

Article

Forecasting Regional COPD Outpatient Visits Using Environmental and Meteorological Data in Pennsylvania

Basema Jarrar ¹, Parv Venkitasubramaniam ^{1,*}  and Hyunok Choi ²¹ Department of Data Science, Lehigh University, Bethlehem, PA 18015, USA; basemajr2003@gmail.com² Population Health, College of Health, Lehigh University, Bethlehem, PA 18105, USA; hyc219@lehigh.edu

* Correspondence: pav309@lehigh.edu

Abstract

Chronic obstructive pulmonary disease (COPD) remains a major cause of respiratory morbidity and healthcare utilization, creating challenges for healthcare planning, resource allocation, and early intervention. This study aimed to develop a forecasting framework for quarterly age-adjusted COPD outpatient visit rates across 762 regions in Pennsylvania from 2019 to 2023, using multi-source data including PHC4 outpatient records, satellite-derived environmental pollutant variables, and meteorological variables. To compare models that capture nonlinear relationships with those that explicitly model temporal dependencies, several machine learning models and a deep learning long short-term memory (LSTM) model, designed to learn sequential patterns and lagged temporal effects, were evaluated. The seasonal naïve baseline achieved $R^2 = 0.570$, classical machine learning models achieved $R^2 = 0.58\text{--}0.62$, and the LSTM model achieved $R^2 = 0.705$ with lower prediction error. Lagged COPD activity was the strongest predictor, while environmental, meteorological, and geographic variables provided additional predictive information. These findings highlight the value of integrating multi-source environmental data with sequence-based modeling for forecasting regional COPD activity and suggest that environmental and meteorological variables can provide additional predictive information beyond historical COPD activity alone.

Keywords: COPD forecasting; environmental health; temporal modeling; LSTM; air quality; meteorology; spatio-temporal prediction

1. Introduction

1.1. Burden of COPD and Healthcare Utilization

Chronic obstructive pulmonary disease (COPD) remains one of the most common and burdensome chronic respiratory diseases in the United States and is a leading cause of morbidity, mortality, and healthcare utilization. COPD contributes substantially to outpatient visits, hospitalizations, and medical expenses, placing a significant burden on patients and healthcare systems. As the occurrence of COPD continues to increase among aging populations, accurate forecasting of COPD-related healthcare activity has become increasingly important for public health planning, resource allocation, and early intervention strategies.

1.2. Environmental and Meteorological Influence on COPD

Although clinical risk factors such as smoking and age are well established, population-level COPD activity is also influenced by environmental exposures, seasonal variation [1,2],



Academic Editor: Ling Tim Wong

Received: 14 April 2026

Revised: 28 June 2026

Accepted: 29 June 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and broader meteorological conditions [3]. Fluctuations in air quality, temperature, and humidity can contribute to respiratory stress, exacerbations, and changes in healthcare utilization [4].

Environmental and atmospheric factors are widely recognized as important contributors of respiratory morbidity. Previous studies have shown that pollutants such as particulate matter and gaseous emissions are associated with increased respiratory admissions and healthcare utilization [1,4], while meteorological variables including temperature, humidity, and atmospheric pressure can modify pollutant behavior and influence airway inflammation. In addition, ecological and geographic characteristics such as elevation, vegetation cover, and urban structure may shape environmental exposure patterns, contributing to spatial variability [5–7] in COPD burden across regions.

1.3. Forecasting Respiratory Healthcare Activity

Despite growing evidence linking environmental conditions to respiratory outcomes, much of the existing literature focuses primarily on associations rather than prediction. Many environmental health studies analyze cross-sectional or short-term relationships using daily or annual data, while fewer studies attempt to forecast respiratory disease activity over intermediate time scales such as quarters. In addition, existing approaches often rely on raw hospitalization or visit counts, which may obscure demographic differences across regions when outcomes are not adjusted for population structure [6]. These limitations make it difficult to develop forecasting systems capable of supporting regional healthcare planning.

Several studies have explored forecasting approaches for respiratory healthcare activity. Peng et al. [8] developed machine learning models to forecast peak outpatient and emergency department visits among patients with chronic respiratory diseases and showed the potential of predictive analytics for healthcare resource planning. Ye et al. [9] evaluated a COPD health forecasting service in Shanghai and reported reductions in outpatient and emergency department utilization, highlighting the practical value of disease forecasting systems. These studies suggest that forecasting models can provide actionable information for healthcare management and public health surveillance; however, most of the existing work has focused on short-term forecasting horizons or healthcare systems outside the United States.

1.4. Machine Learning and Deep Learning Approaches

Recent advances in machine learning provide new opportunities for modeling complex environmental and epidemiological relationships. Ensemble methods such as Random Forest and Gradient Boosting can capture nonlinear interactions among environmental variables and have shown promising performance in respiratory disease prediction [10–12]. These approaches have also been increasingly applied to healthcare forecasting tasks, including respiratory disease surveillance, healthcare utilization prediction, and environmental health modeling. In particular, these models treat observations as independent samples and do not explicitly account for temporal ordering, which limits their ability to capture seasonal recurrence and delayed environmental effects. This limitation is especially important in chronic respiratory diseases, where current disease burden is strongly influenced by past patterns.

However, forecasting respiratory disease activity often requires models that explicitly incorporate temporal structure and sequential dependencies. Deep learning models, particularly long short-term memory (LSTM) networks, offer a powerful framework for modeling sequential processes and long-range temporal relationships. These architectures are capable of learning patterns across multiple time steps and may therefore be well suited

for forecasting disease activity influenced by lagged environmental exposures and seasonal recurrence [13–15].

1.5. Research Gap and Study Objective

Despite these advances, important gaps remain in the literature. Most existing studies focus on short-term forecasting of hospital admissions, emergency department visits, or patient-level outcomes rather than regional forecasting of age-adjusted COPD outpatient visit rates. In addition, relatively few studies have examined intermediate forecasting horizons, such as quarterly prediction, while integrating healthcare utilization, environmental, and meteorological data within a unified framework. Furthermore, direct comparisons between traditional machine learning approaches and sequence-based deep learning models for forecasting age-adjusted COPD outpatient visit rates at a regional scale remain limited. Addressing these gaps may improve understanding of regional COPD dynamics and support public health planning and resource allocation.

This study develops and evaluates a multi-source forecasting framework for age-adjusted COPD outpatient visit rates across 762 regions in Pennsylvania between 2019 and 2023. The proposed approach integrates outpatient health records with environmental and meteorological data, including satellite-derived pollutant variables, weather observations, and environmental indicators. Classical machine learning models are evaluated alongside a deep learning LSTM architecture to assess their ability to capture nonlinear relationships and temporal dependencies in COPD activity. By combining satellite-derived environmental data with modern predictive modeling techniques, this study aims to improve regional forecasting of respiratory disease burden and evaluate the predictive contribution of environmental and meteorological information for public health planning.

2. Materials and Methods

2.1. Study Area and Time Period

The study area is the state of Pennsylvania, United States. The analysis was conducted at the regional level defined by the Pennsylvania Health Care Cost Containment Council (PHC4), which divides the state into 762 regions. Each region represents a geographic unit formed by aggregating multiple ZIP codes (mapped to ZCTAs) into a single PHC4 region. This study was designed as a retrospective regional forecasting study using quarterly panel data. The final analytical dataset covers the period from 2019 through 2023, and each observation corresponds to a PHC4 region observed during a specific calendar quarter.

2.2. Data Sources

The primary health dataset used in this study was obtained from the Pennsylvania Health Care Cost Containment Council (PHC4), which provides outpatient encounter records from healthcare facilities across the state, where each record represents a single outpatient encounter rather than a unique patient. The dataset includes patient ZIP code, diagnosis codes, and encounter-level information. Direct visit dates were not available; therefore, temporal aggregation relied on the quarterly structure provided by PHC4 [16].

Daily air pollutant variables were obtained from Sentinel-5P TROPOspheric Monitoring Instrument (TROPOMI) Level-3 products available through Google Earth Engine [17–22]. Daily satellite products for carbon monoxide (CO), sulfur dioxide (SO₂), nitrogen dioxide (NO₂), and ozone (O₃) were exported at an approximately 3.5 km spatial resolution. Regional pollutant estimates were generated by applying zonal statistics to buffered PHC4 regions, producing daily mean column densities for each region. Daily regional values were subsequently aggregated to quarterly averages to match the temporal

resolution of the health data. Pollutant variables represent atmospheric column number densities and are reported in mol/m².

