Air Pollution Prediction using Machine Learning

Shreyas Simu*,Varsha Turkar*, Rohit Martires*, Vranda Asolkar*, Swizel Monteiro*, Vaylon Fernandes*, and Vassant Salgaoncar†
*ETC Department, Don Bosco College Engineering, Fatorda, Goa, India
†Enviroguru Private Limited, Goa, India

2020 IEEE Bombay Section Signature Conference (IBSSC) | 978-1-7281-8993-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/IBSSC51096.2020.9332184

***Abstract*—Industrial pollution is one of the most serious problems faced today. Long-term exposure to air pollution causes severe health issues including respiratory and lung disorders. Presently laws regarding industrial pollution monitoring and control are not stringent enough. The working dataset includes parameters of air in terms of ambient air as well as of the stack emission. On this data, various Machine Learning (ML) algorithms were applied for prediction of emission rate, and comparative analysis is done. These algorithms were implemented using python and the mean square error of each of these was measured to check for accuracy. It was observed that among all classifiers, the Multi-layer perceptron model was seen to have the least error. The air dispersion models are then applied to the predicted emission rate to calculate the dispersion of pollutants from the source that is at the stack level.**

***Index Terms*—Machine Learning, Air dispersion models, Air Polluton**

I. Introduction

With the rapid growth of the economy, industrial activities are increasing more frequently, leading to a faster rate of pollution. This rate is increasing and if not kept in check can cause harmful effects to mankind and other living organisms. Environmental pollution is a grave problem faced by all humans and other life forms, pollution caused by industries contributing a major share in it. Industrial pollution is a result of factories and other industrial plants which emit harmful by-products and waste into the surroundings. The work of this paper is focused on Air pollution. Air pollution constitutes solid particles and gases which include dust, pollen, and spores. Air pollutants can be largely classified into two categories primary pollutants and secondary pollutants [1]. Primary pollutants are usually produced by processes that directly emit such as vehicle exhausts, and secondary pollutants are both emitted directly and formed from other primary pollutants such as photochemical smog.

A. Motivation

The Ministry of Environment, Forest and Climate Change (MoEFCC) published the Draft Environment Laws (Amendment) Bill on October 7, 2015 [2]. The aim of the Draft bill was to fine or impose a penalty on individuals or groups that violate these laws. Since the ill effect of industrial pollution is alarming, governments worldwide have intervened in the same. The central aim of the Paris agreement is to strengthen the global response to the threat of climate change. Keeping a check on global temperature rise this century to be well below 2 degrees Celsius [3]. Small changes could make a huge difference, like purchasing energy-efficient equipment and recycled materials for industries and having industrial pollution control policies in place, and adhering to them strictly. These could help a lot in reducing industrial pollution to great extent.

B. Objectives of Research

The primary objective of this research is to monitor the air pollutants emitted from an industry/factory and predict the future spread of these pollutants also to output these results in the form of reports for industries/factories. To accomplish this work, two objectives were identified that are:

1) To develop a module that uses Machine Learning (ML) models with pollution data to predict future emission parameter values.
2) To create a module that simulates the movement of fluid particles (pollutants) in the air using air dispersion models with meteorological data.

The rest of the paper is as follows, we have presented literature review in section II, methodology in section III, results and discussion in section IV and conclusion in section V.

II. Literature Survey

The authors of [4] proposed an air quality monitoring system with and an early warning system, including an assessment module and a forecasting module. The model was able to successfully identify the major pollutants in two cities. Their results showed that the proposed model had good accuracy and

stability compared to other neural network systems.

In [5] the authors compared research work on air quality evaluation based on big data analytics, machine learning models, and techniques and also highlighted some future resource issues, needs, and challenges. This paper also mentioned that the accuracy of air quality evaluation and assessment was affected by hardware issues. It was also mentioned that due to this issue there was a strong need for research in data quality modeling and automatic real-time validation.

In [6] the authors proposed a diagnostic model for calculating concentration distribution. This model requires measurements of the wind velocity and direction at a certain reference height above the obstacles. It was effectively able to predict 3-D concentration distributions and was also able to identify concentration accumulation at specific and precise points. It also succeeds in predicting concentration distribution both quantitatively and qualitatively.

