INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Multioutput Ensemble Machine Learning Algorithm: A Prediction Model of Acute Respiratory infection and Pneumonia Occurrence

Bhakti S Pimpale[1]*, **Anala A Pandit**[2]

1 Research Scholar, Department of Computer Application, VJTI, Mumbai, India
2 Ph.D. Supervisor, Department of Computer Application, VJTI, Mumbai, India

OPEN ACCESS

Received: 29-04-2023
Accepted: 02-11-2023
Published: 05-12-2023

Citation: Pimpale BS, Pandit AA (2023) Multioutput Ensemble Machine Learning Algorithm: A Prediction Model of Acute Respiratory infection and Pneumonia Occurrence. Indian Journal of Science and Technology 16(45): 4141-4155. https://doi.org/10.17485/IJST/v16i45.1011

* **Corresponding author**.

bspimpale_p18@mc.vjti.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To forecast daily OPD patients based on air pollution and weather parameters, the objective is to build a robust model that accurately predicts patient volume by considering major missing values and factors such as PM2.5 levels, temperature, humidity, wind speed, and rainfall, etc. thereby improving healthcare planning and delivery. **Methods:** To develop the multioutput ensemble model for forecasting daily OPD (out-patient department), we have used 13 machine learning techniques such as regression analysis, Extra tree regressor, Support vector regressor, etc. We have collected and pre-processed data from multiple sources, including air quality and weather parameters from NASA's website, and historical healthcare data from Shatabdi Hospital, Govandi, Mumbai. We have developed the model using a combination of Gaussian regressor and Extra tree regressor and evaluated its performance using metrics such as FastDTW, RMSE, etc. **Findings:** The prediction result shows that the multioutput ensemble model performed significantly better than other models even with the presence of outliers, multicollinearity, and non-stationarity with Root Mean Squared Error 0.46 and 0.22 for ARI and Pneumonia with lag 7 days and 8 days respectively. Moreover, this model also worked well including Covid-19 period data when there was a negligible correlation between independent and dependent variables. **Novelty:** None of the datasets that have been used for the prediction of time series data have had a significant gap in recorded data in the time domain which has been effectively taken care of in this research. Secondly, all the earlier research work in this domain addresses only a single disease that provides the same lag value irrespective of the disease. The period of expression after the event occurrence may vary for multiple diseases, albeit in one domain that is triggered by similar and /or different air pollutants. This issue has been addressed by ensembling multiple ML algorithms to effectively optimize time complexity.

Keywords: Acute Respiratory Infection; Pneumonia; Gaussian Regressor; Extra Tree Regressor; Weather Data; Air Pollution

1 Introduction

The impact of air pollution on human health, particularly respiratory diseases, is a critical concern worldwide. According to the World Health Organization (WHO), air pollution is recognized as the fifth leading cause of global mortality [(1)]. As a result, extensive research is being conducted to comprehend the relationship between respiratory diseases and environmental factors, such as climatic conditions and air pollution.

For patients with respiratory disorders, Khatri K L, Tamil L S [(2)] used a multi-layered ANN model with a backpropagation algorithm to forecast high occurrences or days of peak demand. The model was built using 8 predictors. The proposed ANN model produced good forecasting results, with the system's overall accuracy coming in at 81%. However, the study population was limited to emergency department visits for respiratory diseases. Moreover, the research did not take into account separating the meteorological and air pollution data individually. Ku Y, Kwon SB, Yoon JH, Mun SK, and Chang M [(3)] compare Gaussian process regressor and gradient boosting methods for patient arrival forecasting for the period of 2014-2019. Both models demonstrated competitive prediction performance, with R2 values of over 0.67 and RMSE values below 13.9. One major drawback of this study is the lack of consideration for different diseases as separate time series. Instead, all respiratory diseases were combined and treated as a single time series. This approach overlooks the potential variations in disease patterns, trends, and factors associated with specific respiratory illnesses. By analyzing each disease as a separate time series, researchers could have gained a more nuanced understanding of the individual dynamics and specific risk factors associated with each respiratory disease. This could have provided valuable insights for targeted interventions and preventive measures tailored to different respiratory conditions. A machine learning strategy was suggested by Khaiwal R, Bahadur S S, Katoch V, Bhardwaj S, Kaur-Sidhu M, Gupta M et al. [(4)] to examine outpatient visits and air pollution. Machine learning models were tested with the lagged effect of exposure and without the lagged effect of exposure. A strong correlation exists between total patient visits and gaseous air pollutants during a 1-day period. Using trained data, the random forest regression model has the best R2 of 0.87. The fundamental flaw in this study is the author's random data division into train and testing which is not appropriate for time series data. The author has not taken into account data from the COVID-19 timeframe. With a forecasting error of roughly 8.17%, Kim M S, Lee J H, Jang Y J, Lee C H, Choi J H, and Sung T E [(5)] suggested an ensemble VAR-DNN model predict the occurrence of asthma. For a very limited set of data, the author used DNN to create an ensemble model. Because DNN has a higher computational cost than machine learning algorithms, it is not advised to implement it for such a small data set.

While various studies have examined the effects of climatic and air pollution factors on respiratory diseases individually, there is a scarcity of research that integrates these factors into predictive models, especially when dealing with major data gaps during holidays, weekends, or periods of limited outpatient admissions like during the COVID-19 pandemic, especially in areas adjacent to dumping grounds for multiple diseases. This study aims to address this research gap by proposing the development of machine learning models that utilize spatial climatic and air pollution datasets to predict the occurrence of respiratory diseases like ARI and Pneumonia. By incorporating spatial climatic and air pollution datasets into machine learning models, it is expected that patterns and correlations can be identified to predict the occurrence of respiratory diseases. This predictive information can be valuable in formulating targeted control measures and preventive strategies. These models can assist hospital management in making informed decisions regarding resource allocation, both in terms of existing resources and potential requirements for additional resources.

In this paper, the main aim is to bridge the research gap by integrating spatial climatic and air pollution datasets into machine learning models for predicting respiratory disease occurrence. By doing so, it seeks to enhance our understanding of respiratory diseases and facilitate the formulation of effective preventive measures to safeguard public health. Additionally, the study acknowledges the need to address data gaps during specific periods, such as holidays, weekends, and times of reduced outpatient admissions like during the COVID-19 pandemic.

The rest of the paper follows the following organization. Section two elaborates on the methodology for regression, evaluation metrics, dataset description, and proposed model. Results and discussion, and conclusion are explained in sections three and four respectively.

2 Method ology

2.1 Methods for Regression

In this study following regression methods were used.

Vector Autoregression (VAR) is a statistical model used in time series analysis. VARs are effective tools for data description and producing trustworthy multivariate benchmark predictions.

A statistical model called linear regression (LR) is used to examine the correlation between a dependent variable and one or more independent variables. Finding the line that reduces the total squared differences between the actual data and the predicted values is the method's key step.

Support Vector Regressor (SVR) is an adaptation of the well-known Support Vector Machine (SVM) technique. The goal of SVM is to identify the hyperplane that divides the data into two classes with the greatest possible margin. A similar idea underlies SVR's approach to regression, which aims to identify the hyperplane with high accuracy and best fits the data.

