Science of the Total Environment 858 (2023) 159509

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections

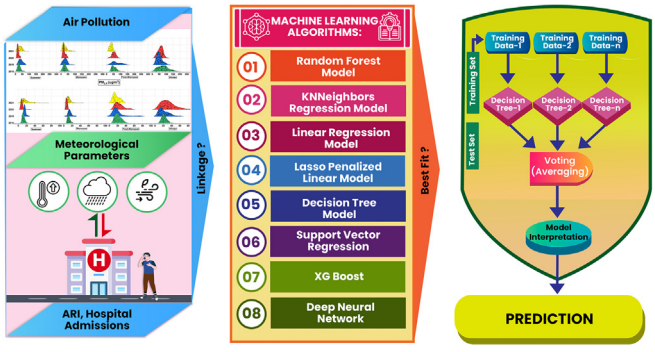
Khaiwal Ravindra [a,*], Samsher Singh Bahadur [a], Varun Katoch [a,b], Sanjeev Bhardwaj [a], Maninder Kaur-Sidhu [a], Madhu Gupta [a], Suman Mor [b]

[a] Department of Community Medicine & School of Public Health, PGIMER, Chandigarh 160012, India
[b] Department of Environment Studies, Panjab University, Chandigarh 160014, India

HIGHLIGHTS

- Machine learning approaches were applied to study outpatient visits & air pollution.
- 8-Machine learning models with & without lagged effect of exposure were compared.
- Total patient visits with 1-day lag strongly correlate with gaseous air pollutants.
- The random Forest regression model provides the best R^2 on trained data.
- We recommend long-term data with a large sample size to establish a better link.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Pavlos Kassomenos

Keywords:
Machine learning programs
Air pollution
ARI
Random forest regression
Risk prediction

ABSTRACT

With a remarkable increase in industrialization among fast-developing countries, air pollution is rising at an alarming rate and has become a public health concern. The study aims to examine the effect of air pollution on patient's hospital visits for respiratory diseases, particularly Acute Respiratory Infections (ARI). Outpatient hospital visits, air pollution and meteorological parameters were collected from March 2018 to October 2021. Eight machine learning algorithms (Random Forest model, K-Nearest Neighbors regression model, Linear regression model, LASSO regression model, Decision Tree Regressor, Support Vector Regression, X.G. Boost and Deep Neural Network with 5-layers) were applied for the analysis of daily air pollutants and outpatient visits for ARI. The evaluation was done by using 5-cross-fold confirmations. The data was randomly divided into test and training data sets at a scale of 1:2, respectively. Results show that among the studied eight machine learning models, the Random Forest model has given the best performance with R^2 = 0.606, 0.608 without lag and 1-day lag respectively on ARI patients and R^2 = 0.872, 0.871 without lag and 1-day lag respectively on total patients. All eight models did not perform well with the lag effect on the ARI patient dataset but performed better on the total patient dataset. Thus, the study did not find any significant association between ARI patients and ambient air pollution due to the intermittent availability of data during the COVID-19 period. This study gives insight into developing machine learning programs for risk prediction that can be used to predict analytics for several other diseases apart from ARI, such as heart disease and other respiratory diseases.

1. Introduction

There is increasing epidemiological evidence relating exposure to air pollution with a rise in respiratory diseases (Landrigan et al.,

* Corresponding author at: Department of Community Medicine and School of Public Health, Post Graduate Institute of Medical Education and Research, Chandigarh 160012, India.
E-mail addresses: khaiwal.ravindra@pgimer.edu.in khaiwal@yahoo.com (K. Ravindra).

http://dx.doi.org/10.1016/j.scitotenv.2022.159509
Received 9 February 2022; Received in revised form 13 September 2022; Accepted 13 October 2022
Available online 17 October 2022

2018, Schraufnagel et al., 2019). Around 90 % of the world population is exposed to very high concentrations of a pollutant that exceed the safe level considered by the U.S. Environmental Protection Agency (US EPA). Air pollution, specifically $PM_{2.5}$, is an essential parameter for air quality regulatory purposes because of its pathogenicity. There are strong evidence that greater exposure to air pollutants can have multiple adverse impacts on the human body (Ravindra et al., 2021a, 2021b).

Several studies have been conducted to study health care visits and their association with ambient air pollution. Air pollutants are mainly linked with an elevated risk of respiratory and cardiovascular ailments (Zhang et al., 2016; Sarizadeh et al., 2020; Brook et al., 2010; Romieu et al., 2002). In general, even in the areas where air quality is better, the consequences and outcomes of air pollution can't be ignored. Comparable studies can yield different results due to differences in air pollution concentration, air pollution impacts on varying age groups, and sensitivity between different regions (Huang et al., 2017; Liu et al., 2019). Associating the air pollutants with data from other areas is incorrect; therefore, studies are required to find the association between the impact of air pollution and patients visiting hospitals in that particular area to better understand the regional consequences (Zhang et al., 2019; Darrow et al., 2014; Kelly et al., 2011).

Air quality modeling plays an essential role in providing information and supporting the decision-making process (Amuthadevi et al., 2021; Lu et al., 2021). However, one air pollution model is insufficient; thus, considering the synergetic effects, multi-pollutant models are often employed to find the relationship between patients' hospital visits and air pollutants. Jiang et al. (2020) conducted a study in China using air quality modeling tools to evaluate the synergetic effects between air pollution and meteorology and its association with hospitalized patients with impaired blood circulation and respiration (Wang et al., 2019; Künzli et al., 2010). Numerous studies have been conducted to predict mortality and morbidity. Still, limited literature is available on the applicability of machine learning models for predicting patients' hospital visits due to respiratory diseases attributable to air pollution. Recently, studies have modeled exposure-lag response association on health outcomes, allowing to better understand the spread of pollutants over multiple days. Thus, lagged effect of exposure was considered in the current study. Although, the degree of prediction accuracy of various machine learning models is ambiguous (Lu et al., 2021).

Few attempts have been made in the past to explore the feasibility of machine learning methods using the Random Forest model, the K-Nearest Neighbors regressor model, the LASSO model, the Support Vector Machine, etc. The method most commonly used for evaluating the machine learning model includes the 3-fold cross-validation technique for predicting the relationship between the number of patients visiting hospitals and its relation to air pollution (Lu et al., 2021; Liao et al., 2021). Cheng et al. (2020) have shown that the random forest regression model outperformed the other models and gave the best accuracy ($R^2 = 0.969$) for linking conjunctivitis with air pollution. Similarly, Kannan and Vasanthi (2019) also developed machine learning algorithms for predicting heart diseases. The study suggests that logistic regression gave the best performance at an accuracy of 87 % compared to other models.

Concerning health studies based on air pollution, there are limited pieces of evidence for computing the feasibility of machine learning models to evaluate the association between 'outpatients' visits due to Acute Respiratory Infection (ARI) and air pollution (Host et al., 2008; Medina-Ramon et al., 2006). In this context, the proposed study was conducted for detailed knowledge about the impact of air pollution on ARIs patients in a Northern Indian city. The present study aims to explore the predictability of eight machine learning models, namely, the Random Forest model, the K-Nearest Neighbors regressor model, the Regression model, the LASSO model, the Decision Tree, Support Vector Machine, X.G. Boost, and Deep Neural Network with 5-layers for predicting the impact of air pollution on outpatient visits for ARI.

2. Methodology

2.1. Study area

The study was conducted in Chandigarh, a beautiful city located in northern India, where the air quality is better than in the other northern states of India (Ravindra et al., 2019). Meteorology and air pollution data were compared for the four seasons: Winter (Jan-Feb), Summer (Mar-May), Monsoon (Jun-Sept) and Post-monsoon season (Oct-Dec) listed, according to the Indian Meteorological Department (IMD).

