International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 6, Nov – Dec 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

DATA-DRIVEN PREDICTION OF RESPIRATORY HEALTH RISKS USING AIR QUALITY AND METEOROLOGICAL FACTORS: AN EXPLAINABLE MACHINE LEARNING APPROACH

Kavya C*, Sowmya G**, Harshini T***
*(M.Sc. Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College Coimbatore,
Email : kavyachinnaduraikavya@gmail.com)
** (M.Sc. Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College Coimbatore,
Email : gsowmi75@gmail.com)
*** (M.Sc. Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College Coimbatore,
Email : harshinithangaraj11@gmail.com)

----------**********----------

Abstract:

Air pollution is an environmental problem that can cause bad breathing problems. When people breathe in dirt particles and gas pollutants and when the weather changes it can make breathing problems happen and get worse. This paper shows a way to use data and clear machine learning to guess the risk of breathing problems from air quality and weather data. The method looks at dust particles like PM2.5 and PM10 gases like nitrogen dioxide, sulfur dioxide, carbon monoxide and ozone. It also uses weather numbers such as temperature, humidity, wind speed, rain and air pressure when those numbers are available. The steps are: get the data clean the data look at the data pick pieces of data build a model test the model and explain the model. The paper tries models such as Logistic Regression, Decision Tree, Random Forest and XGBoost to predict risk. The paper checks how good the models are with accuracy, precision, recall, F1-score and a confusion matrix, for the test. The paper also uses SHAP, a method that shows how each air quality and weather factor pushes the model's guess. This way the method can help people watch the air warn early and teach the public. It does not replace a doctor's diagnosis. A second file gives a Python program that does everything from getting the data to drawing SHAP explanations.

Keywords — Air Quality, Respiratory Health, Machine Learning, Meteorological Factors AI

----------**********----------

I. *INTRODUCTION*

Air pollution is an important environmental and public health concern because people are exposed to a mixture of particulate and gaseous pollutants in outdoor environments. The World Health Organization identifies particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide as key pollutants for which health-based air-quality guidelines are provided. Fine particulate matter is particularly important because small particles can penetrate deeply into the respiratory system and are associated with respiratory and cardiovascular effects

I think respiratory health is affected by pollution exposure and by conditions. Pollutants can come from transportation, industry, burning fuel, construction, farming and other places. Weather conditions like temperature, humidity, wind speed, atmospheric pressure and rain can change how pollutants spread build up and how people are exposed. Therefore looking at air-quality numbers may not fully show the environmental conditions that relate to respiratory health risk.

Traditional statistical analysis can help find links. Environmental and health data might have nonlinear connections and interactions between many variables. Machine learning methods can capture patterns in many-dimensional data sets and can therefore serve as prediction tools. Earlier research has used methods such as boosting, Gaussian process regression, Random Forest, XGBoost and others to predict respiratory health and

 ©IJSRED: All Rights are Reserved

air quality, with climate and pollution data [2–5] [10–15]**.**

A major limitation of machine learning models is interpretability.A machine learning model may give a prediction while giving little information about why the prediction was made.Explainable Artificial Intelligence (XAI) methods can address this problem.SHAP assigns contribution values to features and can be used to examine feature importance and the contribution of features to individual predictions.Recent health and air-quality studies have used SHAP to identify important pollutants and meteorological variables [2–5] [10].

We propose an integrated framework that combines data-driven analysis, machine learning prediction and explainability.

We aim to analyze air-quality and meteorological patterns identify predictors compare machine learning models for respiratory-risk prediction and use SHAP to interpret the contribution of environmental variables.A companion document implements the pipeline in Python to demonstrate reproducibility.The proposed work is intended for research and early-risk assessment, than clinical **diagnosis.2.**

II. LITERATURE SURVEY / REVIEW

Research on air pollution and respiratory health has shifted from statistical connections to using data for predictions. The WHO air-quality guidelines present evidence about major pollutants and highlight how important PM2.5, PM10, O3, NO2, SO2 and CO are for health [1]. This gives a health-focused base for choosing variables in prediction studies.

