IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Air Quality Prediction Using Feature Selection and Deep Learning model

Adhiraj Sud
The Shri Ram School - Aravali
Hamilton Court Complex, DLF Phase IV, DLF City
Gurgaon - 122002

Abstract - ***The growing number of vehicles and industry pollution leads to poor air quality, which brings serious health issues and lung problems for humans. There are safety measures followed by the Government to regulate the air quality, the detection of air quality is challenging due to the implementation cost and computations. The air quality measure is to be continuously monitored under urban areas to ensure healthy living. Artificial intelligence is the growing area in all industries, thus this work is proposed to exploit the advantages of deep learning algorithms for air quality detection. The proposed work used feature selection technique for selecting the important and relevant features brings high accuracy of air quality detection. The wrapper based technique, Recursive feature elimination (RFE) is used for feature selection and deep learning model Deep Neural Network (DNN) is used for predicting the air quality. This prediction gives an alert for the surrounding to work on further measures in controlling pollution. Experimental results showed that the deep learning model outperforms in air quality prediction with the highest accuracy.***

Key Words: Machine Learning, Deep learning, Recursive feature elimination, Feature selection, Air quality, Deep Neural Network

1. INTRODUCTION

Air pollution is generally caused by high levels of PM2.5, which is called the particulate matter having an aerodynamic diameter less than 2.5 micron has been one of the growing threats around the world. Urbanization and industrial growth are the major cause for air pollution and less quality of air leads to the health issues to humans. Governments are taking necessary measures to control air pollution, abiding by the World Health Organization's standards. Air pollution prediction is crucial in modern cities to monitor the air quality and it is also challenging. Thus it has gained research attention these years, to study and overcome these issues. The emergency action to be taken for any high-level pollution events majorly depends on the prediction precision from the data centers and its promptness.

Artificial intelligence is emerging in all industries such as health care, finance, agriculture, automobiles and more. The growth of Internet of Things (IoT) is also beneficial for regular monitoring of air quality throughout various monitoring centers with less hardware and using distributed cloud environments to handle collected data from various centers. The data-driven air pollution prediction is challenging with the collected data. The proposed work aimed at air pollution prediction with artificial intelligence techniques to alleviate accuracy of the prediction. The challenges in these systems is partial observation of attributes/factors that need to be carefully handled. This system has to predict the possible pollution event to handle emergency situations.

The occurrence of various pollutants such as nitrogen dioxide, (NO2), carbon monoxide (CO), sulphur dioxide (SO2) and particulate matter (PM) are the major causes for air pollution. The PM has a diameter less than 2.5 and less than

10 microns are considered as pollutants. These pollution causes lung problems and diseases to human beings. Thus there are monitoring stations available to closely monitor these levels on a regular basis. There are challenges in collecting these pollution information from these stations and identifying any potential threat to the environment. Collecting these information and monitoring involves high human resources in traditional systems. These data need to be monitored regularly to make any quick decision based on alert. The recent growth in IoT technologies has provided a better opportunity to collect data from small sensors at different stations and process the collected data through cloud servers. The growth of AI technologies provides an opportunity for predicting the air pollution types at minimal cost and in the most accurate manner.

The proposed work is a deep learning approach for predicting air pollution. There are several features involved in air pollution data, thus to avoid the problem of over-fitting, the proposed work used a wrapper based feature selection approach. Feature selection technique, Recursive feature elimination selects only high important features from dataset for learning. The accuracy of the air pollution prediction is increased with this proposed model. The objective of study is to implement air quality prediction with machine learning algorithm and deep learning algorithm, with feature selection and compare their results.

The main contributions of this paper are

• Air quality prediction through machine learning model, logistic regression and deep learning model, Deep Neural Network (DNN) and evaluate its performance

• Implement feature selection algorithm for arriving best features from the attributes.

• Minimizing the over-fitting problem with feature selection and choosing best features according to wrapper technique.

This paper is organized as Chapter 2 gives detailed study of literature survey and related work in air quality prediction. Chapter 3 discusses proposed algorithms. Chapter 4 gives detailed discussion on results arrived for air pollution prediction. Chapter 5 draws conclusions based on this work and further enhancement possibilities.

1.1 RELATED WORK

Air pollution leads to serious health issues and lung disease for humans. Traditionally there are only a few monitoring stations for air pollution monitoring. The growth in industries and urbanization brings more pollution. The existing studies are performed based on traditional statistical models. There are few research done based on artificial intelligence are discussed in this chapter.
The adaptive probabilistic prediction is proposed in [1] for prediction for air quality, the model used in generative adversarial network with three layers namely inferences, generators and discriminators. This model is built with three hidden layers and 256 neurons. Experimental results showed that the mean square error (MSE) is 1.05. However, the model is deep learning the classification accuracy is very less thus the MSE loss is very high.

