https://doi.org/10.1038/s44407-025-00035-4

Innovative integration of wind transformation in AI models for real-time carcinogenic risk assessment

Shivam Singh ✉, Pratibha Vishwakarma & Tarun Gupta ✉

Air pollution remains a major public health concern, with fine particulate matter ($PM_{2.5}$) and its associated toxic pollutants contributing substantially to cancer risk. This study introduces a novel machine learning framework to predict Incremental Lifetime Cancer Risk (ILCR) using routine meteorological and air quality data, offering a cost-effective alternative to direct Polycyclic Aromatic Hydrocarbons (PAHs) measurements. In the present study, two modelling strategies were evaluated. The first is the Pollution Source Method (PSM), which incorporates wind parameters transformed according to local pollution source directions, and the second is the Conventional Method (CM), which uses unprocessed meteorological inputs. Artificial Neural Network (ANN) and Extreme Gradient Boosting (XGBoost) models were applied under both strategies and assessed using R^2, MAE, MSE, RMSE, and MAPE. The PSM-ANN model showed the strongest performance ($R^2 = 0.944$; MAE = 0.037), while the CM-XGB model performed the weakest ($R^2 = 0.799$; MAE = 0.061). Error analyses confirmed that PSM-based models produced more stable predictions with reduced uncertainty. By enabling real-time ILCR prediction from low-cost sensors, this framework can support early public health interventions and risk communication. Future work will expand this approach to diverse regions and explore deep learning techniques to further enhance predictive accuracy.

Air pollution is a critical global issue that poses severe threats to public health, environmental sustainability, and economic stability. Industrialization, urbanization, and vehicular emissions have led to deteriorating air quality, resulting in respiratory diseases, cardiovascular complications, and increased mortality rates. Given the need for effective environmental management and public health interventions, accurate air pollution prediction has become an essential research area. Traditional air quality monitoring stations, while precise, are costly to install and maintain, making them infeasible for widespread implementation. To address these limitations, computational models have been developed to estimate air pollutant concentrations across broader spatial regions[1].

Recent advancements in machine learning (ML) have significantly improved the ability to model complex atmospheric interactions while reducing computational costs. ML models have demonstrated high efficiency in predicting air quality indices (AQI) and various air pollutants, thereby offering real-time predictions at a reduced financial burden. Instead of relying on high-end air quality monitoring equipment, ML models can utilize data from low-cost sensors to predict intricate pollution indicators such as Incremental Lifetime Cancer Risk (ILCR), as explored in this study.

Artificial Neural Networks (ANN) have been widely used in air quality forecasting because of their capacity to capture nonlinear relationships. For example, studies in Liaocheng, Shanghai, and Chongqing have successfully applied ANN and wavelet-based ANN frameworks for $PM_{2.5}$ and PM_{10} prediction[2–5]. These works demonstrate the suitability of ANN in short-term pollution forecasting. More recent studies also extend such models to climate and ozone prediction, employing architectures such as CNNs, LSTMs, and hybrid wavelet-deep learning frameworks[6–9]. Such approaches highlight the growing potential of advanced neural networks to model complex atmospheric processes more effectively than traditional methods. Several prior studies have successfully applied ML models to predict air pollutant concentrations[10–13] and AQI[14,15]. For instance, a study comparing four ML models for AQI prediction using air pollutants and weather parameters achieved RMSE values of 24.14, 15.97, and 18.72 for ANN, XGBoost, and decision tree models, respectively, outperforming traditional multilinear regression (MLR) models[14]. Another study evaluating seven regression and seven classification models found that random forest performed the best, achieving an R^2 value of 0.91 and an MSE of 0.0067[15]. These findings highlight the necessity of ML

Department of Civil Engineering, APTL, Centre for Environmental Science and Engineering (CESE), IIT Kanpur, Kanpur, UP, India.
✉e-mail: shivamnov20@iitk.ac.in; tarun@iitk.ac.in

models in air quality prediction, as traditional regression models struggle to capture complex relationships.

Optimizing ML models by selecting the most relevant input variables enhances their predictive performance. Several studies have explored different methods to minimize the number of input variables while maintaining model accuracy. A study on NO_2 prediction using an ANN model reduced weather parameters to two derived stochastic variables[13], while others have employed techniques such as sensitivity analysis, genetic algorithms, principal component analysis (PCA), and correlation coefficient methods to refine input selection[10,11,16,17]. The primary goal of these studies was to ensure that input variables were highly relevant to the target variable, ultimately improving model efficiency. In contrast, this study aims to maintain input relevance across multiple locations, leading to an increase in the number of input variables.

Machine learning models also hold great potential for reducing the cost of air pollution monitoring. For example, a study in China utilized ML models to predict $PM_{2.5}$ concentrations in metropolitan areas, such as Xinzhuang, Sanchong, and Cailiao, based on data from pollution measurement stations. This approach demonstrated the feasibility of reducing the number of expensive monitoring stations[18]. Similarly, other research efforts have sought to minimize reliance on costly air pollution monitoring devices by developing ML models capable of real-time or hourly pollutant predictions. One study trained an ANN model using meteorological data (temperature, relative humidity, wind speed, and wind direction) to predict hourly pollutant levels, achieving R^2 values of 0.87, 0.87, 0.85, 0.77, and 0.92 for PM_{10}, NO_x, NO_2, O_3, and CO, respectively[12].

Efforts to replace traditional air pollution monitoring stations with virtual ones have gained traction. A study investigating the concentration of $PM_{2.5}$, PM_{10}, and NO_2 used five ML models, which include support vector regressor, ridge regressor, random forest, XGBoost, and extra tree regressor, to predict pollution levels with the objective of reducing reliance on physical monitoring stations[19]. Another study leveraged the Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict same-day and one-day-ahead air quality, reducing computational costs and improving prediction accuracy[16]. To achieve high-resolution temporal and spatial predictions of NO_2, O_3, and HCHO, researchers have employed physically informed neural networks (PINNs), which integrate domain knowledge with ML algorithms[20].

Beyond conventional pollutants like PMs and VOCs, ML models have also been utilized for predicting more complex environmental compounds such as polycyclic aromatic hydrocarbons (PAHs). A study using support vector regression (SVR) achieved an R^2 of 0.9468 and an RMSE of 7.3116 when predicting total PAHs based on total petroleum hydrocarbons (TPH) in soil[21]. Another study applied a backpropagation ANN model to predict PAH concentrations in soil, achieving an R^2 of 0.9994[22]. Furthermore, PAH concentrations in the air have been predicted using recurrent neural networks trained on data related to forest fires, air emissions, sea ice cover, and meteorological parameters, with RMSE values ranging from 0.51 to 46.36[23].

This study aims to predict ILCR due to the 16 most hazardous PAHs identified by USEPA (United States Environmental Protection Agency) in the air using $PM_{2.5}$ concentrations and weather parameters. The key motivation is to develop a model capable of utilizing low-cost sensor data for real-time ILCR prediction, which would otherwise require substantial financial and computational resources. Additionally, the study introduces a novel approach to converting wind data into "source factors" (SF), which enhances model transferability across different locations. Since wind direction data from different locations cannot be directly combined or used interchangeably, the proposed method first identifies local pollution sources and then integrates this information with wind direction and speed to generate source factors. These factors serve as input variables for the ML model, improving both predictive accuracy and cross-location applicability. The study systematically compares this novel approach to conventional methods, where raw meteorological data is directly fed into the models. To achieve this, three ML models, Artificial Neural Networks (ANN), eXtreme Gradient Boosting (XGB), and Random Forest (RF) were systematically optimized and trained using data from two locations in India.

Results

PAHs profile and source factor-specific ILCR distribution

Figure 1: PAH profile and source-specific ILCR distribution across different pollution source factors. presents the profile of Polycyclic Aromatic Hydrocarbons (PAHs) and the corresponding Incremental Lifetime Cancer Risk (ILCR) distribution associated with different local pollution sources influenced by wind direction. The analysis reveals that vehicular emissions contribute to the highest PAH concentrations, leading to the highest ILCR

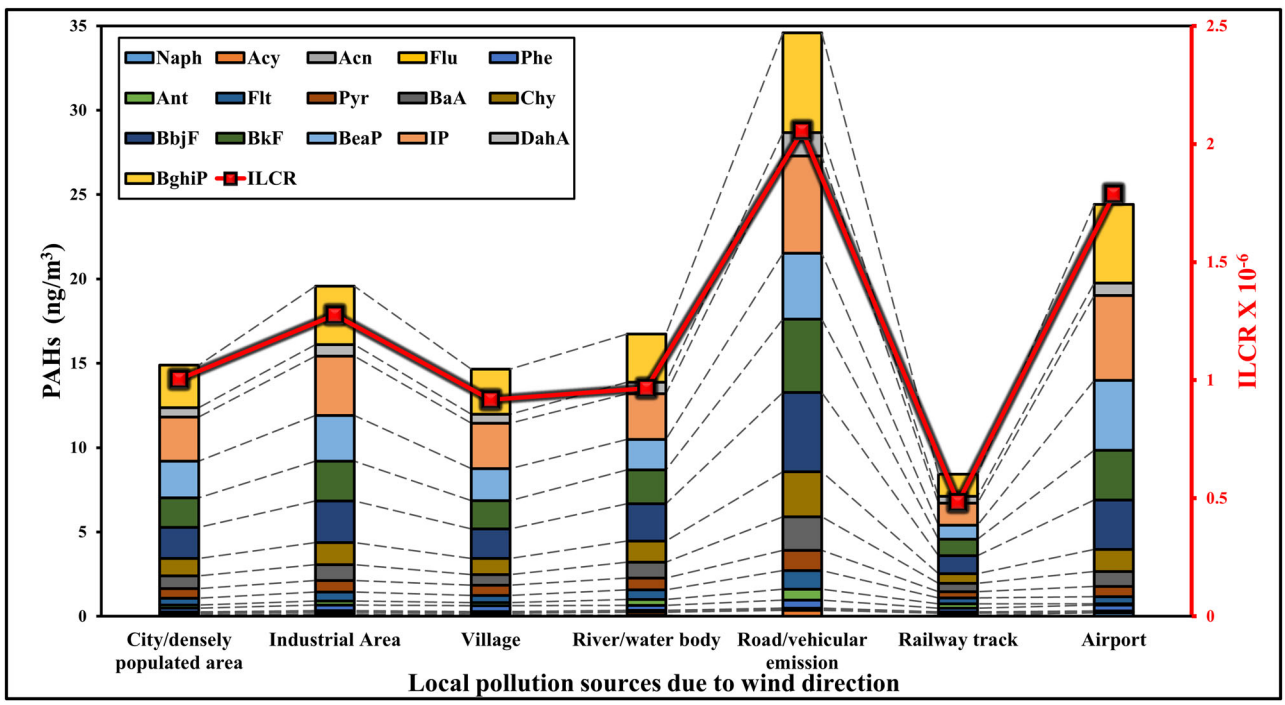


Fig. 1 | PAH profile and source-specific ILCR distribution across different pollution source factors.

