventions during high-risk periods. Future research should extend the temporal scope, incorporate additional covariates, and transition toward real-time, spatially resolved decision-support tools for public health planning.

**Chapter 1: Introduction**

## **1.1 Project Overview**

This dissertation investigates the temporal relationships between ambient air pollutants and acute respiratory health outcomes over a full year (June 2024 to June 2025) in an urban context. It integrates high-resolution environmental data—including conventional pollutants (PM₁₀, PM₂.₅, NO₂, NOₓ, O₃, black carbon) and spectrally-resolved particulate matter—with daily syndromic health surveillance data covering acute respiratory infections, bronchiolitis, influenza-like illness, and scarlet fever. The study employs a multi-method approach combining time series decomposition, lag analysis, and predictive modeling (using Random Forest, XGBoost, and LSTM), while interpretability tools such as SHAP values are used to identify the most influential pollutants. The project also includes the development of interactive visualizations aimed at supporting data-informed public health policy and early-warning decision systems.

**1.2 Background and Context**

Air pollution remains one of the most urgent environmental health challenges worldwide, contributing significantly to global morbidity and mortality. According to the World Health Organization (2022), exposure to pollutants such as particulate matter (PM₁₀ and PM₂.₅), nitrogen oxides (NO, NO₂, and NOₓ), black carbon, and ozone (O₃) is linked to a range of adverse respiratory outcomes, including asthma, bronchitis, chronic obstructive pulmonary disease (COPD), and acute respiratory infections. These pollutants stem from diverse sources—vehicular emissions, industrial activities, domestic heating, and atmospheric photochemical reactions—and their distribution and health effects fluctuate with meteorological conditions, topography, and seasonal cycles.

Urban environments are particularly vulnerable due to high population density and concentrated emission sources. 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 critical for public health preparedness.

Advancements in real-time environmental monitoring and digital health surveillance have created new possibilities for data-driven forecasting and risk modeling. However, much of the current literature still relies on coarse temporal granularity, limited pollutant scope, or basic statistical models that do not fully capture the complex, non-linear, and lagged nature of pollution–health interactions. There remains a need for integrative approaches that combine robust statistical foundations with interpretable machine learning to support both scientific understanding and policy implementation.

**1.3 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.

**1.4 Research Aim and Objectives**

**1.4.1 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.

**1.4.2 Objectives**

* To characterize seasonal patterns and temporal trends in daily respiratory health indicators through time series decomposition techniques.
* To investigate the statistical associations and lagged effects between key air pollutants (e.g., PM₂.₅, NO₂, O₃, black carbon) and acute respiratory illnesses.
* To develop and evaluate predictive models—including Random Forest, XGBoost, and LSTM—to forecast respiratory health outcomes based on environmental exposures.
* To identify and rank the most influential pollutants driving health outcomes using interpretable machine learning techniques such as SHAP and sensitivity analysis.

**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 wealth of epidemiological research underscores the detrimental effects of air pollution on respiratory health. Monoson et al. (2020) highlight the role of pollutants like PM2.5, NO2, and SO2 in increasing susceptibility to and severity of respiratory infections, including bacterial, fungal, and viral pathogens. Notably, exposure to PM2.5 has been linked to heightened risks of severe outcomes during the COVID-19 pandemic, suggesting that air pollution may amplify the impact of infectious diseases (Tian et al., 2021). The World Health Organization estimates that air pollution contributes to approximately 7 million premature deaths annually, with a significant portion attributed to respiratory illnesses (WHO, 2024). Tran et al. (2023) further emphasize that air pollution, exacerbated by climate change, increases respiratory disease risk through mechanisms such as inflammation, oxidative stress, and impaired immune function, particularly affecting vulnerable populations like children and the elderly. Similarly, Lee et al. (2024) conducted a systematic review and meta-analysis, finding that exposure to ambient PM2.5 and NO2 is associated with increased adult asthma incidence, though heterogeneity among studies suggests the need for further research to confirm these associations across diverse populations.

Particulate matter, particularly PM2.5 and black carbon, is a major focus due to its ability to penetrate deep into the respiratory system. Black carbon, a component of fine particulate matter produced by incomplete combustion of fossil fuels and biomass, is associated with increased risks of respiratory infections, asthma exacerbation, and reduced lung function (Monoson et al., 2020). Its small size allows it to reach the alveoli, triggering inflammation and oxidative stress, which can worsen existing respiratory conditions. Studies suggest that black carbon may also act as a carrier for pathogens, facilitating their entry into the lungs and increasing infection risks. Additionally, chronic exposure to black carbon has been linked to cardiovascular issues, highlighting its broader health impacts. The State of Global Air 2024 report indicates that air pollution accounted for 8.1 million deaths globally in 2021, with PM2.5 contributing to 4.7 million deaths, many related to respiratory diseases (Health Effects Institute, 2024).

In the West Midlands, UK, research conducted in the late 1980s examined hospital admission rates for asthma and other respiratory diseases in relation to ambient levels of smoke, SO2, and NO2 (Walters et al., 1995). Using data from the West Midlands Regional Health Authority over a two-year period, the study employed multivariate regression models to assess correlations between standardized hospitalization ratios and pollutant levels across electoral wards. The results indicated a significant association between NO2 levels and increased hospital admissions, particularly for asthma, with variations observed across age groups. Socioeconomic factors, such as deprivation and ethnicity, were found to influence admission rates but did not fully account for the pollution-related effects. This study highlights the importance of considering both environmental and social determinants in assessing health impacts.