Air pollution prediction with a Bayesian deep learning model is proposed in [2], the work used domain specific knowledge integration. The specific pollutant types PM2.5 and PM10 are used as regularization features. The multi step forecast model is used to project the air quality. The data pre-processing handled missing data and data interpolation. The implemented model had two layers one with RNN one time prediction strategy and another with recursive prediction strategy. Experimental results proved that Bayesian mode model reduced the error by 3.7%.

Recurrent neural network type, Long short term memory (LSTM) model for air pollution prediction is discussed in [3]. To increase the accuracy of prediction and forecast, kernel density estimation with sub division tuning is performed. Spatially adjusted multivariate imputation is developed to manage the incoming data from different monitoring stations. The dataset also considered meteorological variables such as wind speed, direction, humidity and rainfall. Experimental results showed that MAE error is 8.12.

IoT sensing based air quality monitoring is widely used, the collection of data is made more reliable with sensor networks. The work [4] studied air quality monitoring through IoT networks. The regression models, support vector regressor, random forest regressor and Gradient Boosting regressor (GBR) are used. RMSE error 7.8 for GBR, which is more effective in air quality prediction.

Deep learning architectures are highly used for air quality prediction, the work [5] implemented ConvLSTM for air quality forecasting. The model is four stacked convLSTM with five hidden states to extract features from historical data. Dropout layer is added in each ConvLSTM block that enhances the speed of learning. The mode is compared with SVR, from experimental results it is shown that deep learning model perform well than machine learning regression.

This literature survey studied air quality prediction using various artificial intelligence models. Some of the work studied machine learning models, some of them used deep learning models. The deep learning model most widely studies is Long Short Term Memory (LSTM) model as it considered the time series dataset. Some of the studied used regression analysis for air quality prediction. The literature review showed that these works are effective in predicting air quality, however mean square error is the evaluation metric used in most of the work used, which is very high. The higher the MSE error, the lower the model accuracy. Thus it is necessary to propose an effective model, which can accurately predict air quality.

2. RESULTS

The proposed work air quality prediction is implemented in Python 3. Logistic regression and deep neural networks are implemented in the air pollution data. The data contains all state pollution information registered from 1990 to 2015 provided in the Indian Government portal. The dataset is visualized with matplotlib, the visualization shows the highest, moderate and lowest pollution states with pollutant parameters specified for each state. RFE feature selection is implemented to reduce the number of attributes to three with highest scoring on accuracy basis. The logistic regression model, which is a classification model, is build for RFE as base classification model. The results shows that the deep learning algorithm has achieved the highest accuracy of air pollution classification.

3. DISCUSSIONS

The below plot shows an evaluation metric computed for a logistic regression model for air pollution prediction. The results show that the accuracy of the model is 62.3%, the error metrics computed for classification are Mean square error is 0.65, mean absolute error and Root mean square error is 0.80.

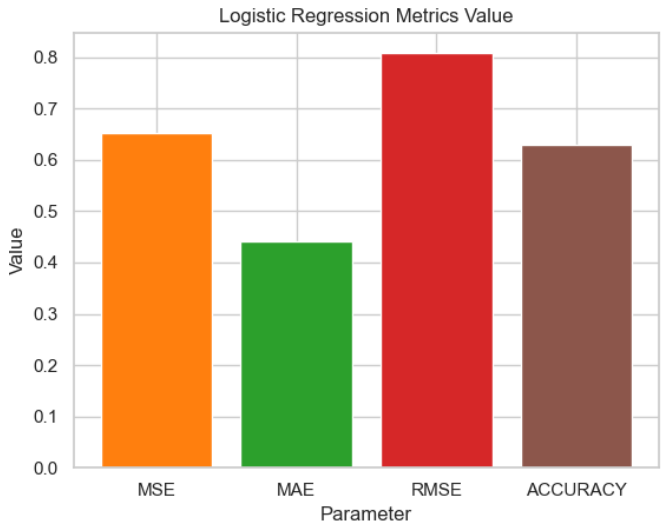

Figure 4: Evaluation metrics of Logistic regression for air pollution prediction

The following plot shows the training accuracy for Deep Neural network algorithm, the plot shows the accuracy is 62.8% with 30 epochs.