2.2. Study data on population exposure

In the National Ambient Air Quality Monitoring program, the Central Pollution Control Board (MoEFCC, India) has set up various continuous Ambient Air Quality Monitoring Stations (CAAQMS) at several places in India (CPCB, 2014). The secondary data about primary air pollutants was obtained from the near real-time CAAQMS situated in Sector 25, Chandigarh, India. The daily air pollutant monitoring data for pollutants over three consecutive years was collected from March 2018 to October 2021 from the air pollution monitoring station, as shown in Fig. 1. Apart from air pollutants, data on meteorological parameters was obtained from IMD.

2.3. Study data on hospital visits

The data collected from the nearest CAAQMS to the Health Wellness Centre (HWC) represented the air quality of that particular area. The daily outpatient visits for ARI were collected from HWC sector-49, Chandigarh, for three consecutive years, i.e., 2018 to 2021, as shown in Fig. 1. These records contained the number of patients visiting daily in the OPD, age, patient name, address and diagnosis of the diseases. ARI diagnosis was classified according to the International Classification of Diseases (ICD-10, Ministry of Health Statistical Information Centre, 2016). The HWC data was also computed to obtain information on the total number of ARI patients.

2.4. Statistical analysis technique

Data collection from HWC was entered in Microsoft Excel and then data analysis was done using the statistical software Python (version 3.2). Eight different machine learning programs were applied to train the retrieval model to investigate the prediction and association of air pollution and hospital peak visits. These include the Decision Tree, Deep Neural Network with 5-layers, K-Nearest Neighbors regressor model, LASSO model, Random Forest model, Regression model, Support Vector Machine, and X.G. Boost. In the individual case, the results were also evaluated using five cross-fold confirmations. The data was randomly divided into test and training data sets at a scale of 1:2, correspondingly.

The prediction accuracy of the listed eight models was assessed based on the coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Pearson correlation (r). R-squared (R^2) is a statistical measure that indicates the fraction of a dependent variable's variance explained by independent variables in a regression model. The MAE quantifies the average magnitude of forecast errors without accounting for their direction. It evaluates the precision of continuous variables. The MSE is the difference between the model's predictions. The actual data is squared and then averaged across the entire dataset. The MSE is an excellent tool to ensure that the trained model does not produce predictions with large outlier errors, as the MSE places a greater emphasis on these errors due to its squared nature. RMSE is the quantification of the standard deviation of the residuals. Residuals show the distance between the data points from the regression line and indicate the goodness of fit. RSME is mainly used to verify the experimental results (Kenney and Keeping, 1962; Barnston, 1992). To assess the correlation between the health impacts of air pollution on patient hospital visits, the Pearson coefficient evaluated the dependency between the population

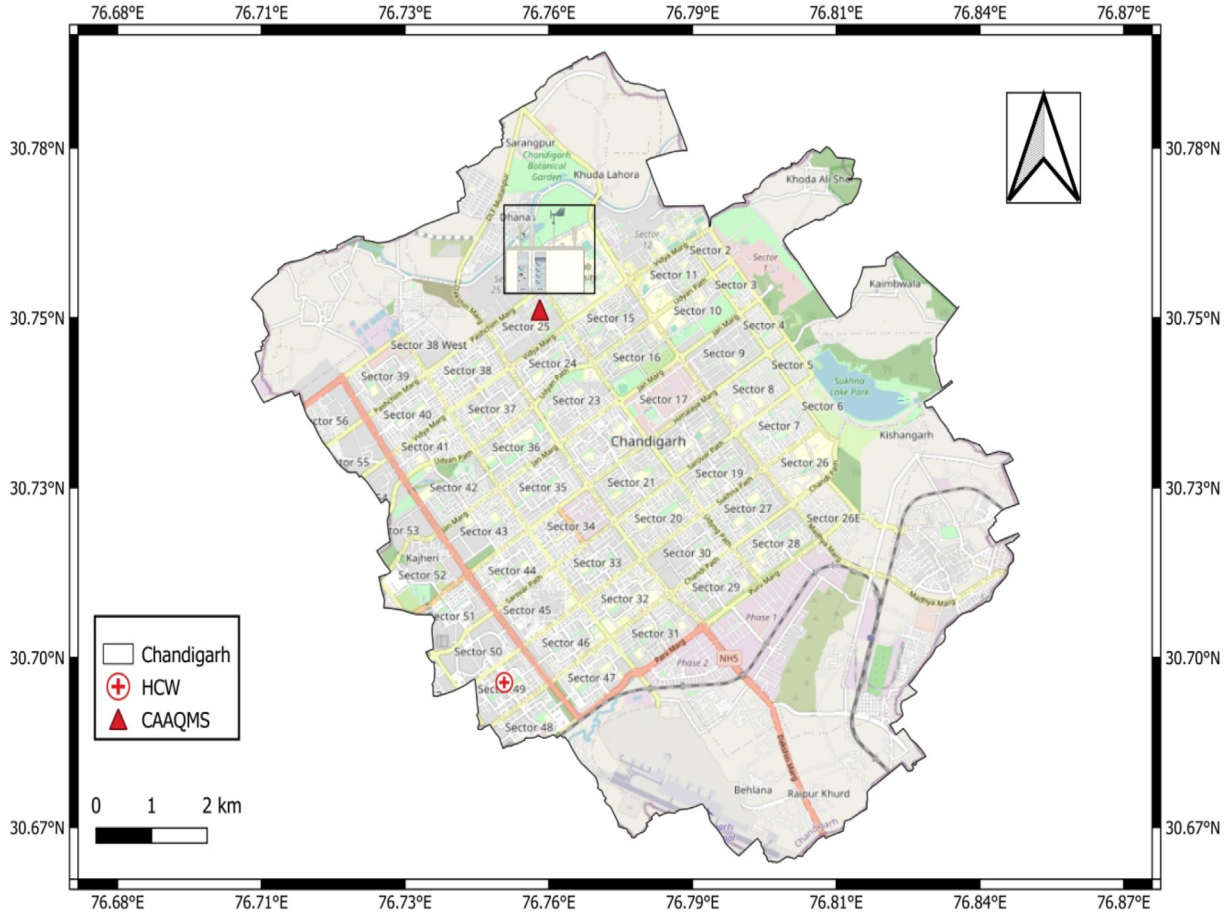


Fig. 1. Map depicting the CAAQM station location in Chandigarh, India.

exposure to 18 pollutant parameters and hospital admissions of ARI. Graphs were obtained using OriginPro software (Version 2020).

3. Machine learning methods

Eight typical machine learning methods, including Decision Tree, Deep Neural Network with 5-layers, K-Nearest Neighbors regressor model, LASSO model, Random Forest model, Regression model, Support Vector Machine, and X.G. Boost, are described in detail in this section. Random Forest and K-Nearest Neighbors algorithm is a supervised machine learning technique for classification and regression. Fig. 2 presents the variable significance measurement (Variable Importance) for the Random Forest Model. Multiple Linear Regression (MLR) is a critical regression approach for simulating the linear relationship between a particular continuous dependent variable and multiple independent variables. MLR is a technique for forecasting a quantitative reaction based on a combination of features as described by the following equation:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta nXn + e$$

LASSO (Least Absolute Shrinkage and Selection Operator) model is to determine the variables and regression coefficients that construct the model with the smallest prediction error. This was done by imposing a constraint on the model parameters, which "shrinks" the regression coefficients close to zero by initiating the total of the absolute value of the regression coefficients to be lesser than a predetermined number. This constrains the model's complexity and as a result, after shrinkage, variables with a regression coefficient of zero are eliminated from the model. Automated k-fold cross-validation is frequently used to determine the machine learning algorithm's performance on a dataset (Fig. 3, Fig. 4 and Fig. 5).