Other pollutants, such as ozone and nitrogen oxides, have also been implicated in respiratory health outcomes. Ozone, a reactive gas formed through photochemical reactions, can irritate the airways, leading to symptoms such as coughing and shortness of breath. Nitrogen oxides, primarily emitted from vehicle exhausts, contribute to airway inflammation and are associated with increased hospital visits for respiratory conditions (Health Effects Institute, 2024). These findings underscore the need for comprehensive studies that examine the combined effects of multiple pollutants, as their interactions may amplify health risks.

**2.3 Geographical Variations in Air Pollution Impacts**

Regional studies provide critical insights into the spatial distribution of air pollution and its health consequences. A report focusing on Birmingham, UK, analyzed the health impacts of PM2.5 using concentration-response functions recommended by the Committee on the Medical Effects of Air Pollutants (Hall et al., 2020). The study estimated that air pollution contributes to approximately 720 early deaths and 7,500 lost life years annually in Birmingham, with around 900 new asthma cases each year among children and adults. The report identified significant geographical disparities, with central wards such as Tyseley & Hay Mills, Holyhead, and Aston experiencing the highest proportions of pollution-related mortality, up to 8.5%. These areas, characterized by high population density and proximity to industrial and traffic-related pollution sources, bear a disproportionate health burden.

In the U.S., a study published in *Frontiers in Public Health* (2025) analyzed the disease burden linked to air pollution across states from 1990 to 2021, finding an 80.5% decline in PM2.5-attributable mortality but variations in health outcomes, with some states like California showing increased diabetes-related disability-adjusted life years (DALYs) (Frontiers Authors, 2025). In Southeast Asia, Tan et al. (2023) found that exposure to particulate matter in urban areas of Malaysia is associated with respiratory symptoms such as coughing and wheezing among children, with diesel vehicle emissions and industrial activities as primary sources. In Africa, Nkosi et al. (2024) highlighted the persistent burden of household air pollution, noting that reliance on biomass fuels like charcoal and kerosene contributes to significant respiratory disease rates, with 1.1 million air pollution-related deaths in Africa in 2019. A multicity study by Zhang et al. (2024) during 2017–2022 showed varying hospitalization and mortality rates due to PM2.5 and O3 exposure, with older populations and socioeconomically disadvantaged areas showing higher vulnerability. The State of Global Air 2024 report further notes that countries like India and China bear a high burden, with 2.1 million and 2.3 million air pollution-related deaths, respectively, in 2021, many linked to respiratory conditions (Health Effects Institute, 2024).

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

The advent of machine learning has revolutionized environmental health research by enabling the analysis 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 the presence of respiratory viruses across 31 regions in China from 2016 to 2021, integrating air quality indices and meteorological factors, achieving an average accuracy of 0.76 and an AUC score of 0.9. Wang et al. (2023) applied eight machine learning algorithms to predict outpatient visits for acute respiratory infections from 2018 to 2021, with the random forest model demonstrating the highest accuracy and identifying NO2 as a significant predictor with a one-day lag effect. Similarly, Kim et al. (2022) developed models using gradient boosting and Gaussian process regression to predict respiratory disease occurrence, achieving R² values of 0.67–0.68, with key factors including temperature, humidity, and pollutants like PM2.5 and SO2.

A longitudinal study by Su et al. (2024) used digital health sensors to monitor medication use in 3,386 asthma and COPD patients in California from 2012 to 2019, finding positive associations between NO2, PM2.5, and O3 exposure and increased rescue medication use, with random forest models confirming these findings. The study estimated a 23.9% increase in daily rescue puffs, translating to significant economic costs. These results highlight the potential of machine learning to enhance predictive accuracy and inform public health strategies. The adaptation of early warning systems (EWS) from infectious disease surveillance to environmental health, as evaluated by Meckawy et al. (2020), offers a promising framework. Integrating real-time air quality data with health surveillance could enable forecasting of pollution-related health events, though challenges like data standardization and accessibility remain.

**2.5 Data Visualization in Environmental Health Research**

Effective communication of complex environmental health data is essential for informing policy and engaging stakeholders. Research indicates that well-designed visualizations, such as interactive dashboards and maps, enhance comprehension and facilitate decision-making (Ramirez et al., 2019). The Atmotube blog (2024) discusses techniques like heat maps, time series plots, and color-coded indices to communicate air quality data, emphasizing the use of intuitive interfaces and colorblind-friendly palettes to reach diverse audiences. For example, the Air Quality Index (AQI) uses a color-coded scale to indicate pollution levels, but Ramirez et al. (2019) found limited evidence that such indices induce recommended behavior changes during poor air quality events, suggesting a need for improved communication strategies.

Brown and Green (2022) conducted a bibliometric review using cluster analysis to visualize trends in air pollution and health research, demonstrating how visualization aids in identifying research hotspots. The current project’s use of Power BI aligns with these principles, aiming to provide an intuitive platform for stakeholders to explore temporal trends and spatial variations in air pollution and health outcomes. However, further evaluation is needed to assess how different audiences interact with and interpret these visual tools to maximize their impact on policy and public awareness.

**2.6 Gaps and Opportunities for Further Research**

Despite significant advancements, several gaps remain in the literature. Many studies focus on specific pollutants or broad health outcomes, with limited emphasis on the temporal dynamics of exposure. The current project’s use of daily measurements over a year offers an opportunity to uncover seasonal patterns and short-term fluctuations in pollutant levels and their health impacts. The inclusion of a wide range of pollutants, including spectral particulate matter measurements, enables a more comprehensive analysis of their combined effects, addressing a gap in multi-pollutant studies.