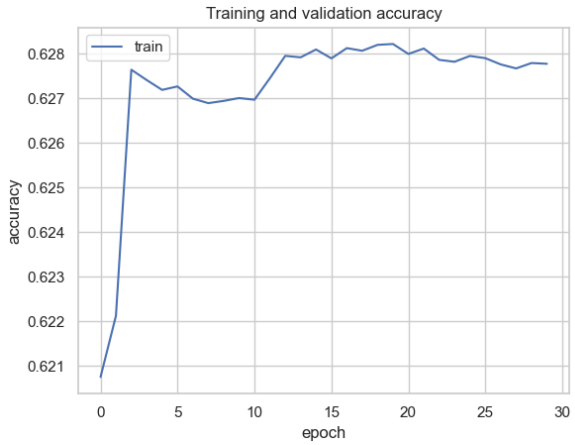

Figure 5: Training accuracy for air pollution prediction using DNN

The following plot shows the training loss for Deep Neural network algorithm, the plot shows the loss is 0.6 with 30 epochs.

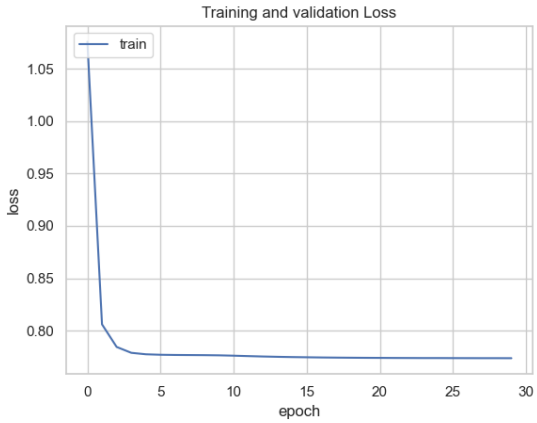

Figure 6: Training loss for air pollution prediction using DNN

This paper proposed air pollution prediction based on machine learning and deep learning models. The proposed work used a feature selection algorithm, RFE for selecting the important feature and reducing the over-fitting problem. The base classification model used for feature selection is Logistic regression, it is a classification algorithm, classifies and gives scores based on accuracy. The top ranked features, three important features are selected by the model. These features are trained with machine learning model Logistic regression and tested the performance and similarly a deep learning model Deep Neural Network is trained and evaluated its performance. Experimental results showed that deep learning has performed better in classification accuracy.

This work can be further enhanced by implementing other deep learning models such as Convolutional Neural Network and Long short term memory model. These deep learning models can be optimized for the parameters such as batch size, epochs and number of layers.

4. METHODOLOGY

Air quality is an important measure to be monitored in every location of the city and other regions such as industrial regions. Air pollution type prediction is considered as the problem in this proposed work. The attributes used for this problem are SO2, NO2, RSPM, SPM are the major pollutants used for air pollution type prediction. Pollution risk and exposure levels can be identified with air quality prediction, however, it is highly challenging due to dispersion of particles and meteorological factors. Existing approach is manually handled, it takes high cost and resources for monitoring at different stations.

4.1 DATASET DETAILS

The air pollution dataset is the Historical Daily Ambient Air Quality Data from the Ministry of Environment and Forests and Central Pollution Control Board of India.

The dataset contains the following features

Features	Description
stn_code	Station code, given to each station for recorded data.
sampling_date	The date when the data was recorded
State	States, which air quality data is measured
Location	City in which data is collected
Agency	Name of the agency that measured the data.
Type	The type of area where the measurement was made.
so2	Sulphur Dioxide measured.
no2	Nitrogen Dioxide measured
Rspm	Respirable Suspended Particulate Matter
Spm	Suspended Particulate Matter
location_monitoring_station	Location of the monitoring area.
pm2_5	Particulate matter less than 2.5 micron
Date	Date of recording

Table 1: Air Pollution Dataset Details

4.2 Exploratory data analysis

Exploratory data analysis (EDA) helps in visualizing data with plots and helps data scientists to understand better on data pattern and its behaviour. This proposed work, data is visualized through some plots. The following plot shows the state-wise air pollution level. This plot helps not only to understand which state has highest pollution and which is lowest pollution, this plot also helps in analysing each pollutant factor such as SO2, NO2 presence in each state.

In the below plot, the air quality among all states are plotted, the colours in each bar represents the pollutants name, for example, blue represents the presence of SO2. From this visualization it is visible that air pollution is high in Andhra Pradesh and the next highest state is Maharashtra.