The Decision Tree model is a supervised machine learning method that manages both categorical and numerical data. This model's output contains horizontal and vertical line splits based on dependent variables' conditions. The dataset was analyzed in tree-shape format; its interior node represents the condition of dependent variables and the Leaf node on which the final analysis was carried out. The Support Vector Machine, a supervised learning algorithm, is used for both regression, classification, and continuous and categorical variables. This model assumes that the relationship between the independent and dependent variables is given by a deterministic function "'f' with additive noise. The model performs the next step to find a functional form for f that can correctly predict new cases that the Support Vector Machine has never accessed. The Support Vector Machine uses an iterative training approach to generate an ideal hyperplane. This algorithm is used to minimize an error function. Support Vector Machines models may be broken down into four unique categories according to the shape of the error function, which are as follows:

- Type 1 Regression SVM Regression (also known as epsilon-SVM regression)
- Type 2 Regression SVM Regression (also known as nu-SVM regression)

XGBoost is an abbreviation for eXtreme Gradient Boosting. It exemplifies the Gradient Boosting Machine (GBM), a technology primarily employed in building regression and classification predictive modeling issues. X.G. Boost is an ensemble method in which new models are generated to correct the residuals or errors of previous models. Deep learning is a kind of machine learning that is descended from the Artificial Neural Network (ANN). Several efforts are made to explore alternative architectures for patient prediction using deep learning, including a deep neural network, long and short-term memory (LSTM), and a convolutional neural network (CNN). Missing values is one of the most difficult challenges in deep

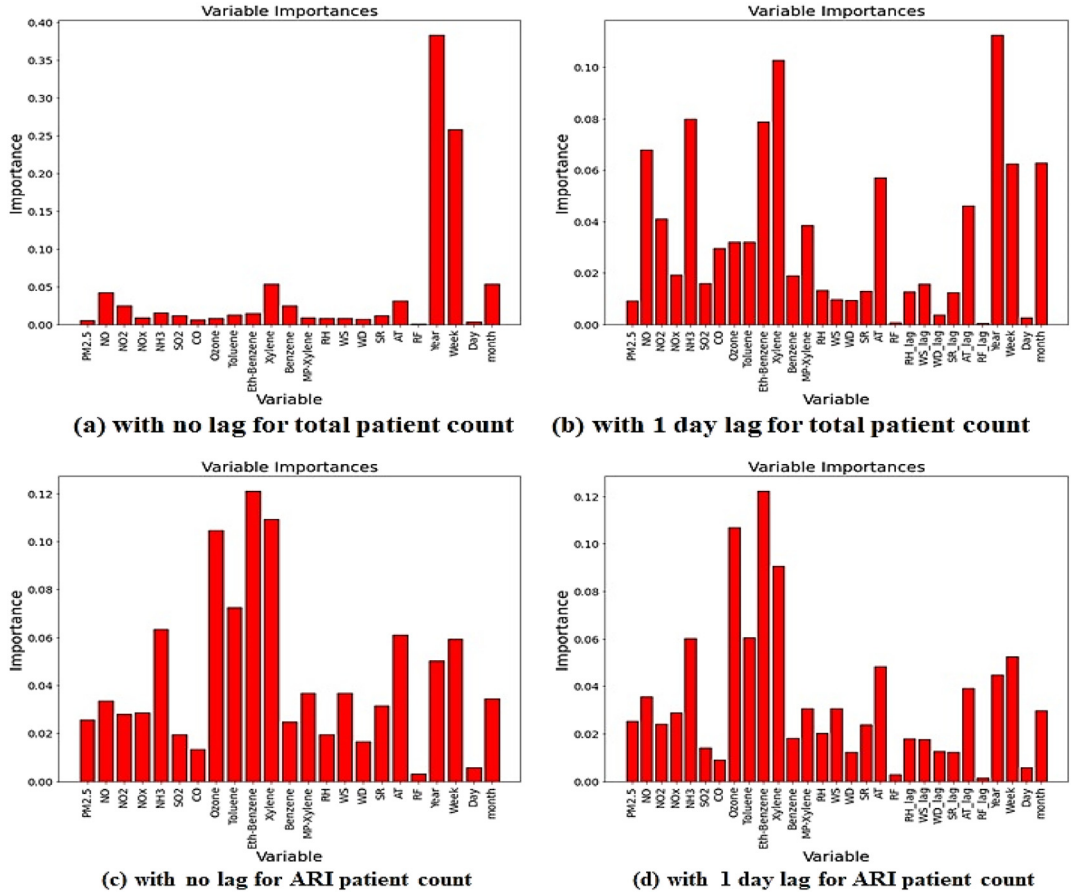


Fig. 2. Random Forest Model Variable Importance with no- and 1-day lag.

learning approaches. Because the number of missing data in our research region was significant, CNN and LSTM could not be applied. As a result, the researchers employed a five-layer neural network with a "relu" activation function and regularisation to minimize overfitting with an Adam optimizer. Regularizers.L2($1e^{-5}$) was added to the layers to prevent the model from over-fitting. The Deep Neural Network parameters are illustrated in Fig. S3, and the configuration of the deep learning layer is specified in Table S2.

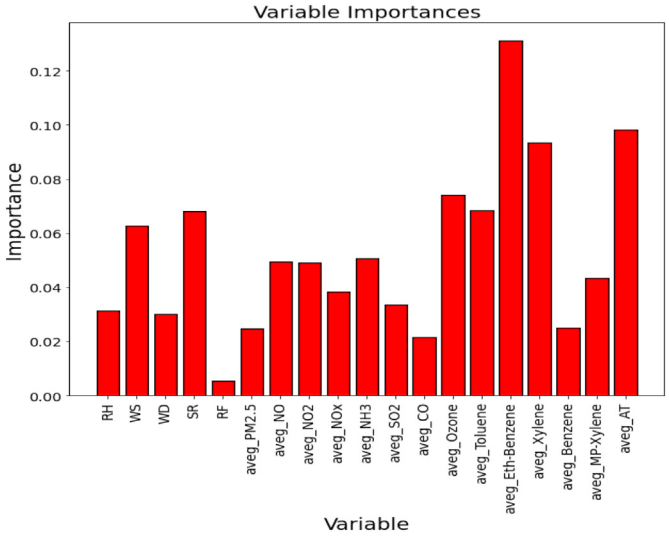


Fig. 3. Random Forest Model Variable Importance with air pollution 0–2 days moving average and temperature 0–21 days moving average.

Poisson regression is a regression analysis technique used to model discrete data and is a suitable method for forecasting non-negative counts. This technique assumes a conditional Poisson distribution for 'y', modeling the expected count logarithm as a linear function of the input variables. This formulation eliminates the possible problem of negative predictions for counts that can arise with least-squares linear regression and an assumption of Gaussian noise.

$$P(y_i = y|x_i'\beta) = \frac{\exp(-\exp(x_i'\beta)) \cdot \exp(x_i'\beta)^y}{y!}$$

The model parameters for Random Forest Model are given in Table 1, the K-Nearest Neighbors regressor model, the Regression model, the LASSO model, the Decision Tree, Support Vector Machine, X.G. Boost and Deep Neural Network are displayed in Table S1. These models compute air pollutant concentration data as independent variables, such as air pollution data collected from CAAQMS ($PM_{2.5}$, NO, NO_2, NOx, NH_3, SO_2, CO, Ozone, Toluene, Ethyl-Benzene, Xylene, Benzene, MP-Xylene) along with meteorological parameters (R.H., W.S., W.D., S.R., AT and R.F.), and to capture the time variation (Year, Week, Day, Month) was added.

Past studies have also suggested that the effect of air pollution exposure may be seen by conducting a multi-day moving average (Hassan, 2021). Therefore, moving average lag models (0–2 days moving average of air pollution and 0–21 days moving average of temperature) were also executed to thoroughly examine the lag effect of air pollution from the time series perspective.

