Air Pollutant Concentration Prediction using Ensemble of Machine Learning Techniques

Amrutha C
Department of Computer Science
BMS College of Engineering, Bengaluru.

Dr. B G Prasad
Department of Computer Science
BMS College of Engineering, Bengaluru

***Abstract*—** An Ensemble of conceptually different machine learning algorithms is proposed for predicting 1 day ahead PM2.5 pollutant concentration. Three individual prediction algorithms ARIMA, Multiple Linear Regression and Multilayer Perceptron are built and the predictions from each of these models are then combined using Ensemble of weighted average technique to obtain final PM2.5 prediction. Historical time series observations of various meteorological parameters such as Temperature, wind, humidity etc. and gaseous pollutants NO2, SO2, CO and NO are used as predictors. The combination of conceptually different machine learning algorithms can effectively capture various nature of air pollution and increases the prediction accuracy as it averages out the error from each model leading to better prediction than individual models.

Keywords*—Machine Learning, Multilayer Perceptron (MLP), ARIMA, Multiple Linear Regression (MLR), Ensemble Technique.*

I. INTRODUCTION

Air Pollution is a result of harmful contaminants and emissions that are present in the ambient air. The atmospheric pollution has a bad impact on living beings causing major health issues. It also brings on environmental risks such as acid rain, fog, ozone depletion and also effects vegetation. Air pollution is increasing at a high rate since few years due to development, urbanization, and increase in vehicular traffic. It has become very essential to monitor the causes of pollution and control the air pollution level to avoid its destructive effects.

Metropolitan air impurity ascend primarily due to the combustion processes and industrialization. The key pollutants affecting metropolitan regions are Particulate matter (PM), Ozone (O3), Nitrogen dioxide (NO2), Carbon monoxide (CO), Sulfur dioxide (SO2). Increment in level of air pollutants above allowable level is influencing wellbeing of individuals and ecological condition.

In adding to the effect of urbanization and speedy population growth, the level of pollution in the cities is to a great extent tweaked by meteorological elements. For instance, wind and precipitation may go about as components that separately ventilate and clean the environment from vaporizers or function as impurity transport. Hence the meteorological parameters such as atmospheric wind speed, wind direction, humidity, temperature etc. also influence pollution concentrations.

Several urban air quality observing projects covering different real Indian urban areas have been underway generating vast databases with the continuous observations on the various atmospheric pollutants and weather. Many works have been going on in analyzing these data collected, to understand the behavior of these pollutants over a period of time. Consequently, to develop approaches for urban air quality supervision, it is basically needed to implement appropriate methods that can be used in predicting the air quality and computing the seasonal behavior on the atmospheric air in a region. Many automatic data analysis and knowledge discovery tools have been developed as data sets are growing in size and complexity, which are generally described as machine learning (ML) methods. Machine learning (ML) approaches have been used for pollution forecast in recent. It involves computational methods that improve the performance of computerizing the acquisition of knowledge from experience. Forecasting is one of the tasks that greatly involves learning in which the forecasting model is constructed by training models from datasets which is usually nonlinear in the case of air quality data. Data Mining and Machine Learning Techniques can be applied to historical time-series air quality and weather-related data in order to analyze the behavior of pollution and use these on building efficient prediction models. Predictions can use to forecast future pollutants concentration which can provide insights and help in taking appropriate decisions to reduce pollution.

The proposed system involves building multiple individual prediction models including ARIMA, MLP, and MLR for PM2.5 pollutant concentration prediction and then combining the predictions from each model into ensemble model using the weighted average technique for final prediction of the pollutant concentration. The Ensemble prediction model aims at increasing the accuracy of prediction compared to individual models prediction as the predictions from conceptually different models are combined, which effectively reduces the prediction errors.

II. LITERATURE REVIEW

Existing literature survey focuses on using various machine learning models in classifying the pollutants level and in time ahead prediction of pollutants concentration. Particulate Matter

(PM10, PM2.5) is one of the dangerous pollutant and is used as a key parameter in measuring the quality of air in an area. Classifiers such as Multilayer Perceptron (MLP), Naive Bayes and Support Vector Machine (SVM) are used in [1] for classifying the PM10 values into two categories "High" with the PM10 values greater than100 µg/m3 and "low" with the values lesser or equal to 100 µg/m3. Various meteorological parameters and weather parameters are also used as input parameters. The performance of Multilayer Perceptron had better accuracy of 98.1% compared to other models, where Naive Bayes and SVM had an accuracy of 91.25% and 92.5% respectively.