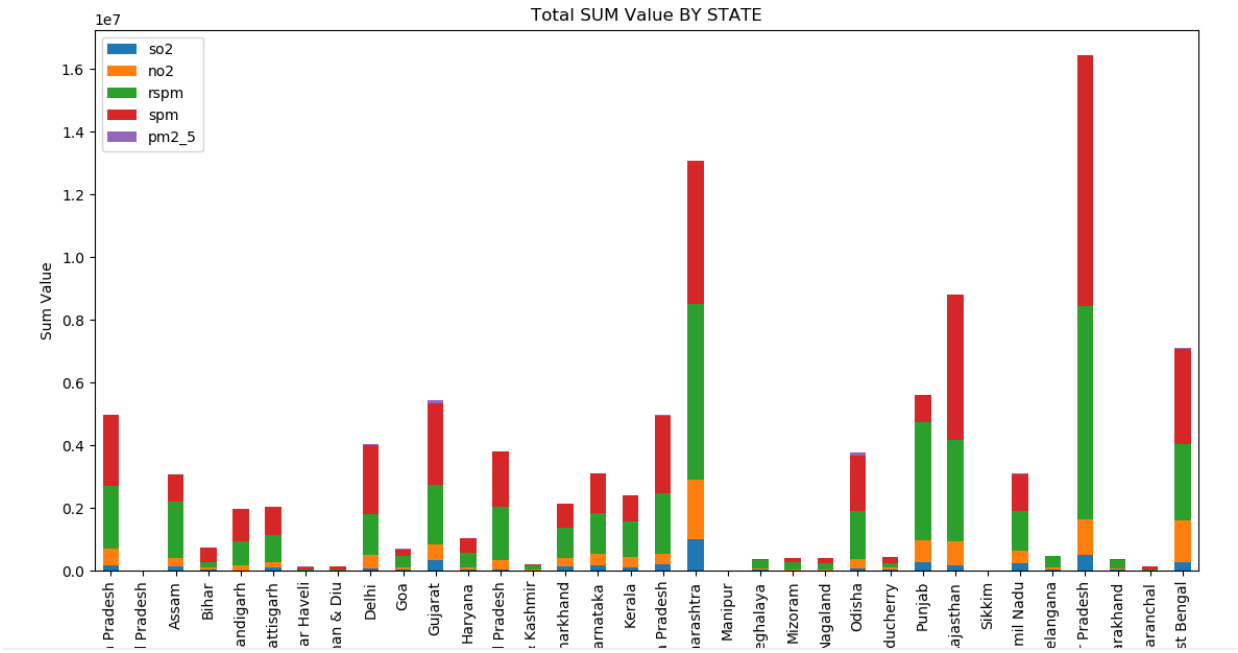

Figure 1: State wise Air Pollution data in India

This work implements a machine learning model, logistic regression and deep learning model, Deep Neural Network (DNN). The dataset with 12 features are taken for this study. This work is classified as a multi class classification problem with four pollution types Residential, Industrial, RIRUO (Residential / industrial / rural / other areas), sensitive. The feature selection technique is used to select important features from the dataset. There are five independent variables, the top three important features are selected for this implementation through a scoring function performed by the RFE algorithm.

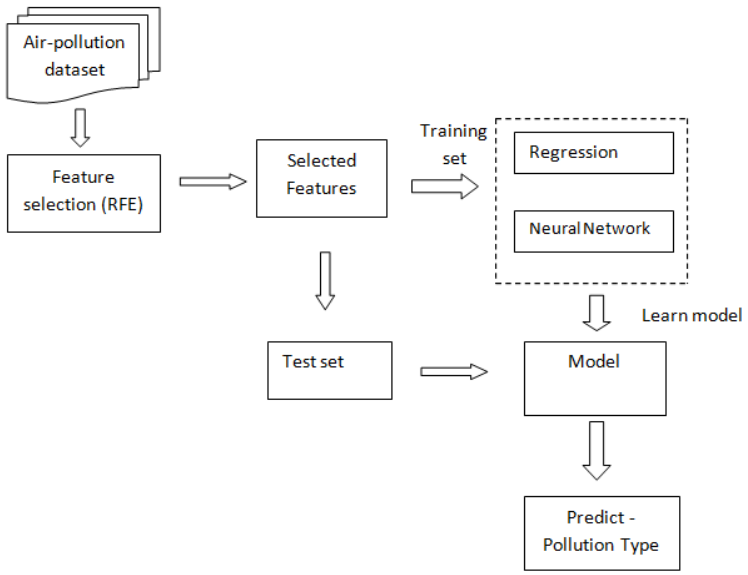

Figure 2: System architecture of Airpollution prediction

The above figure represents the proposed system architecture, the dataset is loaded and first applied RFE technique to select the top three important features from the dataset. There are two models implemented and an air pollution type of four classes is predicted.