The LARS-Lasso method was first introduced by Bradley E, Trevor H, Iain J and Robert T in 2004. The objective function of the LARS-Lasso method is extended from LARS by including a penalty component that promotes the coefficients to be small. This can improve model interpretation and prediction accuracy.

Ridge Regression is used to solve some of the drawbacks of ordinary least squares (OLS) regression. A regularisation method for performing linear regression is called Lasso. The maximum size of the estimated coefficients is constrained by the penalty term in Lasso. As a result, ridge regression is similar. The coefficient estimates generated by the shrinkage estimator Lasso are skewed towards low values. In order to improve the outcome, it applies both variable selection and regularisation.

Bayesian Ridge Regression is a Bayesian linear regression method that incorporates a prior distribution of the model parameters to regularize the regression coefficients. With its ability to reduce regression coefficients to zero and prevent overfitting, Bayesian Ridge Regression is able to significantly minimize the effects of multicollinearity among feature variables.

The Elastic Net algorithm is used to achieve variable selection and high prediction accuracy. The Elastic Net algorithm, a regularised linear regression algorithm, integrates both the L1 and L2 penalties. It has become a common machine learning technique, especially in scenarios where there are many more predictors than observations.

Extreme Gradient Boosting (XGBoost) is a scalable, distributed gradient-boosted decision tree machine learning library. It supports parallel tree boosting and is the top machine-learning library for predictive, classification, and ranking tasks.

The decision tree is a well-known machine-learning technique that is used for classification and regression applications. It is a decision-making paradigm that resembles a tree.

The Random Forest algorithm is an extension of the decision tree algorithm, with the aim of reducing overfitting and improving the accuracy of the predictions. The Random Forest algorithm works by constructing multiple decision trees, where each tree is built using a different subset of the training data and a different subset of the input features. The final prediction is made by combining the forecast of each and every tree.

The Multi-Layer Perceptron (MLP) neural network is a type of feedforward neural network, which consists of multiple layers of interconnected nodes or neurons that can learn to approximate complex non-linear relationships between the input and output variables.

Stochastic Gradient Boosting is a variant of the Gradient Boosting algorithm. The Stochastic Gradient Boosting algorithm works by iteratively adding weak learners to the model, where each learner is trained on a randomly selected subset of the training data and a randomly selected subset of the input features.

The Gaussian process regression algorithm uses a Gaussian distribution to model the distribution of possible functions that can fit the data and uses Bayesian inference to find the most probable function given the observed data. The algorithm defines a prior distribution over the possible functions and then updates this distribution based on the trained data to obtain a posterior distribution over the functions. Its ability to provide uncertainty estimates for its predictions and its flexibility in modeling complex, non-linear relationships make it a powerful tool for many machine learning tasks.

Extra Trees (Extreme Randomized Trees) is an ensemble machine learning algorithm introduced for regression and classification tasks. They created the Extra Trees algorithm as a variant of the Random Forest algorithm, which adds more randomness to the tree-building process by selecting random splits at each node of the decision tree rather than the best split based on a set of pre-defined criteria. This additional level of randomness makes the Extra Trees algorithm less sensitive to noisy or irrelevant features, which can result in improved performance and less overfitting. Furthermore, the Extra Trees algorithm is computationally efficient and capable of dealing with high-dimensional data with a large number of features.

2.2 Evaluation Metrics

To evaluate the accuracy of the various machine learning models employed in this work to predict the occurrence of ARI and Pneumonia, seven statistical parameters were used.

2.2.1 Root mean square error (RMSE)

RMSE is a commonly used metric for comparing the accuracy of prediction for various models. The lower the RMSE number, the better the model's ability to predict in terms of absolute deviation.

It measures the average squared difference between the predicted and actual values of a continuous variable. MSE is used to test the training and testing datasets' performance.

2.2.2 The coefficient of determination (R2)

It indicates the percentage of the variance (of y) that is explained by the independent variables in the model. Indicating the quality of the fit, the fraction of explained variance assesses how well the model is likely to predict unseen samples. The maximum score is 1.0, although it can also be a negative number (because the model can be arbitrarily worse).

2.2.3 Max error (ME)

The maximum residual error, a measure that quantifies the worst-case difference between the predicted and true values, is computed using the max error function. A single output regression model that is perfectly fitted would have a maximum error of 0 on the training set, and while this is highly unlikely in practice, this metric shows the extent of error that the model had when it was fitted. The best possible score is 0.0, and a lower value is preferable. [0, +inf] is the range.

2.2.4 MASE

The mean absolute scaled error (MASE) is a statistical measure of forecast accuracy. It is calculated by dividing the mean absolute error of the predicted values by the mean absolute error of the in-sample one-step naive forecast. This metric is used for non-stationary time series forecasting.

2.2.5 The correlation coefficient (COR)

COR is used to evaluate the association between actual and predicted values.

2.2.7 FastDTW

FastDTW, or Fast Dynamic Time Warping is a variation of the original Dynamic Time Warping (DTW) algorithm, which is a well-liked method for comparing two-time series with various durations and warping forms. We used this evaluation metric to check whether predicted values follow the same trend as actual values. Lower values of FastDTW show a similar trend.

2.3 Dataset

2.3.1 Description

This study utilized a total of 20 endogenous variables, From Govandi which is a suburb in eastern Mumbai, Maharashtra. Meteorological data for Govandi, including daily average temperature (C), relative Humidity (%), rainfall (mm), wind speed (m/s), dew point, UV index, minimum, and maximum temperature were collected from NASA's official website [(6)]. Air pollution data, like SO_2 (kg/m2), NO_2 (kg/m2), PM2.5 (kg/m2), PM10 (kg/m2), CO (kg/m2), O_3(kg/m2), CH_4(kg/m2), and Aerosol optical depth were obtained from Copernicus website which is satellite-based data [(7)].

The data was collected for a daily count of outpatients from Shatabdi Hospital Govandi, Mumbai for the period of Jan 2018 to June 2022. During the period of study, a major uncertain event occurred as covid-19. This created a gap in the availability of data as the OPD was closed for ARI and Pneumonia patients. The research papers which were published after this period either did not considerthis period or created pre covid and post covid datasets and applied the models independently on two datasets. None of the published research to the extent of the author's knowledge has applied a single model to cover this period. Also, none of the published results has considered the days when there are public holidays or weekends (OPDs are normally closed on those days) to eliminate the gaps of unavailability of data on those dates.

To overcome this issue two new features holiday and lockdown were introduced because the Covid-19 period was to be considered for analysis. This created a binary time series. A binary time series is a sequence of data that consists of binary values, which can be either 0 or 1. It is often used to represent the presence or absence of an event. In binary time series, 260 days out of 1673 days were observed as holidays including Sundays and national holidays. Similarly, for lockdown, 400 days were observed as a lockdown period from March 2020 to June 2021. The summary statistics and binary time series data for holidays and lockdown periods are shown in Figure 1 and Table 1.

The summary statistics and graphical plots of the values of Daily Meteorological data are shown in Table 1, Figure 1 respectively. As it is observed from the plots in Figure 1 the time series for all 8 parameters have seasonal components in it. Also, we can see that in the plots of Wind speed and rainfall, there are few outliers.