For the construction of the models, Sklearn was used for constructing all models, and then grid search cv was used for hyperparameter tuning. The whole dataset was de-noted as D = {D1, D2, …, Dn}, in which n represents the length of the times series. For the i^{th} day, Di is de-noted as Di = {x1, i, x2, i, …, xm, i, yi}, in which xm, i is the average of the m^{th} influencing factor

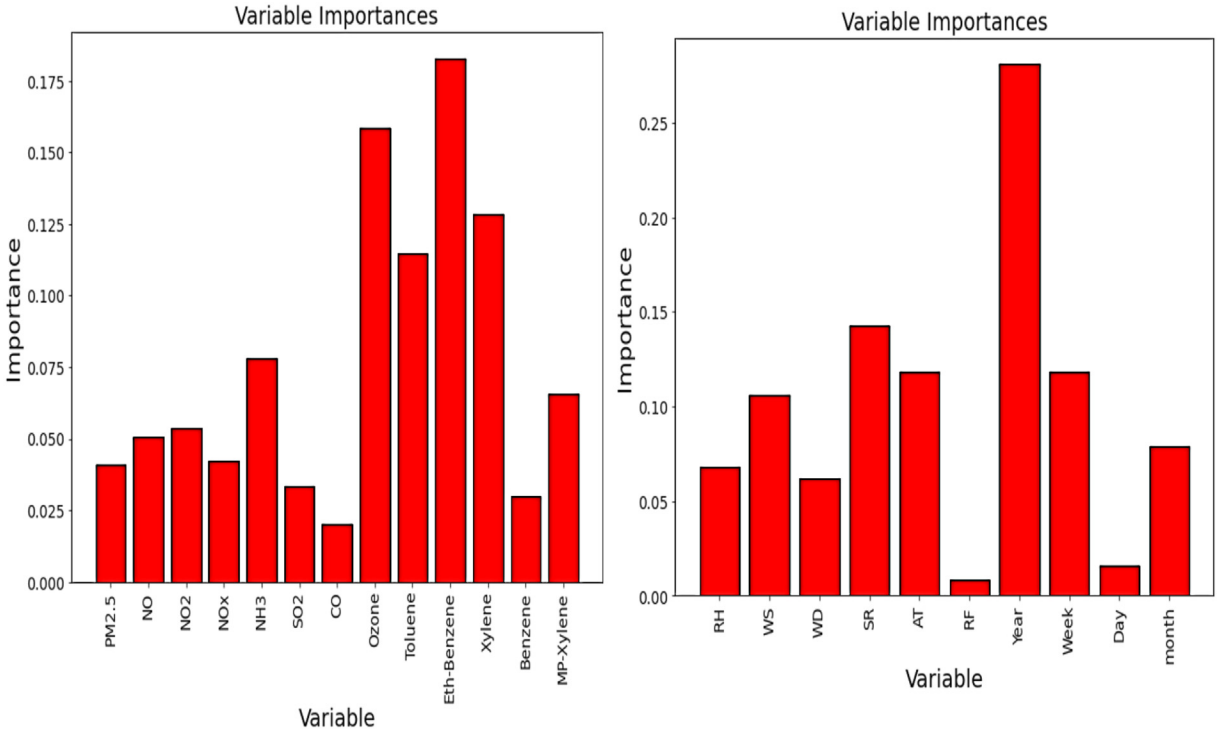


Fig. 4. Random Forest Model Variable Importance comparison between with air pollutants and without air pollutants.

on an i^{th} day, m represents the number of factors, and yi represents the total ARI patients. The study also took into consideration the lag effect; the predictions are made based on 1 to 5-day lag assumptions, naming as timestep t, which means the observed data of the previous several days are used to predict the patient number on a certain day, i.e., {Di − t, …, Di − 2, Di − 1} is used to predict {yi}. In addition, the whole dataset was divided into training and testing data, and the number of patients was predicted using eight different machine learning methods for inter-comparisons.

4. Results

4.1. Trends of air pollutants in Chandigarh

The observed value of both $PM_{2.5}$ and PM_{10} far exceeds the permissible WHO and NAAQ standards for most days. During the winter, elevated concentrations of particulate and gaseous pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NOx, NH_3, O_3, Benzene, Toluene, Ethyl-Benzene and Xylene) were

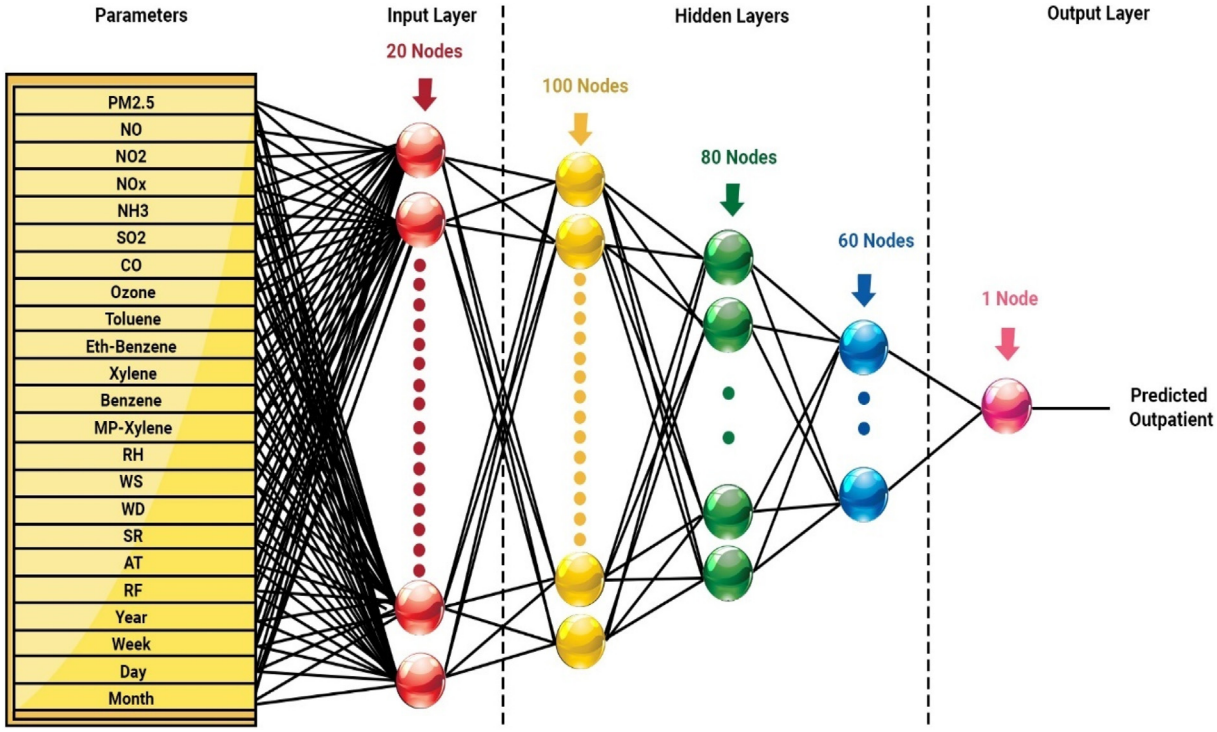


Fig. 5. Deep Neural Network parameters and approaches applied using air pollution and meteorological data.

Table 1
Model parameters.

Random forest model		
Parameter	Range	Optimum Value
bootstrap	[True, False]	true
max_depth	2–40	10
max_features	['auto', 'sqrt']	auto
min_samples_leaf	2–25	3
min_samples_split	2–25	16
n_estimators	100–200	126

observed, as shown in Fig. S1. The average C.O. and SO_2 concentrations were < 1 mg/m^3 and 8.4 μg/m^3, respectively, in Chandigarh from the year 2018 to the year 2021. Benzene concentration over the three studied years shows that the levels remain between 5 and 15 μg/m^3 except for a sudden increase observed during November 2019. An unexpected rise in the daily concentration of Ethyl Benzene was observed in February 2020 and the concentration of Toluene increased in June 2020. Some peaks in the concentration of Toluene were also observed in the months November 2019, May 2021 and September 2021. The concentration of Xylene and MP-Xylene, Benzene and Ethyl Benzene displayed a direct relationship showing a similar increasing and decreasing trend in winter and summer, respectively, over the years.

Unlike other pollutants, the daily concentration of O_3 fluctuated over the year and peak concentrations up to 65 μg/m^3 were observed during summers). The O_3 levels were lower in the winter compared to other seasons. The level of O_3 is significantly influenced by meteorological parameters such as temperature and wind speed. The data shows that O_3 concentrations increased with a rise in temperature during the summer months.