Meteorological variables were obtained from the GRIDMET dataset, which provides high-resolution gridded daily meteorological fields across the continental United States, including temperature, humidity, precipitation, wind speed, radiation, evapotranspiration, fuel moisture indicators, and fire-weather indices such as the burning index and energy release component [23].

Additional geospatial variables were included to capture regional structural variation. Elevation characteristics were derived from the USGS National Elevation Dataset (NED) at approximately 1/3 arc-second resolution (~10 m) [24]. Vegetation density was represented using the Normalized Difference Vegetation Index (NDVI) derived from Landsat 8 surface reflectance imagery at 30 m spatial resolution [25].

To ensure spatial consistency across datasets, individual patient-level outpatient records from PHC4 were first aggregated at the ZIP code level by quarter to obtain COPD encounter counts. ZIP codes were then mapped to Census ZCTAs using the 2021 ZIP–ZCTA crosswalk, and subsequently linked to PHC4 regions using a region–ZCTA aggregation file. This mapping enabled the integration of health, pollutant, meteorological, and geospatial variables within a unified regional framework. Quarterly COPD rates were computed at the regional level using age-specific COPD encounter counts, population estimates from the American Community Survey (ACS), and U.S. standard population weights. This procedure produced age-adjusted rates that improved comparability across regions with different age structures.

2.3. Data Preprocessing and Feature Engineering

Several preprocessing steps were performed before modeling. The PHC4 dataset consists of individual-level outpatient data, where each record represents a single outpatient encounter. These records were filtered to include only visits associated with COPD-related diagnosis codes (ICD-10 J44.*). Records with missing or invalid ZIP codes or diagnosis information were removed. Daily satellite-derived pollutant estimates and meteorological observations were averaged to quarterly values to match the temporal resolution of the health data.

The primary outcome variable was the quarterly mean age-adjusted outpatient COPD visit rate. To account for differences in age structure across regions, age-adjusted rates were calculated using age-specific COPD encounter counts, population estimates from the American Community Survey (ACS), and U.S. standard population weights. The age-adjusted rate was computed as

$$\text{Age-Adjusted Rate} = \sum_{i=1}^k \left(\frac{A_i}{P_i} \times W_i \right) \times 100,000$$

where A_i is the number of COPD encounters in age group i , P_i is the population of age group i , and W_i is the corresponding U.S. standard population weight. This procedure produces an age-adjusted COPD rate per 100,000 population and improves comparability across regions with different demographic compositions.

Feature engineering aimed to capture temporal dynamics, delayed environmental effects, and static regional structure. Calendar year, quarter, and an ordinal quarter index were included as temporal predictors. Lagged outcome and exposure variables were constructed to capture persistence and delayed effects. The lagged COPD rate from the same region in the previous year (lag 4 quarters) was included because it captures annual recurrence patterns and proved to be a strong predictor of future COPD activity. For environmental and meteorological variables, lag 1 and lag 2 quarter features were created.

The meteorological feature set included both conventional weather variables and derived environmental indicators available from GRIDMET, including evapotranspiration, fuel moisture measures, and fire-weather indices. Lag 4 environmental features were evaluated but removed because they introduced substantial missingness and did not improve model performance. The final feature set consisted of 23 predictor variables, including satellite-derived pollutant variables, meteorological variables, geospatial characteristics, temporal indicators, and lagged COPD information.

Static geospatial covariates were also included. Elevation mean and standard deviation summarize topographic characteristics within each region, while the NDVI provides a measure of vegetation cover. These features do not vary over time but help explain systematic geographic variation in COPD burden.

2.4. Modeling Dataset

The final modeling dataset consisted of 15,188 region–quarter observations covering all 762 PHC4 regions from 2019 through 2023. A complete panel of 762 regions observed across 20 quarters would contain 15,240 region–quarter observations; however, 52 region–quarter combinations were not present in the available PHC4-derived quarterly COPD dataset. The original PHC4 data consist of individual-level outpatient records, where each record represents a single outpatient encounter. After preprocessing and aggregation, each observation in the modeling dataset corresponds to one region in one quarter, where regions are defined by mapping multiple ZCTAs to PHC4-defined geographic units. Each observation includes the target variable, environmental and meteorological predictors, temporal indicators, lagged variables, and static regional characteristics.

In addition to quarterly pollutant averages, weekly pollutant means were also calculated from daily Sentinel-5P regional pollutant estimates and summarized at the regional level for each quarter. These features were included to capture shorter-term fluctuations in air quality that may not be fully represented by quarterly averages. Both quarterly pollutant summaries and quarterly summaries of weekly pollutant means were evaluated as candidate predictors.

Missing values occurred primarily in lagged variables created during feature engineering for the classical machine learning models. These missing values arose at the beginning of regional time series because historical observations required for lag construction were unavailable. Missing values were imputed using median values estimated from the training data and subsequently applied unchanged to the validation and test datasets. For the LSTM model, however, the input sequences were constructed using complete quarterly features and past age-adjusted COPD rates; no missing values were present in the variables used to build the final four-quarter sequence windows. Therefore, median imputation was not applied within the LSTM lookback windows and did not interrupt temporal continuity in the sequential inputs. After preprocessing, no missing values remained in the modeling datasets used for training and evaluation.

For models requiring scaled inputs, such as the LSTM network, feature standardization was performed using Z-score scaling based on the training data only. Specifically, the mean and standard deviation of each predictor were calculated from the training set, and these parameters were subsequently applied unchanged to the validation and test sets. This procedure ensured that predictors with different units and scales contributed comparably during neural network training while preventing information leakage from future observations. Tree-based machine learning models (Random Forest, Gradient Boosting, XGBoost, and LightGBM) were trained using the original feature scales because these algorithms are generally insensitive to differences in predictor scaling.

A strict temporal split was used for forecasting:

- Training: 2020–2021;
- Validation: 2022;
- Test: 2023.

This sequential partition prevents future information from influencing model training and reflects a realistic forward-looking forecasting scenario.

2.5. Descriptive Statistics

Table 1 summarizes the distributions of the outcome variable and selected environmental, meteorological, and geospatial predictors included in the analysis. The quarterly age-adjusted COPD outpatient visit rate exhibited substantial regional variation, reflecting differences in disease burden and healthcare utilization across Pennsylvania. Environmental and meteorological variables also showed considerable variability across regions and quarters, supporting their inclusion as potential predictors in the forecasting framework.

Table 1. Summary Statistics for Key Variables in the Final Analytical Dataset.

Variable	Mean	Std. Dev.	Min	Max
Health Outcome				
Age-adjusted COPD rate	192.837	134.294	0.000	3537.127
Pollutants (Quarterly Averages)				
CO (mol/m ²)	0.0340	0.0030	0.0260	0.0420
NO ₂ (mol/m ²)	0.000046	0.000021	0.000017	0.000136
O ₃ (mol/m ²)	0.1430	0.0090	0.1260	0.1590
SO ₂ (mol/m ²)	0.000306	0.000228	−0.000043	0.001315
Meteorological Variables				
Minimum temperature (°C)	6.034	7.615	−9.834	20.225
Maximum temperature (°C)	16.781	8.191	1.207	29.897
Maximum relative humidity (%)	85.477	5.187	71.553	98.053
Minimum relative humidity (%)	46.190	3.591	36.837	61.125
Precipitation amount (mm)	3.372	1.013	1.224	8.532
Wind velocity at 10 m (m/s)	4.104	0.666	2.614	6.500
Surface downward shortwave radiation	172.543	63.817	80.834	260.236
Static Regional Characteristics				
Elevation mean (m)	260.433	156.688	2.502	720.215
Elevation std (m)	45.915	33.180	1.497	151.025
NDVI	0.638	0.138	0.124	1.018

Note: Pollutant variables represent Sentinel-5P satellite-derived regional atmospheric column number densities summarized using zonal statistics. Values are reported in mol/m² and should not be interpreted as ground-level ambient concentrations. Small negative SO₂ values reflect satellite retrieval uncertainty when atmospheric column amounts are close to the detection limit.