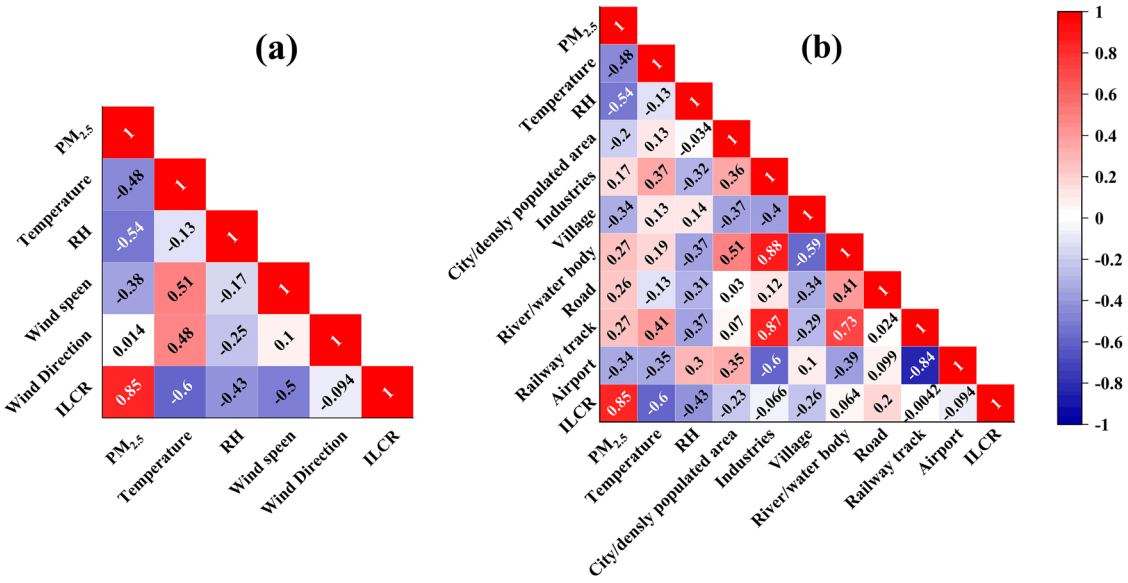


Fig. 2 | Correlation heatmaps of input parameters with ILCR. a Conventional meteorological (CM) inputs, **b** Pollution source method (PSM) inputs.

values. This aligns with existing literature, where road traffic emissions have been identified as a dominant source of PAHs due to incomplete combustion of fossil fuels and vehicular exhaust[24]. Conversely, the lowest PAH levels and ILCR values are observed when wind is flowing from the railway tracks. This can likely be attributed to the increasing electrification of railway networks, reducing dependence on diesel-powered locomotives, which are traditionally known to be significant PAH sources[25]. To provide context on the baseline pollution levels of the study region, the sum of targeted PAHs concentrations was observed to be 16.31 ng/m³ at Jorhat and 21.54 ng/m³ at Shyamnagar. Correspondingly, BaPeq concentrations of all 16 PAHs estimated using the TEF values were 4.10 ng/m³ at Jorhat and 4.40 ng/m³ at Shyamnagar. These values indicate moderate PAH burdens in the ambient air, which serve as the baseline exposure levels for evaluating the Incremental Lifetime Cancer Risk (ILCR) in this study.

The sharp contrast between vehicular and railway emissions highlights the effectiveness of transitioning towards cleaner energy sources in reducing carcinogenic air pollutants.

Industrial areas and densely populated urban regions also show considerable PAH concentrations and moderate ILCR values, suggesting that industrial combustion processes and domestic activities contribute significantly to PAH emissions.

Correlation and sensitivity analysis

The correlation analysis Fig. 2 and sensitivity analysis Fig. 3 provide critical insights into the relationship between input parameters and Incremental Lifetime Cancer Risk (ILCR). The correlation heatmaps (Fig. 2) indicate that $PM_{2.5}$ exhibits the strongest correlation with ILCR (0.85), reinforcing its significance in predicting health risks associated with air pollution exposure. This strong association is further validated by the sensitivity analysis (Fig. 3), where $PM_{2.5}$ demonstrates the highest sensitivity, highlighting its dominant influence on ILCR predictions.

Apart from $PM_{2.5}$, temperature and relative humidity (RH) also show a reasonable correlation with ILCR (−0.6 and −0.43, respectively) and exhibit moderate sensitivity values in Fig. 3. These findings suggest that meteorological parameters, despite being indirect contributors, play a crucial role in influencing pollutant dispersion and human exposure levels. Interestingly, wind direction in the conventional method (CM) presents a very weak correlation (−0.094) with ILCR, which suggests that its direct impact on cancer risk assessment is minimal when used in its traditional form.

When examining pollution source-specific parameters (PSM) in Fig. 2b, it becomes evident that many of these parameters exhibit weak correlations with ILCR. However, despite their poor correlation coefficients, their sensitivity values (Fig. 3) are comparable to those of conventional meteorological inputs. This highlights a crucial aspect of machine learning (ML) modelling, a strong correlation is not a prerequisite for a feature to contribute significantly to predictive models[26]. ML models can capture complex, non-linear interactions among variables, making them superior in handling features that may not exhibit high linear correlations but still influence the outcome through intricate dependencies.

Artificial Neural Network (ANN)

The performance of ANN models trained using the Pollution Source Method (PSM) and Conventional Method (CM) varies significantly depending on the activation function. Figure 4a highlights that while the 'Purelin' activation function ensures consistency, 'Tansig' performs better in terms of accuracy for both methods. Due to this, Fig. 4b focuses on 'Tansig' to explore the influence of neuron count and layer depth. A cyclic pattern in Fig. 4b suggests that model performance increases with the number of neurons per layer. The influence of the number of layers appears marginal, showing a peak at around 5–6 layers before slightly declining. A direct comparison in Fig. 5a along with Fig. 4 demonstrates that PSM-ANN consistently outperforms CM-ANN across various parameter configurations. The selected best model configurations are summarized in Table 1.

Both modelling methods closely follow the general trend of the observed ILCR values, as illustrated in Fig. 5a. However, upon detailed inspection of Fig. 5a and the residual plot presented in Fig. 5b, it becomes evident that the Pollution Source Method (PSM) demonstrates superior performance. Specifically, the PSM approach yields lower residuals, particularly at higher ILCR values, where deviations between the predicted and observed data are more pronounced. This indicates that the PSM provides more accurate predictions during periods of elevated cancer risk compared to the Conventional Method (CM), effectively capturing critical variations that traditional meteorological inputs alone may miss. When analysing the scatter plot of model predictions against actual normalized ILCR values in Fig. S3, it is evident that both models struggle with higher ILCR values. However, the scatter plot of residuals against actual ILCR in Fig. S4, having a closer spread around the zero line for PSM, shows that PSM-ANN exhibits lower residuals than CM-ANN, indicating better overall accuracy. Figure S3 further supports this observation, as PSM-ANN predictions align more closely with the ideal 45-degree reference line. The residual spread in Fig. S4 reveals greater dispersion for CM-ANN, reinforcing that PSM-based modelling yields lower prediction errors and improved stability.

XGBoost

The XGBoost model's performance for different hyperparameter combinations, as illustrated in Fig. 6, follows a cyclic pattern where a higher min_child_weight leads to lower accuracy, while a lower min_child_weight enhances model efficiency. The influence of learning rate and max_depth reduces this disparity, allowing for improved stability in model training. Notably, PSM-XGB consistently outperforms CM-XGB across all parameter combinations, showing superior training and testing R^2 values.

Table 1 presents the final model parameters, while Fig. 7 compare predictions and residuals. Figure 7 reveals that residuals are higher for elevated ILCR values, yet PSM-XGB consistently produces lower errors than CM-XGB. A scatter plot of predicted vs actual values for XGBoost models (Fig. S5) further confirms this, as CM-XGB predictions deviate more from the ideal model line, even for lower ILCR values, highlighting PSM's superiority in capturing the underlying ILCR distribution more effectively. A residual against actual values of the ILCR plot is present in Fig. S6, showing the less spread of PSM residuals around the zero error line.

Furthermore, XGBoost models are often preferred for structured data applications due to their ability to handle complex feature interactions, making them highly effective for environmental predictions[27]. XGBoost has shown promise in environmental applications, such as improving the accuracy of $PM_{2.5}$ predictions in air quality models[28]. The PSM approach provides additional advantages by incorporating refined wind pollution mapping, which enhances the model's ability to capture localized pollution source impacts more effectively than CM models.

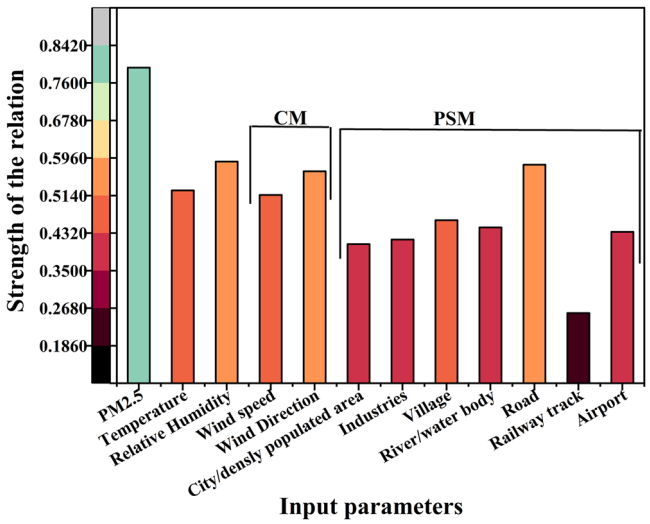


Fig. 3 | Sensitivity analysis of input parameters in ILCR prediction.

