Journal of Air Pollution and Health (Spring 2026); 11(2): 183-200

Original Article

Available online at http://japh.tums.ac.ir

Empowering data analysis and machine learning to predict asthma intensity using air pollutants in conjunction with environmental factors

Priyanshi Kotlia[1,2], *Janmejay Pant*[1,2,*], *Manoj Chandra Lohani*[2,3]

[1] School of Computing, Bhimtal Campus, Graphic Era Hill University, India
[2] Department of Centre for Promotion of Research, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India
[3] School of Computing, Haldwani Campus, Graphic Era Hill University, India

ARTICLE INFORMATION

Article Chronology:
Received 21 February 2026
Revised 07 May 2026
Accepted 02 June 2026
Published 29 June 2026

Keywords:
Air pollutants; Machine learning; Logistic regression; Random forest classifier; Gradient boosting (XGBoost)

CORRESPONDING AUTHOR:
geujay2010@gmail.com
Tel: (+5942) 270015
Fax: (+5942) 270000

ABSTRACT

Introduction: Asthma is a respiratory disease, the severity of which is affected by air pollutants and environmental factors. Predicting asthma severity can help in disease monitoring and control. The objective of this research is to develop a model for predicting the severity of asthma based on environmental and demographical factors using machine learning.

Materials and methods: Data was obtained from different districts in Uttarakhand, India, from government sources. Asthma severity was the output feature or dependent feature, while the input features or independent features were air pollutants such as Particulate Matters ($PM_{2.5}$, PM_{10}), Nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), Ozone (O_3), Carbon monoxide (CO), environmental factors (temperature, humidity, wind speed) and socio-economic factors (age, gender) in addition to a pollution index. Logistic Regression, Random Forest and XGBoost machine learning models were used for multi-class classification. The metrics for model performance were accuracy, precision, recall and F1-score.

Results: Logistic Regression had the highest accuracy (98%) compared to Random Forest and XGBoost (both 89%). It had goo) with an F1-score of 0.00 (support=1).

Conclusion: Our findings show the potential of machine learning models, especially class performance with F1-scores of 0.99 (class 0) and 0.96 (class 1). But all models could not predict the minority class (class 2). Logistic Regression, to predict asthma severity from environmental data. But it has limitations due to the exclusion of various factors like smoking, obesity, genetics, previous asthma, and medication.

Please cite this article as: Kotlia P, Pant J, Lohani MCh. Empowering data analysis and machine learning to predict asthma intensity using air pollutants in conjunction with environmental factors. Journal of Air Pollution and Health. 2026;11(2): 183-200.
Doi: https://doi.org/10.18502/japh.v11i2.21873

Introduction

Asthma, a chronic respiratory condition, is still a major global health issue, affecting millions of individuals and imposing a huge health burden worldwide [1]. Air pollution, for instance, has been studied exhaustively on its relationship with asthma incidence and severity, for it is among the most influential factors outside the body affecting this illness. Particulate matter ($PM_{2.5}$ and PM_{10}), NO_2, SO_2, CO, and O_3 are all associated with the worsening of asthma [2-8]. Studies have shown that these pollutants lead to airway inflammation and hyperresponsivity to develop symptoms that include wheezing, coughing and shortness of breath [9-15]. These pollutants are mainly from automobiles, emissions from industries and garbage dumps, and thus, asthma is common in urban and industrialized regions with dumping sights [16]. Apart from air pollution, other environmental conditions, including temperature, humidity and wind speed, have been known to affect asthmatic conditions' triggering and severity. For instance, high temperatures raise respiratory distress while high humidity fosters mould growth; dust mites and other allergens which are associated to asthma. The same applies to wind speed that also helps in spreading airborne allergens and pollution, worsening asthma situation in the affected regions. Because asthma is a multifaceted condition, environmental factors present known potential to raise the rates of attacks [17-18]. Information and communication technology GIS/RS was found to be suitable in providing a geographic approach to asthma distribution and environmental connection in public health research [19]. Using GIS, the investigator is capable of studying spatial distribution of asthma occurrence and identifying high-risk zones and finding environmental quality factors associated with those zones. On this basis, making use of health data combined with pollution and climate data GIS can mark out asthma-susceptible areas and furnish appropriate information to care providers and legal authorities. RS assists GIS by offering holdings on spatially, large, satellite-derived data on air quality, vegetation, temperature, and other such factors which can be used to measure the effects made over time with references to asthma. Advanced analysis has made drastic changes in risk measures of Asthma using ML since they deal with sizeable data work and establish unproportional relations between environmental characteristics and Asthma occurrence. Other techniques like RF, xgBoost, Bagging, AdaBoost and Stacking techniques have been found useful in asthma related studies because of their high performance and collection of huge amounts of data in environmental health studies [19]. These algorithms can be used to analyze virtually all types of data with respect to asthma risk factors including air parameters, meteorological characteristics, and demographic data such as age and gender, which may not be identified using more traditional analytical techniques. This makes the application of the GIS and ML-based techniques quite special, especially for asthma risk prediction as it obtains or produces a perfect spatial ratio that can distinguish risky areas and forecast future asthma risky zones. Using air quality index, temperature, humidity and other meteorological factors, these models can predict the incidence of asthma. Moreover, it can also consider temporal dependencies as the time-lag effect of pollutant exposure on asthma exacerbations [20]. This skill is essential to help identify looming asthma threats to healthcare practitioners and public health agencies, so action can be taken. Many researchers have established the possibility of applying machine learning in asthma-related research. For example, linear regression models, RF, and LASSO were employed to measure each of the pollutants' contribution to asthma hospital visits while using Bayesian models in mapping