In contrast, NO_x concentration remains lower than 15 μg/m^3 from March 2020 to September 2020. A sudden rise in NO_x concentration was observed during the COVID-19 post-lockdown period, which seems to be strongly linked with increased vehicular emissions, as reported by Mor et al. (2021). NH_3 data was not available until September 2019. NH_3 concentrations also exhibit a similar trend as NO_x during the post-lockdown period. Pollutants concentration for $PM_{2.5}$ and NOx, as shown in Fig. 6, travel speedily upwards along the ridgeline in the winter season. This sudden rise in pollutant levels during winters may be associated with the burning of solid biomass fuels and the seasonal burning of crop residues being observed in the rural areas surrounding Chandigarh, which significantly deteriorates the air quality over the Indo-Gangetic plains (Biswal et al., 2020; Mor et al., 2021; Ravindra et al., 2020a, b, 2022a, b, c; Singh et al., 2021, 2020).

4.2. Trends in meteorological parameters

Chandigarh is located about ~350 m above the Mean Sea Level, and its climate is classified as Koeppen's CWG, which means it has a hot summer, cold winters, and tropical rains (Ravindra et al., 2020a, 2020b). Average annual rainfall is estimated to be 910 mm, with temperatures ranging from 2 °C–10 °C in winter and 38 °C - 43 °C in summers. Fig. S2 depicts the daily mean rainfall, air temperature, solar radiation, wind direction, wind speed and relative humidity. During the winter, winds prevail from the northwest to the southeast and southeast to northwest in the summer.

4.3. Association between air pollutants and ARI patient visits

Approximately 17,000 inhabitants reside in Sector −49, Chandigarh. The average number of patients who visited HWC-49 daily was 79 between March 2018 and April 2019. On average, 85 patients visited between April 2019 to March 2020 and 51 from April 2020 to March 2021. From 2018 to 2019, about a 7 % increase in ARI patients was observed against a 25 % decline in the total number of patients visiting OPD until 2020. This may be due to the COVID-19 lockdown-related restrictions. Even then, there was a 5 % increase in the number of ARI cases against the total number of patients visiting OPD from 2018 to 2019. The data on the total number of patients, who visited HWC-49 after 20th March 2020, shows the impact of restricted visits during the COVID-19 lockdown. An increase in the number of patients visiting OPD was witnessed after lifting the COVID-19 restrictions.

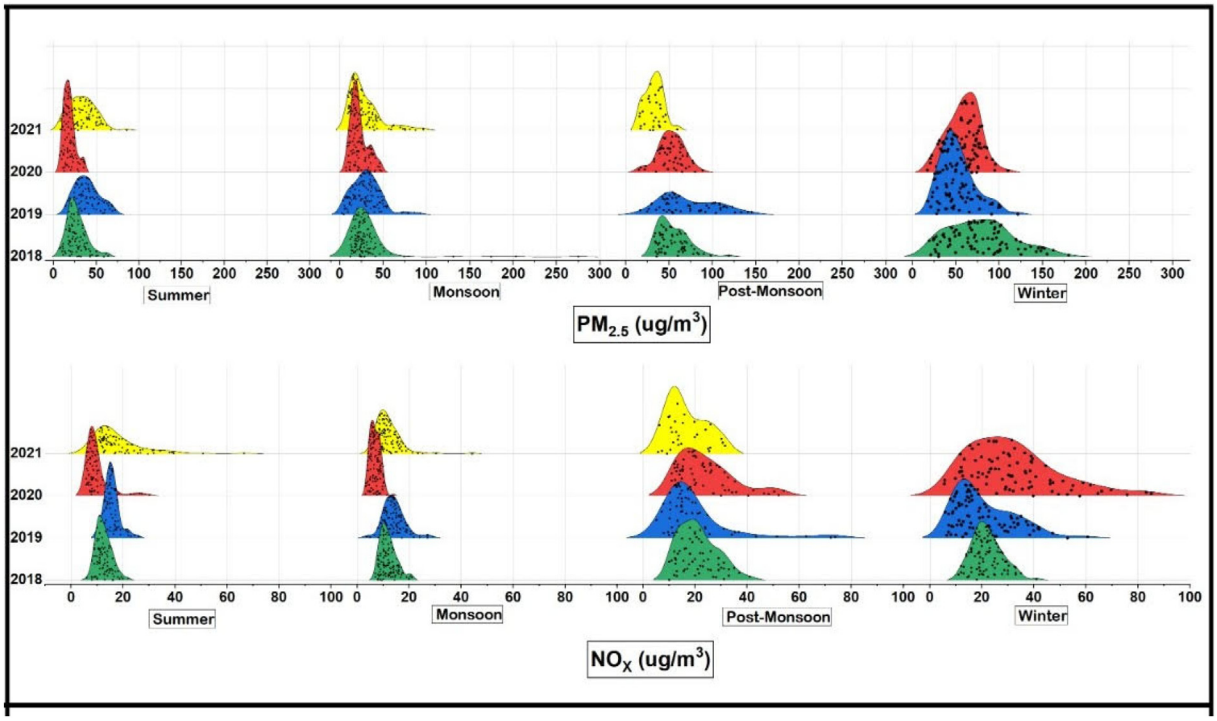


Fig. 6. Ridgeline plots showing the annual variation in different pollutants during 2018–2021 in Chandigarh, India.

4.4. Precision of the eight machine learning algorithms

Multivariate Linear Regression equation is constructed with patients as dependent variable(Y) and PM2.5(X1), NO(X2), NO_2(X3), NO_x(X4), NH_3 (X5), SO_2(X6), CO(X7), Ozone(X8), Toluene(X9), Eth-Benzene(X10), Xylene(X11), Benzene(X12), MP-Xylene(X13), RH(X14), WS(X15), WD (X16), SR(X17), AT(X17), RF(X19), Year(X20), Month(X21), Week(X22), Day(X23), as independent variables. As detailed in the methodology section, eight different machine learning models were used to predict the impact of ambient air pollution on outpatient ARI visits. Their performances are assessed based on R^2, MAE, MSE, RMSE, and r.

4.5. Results for total patients

Without Lag effect: - The results in Table S3(A) exhibit that the performance of the Random Forest Model is the best, with a maximum R^2 value of 0.87 and corresponding MAE, RMSE MSE, and r of 5.66, 7.798, 60.819, 0.934respectively, while model Decision Tree Regressor is the worst with a maximum R^2 value of 0.46 and corresponding MAE, RMSE, MSE and r of 11.71,15.56, 242.17, 0.68. Further, Support Vector Regression, Deep Neural Network and K-Nearest Neighbors regression model models were found to have R^2 values between 0.70 and 0.80 and r between 0.85 and 0.89, while Linear Regression and LASSO performed average with R^2 of 0.48–0.56 and r between 0.69 and 0.75.

With lagged effect: -The results in Table S3(A) show that performance of the X.G. Boost and Random Forest model performed competitively with R^2 values of 0.89; 0.87, and corresponding MAE, RMSE, MSE and r of 5.48, 5.70; 7.44, 7.94 and 55.35, 63.00 respectively. Pearson's correlation coefficient for X.G. Boost and Random Forest model were 0.94 and 0.93 for the 1-day lag effect on total patients. At the same time, Decision Tree Regressor performed the worst among all models with an R^2 value of 0.48 and corresponding MAE, RMSE and MSE of 11.63,15.34 and 235.40, respectively. Pearson's correlation coefficient was 0.69.

4.6. Results for ARI patients

Without Lag effect: - The results clearly highlight that the performance of the Random Forest Model is better than the other seven models, with an R^2 = 0.61 along with an MAE value of 2.74, RMSE = 5.38, MSE = 28.97. The observed value of Pearson correlation for ARI patients without lag was = 0.78. The model performs poorly on ARI patients without lag with Decision Tree Regressor and Lasso havingR^2 = 0.20, 0.19, MAE =5.29, 5.33RMSE =7.52, 7.57, and MSE = 56.52, 57.26, respectively. Additionally, X.G. Boost, Random Forest model, Support Vector Regression, and Deep Neural Network model performed better with an R^2 between 0.53 and 0.60 and r between 0.73 and 0.77, as given in Table S3(A).