2.6. Machine Learning Models

Four tree-based ensemble models were trained and evaluated: Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, and LightGBM Regressor. These models were selected because they can capture nonlinear interactions, handle mixed-scale predictors, and provide interpretable feature importance measures.

All predictor variables were numeric and did not require additional encoding. Tree-based models were trained on the original feature scales. All four tree-based models were initially trained using their default hyperparameters. This approach was adopted to provide a consistent baseline comparison across models and to reduce the risk of overfitting given the limited number of validation quarters available for tuning.

2.7. Deep Learning Model

To better capture temporal dependencies, a sequence-based long short-term memory (LSTM) model was developed. The final architecture was trained on sequences of four consecutive quarters ($T = 4$). A sequence length of four quarters was selected because it captures one complete annual cycle of COPD activity while maintaining sufficient training sequences within the available study period. Each training instance consisted of a four-quarter sequence of predictor variables together with an eight-dimensional embedding vector representing the PHC4 region. Several embedding dimensions were explored during model development, and an embedding dimension of eight produced the strongest validation performance and was therefore used. Each sequence contained 23 predictor variables observed across four consecutive quarters, resulting in an input tensor of shape (4, 23) for each training sample.

The final model architecture included an embedding layer for regional identifiers with an embedding dimension of 8, followed by a single LSTM layer with 64 hidden units. To reduce overfitting, the LSTM layer used a dropout rate of 0.20 and a recurrent dropout rate of 0.10. The default Keras activation functions were used, corresponding to a hyperbolic tangent activation for the cell state and a sigmoid activation for the recurrent gates. The LSTM output was concatenated with the regional embedding vector and passed to a dense layer with 64 units and ReLU activation, followed by an additional dropout layer (dropout = 0.20) and a single linear output node predicting the age-adjusted COPD outpatient visit rate.

The final model contained 33,361 trainable parameters. Training was performed using the Adam optimizer with an initial learning rate of 0.003 and mean squared error as the loss function, while mean absolute error was monitored as an additional evaluation metric. Models were trained for up to 200 epochs with a batch size of 128. Early stopping monitored validation loss with a patience of 12 epochs and restored the best-performing model weights. In addition, a ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 after 6 epochs without validation-loss improvement, with a minimum learning rate of 1×10^{-5} .

Although both LSTM and GRU architectures were initially explored, the LSTM model achieved superior validation and test performance and was therefore selected as the final forecasting model.

2.8. Model Evaluation

Model performance was assessed using the 2023 test set. Three standard regression metrics were used: coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). Validation performance on 2022 was used to monitor generalization.

To interpret model behavior, feature importance was examined using gain-based importance from XGBoost and SHAP (SHapley Additive explanations) values. Feature importance analyses were used to interpret model predictions and predictive contributions of variables rather than to infer causal relationships. For the LSTM model, feature sensitivity was assessed using leave-one-feature-out (LOFO) analysis, which quantified the change in predictive error when individual features were removed. Regional forecast performance was further examined by computing MAE and RMSE separately for each PHC4 region and visualizing their spatial distribution across Pennsylvania.

3. Results

3.1. Performance of Classical Machine Learning Models

A seasonal naïve forecasting baseline and four tree-based ensemble models were evaluated and compared with the proposed LSTM framework. The seasonal naïve baseline predicted each quarter using the observed COPD rate from the same quarter in the previous year. On the 2023 test set, the seasonal naïve model achieved ($R^2 = 0.570$), MAE = 64.29, and RMSE = 97.76. The machine learning models achieved test-set (R^2) values ranging from 0.583 to 0.616 (Table 2), outperforming the seasonal naïve approach. Among the classical machine learning models, XGBoost performed best, achieving a test-set (R^2) of 0.616, followed by Random Forest ($R^2 = 0.604$) and LightGBM ($R^2 = 0.590$). Gradient Boosting showed the weakest performance, with a test-set (R^2) of 0.583.

Table 2. Performance comparison of machine learning and deep learning models.

Model	Split	R^2	MAE	RMSE
Seasonal Naïve	Test	0.570	64.29	97.76
Random Forest	Train	0.952	17.99	26.26
	Validation	0.452	68.10	113.01
	Test	0.604	66.72	93.88
Gradient Boosting	Train	0.733	45.22	62.11
	Validation	0.417	70.19	116.48
	Test	0.583	69.15	96.33
LightGBM	Train	0.912	25.94	35.64
	Validation	0.454	70.01	112.73
	Test	0.590	64.93	95.52
XGBoost	Train	0.888	30.54	40.18
	Validation	0.448	72.34	113.36
	Test	0.616	66.56	92.43
LSTM	Train	0.842	34.34	47.80
	Validation	0.626	54.95	93.36
	Test	0.705	56.13	81.04

Abbreviations: LSTM, long short-term memory; MAE, mean absolute error; RMSE, root mean squared error; R^2 , coefficient of determination.

Although the classical machine learning models captured meaningful nonlinear structure and outperformed the seasonal naïve baseline, their validation performance was noticeably lower than their training performance, suggesting moderate overfitting and weaker generalization across future quarters. The LSTM model achieved the strongest overall performance, indicating that sequence-based modeling provided additional predictive value beyond simple annual recurrence patterns and traditional machine learning approaches.

The lower performance observed on the 2022 validation set compared with the 2023 test set was consistent across all models. This pattern may reflect the transitional nature of healthcare utilization during the post-pandemic recovery period. Following the substantial disruption in outpatient healthcare activity during 2020 and 2021, some regions experienced atypical fluctuations in COPD utilization during 2022 as healthcare services and patient behavior returned toward pre-pandemic patterns. By 2023, utilization patterns appeared more stable and recurrent, which may have contributed to improved predictive performance across both machine learning and deep learning models.

3.2. Feature Importance Analysis

To identify the strongest predictors of COPD activity, feature importance was examined using XGBoost gain-based importance and SHAP values. Because XGBoost provided the highest test-set performance among the classical models, it was selected for feature importance and SHAP analyses.

Figure 1 shows the gain-based feature importance scores. The lagged COPD rate from the same quarter in the previous year `age_adj_lag4` was by far the most influential predictor, highlighting the strong annual recurrence pattern in COPD activity. The ordinal quarter index also ranked highly, confirming the importance of temporal structure. Elevation mean and standard deviation were among the next most important features, suggesting that persistent regional differences were useful for model prediction. Several lagged environmental predictors, including ozone lag-2, humidity lag-1, and shortwave radiation lags, also provided measurable predictive value.