The application of machine learning models to predict specific respiratory conditions based on multi-pollutant exposures requires further exploration. While studies like Wang et al. (2023) and Kim et al. (2022) demonstrate potential, validation across diverse populations and settings is necessary to ensure generalizability. The development of interpretable models, such as those using SHAP analysis, is critical for translating predictions into policy and practice. The adaptation of early warning systems from infectious disease surveillance to environmental health presents a promising avenue, with potential to enhance risk prediction during high-pollution episodes (Meckawy et al., 2020). Research is needed to address challenges such as data integration and accessibility in low-resource settings.

Finally, the effectiveness of data visualization tools in communicating environmental health risks warrants further evaluation. Ramirez et al. (2019) highlight limitations in current visualization strategies, suggesting a need for research on how different audiences interpret visual representations to inform more effective 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 methodology emphasizes reproducibility, analytical rigor, and the use of advanced tools for temporal and statistical insight.

**3.2 Dataset Overview and Collection**

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

**Health Data:**

**Daily records were obtained from the UK Health Security Agency (UKHSA), detailing emergency department (ED) attendances associated with four respiratory syndromes. These indicators serve as proxies for real-time community-level respiratory health trends:**

**• Acute Bronchiolitis Syndromic**

**• Acute Respiratory Illness**

**• Influenza-like Syndromic**

**• Scarlet Fever Syndromic**

**Air Quality Data:**

Environmental exposure data was sourced from the UK Department for Environment, Food & Rural Affairs (DEFRA) Air Quality Archive. It includes daily measurements from air quality monitoring stations across the West Midlands. The pollutants captured fall into two categories:

Conventional pollutants such as:

* Particulate Matter (PM2.5 and PM10)
* Nitrogen Dioxide (NO₂)
* Nitric Oxide (NO)
* Combined Nitrogen Oxides expressed as NO₂ equivalents
* Ozone (O₃)
* Black Carbon

Spectral particulate matter, capturing optical responses across multiple wavelengths, including:

* Blue, Green, Red, and Yellow (590 nm) particulate matter
* Ultraviolet (UV, 370 nm) and Infrared (IR) particulate matter

The resulting dataset spans a continuous 12-month period, comprising 365 daily entries and 18 variables. It enables detailed temporal analysis of air pollution and its potential correlation with short-term spikes in respiratory-related healthcare utilization. This structured, merged dataset lays the foundation for subsequent exploratory, statistical, and predictive analyses.

**3.3 Data Preparation and Preprocessing**

Preprocessing was performed in Python, using structured Jupyter notebooks for both the health and air quality datasets. The objective was to ensure data quality, temporal alignment, and analytical readiness.

**3.3.1 Health Data Preprocessing**

* **Filtering**: Selected only the four respiratory syndromes relevant to the study
* **Date Formatting**: Parsed and standardized using Python's datetime module
* **Missing Value Imputation**: Applied a 7-day rolling mean to fill missing entries, preserving temporal structure
* **Renaming and Validation**: Renamed columns for clarity and validated that all values were numeric and dates were sequential

**3.3.2 Air Quality Data Preprocessing**

* **Metadata Cleaning**: Removed headers, footnotes, and non-numeric rows from DEFRA CSVs
* **Placeholder Handling**: Replaced strings such as "No data" with NaN, then converted to numeric
* **Spatial Aggregation**: Averaged pollutant values across monitoring stations for each day
* **Unit Consistency**: Verified that all readings used micrograms per cubic meter
* **Imputation and Formatting**: Imputed missing values using a 7-day rolling average and rounded values to three decimal places

**3.3.3 Dataset Integration**

* **Merge Operation**: Performed an inner join on the date column
* **Verification**: Ensured complete daily coverage and absence of null values
* **Structure**: Organized data into a wide format with logical grouping of health and pollutant metrics

The datasets were merged using the date column as a key, covering the period from **June 9, 2024, to June 8, 2025**. The final dataset contains:

* **365 daily records**
* **18 variables**, including 1 date field, 4 health indicators, and 13 pollutant measurements
* All pollutant concentrations expressed in **µg/m³**, aligned with 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 characterize each variable’s distribution. For all pollutant and health indicators, the following were calculated:

* Central tendency (mean, median)
* Dispersion (standard deviation, min, max)
* Distribution shape indicators (skewness)

These metrics were used to assess the presence of outliers, variability across time, and to inform data normalization and transformation decisions.

**3.4.2 Time Series Visualization and Rolling Averages**

Time series plots were generated for all variables across the full date range to visually inspect temporal continuity and detect periodic behaviors or disruptions.

* **Raw daily values** were visualized to understand natural variability
* **7-day rolling averages** were applied to each time series to reduce noise and emphasize mid-term patterns
* Visual comparisons between pollutant and health variables were included to support hypothesis generation

This step provided the basis for understanding time-dependent behavior prior to decomposition and lag analysis.

**3.4.3 Seasonal Decomposition**

To separate and analyze different components of each time series, **Seasonal-Trend-Loess (STL) decomposition** was applied to both pollutant and health indicator variables.

* Each series was decomposed into **trend**, **seasonal**, and **residual** components
* This was done using Python’s statsmodels package under additive mode
* The goal was to isolate recurring temporal patterns and distinguish them from underlying trends or noise

The resulting decompositions were later used to support feature engineering and interpretability in modeling.

**3.4.4 Correlation Exploration (Preliminary)**

A preliminary correlation analysis was conducted to identify potential linear relationships among variables.

* **Pearson correlation coefficients** were computed between all pollutant and health indicators
* A full **correlation matrix** and **heatmaps** were created to visualize these relationships
* Intra-group correlations (e.g., among pollutants) were also explored to detect redundancy or shared environmental behavior

These exploratory steps informed the design of the formal correlation analysis described in section 4.5 and guided early feature selection.