the distribution of asthma about environmental risk factors [21]. Research done on a cross-section of cities, starting from New York to Karachi, also identified strong positive associations between the prevalence of asthma and $PM_{2.5}$ levels, while GIS analysis was employed to establish spatial patterns [22]. This paper[23] looked at the role of exposure to air pollution in the prevalence of asthma in adulthood. It was a cross-sectional study between November 2015 and February 2016 with a population of 3,172 persons aged 2044 years in Tehran, Iran. Face-to-face interviews using a standardized European Community Respiratory Health Survey (ECRHS) questionnaire were used in collecting data on asthma prevalence. Data on the concentration of air pollutants throughout Tehran was acquired at the Air Quality Control Company (AQCC), and the individual exposure levels were estimated using the GIS-based spatial analysis.To test the hypothesis of the relationship between asthma symptoms and air pollution, the crude and adjusted logistic regression were performed, and the results were expressed in terms of Odds Ratios (OR) and Confidence Intervals (CI). The results showed that prevalence of asthma in adults in Tehran was 11.73% which is greater than the national prevalence rate of 8.9%. Even though the prevalence rate of asthma among the males (6.43%) seemed to be higher than among the females (5.26%), the difference was not statistically significant (P = 0.29).The aim of the study[24] was to determine the effect of air pollution and environmental noise on the incidence of asthma in children in Tehran, Iran. A standardized questionnaire was used to collect the data; either by parents of children aged 67 years or by adolescents aged 1314 years. It was found that the prevalence of asthma was 8.8% among children aged 67 and 17.44% among children aged 1314. It was found that there is a strong positive relationship between the exposure to certain pollutants and the symptoms associated with asthma. As an example, the association between the higher levels of Carbon monoxide (CO) and the ever wheezing in adolescents (1314 years old) was significant (OR = 1.84; 95% CI: 1.053.25). Increased frequency of wheezing attacks (4 to 12 episodes) was linked to sulfur dioxide (SO_2) (OR = 1.39; 95% CI: 1.04–1.91) and fine particulate matter ($PM_{2.5}$) concentrations (OR=1.38; 95% CI: 1.05–1.98 for ages 6–7, and OR = 1.13; 95% CI: 0.98–1.39 for ages 13–14). Also, the presence of sleep disturbances at least once a week in younger children (67 years) were significantly associated with the level of Nitrogen dioxide (NO_2). The research found that there is a huge interaction between environmental noise and PM_{10} levels, which means that the combined exposure to both factors increases the symptoms of asthma. In general, the findings indicate that not only air pollution but also noise are contributing to the further increase in the prevalence of asthma in children in Tehran, and that the combination of the two is further worsening the situation. Subsequently, environmental pollution and noise containment are crucial actions that would ensure the minimization of the symptoms of asthma. This work has found that asthma can be managed differently using machine learning algorithms and geographical analysis as the world becomes more urbanized and industrialized leading to high pollution levels in different regions [25]. In other words, GIS, RS, and machine learning provide a holistic model of asthma's spatial and environmental aspects [26]. These technologies can affect the planning of the cities, provide information for health treatments, and, in the long run, assist in decreasing morbidity and mortality rates of asthma by predicting and explaining the environmental factors of the disease [27]. With developments in real-time data and model predictions, this merged approach can open new avenues to more preventive asthma care management with critical relevance for global health, especially in asthmatic susceptible cities [28-29].

Geographic Information Systems (GIS) have been widely used to map spatial patterns of asthma and its link to environmental factors in recent studies. Novel research in Israel combined spatial databases with childhood asthma occurrences, and determined air pollution and vegetation as crucial factors. Other GIS-aided research in New York City and Karachi also indicated strong positive links between asthma and air pollution, specifically $PM_{2.5}$ but also negative links with vegetation. Succeeding studies validated that pollutants like PM and CO_2 are reliable input features of asthma occurrence, and mapping techniques have been successful in mapping high-risk areas and urban hotspots. Public health related studies also confirm strong relationships between indoor and outdoor air pollution and asthma. The pollutants $PM_{2.5}$, PM_{10}, NO_2, SO_2 and O_3, especially those associated with the vehicle emissions, have been found to raise asthma risk in various age groups. Chronic exposure, especially in early life, mostly increases the risk. Residential pollutants such as solid fuel burning and passive smoking also play a vital role. Additionally, latest studies on genetic factors and gene-environment interactions suggest that people have different levels of vulnerability to asthma stimulated by pollution, based on their biological attributes. Geographic Information System (GIS) methodologies are useful for mapping spatial patterns and location at risk, but are not significantly predictive. On the other hand, machine learning models, mainly ensemble models like Random Forest, have been shown to address complex relationships between environmental and demographic factors. Such models have shown great capability in predicting environmental patterns and can be applied to asthma. But there is a need for more hybrid prediction models that include air pollutants, environmental and demographic variables to improved predictive asthma severity, which this study focuses. This study develops a new data-driven model for asthma severity prediction that include air pollutant levels ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO), environmental conditions (temperature, humidity, wind speed) and individual features (age and sex).

This research enhances upon preceding works that have mainly specific on GIS-based spatial assessment and/or asthma intensity, by Emphasizing multi-class asthma severity prediction via machine learning. This Approach Investigates and Differentiate various models such as Logistic Regression, Random Forest and XGBoost to interpret the relationships between environmental features and asthma intensity. The initial purpose of this study is to create and examine machine learning models for estimation of asthma intensity levels using environmental and demographic features. Our analytical outcomes indicate that Logistic Regression is the best implementing model with an accuracy of 98% (Random Forest and XGBoost 89%). Moreover, correlation analysis presents a strong link between pollution index and asthma intensity, recommend the role of the environment in the advancement of asthma. The effect of this research is that it can support validation-based decision-making in healthcare, enabling early diagnosis of at-risk groups and novel techniques toward asthma handling. The implementation of a predictive model based on environmental data helps to improve public health monitoring in areas with pollution. But limitation such as class Disproportion and lack of evaluation of confounding variables (e.g., smoking, obesity) point to future research directions.

Materials and methods

The research was carried out in the state of Uttarakhand in India covering various districts that represented urban, semi-urban and rural regions. It combines environmental and demographic information in order to predict the

risk of asthma. Air quality data were acquired via the Uttarakhand Pollution Control Board, which includes major pollutants such as $PM_{2.5}$, PM1.0, SO_2, CO, NO_2 and O_3 in addition to meteorological parameters such as temperature, humidity and the speed of wind. These measurements were taken at hourly and daily intervals and summarized to make analysis. The demographic data were obtained in the form of census and public health records, which included demographic information such as age and gender. The data set is representative of district level spatial coverage, and reflects seasonal changes. Understanding of environmental exposure patterns was achieved by using descriptive statistics like mean, range, and variability of pollutant concentrations. On the whole, this combined data can be a good starting point to create machine learning models to determine the risky areas and populations at risk of asthma.

Data set description

For our research on asthma prediction, we have gathered extensive environmental and demographic data from numerous government sources across multiple districts in Uttarakhand. Fig. 1 describes the daily average levels of air pollutants including PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, and NO_2 and some meteorological conditions with examined threshold levels for respiratory health such as temperature, humidity, and wind speed. Furthermore, information including age and gender was collected to enable a low risk assessment of asthma in the community. Data regarding air quality and other environmental variables were collected from the Uttarakhand Pollution Control Board along with the required meteorological departments. These agencies provided the hourly and daily favoured measurements to enable comparison of fluctuating pollutant and weather variations over the time and across the seasons. The contingency of age and gender distributions was obtained from public health records coupled with census data to match these factors with health parameters pertaining to asthma cases. By constructing a rich dataset, we were able to provide the groundwork for machine learning model training and validation, with the goal of predicting asthma susceptibility in a variety of environments. The precision of this data, together with its large district-level coverage across Uttarakhand, improves the reliability and application of our findings in identifying asthma-prone locations and populations.

