# Forecasting Air Pollution using a Modified Compositional Learning Approach

Samuel A. Ajila, IEEE Senior Member
Sytems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
ajila@sce.carleton.ca

Karthik Dilliraj
Sytems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
KarthikDilliraj@cmail.carleton.ca

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671347

***Abstract*— Major air pollutants, especially fine particles PM2.5, are generally associated with adverse health effects, including cardiac and respiratory morbidity. The aim of this paper is to find the best combination of machine learning techniques to forecast the Air Quality Index (AQI) using the Beijing air quality datasets. The dataset consists [among other] of six air pollutant attributes - PM2.5, PM10, $SO_2$, $NO_2$, CO and $O_3$ that are considered important factors in calculating the Air Quality Index. Our initial results showed that Linear Regression model is not adequate in predicting and forecasting the air pollutants. Random Forest and Random Committee models performed better in terms of MAE and RMSE values compared to Linear Regression. Furthermore, it was noticed that Random Forest performs better in terms of accuracy for certain features but not all while Random Committee performs better in other set of features. This shows that using a "single" machine learning approach to predict or forecast the entire features set may not give the best accuracy. So, as a result, a modified compositional learning model with disentanglement using optimized hyperparameters and search space was designed. The results of this novel network show a marked improvement (3.34% to 78%) in terms of MAE and RMSE values when compared to Random Forest and Random Committee.**

***Keywords*— *Compositional Learning, Air Quality Index Optimized Hyperparameter and Search Space, Random Forest Random Committee***

## I. INTRODUCTION

Air pollution has become a serious issue around the world because of the threat posed to human health and the destruction of the environment. Air pollution is particularly noticeable in some parts of the world – examples are the USA, India, and People Republic of China. In 2016, data collection shows that only 84 of the 338 prefecture-level or high-density cities in China attained the national standard for air quality [1]. Air pollution is considered serious environmental issues because pollutants such as fine particles with diameter less than 2.5 μm (PM2.5) are particularly associated with adverse health effects that include respiratory problems and cardiac. Independent research in environmental impacts shows that air pollution caused more than a million deaths per year [2].

The main proposition in this paper is that working with the Beijing Air-Quality Dataset [1], [2], can we forecast the future trend of air pollutants using a set of machine learning algorithms (i.e. compositional learning) compared to a "single" learning approach. The main goal in this paper is be able to forecast the different components concentration in the air pollution at once and for few months ahead instead of few days or weeks. The results of the forecasts can then be used to mitigate air pollution and thereby improve the environment.

The Beijing dataset [2], [3] includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites from the Beijing Municipal Environmental Monitoring Centre. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. The dataset is divided into four regions of Beijing city - Aotizhongxin, Changping, Dingling and Dongsi. In this paper the entire four regions' datasets are used in section III for data understanding and preprocessing in terms of missing data with data imputation.

The remainder of the paper is organized as follows. Section II presents the background, previous work, and methodology. Section III discusses the dataset and data pre-processing focusing on data imputation methods while section IV presents the initial experiments. Section V presents the novel compositional learning with disentanglement and the experiment results. Conclusions and discussions are given in section VI.

## II. BACKGROUND, PRIVIOUS WORK, AND METHODOLOGY

### A. Linear and Non Linear Models

Linear regression models the relationship between dependent variable and independent variables. [4], [5]. When there is only one independent variable, it is called simple linear regression and multiple linear regression if more than one independent variable. Multivariable linear regression is where multiple correlated dependent variables are predicted compared to a single variable. *Linear regression* can be used easily for numeric attributes. It can also be used for any classification as long as regression is performed for each class by setting to "1" for those instances that belong to the class and "0" for those that do not. The problem with linear regression is that the membership values produced are not proper probabilities,

because they can fall outside the range of 0 and 1. However, a related technique called *Logistic Regression* does not have these problems because it builds a linear model based on the transformed target variable instead of approximating the values 0 and 1 directly [4], [5], [6].

A *Multilayer Perceptron (MLP)* is a network of simple neurons that are called perceptron. A perceptron computes a single output from multiple real valued inputs by forming a linear combination to its input weights and putting the output through a nonlinear activation function [6]. The function is attached to each neuron in the network and determines whether it should be activated ("fired") or not, based on whether each neuron's input is relevant for the model's prediction.

### B. Tree-like Models

Random Forest (RF) algorithm is a supervised machine learning algorithm, which uses the decision tree for prediction. A decision tree [5], [6] is a map of the possible outcomes of a series of related choices. Decision tree is generated by analysing all the features of the data based on the probability of the occurrence. However, decision trees can become excessively complex for large number of features and data. Thus, Random Forest is introduced to mitigate the disadvantages of decision tree. Random Committee algorithm is a supervised machine learning algorithm, which also uses the decision tree to predict the data similar to Random Forest [6], [7]. Random committee involves the construction of a number of base classifiers using unique random number seed values and the final classification result is provided by computing the average of the predictions generated by the individual base classifiers [8].

### C. Performance Measure Metrics

The evaluation of machine learning algorithm is important to decide which algorithm provides more accuracy for the given dataset. In this paper, we have used two common performance measures – mean absolute error and root mean square error to evaluate the accuracy of the different algorithms.