**3.4.5 Lag and Advanced Lag Analysis**

To investigate potential delayed effects of air pollutant exposure on health outcomes, lag analysis was employed.

* **Simple lag shifts** of 1 to 14 days were applied to pollutant variables
* **Cross-correlation functions (CCFs)** were computed to assess the temporal dependencies between pollutants and health indicators
* **Compound lag windows** (e.g., 3-day or 7-day moving averages) were considered to model cumulative exposure
* **Polynomial lag models** (e.g., Almon lags) were optionally explored to structure and interpret distributed lag effects.

The outputs of this analysis contributed to the selection of lagged features for time-aware modeling techniques.

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

A custom proxy for the Air Quality Index (AQI) was constructed by aggregating normalized concentrations of selected pollutants.

* This proxy metric was visualized using a **seasonal area chart** to explore its variation across the study period
* Overlay plots were also developed to compare this AQI proxy with health indicator timelines (without making interpretative claims)
* These visual tools were intended to support further hypothesis development and aid communication in dashboards.

**3.5 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 Define and Describe the Models**

**Random Forest Regressor (Tuned Baseline Model)**

Random Forest is a **bagging-based ensemble learning algorithm** that operates by constructing a multitude of decision trees during training and aggregating their results (via averaging in regression tasks). In this study, a tuned version of the RandomForestRegressor was implemented, with key hyperparameters such as n\_estimators=100 and n\_jobs=-1 to ensure sufficient model complexity and computational efficiency.

This model treats each pollutant-lag pair as an independent feature, allowing it to model non-linear relationships and interactions across features. Despite its lack of temporal sequence modeling, Random Forest performs reliably on tabular data with manually engineered lagged variables.

**XGBoost (Untuned and Tuned Versions)**

XGBoost (Extreme Gradient Boosting) is a **gradient boosting framework** that builds additive models iteratively, each correcting the errors of its predecessor. It is particularly known for its **high efficiency, scalability, and predictive power**. The model was used in two phases:

* **Untuned version**: Deployed with default settings to quickly establish a baseline.
* **Tuned version**: Configured with optimized hyperparameters (max\_depth=6, eta=0.1, subsample=0.8, etc.) to enhance performance.

The model operates on a **tabular matrix of features**, which includes current-day pollutant concentrations and multiple lagged versions. It internally evaluates **feature importance**, enabling insights into which pollutants (and their corresponding lag periods) are most predictive of ARI.

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

LSTM is a specialized type of **recurrent neural network (RNN)** designed to model long-term temporal dependencies in sequential data. In the context of this study, LSTM was employed to handle **sequences of pollutant levels over a 35-day window**, aiming to learn temporal dynamics directly rather than relying solely on engineered lag features.

The model architecture included:

* **Sequential input** of 35-day pollutant data vectors,
* A **multi-output prediction** design for forecasting ARI occurrences over different horizons,
* An **attention mechanism**, highlighting specific time steps (e.g., lag days 17–30) as more influential,
* A **multi-objective loss function**, assigning weights to different outputs based on their predictive relevance.

This makes LSTM particularly suited for detecting **latent patterns across time** that conventional models may miss.

**3.5.2 Rationale Behind Model Selection and Comparative Evaluation**

**Random Forest: Strengths and Suitability**

* **Robustness to Noise and Overfitting**: The ensemble nature guards against high variance, which is critical in environmental datasets prone to fluctuations.
* **Handles Mixed Data Types**: Works well with both continuous variables and categorical encodings (e.g., seasons, weekday).
* **Feature Importance**: Offers straightforward interpretation of which pollutant-lag combinations are driving predictions.

However:

* **Not sequence-aware**: Temporal relationships are not modeled inherently; instead, they rely on manually created lag features.
* **Curse of Dimensionality**: Including many lag features across pollutants can inflate the feature space, potentially diluting model focus.

**XGBoost: Strengths and Suitability**

* **Superior Predictive Accuracy**: Frequently outperforms other machine learning models in structured data settings.
* **Efficient Regularization**: Penalizes complexity through L1/L2 terms, improving generalization.
* **Insight into Lag Effects**: Through feature importance and SHAP values, XGBoost revealed meaningful insights (e.g., NO₂ lag 19 showed strong relevance to ARI).

However:

* **Tuning Complexity**: Requires careful hyperparameter optimization to avoid under/overfitting.
* **Non-sequential modeling**: Like Random Forest, it doesn't natively handle sequences, necessitating lag engineering.

**LSTM: Strengths and Suitability**

* **Captures Sequential Dependencies**: Naturally models the progression of pollutant levels over time, critical for capturing the cumulative burden on respiratory health.
* **No Manual Lag Engineering Needed**: The model learns lag relationships directly from the sequence structure.
* **Attention Mechanism**: Provides interpretability by showing which days in the 35-day sequence contribute most to ARI prediction.

However:

* **High Computational Demand**: Requires more memory, training time, and tuning.
* **Data-Hungry**: Performs best with large, high-quality datasets, which may not always be available in environmental health contexts.
* **Less Transparent**: Even with attention layers, interpretability is inherently more limited compared to 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.

**3.6.1 Root Mean Squared Error (RMSE)**

**Definition**:

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

**Purpose & Justification:**

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.

**3.6.2 Mean Absolute Error (MAE)**

**Definition:**

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

**Purpose & Justification:**

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.

**3.6.3 Median Absolute Error (MedAE)**

**Definition:**

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

**Purpose & Justification:**

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.