Random Forest (RF)

Random Forest models perform well with complex datasets due to their ensemble nature. They are capable of handling high-dimensional data effectively, as at each split, only a random subset of features is considered, reducing computational complexity and preventing overfitting to irrelevant features.

Figure 8 illustrates that model performance improves as minLeafSize decreases, with other parameters showing minimal effect. Notably, PSM-RF exhibits greater stability and higher test R^2 values when minLeafSize is set to 1. Table 1 summarizes the final model parameters.

Prediction accuracy is assessed in Fig. 9, Figs. S7 and S8, with residual distributions in Figs. 9b and S8. Figure 9a shows that both models capture the trend well, though Fig. 9b, suggest that PSM-RF achieves slightly better performance compared to CM-RF. A scatter plot of predicted vs actual values for RF models is presented in Fig. S7 and a residuals against actual values of ILCR plot is shown in Fig. S8. Both of the plots show that the output of PSM is near the ideal line with residuals near zero in comparison to CM. The overall findings reaffirm that PSM-based models yield superior accuracy and generalization capabilities compared to CM-based models.

Model overfitting evaluation

To evaluate the risk of overfitting in the trained models, the study compared the R^2 values obtained during training and testing across a wide range of parameter combinations. As shown in Figs. 4. 6, and 8 for ANN, XGB, and RF, respectively, the general trend of training and testing R^2 remains consistent for all three models, suggesting good generalization capability. Only the ANN model shows occasional deviations where test R^2 drops compared to train R^2, specifically for parameter combinations related to hidden layers 4, 7, and 10. These cases suggest possible overfitting, but they are exceptions rather than the norm.

To further investigate this, we calculated the relative R^2 gap defined as $(R^2_{train} - R^2_{test})/R^2_{train}$, and plotted it for all parameter sets under both the PSM and CM methods. The results are provided in supplementary information as Fig. S9 (ANN), S10 (XGB), and S11 (RF). The XGB and RF models show consistently low relative gaps across all parameter IDs, indicating little to no overfitting. In the case of ANN (Fig. S9), a few parameter

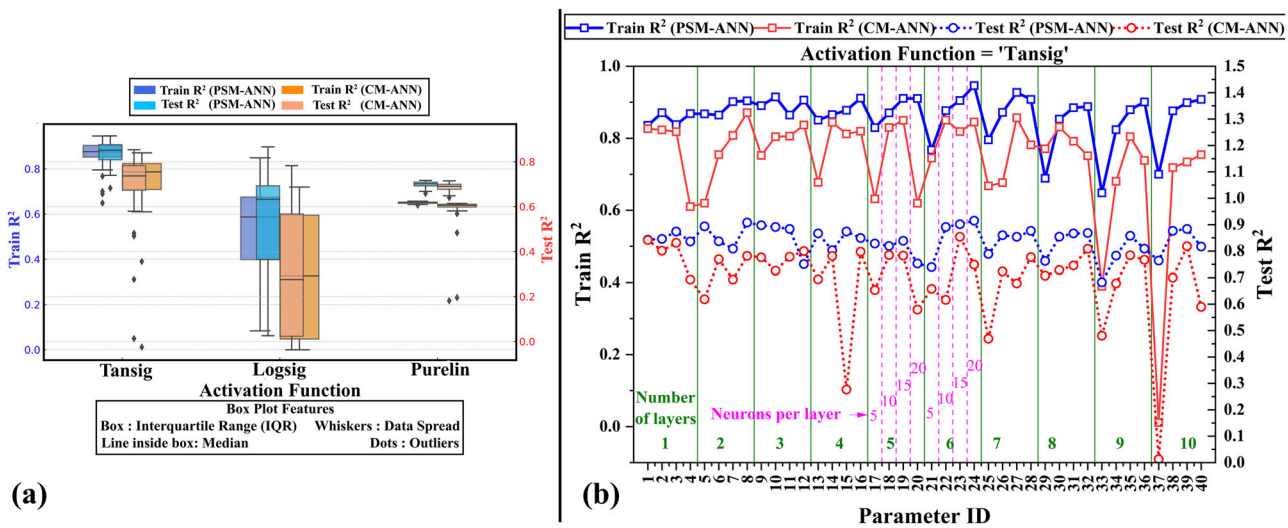


Fig. 4 | ANN model performance for different training parameters. a Training and testing R^2 for different activation functions. **b** Training and testing R^2 for various parameter IDs (see Table S1 for parameter combination for parameter ID).

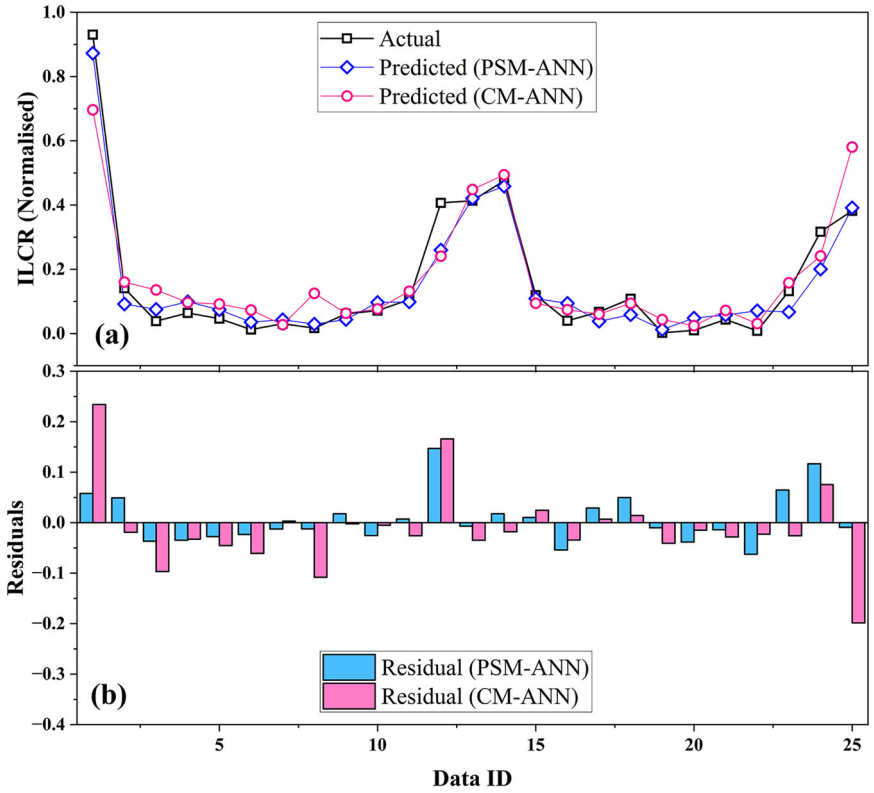


Fig. 5 | Evaluation of ANN model outputs. a Actual ILCR compared with model predictions. **b** Residuals of both ANN models.

Table 1 | Finalised parameters for each model.

Model	Parameter	PSM	CM
ANN	No. of layers	3	6
	Neurons per layer	10	15
	Activation function	tansig	tansig
	Training algorithm	train lm	train lm
	Regularization	0.01	1.01
XGB	Learing rate	0.1	0.3
	max_depth	5	7
	min_child_weight	1	5
	n_estimators	100	100
	subsample	0.8	1
RF	numTrees	50	50
	maxFeatures	5	3
	minLeafSize	5	5

settings show noticeably higher gaps, confirming some level of overfitting in those cases. However, the majority of parameter combinations still maintain low gaps, reinforcing that the model was generally not overtrained. Overall, these assessments confirm that while some overfitting is observed in isolated cases for ANN, the trained models demonstrate stable and generalizable performance.

Model comparison

The statistical distribution patterns of actual normalized ILCR values and those predicted by different models are presented in Fig. 10. Among all models, PSM-ANN aligns most closely with actual data, confirming its effectiveness. Although in the lower range (<0.3), many models are performing well, they fail to match the performance for higher values of ILCR. This may be attributed to ANN's superior ability to capture relationships effectively, even with a limited amount of data.

The evaluation metrics of selected models for the test set are summarized in Table 2. Among both methods used in this study, PSM consistently achieves higher R^2 values while exhibiting lower MAE, MSE, and RMSE than CM models. The Table 2 includes metrics for both normalized and original ILCR values to provide a comprehensive assessment. The high MAPE values observed can be attributed to the small magnitude (near zero) of ILCR data, as values closer to zero tend to inflate MAPE disproportionately. This issue arises because the MAPE calculation involves dividing by the actual value; thus, as the actual value approaches zero, the percentage error approaches infinity[29]. The result of the MLR model for both methods (Table 2) shows that it can capture the part of the variance in ILCR under limited data conditions, but prediction accuracy and error matrix remain less favourable than those obtained using advanced ML models. The results, presented in Table 2, indicate that the PSM-ANN model achieved the highest R^2 value (0.944), demonstrating superior predictive capabilities compared to other models. Followed by PSM-XGB, PSM-RF, CM-ANN, CM-XGB and CM-RF. Therefore, ANN is found to be a better ML technique to develop predictive models related to work similar to this study than XGBoost and random forest.

The Regression Error Characteristic (REC) curve is a graphical evaluation metric used to assess the performance of regression models. It plots the cumulative percentage of predictions that fall within a given error tolerance against the error threshold. Unlike traditional metrics such as RMSE or MAE, the REC curve provides a visual representation of model accuracy across different error levels[30]. The REC curve in Fig. 11 further supports the interpretation of the distribution curve (Fig. 10). As the ideal model's REC curve should be a vertical line at an error threshold of zero, a model with better predictive accuracy will have its REC curve positioned closer to the upper left corner of the plot. The larger the area under the REC curve, the better the model's overall performance. The plot shows that overall PSM-ANN performance is best among other models, followed by PSM-XGB, further reinforcing the advantages of PSM-based models.