The summary statistics and graphical plots of the values of Daily air pollution data are shown in Table 1 and Figure 1 respectively. All air pollutants data was total column data which means that it is measured in a column from the surface to

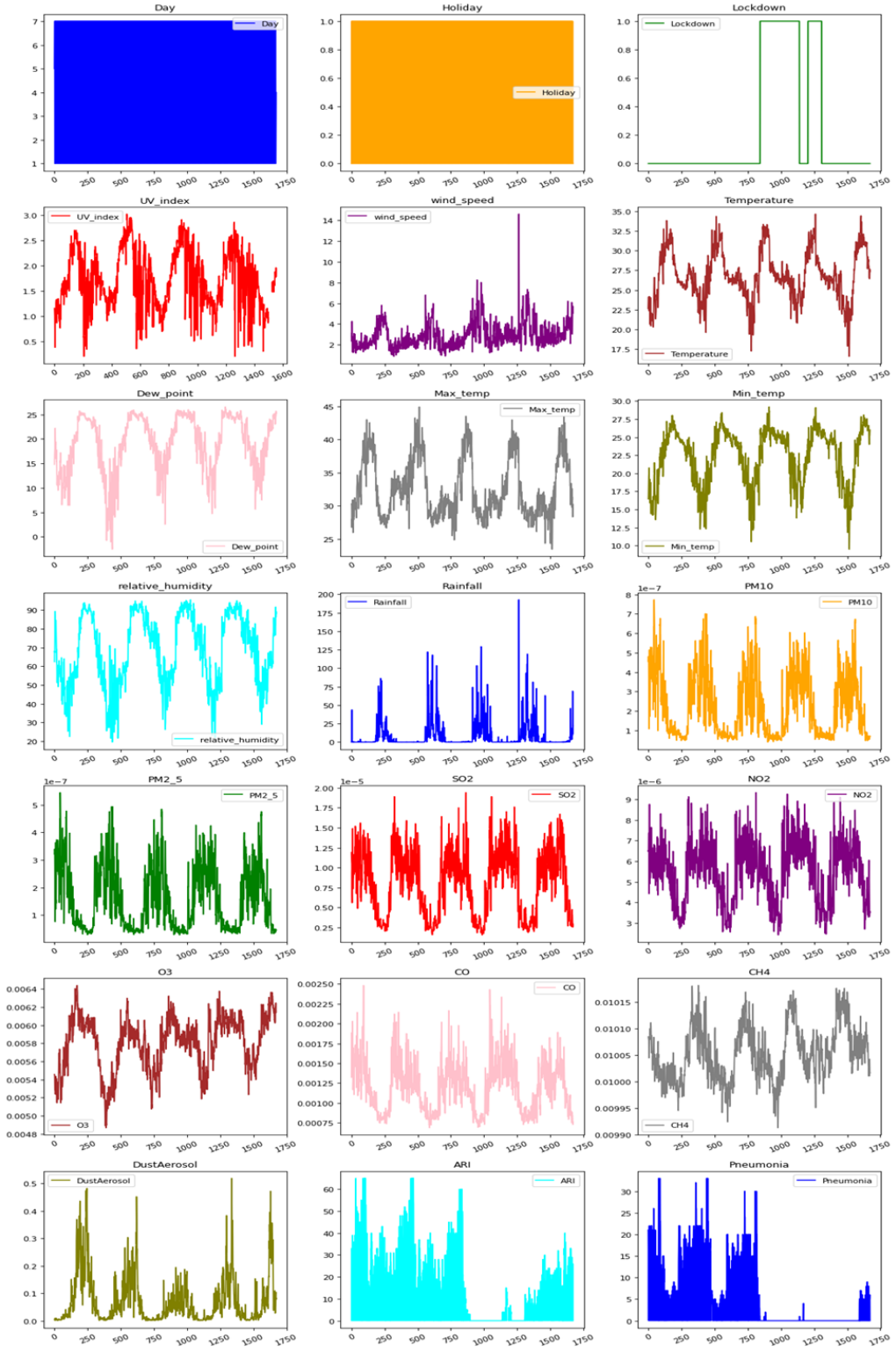

Fig 1. Daily Meteorological, Air pollution and disease data

Table 1. Summary statistics of data

	Coun t	Mea n	st d	min	25%	50%	75%	Ma x
Holiday	1673	0.155	0	0.362	0	0	0	1
Lockdown	1673	0.239	0	0.426	0	0	0	1
UV_inde x	1483	1.726156	0.5666401	0.21	1.3	1.68	2.2	3.02
wind_spee d	1673	2.871722	1.2092507	0.86	2.04	2.6	3.42	14.62
Temperatur e	1673	26.88624	3.1034754	16.58	25.05	26.61	29.12	34.65
Dew_ poin t	1673	19.01087	5.7802494	-2.54	15.21	19.96	24.25	26.53
Max_te mp	1673	32.84157	4.3497513	23.48	29.29	31.56	36.6	44.91
Min_te mp	1673	22.37232	3.7699736	9.48	19.41	23.69	25.28	29.19
relative_ humidity	1673	68.03619	18.745102	19.5	53.25	69.81	86.38	95.44
Rainfa ll	1673	7.262355	17.374571	0	0	0.08	5.36	192.72.
PM 1 O	1673	2.24E-07	1.58E-07	3.97E-08	8.13E-08	1.78E-07	3.52E-07	7.73E-07
PM 2. 5	1673	1.58E-07	1.12E-07	2.80E-08	5.57E-08	1.26E-07	2.48E-07	5.44E-07
SO2	1673	8.13E-06	3.99E-06	1.57E-06	4.04E-06	8.76E-06	1.12E-05	1.94E-05
NO2	1673	5.45E-06	1.41E-06	2.41E-06	4.35E-06	5.70E-06	6.47E-06	9.34E-06
O3	1673	0.005799	0.0002924	0.00487	0.00559	0.00587	0.00602	0.00643852
CO	1673	0.001204	0.0003159	0.00069	0.00092	0.0012	0.00141	0.00247969
CH4	1673	0.010053	5.07E-05	0.00991	0.01001	0.01005	0.01009	0.01018082
Aerosol Optic al Depth	1673	0.07035	0.0872686	9.80E-05	0.00892	0.03674	0.09794	0.51932358
ARI	1656	15.007	16.381	0	0	11	27	65
Pneumonia	1655	5.368	7.404	0	0	0	10	33

the top of the atmospheric level. The troposphere, stratosphere, mesosphere, thermosphere, and exosphere are the primary layers, in order from lowest to highest. while the concentration of all air pollutants except Ozone is the same in all layers. The total atmospheric Ozone is dominated by high concentrations of stratospheric Ozone. Ozone is found high in the atmosphere in the stratospheric layer and is beneficial to human health as it absorbs solar ultraviolet radiation. However, it becomes harmful to human health when it is found at ground level in the air in cities. So, if the concentration of Ozone in the stratospheric layer decreases the number of patients increases. Aerosol optical depth (AOD) is a measure of the number of aerosols, such as dust, smoke, and other tiny particles, present in the atmosphere. It is a measure of how much the aerosols are blocking or scattering the incoming solar radiation. The low value of AOD indicates a large number of dust particles in the atmosphere inclining an increased number of patients.

