Comparative Study of Ambient Air Quality Prediction System Using Machine Learning to Predict Air Quality in Smart City

Gopal Sakarkar, Sofia Pillai, C. V. Rao, Atharva Peshkar and Shreyas Malewar

Abstract It is a herculean task to predict air quality of a particular area due to indefinite characteristics. As air pollution is a complex mixture of toxic air components that include ozone (O_3), particulate matter 2.5_m (PM2:5), SO_2, RSPM, SPM and nitrogen dioxide (NO_2). These small particles penetrate deep into the alveoli as far as the bronchioles, interfering with a gas exchange within the lungs. Though research is being conducted in environmental science to evaluate the severe impact of particulate matters on public health. The capital city of Maharashtra, Nagpur is used as a case study since nearly ten thousand motor vehicles are being registered in Nagpur on a monthly basis contributing exponentially to air pollution. Various machine Learning-based algorithms are checked to compare and to find out the predictive analysis using available dataset. After comparing seven different machine learning algorithms, Boosted Random Forest algorithm was found out to be the most accurate predictive algorithm, with the maximum coefficient of determination and less mean absolute error.

Keywords Air quality · Forecasting system · Machine learning · Cancer · Forecasting · Ensemble methods · Random forest

G. Sakarkar (✉) · S. Pillai · A. Peshkar · S. Malewar
G H Raisoni College of Engineering, Nagpur, India
e-mail: gopal.sakarkar@raisoni.net

S. Pillai
e-mail: sofia.pillai@raisoni.net

A. Peshkar
e-mail: peshkar_atharva.ghrceit@raisoni.net

S. Malewar
e-mail: malewar_shreyas.ghrcecs@raisoni.net

C. V. Rao
Former-Sr. Scientist, NEERI, Nagpur, India

M. Dutta et al. (eds.), *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India*, Lecture Notes in Networks and Systems 116,
https://doi.org/10.1007/978-981-15-3020-3_16

1 Introduction

We humans have been very successful in the domains of technology, commerce and each and every field of human existence, but through a due course of time, we have inflicted an indelible footprint of several sets of polluting factors. Our environment is being deteriorated innumerable factors like Deforestation, Water pollution, Genetic modification, Ozone Layer Depletion, Air pollution, etc. Each and every type of pollution is lethal, but we have considered Air Pollution as the most detrimental type because it gives the least reaction time to a person. Air pollution might cause Pneumonia, Influenza, Bronchitis, etc. Such health complications can be controlled to a certain level by raising the awareness about air quality conditions in urban areas, enabling the citizens to limit their day to day activities in the cases of elevated pollution episodes and planning their routes and schedule to avoid inhaling harmful pollutants, by using machine learning models to forecast air pollution in areas after a certain time duration.

Air pollution is caused for the most part by transportation, fuel ignition in stationary sources, consuming of petroleum derivatives like coal, wood, dry grass and development action. Engine vehicles produce abnormal amounts of Carbon Monoxide (CO) and Hydrocarbons (HC) and Nitrogen Oxides (NO). Development exercises, industrial chimneys, terrible streets and consumption of petroleum products are in charge of dust pollution. Private and commercial exercises additionally add to Air Pollution [1].

2 Literature Review

Following the Rio De Janerio Earth summit in 1992 Environmentalists and Researchers worldwide have been focussing on Air Quality and weather prediction systems. Elia Dragomir proposed a solution to Predict Air Quality using the K Nearest Neighbour Technique. She focussed on SO_2, CO, NO, NO_2 and O_3 pollutants only. She referenced the prediction results to a fuzzy set of Quality Index and concluded that best results are yielded in 10 fold cross-validation. She was unable to reveal the characteristics of the dataset.

Qi Feng in his research paper Improving Neural Network Prediction Accuracy for PM10 Individual Air Quality Index Pollution Levels stressed on pollutants having a diameter less than <10 μm (PM10) in two major cities of China. The reason for the generation of fugitive dust was due to construction activities and was interlinked with Construction Influence Index. His Neural Network Models were based on perceptron, Elman and Support Vector Machine. The dataset was decomposed into wavelet representations and then wavelet representations were predicted.