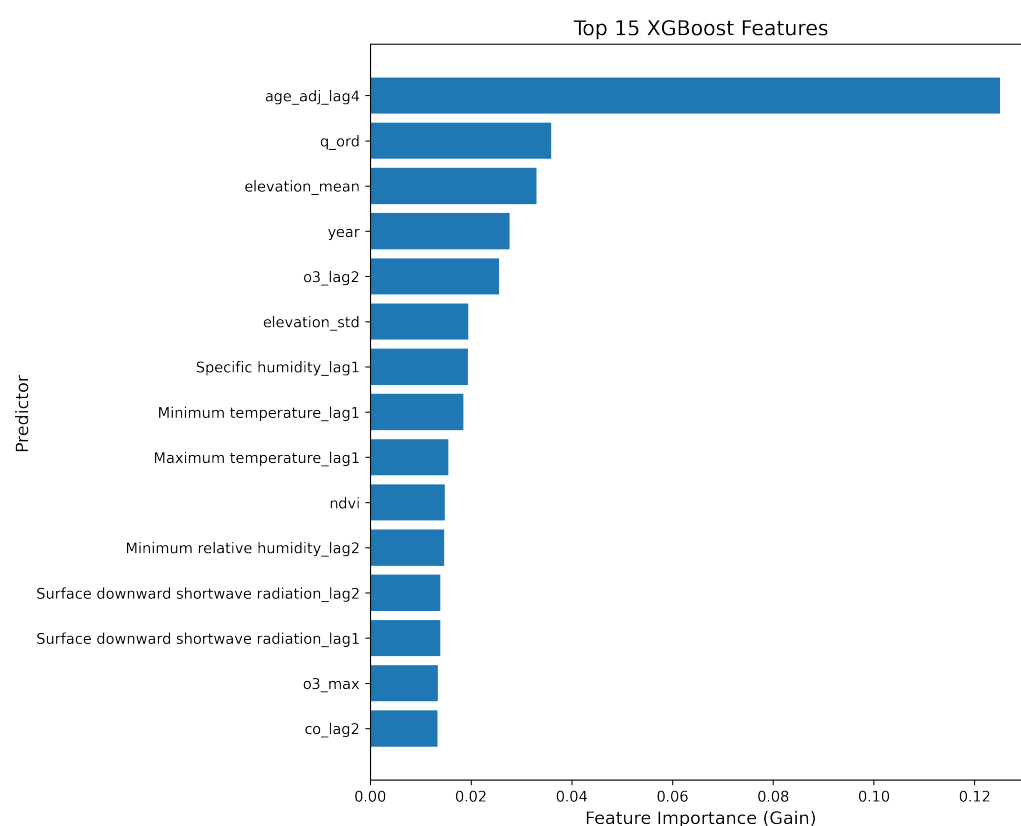


Figure 1. XGBoost gain-based feature importance scores for the final model.

SHAP analysis produced a consistent pattern. As shown in Figure 2, the lagged COPD rate generated the largest shifts in predicted values, while elevation, the NDVI, ozone, humidity, temperature, and radiation contributed moderate but systematic effects.

These results indicate that prior COPD activity was the dominant source of predictive information across the classical models. Environmental and meteorological variables contributed additional predictive value, although their influence was substantially smaller than that of the lagged COPD rate.

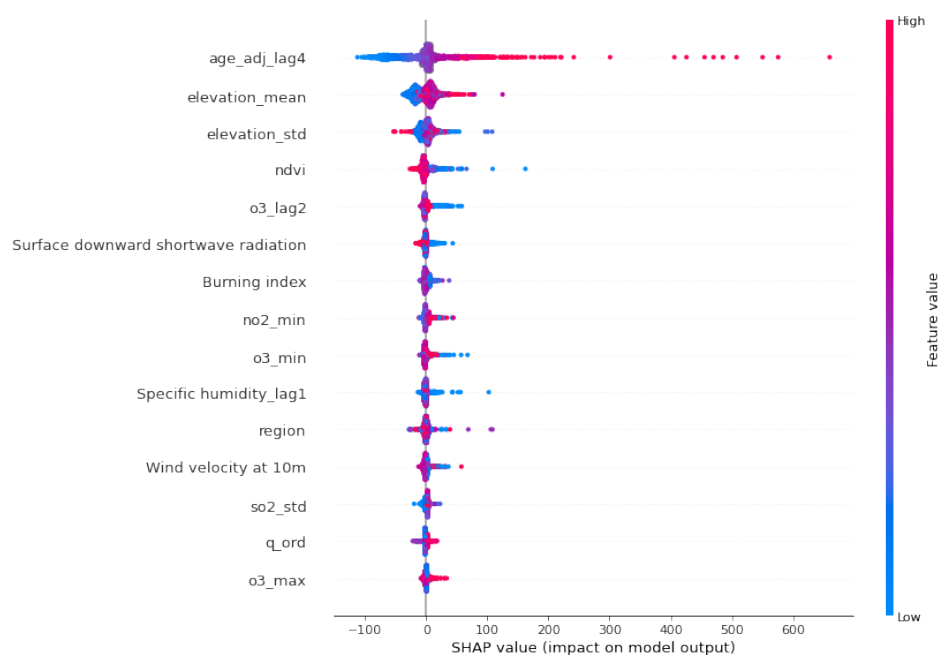


Figure 2. SHAP summary plot showing global feature influence in the XGBoost model.

3.3. Performance of the LSTM Model

The sequence-based LSTM model achieved the strongest overall predictive performance. Compared with the best-performing classical model (XGBoost, $R^2 = 0.616$), the LSTM improved predictive performance by approximately 14% in terms of explained variance. On the 2023 test set, the model obtained $R^2 = 0.705$, MAE = 56.13, and RMSE = 81.04, exceeding all classical machine learning baselines (Table 2).

Training and validation loss curves are shown in Figure 3. The validation loss stabilized after several epochs and followed the training loss closely, indicating controlled overfitting and effective use of early stopping.

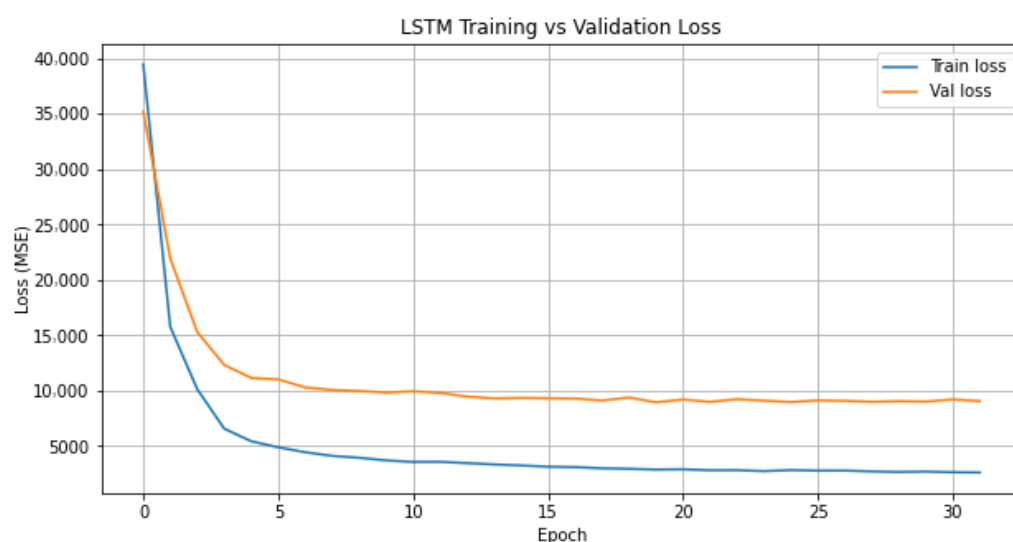


Figure 3. Training and validation loss curves for the LSTM model.

A scatter plot of actual versus predicted COPD rates for the 2023 test set is shown in Figure 4. The model captured low and moderate COPD rates well, while underestimating the highest values, which were relatively rare and exhibited greater variability. This pattern is also clear in the residual plots, where prediction errors increase at higher COPD rates. From a forecasting perspective, this underestimation is important because extreme COPD

activity may represent periods of elevated healthcare demand. The reduced accuracy for these high-burden observations suggests that additional predictors or longer historical time series may be beneficial to improve the prediction of rare but clinically important peaks.

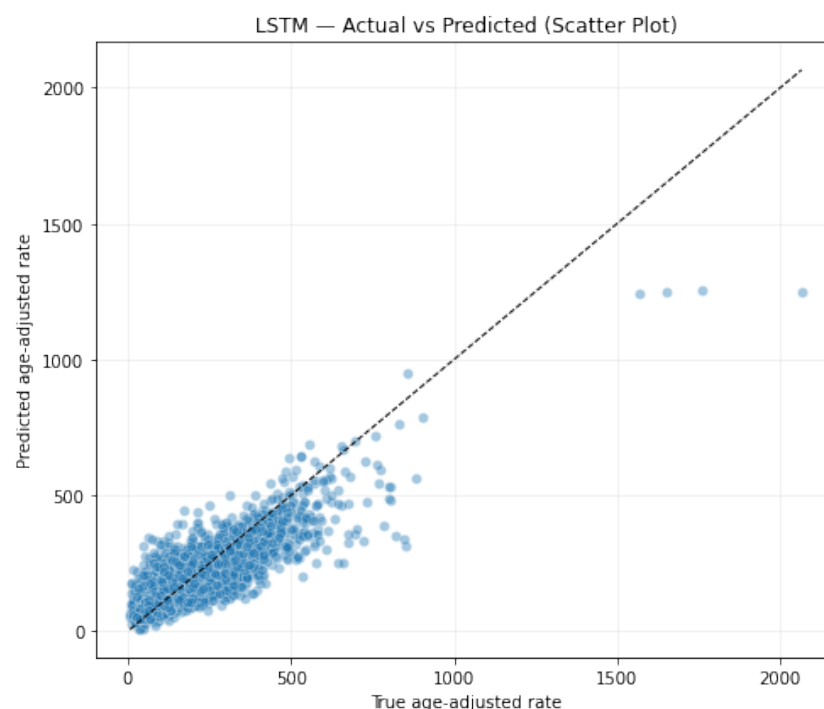


Figure 4. Actual versus predicted age-adjusted COPD rates for the LSTM model on the test set. Blue dots represent individual observations in the 2023 test set, and the dashed line represents the line of perfect agreement between observed and predicted values (1:1 line).