ARI and Pneumonia occurrence data were used as forecasting targets in this study. The summary statistics and graphical plots of the values of the Daily count of ARI and Pneumonia are shown in Table 1, Figure 1 respectively. From Figure 1, it can be observed that during the Covid-19 period daily count of ARI and Pneumonia patients has drastically decreased to almost 0 occurrences per day.

From Table 1 it is observed that a minimum of 0 and a maximum of 65 and 33 patients of ARI and Pneumonia respectively have visited the OPD on any day.

The scatter plot with correlation line is presented in Figures 2 and 3 . The presence of a positive/negative correlation line suggests a positive/negative association between the variables. Figures 2 and 3 are used to show the correlation between the occurrence of ARI and pneumonia with respect to air pollutants and meteorological parameters.

From Figure 2 it is observed that when the concentration of UV index, wind speed, temperature, dew point, O_3, and aerosol optical depth decreases the number of patients of ARI increases. While an increase in the concentration of particulate matter (PM10 and PM2.5), Sulfur dioxide, Nitrogen dioxide, Carbon monoxide, and Methane results in an increase in the daily count of patients of ARI. The max temperature and number of patients with ARI are directly proportionate. The observed direct proportionality between maximum temperature and the number of ARI patients indicates that higher temperatures might contribute to an increased susceptibility to ARI, potentially creating a conducive environment for the transmission and severity of respiratory illnesses.

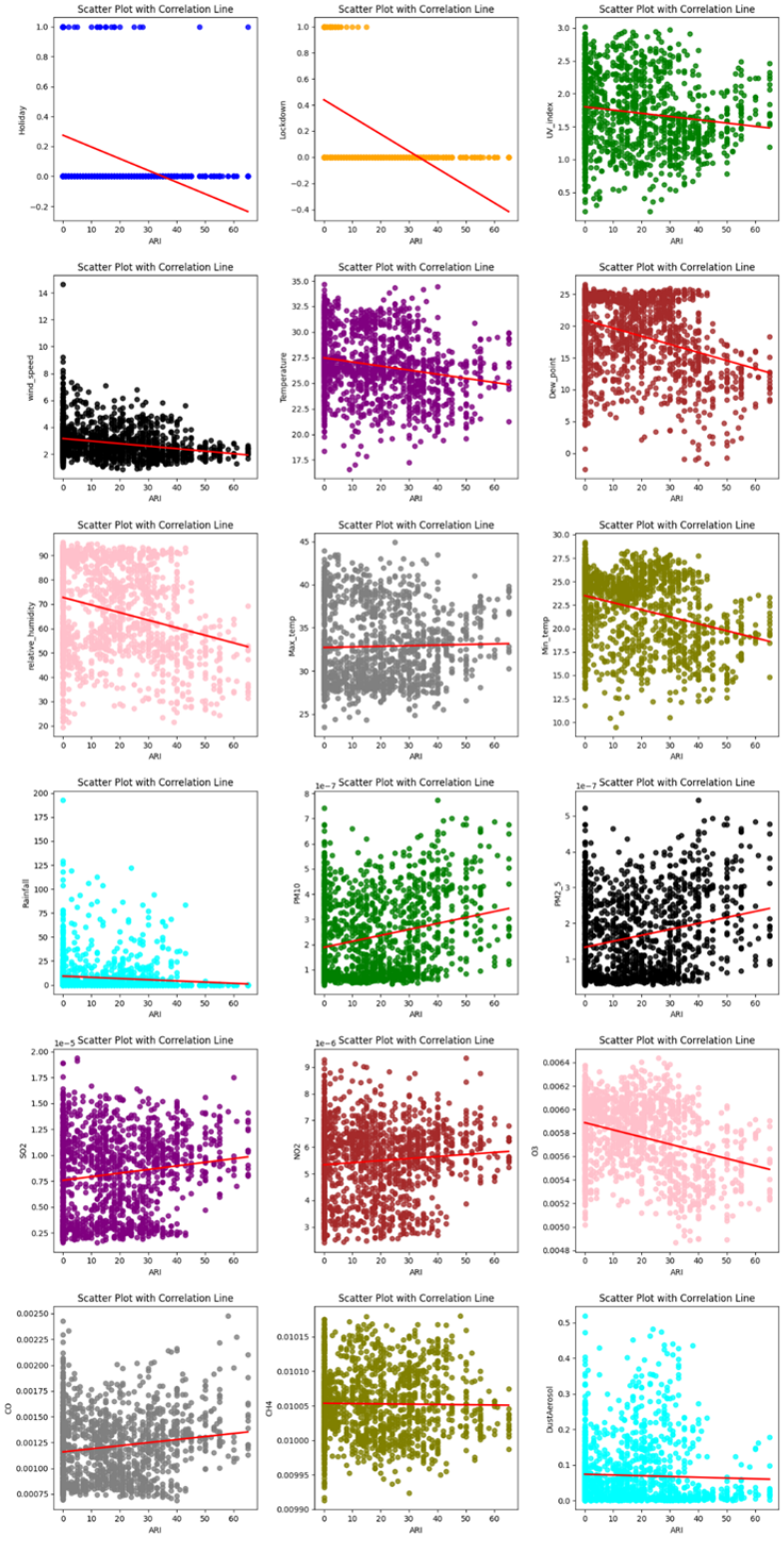

Fig 2. Correlation of daily count of ARI with other features

Figure 3 depicts that when the concentration of UV index, wind speed, temperature, dew point, O_3, and aerosol optical depth decreases the number of patients of Pneumonia increases. While an increase in the concentration of particulate matter (PM10 and PM2.5), Sulfur dioxide, Nitrogen dioxide, Carbon monoxide, and Methane results in an increase in the daily count of patients of Pneumonia. Max temperature and occurrence of pneumonia are indirectly proportionate which indicates that as temperatures rise, the incidence of pneumonia tends to decrease. This observation highlights the potential influence of temperature on the prevalence of pneumonia, suggesting that lower temperatures might create a more favourable environment for the transmission and development of this respiratory illness.

2.3.2 Dataset Pre-processing

Real data frequently needs to go through pre-processing in order to be used for proper time series analysis. Estimating missing values, eliminating outliers, and handling non-stationarity, multicollinearity, and imbalanced data are some examples of data pre-processing.

2.3.2.1 Missing Data. Graph 1 depicts the percentage of missing values and it is observed that overall, only 13% of data is missing. The methods used to handle missing data were Knn-Imputer and interpolation. In comparison prediction accuracy of either of these techniques, there is no noticeable difference. The interpolation method was chosen here because missing values are relatively sparse and the data has a smooth structure also, the time complexity of the interpolation method is O(n) which is comparatively smaller than the Knn-Imputer method $O(n^2)$.

2.3.2.2 Outliers. "In statistics, an outlier is a data point that differs significantly from other observations [8]". From Figure 1, it is observed that Wind speed, Rainfall, and Aerosol Optical Depth had a comparatively large number of outliers. However, in this case, outliers represent extreme weather events, it is important to include them in the analysis to understand their impact on long-term trends.