Various regression models are used in predicting time ahead concentrations of pollutants. Multiple Linear Regression (MLR) and Regression with Time Series Error (RTSE) models are used for 1-day-ahead prediction of daily average PM10 concentrations in [2]. The air quality dataset from 5 stations are used for the study by implementing two variations of MLR models, one with lagged parameters as predictor variables (MLR-1), and other with lagged PM10 concentrations and predictor variables (MLR-2). The performance evaluation of all these models for each station's data showed that MLR1 gave least accurate forecast while MLR2 and RTSE gave better accuracy as multiple parameters were considered in the prediction process.

ARIMA (Auto Regressive Integrated Moving Average) based Air quality prediction model is proposed in [3]. The time series modeling like ARIMA mainly requires the data to be stationary. Augmented Dickey-fuller test (ADF) is used in testing dataset for stationarity. Order of ARIMA is examined based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PCAF). Pollutants are analyzed individually for Industrial, Residential and sensitive monitoring stations. The accuracy of RSPM and NO2 predictions for residential stations is more as the data are stationary. ARIMA model is suitable for short-term predictions.

When a large number of parameters are to be used in prediction modeling, the dimensionality of the dataset can be reduced using Principal Component Analysis (PCA). Principal Components are a set of linearly uncorrelated variables that are obtained by applying an orthogonal transformation on a set of correlated variables. A Neural Network (NN) along with PCA has been implemented in forecasting the 1 day ahead daily Air Quality Index (AQI) on a seasonal basis in [4] where the Principal components are calculated using the AQI and meteorological variables of previous day. Similar work of using PCA with Multiple Linear Regression (MLR) and Feed Forward Back Propagation (FFBP) for PM10 prediction up to three days ahead has been carried out in [5]. The application of PCA to the models have shown in increase in the accuracy of prediction noticeably compared to individual models without PCA.

Various optimization techniques can be used on input parameter to reduce the irrelevant inputs to the model. Optimization techniques such as Genetic Optimization, Forward Selection and Backward Elimination has been used on inputs in [6] before feeding it to neural networks for better accuracy. Artificial Neural Networks has been widely used in modeling as it can handle the non-linearity nature of the air quality data. Three Multi Linear Regression (MLR) models are developed using the same data set as used for ANN model and performance has been analyzed. The inputs to the MLR-1 model included only timescales representing hour, day and month of the year, whereas the inputs to MLR-2 model included wind speed and direction along with timescales, and the MLR-3 model used all meteorological parameters and timescales. The ANN model resulted in the least RMSE value compared to other models which say ANN model is better than MLR.

Univariate and Multivariate modeling approaches have been carried out in [7], where former considers only the target gas concentration value and latter considers different features as predictors. M5P model trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) are implemented using the air quality collected from one Multi-Gas Sensing devices (MGS) station which are then processed through the Machine Learning steps to forecast pollutants concentration values of O3, NO2, and SO2 for 1, 8, 12, and 24 hours ahead. Time Windowing is used where a number of time-lagged features for each input attribute is generated that helps in understanding the time dependency between consecutive data points and allows training the forecasting models to predict counteractions for 1, 8, 12 and 24 hours ahead. M5P has outperformed other algorithms for all gases because of the tree structure efficiency and powerful generalization ability. The multivariate modeling approach used has resulted in enhanced prediction accuracy and reduced error as the dependency between target gases and other features are considered in modeling.

Various factors influence the pollution concentration which include Meteorological factors, vehicular traffic-related parameters, background concentration (lagged parameters), and geographical factors. These factors are considered in different combinations in modelling Back-Propagation Neural Network (BPNN) and the results are compared in [8]. Inputs to the BPNN are chosen using 2 methods namely Known-background-concentrations Prediction (KBCP) where actual measured values of three previous hour concentrations of the target pollutant are used as inputs and Unknown-background-concentrations prediction (UKBCP) where forecasted concentration values are used as background factors.