The authors of [7] proposed a detailed approach to model dispersion which widely aims at combining the advantages of puff models and particle models. This was called Puff Particle Model (PPM). In PPM, a hundred puffs in three-dimensional space are collectively simulated, in comparison to many thousand particles usually required in pure particle models. The overall PPM concept is quite simple, while puff growth is described by the concept of relative dispersion which accounts for eddies smaller than the puff, causing the effect of meandering. The variation between the trajectories of different puffs due to larger eddies, those larger than the actual puff size is simulated by introducing puff center trajectories derived from particle trajectories from a particle model.

III. METHODOLOGY

A. Machine Learning Techniques

Several machine learning algorithms were compared based on accuracy, robustness, and flexibility:

1) Kth Nearest Neighbors (KNN): In this machine learning model, similar things are considered to be in close proximity to each other. The value of K indicates the nearest neighbors that are taken for consideration. This similarity criterion can be based on distance or other similar factors. It is extended as a regression technique wherein the output of the point is the average or median value of its nearest neighbors. In KNN the training data required is very less, which means there is a lack of generalization [8].

2) Support Vector Regression (SVR): SVR is a supervised machine learning algorithm that is being used for both classification and regression problems. In this algorithm, a data point is plotted in n-dimensional space after which it is classified by finding the plane that will differentiate the classes very well. The hyperplane which is considered is a linear separator of any dimension (line, plane, hyperplane). The training points are used in the decision function and are called support vectors [9].

3) Random Forest (RFS): The decision tree is a graph like structure in the form of a tree which shows possible consequences based on the inputs given. Random forest is a collection of such multiple decision trees. The final result is based on the decision of the majority of trees. When a training data set is considered with targets and features, some set of rules are formulated, these rules are used to perform predictions. It identifies the most important features out of all the available features in the training dataset. The larger the no. of trees the more accurate the results will be. Random forest classifier handles missing values and the over-fitting problem does not exist [10].

4) Multilinear Regression (MLR): Several explanatory variables are used in this technique to predict the outcome of a response variable. A linear relationship is modeled between these independent and response variable (dependent). Prediction about one variable is done based on the information about the other variable. Which means a correlation between the independent variables should less [11].

5) Artificial Neural Network (ANN): It is a model that imitates the way a human brain works. The basic entity is a neuron that is based on some mathematical function that collects data/information and classifies it according to some pre-designed architecture. A neural network contains layers of interconnected nodes. Each neural connection is associated with a weight value which is multiplied by the input value. Every neuron also has an activation function that characterizes the output which is used to introduce some non-linearity in the network model [12].

B. Air Dispersion Model

Various air dispersion models were studied and compared based on accuracy and computational complexity:

1) Gaussian Dispersion Model: Gaussian dispersion model is a steady-state system. The concentration of pollution downwind from a source is considered as increasing outward from the stack. The concentration of the pollution is maximum at the source and gradually disperses, where this dispersion follows a Gaussian curve.

2) Lagrangian Model: The Lagrangian model[13] determines the trajectory of air pollutants. The pollutants are tracked as air parcels which move along trajectories determined by wind field, the buoyancy and turbulence effects. The estimation of the concentration field is given by the final distribution of a large number of particles. The particles can be tracked from the source area to the area of reception.

3) Puff - Plume Model: The puff-plume model[14] maintains the advantages of the puff models and also of the plume models. All pollutant particles are grouped in various clusters. These clusters are then treated as Gaussian puffs which are dispersed using the concept of relative diffusion. The center of mass of each puff follows a stochastic trajectory. The particle trajectories of the Lagrangian stochastic dispersion model gives the trajectory of the puffs.

The Gaussian dispersion model has been selected for implementation as it has relatively less computational time as compared to the Lagrangian model in which the computational time increases significantly with an increase in distance. The accuracy of the gaussian model is less compared to other dispersion models, but it is sufficient to estimate the spread of pollutants in the air according to the consulted industry experts.