	PM2.5	PM10	NO2	SO2	O3	CO	Temperature	Humidity	Wind Speed	Age	Gender	Pollution Index	Asthma	Severity
0	81.162623	71.837220	29.862040	54.143537	90.079423	3.996992	25.930278	23.103956	28.810717	41	Male	55.180306	0	0
1	190.635718	171.732265	28.462986	63.937830	122.760526	4.787013	6.895454	34.941802	27.491320	65	Female	97.052723	0	1
2	149.078849	264.424834	91.094185	20.786964	116.422530	8.560019	34.895783	86.499664	3.830168	38	Male	108.394564	1	1
3	123.745112	225.022968	28.706889	50.365054	31.545987	3.466043	24.524650	81.341469	36.902896	74	Female	77.142009	0	0
4	39.643542	245.837121	30.835224	46.167933	30.894926	8.709532	6.288155	48.051415	22.738888	21	Male	67.014713	0	0

Fig. 1. Data set description

Machine learning algorythms

Machine learning methods for asthma prediction employ statistical and computational tools to examine complicated data sets containing environmental and demographic elements that influence asthma risk. In our research we have used 3 algorithms -

Logistic regression- Logistic regression is a type of statistical technique used in machine learning to predict one of two possible outcomes (e.g., "yes" or "no," "disease" or "no disease"). Logistic regression, unlike linear regression, predicts probabilities rather than continuous values. It applies the logistic function, often known as the sigmoid function, to map predictions to a probability scale ranging from 0 to 1. The model assigns the outcome to one of two categories based on a predetermined threshold (often 0.5). Logistic regression identifies the best-fitting line (or decision boundary) that separates the data points of the two classes by estimating coefficients for each feature in the dataset, making it frequently used in applications such as medical diagnosis, spam detection, and credit scoring [30-35].

Random Forest: Random Forest is an extension of the bagging model of ensemble learning systems that uses multiple decision trees in order to make better predictions on the results. They are all trained on a random sample of the data and make an independent prediction on the test data. In the case of classification, the Random Forest algorithm arrives at an aggregate conclusion by casting the votes and choosing the most common [36-39]. For regression problem it simply calculates average of all trees prediction. The influence of a least squares method is that it results in overfitting because each tree is trained to different sections of the data set, gather various pattern while cutting out outlier. As said earlier, Random Forest is widely used for disease prediction such as Asthma because of its ability to work well even when dealing with large data set and also because variation is reduced.

XGBoost: XGBoost or Extreme Gradient Boosting is the state-of-the-art machine learning algorithm from the gradient boosting family. This ensures that it supports high, light and fast model creating and is perfect for modeling tasks such as classification and regression [40-46]. XGBoost builds decision trees in teams in a sequent manner, which in turns tries to correct the errors of the previous tree. It also uses this repeated approach to develop an ability that helps it identify complex patterns within the data. Regularization techniques is used to curb overfitting, which boosts the algorithms ability to generalize. The speed with which XGBoost deals with very large data sets and its integrated cross validation have made it very popular in data science competitions and practical data analysis especially where the problem involves predicting outcomes given several characteristics of the data [47-48].

Proposed methodology

The primary purpose of this study is to assess asthma risk using various ensemble learning algorithms. The technique begins with collecting air quality and demographic data, which is then preprocessed to remove any attributes with missing values, errors, noise, or discrepancies. To construct a classifier model and evaluate its performance, we divide the improved dataset into two parts: a training set and a test set. Ensemble learning techniques are applied to improve asthma prediction accuracy by revealing underlying patterns in the data. The findings are then analyzed. Figure 1 displays the research framework, methods for forecasting asthma risk, and the method's

step-by-step procedure. Because the data came from various sources, the primary goal was to organize it in a way that could be analyzed with Python. This section concisely summarizes the steps outlined in the technique [49-53].

Preprocessing- Data Preprocessing scans the data for non-numeric values as well as the unknown values and negated that there were no abnormalities in the asthma data set. For example, any missing variable can be replaced with mean value. Standard normalization of the data can be done using standard normal equations [54-58]. This data was then pre-processed and cut into training and testing datasets.

Construct the prediction model - Section 3.2 described the Ensemble Learning prediction models, which used data from the previous stage. Ensemble learning models were used to predict asthma risk, accounting for $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO, temperature, humidity, wind speed, age, and gender [59-64].

Assessment of the prediction model: The expected outcomes from the preceding phase were compared to real asthma incidence data. The predictive model's performance was assessed using several metrics, including precision, recall, F1-score, support, and accuracy [65-66].

The accuracy and efficiency of the proposed machine learning models were measured using standard classification metrics, including accuracy, precision, recall, and F1-score. The input data was divided into training and testing sets using a hold-out validation technique to assess model generalization. Each model—Logistic Regression, Random Forest, and XGBoost—was trained on the training set and evaluated on the unobserved test data.

The final assessment of models was based initially on accuracy, supported by class-wise precision, recall, and F1-score to ensure stable assessment across different severity classes. Logistic Regression achieved the highest overall accuracy of 98%, while Random Forest and XGBoost achieved 89%. However, class-wise assessment showed that all models failed to predict the minority class due to severe class instability, which affects the reliability of the assessment.

The pair plot evaluation was conducted and demonstrated in Fig. 2 to examine the distribution and pairwise associations of air quality features across distinct severity levels. The diagonal concentration plots show that maximum pollutant features such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, and O_3 have significant overlap among severity classes, shows that their individual distributions are not appropriately separated to independently classify severity levels. Correspondingly, the off-diagonal scatter plots explain weak or complex relationships among major feature pairs, with data points from different classes showing highly combined and lacking clear separation margins. These results suggest that the dataset is irregularly separable and that individual features possess limited selective ability. Hence, severity prediction is likely dependent on the consolidated effect and interaction of different features rather than any single pollutant measure, Validating the need for multivariate and complex machine learning models for impactful classification.