Discussion

From an ML perspective, merely supplying wind data from multiple locations and expecting accurate predictions related to air pollution (in this

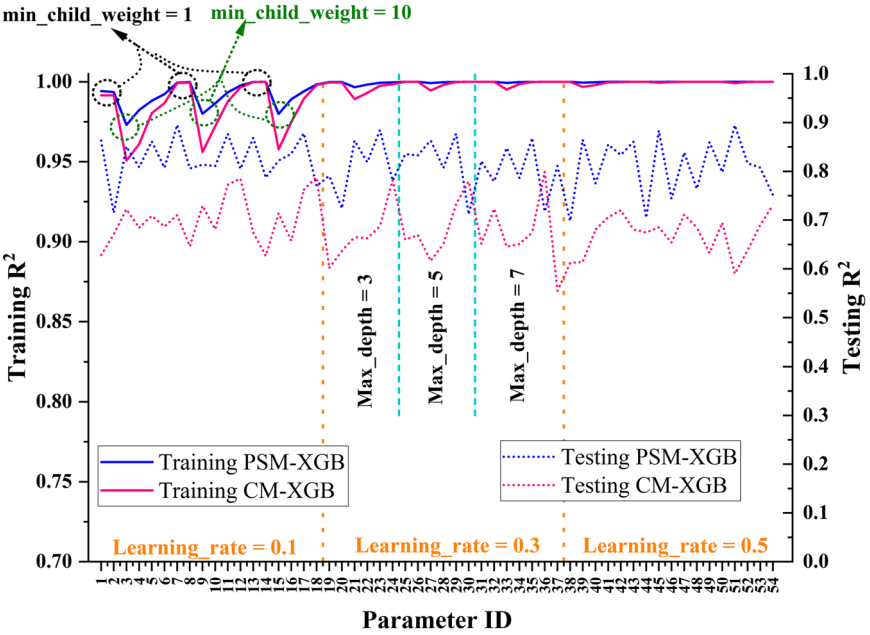


Fig. 6 | Training and testing R^2 of both XGBoost models for different parameter combinations (for parameter combinations of each parameter ID, see table S2).

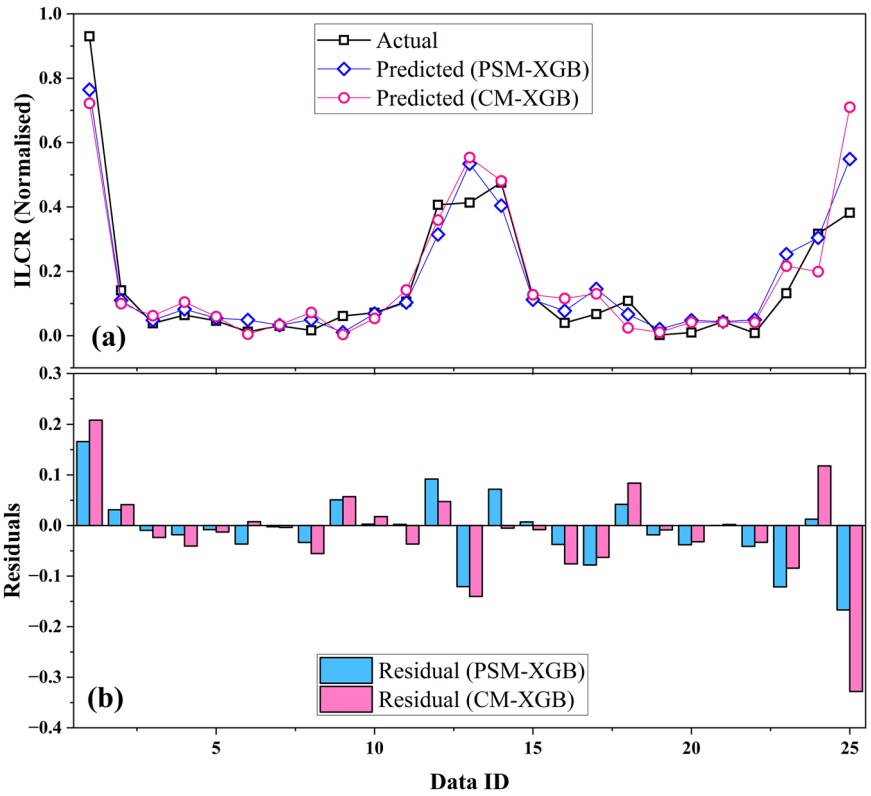


Fig. 7 | Evaluation of XGBoost model outputs. **a** Actual ILCR compared with XGBoost model predictions. **b** Residuals of both XGBoost models.

study, ILCR) might confuse the machine due to the presence of different local sources in different directions at various locations. Converting wind parameters into pollution source factors using PSM provides a more structured and informative input for ML models, resulting in superior predictions.

The distribution of errors across different models provides insight into the reliability of each approach. Residual plots and histograms reveal that PSM-based models demonstrate a tighter clustering of errors around zero, indicating less bias and more precise predictions. CM-based models, on the other hand, exhibit a broader error distribution, highlighting increased uncertainty.

All three ML models evaluated in this study, ANN, XGBoost, and Random Forest, demonstrate a clear advantage when trained using PSM over CM. Despite having a lower correlation coefficient and similar sensitivity of PSM features to those of CM, PSM consistently delivers improved model performance, reducing residual errors and achieving higher predictive accuracy. Moreover, the application of PSM not only enhances individual model performance but also ensures broader applicability across

multiple locations, making it a more generalizable and robust approach for air pollution modelling. Our results are consistent with earlier findings that ANN and hybrid models can effectively capture nonlinear interactions between meteorology and air pollutants[2–5,9,31,32]. However, unlike prior studies that mainly predicted pollutant concentrations, our work shows that using transformed wind parameters for the pollution source method significantly enhances the reliability of ILCR predictions. This aligns with recent studies in applying advanced models such as LSTM and CNN for climate and air quality forecasting[6–8,33], while addressing a novel application in health risk assessment.

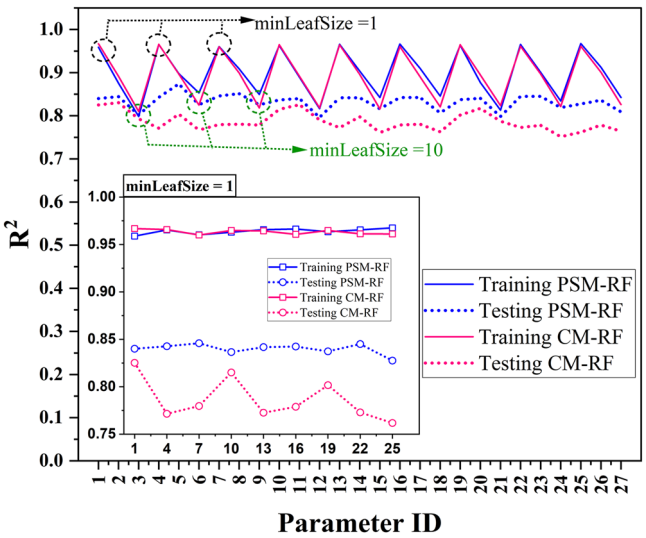


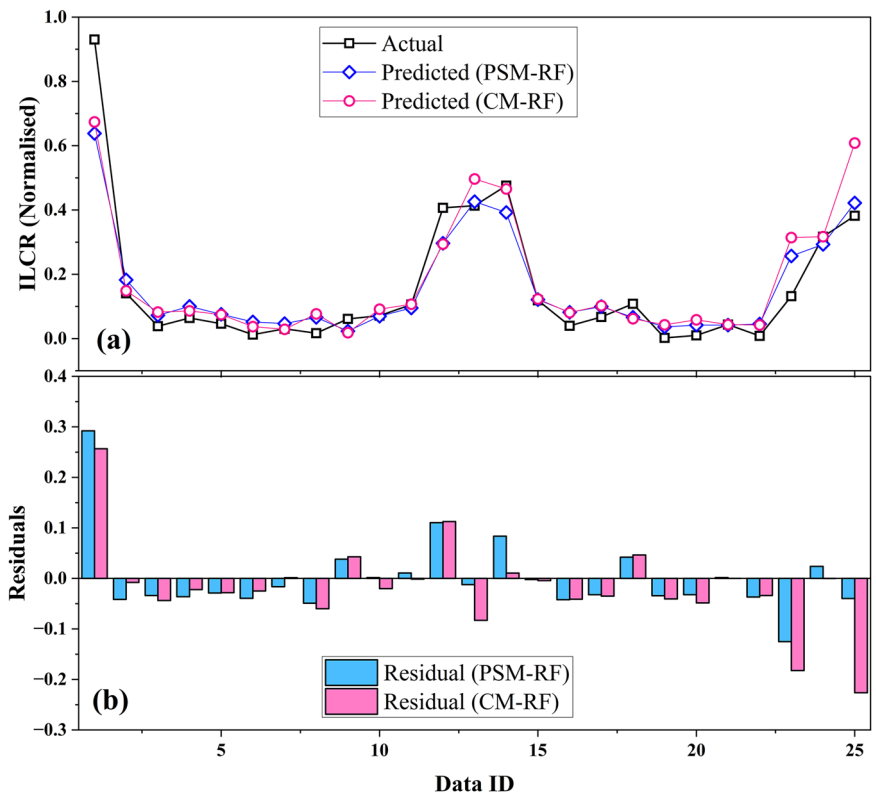
Fig. 8 | Training and testing R^2 of both RF models for different parameter combinations (for parameter combinations of each parameter ID, see Table S3).

This study presents strong preliminary evidence supporting the use of transformed wind parameters to enhance ILCR prediction accuracy through machine learning models. Among the evaluated models, the PSM-ANN model demonstrated the best predictive performance, achieving an R^2 value of 0.944 with a low Mean Absolute Error (MAE) of 0.037. Additionally, the multiple linear regression (MLR) model exhibited significantly lower predictive performance, indicating its limitation in capturing the complex, nonlinear relationships inherent in the data and reinforcing the need for more advanced machine learning approaches. In contrast, the CM-XGB model showed the weakest performance among advanced ML models, with an R^2 of 0.799 and the highest MAE of 0.061, suggesting that conventional meteorological inputs are insufficient to fully represent pollution dispersion dynamics. Across all error metrics, PSM-based models consistently outperformed CM-based models, with notably lower RMSE and MAPE values. The stability of the PSM-ANN model in high ILCR concentration scenarios highlights the effectiveness of using transformed wind features to improve the reliability of health risk predictions.