A model to estimate real-time roadside CO and NO2 concentrations based on Neural Network (NN) using meteorological condition and vehicular traffic data has been implemented in [9]. The traffic data has been collected using a traffic monitoring system called SCOOT (Split Cycle and Offset Optimization Technique) which included vehicle flow, Stops, Delay and Congestion time. Six NNs - MLP, RBF (Radial Basis Function), and MNN (Modular Neural Network) for each pollutants CO and NO2 are implemented. The RBF showed more accuracy in capturing and predicting extreme values of pollutants than the MLP.

Instead of training individual models, multiple base learners can be trained and the strength of each model can be combined to obtain better prediction models. Such type of models is referred to as an Ensemble model. An architecture for ensembles of Adaptive Neuro-Fuzzy Inference System

(ANFIS) for Air Quality Index forecasting has been proposed in [10]. The construction of EN-ANFIS model included five layers made up of Input layers, Sample layer where training dataset is created using different sampling methods, training layer which contains multiple ANFIS that are trained using training dataset, testing layer contains trained ANFIS models where the testing data is inputted to each model at the same time, and the output layer where the final prediction is calculated by uniform weighting of output from each Sub-ANIFS unit. Different sampling techniques such as Random sampling without replacement, Bootstrap sampling with replacement have been used to create training datasets to the subsystems. The comparative results showed that EN-ANFIS with Bootstrap and Random sampling techniques had lesser RMSE and training time than individual ANFIS with and without sampling techniques.

From the previous works done, it can be concluded that Neural Networks works well with air quality data because of its nonlinear nature. Including various meteorological factors and traffic-related data would help in modeling more accurate predictions. Considering the lagged target pollutant concentrations would also help in understanding the relation between consecutive data.

The current air pollution prediction systems are based on individual models. Commonly used algorithms for building prediction models include Multiple Linear Regression, Neural Networks, SVM and time series model like ARIMA. There exist few Ensemble models that are built by combining multiple same base algorithms with techniques like boosting, bagging or basic averaging. Each algorithm has its own pros and cons, where some may handle non-linearity in the data well and some may require stationary data to perform better. Combining similar base models might not significantly improve the performance. Hence an ensemble of hybrid prediction models is proposed in this paper.

III. DATA COLLECTION

The Dataset is collected from Central Pollution Control Board (CPCB) for the period of 3 years from March 2015 to May 2018 for 2 stations Peenya and BTM layout representing industrial and residential area respectively. The dataset includes daily average concentrations of various gaseous and meteorological parameters as shown in Table 1.

The proposed system involves building multiple individual prediction models including ARIMA, MLP, and MLR for PM2.5 pollutant concentration prediction and then combining the predictions from each model into ensemble model using the weighted average technique for final prediction of the pollutant concentration. The Ensemble prediction model aims at increasing the accuracy of prediction compared to individual models prediction as the predictions from conceptually different models are combined, which effectively reduces the prediction errors.

Table 1. Air quality and meteorological parameters.

Parameters	**Unit**
Particulate Matter (PM2.5)	µg/m3
Relative Humidity (RH)	%
Temperature (Temp)	°C
Carbon-Monoxide (CO)	mg/m3
Nitrogen-dioxide (NO2)	µg/m3
Nitrogen-oxide (NO)	µg/m3
Sulphur-dioxide (SO2)	µg/m3
Wind Speed (WS)	m/s
Wind Direction (WD)	degree
Solar Radiation (SR)	W/mt2
Ozone (O3)	µg/m3

IV. PROPOSED SYSTEM

A. Auto-Regressive Integrated Moving Average (ARIMA)

ARIMA is a popular class of forecasting models that can be fitted to time series data. It is most suitable for long, stable series of data as it directly depends on the past values for forecasting. ARIMA models can be used to model both seasonal and non-seasonal time series data. ARIMA is defined by three parameters p, d, q representing AR, I, MA components respectively.

• AR - Auto-Regressive component (p), uses the past values as predictors in order to forecast future data point. It incorporates the effect of history into the future predictions by using the autocorrelation or lags as predictors.

• I – Integrated component (d), represents the order of differencing that is required to make the time series data non-stationary. The differencing process involves subtracting current value with the d previous value. This makes the data to have constant mean and variance and auto covariance over time and helps in building the stable model.

• MA – Moving-Average component (q), represents the regression error of the model which is a linear combination of previously occurred error terms. The order q represents the number of error terms to include in the modeling of time series.

By analyzing and estimating the appropriate order of p, d, q and using them to fit ARIMA to time series air quality data, would help in forecasting pollutant concentration.

B. Multiple Linear Regression (MLR)

Regression refers to finding the correlation between the values of one variable with the corresponding values of other variables. It is used to determine a mathematical relationship among a number of random variables. In other terms, MLR

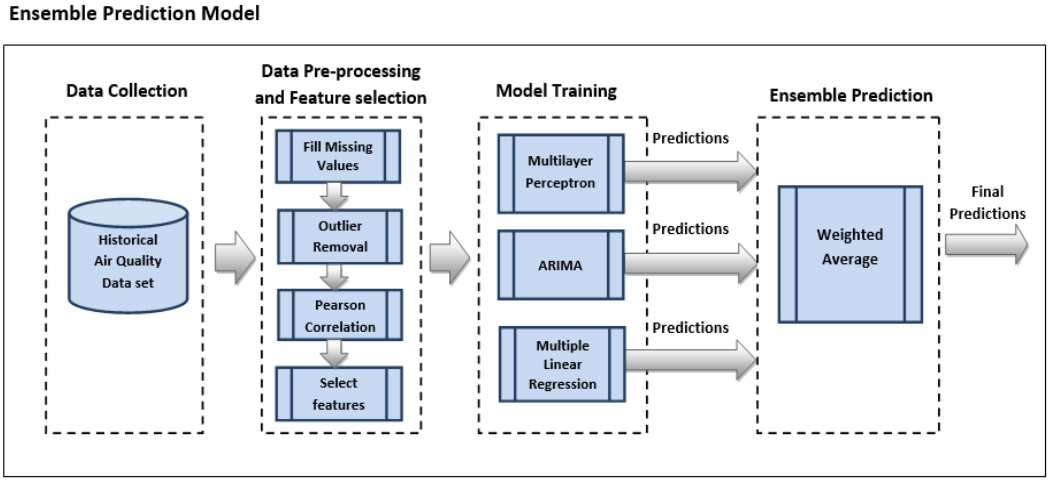

Fig. 1. The Architecture of Ensemble prediction model

examines how multiple independent variables are related to one continuous dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables learned can be used on the similar set of new variables to predict their respective output. The model creates a function in the form of a straight line (linear) that best approximates all the individual data points. The model assumes that there exists a linear relationship between the set of input parameters and real-valued output parameter.

The linear relationship among multiple input variables X1, X2, …, Xn and the output variable Y, can be mathematically expressed as

$$Y = Coeff\_1 * X1 + Coeff\_2 * X2 + \ldots + Coeff\_n * Xn$$

Where Coeff_1, Coeff_2, . . , Coeff_n are the coefficients learned using the dataset.

C. Multilayer Perceptron (MLP)

MLP is a feed-forward artificial neural network model that takes in a set of input data and maps it onto a set of outputs. They consist of a system of simple interconnected neurons or nodes with the non-linear mapping between an input vector and an output vector. A perceptron is a unit which calculates single output from multiple real-valued inputs by a linear combination of inputs based on their weights and then applying a non-linear function called activation function on output. MLP make use of back-propagation for training a network, which is a supervised learning technique. MLP's have the ability to describe highly non-linear and complex relations among the data and they are universal approximators. Therefore they are commonly used in modeling and studying atmospheric data and forecasting. MLP can be trained on air quality and meteorological features in order to predict target pollutant concentration with good accuracy.

D. Ensemble Model

As each individual prediction algorithms may have their own limitations, a more advanced Machine learning technique, called ensemble learning method can be used to combine the prediction from individual models. Rather than a prediction algorithm, this method works as a framework which aims to reduce prediction errors by combining different prediction algorithms together. Ensemble learning is primarily used for improving classification, prediction and performance of a model.

The overall system architecture of ensemble prediction model for air pollutant concentration is as shown in Fig. 1. A brief description of the system architecture is as follows:

The historical gaseous pollutants concentrations and meteorological factors collected from pollution monitoring station form the input data set. The input dataset includes day t-1 values of each parameter as predictors and PM2.5 concentration of day t as target value.