4.3 FEATURE SELECTION

Recursive feature elimination (RFE), a Wrapper based feature selection technique is implemented to select top three important features based on the scoring function. The machine learning algorithm to be given as base classification model for RFE, logistic regression is given for RFE for classification and get scoring. RFE scoring function computes

score for every attribute and the least important attributes are removed from the list, thus giving the highly important features. The algorithm chosen top three features are SO2, SPM and PM2.5. The following tables shows the attributes and its ranking given by RFE and top ranked features are selected for training.

Feature	Rank
SO2	1
SPM	1
RSPM	2
NO2	3
PM2.5	1

Table 2: RFE Scoring given to air pollution dataset

For classification problems, the scoring metric used by RFE is accuracy. Accuracy is computed based on the proportion of correctly classified instances. The formula to calculate accuracy is given in equation 1.

$$\text{Accuracy} = \text{Number of Correctly Classified Instances} / \text{Total Number of Instances} \quad \text{--(1)}$$

The selected features are given input to the machine learning model Logistic regression, this is a classification algorithm, which takes independent features from the dataset and classifies the air pollution into four types. This classification is a multi -class problem with 4 types are Residential, Industrial, RIRUO and Sensitive.

4.4 DEEP NEURAL NETWORK

Deep neural network with three layers as shown in the below figure is added. The first layer is a dense layer with input dimension 2, units 4 and a 'uniform' kernel. The second layer added is dense layer with 'softmax' activation, the third layer is dense layer and the model is compiled with adam optimizer and loss metric computed using categorical cross entropy and accuracy metric. The number of epochs trained is 10 with batch size 1000.

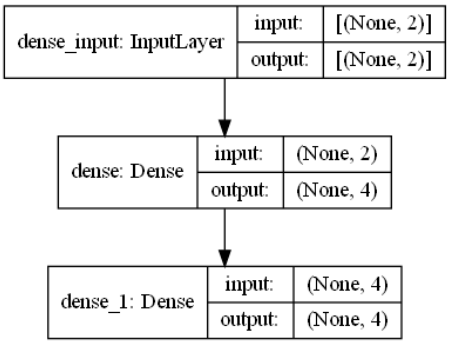

Figure 3: Deep Neural Network Layers for Air pollution prediction

5. ACKNOWLEDGEMENTS

I would like to thank all my teachers at The Shri Ram School for encouraging me to write the paper particularly during the difficult times.

6. REFERENCES

[1] Z. Wu, N. Liu, G. Li, X. Liu, Y. Wang and L. Zhang, "Learning Adaptive Probabilistic Models for Uncertainty-Aware Air Pollution Prediction," in IEEE Access, vol. 11, pp. 24971-24985, 2023, doi: 10.1109/ACCESS.2023.3247956.

[2] Y. Han, J. C. K. Lam, V. O. K. Li and Q. Zhang, "A Domain-Specific Bayesian Deep-Learning Approach for Air Pollution Forecast," in IEEE Transactions on Big Data, vol. 8, no. 4, pp. 1034-1046, 1 Aug. 2022, doi: 10.1109/TBDATA.2020.3005368.

[3] H. A. D. Nguyen et al., "Long Short-Term Memory Bayesian Neural Network for Air Pollution Forecast," in IEEE Access, vol. 11, pp. 35710-35725, 2023, doi: 10.1109/ACCESS.2023.3265725.

[4] D. Zhang and S. S. Woo, "Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network," in IEEE Access, vol. 8, pp. 89584-89594, 2020, doi: 10.1109/ACCESS.2020.2993547.

[5] I. Mokhtari, W. Bechkit, H. Rivano and M. R. Yaici, "Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction," in IEEE Access, vol. 9, pp. 147 65-14778, 2021, doi: 10.1109/ACCESS.2021.3052429.

[6] Y. Yang, G. Mei and S. Izzo, "Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning," in IEEE Access, vol. 10, pp. 50755-50773, 2022, doi: 10.1109/ACCESS.2022.3173734.

[7] Y. Huang, Y. Xiang, R. Zhao and Z. Cheng, "Air Quality Prediction Using Improved PSO-BP Neural Network," in IEEE Access, vol. 8, pp. 99346-99353, 2020, doi: 10.1109/ACCESS.2020.2998145.

[8] Y. Yu, J. J. Q. Yu, V. O. K. Li and J. C. K. Lam, "A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data," in IEEE Access, vol. 8, pp. 74291-74305, 2020, doi: 10.1109/ACCESS.2020.2988684.

[9] Y. Zhou, S. De, G. Ewa, C. Perera and K. Moessner, "Data-Driven Air Quality Characterization for Urban Environments: A Case Study," in IEEE Access, vol. 6, pp. 77996-78006, 2018, doi: 10.1109/ACCESS.2018.2884647.