The implementation of this model in real-time applications presents considerable potential for public health and environmental decision-making. By integrating real-time $PM_{2.5}$ and meteorological data from affordable sensors, cities can estimate ILCR continuously and dynamically, enabling individuals to make better-informed decisions about outdoor exposure. This approach is not only practical but also cost-effective, as traditional ILCR estimation methods based on PAH measurements involve time-consuming sampling and expensive laboratory analysis. Machine learning-based prediction models like PSM-ANN offer a scalable solution for real-time pollution risk assessment at a fraction of the cost.

While the proposed methodology significantly improves prediction accuracy, certain limitations must be acknowledged. In this context, it is worth noting that physics-based dispersion models such as AERMOD have also been successfully integrated with machine learning frameworks, as demonstrated in a recent study[34]. Integrating such models could further enhance the spatial and physical representativeness of ILCR predictions. However, due to the unavailability of detailed emission inventory data in the

Fig. 9 | Evaluation of RF model outputs. a Actual ILCR compared with RF model predictions. **b** Residuals of both RF models.



present study region, this integration was not feasible. Future research could combine AERMOD-based dispersion outputs with machine learning approaches to improve model interpretability and predictive performance. The dataset primarily focuses on two Indian cities, and further studies are required to validate the approach in different geographical regions. Additionally, integrating real-time air quality monitoring data could enhance model responsiveness and adaptability. While the present analysis focuses on external exposure through ambient PAH concentrations, future studies integrating both internal and external exposure pathways, as demonstrated in recent literature[35–37], would provide a more comprehensive assessment of cumulative health risks. Although the dataset used in this study was relatively limited in size, it was adequate to train and validate the models and demonstrate their feasibility for ILCR prediction. However, we recognize that a larger and more diverse dataset would allow the models to capture broader variability in emission patterns and meteorological conditions, thereby further enhancing their generalizability. The framework developed in this study can be easily expanded, and future studies can use techniques such as transfer learning and domain adaptation to strengthen robustness across different geographic and environmental contexts. Future work may also explore deep learning techniques to further optimize predictive performance.

Methods

Study sites

For the development of an ANN model, aerosol sampling and data acquisition for weather parameters were done for two sites, Jorhat and Shyamnagar. Jorhat is a city in Assam located in north-east India at 26°45' N 94°13' E and an average elevation of 116 m. Jorhat has a population of approximately 1.26 lakhs as of the 2011 census[38]. The sampling was done at the Council of Scientific and Industrial Research - North East Institute of Science and Technology (CSIR-NEIST) in the west direction of Jorhat. Another sampling site, Shyamnagar, is a semi-urban town in West Bengal, India, which is located at 22°49' N 88° 23' E with an average elevation of 2 m. The location of both sites is shown in Fig. 12.

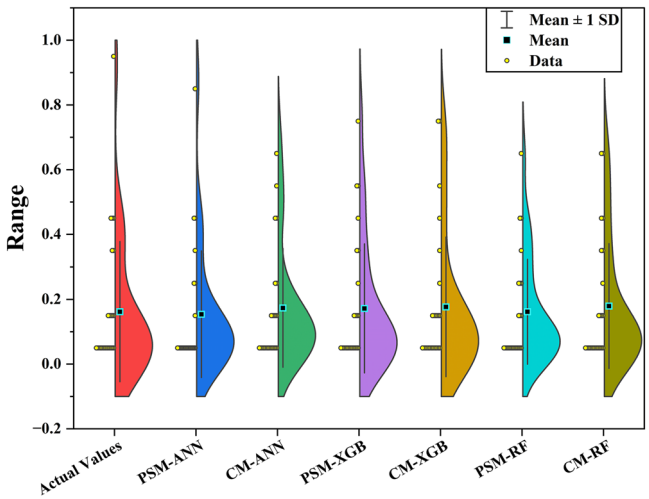


Fig. 10 | Statistical distributions, means, and standard deviations of actual values and predicted by various models developed.

Sampling and PAHs analysis

Air samples were obtained using the Speciation Air Sampler System (SASS) from Met One Instruments, operating at an average flow rate of 6.72 L/min. This system features multiple channels designed for sample collection on various substrates, including Quartz, Teflon, and Nylon. Following collection, the samples were preserved at −19°C until further chemical analysis. Meteorological parameters such as temperature, rainfall, and humidity were recorded using the AIO2 weather station (Met One Inc., OR, USA) installed at the sampling site. The study sites adhered to a 24-hour time-integrated ambient aerosol sampling schedule, conducted every alternate day from January 1 to December 31, 2019.

$PM_{2.5}$ samples were collected on quartz filter paper (47 mm diameter) over a 24-hour period on alternate days of the sampling period for the analysis of polycyclic aromatic hydrocarbons (PAHs). Additionally, 47 mm Teflon filters were used to measure gravimetric mass using a microbalance. PAHs in samples were extracted in a 1:1 mixture by volume of DCM and acetone solvents. This solvent combination was selected as it provided maximum extraction efficiency when compared to other common solvents such as toluene and hexane (Rajeev et al. 2021). A recorded number of punches (each of dia. 3.14 cm^2) were taken in Q cups based on $PM_{2.5}$ concentration of the samples and loaded in the energized dispersive extractor (EDGE, CEM Corporation, USA). Extraction was done with DCM and acetone (1:1 v/v) at 120 °C and 60–70 psi pressure with a holding time of 4 min for each sample. The samples were extracted in 30 ml solvent by the method of adding 10 ml top volume, 10 ml bottom volume, and 10 ml rinse volume of solvent. After the extraction of the sample in DCM and acetone, the samples were concentrated to 1–2 drops by the CentriVap Concentrator. The concentrator was programmed to vaporize the solvent at 30 °C for the first 30 min and at 50 °C for the next 80 min. Vaporization temperatures were selected based on the normal boiling points of DCM (boiling point 39.6 °C) and acetone (boiling point 56 °C), Toluene was added to the concentrated samples to make up the final volume to 2 ml and the resulting solutions were sonicated for 20–25 min for proper dissolution of PAHs. The amount of each PAH in the extracted sample was determined using Gas Chromatography Mass Spectrometer (GC–MS., Agilent technologies; GC: 7890B; MSD: 5977B). A column (DB-5 capillary column) of fused silica with polyamide coating was employed for this analysis. External standards for 5-point calibration were prepared by serial dilution of a 16-PAHs mix solution (EPA 610 PAH Kit 16 analyte in methanol, Sigma-Aldrich) in concentrations of 5 ppb, 10 ppb, 25 ppb, 50 ppb, 100 ppb for quantification of PAH compounds. Helium flow rate was kept at 1 ml/min. In order to

Table 2 | Evaluation parameters for all the trained models.

Model	R^2	MAE	MSE	RMSE	MAPE				
		ILCR normalised	ILCR	ILCR normalised	ILCR	ILCR normalised	ILCR	ILCR normalised	ILCR
PSM-ANN	0.9440	0.0374	3.74E−08	0.0025	2.51E−15	0.0501	5.01E−08	105.6106	105.6108
CM-ANN	0.8539	0.0536	5.36E−08	0.0066	6.55E−15	0.0810	8.10E−08	171.1157	171.1157
PSM-XGB	0.8950	0.0483	4.83E−08	0.0047	4.71E−15	0.0686	6.86E−08	115.9011	115.9010
CM-XGB	0.7993	0.0614	6.14E−08	0.0090	9.00E−15	0.0949	9.49E−08	100.0139	100.0138
PSM-RF	0.8739	0.0483	4.83E−08	0.0057	5.66E−15	0.0752	7.52E−08	148.9406	148.9407
CM-RF	0.8306	0.0550	5.50E−08	0.0076	7.60E−15	0.0872	8.72E−08	165.4630	165.4630
PSM-MLR	0.7114	0.0705	7.05E−08	0.0129	1.29E−14	0.1138	1.14E−07	130.4747	130.4748
CM-MLR	0.6994	0.0725	7.25E−08	0.0135	1.35E−14	0.1161	1.16E−07	107.7413	107.7413

separate compounds based on their boiling points and to reduce total run time, oven temperature ramping was provided. Oven temperature ranged from 90 °C–200 °C, 200 °C–260 °C, and 260 °C–310 °C with temperature ramping of 15 °C/min, 4 °C/min, and 9 °C/min, respectively (Rajeev et al. 2021). Identification of peaks was carried out with the help of the retention time of each PAHs. Replicate samples of field blank were run on the instrument, and the results obtained were used for the field sample correction. This methodology has been published by our group in a previous study[24]. Out of 175 data points, a random 25 data points were kept separate for testing and not used in the training and validation.

Quality assurance and quality control (QA/QC)

Extraction efficiency was evaluated by repeated extraction and analysis of selected aerosol samples ($n = 10$), which confirmed recovery of ~95%. Also, internal standards (phenanthrene d-10 and perylene d-12) response was found to be within ± 5% as a quality control measure.

Clean quartz microfiber filters were extracted and analysed with every 15 samples as field blanks, while solvent blanks were also included and analysed every 6 samples on the GC–MS. The PAHs detected in blanks were subtracted from sample concentrations.

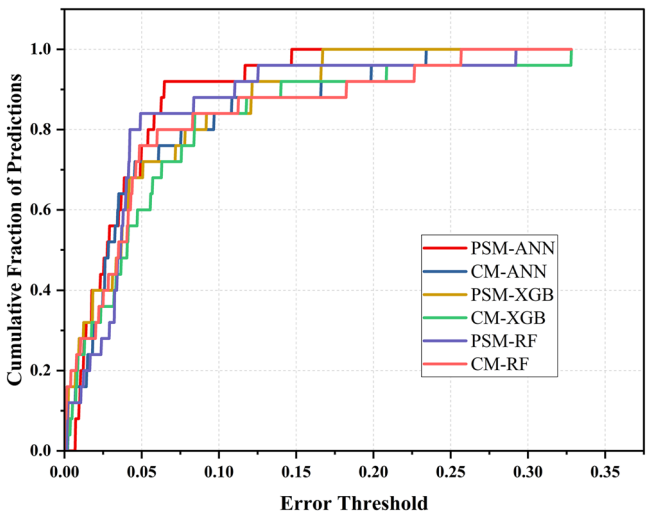


Fig. 11 | Regression Error Characteristic (REC) curve of all the models developed.