**3.6.4 Mean Absolute Percentage Error (MAPE)**

**Definition:**

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

**Purpose & Justification:**

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.

**3.6.5 Coefficient of Determination (R² Score)**

**Definition:**

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

**Purpose & Justification:**

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.

**3.6.6 Feature Importance Analysis**

In addition to quantitative performance metrics, feature importance analysis was conducted to improve the interpretability of the models and understand which input variables most significantly influenced the predictions.

**Definition:**

Feature importance scores indicate the relative contribution of each feature in building the model. For tree-based models like Random Forest and Gradient Boosting, importance is typically derived from the reduction in impurity (e.g., Gini importance).

**Purpose & Justification:**

While error metrics quantify how well a model performs, feature importance helps uncover why the model makes certain predictions. This is crucial in environmental modeling, where understanding the drivers of air pollution can inform 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**

* **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) Strategy**

**4.1.1 Descriptive Statistics: Results**

Descriptive statistical analysis revealed important characteristics of both pollutant concentrations and health indicator variables over the observed 365-day period. The central tendency, dispersion, and distributional asymmetries were assessed to understand variability, potential outliers, and the necessity for normalization prior to modeling.

**Health Indicators.** Among the health-related variables, *acute respiratory illness* exhibited the highest daily mean count (μ = 214.44, σ = 73.72), with values ranging from 105 to 470, indicating substantial inter-day variability. In contrast, *acute bronchiolitis syndromic cases* had a lower mean (μ = 14.45, σ = 12.26), but a highly skewed distribution, as suggested by a maximum count of 65 juxtaposed against a median of only 11. *Influenza-like illness* showed a pronounced skew as well, with a minimum of 0 and a maximum of 83 cases, while its median (3) was far below the mean (10.86), suggesting episodic outbreaks. The *scarlet fever syndromic indicator* maintained a near-zero median and mean (μ = 0.50, σ = 0.80), indicative of rare occurrences.

**Particulate Matter and Aerosol Constituents.** Across the optical particulate measurements, *UV-sensitive particulate matter at 370 nm* (mean = 1.14, max = 6.78) and *yellow-channel particulate matter at 590 nm* (mean = 1.01, max = 5.93) displayed the highest values among spectrally resolved PM signals. These variables exhibited right-skewed distributions, highlighting intermittent high-aerosol events. *Red-channel PM* and *infrared PM* showed similar distributional characteristics, with relatively low medians compared to their maximums (e.g., IR PM max = 5.6, median = 0.777), again suggesting episodic spikes in fine particulate matter concentrations.

**Gaseous Pollutants.** Among traditional gaseous pollutants, *ozone (O₃)* showed a consistent profile with a relatively high mean (50.59 ppb) and a controlled variance (σ = 16.99), ranging from 4.27 to 96.36 ppb. *Nitrogen dioxide (NO₂)* and *nitric oxide (NO)* demonstrated typical urban profiles. NO₂ exhibited a mean of 17.36 ppb, while NO had a lower mean (8.16 ppb) but a disproportionately high maximum value (88.25 ppb), indicative of transient traffic or combustion-related spikes. The composite metric *NOₓ as NO₂* had a wide range (7.94 to 192.31 ppb), reflecting cumulative nitrogen-based pollution episodes.

**Standard Particulate Indicators.** *PM₁₀* and *PM₂.₅* demonstrated relatively moderate concentrations with mean values of 13.83 µg/m³ and 8.38 µg/m³, respectively. Their respective maximums (55.33 and 45.15 µg/m³) suggest occasional pollution events, likely of anthropogenic origin or weather-induced entrainment of particulates. Interquartile ranges indicated moderate day-to-day variability, with a right-tailed distribution, particularly pronounced for PM₁₀.

Overall, the descriptive statistics underscored the heterogeneity of the dataset, both in terms of pollutant concentrations and health outcomes. Several variables demonstrated heavy-tailed distributions, motivating the use of normalization or transformation techniques in subsequent analyses. Moreover, the wide ranges and standard deviations observed reinforced the necessity for robust modeling techniques capable of accounting for non-linearity and outlier presence.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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: Results**

Temporal inspection of both health indicators and pollutant concentrations was conducted using line plots of daily values and their 7-day rolling averages. This visualization strategy enabled the identification of seasonal fluctuations, episodic peaks, and potential co-evolution patterns between health outcomes and environmental exposures.

**Temporal Dynamics of Health Indicators**

Daily time series of the four health indicators demonstrated distinct temporal signatures across the one-year observation window. As shown in **Figure 1**, *acute respiratory illness* consistently maintained the highest incidence rates, peaking sharply between November 2024 and January 2025 with daily case counts approaching 450. This peak aligns temporally with increases observed in both *acute bronchiolitis* and *influenza-like illness*, which also rose markedly during this period. The onset of these increases was notably synchronous across indicators, suggesting a shared seasonal or environmental driver. *Scarlet fever syndromic cases*, on the other hand, remained sparse and largely invariant throughout the timeline.

**A graph showing different colored lines

AI-generated content may be incorrect.**

***Figure 1: Raw Daily Time Series of Health Indicators***

Averaging monthly case counts further elucidated these temporal patterns (see **Figure 2**). *Acute respiratory illness* showed a distinct wintertime escalation, with a maximum monthly average exceeding 370 cases in December 2024. Similarly, *influenza-like illness* exhibited a sharp December peak, whereas *bronchiolitis* peaked in November before declining. These trends collectively suggest a strong seasonal influence, possibly linked to ambient environmental or viral transmission conditions.