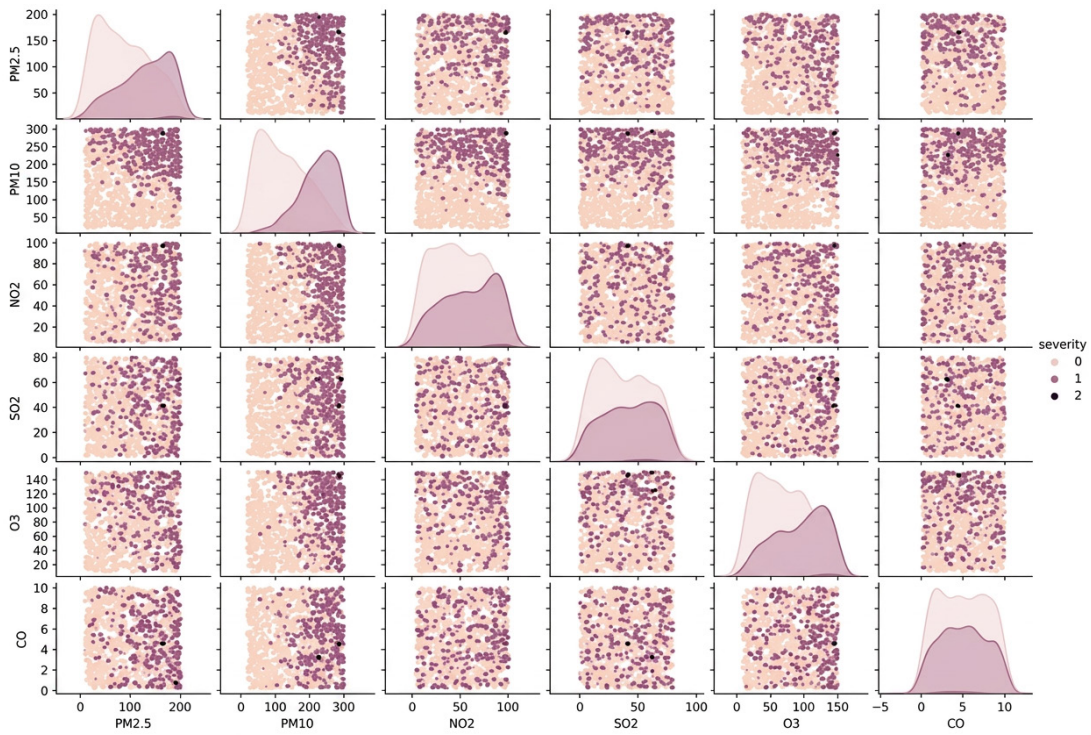

Fig. 2. Pairwise relationships among air pollutants (PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2) showing their distributions and inter-correlations

Fig. 3 demonstrates the correlation matrix explains the relationships between different air pollutants, environmental features, asthma, and severity levels. A strong positive correlation is monitored between Pollution Index and Severity (r=0.80), implying that overall pollution feature is the most impactful factor in evaluating severity. Among individual pollutants, PM_{10} (r=0.61) and $PM_{2.5}$ (r = 0.39) show moderate to strong relationship with severity, followed by O_3 (r=0.31), recommending their considerable contribution to unfavorable outcomes. In contrast, variables such as CO, temperature, humidity, wind speed, age, and gender exhibit negligible correlations ($r \approx 0$), show limited direct influence. Moreover, asthma shows a moderate correlation with severity (r =0.40) and with Pollution Index (r=0.53), indicating that individuals with asthma are more likely to face higher severity under elevated pollution vulnerability. It is important to differentiate that asthma is a input features, whereas severity denotes the output features indicating the intensity of health impact. Overall, the analysis emphasis that pollution-related features, usually composite indices and particulate matter, perform a leading role in severity prediction.

Fig. 3. Correlation matrix illustrating the relationships between air pollutants, environmental factors, health indicators (asthma), and severity levels

Fig. 4 consists of three graphs of different air pollutant characteristics ($PM_{2.5}$, PM_{10}, NO_2, SO_2, etc). These plots are used in EDA to visualize the mean, median, mode, standard deviation, normality and data spread. Here's a breakdown of the graphs:

Density plot with histogram (Leftmost column) – This indicates the density of each feature. The horizontal bar (blue bars) shows the distribution of the data in the form of frequency. The density curve shown by the blue line smooths it by giving the points' density. The value obtained shows where the distribution lies, if it is normal (close to zero), positively skewed (>) or negatively skewed (<). Example: The results show that $PM_{2.5}$ has a skewness of 0.04, which suggests that the data is normally distributed.

Boxplot (Middle column) displays data distribution, including the median, the range of values, and other observations of data outside the range of interquartile range limits. The box represents the interquartile range, the difference between the first quartile, Q1 and the third quartile, or Q3. The one drawn in the line in the box is the median. Outliers have data points beyond 1.5×IQR; whiskers extend to show them. Any point outside this range is regarded as an outlier. Example: A few more features could be identified as having outliers, but none of the selected features seem to exhibit these strongly.

Probability Plot (Q-Q Plot) (Rightmost column) Checks if the feature follows a normal distribution. The x-axis represents theoretical quantiles from a normal distribution, while the y-axis represents observed quantiles. Points falling close to the red diagonal line indicate normality.

We also observe reasonably low skewness for all the features, suggesting that the features are symmetrically distributed. No feature has it shown that is particularly above or below the mean or significantly spiked.