2.3.2.3 Multicollinearity. The correlation analysis revealed a strong correlation between PM10, PM2.5, SO2, and NO2, indicating a potential issue of multicollinearity in the dataset. To address this, we employed two methods: Variance Inflation Factor (VIF) [9] and Principal Component Analysis (PCA).

The VIF method is commonly used to assess multicollinearity by calculating the VIF values for each predictor variable. Higher VIF values indicate a stronger correlation between variables. In our case, applying the VIF method resulted in a reduction of multicollinearity; however, surprisingly, it led to a decrease in the accuracy of the final model.

To further explore the impact of multicollinearity, we also applied PCA, which is a dimensionality reduction technique. PCA transforms the original variables into a new set of uncorrelated variables called principal components. By selecting a subset of these components, we aimed to capture the majority of the variance in the data while minimizing the impact of multicollinearity. However, similar to the VIF approach, the accuracy of the final model decreased instead of improving.

Considering the unexpected decrease in model accuracy after handling multicollinearity through VIF and PCA, we made the decision to retain all the weather and air pollutants in the dataset.

2.3.2.4 Imbalance in the dataset. In regression, an imbalanced output variable refers to a situation in which the distribution of the target variable is not evenly distributed across the range of values. This can occur in regression when there is a significant number of observations with values clustered around one end of the range, leading to a skewed distribution. Figure 4 depicts the histogram and kernel density of ARI and Pneumonia respectively.

As Figure 4 indicates very clearly that both output variables are imbalanced. Transformation of data can be one way to address the problem of imbalance in regression. Most of the methods do not work well when there are many zero values. Since the data has more zero values Yeo-Johnson transformation method [10] was used.

2.3.2.5 Stationarity and trend. For statistical methods, it is necessary to check whether the time series is stationary or non-stationary because it affects the accuracy of the final model but most machine learning models can handle non-stationary data. To check whether the time series is stationary or non-stationary two statistical tests KPSS (Kwiatkowski-Phillips-Schmidt-Shin) and ADF (Augmented Dickey-Fuller) were applied.

From Table 2 it is observed that time series CH_4 and binary time series Lockdown are non-stationary, while all other time series are either stationary or trend stationary. It is important to have stationary data to ensure results will be consistent. Extra Tree Regressor has been found effective in handling non-stationary data [11]. This happens due to the fact that in an Extra tree regressor, there is a random selection of features and training samples that helps to reduce overfitting and capture more diverse data patterns. This can be especially useful in non-stationary environments where the data distribution can change over time.