With lagged effect: - The model performance on ARI patients with lagged was worst given by LASSO regression and K-Nearest Neighbors regression model with R^2 = 0.21, 0.14, MAE = 5.31, 4.16, RMSE =7.46, 7.77, and MSE = 55.59, 60.42, respectively. In contrast, X.G. Boost, Random Forest model, Support Vector Regression, and Deep Neural Network model perform better with an R^2 between 0.46 and 0.60 and r between 0.68 and 0.78, as given in Table S3(A). It reveals that all the models applied did not perform well in predicting the association of ambient air pollution with outpatient visits for ARI. The Scatter plot between actual and predicted patients with lag effect for total and ARI patients is depicted in Fig. 7 (a, b). The Scatter plot between actual patients and predicted patients with no lag effect for total patients and ARI patients is depicted in Fig. 7 (c,d).

The results show that the performance of Random Forest model is the best, with a R^2 = 0.87, 0.61, 0.87, 0.61, MAE = 5.66, 2.74, 5.70, 2.79 RMSE = 7.80, 5.38, 7.94, 5.37 and MSE = 60.82, 28.97, 62.35, 28.84 and Pearson correlation = 0.93, 0.78, 0.93, 0.78 for total no of patients without lag, ARI patients without lag, total no of patients with 1 day lag and ARI patients with 1 day lag respectively. In contrast, XGBoost model perform better with a R^2 = 0.89, MAE = 5.48, RMSE = 7.44, and MSE = 55.35 having Pearson's correlation = 0.94 only in case of total no of patients with 1-day lag. It discloses that the Random Forest model performs well for predicting ambient air pollution during outpatient visits for total patients.

4.7. Prediction accuracy based on random forest

The correlation was evaluated using the Random Forest method between the predicted and actual number of patients. Pearson's correlation was computed at 0.93 for total patients who visited HWC-49 and 0.78 for ARI patients who visited HWC-49. The Random Forest model showed the highest accuracy on the dataset and performed better than other models.

5. Discussion

The accelerated worldwide urbanization has resulted in increased air pollutant emissions from stationary and mobile sources. The population is being exposed to a gradually higher level of air pollutants (Cohen et al., 2017). The present study is planned to examine a relationship between outpatient visits for ARI and levels of ambient air pollutants in Chandigarh. The observed value of both $PM_{2.5}$ and PM_{10} far exceeds the permissible WHO and NAAQ standards for most days. Over the past three years, the highest and lowest concentrations for the majority of air pollutants were recorded during the winter and monsoon season, respectively (Ravindra et al., 2020a, 2020b). Lower levels of air pollutants observed during the monsoon reflect the impact of precipitation's washout phenomenon (Ravindra et al., 2003).

A decline in PM concentration was observed from March 2020 to September 2020, possibly due to the COVID-19 lockdown, which restricted transport and industrial activities. Similarly, distinct seasonal differences were observed for gaseous pollutants and meteorological conditions (Roberts-Semple et al., 2012). The data analysis indicates that the average daily variation in O_3 is opposite to NOx (Pancholi et al., 2018) and directly related to meteorological parameters, including temperature, W.S. and S.R. (Pudasainee et al., 2006). Eight machine learning algorithms (Random Forest model, K-Nearest Neighbors regression model, Linear regression model, LASSO regression model, Decision Tree Regressor, Support Vector Regression, X.G. Boost and Deep Neural Network with 5-layers) were utilized to find the association among daily air pollutants and outpatient visits for ARI. These programs gave the insight to explore the potential and unveil the response of various machine learning models to the dataset (Amin et al., 2019).

To estimate the accuracy of derived values from different machine learning models, we used the R^2 value. The outcome of the analysis of air pollutants concentration on a daily basis for three years and the outpatient who visits for ARI in the HWC-49, Chandigarh, showed that there is no significant linkage between the air pollutants concentration and outpatient visits of ARI. The data collected during the COVID-19 period may be limited for concluding the significant association between air pollution and the number of hospital visits, especially for predicting the risk of ARI patients. The dataset of total patients exhibited better performance of models with a 1-day lag effect. Though a comparatively small difference was exhibited with and without lagged effect, the Random Forest model gave the best performance in both cases, followed by X.G. Boost.

As depicted in Fig. 2a, it has been found that the total patient visits were associated with the time variable (f score: 0.40, at lag 0). Total patient visits were more associated with the time variables rather than air pollutant parameters. However, no such association was found among time variables, air pollution & meteorological parameters, at lag 1, in the case of ARI patients (Fig. 2 b,c,d). In 0–2 days moving average of air pollution and 0–21 days moving average of temperature, the results have shown that the Random Forest Regressor model performed better (R^2–0.545, MAE-3.280, RSME 5.8, MSE 33.784 and r value 0.738) than other models (Random Forest Regressor>X.G. Boost>Support Vector Regression>Deep Neural Network>Poisson Regressor>K-Nearest Neighbors Regression Model>Linear Regression Model>LASSO Regression Model>Decision