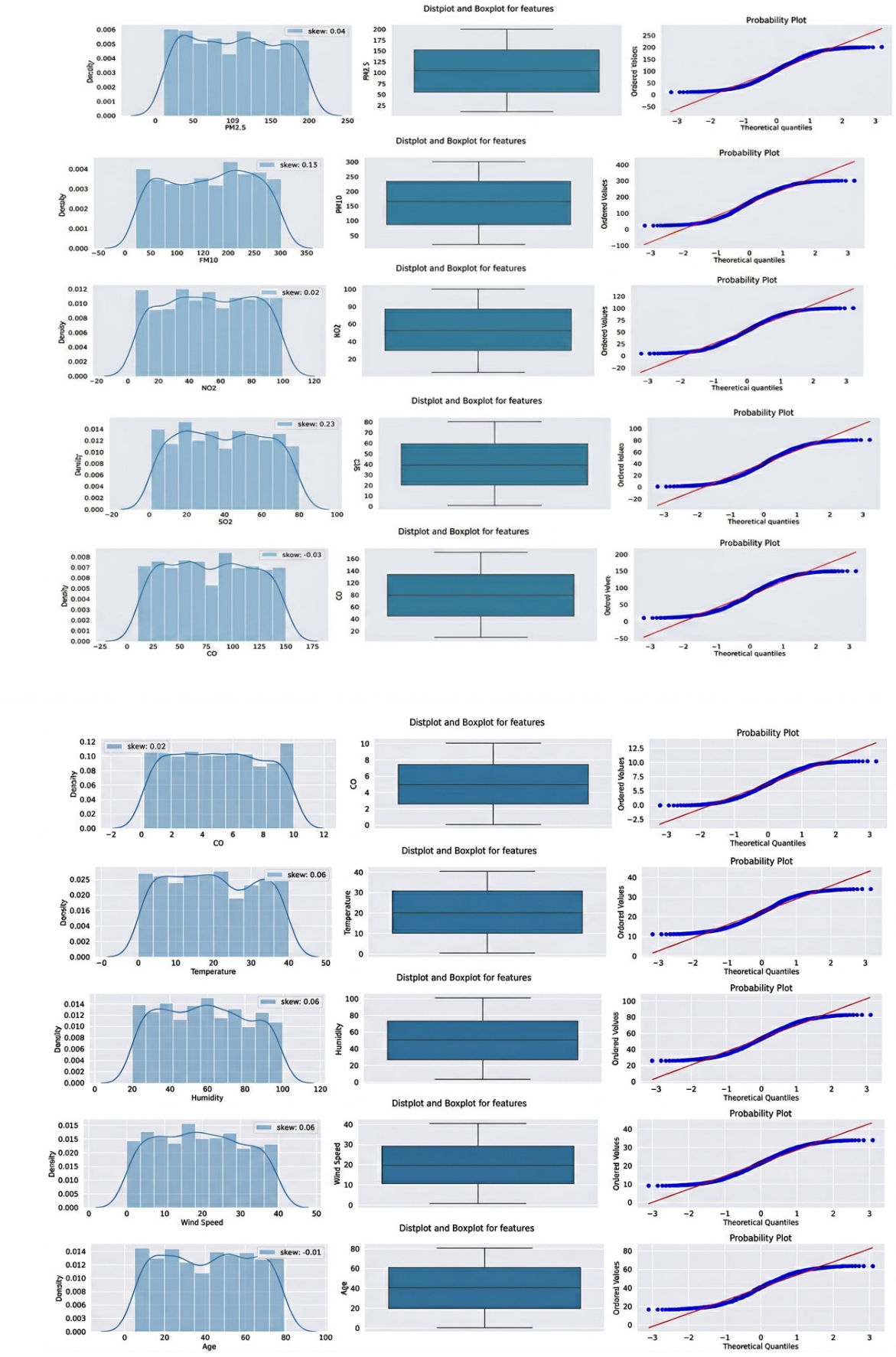

Fig. 4. Feature Analysis to understand distribution

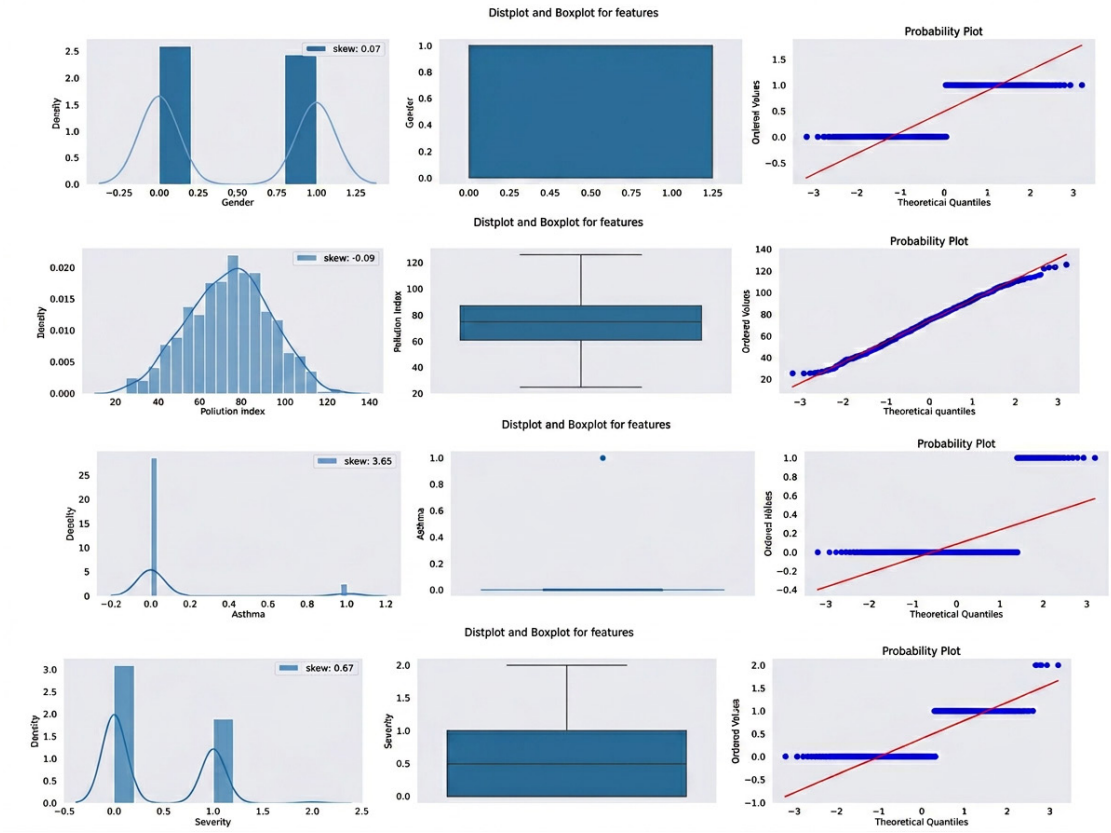

Fig. 4. Continued

Results and discussion

In the above section, Figs. 1-4 illustrate descriptive statistics, data visualization, and correlation. The following step includes a detailed data analysis of how air pollution and environmental factors affect asthma. We used Jupyter Notebook to develop Machine Learning models. The data set is separated into training and testing segments using a 0.7 split factor. 70% of the data is required for model training, with the remaining 30% for testing. We created the model using logistic regression, random forest, and XGBoost machine learning approaches. Random Forest and XGBoost obtained 89% and 89% accuracy, respectively, whereas Logistic Regression achieved 98% accuracy on our dataset. Figs. 5, 6, and 7 show the confusion matrix and the classification reports for Logistic Regression, Random Forest and XGBoost, respectively.