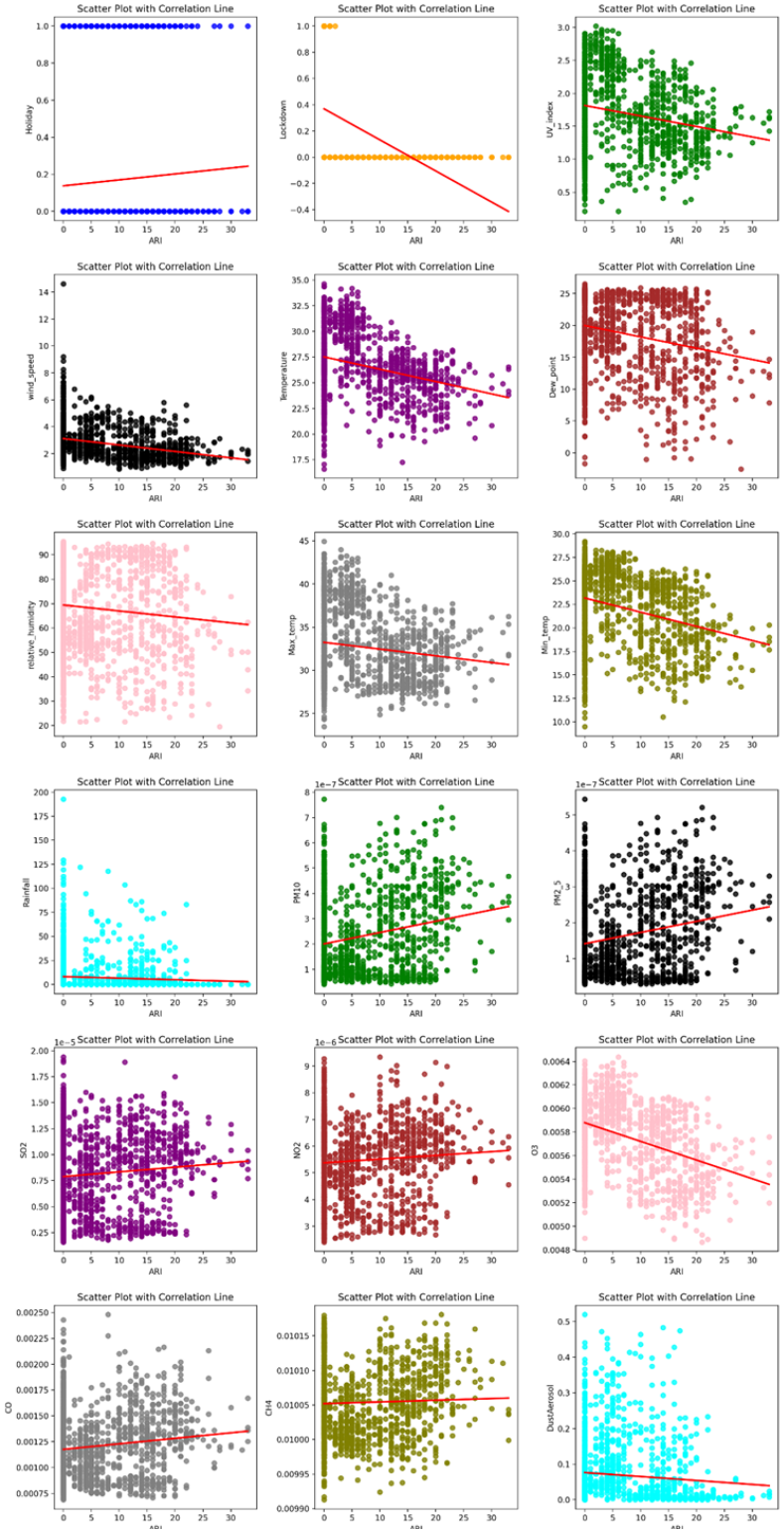

Fig 3. Correlation of daily count of Pneumonia with other features

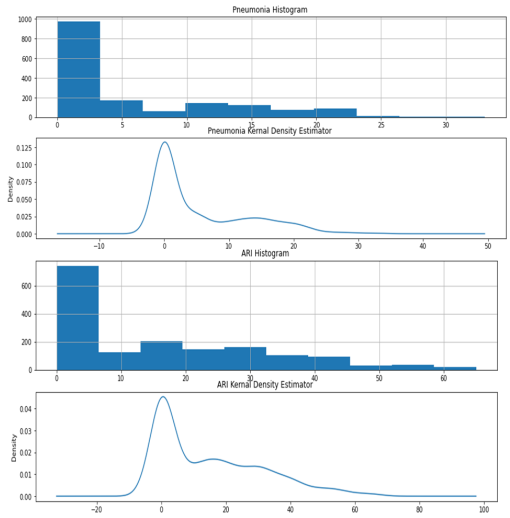

Fig 4. Histogram and Kernel Density of ARI and Pneumonia

Table 2. Stationarity and trend

Column1	KPSS	ADF	Conclusion
Holiday	Non Stationary	Stationary	Trend Stationary
Lockdown	Non Stationary	Non-Stationary	**Non Stationary**
UV_index	Stationary	Stationary	Stationary
wind_speed	Non Stationary	Stationary	Trend Stationary
Temperature	Stationary	Stationary	Stationary
Dew_point	Stationary	Non Stationary	Trend Stationary
Max_temp	Stationary	Non Stationary	Trend Stationary
Min_temp	Stationary	Stationary	Stationary
relative_humidity	Stationary	Non Stationary	Trend Stationary
Rainfall	Stationary	Stationary	Stationary
PM10	Stationary	Stationary	Stationary
PM2_5	Stationary	Stationary	Stationary
SO2	Stationary	Non Stationary	Trend Stationary
NO2	Stationary	Stationary	Stationary
O3	Non Stationary	Stationary	Trend Stationary
CO	Stationary	Stationary	Stationary
CH4	Non Stationary	Non Stationary	**Non Stationary**
Aerosol_Optical_Depth	Stationary	Stationary	Stationary
ARI	Non Stationary	Stationary	Trend Stationary
Pneumonia	Non Stationary	Stationary	Trend Stationary

2.4 Proposed Model

The objective of this research was to predict the occurrence of respiratory disease with a focus on ARI and Pneumonia due to air pollution, especially near the dumping ground. Initially, data was collected from various sources as mentioned in section 2.3, pre-processed, and combined to create a complete dataset. Two activities are necessary before prediction. Those activities are hyperparameter tuning and determining the lag (in days). A hyperparameter is a parameter whose value is predetermined before the machine learning procedure starts. The speed and effectiveness of the learning process are influenced by algorithm hyperparameters. Hyperparameters are crucial since they can significantly affect the performance of the model being trained as well as the way the training algorithm behaves. The GridSearchCV method, a widely used technique in machine learning, was employed for hyperparameter tuning. This method systematically searches for the optimal combination of hyperparameters through an exhaustive grid search, allowing us to fine-tune the model's parameters and improve its performance. By meticulously exploring the specified parameter values, the GridSearchCV method assists in identifying the most suitable set of hyperparameters that yield the best results for our predictive model. This rigorous approach not only ensures the robustness of our model but also enhances its predictive capabilities, enabling us to achieve superior accuracy in forecasting ARI and Pneumonia cases.

Respiratory diseases express themselves after a certain period of time. Hence initial experiments were performed to identify the amount of time after (in days) which the diseases, considered ARI and Pneumonia, express themselves in the patients i.e., lag.

Basic statistical techniques like VAR and VARMAX were used to predict the occurrence. However, the results indicated that the period after which this disease express themselves was the same. Hence it was decided to use machine learning algorithms to identify the lag. Fifteen machine-learning models were applied on the pre-processed dataset. The observations which are reported and discussed in detail in the result and discussion section showed that the Gaussian regressor and Extra tree regressor were the two best-performing algorithms. However, the execution time increased by 900secs. in the Extra tree regressor. Additional time was required to find the lag and hyperparameter tuning.

Hence Multioutput ensemble model was proposed using a Gaussian regressor and an Extra tree regressor. This ensemble model is comprised of two phases. In the first phase, a Gaussian regressor was used to extract lag (the number of days required to affect meteorological parameters and air pollutants on the respiratory disease) for the occurrence of ARI and Pneumonia in the least possible time. In the next phase, an Extra tree regressor was used to predict the daily occurrence of ARI and Pneumonia for the lag values obtained from phase 1. Figure 5 pictorially represents this ensemble modeling process.

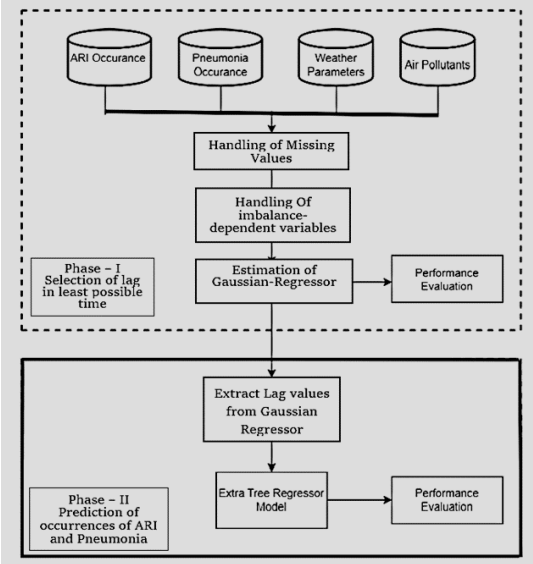

Fig 5. Structure of the proposed ensemble machine learning model

Since our data was time series data, instead of randomly splitting data into training and testing time series cross-validation method was used. Nested cross-validation integrates feature selection and parameter adjustment into machine learning model optimization to improve model accuracy while lowering the effects of overfitting.

2.5 Experiment Setting

The machine used for conducting experiments with the following configuration, an Intel Core i5 @ 1.19 GHz and 8 GB of memory. Python software is used for model creation and prototyping because it includes publicly accessible library sets for machine learning and statistical methods like Scikit-learn, and Matplotlib. Modeling tests were run to confirm the efficacy of the proposed Ensemble Gaussian regressor and Extra tree regressor algorithm. Data and results were plotted using the Python 2D graphing tool included in the Matplotlib. The effectiveness of the model was examined using PerMetrics, a Python module for performance measurements of machine learning models. All experiments were also performed on Google colab and got similar results.

3 Results and Discussion

Initially, a statistical technique VAR was used. To handle the non-stationarity difference method was used but got comparatively better results without using the difference method. It gave the resultant lag same for both diseases. However, according to domain experts and literature, the lag is generally different for both diseases where it is more for Pneumonia as compared to ARI[12,13]. So, additional machine learning models were tried to find the appropriate lag. Those results are shown in Table 3.

As can be seen from Table 3 results for various evaluation metrics were found however for further analysis only RMSE, FastDTW, and execution time have been considered. For ARI and Pneumonia is it observed that RMSE, as well as FastDTW, were least in Gaussian regressor and Extra tree regressor. Also, the Execution time of the Extra tree regressor was comparatively larger than the Gaussian regressor. Our objective was to minimize both error and execution time. Since an Ensemble model was proposed as discussed in the earlier section. For optimizing the time, the Gaussian regressor was used and for reducing the error Extra tree regressor was used.