ILCR calculation

PAHs can impact human health in several ways, including toxic, cancer-causing, birth defect-inducing, and gene-altering effects. People may be exposed to these compounds through multiple pathways, such as inhaling polluted air, consuming contaminated food or water, or contact with soil. According to USEPA (United States Environmental Protection Agency) list, 16 PAHs (naphthalene, acenaphthylene, acenaphthene, fluorene, phenanthrene, anthracene, fluoranthene, pyrene, benzo(a)anthracene, chrysene, benzo(b,j)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, dibenzo(a,h)anthracene, indeno(1,2,3-cd)pyrene, and benzo(g,h,i)perylene.) have been identified as compounds of grave concern out of which seven PAHs have been marked as most probable human carcinogens like benzo(a) anthracene, benzo(a)pyrene, benzo(b,j)fluoranthene, benzo(k)fluoranthene, chrysene, dibenzo(a,h)anthracene, and indeno (1,2,3-cd) pyrene[24]. Benzo(a)pyrene is one of the most potent carcinogens among all 16 PAHs and is used as a marker for all PAHs in determining the carcinogenic potency. In the current study, the applied formulas follow USEPA risk assessment guidelines[39,40]. Benzo(a)pyrene equivalent is the parameter which is calculated for risk assessment as follows[41]:

$$\mathrm{B[a]Peq} = C_i \, X \, TEF_i \quad (1)$$

C_i is the concentration of the ith species, and TFE is the Toxic Equivalent Factor (TEF) of the i_{th} species. TEF values for each of the 16 PAHs were taken from the previous study[41].

Incremental lifetime cancer risk (ILCR) represents the additional risk of cancer-related mortality beyond the natural background risk due to prolonged exposure to carcinogenic substances such as polycyclic aromatic hydrocarbons (PAHs). It is determined by calculating the lifetime average daily dose (LADD), which quantifies the daily intake of a chemical per

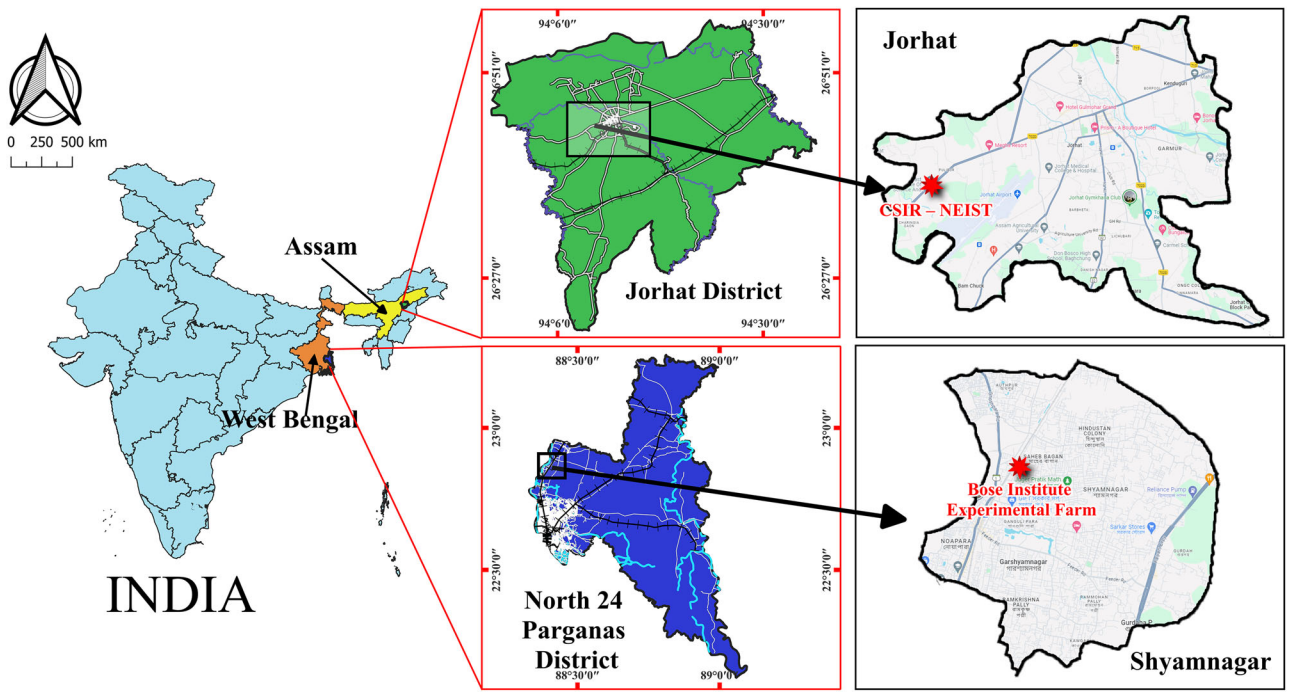


Fig. 12 | Showing sampling sites Jorhat (https://github.com/sssmartsearch/India_Boundary_Updated) and Shyamnagar (https://maps.google.com/). The red star represents the sampling location inside the city/town.

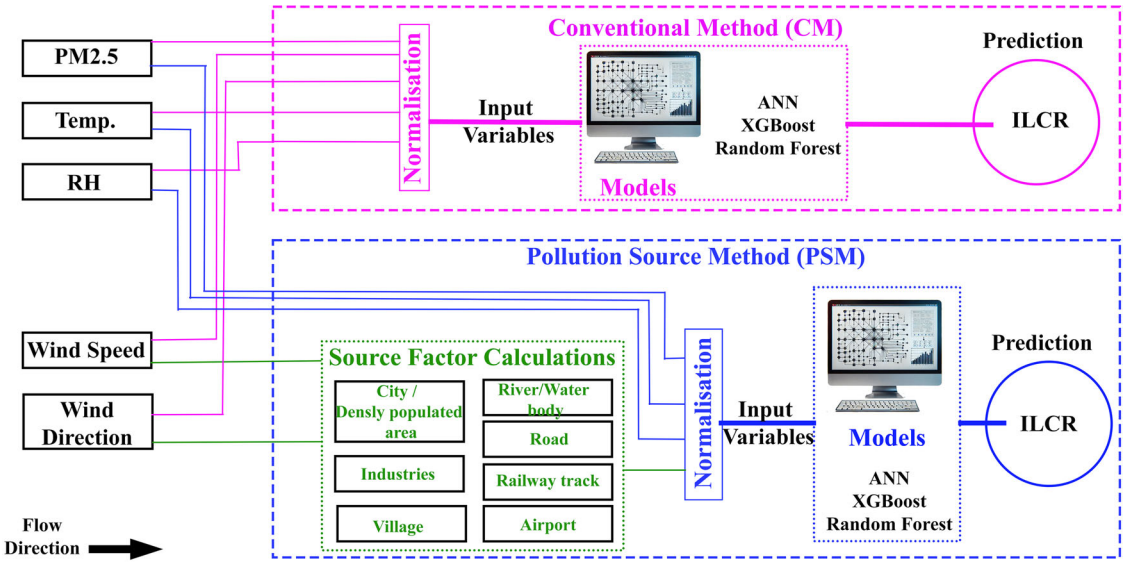


Fig. 13 | Flow chart of Model development using CM and PSM methods.

kilogram of body weight. This measure helps evaluate potential health hazards associated with specific compounds. The formulas for LADD and ILCR, are as follows[42]:

$$\mathrm{LADD}\left(\mathrm{mgkg^{-1}day^{-1}}\right) = \frac{(\mathrm{CP} \times \mathrm{AIR} \times \mathrm{UCF} \times \mathrm{EF} \times \mathrm{LED})}{(\mathrm{BW} \times \mathrm{AT})} \quad (2)$$

$$\mathrm{ILCR} = \mathrm{LADD} \times \mathrm{Cancer\ slope\ factor\ (CSF)} \quad (3)$$

CP represents the BaPeq concentration of individual PAHs in ng/m³. To estimate the Incremental Lifetime Cancer Risk (ILCR), the concentration of each PAH was first converted into its benzo(a)pyrene-equivalent (BaPeq) concentration by applying the respective Toxic Equivalency Factor (TEF), with benzo(a)pyrene (BaP) taken as the reference compound (Eq. 1). This study reports LADD and ILCR for adults. AIR refers to the air inhalation rate, set at 20 m³/day. UCF is the unit conversion factor from ng to mg (10^{-6}). EF denotes the emission frequency, standardized at 350 days per year. LED represents lifetime exposure duration, calculated as 24 years. BW corresponds to body weight, with values of 70 kg. AT signifies the average lifespan, estimated at 25,550 days (70 × 365) (Rajeev et al., 2021; Singh & Gupta, 2016b).CSF, or cancer slope factor, is the key parameter for assessing carcinogenic hazards, with risk determined by the equation:

$$CSF = risk\ per\ unit\ dose = risk\ per\ mgkg^{-1}day^{-1} \quad (4)$$

Previous research has reported the CSF value for benzo(a)pyrene as 3.1, with a geometric standard deviation of 1.8 for risk assessment[42–44].

Model development

In this study, two approaches were employed to develop machine learning (ML) models using $PM_{2.5}$ and weather data as input. The first approach, referred to as the Conventional Method (CM), utilized the meteorological parameters in their original form. Atmospheric temperature, relative humidity (RH), wind direction, and wind speed were commonly used in traditional ML model development. In the second approach, termed the Pollution Source Method (PSM), wind direction and wind speed data were transformed into novel variables called 'source factor,' while the other meteorological parameters, such as atmospheric temperature and RH, were retained in their original form. Figure 13 shows the methodology flow chart for both the CM and PSM methods. To understand the need for advanced machine learning models, multiple linear regression (MLR) models were also developed for both methods and compared to the other models' results.

Source Factor (SF) calculations

For this study, eight common air pollution sources were identified for both cities: urban/densely populated areas, industries, villages, forests, rivers/water bodies, roads (vehicular emissions), railway tracks, and airports. The maps of the sampling locations were divided into 16 equal sectors, and the locations of these sources were determined using a combination of Google Maps data and local surveys.