Ku et al. Created machine learning models to predict the number of people with respiratory diseases by using weather and air-pollution factors in Seoul, South Korea [2]. The study tested gradient boosting and Gaussian process regression used a relief-based method to select features and used SHAP to understand the models. Their findings showed that lower temperatures, daylight and less humidity along with pollutants like SO2, CO and PM2.5 were linked to more respiratory disease cases. This shows the importance of mixing weather and pollution data.

Ma et al. Looked into machine learning that's easy to understand for predicting visits for respiratory diseases, in Lanzhou, China using daily weather and air-quality data [3]. The study compared Generalized Additive Model, Random Forest and XGBoost methods. Used SHAP to explain the results. The researchers found that looking at data from three days before gave the performance. They also found that O3, minimum temperature, maximum temperature and NO2 were factors.

I read a Scientific Reports study conducted in Penang, Malaysia. The study looked at how respiratory-disease happens by using environmental factors and machine learning. Decision Tree, Random Forest and XGBoost were compared with feature-selection approaches. The study found that Random Forest combined with a genetic-algorithm feature-selection approach worked well. Random Forest highlighted SO2, CO, temperature, PM10, wind speed, O3 and rainfall as important variables. SHAP helped make these results easier to understand.

Other recent work also uses machine learning for air pollution and respiratory-health prediction on larger geographic scales. These studies show that PM2.5 can be the predictor in some places. When key pollutant variables are missing, meteorological and socioeconomic variables can become more important. Feature importance can also change across income levels and regions. These findings show how important dataset composition, location, time resolution and feature availability are.

Several studies from India support these findings. Kunjir and colleagues used MODIS Fine Aerosol Optical Depth and ERA5 meteorological reanalysis data with Random Forest regression to estimate concentrations across Maharashtra. They found that Random Forest outperformed linear, linear and lasso regression baselines. They also found that humidity, rainfall and wind speed were negatively correlated with PM2.5. Kumar and colleagues monitored emissions from biomass cooking in rural Rajasthan and Haryana. They combined linear Random Forest and XGBoost regression with variables to show how exposure changes with seasons. Singh and Suthar compared linear, ensemble and deep-learning approaches—including Random Forest, KNN, GRU and CNN—, for predicting PM2.5 in Jaipur.

 ©IJSRED: All Rights are Reserved

Gaikar and others used Random Forest, Support Vector Regression and Artificial Neural Networks to predict the Air Quality Index in India. They found that Random Forest worked best for helping people make decisions about air quality.

Beyond predicting air quality explainable machine learning has been applied directly to conditions. Mahmood and co-workers used AutoML, SHAP and LIME to build a model that can be understood by doctors. The model used age, gender, medical history and breathing test results to predict asthma outcomes.

Atzeni and colleagues created a machine-learning framework that used air-quality monitors and lifestyle data. They used SHAP to explain the predictions. They found that when people are exposed to pollution and have had symptoms before the system can predict short-term flare-ups in chronic obstructive pulmonary disease for some patients. Although these studies show that machine learning and explainable methods are useful the prediction performance and feature importance can change a lot across locations and datasets. A model trained in one region is not automatically able to work in another region. The present study therefore stresses a workflow with a ready-made Python pipeline, in which local or selected air-quality and meteorological data are cleaned, analysed, modelled evaluated and interpreted. The research gap is not the lack of studies on this relationship but the need, for a transparent location-aware, explainable and reproducibly implemented analysis that looks at pollutant and meteorological variables together.

III. METHODOLOGY / FINDINGS

3.1 Proposed Framework

The proposed framework has seven steps: data collection, preprocessing, exploratory data analysis, feature selection, model development, performance evaluation and explainability. The whole process is set up so that the same cleaned dataset and evaluation method are used for all the machine learning models being tested.

Air-quality and weather data are first gathered from a public or official source, like the Central Pollution Control Board (CPCB) air-quality data website and the India Meteorological Department (IMD) or ERA5 reanalysis data. Depending on what's available the dataset might include PM2.5, PM10, CO, NO2, SO2, O3, AQI, temperature, relative humidity, wind speed, rainfall and atmospheric pressure. A respiratory-health target is needed for learning. The target could be a health result or a specific risk group that is created using a clear and scientifically supported method. The target shoul.d not be given as a diagnosis.