**A graph showing the number of health indexes

AI-generated content may be incorrect.**

***Figure 2: Monthly Averages of Health Indicators***

**Temporal Dynamics of Pollutants**

Monthly-averaged pollutant concentrations (shown in **Figure 3**) highlighted variable behaviors across gaseous and particulate species. *Ozone (O₃)* displayed a classical seasonal inversion, reaching its nadir in December–January and peaking in late spring (April–May 2025), consistent with photochemical production dynamics. Conversely, *nitrogen oxides* (NO, NO₂, and total NOₓ) exhibited maxima during colder months, reflecting heightened vehicular and heating emissions, as well as reduced atmospheric dispersion. Notably, *PM₂.₅* and *PM₁₀* concentrations also rose during the late autumn and winter months, likely driven by combustion sources and atmospheric stagnation.

**A graph of different colored lines

AI-generated content may be incorrect.**

***Figure 3: Monthly Averages of Key Pollutants***

**Rolling Average Smoothing and Pattern Amplification**

Application of a 7-day rolling average smoothed out high-frequency noise, making mid-term patterns more discernible. In the smoothed health indicator series (**Figure 4**), the composite seasonal wave across all variables became pronounced, with *influenza-like* and *bronchiolitis* episodes co-peaking during late 2024 and then gradually tapering off.

**A graph showing the number of health indicators

AI-generated content may be incorrect.**

***Figure 4: 7-Day Rolling Averages of Health Indicators***

Similarly, the rolling average curves for pollutants (**Figure 5**) revealed episodic surges, particularly for *NOₓ as NO₂*, which spiked significantly in late December 2024—potentially corresponding to the seasonal confluence of increased emissions and meteorological stagnation. These pollutant spikes appear to temporally align with peaks in respiratory morbidity, warranting deeper lag analysis in subsequent sections. Notably, *ozone* continued to oscillate with a clear semi-regular periodicity, further reinforcing its strong seasonal modulation.

**A graph showing the number of pollutants

AI-generated content may be incorrect.**

***Figure 5: 7-Day Rolling Averages of Pollutants***

**Interpretation and Implications**

These visual diagnostics collectively underscore the temporally dynamic nature of both air quality and public health indicators. The strong seasonal signatures, along with the apparent temporal proximity of pollutant and morbidity peaks, support the hypothesis of delayed or cumulative pollutant impacts on respiratory health. This forms the empirical basis for the decomposition and lag modeling strategies detailed in the following sections.

**4.1.3 Seasonal Decomposition: Results**

To disentangle temporal complexity in the dataset, each time series was decomposed into its constituent components—trend, seasonal, and residual—using Seasonal-Trend-Loess (STL) under the additive decomposition assumption. This approach facilitated the identification of persistent trends, periodic signals, and short-term anomalies that were not evident in raw or smoothed series alone.

**Health Indicators: Component Analysis**

The STL decompositions of health indicator time series are presented in **Figure 6**. Each subplot contains three panels: the raw signal (observed), long-term trajectory (trend), and detrended irregular fluctuations (residual).

* **Acute Respiratory Illness** exhibited a well-defined seasonal peak centered around the 180–200 day mark, consistent with winter months. Its trend component increased steadily until early January before declining, reflecting the central wave of respiratory morbidity.
* **Acute Bronchiolitis Syndromic** followed a similar pattern, peaking marginally earlier, around day 170, and then subsiding.
* **Influenza-like Illness** displayed the sharpest and most pronounced seasonal effect among the indicators, with a tightly peaked trend spanning fewer days, reinforcing its epidemic-like dynamics.
* **Scarlet Fever Syndromic** lacked a discernible seasonal trend and remained dominated by stochastic residuals and sparse event spikes, confirming its low prevalence and unpredictability.

**A screenshot of a graph

AI-generated content may be incorrect.**

***Figure 6: STL Decomposition of Health Indicators (Observed, Trend, Residual)***

The residual components of all three major indicators showed mild volatility, with spikes that often co-occurred, potentially signaling shared environmental or transmission drivers. The low residual amplitude in *influenza-like illness* further confirmed that its variability is strongly governed by seasonal and trend components.

**Pollutants: Component Analysis**

A parallel decomposition of key air pollutants is shown in **Figure 7**, highlighting distinct seasonal behaviors across pollutant types:

* **Nitrogen-Based Pollutants (NO, NO₂, NOₓ)** exhibited strong seasonal modulation, with upward-trending values from early autumn to winter, peaking near day 200. The trend components then declined into spring. These patterns mirror increased heating and traffic emissions during colder months.
* **Black Carbon**, a proxy for combustion-derived particulate pollution, followed a nearly identical trajectory, further supporting its emission source similarity with NOₓ gases.
* **PM₂.₅ and PM₁₀** trends also peaked in winter, although with a slightly flatter seasonal profile, suggesting additional contributions from resuspension or non-combustion sources.
* **Ozone**, in contrast, followed an inverse seasonal pattern, with a trough in the winter and a gradual increase toward spring and early summer (day 250 onward), as expected from its photochemical formation process.

**A screenshot of a graph

AI-generated content may be incorrect.**

***Figure 7: STL Decomposition of Pollutants (Observed, Trend, Residual)***

Residual plots revealed episodic spikes in **NOₓ**, **PM**, and **Black Carbon**, likely corresponding to acute pollution episodes (e.g., festivals, weather inversions). These transient deviations underscore the importance of retaining residual components for event-level modeling or anomaly detection.