For each sector, when wind originated from a specific direction corresponding to that sector, all pollution sources within the sector were assigned a value of 1 multiplied by the wind speed (to account for its weightage), while all other source factors were assigned a value of 0. Equation (4) calculates the source factor (SF_i) for each pollution source by summing the product of the wind direction factor (WDF_i) and wind speed (WS_t) over a 24-hour period. Equation (5) defines the wind direction factor (WDF_i). Which are given below:

$$SF_i = \sum_{t=0}^{1440} \left(WDF_i \times WS_t\right) \quad (5)$$

$$WDF_i = \begin{cases} 1, \textit{When wind is blowing from sector of source i} \\ 0, \textit{When wind is not blowing from sector of source i} \end{cases} \quad (6)$$

Where SF_i is the source factor for pollution source i, WDF_i is the wind direction factor, WS_t is the wind speed at time t, and t represents time in minutes (0–1440 minutes in a day).

The sector divisions and the spatial distribution of pollution sources are illustrated in Fig. 14 for Jorhat, and that of Shyamnagar is presented in Figs. S1 and S2 for Shyamnagar. For example, if a wind with a speed of 1.2 m/s was blowing from the east toward the sampling location in Jorhat, the source factors for the urban/densely populated area, industries, and airport within the corresponding sector were assigned a value of 1.2, while the other source factors were assigned a value of 0. These calculations were performed every minute based on real-time data from the weather station. To determine daily source factor values, the minute-wise values for each factor were summed up. Before training the models, the data were normalized along with the other input variables.

Artificial Neural Network (ANN) model development

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of the human brain, widely used for solving complex nonlinear problems in various fields, including environmental

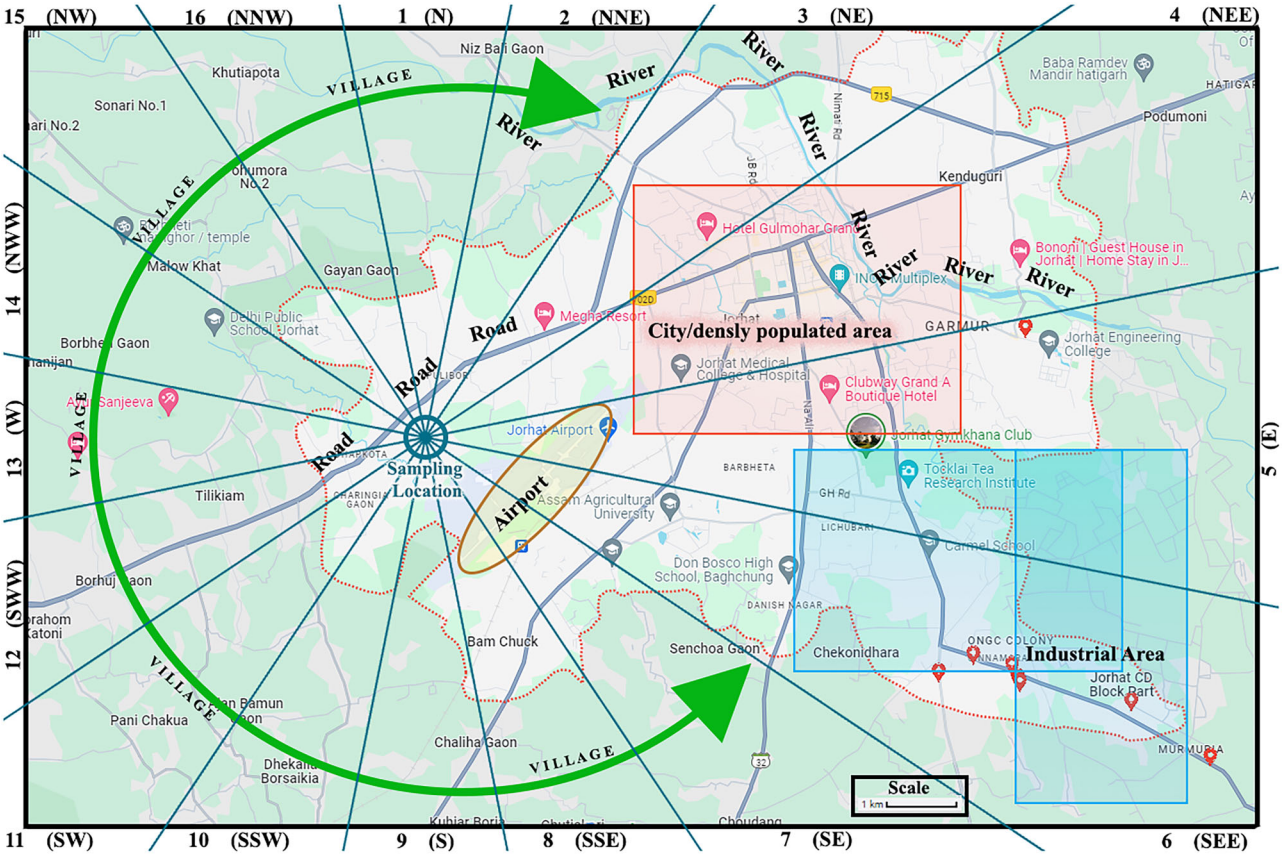


Fig. 14 | Sector division based on direction and pollution source locations of Jorhat city (https://maps.google.com/).

science and air pollution studies. ANNs consist of interconnected layers of nodes, also known as neurons, that process input data through weighted connections, non-linear activation functions, and bias terms to produce output predictions. In this study, a feedforward neural network architecture was employed, trained using the backpropagation algorithm[45] to predict the incremental lifetime cancer risk (ILCR) associated with $PM_{2.5}$ exposure. The ANN model was trained using a systematic approach to optimize its architecture. The number of hidden layers was varied from 1 to 10, while the number of neurons per layer was tested from 5 to 20 in increments of 5. Three activation functions, such as 'logsig,' 'tansig,' and 'purelin,' were evaluated to determine their impact on model performance. Each unique combination of these parameters was trained and validated 10 times to ensure consistency and reliability. The best-performing configuration for each combination was recorded and compared to identify the optimal architecture for the study. Input variables included conventional meteorological data (e.g., temperature, relative humidity) and transformed pollution source factors derived from wind parameters, as described earlier. The ANN model developed using the Conventional Method (CM) was designated as 'CM-ANN', while the model developed using the Pollution Source Method (PSM) was designated as 'PSM-ANN'.

XGBoost (eXtreme Gradient Boosting) model development

Extreme Gradient Boosting (XGBoost) is an advanced machine learning algorithm based on decision-tree ensembles, widely recognized for its speed and accuracy in predictive modelling tasks. XGBoost incorporates a gradient boosting framework that optimizes model predictions by iteratively minimizing the loss function and updating weights for misclassified samples[46]. Both methods i.e., CM and PSM, were used to develop two models designated as CM-XGB and PSM-XGB respectively.

The XGBoost models were trained using grid search approach to optimize hyperparameters, including maximum tree depth (max_depth) varied between 3 and 7, learning rate (learning_rate) tested at 0.1, 0.3, and 0.5, minimum child weight (min_child_weight) evaluated at 1, 5, and 10, and subsample ratio (subsample) set at 0.8 and 1.0 to control the fraction of data used in each boosting iteration. Each unique combination of these parameters was trained and validated 10 times to ensure consistency and robustness. The best-performing configuration for each parameter combination was recorded, and the corresponding results were analysed to identify the optimal model configuration.

Random Forest (RF) model development

Random Forest (RF) model, a widely used ensemble learning method known for its robustness and predictive accuracy in regression and classification tasks. RF constructs multiple decision trees during the training phase, each using a randomly selected subset of features and data samples. By averaging the outputs of these trees, RF minimizes overfitting and enhances generalization, making it particularly effective for datasets with complex, non-linear relationships and high-dimensional feature spaces[47]. The inherent randomness in RF also provides a built-in mechanism for estimating the importance of individual features, adding interpretability to the model's predictions. MATLAB's 'TreeBagger' function was employed for regression modelling, offering flexibility in customizing critical hyperparameters to achieve optimal performance.

A grid search optimization approach was applied to tune three key hyperparameters: the number of trees in the ensemble (optimized over the range [50, 100, 200]), the number of features considered at each split (optimized over the range [3, 5, 7]), and the minimum leaf size for terminal nodes (optimized over the range [1, 5, 10]). Each hyperparameter configuration was evaluated 10 times to account for the stochastic nature of the RF algorithm, with the best model selected based on the highest coefficient of determination (R^2) value achieved on the test data. Two RF models trained using CM and PSM were designated as CM-RF and PSM-RF, respectively.

Table 3 | Ideal value and range of statistical parameters used for evaluation.

Parameter	R^2	MAE	MSE	RMSE	MAPE
Ideal Value	1	0	0	0	0
Range	0–1 (can be negative)	0–∞	0–∞	0–∞	0–∞ (can be >100%)

Overall, model optimization was performed through backpropagation-based architecture tuning for ANN and grid search-based hyperparameter tuning for XGBoost and Random Forest.

Model evaluation

This study evaluates the accuracy, robustness, and generalizability of the models using five widely accepted performance metrics. The coefficient of determination (R^2) indicates the proportion of variance in the target variable explained by the model, providing a measure of goodness-of-fit. Mean Absolute Error (MAE) quantifies the average magnitude of errors without considering their direction, offering an intuitive measure of prediction accuracy. Mean Squared Error (MSE) penalizes larger errors more than smaller ones by squaring the differences, making it sensitive to outliers. Root Mean Squared Error (RMSE), the square root of MSE, expresses the error in the same unit as the target variable, facilitating better interpretability. Mean Absolute Percentage Error (MAPE) measures the percentage error relative to the actual values, making it useful for assessing relative prediction accuracy across different scales. The mathematical formulas of these parameters are given by Eqs. 6–10[48] as below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

$$RMSE = \sqrt{MSE} \quad (10)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (11)$$

where y_i represents the actual observed values, $\hat{y}_i$ denotes the predicted values, $\bar{y}$ is the mean of the observed values, and n is the total number of observations. Table 3 provides the range and ideal value of these statistical parameters.

Sensitivity calculation

To assess the sensitivity of input variables in predicting ILCR, the study employed the cosine amplitude method for sensitivity analysis using the correlation strength equation. The relationship between an input variable (x_i) and the target variable (x_j) is quantified by the sensitivity coefficient (R_{ij}), calculated using the equation:

$$R_{ij} = \frac{\sum_{k=1}^{m} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}} \quad (12)$$

where R_{ij} represents the strength of the relationship between the input and output variables across m observations. A higher R_{ij} value indicates a stronger influence of the respective input variable on ILCR predictions. This analysis was conducted separately for both the conventional method (CM) and the proposed pollution source method (PSM) to evaluate how the transformation of meteorological inputs affects model sensitivity.