His predictions were tested between 1 January 2005, and 31 December 2011, at six monitoring stations situated within the urban area of the city of Wuhan, China.

It yielded better results than previous models but he only focussed on pollutants <10 μm (PM10).

Ozone and PM10 were two pollutants which were emphasized Giorgio Corani in research paper Air quality prediction in Milan: feedforward neural networks, pruned neural networks and lazy learning. Feedforward Neural Networks (FFNNs), Pruned Neural Networks (PNNs) and Lazy Learning (LL) were the foundation of entire statistical prediction. Lazy learning provided the best results on the basis of evaluation metrics such as correlation and mean absolute error.

3 Methodology

3.1 Dataset

The data used for comparing the predictive accuracy of the models was accessed from Kaggle uploaded by Shruti Bhargava as 'India Air Quality Data' (https://www.kaggle.com/shrutibhargava94/india-air-quality-data), it is a highly cleaned and compiled version of the 'Historical Daily Ambient Air Quality Data' released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

The columns of the data include Station Code, State, City, Agency, Type of Area, concentrations of Sulphur Dioxide, Nitrogen Dioxide, Respirable Particulate Matter, Suspended Particulate Matter, the Location of the monitoring area, PSI 2.5, Date of recording. The columns in the data are of numeric and string type and contain categorical variables. Thus, it is essential to appropriately encode categorical variables. For that purpose, we have used the LabelEncoder function from Scikit Learn Machine Learning Library.

Since the missing values and outliers can have a great negative impact on the predictive accuracy of the models, the authors have imputed the missing values in the numeric columns (conc. of the pollutants) with the mean values. To identify and remove outliers, the ZScore has been calculated, with −3 and 3 being the threshold to retain the values in the dataset.

3.2 Evaluation Metric

3.2.1 Mean Absolute Error

In statistics, the mean absolute error is the quantity used to measure how close forecasts and predictions are to the actual outcomes. Mean absolute error performs in ways that disregard the directions of over or under prediction [11].

Table 1 Boosted random forest results

Pollutants	MAE
SO_2	1.7091
NO_2	3.8402
RSPM	21.2891
SPM	19.8359
PM2.5	0.0426

A lower value of MAE indicates a small difference between the pollutant concentration predictions by the model and the actual concentrations, averaged over the entire dataset, making the model a good fit for the dataset, whereas a higher value of MAE proves the opposite. MAE is an important metric for this study due to the highly fluctuating nature of the data that we are dealing with.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} \left| y_j - \hat{y}_j \right|$$

3.3 Fitting Data to Various Regression Models

3.3.1 Boosted Random Forest Regression

An ensemble learning method is implemented for classification and regression by creating a multitude of decision trees during training. Applying a boost algorithm to Forest Regression reduces the bias and variance.

Implementation of Boosted Random Forest regression on the dataset used in the study yielded the following results (Table 1) (Fig. 1).

The spike in the performance can be attributed to the ensemble method, where the final prediction depends on the predictions of individual decision trees while the boosting corrects the wrong predictions. This makes the model much more robust to fluctuations also helping avoid the model from developing bias and variance.

3.3.2 Deep Neural Network

It is a neural network with multiple neural layers which process data by advanced mathematical modelling wherein each mathematical manipulation correlates to a single layer. An object is shown as a layered composition of primitives in a compositional model of DNN architecture.

Implementation of Deep Neural Network on the dataset used in the study yielded the following results (Table 2) (Fig. 2).