Residual diagnostics further supported model stability. The residual distribution was approximately symmetric and centered near zero (Figure 5), although residual magnitude increased for very high COPD rates (Figure 6), which is expected when extreme outcomes are rare.

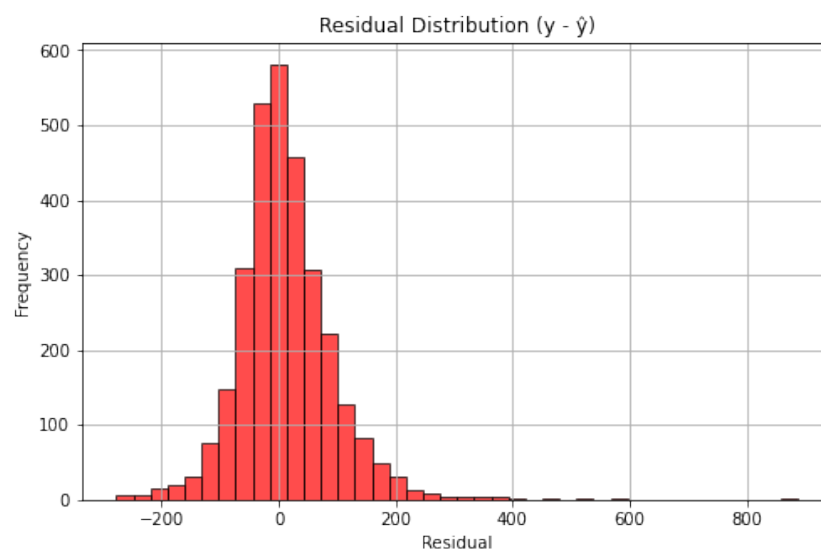


Figure 5. Distribution of residuals (true minus predicted values) for the LSTM model.

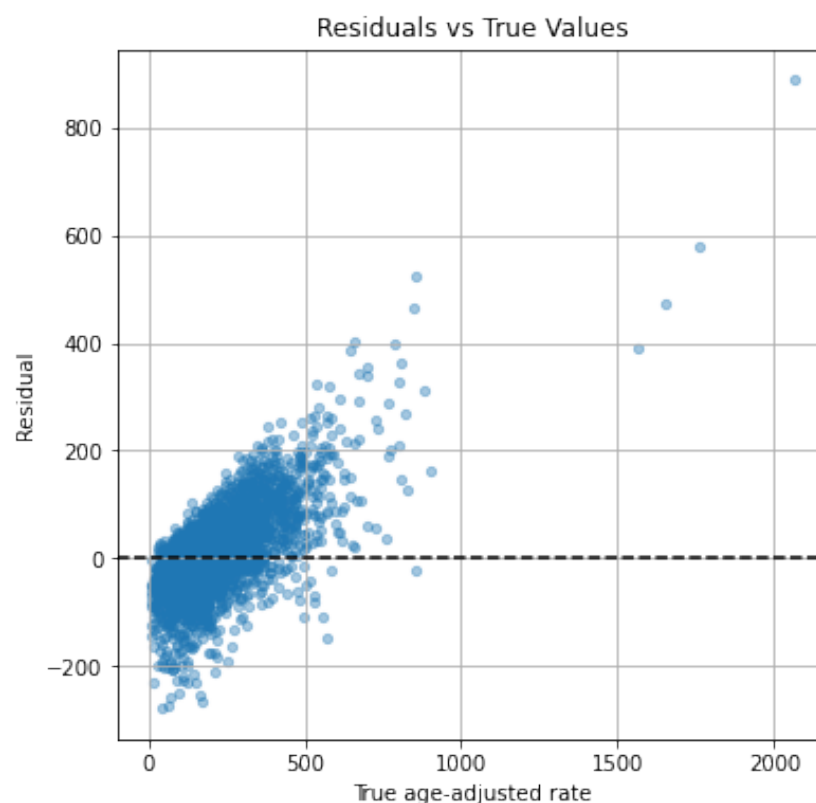


Figure 6. Residuals (observed minus predicted values) versus the true age-adjusted COPD rates for the 2023 test set. Each point represents one observation.

Quarter-specific performance in 2023 remained relatively stable, with slightly higher MAE in the second quarter (Figure 7).

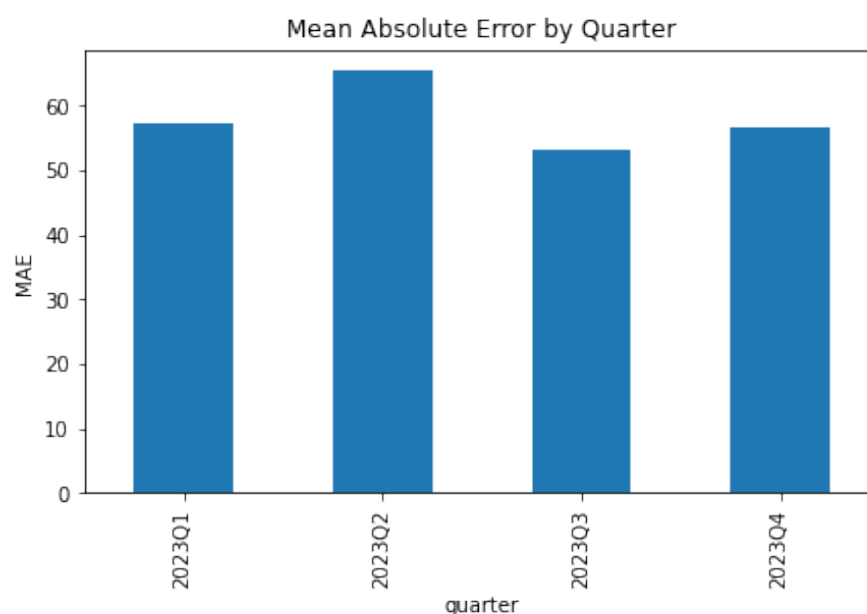


Figure 7. Mean absolute error for each quarter of the 2023 test set.

3.4. Feature Sensitivity of the LSTM Model

LOFO analysis was used to assess the sensitivity of the LSTM model to individual predictors. The age-adjusted COPD rate from previous quarters was by far the most important predictor, confirming the dominant role of temporal persistence and seasonal recurrence

in COPD forecasting (Table 3). Among the environmental and geographic variables, the NDVI, CO, elevation mean, and O₃ showed the largest positive contributions to predictive performance. Most remaining meteorological and pollutant variables produced relatively small or negative LOFO scores, indicating that their predictive information was largely captured by stronger temporal and regional signals already present in the model.

Table 3. LOFO sensitivity values for selected features in the LSTM model.

Feature	LOFO Score
Age-adjusted rate	32.17
NDVI	1.74
CO	1.51
Elevation mean	1.02
O ₃	0.65
100-hour dead fuel moisture	0.23
Maximum relative humidity	−0.16
1000-hour dead fuel moisture	−0.38
Minimum relative humidity	−0.43
Energy release component	−0.44
Burning index	−0.60
Elevation std	−0.71
SO ₂	−1.22
Wind velocity at 10 m	−1.27
Wind direction	−1.37
Precipitation amount	−1.38
Maximum temperature	−1.51
Mean vapor pressure deficit	−1.56

LOFO = leave-one-feature-out. Positive values indicate performance degradation when a feature is removed, suggesting greater importance. Negative LOFO scores do not necessarily indicate that a variable is uninformative; they may arise from redundancy among correlated predictors or variability introduced during model retraining.