Data availability

All the data are available from the corresponding authors upon request.

Code availability

Code sources are available from the corresponding authors upon request.

Received: 12 July 2025; Accepted: 21 October 2025;
Published online: 25 November 2025

References

1. Yadav, V., Yadav, A. K., Singh, V. & Singh, T. Artificial neural network an innovative approach in air pollutant prediction for environmental applications: A review. *Results Eng.* **22**, 102305 (2024).
2. He, Z., Guo, Q., Wang, Z. & Li, X. Prediction of monthly $PM_{2.5}$ concentration in Liaocheng in China Employing Artificial Neural Network. *Atmosphere* **13**, 1221 (2022).
3. Guo, Q., He, Z. & Wang, Z. Predicting of daily $PM_{2.5}$ concentration employing wavelet artificial neural networks based on meteorological elements in Shanghai, China. *Toxics* **11**, 51 (2023).
4. Guo, Q. et al. Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol Air Qual. Res.* **20**, 1429–1439 (2020).
5. Guo, Q., He, Z. & Wang, Z. Prediction of hourly $PM_{2.5}$ and PM_{10} concentrations in Chongqing City in China based on artificial neural network. *Aerosol Air Qual. Res.* **23**, 220448 (2023).
6. Guo, Q., He, Z. & Wang, Z. Prediction of monthly average and extreme atmospheric temperatures in Zhengzhou based on artificial neural network and deep learning models. *Front. Glob. Change* **6**, 1249300 (2023).
7. Guo, Q., He, Z. & Wang, Z. Monthly climate prediction using deep convolutional neural network and long short-term memory. *Sci. Rep.* **14**, 17748 (2024).
8. Guo, Q. et al. A performance comparison study on climate prediction in Weifang City using different deep learning models. *Water* **16**, 2870 (2024).
9. He, Z. & Guo, Q. Comparative analysis of multiple deep learning models for forecasting monthly ambient $PM_{2.5}$ concentrations: A case study in Dezhou City, China. *Atmosphere* **15**, 1432 (2024).
10. Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A. & Samarasinghe, S. Complex time series analysis of PM_{10} and $PM_{2.5}$ for a coastal site using artificial neural network modelling and k-means clustering. *Atmos. Environ.* **94**, 106–116 (2014).
11. Elangasinghe, M. A., Singhal, N., Dirks, K. N. & Salmond, J. A. Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmos. Pollut. Res* **5**, 696–708 (2014).
12. Arhami, M., Kamali, N. & Rajabi, M. M. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environ. Sci. Pollut. Res.* **20**, 4777–4789 (2013).
13. Russo, A., Raischel, F. & Lind, P. G. Air quality prediction using optimal neural networks with stochastic variables. *Atmos. Environ.* **79**, 822–830 (2013).
14. Anurag, N. V., Burra, Y., Sharanya, S. & Gireeshan, M. G. Air quality index prediction using meteorological data using featured based weighted xgboost. *Int. J. Innov. Technol. Explor. Eng.* **8**, 1026–1029 (2019).
15. Christian, M. M. & Choi, H. Air Quality Forecasting Using Machine Learning: A Global perspective with Relevance to Low-Resource Settings. In *SIBR 2024 (Seoul) Conference on Interdisciplinary Business and Economics Research, 5th-6th January 2024, Seoul, S. Korea* (Seoul, 2024).

16. Prasad, K., Gorai, A. K. & Goyal, P. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmos. Environ.* **128**, 246–262 (2016).
17. Yadav, V. & Nath, S. Novel hybrid model for daily prediction of PM 10 using principal component analysis and artificial neural network. *Int. J. Environ. Sci. Technol.* **16**, 2839–2848 (2019).
18. Wang, S. Y., Lin, W. B. & Shu, Y. C. Design of machine learning prediction system based on the Internet of Things framework for monitoring fine pm concentrations. *Environments - MDPI* **8**, 99 (2021).
19. Samad, A., Garuda, S., Vogt, U. & Yang, B. Air pollution prediction using machine learning techniques - An approach to replace existing monitoring stations with virtual monitoring stations. *Atmos. Environ* **310**, 119987 (2023).
20. Li, B. et al. Physical informed neural network improving the WRF-CHEM results of air pollution using satellite-based remote sensing data. *Atmos. Environ.* **311**, 120031 (2023).
21. Akinpelu, A. A. et al. A support vector regression model for the prediction of total polyaromatic hydrocarbons in soil: an artificial intelligent system for mapping environmental pollution. *Neural Comput. Appl.* **32**, 14899–14908 (2020).
22. Olawoyin, R. Application of backpropagation artificial neural network prediction model for the PAH bioremediation of polluted soil. *Chemosphere* **161**, 145–150 (2016).
23. Zhao, Y. et al. Deep learning prediction of polycyclic aromatic hydrocarbons in the High Arctic. *Environ. Sci. Technol.* **53**, 13238–13245, https://doi.org/10.1021/acs.est.9b05000 (2019).
24. Vishwakarma, P. et al. Wintertime trends of particulate-bound polycyclic aromatic hydrocarbons (PAHs) at north-east site of India: chemical characterization and source identification. *J. Atmos. Chem.* **80**, 251–269 (2023).
25. Jamhari, A. A. et al. Concentration and source identification of polycyclic aromatic hydrocarbons (PAHs) in PM10 of urban, industrial and semi-urban areas in Malaysia. *Atmos. Environ.* **86**, 16–27 (2014).
26. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28 (2014).
27. Jain, P. K. Predicting bentonite swelling pressure: optimized XGBoost versus neural networks. *Sci. Rep.* **14**, 17533 (2024). Utkarsh.
28. Ma, J., Yu, Z., Qu, Y., Xu, J. & Cao, Y. Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai. *Aerosol Air Qual. Res.* **20**, 128–138 (2020).
29. Kim, S. & Kim, H. A new metric of absolute percentage error for intermittent demand forecasts. *Int J. Forecast.* **32**, 669–679 (2016).
30. Bi, J. & Bennett, K. P. Regression error characteristic curves. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* 43–50 (2003).
31. Guo, Q. & He, Z. Prediction of the confirmed cases and deaths of global COVID-19 using artificial intelligence. *Environ. Sci. Pollut. Res.* **28**, 11672–11682 (2021).
32. He, Z., Guo, Q., Wang, Z. & Li, X. A hybrid wavelet-based deep learning model for accurate prediction of daily surface $PM_{2.5}$ concentrations in Guangzhou city. *Toxics* **13**, 254 (2025).
33. Guo, Q., He, Z. & Wang, Z. Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in Liaocheng City. *Sci. Rep.* **15**, 6798 (2025).
34. Karimi, S., Asghari, M., Rabie, R. & Emami Niri, M. Machine learning-based white-box prediction and correlation analysis of air pollutants in proximity to industrial zones. *Process Saf. Environ. Prot.* **178**, 1009–1025 (2023).
35. Rabie, R., Mirzaei, M. & Karimi, S. Monte Carlo simulation framework for assessing heavy metal exposure and adverse health effects in fly-in fly-out workers. *J. Environ. Manag.* **380**, 125074 (2025).
36. Shams, S. R. et al. Assessing the effectiveness of artificial neural networks (ANN) and multiple linear regressions (MLR) in forcasting AQI and PM10 and evaluating health impacts through AirQ+ (case study: Tehran). *Environ. Pollut.* **338**, 122623 (2023).
37. Shams, S. R., Jahani, A., Kalantary, S., Moeinaddini, M. & Khorasani, N. Artificial intelligence accuracy assessment in NO_2 concentration forecasting of metropolises air. *Sci. Rep.* **11**, 1805 (2021).
38. Directorate of Census Operations, A. *Census of India 2011 Assam Series-19 Part XII-a District Census Handbook Jorhat Village and Town Directory Directorate of Census Operations Assam*. (2014).
39. USEPA. *USEPA: OSWER: Risk Assessment Guidance for Superfund - Volume I - Human Health Evaluation Manual (Part A) Interim Final, December 1989*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.epa.gov/system/files/documents/2024-10/rags_a_508.pdf (1989).
40. USEPA. *Guidelines for Carcinogen Risk Assessment*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.epa.gov/sites/default/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf (2005).
41. Nisbet, I. C. T. & LaGoy, P. K. Toxic equivalency factors (TEFs) for polycyclic aromatic hydrocarbons (PAHs). *Regul. Toxicol. Pharmacol.* **16**, 290–300 (1992).
42. Singh, D. K. & Gupta, T. Effect through inhalation on human health of PM1-bound polycyclic aromatic hydrocarbons collected from foggy days in the northern part of India. *J. Hazard Mater.* **306**, 257–268 (2016).
43. Rajeev, P., Singh, A. K., Singh, G. K., Vaishya, R. C. & Gupta, T. Chemical characterization, source identification and health risk assessment of polycyclic aromatic hydrocarbons in ambient particulate matter over central Indo-Gangetic Plain. *Urban Clim.* **35**, 100755 (2021).
44. Chen, S.-C. & Liao, C.-M. Health risk assessment on human exposed to environmental polycyclic aromatic hydrocarbons pollution sources. *Sci. Total Environ.* **366**, 112–123 (2006).
45. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
46. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939785.
47. Breiman, L. Random Forests. *Mach. Learn* **45**, 5–32 (2001).
48. Asteris, P. G., Skentou, A. D., Bardhan, A., Samui, P. & Pilakoutas, K. Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cem. Concr. Res*. **145**, (2021).

Acknowledgements

The authors gratefully acknowledge the Ministry of Environment, Forest and Climate Change (MoEFCC), Government of India, under the National Carbonaceous Aerosols Programme (NCAP-COALESCE), and the Indian Institute of Technology Kanpur for providing institutional support and research facilities that contributed to this work. No external funding was used for this study.

Author contributions

S.S. has contributed to conceptualization, investigation, formal analysis, validation, interpretation, and writing the original draft. P.V. has contributed to data acquisition, investigation, interpretation, review, and editing. T.G. has contributed to conceptualization, supervision, resources, review, and editing.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44407-025-00035-4.

Correspondence and requests for materials should be addressed to Shivam Singh or Tarun Gupta.