C. Dataset and Study area

The objective of the prediction module is to predict the emission rate at the stack/chimney. To build and test the prediction model, 1000 dummy data points were generated using Gaussian distribution under the guidance of an industry expert. The feature set is chosen based on what factors will correlate to the emission rate and the type of pollutants emitted from the stack. The feature set consists of independent variables: day, month, type of industry, size of the industry, and output efficiency of industry and dependent variable: emission rate.

- **Type of industry**: what the industry/factory produces which will correlate to what gases are emitted out of the stack/chimney, this will be in the form of labeled classes.
- **Size of the industry**: how big is the industry/factory which will correlate to how much the maximum is outputted, this will be represented in the form of a scale from 1 to 10.
- **Output efficiency**: the amount of output it produces each day divided by total amount of output it can ideally produce.

D. Implementation

The block diagram of air pollution prediction module is shown in Fig. 1 and each stage is described below:

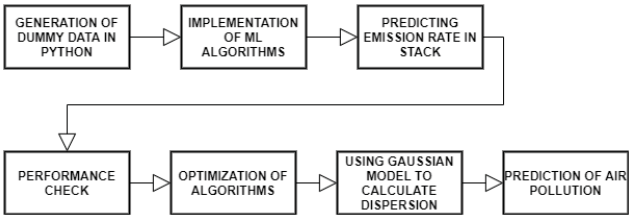

Fig. 1. Block Diagram of Air Pollution Prediction module

1) **Generation of dummy data in python:** 1000 samples of pollution data, is generated using python based on the Central Pollution Control Board (CPCB) of India guidelines [15]. The dummy data was not truly random, it was correlated with various meteorological air parameters such as wind speed, wind direction, and geographical parameters like longitude and latitude. Also taking into account industry expert's opinions so that the machine could be trained well. It was observed that by adding more meteorological factors such as day, time, and season, the prediction performance is greatly improved.
2) **Implementation of Machine Learning (ML) algorithms:** ML algorithms were implemented on the created dummy data to predict the value of Q-emission rate and v-velocity of the wind. For this purpose, the data was divided into a training set (80%) and a test set (20%).
3) **Performance check:** Accuracy measurements were then conducted on the predicted emission rate. The mean square error was measured in each case to check for accuracy.
4) **Optimization of algorithms:** Various algorithms were optimized to reduce the error thereby increasing the accuracy of prediction. On this basis, the best algorithm was selected.

5) **Prediction of Air Pollution Dispersion:** Using Gaussian model [16] to calculate dispersion, the extent of pollution the spread was calculated. We chose the Gaussian dispersion model since it was the most optimal model in terms of implementation and computing power required.

IV. Results

The prediction module and air dispersion module was implemented in the python programming language using Spyder integrated development environment. In an attempt to optimize the models the parameters of the models were tuned and the response (root mean squared error) was plotted. Also, the air dispersion module was implemented to calculate the concentration from a stationary steady-state point source i.e. stack and the results were mapped onto a gray-scale bitmap image where the pixel color represents the concentration.

A. Machine Learning Module

Different machine learning algorithms were studied and implemented on the dataset and results were compared. Table I shows the root mean squared error (RMSE) in the respective models after the parameters were tuned. The error rate is given in meter cube per hour units.

TABLE I
Results of Machine Learning algorithms

Sr no.	Machine Learning Model	RMSE $\frac{m}{hr}^3$
1	K-Nearest Neighbour	3
2	Support Vector Regression	9
3	Random Forest	4
4	Multi-linear Regression	10
5	**Multi-layer Perceptron Regression**	**2.1**

1) K-Nearest Neighbour (KNN): Fig. 2 shows that the KNN model was tuned on its algorithm and weight parameter used to cluster data points.

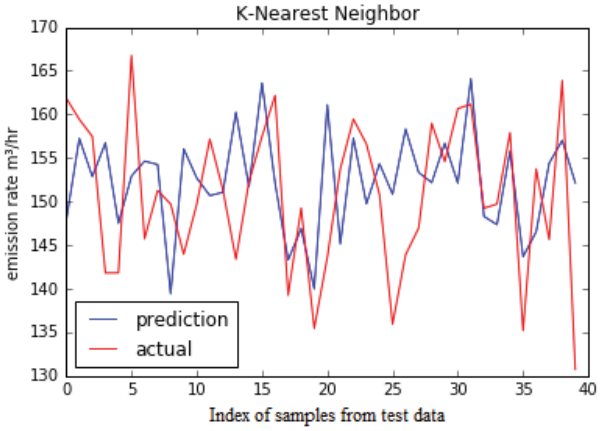

Fig. 2. Results of K-Nearest Neighbour

2) Support Vector Regression (SVR): Fig. 3 shows the prediction using SVR model that was tuned on its kernel which was Radial Basis Function (RBF).