The confusion matrix examines the model's performance across asthma severity classes, highlighting areas with accurate and inaccurate predictions. The categorization report assesses the model's accuracy, recall, and F1 scores at each level of severity. The linear regression, Random Forest, and XGBoost models are all excellent at predicting asthma risk based on air pollutants and environmental factors, with linear regression performing particularly well. The slight discrepancy in accuracy between the three models reveals variances in their predictive capabilities, which can be investigated further via feature significance analysis and hyperparameter adjustment. The successful application of machine learning algorithms for asthma risk prediction has far-reaching consequences for healthcare and policy. Accurately predicting asthma risk from pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO) and environmental factors (temperature, humidity, wind speed) allows health agencies and policymakers to improve preventative measures, allocate resources efficiently, and reduce health risks from air pollution.

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.99	0.98	0.99	202
1	0.95	0.98	0.96	97
2	0.00	0.00	0.00	1
accuracy			0.98	300
macro avg	0.65	0.65	0.65	300
weighted avg	0.97	0.98	0.98	300

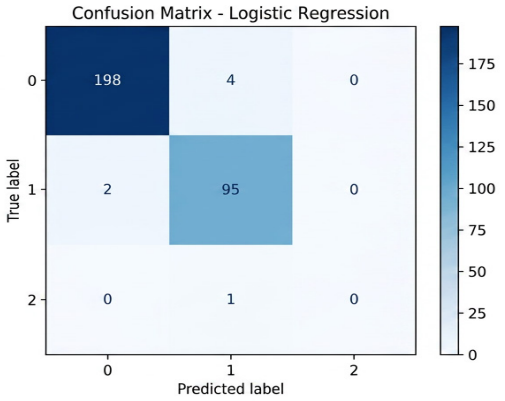

Fig. 5. Classification report and confusion matrix of logistic regression

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.92	0.92	0.92	202
1	0.82	0.82	0.82	97
2	0.00	0.00	0.00	1
accuracy			0.89	300
macro avg	0.58	0.58	0.58	300
weighted avg	0.88	0.89	0.89	300

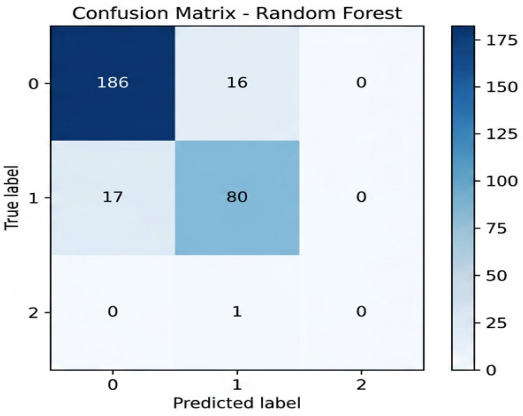

Fig. 6. Classification report and confusion matrix of random forest

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.93	0.92	0.92	202
1	0.82	0.86	0.84	97
2	0.00	0.00	0.00	1
accuracy			0.89	300
macro avg	0.58	0.59	0.59	300
weighted avg	0.89	0.89	0.89	300

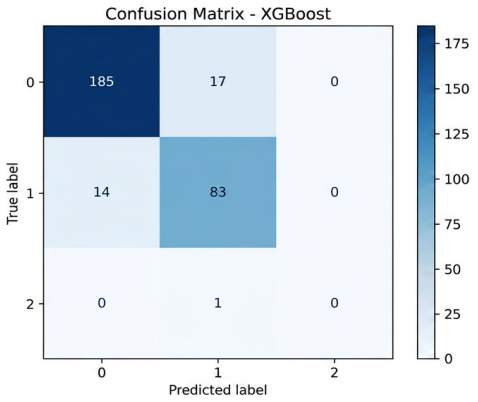

Fig. 7. Classification report and confusion matrix of XGBoost

Our data set has three different target classes. Table 1 describes the comparison of Logistic Regression, Random Forest and XGBoost.

As for the results produced by the models developed in this work, Graphs of bars and line plots have been used as the Python visualization technique required to analyze trends in data pertaining to investigations of Asthma and factors in air quality and the environment.

Table 1. Comparison of proposed models

	class	Precision	Recall	F1-score	Support
Logistic Regression	0	0.99	0.98	0.99	202
	1	0.95	0.98	0.96	97
	2	0.00	0.00	0.00	1
ACCURACY- 98%					
	class	**Precision**	**Recall**	**F1-score**	**Support**
Random Forest	0	0.92	0.92	0.92	202
	1	0.82	0.82	0.82	97
	2	0.00	0.00	0.00	1
ACCURACY- 89%					
	class	**Precision**	**Recall**	**F1-score**	**Support**
XGBoost	0	0.93	0.92	0.92	202
	1	0.82	0.86	0.84	97
	2	0.00	0.00	0.00	1
ACCURACY- 89%					

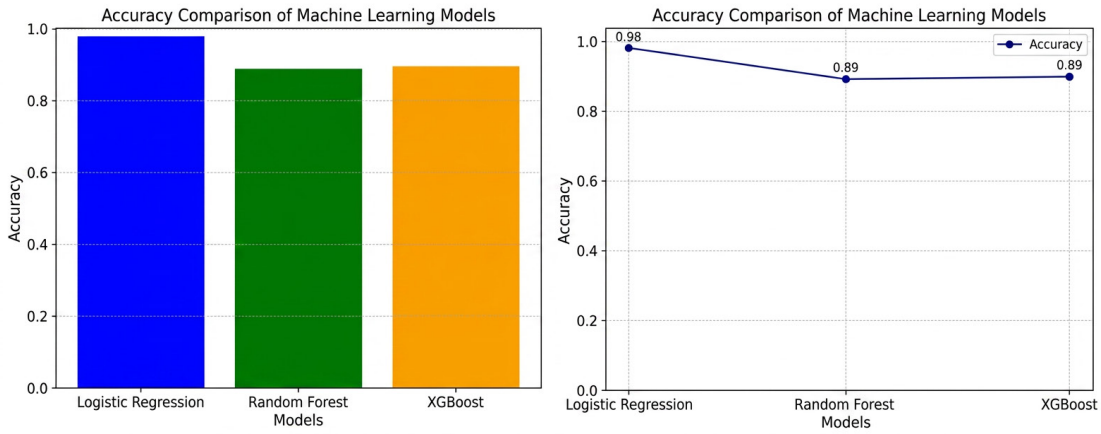

Fig. 8. Accuracy comparison of machine learning models

Conclusion

This research explains the capability of machine learning approaches in predicting asthma severity based on air pollution and environmental features. Logistic Regression, Random Forest, and XGBoost models were executed on a preprocessed dataset, attaining classification accuracies of 98%, 89%, and 89%, respectively, with Logistic Regression illustrates relatively excellent performance. The evaluation metrics, incorporating confusion matrices and classification reports, shows that the models are effective in differentiating between severity classes, although performance may be impacted by class imbalance. The results highlight the notable contribution of key pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, and CO, along with environmental features like temperature, humidity, and wind speed, in determining asthma severity levels.