Table 4 shows that the performance of the Gaussian regressor was best in terms of finding the lag value in a short duration with minimum RMSE for ARI and Pneumonia as compared to other algorithms except for the Extra Tree regressor. Hence Gaussian regressor was selected for phase 1. However, the Extra tree regressor took excess time to find lag (in days) and perform hyperparameter tuning. But the performance of the Extra tree regressor was best in all aspects excluding time criteria. Hence Extra tree regressor was chosen for phase 2 i.e., for hyperparameter tuning and forecasting of the occurrence of ARI and Pneumonia for the lag values obtained from the execution of the Gaussian regressor.

Figure 6 depicts the relationship between the actual values and the predicted values obtained from our Extra Tree regression model for the occurrence of ARI and Pneumonia, respectively. These Figures visually demonstrate the association between the true values and the corresponding predictions. The plot reveals a positive correlation between the actual and predicted values, indicating that the model captures the underlying patterns and trends in the data.

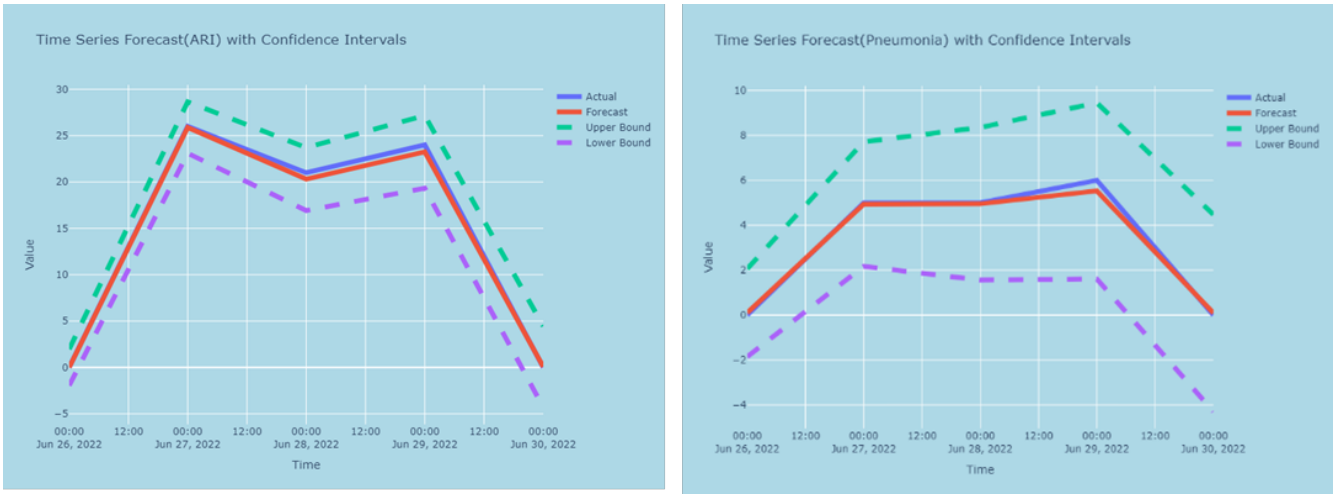

Fig 6. Actual Vs. Predicted (ARI and Pneumonia) with a confidence interval

Furthermore, the Figure 6 also incorporate 95% confidence intervals around the predicted values. These confidence intervals provide a measure of uncertainty and indicate the range within which we can be reasonably confident that the true values lie. By including these intervals, we can assess the reliability and precision of our predictions.

Table 3. Comparison of performance of Machine Learning algorithms for the occurrence of ARI and Pneumonia

Algorithm	lag	Performance Evaluation (ARI)						
		ME	RMSE	MASE	R2	CO	FastDTW	MSE
SGD Regressor	7	17.6165	9.3658	0.501	0.3595	0.5999	36.3255	87.7175
MLP Regressor	7	12.9843	9.7909	0.5935	0.3001	0.8756	43.02742	95.8608
Linear	7	20.3455	13.5556	0.8061	-0.3417	0.4127	58.44385	183.7531
Random Forest	7	14.4786	10.1394	0.5903	0.2494	0.6692	42.79861	102.8077
XGBOOST	7	12.8262	9.412	0.5541	0.3532	0.6916	40.17455	88.5867
Decision Tree	7	15.5984	9.0749	0.5097	0.3987	0.7068	36.95026	82.3543
Support Vector	7	15.7222	11.0098	0.6744	0.115	0.5656	48.89663	121.2147
Ridge	7	20.3184	13.5458	0.8044	-0.3397	0.4331	58.3219	183.4898
Lasso	7	14.2787	10.8415	0.6847	0.1418	0.5626	49.63858	117.5374
Lars Lasso	7	20.3455	13.5556	0.8061	-0.3417	0.4127	58.44385	183.7531
Bayesian Ridge	7	20.3634	13.5623	0.8041	-0.343	0.4473	58.29509	183.9358
Elastic Net	7	19.074	12.7409	0.7664	-0.1852	0.4349	55.56053	162.3304
VAR_DIFF	6	10.1242	4.9153	0.2319	0.8236	0.9687	16.81588	24.16
VAR	7	7.6388	3.8122	0.198	0.8939	0.9785	14.35239	14.5327
Gaussian	7	3.9719	2.1789	0.1141	0.9653	0.9972	8.26983	4.7475
Extra tree	7	0.7521	0.4601	0.0225	0.9985	0.9998	1.63346	0.2117
Algorithm	lag	Performance Evaluation (Pneumonia)						
		ME	MSE	RMSE	MASE	R2	CO	FastDTW
SGD Regressor	7	4.0406	4.4184	2.102	0.5348	0.3652	0.6044	8.02251
MLP Regressor	7	4.1665	8.3124	2.8831	0.8021	-0.1943	0.9019	12.03168
Linear	7	4.6232	12.4294	3.5255	0.9942	-0.7858	0.3415	13.78085
Random Forest	7	4.6115	8.5424	2.9227	0.819	-0.2274	0.5142	11.87541
XGBOOST	8	2.8347	5.6584	2.3787	0.7803	0.187	0.9904	11.7049
Decision Tree	7	3.4314	4.6635	2.1595	0.6196	0.33	0.6065	9.293791
Support Vector	7	3.2317	6.9603	2.6382	0.7959	0	0.5496	11.93915
Ridge	7	4.6152	12.4584	3.5296	0.9926	-0.79	0.3676	13.80266
Lasso	7	4.0057	9.2319	3.0384	0.9074	-0.3264	0.3551	13.61136
Lars Lasso	7	4.6232	12.4294	3.5255	0.9942	-0.7858	0.3415	13.78085
Bayesian Ridge	7	4.6175	12.5478	3.5423	0.9933	-0.8028	0.3851	13.88288
Elastic Net	7	4.3132	10.8187	3.2892	0.9437	-0.5544	0.3345	13.41909
VAR_DIFF	6	2.7775	2.0231	1.4224	0.3753	0.7093	0.9262	5.629989
VAR	7	2.6637	2.0509	1.4321	0.3922	0.7053	0.9348	5.883582
Gaussian	8	1.1637	0.4685	0.6844	0.1906	0.9327	0.9941	2.858663
Extra tree	8	0.4703	0.0491	0.2217	0.0512	0.9929	0.9986	0.76824