Given n set of predictions $y_1, \ldots, y_n$, made by a model *M*, we can define the following:

Mean Absolute Error $(MAE) = \frac{\sum_{j=1}^{n} |y_j - M(d_j)|}{n}$

Root Mean Square Error $(RMSE) = \sqrt{\frac{\sum_{j=1}^{n} (y_j - M(d_j))^2}{n}}$

Where $M(d_1), \ldots, M(d_n)$ is a set of n predictions for a set of test instances $d_1, \ldots, d_n$.

The lower the value of MAE the better the model as the error produced by the model will be low. Unlike MAE, RMSE accounts for both the directions (positive or negative). RMSE provides an estimate of how large the errors are being dispersed. Since the square value is taken between the errors, RMSE tends to be higher than MAE when the sample size and the error increases.

### D. Air Quality Index

Air Quality Index (AQI) is a metric used to communicate how polluted the current air is or forecast to become in the nearest future. In this paper, AQI value is calculated per 24-hour period and as such $SO_2$, $NO_2$, CO, PM2.5 and PM10 concentrations are measured as average per 24-hour. However, $O_3$ concentration is measured as the maximum of 24-hour moving average. AQI value calculated can be classified into different pollution levels [9], [10], [11]. Pollution levels and its descriptions are given in Table I.

TABLE I. AQI POLLUTION LEVELS

| Levels of Concern | Values of AQI | Description |
|---|---|---|
| Good | 0 to 50 | Air quality is satisfactory, and air pollution poses little or no risk. |
| Moderate | 51 to 100 | Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution. |
| Unhealthy for Sensitive Groups | 101 to 150 | Members of sensitive groups may experience health effects. The general public is less likely to be affected. |
| Unhealthy | 151 to 200 | Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects. |
| Very Unhealthy | 201 to 300 | Health alert: The risk of health effects is increased for everyone. |
| Hazardous | 301 and higher | Health warning of emergency conditions: everyone is more likely to be affected. |

### E. Previous Work

Many researches [1, 2, 12, 13, 14] have been done using the Beijing Air Quality Dataset. Our aim in this section is not to summarize the different research works but just to discuss the commonality between these researches. So, the common thing about previous research works using the dataset is that the they all exclusively focused on one air particle concentration – specifically PM2.5 particle. Other particle concentrations ($SO_2$, $NO_2$, CO, $O_3$, and PM10) captured by the dataset are conveniently ignored. It is easy to understand the reason why many researches focused on one particle concentration – firstly PM2.5 is a fine grain particle and the notion is that it carries most of the pollutants but there is no proof of this fact. Secondly, it is relatively easy to model a single attribute for timeseries forecasting over time. The tendency is that the different algorithms can be trained effectively based on single attribute. However, it can be very difficult to deal with models for forecasting multi-attributes. Our approach in this paper is to look at the entire dataset not just one or few particle concentrations but all the six pollutants ($SO_2$, $NO_2$, CO, $O_3$, and PM2.5, PM10). To the best of our knowledge, this is the first time that a forecasting approach will address all the six particles in the dataset at once using a novel architecture. Additionally, we want to start by doing a thorough analysis of the dataset – understanding the data and prepossessing.

### F. Methodology

Our methodology is to first of all pre-processed the dataset using scatter plots and imputation methods. Imputation methods are generally used to fill-in the missing data. Secondly, we apply Linear regression, Random Forest, and Random Committee machine learning algorithms to the pre-processed dataset and use MAE and RMSE as performance indicators.

Thirdly, we develop a novel architecture using compositional learning with disentanglement. Compositional learning is based on the idea that more than one machine learning models are needed to effectively recognize the patterns in a dataset. Disentanglement learning is an unsupervised learning technique that can be used to breakdown or disentangle dataset features. The use of disentanglement in this paper is different from the traditional use of the term in the sense that we are interested in addressing each feature with respect to the best hyperparameters setting and search space that will produce the best performance accuracy. So, the dataset is disentangled to find the best hyperparameters and search space.

## III. Dataset and Data Preprocessing

### A. Beijing Air Quality Dataset

Beijing Air Quality Dataset [10], [11], [12] is partitioned into four regions Aotizhongxin, Changping, Dingling and Dongsi. All of the regions are initially analysed in terms of missing data. The best imputation method and scatter diagram are then applied specifically to Aotizhongxin region. The time period of the dataset is from March 1st, 2013 to February 28th, 2017. The attributes of the dataset are given in Table II.

TABLE II. DATASET ATTRIBUTES

| Attributes | Description |
|---|---|
| No | Row Number |
| year | Year of the data |
| month | Month of the data |
| day | Day of the data |
| hour | Hour of the data |
| PM2.5 | PM2.5 Concentration (ug/$m^3$) |
| PM10 | PM10 Concentration (ug/$m^3$) |
| $SO_2$ | $SO_2$ Concentration (ug/$m^3$) |
| $NO_2$ | $NO_2$ Concentration (ug/$m^3$) |
| CO | CO Concentration (ug/$m^3$) |
| $O_3$ | $O_3$ Concentration (ug/$m^3$) |
| TEMP | Temperature (degree Celsius) |
| PRES | Pressure (Pa) |
| DEWP | