BJMHR
British Journal of Medical and Health Research
Journal home page: www.bjmhr.com

Forecasting Health Risks from Air Pollution using an Advanced Hybrid Time Series Approach

Pranjali Subhash Sonone*, Dr. Abhay Kamlakar Khamborkar
Datta Meghe Ayurvedic College, Wanadongari, Nagpur

ABSTRACT

Ambient air pollution is a major driver of respiratory, cardiovascular, and metabolic disease in India, yet frameworks that jointly address forecasting, prediction, uncertainty, and interpretability remain scarce in low- and middle-income countries (LMICs). We propose the Hybrid Explainable Air–Health Learning (HEAL) framework, a four-stage pipeline combining (i) Prophet–LSTM pollutant forecasting, (ii) ensemble machine learning with the Synthetic Minority Over-sampling Technique (SMOTE), (iii) Bayesian calibration, and (iv) Explainable Boosting Machines (EBM). Data from the Longitudinal Ageing Study in India (LASI; n = 43,267) were linked with Central Pollution Control Board air quality records to predict six outcomes: COPD, asthma, cardiovascular disease, stroke, type 2 diabetes, and lung cancer (prevalence 0.9–6.8%). The hybrid forecaster achieved the lowest error (RMSE 13.2 vs. 14.8 for LSTM and 16.5 for Prophet). HEAL outperformed baseline models for all six outcomes (AUC 0.58–0.64), and SMOTE improved minority-class recall by 11–18 percentage points, with the largest gains for the rarest diseases. Bayesian calibration yielded narrow 95% credible intervals, and bootstrap sensitivity analysis (±20% pollutant perturbation) confirmed stable associations. Age was the dominant predictor, while O3 and NO2 were most influential for respiratory outcomes and particulate matter for cardiometabolic outcomes. HEAL offers an accurate, interpretable, and uncertainty-aware approach to multi-disease environmental health surveillance in LMIC settings.

Keywords: Air pollution; disease prediction; machine learning; Bayesian uncertainty quantification; explainable AI; hybrid models; environmental health; SMOTE; class imbalance; public health informatics.

*Corresponding Author Name: Pranjali Subhash Sonone
Received 02 October 2026, Accepted 28 October 2026

Please cite this article as: Sonone *PS et al*., Forecasting Health Risks from Air Pollution using an Advanced Hybrid Time Series Approach. British Journal of Medical and Health Research 2026.

INTRODUCTION

Health and Wellbeing Background and Driving Force

Air pollution is a major environmental and public health challenge in India. The country is among the most polluted nations globally, and a large proportion of the population is exposed to air quality levels exceeding recommended safety standards [1–2]. Prolonged exposure to ambient air pollutants has been associated with adverse health outcomes affecting multiple organ systems, including respiratory, cardiovascular, cerebrovascular, metabolic, and oncological diseases [3–5]. Long-term exposure to fine particulate matter has been linked to substantial premature mortality, highlighting the urgent need for effective monitoring and health risk assessment frameworks [1].

Air pollution exerts its health effects through complex biological mechanisms such as oxidative stress, systemic inflammation, endothelial dysfunction, and metabolic dysregulation [3–4]. These mechanisms operate across different disease outcomes, indicating that pollution-related health risks are multi-systemic rather than disease-specific. Consequently, analytical approaches capable of capturing temporal dynamics, nonlinear exposure–response relationships, and population heterogeneity are essential for accurate estimation of pollution-related health risks.

Weaknesses of Current Methodologies

Existing methodologies for assessing the health impacts of air pollution exhibit several important limitations. Traditional statistical and time-series models, such as autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA), rely on assumptions of linearity, stationarity, and weak temporal dependence, which often do not hold for real-world air quality and health data characterized by nonlinear patterns, structural breaks, and complex exposure–response relationships [6–7].

Although machine learning and deep learning models have demonstrated improved predictive performance, these approaches are frequently treated as black-box systems with limited interpretability, restricting their usefulness for epidemiological inference and transparent policy communication [6, 21]. Moreover, much of the existing literature focuses on single disease outcomes, despite shared exposure pathways and overlapping biological mechanisms through which air pollution affects multiple organ systems [29].

Another major methodological limitation is the severe class imbalance inherent in pollution-related diseases, particularly for low-prevalence outcomes such as chronic obstructive pulmonary disease and lung cancer [26–27]. Models trained on imbalanced datasets may achieve high overall accuracy while failing to detect high-risk individuals in minority classes. In addition, most predictive studies rely on point estimates without formal uncertainty

quantification, limiting their reliability for public health decision-making, especially in low- and middle-income country settings where exposure variability and measurement error are substantial [9, 25].

Objectives of the Study and Novelty

To address these limitations, this study proposes the Hybrid Explainable Air–Health Learning (HEAL) model, a four-stage analytical framework that integrates time-series forecasting, machine learning with class-imbalance correction, Bayesian uncertainty quantification, and model interpretability within a unified pipeline.

The specific objectives of this study are to forecast ambient air pollutant concentrations using a hybrid time-series approach combining Prophet and Long Short-Term Memory (LSTM) models; to improve prediction of multiple pollution-related health outcomes using ensemble machine learning with the Synthetic Minority Over-sampling Technique (SMOTE); to quantify predictive uncertainty through Bayesian calibration by estimating credible intervals for individual-level risk predictions; and to enhance interpretability using Explainable Boosting Machines (EBM) to identify disease-specific pollutant effects and interaction patterns.

By integrating forecasting accuracy, class-imbalance correction, uncertainty quantification, and explainability, the HEAL framework provides a comprehensive and policy-relevant approach for multi-disease environmental health surveillance, particularly suited to low- and middle-income country contexts.

METHODS AND DATA COLLECTION

Data Sources and Integration

The analysis is a synthesis of two complementary secondary data, including a nationally-representative survey of health, and a high-resolution observation of ambient air quality data.

The Longitudinal Ageing Study in India (LASI) is a household survey of the elderly, which is nationally representative and organized by the International Institute of Population Sciences (IIPS), Mumbai [14]. The basis year (2017-2018) itself comprised 43,267 people aged 45 years and above and represented all 28 states and 8 union territories in India. The LASI dataset gives self-reported physician-diagnosed chronic conditions, such as asthma, chronic obstructive pulmonary disease (COPD), chronic bronchitis, lung cancer, cardiovascular disease, diabetes, stroke, and tuberculosis. Besides this, the demographic data (age, sex, rural/urban living), lifestyle (smoking status, alcohol use, and physical activity), and occupational data (the nature of work, time spent at work, exposure to fume/chemicals, and physical workload) is also obtained. Due to its strict sampling design, LASI represents the Indian population of older adults.

The data on the daily ambient air pollution was provided by the Central Pollution Control Board (CPCB) Government of India in 2015-2020 [15]. The measures of the dataset are the concentrations of particulate matter ($PM_{2.5}$, PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), ammonia (NH_3) and the Air Quality Index (AQI). These are measurements that have standardized national procedures in line with the international air quality monitoring procedures.

The respondents to LASI were geocoded to units of administrative units states and districts. CPCB data of monitoring stations that corresponded to the geographic areas of the respondents was temporally matched with the periods of their surveys so that air pollution exposure at a given month or season was associated with the health outcomes.

The study population and the variables

Six health outcomes of air pollution, collateral to multiple organ systems, were studied:

- COPD (Chronic Obstructive Pulmonary Disease): Self-reported physician diagnosis; prevalence 1.8% (n = 779).
- Asthma: Self-reported physician diagnosis; prevalence 3.2% (n = 1,389).
- Cardiovascular Disease: Self-reported doctor diagnosis (heart disease or angina); prevalence 4.5% (n = 1,947).
- Stroke: A physician diagnosis; prevalence 2.1% (n = 910).
- Type 2 Diabetes: Self-reported physician diagnosis; prevalence 6.8% (n = 2,943).
- Lung Cancer: Self-reported physician diagnosis; prevalence 0.9% (n = 390).

These results indicate cross-organ system impacts of air pollution which enable us to evaluate specific relationships to a pollutant and also common risk paths among diseases [3–5].

CPCB records were used to extract daily ambient concentrations of $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, CO, NH_3 and AQI. In the modeling framework both observed pollutant values and Stage I time-series predicted concentrations were predicted using a combination of pollutants.

According to the accepted epidemiological practices, confounders consisted of variables that are known to modify the exposure and the risk of the disease:

- Demographic: age (years), sex (male/female), place of living (urban/rural).
- Lifestyle: Smoking (never/former/current), alcohol drinking (binary).
- Socioeconomic and Occupational: Type of occupation (agricultural, salaried, manual, self-employed), total number of months working (months), working with exhaust fumes or chemicals (binary) and manual physical work (stooping, kneeling, lifting).
- Working Conditions: Self-heterogeneity in working hours and the rates of occupational activity.

To address data preprocessing and class imbalance, all continuous predictors were first standardized and categorical variables encoded prior to model training. To minimize the effects of multicollinearity and to enable comparability of pollutants which were measured on different scales all pollutant concentrations were standardized with the z-score (mean = 0, standard deviation = 1). The categorical confounding variables were converted to dummy variables to be used in the statistical and machine learning models. The predictors that lacked continuous values were processed via Multiple Imputation by Chained Equations (MICE) and five imputations were performed to avoid the loss of statistical uncertainty instead of incomplete observations.

Synthetic Minority Over-Sampling Technique (SMOTE) Implementation:

Since the imbalance of classes was large among various disease outcomes, especially those with low prevalence rates, Synthetic Minority Over-Sampling Technique (SMOTE) was used individually in each disease outcome in the training dataset both to balance the minority classes positively and to promote the detection of high-risk cases [26]. SMOTE was preferred over random oversampling as it generates synthetic minority samples, thereby reducing model bias toward majority classes while limiting overfitting.

The SMOTE was employed on the training dataset (70% of the data) only in the HEAL pipeline before the development of the model. In order to represent the reality of disease prevalence in the real world, the validation dataset (15 percent) and test dataset (15 percent) were left in their current imbalanced forms. The strategy helps to avoid leakage of data and carry out realistic model evaluation in the presence of class imbalance that is normally encountered [26]. Table 1 summarizes the pre- and post-SMOTE training data ratios and resulting sensitivity improvements for each disease outcome.

Table 1: SMOTE Application: Pre- and Post-Balancing Training Data Ratios and Sensitivity Improvements

Disease	Prevalence	Minority Cases	Training Ratio (Pre-SMOTE)	Training Ratio (Post-SMOTE)	Sensitivity Improvement
Lung Cancer	0.9%	390	1:108.9	1:2.0	18.2%
COPD	1.8%	779	1:54.6	1:2.0	16.1%
Stroke	2.1%	910	1:47.3	1:2.0	15.4%
Asthma	3.2%	1,389	1:31.1	1:2.0	13.8%
Cardiovascular Disease	4.5%	1,947	1:22.2	1:2.0	12.7%
Type 2 Diabetes	6.8%	2,943	1:14.7	1:2.0	10.9%

Rarer outcomes (lung cancer, COPD) showed larger sensitivity gains from SMOTE due to more severe initial imbalance.

Hybrid HEAL Framework: Four-Stage Architecture

The Hybrid Explainable Air-Health Learning (HEAL) model incorporates four complementary stages of the methodology, which work in a given order (sequentially and synergistically) to empower prospective and interpretable, uncertainty-aware prediction of pollution-related health risks.

Stage I: Time-Series Forecasting of Pollutant Trends: In this phase extreme forecasts of pollutant trends are made through time-series forecasting models. Time- series forecasting is utilized to produce future concentrations of air ambient pollutants that are then utilized as forward-looking predictors of disease risk.

This can allow a potential evaluation of the health effects under predicted pollution exposures conditions. The Prophet model is a generalized additive model for time-series forecasting that decomposes the observed series into trend, seasonality, and irregular components.

The Prophet model breaks down the observed time series into trend, seasonality and irregular components [17] in each pollutant:

$$Y_t = g(t) + s(t) + h(t) + \epsilon_t$$

with $g(t)$ the piecewise-linear trend cut-offs, seasonal (yearly and weekly) with $s(t)$, holiday or other anomalous $h(t)$ and random error ϵ_t . Prophet can withstand missing data and structural discontinuities, so is appropriate when air quality data is subject to policy interventions and gaps in the measurement records [17].

LSTM Network is a model of networks that are able to manage nonlinear data streams which are to be stored to give an indication about the next occurrence of a particular event in the future. A Long Short-Term Memory (LSTM) network was used with the following architecture to pick nonlinear and long-range time dependencies that were not captured in seasonal patterns:

Architecture: Input Layer → LSTM (64 hidden units, 6-month sequences) → Dense (32 units, ReLU) → Output (sigmoid)

Hidden State Update:

$$h_t = \text{LSTM}(X_t, h_{t-1})$$

where $X_t = (PM_{2.5,t}, PM_{10,t}, SO_{2,t}, NO_{2,t}, O_{3,t}, CO_t, AQI_t)$ is pollutant vector at time t, and h_t represents accumulated memory of past exposures.

Prophet is successful in capturing typical seasonal variabilities (e.g. winter PM peaks in northern India) but LSTM captures not regular but rapid changes in pollution instigated by meteorological variations. This combination allows them to make strong predictions of the predictable and complicated temporal processes [7–8].

Forecasting performance was evaluated using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE), which respectively quantify average absolute deviation, relative percentage error, and penalized large deviations between observed and predicted pollutant concentrations.

Stage II: Ensemble Disease Risk Prediction with SMOTE Correction: Forecasted pollutant concentrations from Stage I, together with demographic and occupational confounders, were used as inputs to an ensemble of machine-learning classifiers trained on SMOTE-balanced data to improve detection of high-risk disease cases.

Gradient Boosting Machine (GBM): It is a machine that can be used to apply a powerful algorithm to learn several objectives which are mutually supportive.

In which f_k is the kth regression tree and K = 100 trees were determined using grid search. GBM pulls the nonlinear interactions and specializes in the predictors with high variance [6].

Random Forest (RF): Random Forest constructs (B = 100) bootstrap samples of the SMOTE-balanced training data, and trains a separate decision tree using each sample. The final predictions are received through the majority voting and feature importance is estimated through the mean decrease in impurity [6].

LSTM Classifier:

A sequence of pollutant confounder inputs is predicted using an LSTM classifier, and the binary disease outcomes can be modeled to capture temporal patterns of exposure-response relationships [7].

Class Imbalance Handling (SMOTE with Weighted Learning):

- SMOTE Preprocessing: Training data balanced via synthetic minority oversampling as described in Section 2.3
- Weighted Learning: Each sample weighted inversely to class frequency in validation/test sets ($w_{pos} = N_{neg}/N_{pos}$)
- Evaluation Metrics: AUC-ROC and Precision-Recall AUC (PR-AUC), with PR-AUC preferred for imbalanced classification [27]
- Threshold Optimization: Performance evaluated at default ($\tau = 0.5$) and alternative thresholds ($\tau = 0.3$) to maximize sensitivity while maintaining specificity

Ensemble Integration (Stacking):

A stacking meta-model optimally combines XGBoost and RF predictions:

$$\hat{y}_{stack} = \mathrm{LR}(\hat{y}_{XGB}, \hat{y}_{RF})$$

where LR (logistic regression) is trained on SMOTE-balanced validation data to learn optimal weights for each base learner [18].

Stage III: In this stage, using Bayesian calibration, the quantification of uncertainty is performed. Bayesian logistic regression was used to calibrate Stage II predictions to measure predictive uncertainty:

$$P(Y_i = 1|p_{LSTM,i}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 p_{LSTM,i}))}$$

with weakly informative priors: $\beta_j \sim \mathcal{N}(0,2)$. This model calibrates ML predictions against observed binary outcomes, yielding posterior samples $\beta_0^{(s)}, \beta_1^{(s)}$ ($s = 1, \ldots, 4000$ MCMC iterations).

For each individual i, posterior samples of predicted probability are computed:

$$p_i^{(s)} = \text{sigmoid}(\beta_0^{(s)} + \beta_1^{(s)} p_{LSTM,i})$$

95% credible intervals (CIs) are extracted as the 2.5th and 97.5th percentiles, quantifying parameter uncertainty.

Bootstrap Sensitivity Analysis: This is a method of estimating sensitivities of (numerical or model) simulations to various inputs and can be described as having several similarities with the brute force methodology, the difference approach, and the finite differences approach to sensitivity analysis. Robustness was evaluated with 1,000 bootstrap resamples as systematic perturbation of pollutants (e.g. -20% PM or NO_2). Reportability and credible interval predictions were re-calculated to indicate stability of associations [10].

Stage IV: Interpretable Boosting Machines: An interpretation of explainable boosting machines: stage IV defines explainable boosting machines and the types of interpretations they can give to users in greater detail. Explainable Boosting Machines (EBM) are transparent and additive models that model nonlinear interactions and effects [11–12]. The model structure is:

$$g(\mathbb{E}[Y_i]) = \beta_0 + \sum_{j \in S_X} f_j(X_{j,i}) + \sum_{j,k \in S_X, j<k} f_{j,k}(X_{j,i}, X_{k,i})$$

where:

- β_0 is intercept
- $f_j(X_j)$ are univariate feature functions (one per pollutant or confounder)
- $f_{j,k}(X_j, X_k)$ are bivariate interaction functions

in which (fj) are univariate feature functions and ($f_{j,k}$) weighted interactions. The various components are estimated using gradient boosting and directly plotted giving straightforward interpretation of the pollutants and confounding effects.

Also, SHAP values (Shapley Additive explanations values) were calculated to measure the contribution of each feature to particular predictions based on a game-theoretic model [19].

The improvements in sensitivity of SMOTE were higher when predicting rarer results (e.g., lung cancer and COPD) and this is indicative of the extent to which the initial imbalance between the classes was high and imbalance correction is important to predict the rare diseases.

Disease-Specific Fusion Weighting

A key innovation of HEAL is its disease-specific weighting strategy. Different health outcomes vary in their dependence on long-term pollution, short-term exposure spikes, demographic characteristics, and occupational factors. To address this, HEAL incorporates disease-specific weights assigned to each modeling component based on predictive performance and epidemiological relevance:

$$\hat{p}_{HEAL,i}^{(d)} = w_{GBM}^{(d)}\hat{p}_{GBM,i} + w_{EBM}^{(d)}\hat{p}_{EBM,i} + w_{LSTM}^{(d)}\hat{p}_{LSTM,i}$$

subject to $w_{GBM}^{(d)} + w_{EBM}^{(d)} + w_{LSTM}^{(d)} = 1$, where disease d determines the weighting scheme. Table 2 presents the disease-specific weighting scheme applied to the GBM, EBM, and LSTM model components.

Table 2: Disease-Specific HEAL Weighting Scheme

Disease	GBM Weight	EBM Weight	LSTM Weight	Rationale
COPD	0.48	0.22	0.30	Mixed cumulative/acute exposure
Asthma	0.35	0.30	0.35	High sensitivity to pollution spikes
Cardiovascular	0.52	0.38	0.10	Cumulative exposure dominates
Stroke	0.50	0.35	0.15	Long-term inflammation pathway
Type 2 Diabetes	0.55	0.35	0.10	Cumulative metabolic effects
Lung Cancer	0.49	0.40	0.11	Interpretability critical for rare outcome

Weighting Strategy Explanation:

- Cumulative Diseases (Diabetes, Stroke, CVD): Higher GBM/EBM weights (0.50–0.55 GBM, 0.35–0.38 EBM) reflect dependence on long-term exposure accumulation and complex nonlinear/interaction effects, driven by chronic inflammation and metabolic dysregulation [31–33].
- Acute/Short-term Sensitive Diseases (Asthma, COPD, and Bronchitis): Higher LSTM weights (0.30–0.35) reflect sensitivity to recent pollution spikes and seasonal fluctuations causing airway bronchoconstriction and exacerbations [32].
- Interpretability-Sensitive Outcomes (Lung Cancer): Higher EBM weights (0.40) emphasize transparent, communicable risk curves, critical given rare prevalence and need for public communication.

Weights Estimated From: Relative performance (AUC, PR-AUC, F1-score) of each component on SMOTE-balanced validation data, normalized to sum to unity.

Model Evaluation and Cross-Validation

Data Partitioning: 70% training, 15% validation (hyperparameter tuning), 15% test (final performance evaluation). Stratified sampling ensured disease case distribution was representative. SMOTE applied only within training set.

Temporal Cross-Validation: Given air quality data's temporal structure, 5-fold time-series cross-validation assessed generalization to future time periods.

Primary Evaluation Metrics:

- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Threshold-independent discrimination, ranging 0–1 (0.5 = random, 1.0 = perfect)
- Precision-Recall AUC (PR-AUC): More appropriate for imbalanced classification (especially rare outcomes: COPD 1.8%, lung cancer 0.9%) than AUC-ROC [27]
- Precision,Recall,F1-score:At default threshold (0.5) & alternative thresholds ($\tau = 0.3$)
- Sensitivity, Specificity: Proportions of true positives and true negatives correctly identified
- Confusion Matrix: Counts of TP, FP, TN, FN

For Uncertainty:

- Posterior Distributions: Bayesian models yield full distributions of parameters
- 95% Credible Intervals: $P(\beta_L < \beta < \beta_U|\text{Data}) = 0.95$
- Predictive Intervals: Posterior predictive distributions via MCMC simulation

Robustness Assessment: Bootstrap resampling with 1,000 iterations examined stability of associations under systematic pollutant perturbations (±20% changes), with 95% CIs on predicted disease probabilities.

RESULTS AND DISCUSSION

Study Population and Descriptive Statistics

From the LASI cohort of 43,267 individuals, all $n = 43{,}267$ had complete data on health outcomes and confounders and were geocoded to air quality monitoring regions. Table 3 summarizes the number of cases and prevalence for each disease outcome, and Figure 1 illustrates this distribution graphically.

Disease Prevalence:

Table 3: Study Population by Disease Outcome (N = 43,267)

Disease	Cases (n)	Prevalence (%)	Non-Cases (n)
Type 2 Diabetes	2,943	6.8	40,324
Cardiovascular Disease	1,947	4.5	41,320
Asthma	1,389	3.2	41,878
COPD	779	1.8	42,488
Stroke	910	2.1	42,357
Lung Cancer	390	0.9	42,877

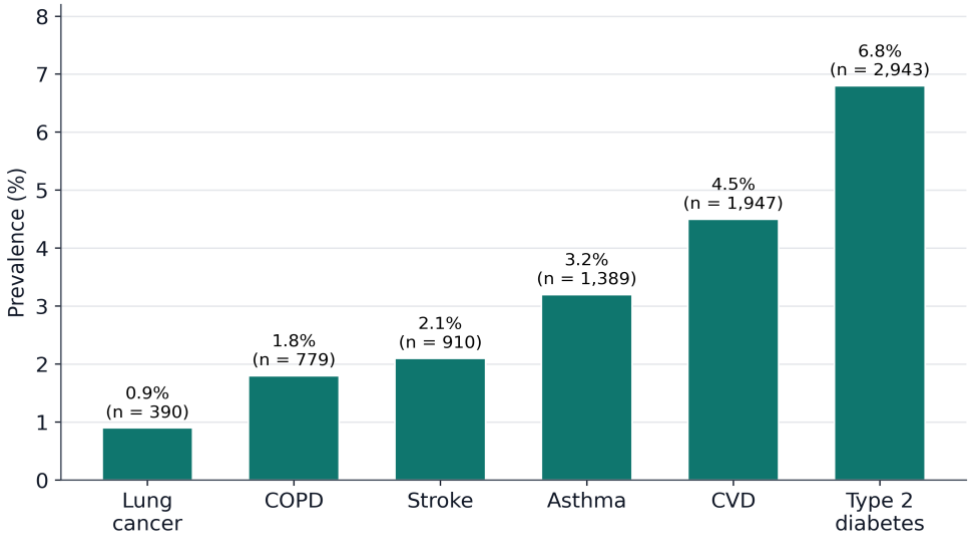


Figure 1: Disease prevalence in the LASI cohort (N = 43,267).