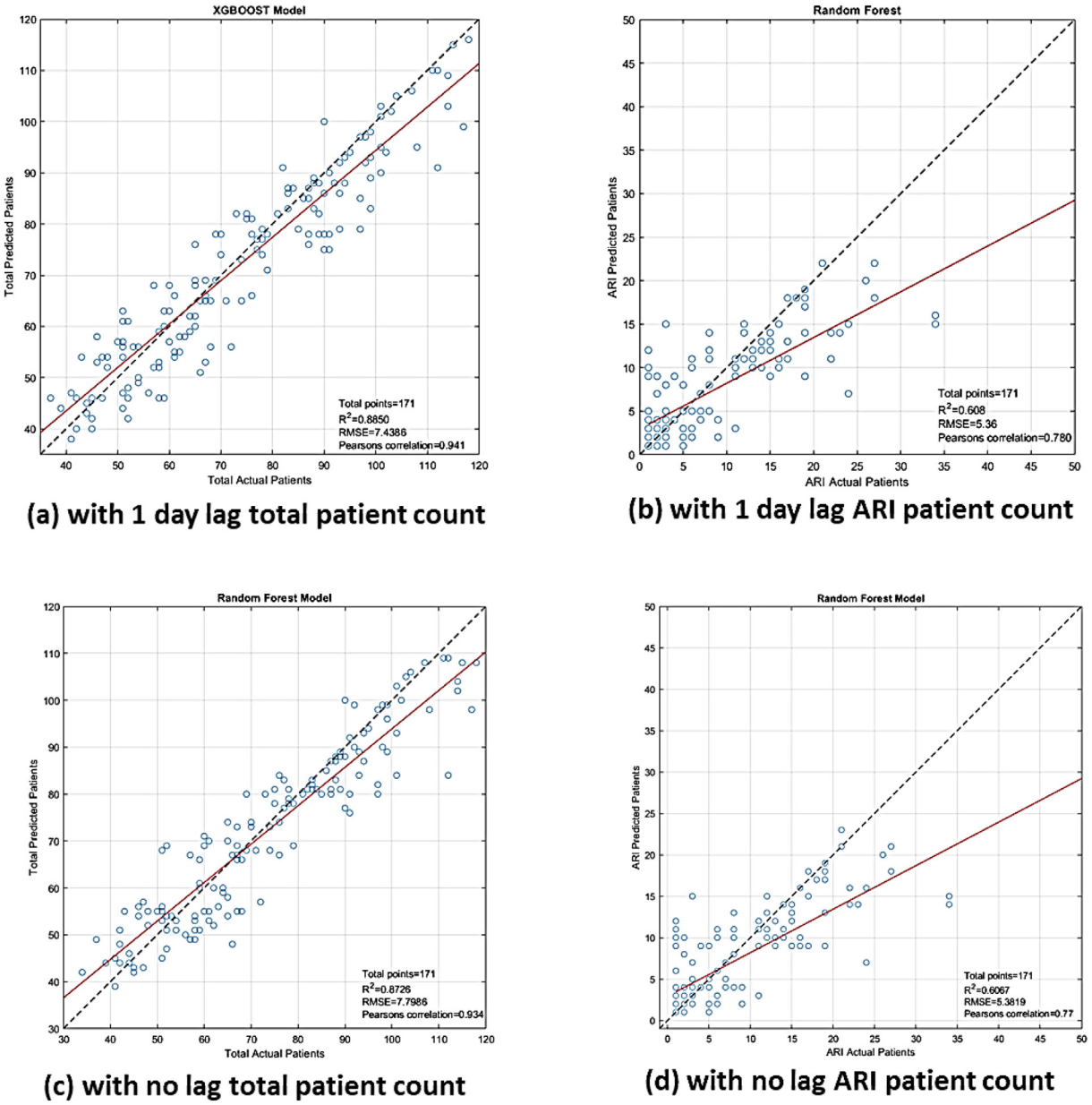


Fig. 7. The results of the Scatter plot between actual- and predicted patients on modeled data.

Tree Regressor) as shown in Fig. S3c. However, no significant association was observed between air pollution and ARI patients. Thus, the association between air pollutants and total patients does exist in Chandigarh.

Further, the analysis of the interrelation between air pollutants concentration on a daily basis and the total number of patients who visited HWC-49, Chandigarh, showed that there is a linear interrelation between the daily air pollutants concentration and the total patient visits. Earlier studies suggested machine learning techniques for predicting respiratory disease-related hospitalization have found that region-specific air pollutants and weather parameters influence the results in different geographical regions (Peng et al., 2020). In agreement with the current study, Peng et al. (2020) also reported the superiority of the Random Forest technique as a prediction model. However, they used only 2-years of data on the daily number of patients with respiratory disorders visits to emergency and OPD. The dataset was further employed to predict the rise in respiratory disease patients due to the variation in meteorological and nine air quality parameters.

The present study brings new geographical evidence from Northern India that may assist in the planning of health care resources according to the prediction of the peak number of patients visiting hospitals. The study also provides some important observations and data for future intervention studies.

5.1. Limitation of the study

Due to the COVID-19 outbreak, the lockdown was implemented in India. Activities contributing to air pollution such as vehicular movement, industrial activities and other anthropogenic sources were minimal. Hence, Chandigarh's air quality during that period was comparatively better. Also, during COVID-19, outpatient services were terminated and telemedicine services were introduced, affecting patients' visits to the HWC. This could have potentially affected the visits of patients in hospitals during the lockdown period. Thus, the data collected in the present study may not be sufficient for concluding the significant linkage between air pollution and the number of hospital visits, especially for predicting the risk of ARI patients. Future studies are recommended to consider other health care centers in the study area. Data may be collected for a longer duration to develop a time-series trend.

6. Conclusion

The marked seasonal differences for most pollutants exhibit significantly higher concentrations during winters and lower levels during monsoon seasons. The current study provides evidence on the application of machine learning techniques to predict respiratory illnesses related to outpatient hospital visits and their association with air pollutants and meteorological parameters. No significant association was found between the air pollutants concentration and outpatient visits of ARI. The studied models did not perform well in predicting ambient air pollution on outpatient visits for ARI. However, the Random Forest model reacted better to single and multi-day moving averages among the eight machine learning algorithm models. This model is marginally more accurate and robust than X.G. Boost, Support Vector Regression, Deep Neural Network, Poisson Regressor, K-Nearest Neighbors Regression Model, Linear Regression Model, LASSO Regression Model, and Decision Tree Regressor. Future studies may also apply this distinguished model to predict ambient air pollution's impact on outpatient acute respiratory infections visits. In conclusion, the study highlighted the application of machine learning models, specifically the Random Forest regression model, to predict analytics for several other diseases apart from ARI, such as heart disease, other respiratory disorders, eye disease, etc. Using such regression models offers vast opportunities for improving and predicting healthcare in hospital settings. The study looks at external environmental factors related to respiratory health to provide critical observations and data for future intervention studies.

CRediT authorship contribution statement

Conceptualization & Methodology: KR, VK, SM.
Software & Validation: KR, VK, SB, SBS.
Resources & Data Curation: KR, VK, MKS, SB, SM, MG.
Writing - Original Draft Preparation: KR, SBS, VK, SM
Writing - Review & Editing: KR, SB, VK, MKS, SBS, MG, SM.
Visualization & Supervision: SM, KR, MKS, MG.
Funding acquisition: NA.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

SM and KR acknowledge the HCWH for the Climate, Health and Air Monitoring Project (CHAMP) project. KR would like to thank the National Programme on Climate Change and Human Health (NPCCHH) under the Ministry of Health and Family Welfare (MoHFW) for designating his institute as a Center of Excellence (CoE) on Climate Change and Air Pollution Related Illness. SM and KR also acknowledge the Ministry of Environment, Forest & Climate Change, for identifying their institute as an Institute of Repute (IoR) under the National Clean Air Program (NCAP).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.159509.

References

Amin, M.S., Chiam, Y.K., Varathan, K.D., 2019. Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform. 36, 82–93.

Amuthadevi, C., Vijayan, D.S., Ramachandran, V., 2021. Development of air quality monitoring (AQM) models using different machine learning approaches. J. Ambient. Intell. Humaniz. Comput. 1–13.

Barnston, A.G., 1992. Correspondence among the correlation, RMSE, and heidke forecast verification measures; refinement of the heidke score. Weather Forecast. 7 (4), 699–709.

Biswal, A., Singh, T., Singh, V., Ravindra, K., Mor, S., 2020. COVID-19 lockdown and its impact on tropospheric NO2 concentrations over India using satellite-based data. Heliyon 6 (9), e04764.

Brook, R.D., Rajagopalan, S., Pope III, C.A., Brook, J.R., Bhatnagar, A., Diez-Roux, A.V., Holguin, F., Hong, Y., Luepker, R.V., Mittleman, M.A., Peters, A., 2010. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation 121 (21), 2331–2378.

Cheng, Y.R., Feng, Z.H., Zhou, M.Y., Wang, N., Wang, M.W., Ye, L., Chen, J., 2020. Machine Learning Prediction on Number of Patient due to Conjunctivitis Based on Air Pollutants: A Preliminary Study.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. Lancet 389 (10082), 1907–1918.

CPCB, 2014. National Air Quality Index. Cent Pollution Control Board, pp. 1–44 (January).

Darrow, L.A., Klein, M., Flanders, W.D., Mulholland, J.A., Tolbert, P.E., Strickland, M.J., 2014. Air pollution and acute respiratory infections among children 0–4 years of age: an 18-year time-series study. Am. J. Epidemiol. 180 (10), 968–977.

Hassan, M.Y., 2021. The deep learning LSTM and MTD models best predict acute respiratory infection among under-five-year old children in Somaliland. Symmetry 13 (7), 1156.

Host, S., Larrieu, S., Pascal, L., Blanchard, M., Declercq, C., Fabre, P., Jusot, J.F., Chardon, B., Le Tertre, A., Wagner, V., Prouvost, H., 2008. Short-term associations between fine and coarse particles and hospital admissions for cardiorespiratory diseases in six French cities. Occup. Environ. Med. 65 (8), 544–551.

Huang, C., Moran, A.E., Coxson, P.G., Yang, X., Liu, F., Cao, J., Chen, K., Wang, M., He, J., Goldman, L., Zhao, D., 2017. Potential cardiovascular and total mortality benefits of air pollution control in urban China. Circulation 136 (17), 1575–1584.

Jiang, Y., Chen, J., Wu, C., Lin, X., Zhou, Q., Ji, S., Yang, S., Zhang, X., Liu, B., 2020. Temporal cross-correlations between air pollutants and outpatient visits for respiratory and circulatory system diseases in Fuzhou, China. BMC Public Health 20 (1), 1–13.