Though, it is essential to note that asthma severity is impacted by multiple features beyond environmental exposure. Critical features such as obesity, genetic predisposition, prior medical history of asthma, and medication usage were not incorporated in the present research, which may act as confounding features and affect the predictive capability of the models. Moreover, an important weakness of this research is the elimination of smoking-related features, containing smoking history, intensity, and passive exposure, due to the lack of availability of data. Smoking is a well-known confounding features in asthma-associated research and may considerably affect both the occurrence and severity of the disease. This missing feature may affect the reliability and applicability of the model outcomes.

Future research should concentrate on integrating these clinical and lifestyle features, managing class imbalance, and examining advanced modeling approaches. Moreover, combination of real-time air quality and meteorological features could elevate the applicability of the model for early warning systems and public health decision-making. With enhanced availability of data and model optimization, these techniques can contribute to more accurate evaluation and management of asthma severity across assorted populations.

Financial supports

The authors declare that no financial support,

funding, grant, or sponsorship was received for conducting or completing this research work.

Competing interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgements

The authors would like to express their sincere gratitude to Graphic Era Hill University, Bhimtal Campus, for providing the resources and academic environment necessary for completing this research work.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely ob-served by the authors.

References

1. Network TGA. The global asthma report 2022. Int J Tuberc Lung Dis. 2022;26:1–104.

2. Pereira AA, Pollard SL, Locke R, Romero K, Lima JJ, Hansel NN, et al. Association between exhaled carbon monoxide and asthma outcomes in Peruvian children. Respir Med. 2018;145:212–216.

3. Zhang L, Yi H, Sang N. Sulfur dioxide-induced exacerbation of airway inflammation via reactive oxygen species production and the TLR4/NF-κB pathway in asthmatic mice. Toxicol Ind Health. 2021;37:564–572.

4. Soares AR, Silva C. Review of ground-level ozone impact in respiratory health deterioration for the past two decades. Atmosphere. 2022;13:434.

5. Gryech I, Ghogho M, Mahraoui C, Kobbane A. An exploration of features impacting respiratory diseases in urban areas. Int J Environ Res Public Health. 2022;19:3095.

6. Nishida C, Yatera K. The impact of ambient environmental and occupational pollution on respiratory diseases. Int J Environ Res Public Health. 2022;19:2788.

7. Czechowski PO, Piksa K, Dąbrowiecki P, Oniszczuk-Jastrząbek AI, Czermański E, Owczarek T, et al. Financing costs and health effects of air pollution in the tri-city agglomeration. Front Public Health. 2022;10:831312.

8. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet. 2014;383:1581–1592.

9. Rodrigo GJ, Rodrigo C, Hall JB. Acute asthma in adults: A review. Chest. 2004;125:1081–1102.

10. Brusselle GG, Koppelman GH. Biologic therapies for severe asthma. N Engl J Med. 2022;386:157–171.

11. Lambrecht BN, Hammad H. The immunology of asthma. Nat Immunol. 2015;16:45–56.

12. Tiotiu AI, Novakova P, Nedeva D, Chong-Neto HJ, Novakova S, Steiropoulos P, et al. Impact of air pollution on asthma outcomes. Int J Environ Res Public Health. 2020;17:6212.

13. Cheng CY, Tseng YL, Huang KC, Chiu IM, Pan HY, Cheng FJ. Association between ambient air pollution and emergency visits for respiratory diseases. Toxics. 2022;10:247.

14. Tornevi A, Olstrup H, Forsberg B. Short-term associations between PM_{10} and respiratory health effects. Toxics. 2022;10:333.

15. Chittrakul J, Sapbamrer R, Sirikul W. Insecticide exposure and risk of asthmatic symptoms: A systematic review. Toxics. 2021;9:228.

16. Hwang H, Jang JH, Lee E, Park HS, Lee JY. Prediction of the number of asthma patients using

environmental factors based on deep learning algorithms. Respir Res. 2023;24(1):302.

17. Lin CC, Chiu CC, Lee PY, Chen KJ, He CX, Hsu SK, et al. The adverse effects of air pollution on the eye: A review. Int J Environ Res Public Health. 2022;19:1186.

18. Gaffin JM, Hauptman M, Petty CR, Sheehan WJ, Lai PS, Wolfson JM, et al. Nitrogen dioxide exposure in school classrooms of inner-city children with asthma. J Allergy Clin Immunol. 2018;141:2249–2255.e2.

19. Razavi-Termeh SV, Sadeghi-Niaraki A, Choi SM. Asthma-prone areas modeling using a machine learning model. Sci Rep. 2021;11(1):1912.

20. Hehua Z, Qing C, Shanyan G, Qijun W, Yuhong Z. Prenatal air pollution exposure and childhood asthma. Environ Res. 2017;159:519–530.

21. Yan W, Wang X, Dong T, Sun M, Zhang M, Fang K, et al. Prenatal $PM_{2.5}$ exposure and childhood asthma. Environ Sci Pollut Res Int. 2020.

22. Lee S, Ku H, Hyun C, Lee M. Machine learning-based analyses of the effects of various types of air pollutants on hospital visits by asthma patients. Toxics. 2022;10(11):644.

23. Faraji, M., Mohammadi, A., Najmi, M. et al. Exposure to ambient air pollution and prevalence of asthma in adults. Air Qual Atmos Health 2021;14, 1211–1219. https://doi.org/10.1007/s11869-021-01011-z

24. Faraji M, Najmi M, Kazemnejad A, Shokouhi Shoormasti R, Fazlollahi MR, Pourpak Z, Moin M. Effect of Air Pollutants and Environmental Noise on the Childhood Asthma Prevalence in Tehran, Iran. Iran J Allergy Asthma Immunol. 2024 Oct 6;23(5):489-501. doi: 10.18502/ijaai.v23i5.16745. PMID: 39586743.

25. Thurston GD, Balmes JR, Garcia E, Gilliland FD, Rice MB, Schikowski T, et al. Outdoor air pollution and airway disease. Ann Am Thorac Soc. 2020;17:387–398.

26. Lin W, Brunekreef B, Gehring U. Indoor NO_2 and asthma in children. Int J Epidemiol. 2013;42:1724–1737.

27. Hüls A, Vanker A, Gray D, Koen N, MacIsaac JL, Lin DTS, et al. Genetic susceptibility and air pollution. Eur Respir J. 2020;55:1901831.

28. Kim KH, Jahan SA, Kabir E. A review on human health perspective of air pollution with respect to allergies and asthma. Environ Int. 2013;59:41–52.