3.5. Regional Forecast Performance

Regional forecast error was evaluated by computing MAE and RMSE separately for each PHC4 region in the 2023 test set. Figures 8 and 9 show the spatial distribution of these errors.

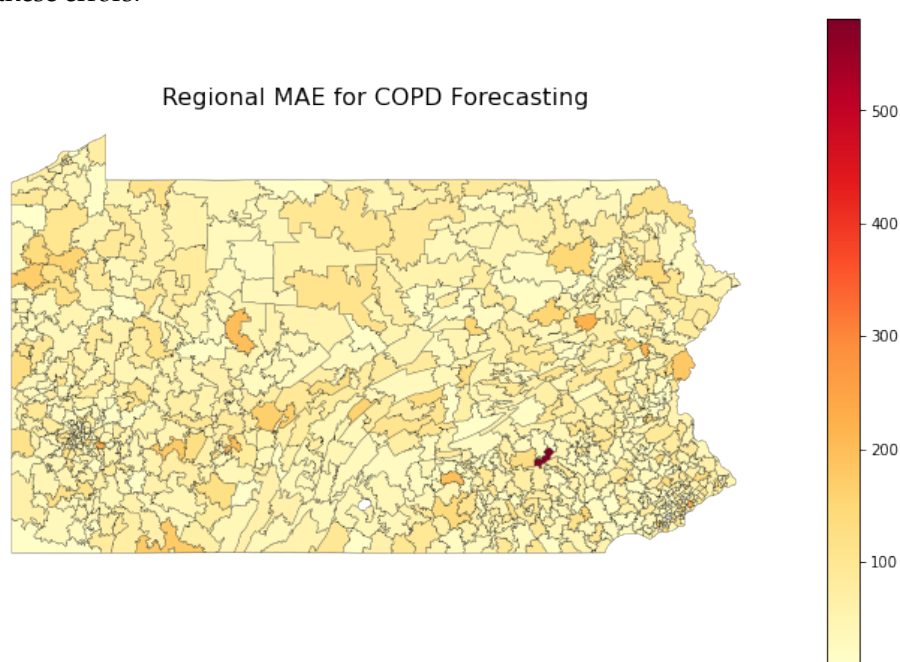


Figure 8. Regional MAE for COPD forecasting in the 2023 test set.

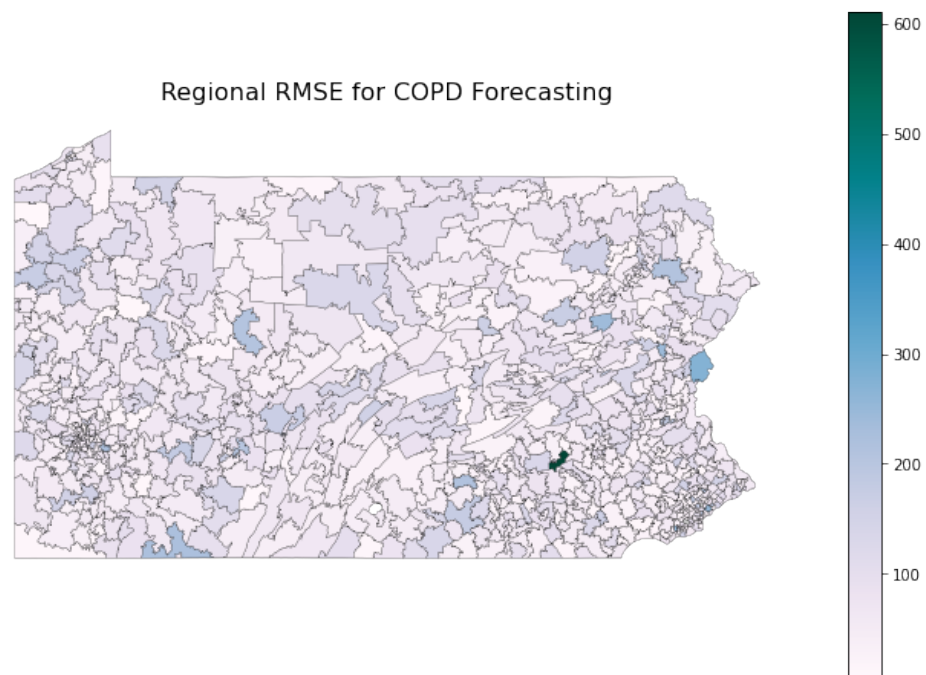


Figure 9. Regional RMSE for COPD forecasting in the 2023 test set.

Each observation represents one PHC4 region in one quarter. Lower values indicate better predictive performance.

Most regions showed relatively low forecast error, suggesting that the LSTM model captured local temporal patterns consistently across the state. However, a smaller subset of regions exhibited substantially higher error. These included several high-volume urban areas, such as parts of Philadelphia (region 696), Delaware County (region 740), and eastern Allegheny County (region 624), where large outpatient volumes and stronger quarter-to-quarter variability make accurate prediction more challenging. Higher errors were also observed in several rural or semi-rural regions, including Somerset and Fayette (region 368), the Poconos area of Monroe County (regions 376 and 698), Luzerne County (region 687), and Wayne/Lackawanna (region 459). In these lower population regions, relatively small changes in COPD visit counts can produce larger relative forecast errors, particularly during the post-pandemic rebound observed in 2023. Overall, the spatial analysis indicates that the LSTM model generalized well across most of Pennsylvania but tended to underestimate COPD activity in a small number of high-burden or highly variable regions.

4. Discussion

This study developed a forecasting framework for quarterly age-adjusted COPD outpatient visit rates across Pennsylvania by integrating healthcare utilization data with environmental, meteorological, and geospatial predictors. The results indicate that historical burden, regional heterogeneity, and environmental conditions all contribute to model predictions of COPD activity, although their relative importance differs substantially.

4.1. Model Performance and Temporal Dynamics

The comparison between classical machine learning models and the sequence-based LSTM architecture highlights the importance of explicitly modeling temporal structure. Tree-based ensemble models, including Random Forest, Gradient Boosting, XGBoost, and LightGBM, captured nonlinear relationships among environmental and meteorological variables and achieved test-set performance in the range of $R^2 = 0.58$ – 0.62 . However, their

lower validation performance suggests moderate overfitting and difficulty representing seasonal recurrence when observations are treated independently.

In contrast, the LSTM model achieved substantially stronger generalization, reaching a test-set $R^2 = 0.705$. By learning from sequences of quarterly observations and incorporating regional embeddings, the model was able to capture both seasonal recurrence and region-specific heterogeneity. These results suggest that sequence-based architectures are particularly well suited for forecasting chronic respiratory outcomes that exhibit recurring seasonal patterns and gradual temporal evolution.

Although direct baseline studies for quarterly COPD forecasting are limited, comparable studies in respiratory disease prediction typically report moderate predictive performance using conventional machine learning models, often in the range of $R^2 \approx 0.4\text{--}0.6$. In this context, the LSTM model's performance ($R^2 = 0.705$) represents a meaningful improvement, highlighting the benefit of explicitly modeling temporal dependencies while adding additional environmental information.

4.2. Environmental and Meteorological Predictors

Although environmental and meteorological variables contributed to model performance, their influence remained secondary to the temporal persistence of COPD activity. Across both feature importance and LOFO analyses, the lagged age-adjusted COPD rate consistently emerged as the dominant predictor. In the LOFO analysis, the lagged COPD rate produced a score of 32.17, which is approximately 18 times larger than that of the next strongest predictor, emphasizing the dominant role of historical COPD activity in forecasting future outpatient visit rates.

Among the remaining predictors, the NDVI, CO, elevation mean, and O₃ showed the largest positive contributions to predictive performance. In the XGBoost model, lagged ozone variables also ranked among the most important environmental predictors, suggesting that satellite-derived air-quality indicators contributed useful predictive information beyond historical COPD activity. Most meteorological variables exhibited relatively small or negative LOFO scores once temporal and regional information were included in the model, indicating that much of their predictive value may already be captured through temporal patterns and interactions with other variables.