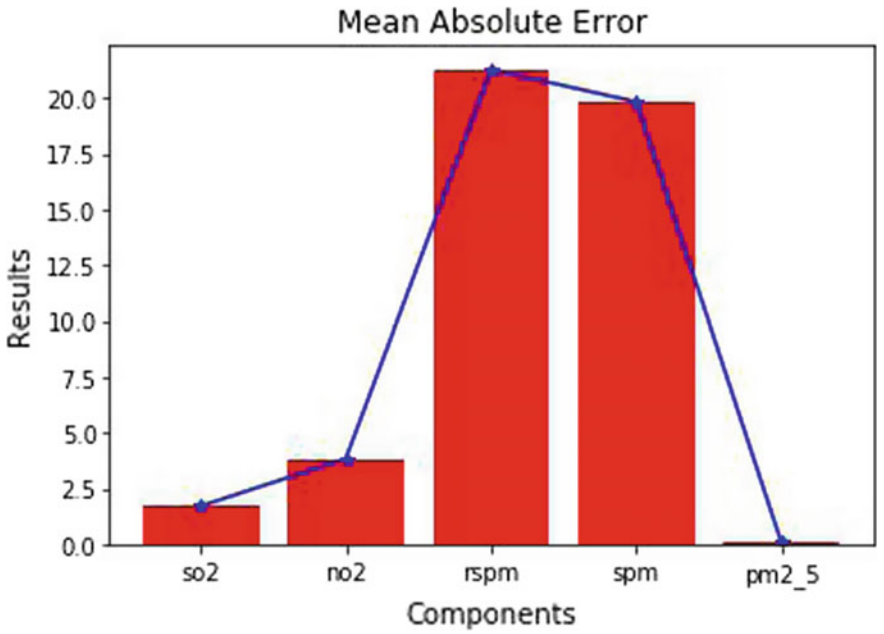

Fig. 1 MAE of BRF

Table 2 Results of deep neural network

	MAE
SO_2	5.9781
NO_2	12.1681
RSPM	94.3660
SPM	162.8340
PM2.5	0.5048

The highly fluctuating nature of the data causes the model to overfit the dataset, thereby giving subpar results on all the pollutants.

3.3.3 Stochastic Gradient Descent (SGD)

Gradient descent is used in Machine Learning algorithms to minimize a cost function to global minima. It is iterated innumerable times to attain the optimal value of desired parameters. But when considering significantly larger datasets the standard gradient descent algorithm ceases to work efficiently, therefore, a batch of values is randomly selected from the entire dataset and then Gradient Descent is applied on the values which prove to be relatively more efficient.

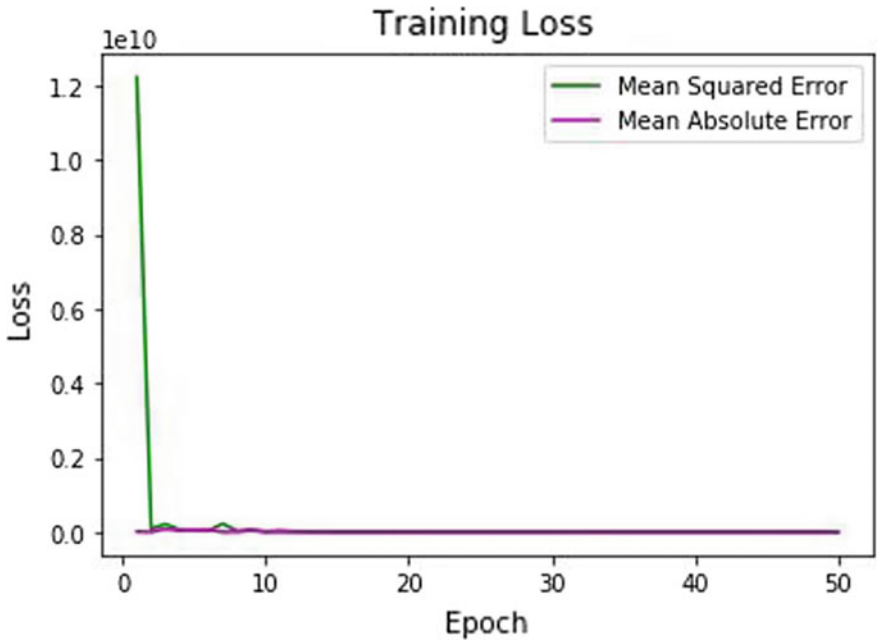

Fig. 2 MAE of DNN

Table 3 Results of stochastic gradient descent

	MAE
SO_2	5.3493
NO_2	1.3719×10^{19}
RSPM	1.24341×10^{19}
SPM	7.7897×10^{19}
PM2.5	7.0088

Implementation of Stochastic Gradient Descent on the dataset used in the study yielded the following results (Table 3) (Fig. 3).

The SGD is unable to learn from the data giving performing the worst at predicting the concentrations of pollutants, which can be inferred by the negative and exponentially large values of COD.