**Summary and Implications**

The STL decomposition confirmed that both pollutant and health indicator time series exhibit meaningful and interpretable seasonality. Health metrics—especially *acute respiratory* and *influenza-like illnesses*—appear to co-evolve temporally with wintertime peaks in nitrogen-based and particulate pollutants. This provides empirical justification for incorporating seasonally lagged pollutant variables into time-aware predictive models. Furthermore, decomposed trend components offer a pathway for feature engineering that preserves long-range dynamics while filtering short-term noise.

**4.1.4 Correlation Exploration (Preliminary): Results**

To assess linear associations between environmental and health variables, Pearson correlation coefficients were computed for all pairwise combinations of the dataset’s variables. These coefficients were visualized as a full correlation matrix, as presented in **Figure 8**, allowing for both inter-group (pollutant–health) and intra-group (within pollutants or health indicators) comparisons.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

***Figure 8: Pearson Correlation Matrix of Health Indicators and Air Pollutants***

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

Among the health indicators, **acute bronchiolitis syndromic** and **acute respiratory illness** showed the strongest correlations with pollutant concentrations. Specifically:

* *Acute bronchiolitis* was positively associated with multiple optical and physical particulate matter measures, including **black carbon**, **blue PM**, **green PM**, and **red PM**, with correlation coefficients ranging from 0.22 to 0.25.
* *Acute respiratory illness* exhibited moderate positive correlations with **PM10** and **PM2.5**, as well as select color-channel PM measures.
* *Influenza-like illness* showed weaker correlations overall, though still displayed mild positive associations with **black carbon** and **UV-sensitive PM**.
* In contrast, *scarlet fever syndromic* appeared largely uncorrelated with any pollutant metric, suggesting it may be influenced by non-environmental factors or subject to low reporting rates.

Notably, **ozone** displayed a *negative correlation* with most health indicators, particularly with *acute bronchiolitis* (r ≈ –0.35), potentially reflecting seasonal inverse alignment (e.g., high ozone in summer, low respiratory morbidity) rather than a causal link.

**Intra-Group Correlations: Within Pollutants**

The matrix also revealed significant redundancy among pollutant variables, especially across spectrally resolved particulate measures:

* **Black carbon**, **UV PM**, and **red/yellow/blue/green PM** demonstrated consistently high inter-correlations (r > 0.90 in many cases), indicating that these variables are likely capturing overlapping dimensions of aerosol presence.
* **NO**, **NO₂**, and **NOₓ** formed a highly collinear block, with NOₓ essentially representing a composite of the two.
* **PM10** and **PM2.5** shared moderate to strong positive correlation (r ≈ 0.7), as expected due to their compositional overlap.

These findings suggest the presence of multicollinearity among pollutant features, which has important implications for downstream modeling. Dimensionality reduction, regularization, or informed feature pruning may be necessary to mitigate redundancy.

**Summary and Implications**

The preliminary correlation exploration provided initial evidence of meaningful associations between pollutant concentrations and respiratory health indicators. While the observed linear relationships were generally modest in magnitude, their consistency—particularly with particulate matter—supports the hypothesis of air quality impacts on health outcomes. Furthermore, the identification of strongly inter-correlated pollutant features provides a foundational basis for parsimonious feature selection in later modeling phases (see Section 5.2).

**4.1.5 Lag and Advanced Lag Analysis: Results**

To elucidate the temporal dependencies between pollutant exposure and health outcomes, both simple and advanced lag techniques were employed. This section presents the empirical evidence for delayed pollutant effects, particularly focusing on **NO₂** as a leading predictor of **acute respiratory illness**. Analyses span from brute-force cross-correlation calculations to structured regression using Almon polynomial lag models and cumulative exposure metrics.

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

Initial cross-correlation analysis assessed Pearson correlations between lagged pollutant values (1 to 30 days) and each health indicator. The strongest lagged correlation was observed between **nitric dioxide (NO₂)** and **acute respiratory illness**, peaking at a **24-day lag** (r = 0.37). Similar moderate correlations were found across multiple lags and 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 comparison of daily NO₂ levels (with 7-day and 30-day rolling averages) and acute respiratory illness cases confirms their co-evolution, with illness peaks consistently trailing NO₂ surges by approximately 3–4 weeks. This visually substantiates the temporal lag inferred statistically.

A graph showing a number of different colored lines

AI-generated content may be incorrect.

***Figure 9: 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, derived features were constructed using rolling windows of 7, 14, 21, and 30 days. These included moving averages, cumulative sums, peaks, and exponentially weighted moving averages (EWMA). Their respective Pearson correlations with acute respiratory illness are shown in **Figure 10**.

* The **30-day cumulative average and peak NO₂** features achieved the highest correlation (r = 0.59), followed closely by 21-day features (r ≈ 0.55).
* Even short-term metrics like 7-day peaks and EWMAs maintained moderate correlations (r ≈ 0.43–0.49).

This suggests a dose-response accumulation effect where longer periods of sustained exposure yield stronger associations with respiratory morbidity.

**A graph with green and white bars

AI-generated content may be incorrect.**

***Figure 10: Bar Chart of NO₂-Based Feature Correlations with Acute Respiratory Illness***

**Polynomial Lag Structure: Almon Lag Model**

To model the distributed lag effect more formally, a **degree-2 Almon polynomial lag model** was fit to the NO₂ series. The estimated lag weights (shown in **Figure 11**) follow a concave profile, peaking between **day 24 and day 28**, confirming the findings of the empirical lag scan.

**A graph of a normalized number of patients

AI-generated content may be incorrect.**

***Figure 11: Almon Lag Weight Profile for NO₂ Effect on Acute Respiratory Illness***

**Parametric Lag Modeling: Almon Regression**

To validate and structure the distributed temporal effects of NO₂ exposure on acute respiratory illness, a polynomial lag model of degree 2 was employed. Using an Almon lag specification, the model approximated 30 individual lag effects through two polynomial basis terms (Z₁ and Z₂). The model achieved an **R² of 0.356**, comparable to the full 30-lag regression model (R² = 0.383), but with vastly reduced dimensionality.

Only the first polynomial term (Z₁) was statistically significant (**p < 0.001**), suggesting a dominant concave lag shape peaking near the 24–28 day window. This confirms that pollutant influence on morbidity is not uniformly distributed but rather concentrated in a temporally focused window—consistent with the Almon weights and cross-correlation results reported earlier.

This parsimonious structure offers both interpretability and modeling efficiency, and was later adopted for lagged feature engineering in forecasting models.

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

To provide a standardized representation of air quality trends throughout the study period, a custom AQI proxy was developed by aggregating the normalized concentrations of seven key pollutants: black carbon, NO, NO₂, NOₓ, ozone, PM₁₀, and PM₂.₅. This composite index was designed to capture overall pollution load without relying on official AQI scales. The proxy was then examined across temporal and contextual axes—including weekdays versus weekends, and seasonal groupings—to support exploratory hypothesis generation and visual communication.

The seasonal overlay visualization of the AQI proxy, shown in **Figure 12**, revealed clear fluctuations across the annual cycle. Notably, the **winter months (highlighted in blue)** demonstrated the **highest variability and peak AQI proxy values**, including the absolute maximum near day 210. This aligns with known meteorological phenomena such as atmospheric inversion and reduced dispersion, as well as increased combustion-related emissions. In contrast, the **summer and spring periods exhibited relatively low and stable AQI proxy levels**, with summer showing the lowest average baseline. The **autumn season displayed a gradual rise in the proxy**, signaling the transition into the high-pollution winter months.

A graph showing a number of different colored lines

AI-generated content may be incorrect.

***Figure 12: AQI Proxy Over Time with Seasonal Overlay***

Complementary to the seasonal trend, weekday–weekend comparisons revealed higher concentrations of traffic-associated pollutants (e.g., NO₂, NO) during weekdays, yet marginally higher respiratory morbidity rates on weekends. Despite slightly lower average AQI proxy values on weekends (0.19 vs. 0.22), conditions such as **influenza-like illness** and **bronchiolitis** remained elevated. This divergence between exposure and outcome may reflect lagged health effects, behavioral differences, or confounding health system factors. The AQI proxy thus serves as a useful composite variable for both pattern recognition and as a feature in downstream health-outcome modeling.

**4.2 Models**

### **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 13: 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 14: 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 15: 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.**

**Insert Figure 5.4.1:** *Training and validation loss over 24 epochs for the initial LSTM model.*

Despite general downward convergence in training and validation loss, the presence of several volatility spikes in the validation curve around epochs 10–15 indicated **instability in learning**, likely due to sensitivity to input variance and the absence of dropout regularization.

#### **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.**

**Insert Figure 5.4.2:** *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.**

**Insert Figure 5.4.3:** SHAP summary plot showing the top 10 most impactful features in the LSTM model.

The top-ranked feature was **black\_carbon\_lag4**, exhibiting a **maximum SHAP value of 0.5554**, making it the most influential single input to the LSTM model. This underscores the model’s ability to recognize **localized pollutant spikes as immediate respiratory risk factors**, aligning with prior literature on the acute effects of black carbon exposure.

Following closely were **PM2.5-related features**, including lag12, lag7, and lag1, with SHAP values ranging from 0.425 to 0.455. These findings reinforce the clinical and epidemiological relevance of fine particulate matter in driving respiratory distress, particularly when exposures are sustained across multiple days.

Notably, **lagged ARI features** such as acute\_respiratory\_illness\_lag5, lag10, lag11, and lag12 also appeared within the top 10, though with **less dominance than in Random Forest or XGBoost** models. This reflects a key characteristic of LSTM architectures: they **integrate autoregressive memory without over-dependence on any single lag**.

Additional important contributors included **nitric oxide lags** (lag1, lag12, lag13), further confirming the LSTM's sensitivity to **gaseous pollutants with delayed inflammatory effects**.

This profile of feature importance emphasizes that the LSTM model successfully leverages **both exogenous pollutant exposures and endogenous health progression signals**, offering a nuanced understanding of ARI triggers over time.

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

While the LSTM did not outperform the tuned XGBoost model in error-based metrics, it showed several conceptual strengths:

* **Temporal generalization**: It learned lagged relationships directly from raw sequences without requiring explicit feature engineering.
* **Pollutant integration**: SHAP analysis confirmed the role of pollutants like black carbon and PM2.5 in shaping model output, a signal less visible in Random Forest outputs.
* **Reduced autoregressive bias**: Unlike tree models, the LSTM architecture balanced its attention between environmental and endogenous signals.

However, the model also exhibited several limitations:

* **Higher prediction error variance** (especially MAPE), due to sensitivity to sequence quality and limited training samples.
* **Longer training time and tuning requirements**, making deployment more resource-intensive.
* **Interpretability challenges**, although partially mitigated via SHAP.

In summary, the LSTM model offers **valuable complementary insights** to tree-based models, particularly in recognizing **nonlinear, delayed pollutant effects** on respiratory illness. Future work could benefit from integrating LSTM with attention mechanisms or hybrid ensemble architectures to combine the strengths of deep learning and gradient boosting.

## **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 6.1. 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 |

**Insert Table 6.1:** *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 Predictive Modeling with Machine Learning**

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: Conclusion & Future Word**

## **6.1 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.

## **6.2 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.

## **6.3 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.
* **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.

**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>