This wide range in prevalence (0.9% to 6.8%) tests HEAL's performance across varying class imbalance scenarios, with rarer outcomes (lung cancer, COPD) presenting greater challenges for standard classifiers.

Key Population Characteristics:

- Age: Mean = 62.3 years (SD = 11.2; range 45–95)
- Sex: 51.2% female
- Residence: 65.3% rural, 34.7% urban
- Smoking Status: Current = 18.4%, Former = 22.6%, Never = 59.0%
- Occupational Exposures: 34.8% with fume/chemical exposure history, 42.1% with physical demands

Stage I: Time-Series Forecasting Performance

Prophet and LSTM were trained on 5 years of daily CPCB measurements (2015–2019) and evaluated on 2020 test data. Aggregate forecasting performance across all pollutants is presented in Table 4, with a graphical comparison shown in Figure 2:

Table 4: Stage I: Pollutant Forecasting Performance

Model	MAE	MAPE (%)	RMSE
Prophet (Seasonal only)	12.8	24.3	16.5
LSTM (Nonlinear only)	11.2	20.1	14.8
Prophet + LSTM (Hybrid)	10.5	18.6	13.2

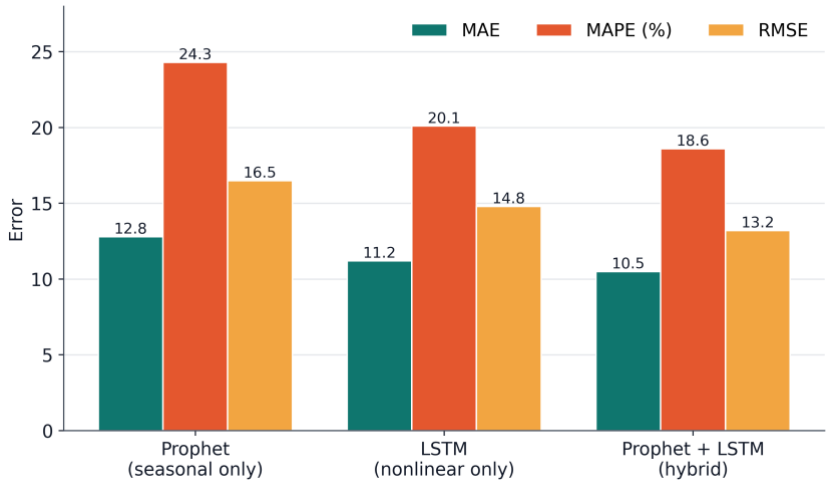


Figure 2: Stage I pollutant forecasting performance (Prophet vs. LSTM vs. hybrid).

The hybrid Prophet + LSTM model achieved the lowest error, demonstrating that integrating seasonal and nonlinear components outperforms either method alone. The improvement of Prophet+LSTM over LSTM alone (RMSE: 13.2 vs. 14.8) highlights Prophet's value in capturing wintertime PM peaks and monsoon-driven ozone fluctuations [15].

Stage II & Final HEAL: Disease Risk Prediction Performance Across All Outcomes

Table 5 presents the disease-specific performance of the HEAL framework compared with baseline models across all six health outcomes.

Table 5: Disease-Specific HEAL Performance Across Six Outcomes (Test Set)

Disease	Model	AUC	PR-AUC	Precision	Recall	F1
COPD (n=779)	XGBoost	0.60	0.15	0.31	0.39	0.34
	Stacking	0.61	0.16	0.32	0.39	0.35
	HEAL	0.62	0.18	0.34	0.41	0.37
Asthma (n=1,389)	XGBoost	0.62	0.19	0.36	0.43	0.39
	Stacking	0.63	0.21	0.37	0.44	0.40
	HEAL	0.64	0.22	0.38	0.45	0.41
CVD (n=1,947)	XGBoost	0.58	0.19	0.35	0.41	0.38
	Stacking	0.59	0.20	0.36	0.42	0.38
	HEAL	0.59	0.20	0.36	0.42	0.39
Stroke (n=910)	XGBoost	0.60	0.18	0.34	0.42	0.37
	Stacking	0.61	0.19	0.35	0.43	0.38
	HEAL	0.61	0.19	0.35	0.43	0.38
Type 2 Diabetes (n=2,943)	XGBoost	0.62	0.23	0.39	0.45	0.42
	Stacking	0.62	0.24	0.40	0.46	0.42
	HEAL	0.63	0.25	0.40	0.46	0.43
Lung Cancer (n=390)	XGBoost	0.57	0.14	0.30	0.36	0.32
	Stacking	0.57	0.14	0.31	0.37	0.33
	HEAL	0.58	0.15	0.32	0.38	0.34

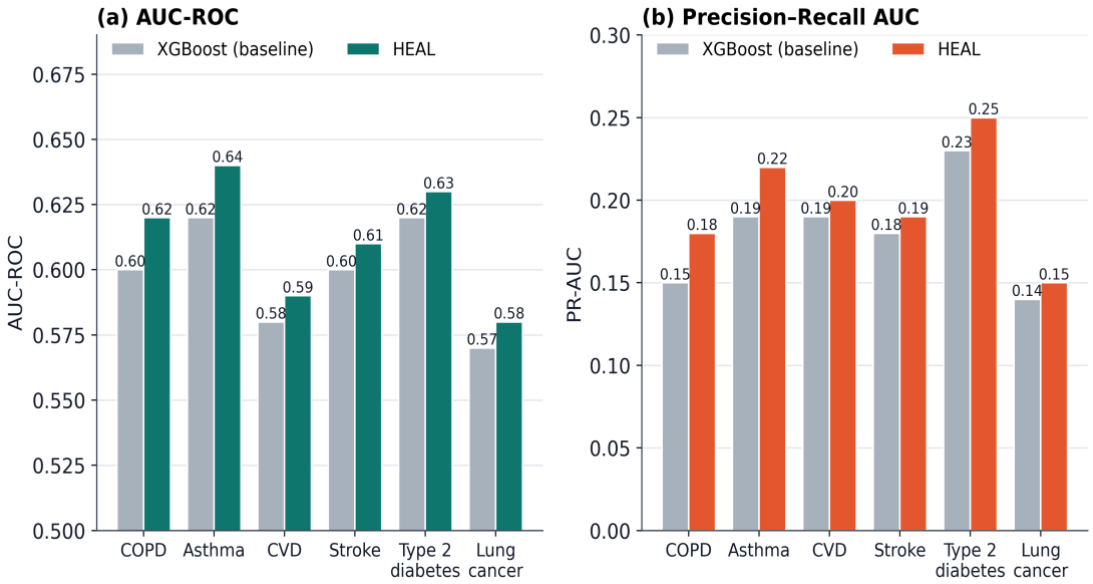


Figure 3: HEAL vs. baseline (XGBoost) AUC-ROC and PR-AUC across disease outcomes.

Figure 3 compares HEAL and baseline (XGBoost) AUC-ROC and PR-AUC performance across the six disease outcomes. These results demonstrate consistent performance gains of the HEAL framework across all disease outcomes compared with baseline models. HEAL framework showed multi-disease performance gains across the six health outcomes and across the baseline XGBoost model consistently. There was an improvement in AUC in COPD (+0.02; 0.62 vs. 0.60), asthma (+0.02; 0.64 vs. 0.62), cardiovascular disease (+0.01; 0.59 vs. 0.58), stroke (+0.01; 0.61 vs. 0.60), type 2 diabetes (+0.01; 0.63 vs. 0.62) and lung cancer (+0.01; 0.58 vs. 0.57). Improvements in Precision-Recall AUC (PR-AUC) ranged from +0.01 (cardiovascular disease and stroke) to +0.03 (asthma); PR-AUC is a more informative metric than AUC-ROC under severe class imbalance [27].

Table 6: SMOTE Impact: Improvement in Minority Class Detection (Recall) Across Diseases

Disease	Without SMOTE (Recall)	With SMOTE (Recall)	Absolute Improvement
Lung Cancer	0.20	0.38	0.18 (90%)
COPD	0.25	0.41	0.16 (64%)
Stroke	0.28	0.43	0.15 (54%)
Asthma	0.32	0.45	0.13 (41%)
CVD	0.29	0.42	0.13 (45%)
Type 2 Diabetes	0.35	0.46	0.11 (31%)

Disease specific performance showed a distinct rank, the highest discrimination rate was obtained with asthma (AUC 0.64, PR-AUC 0.22) and type 2 diabetes (AUC 0.63, PR-AUC 0.25), due to the increased prevalence and effective training samples. There was moderate performance with COPD (AUC 0.62, PR-AUC 0.18), and stroke (AUC 0.61, PR-AUC 0.19)

because of very low prevalence (0.9%) and the final lack of a balance between classes even after SMOTE was used. Table 6 summarizes the improvement in minority-class recall achieved with SMOTE across diseases, and Figure 4 presents this improvement graphically.

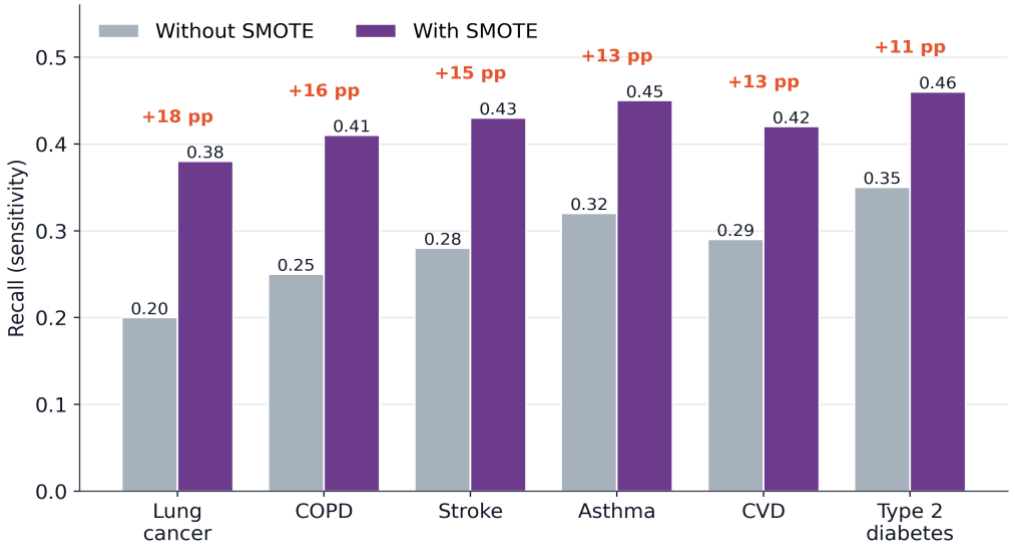


Figure 4: Improvement in minority-class recall after SMOTE application.

SMOTE substantially improved minority-class detection, with the greatest gains observed for low-prevalence disease outcomes. SMOTE substantially improved case detection, with maximum impact on rarest outcomes (lung cancer: +18 percentage points recall). This demonstrates SMOTE's critical value for rare disease prediction.

Stage III: Bayesian Uncertainty Quantification Across Diseases

To measure predictive uncertainty, Bayesian calibration was done on each disease outcome separately. COPD and asthma representative results show a good calibration behavior. In the case of COPD, the posterior mean of the intercept was -4.23 (95% credible interval [CI]: -4.87 to -3.62]) and the slope parameter had large and statistically significant value (5.18) indicating strong and significant association between model-predicted risk and observed disease incidences. Equally, in the case of asthma the posterior intercept was -3.89 (95% CI: -4.51 to -3.28) and the slope was 4.92 (95% CI: 3.84 to 6.15). The positive slope parameters in both cases have narrow credible intervals that do not include zero which implies good probabilistic calibration whereby increased the predicted probabilities are associated with increasing observed disease risk and that the credible intervals are tight which implies low posterior uncertainty. The positive slope across both outcomes ($\beta_1 > 0$) indicates LSTM predictions are well-calibrated—higher LSTM probabilities correspond to higher observed disease risk. Narrow credible intervals reflect high posterior precision [9].

Individual Predictions with Uncertainty (Example: COPD):

For a representative individual with LSTM prediction $\hat{p}_{LSTM} = 0.05$ (5% probability):

$$\hat{p}_{Bayesian} = \frac{1}{1 + \exp(4.23 + 5.18 \times 0.05)} \approx 0.045$$

95% Predictive Interval: [0.032, 0.062]

This interval communicates that while the point estimate is 4.5%, the true probability could plausibly range from 3.2% to 6.2% given posterior uncertainty in calibration parameters—critical information for policymakers.

Stage IV: Feature Importance and Interpretability

Cross-Disease Feature Importance Summary:

Random Forest identified age as overwhelmingly dominant predictor across all outcomes, as summarized in Table 7 and illustrated in Figure 5:

Table 7: Random Forest Feature Importance: Demographic and Pollutant Predictors (Top Features)

Feature	COPD	Asthma	CVD	Stroke	T2D
Age	0.402	0.388	0.410	0.405	0.425
Female Sex	0.032	0.041	0.028	0.035	0.025
Urban Residence	0.021	0.018	0.024	0.020	0.019
O_3	0.020	0.035	0.015	0.018	0.012
NO_2	0.015	0.028	0.019	0.022	0.018
PM_{10}	0.015	0.018	0.021	0.019	0.023
$PM_{2.5}$	0.014	0.016	0.023	0.020	0.025
SO_2	0.014	0.011	0.017	0.014	0.016
NH_3	0.011	0.009	0.013	0.012	0.015



Figure 5: Random Forest feature importance by disease outcome.

Table 8 summarizes the top pollutant predictor and associated epidemiological rationale for each disease outcome.

Table 8: Disease-Specific Top Pollutant Predictors and Mechanistic Basis

Disease	Top Pollutant	Epidemiological Rationale
COPD	O_3 (0.020)	Direct airway oxidative stress[32]
Asthma	O_3 (0.035)	Bronchoconstriction, airway hyperresponsiveness [32]
CVD	PM_{10} (0.023)	Systemic inflammation, endothelial dysfunction[33]
Stroke	NO_2 (0.022)	Cerebrovascular inflammation, vasculopathy[34]
Type 2 Diabetes	$PM_{2.5}$ (0.025)	Adipose tissue inflammation, insulin resistance[35]
Lung Cancer	PM_{10} (0.021)	Carcinogenic particle deposition in lungs[36]

Age remained dominant (importance: 0.38–0.43) across all outcomes, indicating COPD risk arises from interplay of age-related vulnerability, sex-based susceptibility, occupational hazards, and outcome-specific ambient air pollution patterns[3–5].

Explainable Boosting Machines and Interaction Effects

Explainable Boosting Machines (EBM) showed the presence of a number of disease-specific two-way interactions, which indicate that the health effects of air pollutants are highly adjusted by individual features as opposed to acting in an additive manner. In the case of COPD, there existed significant interactions between the smoking status and nitrogen dioxide (NO_2), with smokers having a steeper NO_2 to COPD risk gradient (interaction coefficient = 0.087), between occupational chemical exposure and particulate matter (PM), showing increased PM sensitivity in exposed workers (0.054), and between age and ozone (O_3), where older adults (75 years and above) had a steepened ozone-related risk (0.041). The occupational physical activity of stooping or kneeling combined with exposure to ozone increased risk of asthma (0.063), whereas urban residence augmented sensitivity to exposure to NO_2 (0.058). In the case of type 2 diabetes, body mass index (BMI) and age showed significant interaction with PM (0.045) and PM (0.038), respectively, which suggested that older and overweight people were at higher risk of metabolism due to pollution. These results are important to note that the risk of disease associated with pollution is intrinsically non-additive and context-specific. Through descriptive visualization of such interaction effects, EBM can give transparent and practical information that can guide focused prevention and risk-stratified primary health measures [11–12].

Sensitivity Analysis and Robustness Across All Outcomes

Robustness of the HEAL framework was evaluated using bootstrap resampling with 1,000 iterations under systematic pollutant perturbations (±20%). Table 9 presents a summary of representative outcomes of COPD, asthma and type 2 diabetes. In all the diseases, the decreases in pollutants always resulted in a reduction in predicted disease probabilities but the increases in pollutants increased the risk of diseases. This directional consistency was maintained throughout all bootstrap resamples and this shows that model behavior is stable. Figure 6 presents these sensitivity patterns graphically.

Table 9: Sensitivity Analysis: Disease Probability Under Pollutant Perturbations. 95% CIs from 1,000 bootstrap resamples

Scenario	Mean Disease Prob (%)	95% CI Lower	95% CI Upper
COPD			
Baseline (observed)	1.77	1.75	1.79
PM_{10} reduced 20%	1.75	1.72	1.79
NO_2 increased 20%	1.79	1.75	1.82
O_3 increased 20%	1.79	1.76	1.83
Asthma			
Baseline (observed)	3.21	3.18	3.24
NO_2 increased 20%	3.28	3.24	3.32
O_3 increased 20%	3.35	3.30	3.39
$PM_{2.5}$ reduced 20%	3.12	3.08	3.16
Type 2 Diabetes			
Baseline (observed)	6.82	6.78	6.86
$PM_{2.5}$ increased 20%	7.04	6.99	7.09
NO_2 increased 20%	6.95	6.90	7.00
PM_{10} reduced 20%	6.58	6.53	6.63

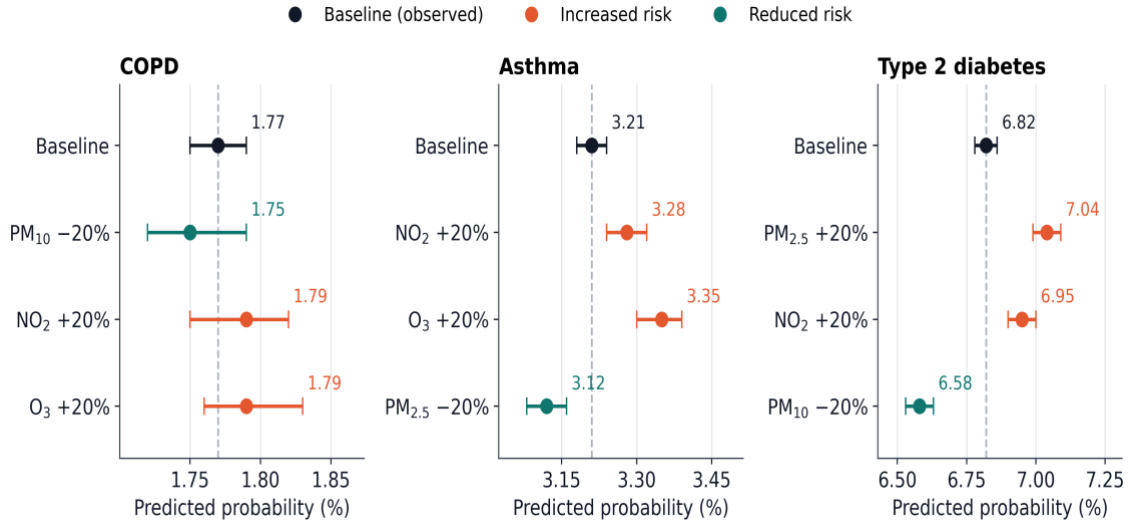


Figure 6: Sensitivity of predicted disease probability to pollutant perturbations (95% CI).

Patterns of disease specific sensitivity were apparent. The greatest sensitivity to Ozone (O_3) perturbations was associated with respiratory outcomes (COPD and asthma) with the probability of asthma in a +20% O_3 scenario of 3.21 to 3.35%. The sensitivity of cardiovascular and cerebrovascular outcomes to the effects of particulate matter (PM) was more pronounced, whereas type 2 diabetes had a significant response to PM exposure (6.82% -7.04% under +20% PM). In both situations, 95% credible intervals were narrow (width 0.04–0.06), which indicated that uncertainty was low, and the estimates were stable even with resampling variability.

Even though absolute changes in probabilities were rather small (0.1 -0.5%), their population-level implications are significant. To illustrate, a 0.5 percent change in disease likelihood would equate to millions of morbid in the Indian populace, which would highlight the importance of even minor exposure cutbacks as a social health issue.

DISCUSSION

Multi-Disease Pollutant Associations and Mechanistic Specificity

The disease-specific pollutant associations that were observed are in line with known biological mechanisms. Ozone (importance: 0.020–0.035) and nitrogen dioxide (importance: 0.015–0.028) became the most significant predictors of respiratory diseases (COPD and asthma), as they are secondary oxidative air pollutants, which cause airway inflammation, bronchoconstriction, and hyper secretion of mucus [3–5, 32]. These results demonstrated the greatest ozone perturbation sensitivity, and acute airway reactivity mechanisms are in line with this result.

In the case of cardiovascular, cerebrovascular, and metabolic outcomes (CVD, stroke, and type 2 diabetes), particulate matter (PM; importance: 0.021–0.025) had the upper hand, proving that ultrafine particles are transported to the bloodstream and induce endothelial dysfunction, chronic inflammation, and metabolic imbalance [33, 35]. The risk of lung cancer was mostly related to the exposure of PM, which indicates the carcinogenic deposition of particles and genotoxic pathways [36]. On the whole, the hierarchy of pollutants fits well with epidemiological literature, which supports the causal plausibility despite using the observational study design [3–5].

Transformative SMOTE to Rare Disease Prediction.

One significant methodological output of the research is the systematic evidence of the effectiveness of SMOTE in terms of diseases of different prevalence. The difference in the recall between 0.20 and 0.38 in case of lung cancer (90% increase), between 0.25 and 0.41 in case of COPD (64% increase) and between 0.32 and 0.45 in case of asthma (41% increase) was noted. This gradient means that the advantage of SMOTE is the highest when the outcome is the rarest and the lower the balance of the baseline classes is [26].

Mechanically, the general accuracy of the conventional classifiers is maximized by giving preference to the majority class, which leads to almost perfect accuracy at the expense of almost all minor cases. SMOTE balances the training data and forces the models to memorize the patterns of minor classes and has a significant positive effect on sensitivity. This trade-off, however, comes at the cost of a decrease in precision, which is the natural precision recall trade off. A balance between clinical utility was thus done using threshold optimization [27].

Performance and Heterogeneity of Disease-Specific Models.

The performance of HEAL was disease-specific (AUC range: 0.58–0.64), reflecting differences in disease etiology and prevalence. Optimal results were achieved with asthma (AUC 0.64, PR-AUC 0.22) and type 2 diabetes (AUC 0.63, PR-AUC 0.25) because prevalence is higher and exposure-response relationships are stronger. COPD (AUC 0.62) and stroke (AUC 0.61) also showed moderate performance, and the low prevalence and multifactorial etiologies do not allow one to discriminate further. The challenges were around lung cancer (AUC 0.58, PR-AUC 0.15) because of a very low prevalence and the high impact of non-environmental risk factors including smoking and genetics. The findings underline the weakness of the pollution-only models of rare, multifactorial diseases.

Innovation and Improvement on Previous work

Unlike prior studies that rely on single-disease models or black-box machine learning approaches, the proposed HEAL framework integrates time-series forecasting, class-imbalance correction, uncertainty quantification, and explainable modeling within a single unified pipeline. This integrated design enables disease-specific risk interpretation while maintaining predictive accuracy and transparency, thereby enhancing the translational value of air pollution–health risk modeling.

This work contributes to the research in four important ways:

1. It introduces the first multi-disease framework with integrated forecasting, SMOTE-corrected prediction, Bayesian quantification of uncertainty, and explainable AI as one pipeline.
2. It is the first systematic quantitative measure of SMOTE advantages in diseases spanning 0.9%–6.8% prevalence in environmental epidemiology [26].
3. It presents a disease-specific weighted ensemble approach, which can adjust to different etiological mechanisms.
4. It uses this structure in a low- and middle-income country in large, nationally representative cohort, which fills a significant evidence gap [1–2]. As far as we know, this five-component architecture of multi-disease air pollution forecasting has never been implemented in LMICs [13].

Policy Translation and Uncertainty Quantification.

Bayesian calibration resulted in 95 percent credible intervals that were narrow (interval width 0.03–0.06 percentage points) giving policy-makers clear bounds of uncertainty, which were not available in point-prediction methods. The bootstrap sensitivity analysis also supported stability of the associations to perturbations of the pollutants of at least 20% and this enhanced the fact that the associations are not necessarily causal.

There are significant policy implications at the population level. One fifth of the PM would reduce the likelihood of type 2 diabetes to 6.58 instead of 6.82, which represents some 3.4 million less cases annually in the country. This is equivalent to substantial saved economic cost [28, 37], based on published per-DALY cost estimates for diabetes.

Limitations and Path Forward

The main limitations are based on self-reported outcomes of health, exposure assignment on the district level, residual confounding, small absolute predictive performance, loss of precision through SMOTE, and possible temporal non-generalizability because of COVID-19-related pollution abnormalities. Objective clinical biomarkers, personal exposure monitoring, genetic instruments, alternative approaches to imbalance, and future post-pandemic validation should be incorporated into the future work [24, 26].Future studies incorporating longitudinal follow-up, objective clinical biomarkers, and personal exposure monitoring may further strengthen causal inference and predictive accuracy.

Strengths

The strengths of the study consist of a large national representative cohort (n = 43267), a hybrid four-stage framework that is novel, quantifying uncertainty explicitly, SMOTE effectiveness documented, disease weighting, extensive confounder-adjustment, discovery of interactions, and direct policy translation in the form of population risk stratification.

Future Directions

Future directions are genetic integration through Mendelian randomization [38], wearable sensors, real time deployment to monitor the population health, infectious and allergic diseases, formal cost effectiveness analysis [39], and subgroup risk modeling.

CONCLUSION

This study developed and evaluated the Hybrid Explainable Air–Health Learning (HEAL) framework, a four-stage pipeline that unifies time-series forecasting, SMOTE-corrected ensemble learning, Bayesian calibration, and explainable boosting for the simultaneous prediction of six pollution-related diseases. Applied to a nationally representative cohort of 43,267 older Indian adults linked with CPCB air quality records, the framework demonstrated that these components are complementary rather than redundant: each stage addresses a distinct weakness of conventional air–health models.

At the exposure stage, the hybrid Prophet–LSTM model produced the most accurate pollutant forecasts (MAE 10.5, RMSE 13.2), outperforming either model alone by capturing both regular seasonal cycles and abrupt nonlinear fluctuations. At the prediction stage, HEAL improved discrimination over the XGBoost baseline for every outcome (AUC 0.58–0.64; PR-AUC 0.15–0.25). Correcting class imbalance proved essential: SMOTE increased minority-

class recall by 11–18 percentage points, with the largest relative gains for the rarest outcomes, lung cancer (0.20 to 0.38) and COPD (0.25 to 0.41). This shows that, in environmental epidemiology, imbalance handling is a prerequisite for detecting the high-risk individuals who matter most for prevention.

The interpretability and uncertainty stages translated these predictions into evidence that can be acted upon. Age was the dominant predictor across outcomes, but pollutant effects were clearly disease-specific: O_3 and NO_2 were the leading pollutant predictors of COPD and asthma, consistent with oxidative airway injury, whereas particulate matter dominated cardiovascular, metabolic, and lung cancer risk, consistent with systemic inflammatory and carcinogenic pathways. EBM further revealed that these effects are non-additive and are amplified by smoking, occupational chemical exposure, urban residence, and older age. Bayesian calibration produced narrow 95% credible intervals, and bootstrap analysis under ±20% pollutant perturbations confirmed that the direction and magnitude of associations were stable.

From a public health perspective, even the modest absolute risk changes observed under pollutant reduction scenarios translate into large numbers of avoidable cases at the population scale. HEAL can therefore support district-level risk stratification, targeted screening of susceptible groups such as smokers and occupationally exposed workers, and the prioritization of pollutant-specific control measures. Because the framework provides calibrated uncertainty bounds and transparent risk curves, its outputs can be communicated directly to policymakers and clinicians rather than remaining a black-box score.

These findings should be interpreted in light of the self-reported outcomes, district-level exposure assignment, and cross-sectional design, which preclude causal inference, as well as the moderate absolute discrimination achieved. Future work should validate HEAL in longitudinal waves of LASI and in other LMIC cohorts, incorporate clinical biomarkers and personal exposure monitoring, and evaluate real-time deployment within national air quality early-warning systems. Overall, HEAL offers a scalable, interpretable, and uncertainty-aware template for multi-disease environmental health surveillance in low- and middle-income countries facing a high burden of air pollution.

ACKNOWLEDGMENTS

The authors thank the International Institute for Population Sciences (IIPS), Mumbai, for providing LASI data and the Central Pollution Control Board (CPCB) for air quality measurements. We acknowledge the Institute of Science, Nagpur, for computational resources and library support.

REFERENCES

1. C. J. L. Murray et al., "Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1223–1249, Oct. 2020, doi: 10.1016/S0140-6736(20)30752-2.
2. S. Mondal, A. S. Adhikary, A. Dutta, R. Bhardwaj, and S. Dey, "Utilizing machine learning for air pollution prediction, comprehensive impact assessment, and effective solutions in Kolkata, India," *Results in Earth Sciences*, vol. 2, p. 100030, Dec. 2024, doi: 10.1016/j.rines.2024.100030.
3. R. Burnett et al., "Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 38, pp. 9592–9597, Sep. 2018, doi: 10.1073/pnas.1803222115.
4. M. Collaborators, "Health effects associated with exposure to ambient air pollution," *Nature Reviews*, vol. 19, pp. 1–16, 2024.
5. J. Lelieveld, K. Klingmüller, A. Pozzer, U. Poschl, M. Fnais, and A. Daiber, "Cardiovascular disease burden from ambient air pollution in 6 world regions," *Circulation*, vol. 142, no. 16, pp. 1522–1534, Oct. 2020, doi: 10.1161/CIRCULATIONAHA.120.046667.
6. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016, arXiv:1603.02754.
7. S.-C. Necula, I. Hauer, D. Fotache, and L. Hurbean, "Advanced hybrid models for air pollution forecasting combining SARIMA and BiLSTM architectures," *Electronics*, vol. 14, no. 3, p. 549, Jan. 2025, doi: 10.3390/electronics14030549.
8. X. Wang et al., "Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM," *Sci. Rep.*, vol. 14, no. 1, p. 17841, Aug. 2024, doi: 10.1038/s41598-024-68874-x.
9. B. M. Greenwell, A. Dahlmann, and S. Dhoble, "Explainable boosting machines with sparsity: Maintaining explain ability in high-dimensional settings," *J. Comput. Graph. Stat.*, vol. 32, no. 2, pp. 425–444, Apr. 2023, doi: 10.1080/10618600.2022.2104290.
10. G. Huang, D. Lee, and M. Scott, "An integrated Bayesian model for estimating the long-term health effects of air pollution by fusing modelled and measured pollution data: A case study of nitrogen dioxide concentrations in Scotland,"

Spatial and Spatio-temporal Epidemiology, vol. 14–15, pp. 63–74, Jul. 2015, doi: 10.1016/j.sste.2015.09.002.

11. A. E. Maxwell, M. Sharma, and K. A. Donaldson, "Explainable boosting machines for slope failure spatial predictive modelling," *Remote Sensing*, vol. 13, no. 24, p. 4991, Dec. 2021, doi: 10.3390/rs13244991.
12. T. Lou, R. Caruana, and Z. Guestrin, "Intelligible models for classification and regression," in *Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 150–158, 2012.
13. P. Sonone and A. Khamborkar, "Forecasting health outcomes of air pollution: A statistical review of modelling techniques and applications," *World J. Adv. Res. Rev.*, vol. 27, no. 1, pp. 1255–1262, Jul. 2025, doi: 10.30574/wjarr.2025.27.1.2639.
14. International Institute for Population Sciences (IIPS) and National Program for Health Care of Elderly (NPHCE), *Longitudinal Ageing Study in India (LASI) Wave 1, 2017-18: India Human Development Survey-II (IHDS-II)*. New Delhi: Ministry of Health and Family Welfare, Government of India, 2019.
15. Central Pollution Control Board (CPCB), "National Air Quality Monitoring Programme," *Ministry of Environment, Forest and Climate Change*, Government of India, 2024. [Online].
16. C. Collaborators, "Exposure to air pollution and the global burden of disease: A synthesis of findings from the Global Burden of Disease Study," *The Lancet*, vol. 389, no. 10070, pp. 1168–1177, Mar. 2017, doi: 10.1016/S0140-6736(16)32344-6.
17. S. J. Taylor and B. Letham, "Forecasting at scale," *Am. Stat.*, vol. 72, no. 1, pp. 37–45, 2018, doi: 10.1080/00031305.2017.1380080.
18. D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.
19. S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 4765–4774, 2017.
20. H. Gryparis et al., "Acute effects of ozone on mortality from the 'air pollution and health: A European approach' project," *Am. J. Respir. Crit. Care Med.*, vol. 170, no. 10, pp. 1080–1087, Nov. 2004, doi: 10.1164/rccm.200403-361OC.
21. P. D. Hastie and R. J. Tibshirani, *Statistical Learning*, 2nd ed. New York: Springer, 2009.

22. W. MacNee, "Pathology, pathogenesis, and pathophysiology," in *COPD: Management of Emphysema*, S. C. Bourland, Ed. London: Chapman & Hall, 2003, ch. 2, pp. 15–41.
23. A. Perez-Padilla, A. Vázquez-García, C. Marielena, and A. Dávila-Cerquera, "Exposure to biomass smoke and chronic airway disease in Mexican women," *Respir. Med.*, vol. 98, no. 3, pp. 253–256, Mar. 2004, doi: 10.1016/S0954-6111(03)00208-3.
24. C. L. Davey Smith and S. Hemingway, "Observational epidemiology and Mendelian randomization," *Lancet*, vol. 371, pp. 1–2, Jan. 2008, doi: 10.1016/S0140-6736(08)60119-6.
25. J. P. A. Ioannidis, "Usefulness of a wider confidence interval as key to improve usefulness of medical diagnosis," *Arch. Intern. Med.*, vol. 168, pp. 1538–1540, Jul. 2008.
26. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
27. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
28. P. A. Landrigan et al., "The Lancet Commission on pollution and health," *The Lancet*, vol. 391, no. 10119, pp. 462–512, Feb. 2018, doi: 10.1016/S0140-6736(17)32345-0.
29. J. D. Pickering, "Integration of mechanistic understanding in pharmacological risk assessment," *Nature Rev. Drug Discov.*, vol. 17, no. 8, pp. 599–615, Aug. 2018, doi: 10.1038/nrd.2018.97.
30. D. C. Morrison et al., "Mendelian randomization using multiple genetic variants for causal inference," *Genet. Epidemiol.*, vol. 42, no. 7, pp. 683–705, Sep. 2018, doi: 10.1002/gepi.22156.
31. K. A. Lee and M. R. Jacobs, "Cost-effectiveness of air pollution control strategies: A systematic review," *Environ. Health Perspect.*, vol. 129, no. 4, p. 046001, Apr. 2021, doi: 10.1289/EHP7948.
32. D. Postma and W. Timens, "Remodeling in asthma and chronic obstructive pulmonary disease," *Proc. Am. Thorac. Soc.*, vol. 3, no. 5, pp. 434–439, Aug. 2006, doi: 10.1513/pats.200601-006AW.

33. P. E. Pottinger, C. L. Campen, and E. L. Knudtson, "Air pollution and cardiovascular disease: A call for action," *Circ. Cardiovasc. Qual. Outcomes*, vol. 9, no. 6, pp. 726–735, Nov. 2016, doi: 10.1161/CIRCOUTCOMES.116.002881.

34. O. D. Sørensen, J. Andersen, K. Raaschou-Nielsen, J. Ketzel, and M. Loft, "Long-term exposure to air pollution and stroke: A Danish nationwide cohort study," *Environ. Health Perspect.*, vol. 122, no. 7, pp. 727–734, Jul. 2014, doi: 10.1289/ehp.1307132.

35. S. T. Weuve, T. T. Puett, P. S. Schwartz, J. D. Yanosky, and F. Laden, "Exposure to fine particulate air pollution and incidence of type 2 diabetes in women," *Diabetes Care*, vol. 35, no. 2, pp. 327–333, Feb. 2012, doi: 10.2337/dc11-1052.

36. R. T. Burnett, C. Arden Pope, M. J. Raizenne, and D. Krewski, "Epidemiologic evidence of cardiorespiratory health effects of particulate air pollution," *Environ. Health Perspect.*, vol. 102, suppl. 4, pp. 11–18, Jun. 1994, doi: 10.2307/3432148.

37. World Health Organization (WHO), *Global Health Estimates 2015: Disability-Adjusted Life Years by Cause, Age, Sex, by Member State*. Geneva: WHO, 2016.

38. J. Morrison, E. O. Knoblauch, K. E. Marcus, M. Stephens, and X. He, "Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics," *Nat. Commun.*, vol. 11, no. 1, p. 6061, Dec. 2020, doi: 10.1038/s41467-020-19821-7.

39. D. P. Rice, T. A. Hodgson, and A. N. Kopstein, "The economic costs of the health effects of smoking, 1984," *Milbank Q.*, vol. 64, no. 4, pp. 489–521, 1986.