Data-preprocessing techniques are then applied to this dataset in order to clean and fill in the missing values in it which involves removing the outlier values for each feature and linear imputation technique for filling in the missing values. Linear interpolation involves fitting a linear curve between the preceding and succeeding data points of the missing value and imputing the missing value.

Correlation between various parameters in the dataset and the target pollutant PM2.5 is analyzed using Pearson Correlation coefficient which tells how much correlation does each predictor features exhibits with the target pollutant., based on which the key parameters are selected to feed as input to each of the prediction models. The features that exhibit positive correlation with the target pollutant is selected for training MLP and MLR whereas for the ARIMA models the target pollutant alone is considered as an input feature. The data frame with selected features is then split into training and testing dataset in the ratio 80% and 20% respectively. The time-series k-fold cross validation is used on the training set with K=10 folds. The training data is divided into 10 folds, where in each iteration a different fold of data is held out as validation set and the model is trained using remaining k-1 folds and model is validated against the validation set. This reduces the model variance as every portion of data is used for both training and testing and also avoids the problem of the model over fitting.

The three prediction models namely ARIMA, Multilayer Perceptron (MLP), and Multiple Linear Regression (MLR) are trained, validated and tested individually to obtain target pollutant concentration prediction. To train and fit ARIMA model, the p, d, q values are estimated based on Auto-Correlated Function (ACF) and Partial Auto-Correlated Function (PACF). The MLP model is built using the following parameters:

The solver used for weight optimization is 'lbfgs' as it can converge faster and perform better for less dimensional data. It gives better results compared to stochastic gradient descent optimizer. The activation function 'relu' is used which stands for Rectified Linear units (ReLu) function. It avoids the problem of vanishing gradient.

The predictions from each model are then combined into a final prediction using weighted average ensemble technique. The Weighted Average Ensemble is a method where the prediction of each model is multiplied by the weight and then their average is calculated. The weights for each base model is adjusted based on the performance ability of each model. The predictions from each model are combined using the weighted average technique, where each model is given different weights based on its performance. The model with better performance is given more weight. The weights are assigned such that the sum of weights must be equal to 1. The weighted average of prediction is calculated as follows:

Final prediction = (w1*P1 + w2*P2 + w3*P3) / total weight

where w1, w2, w3 are the weights assigned for each model and P1, P2, P3 are the predictions from the model ARIMA, MLR, MLP respectively.

V. RESULTS AND DISSCUSIONS

The Fig. 2-4, 6-8 shows the actual PM2.5 concentrations versus the predicted PM2.5 concentrations of the day from the individual models ARIMA, Multiple Linear Regression (MLP) and Multi-Layer Perceptron (MLP) for BTM and Peenya stations respectively. It can be observed that the predictions from MLR and MLP models have crossed peak values at some points, whereas ARIMA model has tried to reach peak values to some extent. Fig. 5, 9 shows the predictions obtained by combining the predictions from respective individual model using ensemble weighted average technique and is compared with the actual observations on that day.

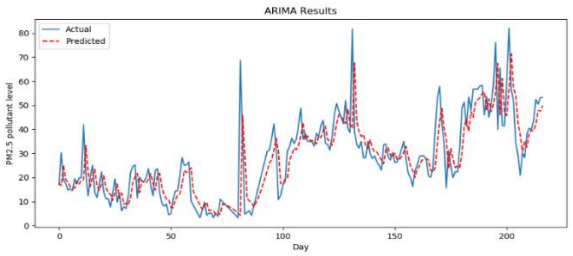

Fig. 2. Actual observations versus predicted concentration of PM2.5 by ARIMA model for the number days in the test data for BTM dataset

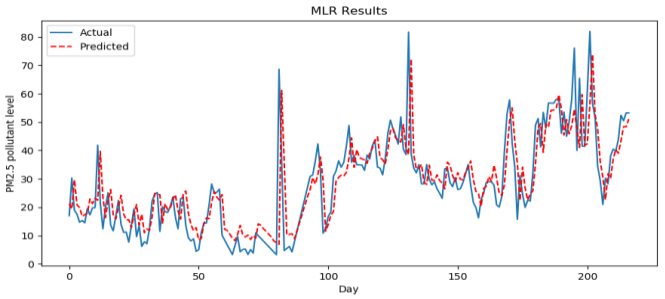

Fig. 3. Actual observations versus predicted concentration of PM2.5 by MLR model for the number days in the test data for BTM dataset