Kannan, R., Vasanthi, V., 2019. Logistic regression and KNN algorithm experimental diagnosis to reduce the impact of cardiac arrest. Int. J. Simul. Syst. Sci. Technol. 20 (1).

Kelly, F.J., Fussell, J.C., 2011. Air pollution and airway disease. Clin Exp Allergy 41 (8), 1059–1071.

Kenney, J.F., Keeping, E.S., 1962. Root mean square. Mathematics of Statistics. 1, pp. 59–60.

Künzli, N., Perez, L., Rapp, R., 2010. Air Quality and Health. 89. European Respiratory Society. World Health Organ, Lausanne.

Landrigan, P.J., Fuller, R., Acosta, N.J., Adeyi, O., Arnold, R., Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., Chiles, T., 2018. The lancet commission on pollution and health. Lancet 391 (10119), 462–512.

Liao, K.M., Liu, C.F., Chen, C.J., Shen, Y.T., 2021. Machine learning approaches for predicting acute respiratory failure, ventilator dependence, and mortality in chronic obstructive pulmonary disease. Diagnostics 2021 (11), 2396.

Liu, C., Liu, Y., Zhou, Y., Feng, A., Wang, C., Shi, T., 2019. Short-term effect of relatively low level air pollution on outpatient visit in Shennongjia, China. Environ. Pollut. 245, 419–426.

Lu, J., Bu, P., Xia, X., Lu, N., Yao, L., Jiang, H., 2021. Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases. Environ. Sci. Pollut. Res. 28 (23), 29701–29709.

Medina-Ramon, M., Zanobetti, A., Schwartz, J., 2006. The effect of ozone and PM10 on hospital admissions for pneumonia and chronic obstructive pulmonary disease: a national multicity study. Am. J. Epidemiol. 163 (6), 579–588.

Ministry of Health Statistical Information Centre, 2016. International statistical classification of diseases and related health problems, Instruction manual 2. https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf.

Mor, S., Kumar, S., Singh, T., Dogra, S., Pandey, V., Ravindra, K., 2021. Impact of COVID-19 lockdown on air quality in Chandigarh, India: understanding the emission sources during controlled anthropogenic activities. Chemosphere 263, 127978.

Pancholi, P., Kumar, A., Bikundia, D.S., Chourasiya, S., 2018. An observation of seasonal and diurnal behavior of O3–NOx relationships and local/regional oxidant (OX= O3+ NO2) levels at a semi-arid urban site of western India. Sustain. Environ. Res. 28 (2), 79–89.

Peng, J., Chen, C., Zhou, M., Xie, X., Zhou, Y., Luo, C.H., 2020. Peak outpatient and emergency department visit forecasting for patients with chronic respiratory diseases using machine learning methods: retrospective cohort study. JMIR Med. Inform. 8 (3), e13075.

Pudasainee, D., Sapkota, B., Shrestha, M.L., Kaga, A., Kondo, A., Inoue, Y., 2006. Ground level ozone concentrations and its association with NOx and meteorological parameters in Kathmandu valley, Nepal. Atmos. Environ. 40 (40), 8081–8087.

Ravindra, K., Mor, S., Kamyotra, J.S., Kaushik, C.P., 2003. Variation in spatial pattern of criteria air pollutants before and during initial rain of monsoon. Environ. Monit. Assess. 87 (2), 145–153.

Ravindra, K., Kaur-Sidhu, M., Mor, S., 2020. Air pollution in rural households due to solid biomass fuel use and its health impacts. Indoor Environmental Quality. Springer, Singapore, pp. 27–33.

Ravindra, K., Singh, T., Mor, S., Singh, V., Mandal, T.K., Bhatti, M.S., Gahlawat, S.K., Dhankhar, R., Mor, S., Beig, G., 2019. Real-time monitoring of air pollutants in seven cities of North India during crop residue burning and their relationship with meteorology and transboundary movement of air. Sci. Total Environ. 690, 717–729.

Ravindra, K., Singh, T., Pandey, V., Mor, S., 2020. Air pollution trend in Chandigarh city situated in indo-Gangetic Plains: understanding seasonality and impact of mitigation strategies. Sci. Total Environ. 729, 138717.

Ravindra, K., Singh, T., Sinha, V., Sinha, B., Paul, S., Attri, S.D., Mor, S., 2021. Appraisal of regional haze event and its relationship with PM2. 5 concentration, crop residue burning and meteorology in Chandigarh, India. Chemosphere 273, 128562.

Ravindra, K., Chanana, N., Mor, S., 2021. Exposure to air pollutants and risk of congenital anomalies: a systematic review and metaanalysis. Sci. Total Environ. 765, 142772.

Ravindra, K., Singh, T., Mor, S., 2022. Preventable mortality attributable to exposure to air pollution at the rural district of Punjab, India. Environmental Science and Pollution Research 29 (21), 32271–32278.

Ravindra, K., Singh, T., Mandal, T.K., Sharma, S.K., Mor, S., 2022. Seasonal variations in carbonaceous species of PM2. 5 aerosols at an urban location situated in indo-gangetic plain and its relationship with transport pathways, including the potential sources. J. Environ. Manag. 303, 114049.

Ravindra, K., Singh, T., Vardhan, S., Shrivastava, A., Singh, S., Kumar, P., Mor, S., 2022. COVID-19 pandemic: what can we learn for better air quality and human health? J. Infect. Public Health 15 (2), 187–198.

Roberts-Semple, D., Song, F., Gao, Y., 2012. Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern New Jersey. Atmos. Pollut. Res. 3 (2), 247–257.

Romieu, I., Samet, J.M., Smith, K.R., Bruce, N., 2002. Outdoor air pollution and acute respiratory infections among children in developing countries. J. Occup. Environ. Med. 640–649.

Sarizadeh, G., Jaafarzadeh, N., Roozbehani, M.M., Tahmasebi, Y., Moattar, F., 2020. Relationship between the number of hospitalized cardiovascular and respiratory disease and the average concentration of criteria air pollutants (CAP) in Ahvaz. Environ. Geochem. Health 42 (10), 3317–3331.

Schraufnagel, D.E., Balmes, J.R., Cowl, C.T., De Matteis, S., Jung, S.H., Mortimer, K., Perez-Padilla, R., Rice, M.B., Riojas-Rodriguez, H., Sood, A., Thurston, G.D., 2019. Air pollution and noncommunicable diseases: a review by the forum of international respiratory 'Societies'Environmental committee, part 2: air pollution and organ systems. Chest 155 (2), 417–426.

Singh, T., Ravindra, K., Sreekanth, V., Gupta, P., Sembhi, H., Tripathi, S.N., Mor, S., 2020. Climatological trends in satellite-derived aerosol optical depth over North India and its relationship with crop residue burning: rural-urban contrast. Sci. Total Environ. 748, 140963.

Singh, T., Ravindra, K., Beig, G., Mor, S., 2021. Influence of agricultural activities on atmospheric pollution during post-monsoon harvesting seasons at a rural location of indo-gangetic plain. Sci. Total Environ. 796, 148903.

Wang, C., Feng, L., Chen, K., 2019. The impact of ambient particulate matter on hospital outpatient visits for respiratory and circulatory system disease in an urban chinese population. Sci. Total Environ. 666, 672–679.

Zhang, S., Li, G., Tian, L., Guo, Q., Pan, X., 2016. Short-term exposure to air pollution and morbidity of COPD and asthma in east asian area: a systematic review and meta-analysis. Environ. Res. 148, 15–23.

Zhang, Z., Chai, P., Wang, J., Ye, Z., Shen, P., Lu, H., Jin, M., Gu, M., Li, D., Lin, H., Chen, K., 2019. Association of particulate matter air pollution and hospital visits for respiratory diseases: a time-series study from China. Environ. Sci. Pollut. Res. 26 (12), 12280–12287.