29. Norbäck D, Lu C, Zhang Y, Li B, Zhao Z, Huang C, et al. Sources of indoor particulate matter and outdoor air pollution in China in relation to asthma and allergies. Environ Int. 2019;125:252–260.

30. Brooks CC, Martin LJ, Pilipenko V, He H, LeMasters GK, Lockey JE, et al. NAT1 genetic variation increases asthma risk. J Asthma. 2019.

31. Silvestri M, Franchi S, Pistorio A, Petecchia L, Rusconi F. Smoke exposure and asthma development: A meta-analysis. Pediatr Pulmonol. 2015;50:353–362.

32. Burke H, Leonardi-Bee J, Hashim A, Pine-Abata H, Chen Y, Cook DG, et al. Prenatal and passive smoke exposure and asthma incidence. Pediatrics. 2012;129:735–744.

33. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA methylation and maternal smoking. Am J Hum Genet. 2016;98:680–696.

34. Neophytou AM, Oh SS, Hu D, Huntsman S, Eng C, Rodríguez-Santana JR, et al. In utero tobacco exposure and asthma. Environ Epidemiol. 2019;3:e048.

35. Magnus MC, Håberg SE, Karlstad Ø, Nafstad P, London SJ, Nystad W. Grandmother smoking and asthma risk. Thorax. 2015;70:237–243.

36. Accordini S, Calciano L, Johannessen A, Portas L, Benediktsdóttir B, Bertelsen RJ, et al. Three-generation study on smoking and asthma. Int J Epidemiol. 2018;47:1106–1117.

37. Wu CC, Hsu TY, Chang JC, Ou CY, Kuo HC, Liu CA, et al. Paternal smoking and offspring asthma. Front Genet. 2019;10:471.

38. Chen Y, Dales R, Krewski D, Breithaupt K. Increased effects of smoking and obesity on asthma among female Canadians. Am J Epidemiol. 1999;150:255–262.

39. Torén K, Olin AC, Hellgren J, Hermansson BA. Rhinitis increases the risk for adult-onset asthma. Respir Med. 2002;96:635–641.

40. Piipari R, Jaakkola JJK, Jaakkola N, Jaakkola MS. Smoking and asthma in adults. Eur Respir J. 2004;24:734–739.

41. Plaschke PP, Janson C, Norrman E, Björnsson E, Ellbjär S, Järvholm B. Onset and remission of allergic rhinitis and asthma. Am J Respir Crit Care Med. 2000;162:920–924.

42. Polosa R, Knoke JD, Russo C, Piccillo G, Caponnetto P, Sarvà M, et al. Cigarette smoking and risk of incident asthma. J Allergy Clin Immunol. 2008;121:1428–1434.

43. Vignoud L, Pin I, Boudier A, Pison C, Nadif R, Le Moual N, et al. Smoking and asthma: A longitudinal approach. Respir Med. 2011;105:1805–1811.

44. Torén K, Hermansson BA. Incidence of adult-onset asthma in relation to smoking. Int J Tuberc Lung Dis. 1999;3:192–197.

45. Verlato G, Nguyen G, Marchetti P, Accordini S, Marcon A, Marconcini R, et al. Smoking and new-onset asthma in adults. Int Arch Allergy Immunol. 2016;170:149–157.

46. Huovinen E, Kaprio J, Koskenvuo M. Lifestyle factors and risk of adult-onset asthma. Respir Med. 2003;97:273–280.

47. Vesterinen E, Kaprio J, Koskenvuo M. Asthma in relation to smoking habits. Thorax. 1988;43:534–539.

48. Becklake MR, Lalloo U. The "healthy smoker" phenomenon. Respiration. 1990;57:137–144.

49. Peled R, et al. Defining localities of inadequate asthma treatment: A GIS approach. Int J Health Geogr. 2006;5:3.

50. Maantay J. Asthma and air pollution in the Bronx. Health Place. 2007;13:32–56.

51. Khan IA, Arsalan MH, Siddiqui MF, Zeeshan S, Shaukat SS. Spatial association of asthma and vegetation. Pak J Bot. 2010;42:3547–3554.

52. Gorai AK, Tuluri F, Tchounwou PB. GIS-based assessment of air pollution and asthma. Int J Environ Res Public Health. 2014;11:4845–4869.

53. Chang TS, et al. Sparse modeling of environmental variables and asthma. J Biomed Inform. 2015;53:320–329.

54. Skarková P, et al. Asthma prevalence and environmental factors using GIS. Cent Eur J Public Health. 2015;23:258.

55. Douglas JA, Archer RS, Alexander SE. Environmental determinants of respiratory health. Prev Med Rep. 2019;14:100855.

56. Bontinck A, Maes T, Joos G. Asthma and air pollution: Pathogenesis insights. Curr Opin Pulm Med. 2020;26:10–19.

57. Heijink I, van Oosterhout A, Kliphuis N, Jonker M, Hoffmann R, Telenga E, et al. Corticosteroid unresponsiveness in asthma. Thorax. 2014;69:5–13.

58. Esposito S, Tenconi R, Lelii M, Preti V, Nazzari E, Consolo S, et al. Air pollution and asthma in children. BMC Pulm Med. 2014;14:31.

59. Sompornrattanaphan M, Thongngarm T, Ratanawatkul P, Wongsa C, Swigris JJ. Particulate matter and respiratory allergy. Asian Pac J Allergy

Immunol. 2020;38:19–28.

60. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and health impacts of air pollution. Front Public Health. 2020;8:14.

61. Obeidat M, Li X, Burgess S, Zhou G, Fishbane N, Hansel NN, et al. Surfactant protein D and COPD risk. Eur Respir J. 2017;50:1700657.

62. Kim BG, Lee PH, Lee SH, Park CS, Jang AS. Impact of ozone on lung function. Environ Toxicol. 2018;33:798–806.

63. Achakulwisut P, Brauer M, Hystad P, Anenberg SC. NO_2 pollution and pediatric asthma burden. Lancet Planet Health. 2019;3:e166–e178.

64. Kelly FJ, Fussell JC. Toxicity of particulate matter. Atmos Environ. 2012;60:504–526.

65. Jiménez-Ruiz CA, Andreas S, Lewis KE, Tonnesen P, van Schayck CP, Hajek P, et al. Smoking cessation in pulmonary diseases. Eur Respir J. 2015;46:61–79.

66. Pant J, Joshi MC, Singh D, Pant HK, Bhatt A, Pant D. Ensemble learning for soil fertility evaluation. Procedia Comput Sci. 2024;235:1998–2008.