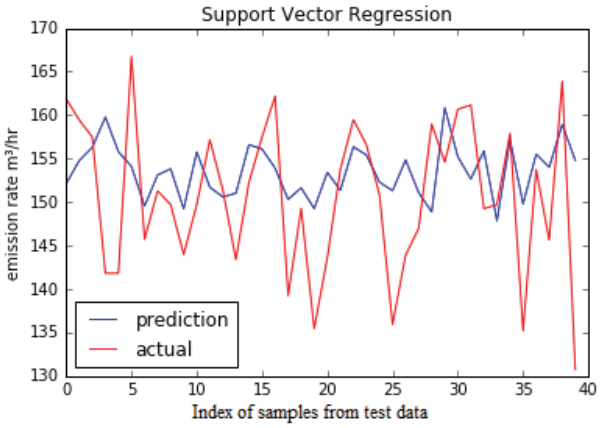

Fig. 3. Results of Support Vector Regression

3) Random forest (RF): Fig. 4 shows predicted values using random forest algorithm. The following parameters were optimised and the corresponding values were used to calculate the final error. (a) Maximum depth = 40, (b) Maximum features= 3 and (e) N estimators = 50 (c) Minimum sample leaf = 20 and (d) Minimum sample split = 20.

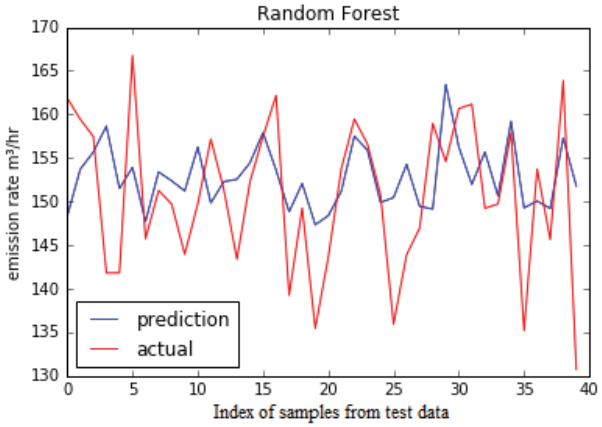

Fig. 4. Results of Random Forest algorithm

4) Multi-linear Regression (MLR): Fig. 5 shows that the plot of predicted values and actual values for Multi-linear Regression algorithm.

Fig. 5. Results of Multi-linear Regression

5) Multi-Layer Perceptron (MLP): Fig. 6 shows predicted values using Multi-layer Perceptron (MLP) after optimization of following parameters to the corresponding values (a) Activation function = 'logistic', (b) Number of neurons = 350 and (c) Type of solvers = 'adam'

B. Air Dispersion module

To calculate the concentration of the pollutant after emission from the stack, Gaussian air dispersion model was implemented.

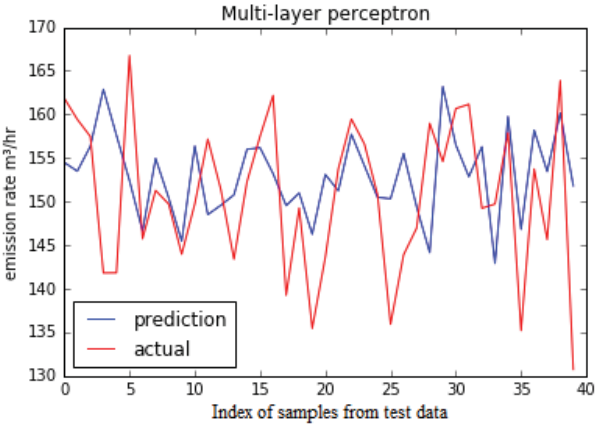

Fig. 6. Results of Multi-Layer Perceptron

The results were then represented using a threshold functions to map different concentration levels with a color level, and this was outputted as an image. Fig. 7 shows the different levels of concentration, white color represents regions with high concentration and black with low concentration.