The results of our study revealed significant improvements in the efficacy of the proposed ensemble model compared to previous works in the field. We conducted experiments on a dataset consisting of 1673 records. Our algorithm achieved an R2 value of 99%, outperforming the approach proposed by Khatri K L, Tamil L S [(2)] and Khaiwal R, Bahadur S S, Katoch V, Bhardwaj S, Kaur-Sidhu M, Gupta M, et al. [(4)] by a margin of 17% and 12% respectively. Also, our algorithm demonstrated higher R2 and RMSE values, with R2 reaching 99% and RMSE 0.46 and 0.22, compared to the R2 of 67% and RMSE of 13.9 reported by Ku Y, Kwon SB, Yoon JH, Mun SK, and Chang M [(3)]. In addition to the above findings, our algorithm demonstrates remarkably low forecasting errors for ARI and Pneumonia, with rates of 2.25% and 5.12% respectively. These values are substantially lower compared to the 8.17% reported by Kim M. S., Lee J. H., Jang Y. J., Lee C. H., Choi J. H., and Sung T. E. in their study [(5)]. These findings suggest that our approach effectively enhances the accuracy of ARI and Pneumonia forecasting, showcasing its potential as a robust tool for improved healthcare management and proactive disease prevention strategies.

Table 4. Performance evaluation of the models

Algorithm	Gaussian Regressor		Extra tree Regressor		Multioutput Ensemble Model	
Lag	7	8	7	8	7	8
Disease	ARI	Pneumonia	ARI	Pneumonia	ARI	Pneumonia
ME	3.972	1.1637	0.752	0.4703	0.7521	0.4703
MSE	4.748	0.4685	0.212	0.0491	0.2117	0.0491
RMSE	2.179	0.6844	0.46	0.2217	0.4601	0.2217
MASE	0.114	0.1906	0.023	0.0512	0.0225	0.0512
R2	0.965	0.9327	0.999	0.9929	0.9985	0.9929
CO	0.997	0.9941	0.9998	0.9986	0.9998	0.9986
FastDTW	8.27	2.859	1.633	0.768	1.6335	0.7682
Execution time	56.93427896		40937.34948		956	

However, there were some limitations to this study. Only the pediatric outpatient department was studied excluding general OPD considering only two diseases ARI and Pneumonia. Finally, the selected models should be updated periodically, in order to improve the forecasting accuracy of the occurrence of ARI and Pneumonia. Also, it is observed that hyperparameter tuning requires more time than forecasting. In future work, we will focus on minimizing the time complexity.

4 Conclusion

The proposed work encompasses the utilization of various Machine Learning techniques, making the selection of a specific method a challenging task. To enhance the efficiency and maximize the output of the Machine Learning algorithm, several steps, such as normalization, cross-validation, and hyper-parameter tuning, have been implemented.

Yeo-Johnson method was successfully employed to handle the issue of data imbalance. Cross-validation has significantly improved both the dataset and the algorithm. By addressing the issue of over-fitting, cross-validation enables the Machine Learning model to generalize independent samples and accurately predict outcomes. Additionally, hyper-parameter tuning is conducted to regulate the behavior of the Machine Learning algorithm, ensuring optimal performance before the training process commences.

The ensemble model (extra tree regressor, gaussian regressor) achieved RMSE of 0.46 and 0.22 resp. for ARI and Pneumonia. These low RMSE values indicate that the ensemble model is performing well in predicting the outcomes for these conditions. The fact that the ensemble model outperformed other models despite considerable data gaps in the dataset highlights its robustness and effectiveness in handling imbalanced data, outliers, multicollinearity, and data with non-stationarity

While the current model focuses on predicting ARI and Pneumonia cases of pediatric OPD, future enhancements will involve incorporating a Deep Learning algorithm to enable the prediction of various other diseases. This expansion will further enrich the model's capabilities and increase its utility.

Furthermore, ensemble modeling proved to be more effective compared to using a single Machine Learning algorithm. Each time series exhibited a unique pattern, highlighting the need for a distinct prediction model for each time series with different lags. By leveraging these predicted values, pediatric OPDs can optimize resource allocation in advance, leading to improved efficiency and better patient care.

It is important to acknowledge that future studies should address the limitations highlighted in the Discussion section to further refine and enhance the proposed model. By addressing these limitations, the model's performance and applicability can be improved, contributing to more accurate predictions and better resource allocation in healthcare settings.

Acknowledgment

The authors express their sincere gratitude to Dr. Priyanka Kulkarni BAMS, MD - Ayurveda, and, Dr. Sonali Renushe BAMS, MD - Ayurveda for their essential guidance and support in assisting them to comprehend the medical component of this research work which has been substantially enhanced by their insights into it.

References

1) World health statistics 2023: Monitoring health for the SDGs, sustainable development . . Available from: https://www.who.int/publications/i/item/9789240074323.

2) Khatri KL, Tamil LS. Early Detection of Peak Demand Days of Chronic Respiratory Diseases Emergency Department Visits Using Artificial Neural Networks. *IEEE Journal of Biomedical and Health Informatics*. 2018;22(1):285–290. Available from: https://doi.org/10.1109/JBHI.2017.2698418.
3) Ku Y, Kwon SB, Yoon JH, Mun SK, Chang M. Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors. *Clinical and Experimental Otorhinolaryngology*. 2022;15(2):168–176. Available from: https://doi.org/10.21053/ceo.2021.01536.
4) Ravindra K, Bahadur SS, Katoch V, Bhardwaj S, Kaur-Sidhu M, Gupta M, et al. Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections. *Science of The Total Environment*. 2023;858:159509. Available from: https://doi.org/10.1016/j.scitotenv.2022.159509.
5) Kim MS, Lee JH, Jang YJ, Lee CH, Choi JH, Sung TE. Hybrid Deep Learning Algorithm with Open Innovation Perspective: A Prediction Model of Asthmatic Occurrence. *Sustainability*. 2020;12(15):6143. Available from: https://doi.org/10.3390/su12156143.
6) Nasa Power. . Available from: https://powerlarcnasagov/data-access-viewer/.
7) Inness A, Ades M, Agustí-Panareda A, Barré J, Benedictow A, Blechschmidt AM, et al. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*. 2019;19(6):3515–3556. Available from: https://doi.org/10.5194/acp-19-3515-2019.
8) Outliers. . Available from: https://en.wikipedia.org/wiki/Outlier.
9) Variance inflation factor. . Available from: https://en.wikipedia.org/wiki/Variance_inflation_factor.
10) Sun L, Hu N, Ye Y, Tan W, Wu M, Wang X, et al. Ensemble stacking rockburst prediction model based on Yeo–Johnson, K-means SMOTE, and optimal rockburst feature dimension determination. *Scientific Reports*. 2022;12(1):15352. Available from: https://doi.org/10.1038/s41598-022-19669-5.
11) Chowdhury S, Lin Y, Liaw B, Kerby L. Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance. *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. 2022;p. 17–25. Available from: https://doi.org/10.1109/IDSTA55301.2022.9923169.
12) Respiratory Infection . . Available from: https://www.oakgov.com/health/information/Pages/Upper-Respiratory-Infection.
13) American Lung Association. Available from: https://www.lung.org/lung-health-diseases/lung-disease-lookup/pneumonia/symptoms-and-diagnosis.