These findings suggest that historical COPD activity remains the primary source of forecasting performance, while environmental, meteorological, and geographic variables provide supplementary predictive information that improves regional forecasting accuracy.

4.3. Spatial Forecasting Patterns

The spatial evaluation of forecasting error revealed that most regions across Pennsylvania exhibited relatively low MAE and RMSE values, indicating that the LSTM model captured locally stable temporal patterns. However, a subset of regions showed higher forecast errors.

Higher errors were observed in several large urban regions, including parts of Philadelphia (region 696), Delaware County (region 740), and eastern Allegheny County (region 624), where large outpatient volumes and strong quarter-to-quarter variability increase forecasting difficulty. Elevated errors were also found in some rural or semi-rural regions, including Somerset and Fayette (region 368), Monroe County and the Pocono region (regions 376 and 698), Luzerne County (region 687), and Wayne/Lackawanna (region 459). In these lower-population areas, relatively small fluctuations in visit counts can produce larger relative prediction errors.

These spatial patterns indicate that forecasting performance is partly determined by the underlying variability in regional COPD activity and highlight the importance of region-aware modeling strategies.

4.4. Impact of the COVID-19 Pandemic

The COVID-19 pandemic introduced a structural break in outpatient healthcare utilization. During 2020 and 2021, COPD outpatient visits were substantially suppressed due to changes in healthcare access, clinical practice, and public health restrictions. These suppressed values were included in the model training period and likely influenced the learned temporal patterns.

As outpatient utilization rebounded during 2022 and 2023, some regions experienced sudden increases in COPD activity that diverged from the pandemic-era trends. Because the LSTM model learned sequences that included the pandemic downturn, it occasionally underestimated the magnitude of this rebound. This effect was most pronounced in regions with high COPD burden or unusually sharp recovery patterns. The pandemic therefore represents an important limitation in temporal forecasting models trained across periods of structural disruption.

The pandemic-related disruption can also affect model generalizability because relationships learned during periods of reduced healthcare utilization may not fully reflect post-pandemic utilization patterns. Although a formal sensitivity analysis excluding the pandemic period was not performed in the current analysis, the observed rebound in 2022 and 2023 suggests that structural changes in healthcare utilization should be considered when interpreting forecasting performance. Future work could evaluate alternative modeling strategies that explicitly account for pandemic-related disruptions or assess model robustness using non-pandemic training periods.

The underestimation of high COPD rates observed in Figure 4 is consistent with this interpretation. Because the model was trained primarily on data from 2020 and 2021, when outpatient utilization was substantially reduced, the learned baseline may not fully reflect the higher levels of COPD activity observed during the post-pandemic recovery period. As a result, the model tended to underestimate some of the highest-burden observations in 2022 and 2023. This limitation is important when considering operational deployment, as systematic underprediction during periods of elevated healthcare demand could lead to underestimation of resource requirements. Future work could evaluate the inclusion of pandemic-period indicator variables or other intervention-based modeling approaches to explicitly account for pandemic-related structural disruptions and improve forecasting robustness.

4.5. Strengths and Limitations

This study has several methodological strengths. First, it integrates multiple independent datasets, including PHC4 outpatient records, satellite-derived pollutant variables, GRIDMET meteorological variables, and geospatial environmental features. The resulting dataset incorporates a wide range of environmental predictors, including satellite-derived pollutant variables, meteorological variables, evapotranspiration measures, fuel moisture indicators, fire-weather indices, elevation metrics and vegetation characteristics. Harmonizing these diverse data sources enabled the construction of a consistent regional panel dataset that captures both environmental conditions and COPD-related healthcare utilization throughout Pennsylvania.

Second, the feature engineering strategy incorporated both lagged outcomes and lagged environmental variables, reflecting the delayed and cumulative nature of respiratory disease processes. The use of age-adjusted COPD outpatient visit rates improved

comparability between regions with different population age structures, while the state coverage of 762 PHC4 regions enabled a comprehensive assessment of regional variation. Furthermore, the temporal train–validation–test split provided a realistic evaluation of forecasting performance, and the use of SHAP and LOFO analyses improved model interpretability by identifying the relative predictive contributions of individual features. Finally, the study addresses a practical public health problem by supporting regional respiratory disease forecasting and healthcare planning.

Despite these strengths, several limitations should be considered. First, the COVID-19 pandemic introduced substantial structural changes in healthcare utilization that may affect model generalizability. Outpatient COPD visits were suppressed during 2020 and 2021 and rebounded during 2022 and 2023, potentially influencing the temporal patterns learned by the models.

Second, the analysis relied on quarterly aggregation, which smooths short-term variation in both environmental exposures and healthcare utilization. The lack of direct visit dates in the PHC4 dataset required aggregation to quarterly time intervals and may have reduced the ability to capture shorter-term fluctuations in COPD activity. Additionally, COPD identification relied on administrative outpatient encounter records and diagnosis coding and may therefore be subject to coding errors or misclassification.

Third, environmental exposure measurements may be imperfectly captured in some regions. The pollutant variables used in this study were derived from Sentinel-5P satellite observations, which measure atmospheric column densities rather than ground-level pollutant concentrations. Consequently, these variables should be interpreted as regional atmospheric indicators rather than direct measures of surface-level exposure. Although satellite observations provide complete spatial coverage, they may not fully capture near-surface pollutant concentrations, and converting column densities to ground-level concentrations would require additional pollutant-specific assumptions that were beyond the scope of this study.

Fourth, the outcome represents outpatient healthcare utilization rather than the underlying biological burden of COPD. Consequently, observed rates may be influenced by healthcare access, care-seeking behavior, coding practices, and other healthcare system factors in addition to disease activity. Although age adjustment improves comparability across regions, it does not fully account for other population-level determinants of COPD burden.

Finally, several potentially important predictors were not included in the modeling framework. These include smoking prevalence, socioeconomic status, poverty, healthcare access, comorbidity burden, viral circulation, and particulate matter exposure (PM_{2.5} and PM₁₀). Although wildfire-related environmental indicators such as the burning index, fuel moisture, and energy release component were evaluated, these variables should not be interpreted as direct measures of smoke exposure. The absence of PM_{2.5} and PM₁₀ data limits the ability to directly assess the contribution of particulate matter and wildfire-related air-quality exposures to COPD activity. The study also did not explicitly evaluate spatial autocorrelation between neighboring regions. Adjacent regions may share similar environmental exposures, healthcare systems, and demographic characteristics, which could introduce spatial dependence in both the outcome and predictor variables. Future work could incorporate spatial statistical methods or spatially aware deep learning approaches to explicitly model these relationships. In addition, external validation was not performed in another state or healthcare system, and the relatively short study period may limit the stability and long-term generalizability of deep learning models trained on quarterly data.

4.6. Public Health Implications

The ability to forecast regional COPD activity has several practical implications for healthcare planning and public health policy. Reliable forecasts can support early identification of regions likely to experience seasonal increases in respiratory burden, allowing health systems to better anticipate potential increases in respiratory healthcare demand and plan staffing, medication supplies, and outpatient care resources accordingly.

In addition, identifying regions with both high COPD burden and higher forecast uncertainty may help guide targeted surveillance efforts and environmental monitoring. Strengthening environmental monitoring coverage, including integration of satellite-derived indicators with ground-monitor observations, could further improve future forecasting systems. More broadly, these results demonstrate the potential value of combining environmental monitoring with modern machine learning approaches to support respiratory health surveillance and planning.

5. Conclusions

This study developed a multi-source forecasting framework for quarterly age-adjusted COPD outpatient visit rates across 762 regions in Pennsylvania from 2019 to 2023 by integrating PHC4 outpatient records with environmental, meteorological, and geospatial datasets. The analysis compared several classical machine learning models with a sequence-based LSTM architecture to evaluate their ability to capture temporal and environmental patterns associated with COPD activity.

The results indicate that explicit temporal modeling substantially improves forecast performance. Tree-based machine learning models achieved moderate predictive accuracy ($R^2 \approx 0.58\text{--}0.62$), while the LSTM model achieved the strongest performance with a test-set $R^2 = 0.705$. Across all models, the lagged COPD rate from the previous year was the most influential predictor, highlighting the strong temporal persistence of COPD activity. Environmental, meteorological, and geographic variables, particularly the NDVI, CO, elevation, and O₃, provided additional predictive information but played a secondary role relative to temporal structure.