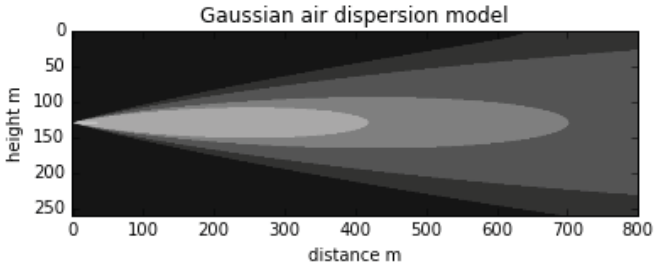

Fig. 7. Gaussian air dispersion model

V. CONCLUSION

The paper has analyzed the application of machine learning algorithms in the prediction of air pollution. The paper also analyses the spread of pollutants and their concentration at distances away from the polluting source. The air dispersion module used the Gaussian air dispersion model which was implemented using python in spyder IDE. For air pollution prediction, 5 different machine learning algorithms were implemented on the data and compared, which include Random Forest, Multi-layer Perceptron, K-Nearest Neighbour, Support Vector Regression, and Multi-linear Regression. The results have shown that the

Multi-layer Perceptron algorithm gives the least mean squared error compared to the other algorithms. Future work can be done in comparing various air dispersion models to predict the spread of air pollution.

REFERENCES

[1] W. W. Nazaroff and C. J. Weschler, "Cleaning products and air fresheners: exposure to primary and secondary air pollutants," *Atmospheric environment*, vol. 38, no. 18, pp. 2841–2865, 2004.

[2] F. Ministry of Environment and G. o. I. Climate Change, "Environment laws (amendment) bill, 2015," 2015.

[3] C. C. United Nations, "Paris agreementl, 2015," 2015.

[4] Z. Yang and J. Wang, "A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction," *Environmental research*, vol. 158, pp. 105–117, 2017.

[5] M. R. Delavar, A. Gholami, G. R. Shiran, Y. Rashidi, G. R. Nakhaeizadeh, K. Fedra, and S. Hatefi Afshar, "A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of tehran," *ISPRS International Journal of Geo-Information*, vol. 8, no. 2, p. 99, 2019.

[6] C. Xiaojun, L. Xianpeng, and X. Peng, "Iot-based air pollution monitoring and forecasting system," in *2015 International Conference on Computer and Computational Sciences (ICCCS)*. IEEE, 2015, pp. 257–260.

[7] K. Singh, "Air pollution modeling," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, no. 3, 2018.

[8] m. F. Erturul and M. E. Taluk, "A novel version of k nearest neighbor," *Appl. Soft Comput.*, vol. 55, no. C, p. 480490, Jun. 2017. [Online]. Available: https://doi.org/10.1016/j.asoc.2017.02.020

[9] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. MIT Press, 1997, pp. 155–161. [Online]. Available: http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf

[10] A. Liaw and M. Wiener, "Classification and regression by randomforest," *Forest*, vol. 23, 11 2001.

[11] T. L. Youd, C. M. Hansen, and S. F. Bartlett, "Revised multilinear regression equations for prediction of lateral spread displacement," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 128, no. 12, pp. 1007–1017, 2002.

[12] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classifiaction," 1992.

[13] H. Kaplan and N. Dinar, "A lagrangian dispersion model for calculating concentration distribution within a built-up domain," *Atmospheric Environment*, vol. 30, no. 24, pp. 4197–4207, 1996.

[14] P. De Haan and M. W. Rotach, "A novel approach to atmospheric dispersion modelling: The puff-particle model," *Quarterly Journal of the Royal Meteorological Society*, vol. 124, no. 552, pp. 2771–2792, 1998.

[15] F. Central Pollution Control Board, Ministry of Environment and G. o. I. Climate Change. (2019) Central pollution control board. [Online]. Available: https://cpcb.nic.in/

[16] A. Abdel-Rahman, "On the atmospheric dispersion and gaussian plume model," 10 2008.