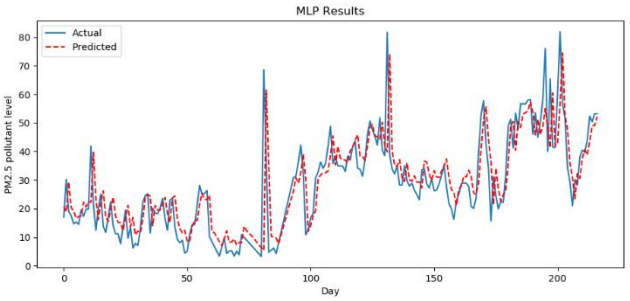

Fig. 4. Actual observations versus predicted concentration of PM2.5 by MLP model for the number days in the test data for BTM dataset

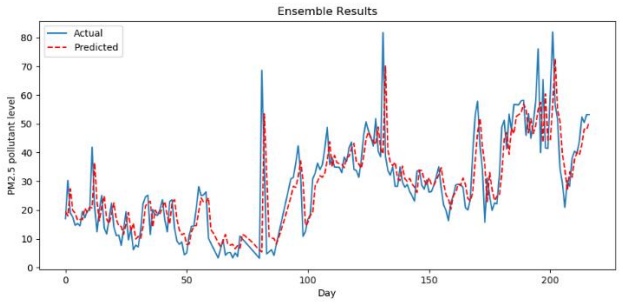

Fig 5 Actual observations versus predicted concentration of PM2.5 by Ensemble model for the number days in the test data for BTM dataset

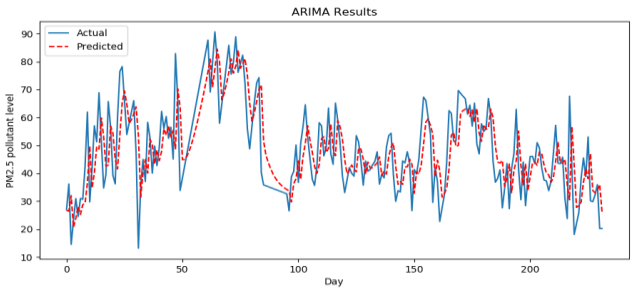

Fig. 6. Actual observations versus predicted concentration of PM2.5 by ARIMA model for the number days in the test data for Peenya dataset

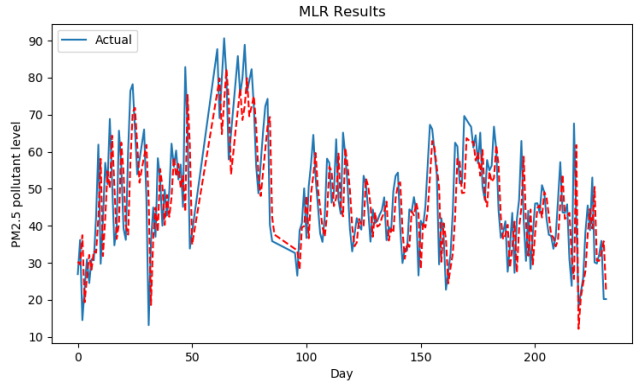

Fig. 7. Actual observations versus predicted concentration of PM2.5 by MLR model for the number days in the test data for Peenya dataset

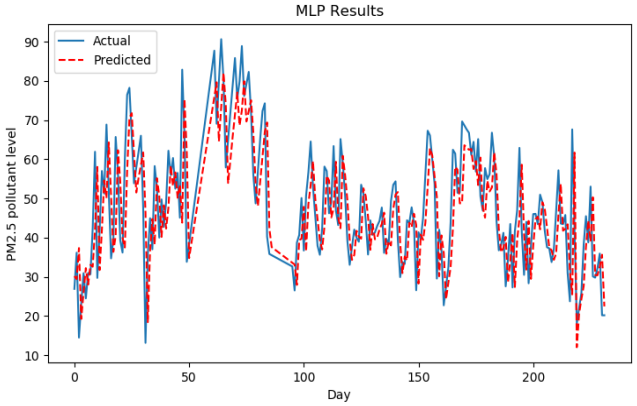

Fig. 8. Actual observations versus predicted concentration of PM2.5 by MLP model for the number days in the test data for Peenya dataset