3.2 Data Preprocessing

Data preprocessing is done to make sure the data is good and works together. All the same records are taken out. The missing parts are looked at. If there are numbers that are missing they can be filled in using math methods or if it makes sense the whole part with too many missing pieces can be removed. Any strange numbers that don't fit are checked using math and charts. The dates and times are made to follow a way so that patterns over days, weeks or other times can be seen. If there are categories they are turned into numbers. Some methods need numbers to be in the size range, like Logistic Regression so they are adjusted.. Models that are based on trees usually don't need this step.

3.3 Exploratory Data Analysis

Exploratory data analysis helps us understand how pollutants and weather-related variables are distributed and how they change over time. We calculate numbers like the average middle value, spread, lowest and highest values. Line charts show how things change from one time to another. Bar charts help us compare pollution levels across different times or places. Scatter plots let us see how pollutants relate to the thing we are trying to predict. A correlation heatmap shows which variables are connected and if there might be much overlap, between them. These steps help us choose the features and better understand the model we build.

3.4 Feature Selection

Feature selection is performed to keep variables that hold useful prediction information and to cut out extra duplicate data. Feature selection involves looking at candidate features such as PM2.5, PM10, CO, NO2, SO2, O3, temperature, humidity, wind speed, rainfall and pressure. Feature selection can use correlation analysis, domain knowledge,

 ©IJSRED: All Rights are Reserved

permutation importance or wrapper-based methods to decide which variables are best. Feature selection must be done carefully so that information, from the test set does not leak into training the model.

3.5 Risk Classification

For a classification-based implementation respiratory health risk can be represented using three categories: Low, Moderate and High. The definition of these classes must be explicitly. Should be based on the available health outcome variable or a validated risk-labeling procedure. If an actual respiratory-health outcome is available the target should be derived from that outcome than arbitrarily assigning risk categories from the same predictors. This distinction is important because directly converting AQI into a health risk label can create circularity when AQI is also used as an input feature.

3.6 Machine Learning Models

Four candidate models are proposed. Logistic Regression offers an interpretable baseline for classification tasks. Decision Tree delivers rule-based predictions that're easy to understand and visualize. Random Forest builds on decision trees and can capture nonlinear relationships while also giving feature-importance measures [9]. XGBoost uses a gradient-boosting method that's effective, at modeling complex interactions and nonlinear patterns [8]. Each model is trained on the training set. Then evaluated on a separate test set. Cross-validation can be applied within the training data to help select the model and tune hyperparameters.

3.7 Model Evaluation

In a classification problem accuracy alone is not enough when class distributions are imbalanced. Model evaluation must look beyond accuracy. Therefore precision, recall, F1-score and a confusion matrix must be reported. Macro-averaged and weighted metrics can be added to give a view across risk classes. If the study uses regression to predict the number of cases instead of classification suitable metrics are R^2, MAE and RMSE. The final model should be chosen based on the evaluation design and not just on training performance.

3.8 Explainable Machine Learning Using SHAP

SHAP (SHapley Additive exPlanations) is used to interpret the selected model [6]. Global SHAP analysis ranks variables based on their impact on model predictions. A SHAP summary plot displays both how important each feature is and whether it increases or decreases the prediction. Local SHAP explanations, similar in idea to -agnostic methods, like LIME [7] help explain a prediction by showing which environmental factors raised or lowered the predicted risk. This makes the model more transparent even if it otherwise acts like a box.

3.9 Expected Analytical Outcomes

The research and the new way of looking at things show that predicting the risk of breathing problems probably depends on things together like tiny particles in the air gases in the air and weather conditions. It's not about one single thing. Earlier studies have found that PM2.5 SO2, CO, O3, NO2, temperature, humidity, wind and rain might all be important to predict. How much each one matters can be different depending on where you are and what data you have [2–5] [10–12].

A key thing to expect is that models based on trees like Random Forest and XGBoost might be able to see patterns and connections that simpler models like linear ones can't. This has been seen in studies about predicting PM2.5 levels in areas, where Random Forest did better than other models like linear or lasso regression [10].. The best model needs to be found from the real data not just guessed before. SHAP analysis is expected to explain which environmental factors are most important, in the predictions. The order of these factors can help tell which ones are important in the area being studied.