Regional diagnostics indicated that the model generalized well across most of Pennsylvania, with larger errors concentrated in high-burden or high-variability regions. The COVID-19 pandemic introduced a structural break in healthcare utilization that likely contributed to underestimation in some areas during the post-pandemic rebound.

To the best of our knowledge, this study is among the first to evaluate regional forecasting of quarterly age-adjusted COPD outpatient visit rates using integrated healthcare, environmental, meteorological, and geospatial data within a unified machine learning and deep learning framework.

Overall, the findings demonstrate the value of sequence-based modeling for regional COPD forecasting, with environmental and meteorological variables providing additional predictive information beyond historical COPD activity. Such approaches may support public health planning by improving the ability to anticipate seasonal changes in COPD burden and identify regions where enhanced surveillance and healthcare planning efforts may be warranted.

Author Contributions: Conceptualization, B.J., H.C.; Methodology, B.J., P.V., H.C.; Software, B.J.; Formal analysis, B.J.; Writing—original draft, B.J.; Writing—review and editing, P.V., H.C.; Supervision, P.V., H.C.; Funding acquisition, P.V., H.C.; Resources, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by PA CURES (SAP # 4100088552) and the National Institute of Environmental Health Sciences (NIEHS) under grant 1R01ES037163-01.

Institutional Review Board Statement: Not applicable. The primary health dataset used in this study was obtained from the Pennsylvania Health Care Cost Containment Council (PHC4), which provides outpatient encounter records from healthcare facilities across the state. Detailed information could be found in Section 2.

Informed Consent Statement: Not applicable. Patient consent was waived because this research was a retrospective study.

Data Availability Statement: The data used in this study are not publicly available due to data use agreements but are available from the corresponding author on reasonable request.

Acknowledgments: The authors thank Masoud Yari for academic support during the master's program. Additional appreciation is extended to Xiang Gao for technical assistance during the development of the project. The authors also thank Nicole DiRado and Shakuntala Jain for administrative and program support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Slama, A.; Kubajek, J.; Śliwczyński, A.; Turzańska-Wieczorek, O.; Woźnica, J.; Zdrolik, M.; Wiśnicki, B.; Gozdowski, D.; Wierzba, W.; Franek, E. Impact of air pollution on hospital admissions with a focus on respiratory diseases: A time-series multi-city analysis. *Environ. Sci. Pollut. Res.* **2019**, *26*, 16998–17009. [CrossRef] [PubMed]
- Neisani Samani, Z.; Karimi, M.; Alesheikh, A. Environmental and Infrastructural Effects on Respiratory Disease Exacerbation: A LBSN and ANN-Based Spatio-Temporal Modelling. *Environ. Monit. Assess.* **2020**, *192*, 641. [CrossRef] [PubMed]
- Hu, Z.; Rao, K.R. Particulate air pollution and chronic ischemic heart disease in the eastern United States: A county-level ecological study using satellite aerosol data. *Environ. Health* **2009**, *8*, 26. [CrossRef] [PubMed]
- Madani, N.A.; Carpenter, D.O. Patterns of Emergency Room Visits for Respiratory Diseases in New York State in Relation to Air Pollution, Poverty and Smoking. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3267. [CrossRef] [PubMed]
- Alvarez-Mendoza, C.I.; Teodoro, A.C.; Freitas, A.; Fonseca, J. Spatial estimation of chronic respiratory diseases using machine learning and remote sensing data in Quito, Ecuador. *Appl. Geogr.* **2020**, *123*, 102273. [CrossRef]
- Kumarihamy, R.M.K.; Tripathi, N.K. Geostatistical predictive modeling for asthma and COPD using socioeconomic and environmental determinants. *Environ. Monit. Assess.* **2019**, *191*, 366. [CrossRef] [PubMed]
- Ren, H.; Cao, W.; Chen, G.; Yang, J.; Liu, L.; Wan, X.; Yang, G. Lung Cancer Mortality and Topography: A Xuanwei Case Study. *Int. J. Environ. Res. Public Health* **2016**, *13*, 473. [CrossRef] [PubMed]
- Peng, J.; Chen, C.; Zhou, M.; Xie, X.; Zhou, Y.; Luo, C.H. Peak Outpatient and Emergency Department Visit Forecasting for Patients with Chronic Respiratory Diseases Using Machine Learning Methods: Retrospective Cohort Study. *Jmir Med. Inform.* **2020**, *8*, e13075. [CrossRef] [PubMed]
- Ye, X.; Li, Z.; Zhou, X.; Ruan, X.; Lin, T.; Zhou, J.; Yang, D.; Yang, S.; Chen, X.; Wu, K.; et al. The Impact of a Health Forecasting Service on the Visits and Costs in Outpatient and Emergency Departments for COPD Patients: Shanghai Municipality, China, October 2019–April 2020. *China CDC Wkly.* **2021**, *3*, 495–499. [CrossRef] [PubMed]
- Lu, J.; Bu, P.; Xia, X.; Lu, N.; Yao, L.; Jiang, H. Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases. *Environ. Sci. Pollut. Res.* **2021**, *28*, 29701–29709. [CrossRef] [PubMed]
- Ku, Y.; Kwon, S.B.; Yoon, J.H.; Mun, S.K.; Chang, M. Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors. *Environ. Res.* **2020**, *191*, 110184. [CrossRef] [PubMed]
- Zein, J.G.; Wu, C.P.; Attaway, A.H.; Zhang, P.; Nazha, A. Novel Machine Learning Can Predict Acute Asthma Exacerbation. *Chest* **2021**, *159*, 1747–1757. [CrossRef] [PubMed]
- Nikparvar, B.; Rahman, M.M.; Hatami, F.; Thill, J.C. Spatio-temporal prediction of the COVID-19 pandemic in US counties using a deep LSTM neural network. *Sci. Rep.* **2021**, *11*, 21715. [CrossRef] [PubMed]
- Fritz, C.; Dorigatti, E.; Rügamer, D. Combining graph neural networks and spatio-temporal disease models to improve prediction of weekly COVID-19 cases in Germany. *Sci. Rep.* **2022**, *12*, 3930. [CrossRef] [PubMed]
- Ward, T.; Johnsen, A.; Ng, S.; Chollet, F. Forecasting SARS-CoV-2 transmission and clinical risk at small spatial scales by the application of machine learning architectures to syndromic surveillance data. *Nat. Biomed. Eng.* **2022**, *4*, 814–827. [CrossRef]
- Pennsylvania Health Care Cost Containment Council (PHC4) Data. Available online: <https://www.phc4.org/> (accessed on 15 September 2025).

17. Veefkind, J.P.; Aben, I.; McMullan, K.; Förster, H.; De Vries, J.; Otter, G.; Claas, J.; Eskes, H.J.; de Haan, J.F.; Kleipool, Q.; et al. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition. *Remote Sens. Environ.* **2012**, *120*, 70–83. [CrossRef]
18. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
19. European Union/ESA/Copernicus. Sentinel-5P OFFL NO2: Offline Nitrogen Dioxide. Available online: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2 (accessed on 26 June 2026).
20. European Union/ESA/Copernicus. Sentinel-5P OFFL CO: Offline Carbon Monoxide. Available online: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_CO (accessed on 26 June 2026).
21. European Union/ESA/Copernicus. Sentinel-5P OFFL O3: Offline Ozone. Available online: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_O3 (accessed on 26 June 2026).
22. European Union/ESA/Copernicus. Sentinel-5P OFFL SO2: Offline Sulfur Dioxide. Available online: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_SO2 (accessed on 26 June 2026).
23. Abatzoglou, J.T. GRIDMET: Daily Surface Weather Data. 2021. Available online: https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET (accessed on 16 September 2025).
24. U.S. Geological Survey. National Elevation Dataset (NED) 1/3 arc-Second. 2023. Available online: <https://www.usgs.gov/the-national-map-data-delivery> (accessed on 16 September 2025).
25. United States Geological Survey. Landsat 8 OLI/TIRS Collection 2 Level-2 Surface Reflectance. 2023. Available online: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2 (accessed on 16 September 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.