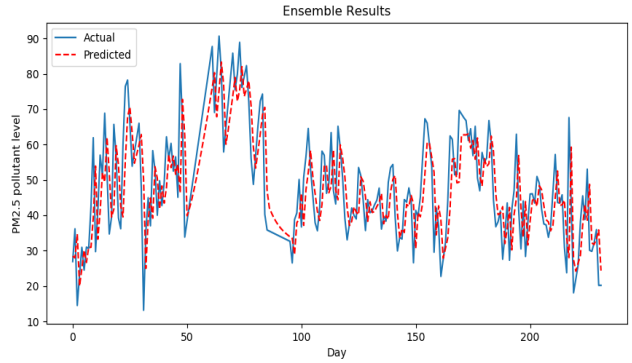

Fig. 9. Actual observations versus predicted concentration of PM2.5 by Ensemble model for the number days in the test data for Peenya dataset

Table 2 shows the comparison of ARIMA, MLR, MLP and Ensemble prediction performance for BTM and Peenya stations with respect to performance metric RMSE. It can be observed that ARIMA has relatively better performance than MLR and MLP, while Ensemble has better predictions compared to all other individual models for both stations.

		ARIMA	MLR	MLP	ENSEMBLE
BTM Dataset	**RMSE**	9.521	9.525	9.554	9.361
Peenya Dataset	**RMSE**	10.870	10.950	10.960	10.680

Table 2. RMSE values obtained for each model.

VI. CONCLUSION

An ensemble prediction system for forecasting 1 day ahead particulate matter concentration using various meteorological factors such as temperature, solar radiation, wind direction and other gaseous pollutants including PM2.5 concentration on previous day has been implemented. Three conceptually different models including ARIMA which is a time series model, Multiple Linear Regression (MLR) which models linear relationship among predictors and Multilayer perceptron which can handle non-linearity in the data are implemented. The predictions from each of these models are then combined using ensemble technique, which performed relatively better than all other models. Hence ensemble models can increase the accuracy of prediction when compared to individual models.

The air pollution is also majorly influenced by vehicular and industrial emissions. In future, traffic-related data along with meteorological parameters could be considered in predictions. The pollution prediction system can be implemented to predict in real-time and for long-term forecasting.

ACKNOWLEDGMENT

The work reported in this paper is supported by the college [BMSCE, Bengaluru] through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-III] of the MHRD, Government of India.

REFERENCES

[1] Akhtar, Aly, et al. "Prediction and Analysis of Pollution Levels in Delhi Using Multilayer Perceptron." Data Engineering and Intelligent Computing. Springer, Singapore, 2018. 563-572.

[2] Ng, Kar Yong, and Norhashidah Awang. "Multiple linear regression and regression with time series error models in forecasting PM10 concentrations in Peninsular Malaysia." Environmental monitoring and assessment 190.2 (2018): 63.

[3] Abhilash, M. S. K., et al. "Time Series Analysis of Air Pollution in Bengaluru Using ARIMA Model." Ambient Communications and Computer Systems. Springer, Singapore, 2018. 413-426.

[4] Kumar, Anikender, and P. Goyal. "Forecasting of air quality index in Delhi using neural network based on principal

component analysis." Pure and Applied Geophysics 170.4 (2013): 711-722.

[5] Ul-Saufie, Ahmad Zia, et al. "Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA)." Atmospheric Environment 77 (2013): 621-630.

[6] Elangasinghe, Madhavi Anushka, et al. "Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis." Atmospheric pollution research 5.4 (2014): 696-708.

[7] Shaban, Khaled Bashir, Abdullah Kadri, and Eman Rezk. "Urban air pollution monitoring system with forecasting models." IEEE Sensors Journal 16.8 (2016): 2598- 2606.

[8] Cai, Ming, Yafeng Yin, and Min Xie. "Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach." Transportation Research Part D: Transport and Environment 14.1 (2009): 32-41.

[9] Zito, P., H. Chen, and M. Bell. "Predicting Real-Time Roadside CO and NO2 Concentrations using Neural Networks." IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 9 (2008): 514-522

[10] Lei, Kin Seng, and Feng Wan. "Applying ensemble learning techniques to ANFIS for air pollution index prediction in Macau." International Symposium on Neural Networks. Springer, Berlin, Heidelberg, 2012.