4 Conclusion

The algorithm performing the best at predicting the target pollutant concentration is Boosted Random Forest which can be devised from the high values of coefficient of determination for all the pollutants considered in this study. The performance of

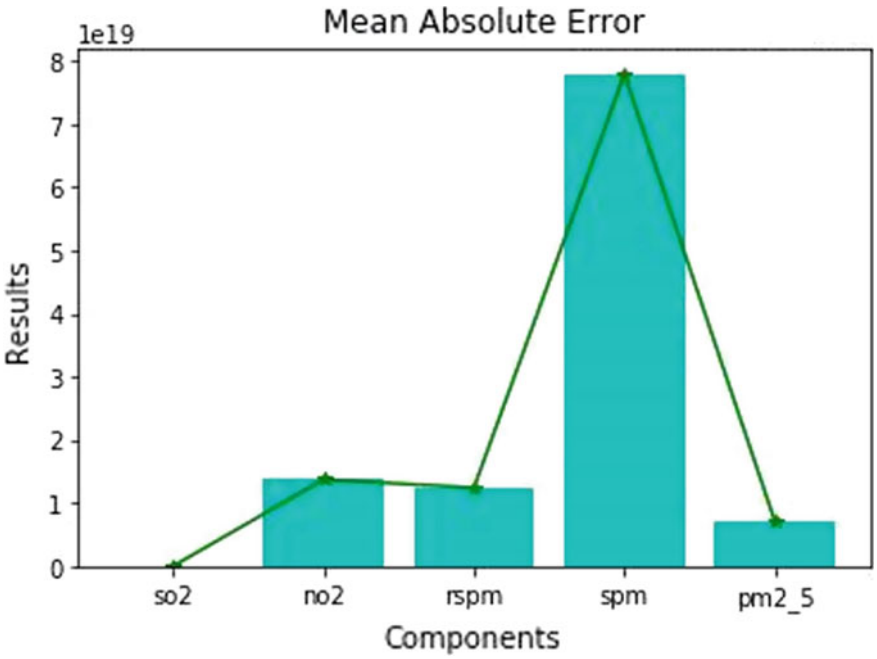

Fig. 3 MAE of SGD

a model can be generalized to all the pollutants. Thus a single well-tuned model performs well on all the pollutants. The model captures the patterns in the gases in a much better way compared to the particulate pollutants like SPM, RSPM and PM2.5. The Boosted Random Forest is a much more robust model for future work.

Acknowledgements This study was supported by Microsoft under the Microsoft AI for Earth Grant.

References

1. T.-C. Bui, V.-D. Le, S.-K. Cha, A deep learning approach for forecasting air pollution in South Korea using LSTM. Environ. Sci., Comput. Sci., Math. (2018)
2. S.B. Kotsiantis, D. Kanellopoulos, P.E. Pintelas, Data preprocessing for supervised learning. Int. J. Comput. Sci. **1**(2), 111–117 (2006)
3. R.G.D. Steel, J.H. Torrie, in *Principles and Procedures of Statistics with Special Reference to the Biological Sciences* (McGraw Hill, 1960)
4. C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**, 79–82, 19 Dec 2005
5. E.L. Lehmann, G. Casella, *Theory of Point Estimation*, 2nd edn. (Springer, New York 1998). ISBN 978-0-387-98502-2. MR 1639875

6. H.L. Seal, The historical development of the Gauss linear model. Biometrika **54**(1/2), 1–24 (1967)
7. R. Quinlan, Learning efficient classification procedures. in *Machine Learning: an Artificial Intelligence Approach*, Michalski, Carbonell, Mitchell eds. by (Morgan Kaufmann, 1983), pp. 463–482, https://doi.org/10.1007/978-3-662-12405-5_15
8. L. Breiman, Bias, variance, and arcing classifiers
9. T.K. Ho, The random subspace method for constructing decision forests (PDF). IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
10. Y. Bengio, Learning deep architectures for AI, Trends Mach. Learn. **2**(1), 1–127 (2009), CiteSeerX/10.1.1.701.9550
11. S. Mei, A mean field view of the landscape of two-layer neural net-works. Proc. Natl. Acad. Sci. **115**(33), E7665–E7671 (2018)