This paper does not report any accuracy, precision, recall, F1-score or feature-ranking result as an experimental finding until the proposed models are trained and tested on the final dataset. The numerical results section of the conference submission should be filled by running the pipeline described in the companion implementation document on the collected dataset. This prevents the presentation of assumed or fabricated performance values, as research findings.

IV. CONCLUSION

 ©IJSRED: All Rights are Reserved

The paper does not include any figures relating to accuracy, precision, recall, F1-score or feature ranking as experimental results until the models suggested in the paper have been trained and tested on the final dataset. The numerical results section of the conference submission must be completed by carrying out the pipeline described in the companion implementation document on the collected dataset. This action avoids the inclusion of assumed or faked performance values, since such values would not be genuine research findings.

It is important to use explainable machine learning since merely making a prediction does not show why a particular risk level was arrived at. SHAP is capable of providing both global and individual explanations and can therefore be used to determine the relative contribution of pollutants and meteorological factors. Earlier studies have shown that such methods are useful for predicting respiratory health and air quality [2–5], [10–14].

The framework could be extended by including longer time periods, various locations, lagged environmental exposures, demographic variables, and other validated health outcomes. Future research could also look into model calibration, external validation, temporal forecasting, and the deployment of the established pipeline via an environmental health dashboard, for example by using a lightweight web framework such as Streamlit. The system should be regarded as a research and early-warning support tool and not as a substitute for medical diagnosis or professional clinical decision-making.

V. REFERENCES

[1] World Health Organization, “WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide,” World Health Organization, Geneva, 2021.

[2] Y. Ku, S. Bin Kwon, J.-H. Yoon, S.-K. Mun, and M. Chang, “Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors,” Clinical and Experimental Otorhinolaryngology, vol. 15, no. 2, pp. 168–176, 2022. DOI: 10.21053/ceo.2021.01536.

[3] Y. Ma, W. Wang, Z. Liu, Y. Zhao, Z. Wan, P. Qin, and B. Cheng, “Interpretable machine learning prediction of visits for respiratory diseases in Lanzhou, China: impact of air pollution and meteorological variables,” Urban Climate, 2025. DOI: 10.1016/j.uclim.2025.102463.

[4] “Prediction of the respiratory disease incidence based on environmental factors using machine learning techniques in Penang, Malaysia,” Scientific Reports, 2026. DOI: 10.1038/s41598-026-48762-2.

[5] M. A. Hassan Abadi and S. Wang, “Interpretable Machine Learning for Air Pollution and Respiratory Health Prediction: A Socioeconomic Subgroup Analysis,” arXiv preprint, 2026. DOI: 10.48550/arXiv.2607.17024.

[6] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[8] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[9] L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.

[10] G. M. Kunjir, S. Tikle, S. Das, M. Karim, S. K. Roy, and U. Chatterjee, “Assessing particulate matter (PM2.5) concentrations and variability across Maharashtra using satellite data and machine learning techniques,” Discover Sustainability, vol. 6, art. 238, 2025. DOI: 10.1007/s43621-025-01082-3.

[11] P. Kumar, E. Upadhyay, and A. Yadav, “Spatiotemporal assessment and machine learning-based prediction of PM2.5 emissions from biomass combustion in Rural India,” Bulletin of Atmospheric Science and Technology, 2026.

[12] S. Singh and G. Suthar, “Machine learning and deep learning approaches for PM2.5 prediction: a study on urban air quality in Jaipur, India,” Institutional Digital Repository, IIT (BHU) Varanasi, 2025.

[13] S. Mahmood, R. Hasan, S. Hussain, and R. Adhikari, “An interpretable and generalizable machine learning model for predicting asthma outcomes: integrating AutoML and explainable AI techniques,” World, vol. 6, no. 1, p. 15, 2025.

[14] M. Atzeni, G. Cappon, and M. Vettoretti, “A machine learning framework for short-term prediction of chronic obstructive pulmonary disease exacerbations using personal air quality monitors and lifestyle data,” 2025.

[15] D. Gaikar, U. Patel, O. Vispute, S. Singh, and T. Sanghvi, “Prediction of Air Quality Index using Random Forest Algorithm,” International Research Journal of Engineering and Technology (IRJET), vol. 10, no. 4, 2023.

 ©IJSRED: All Rights are Reserved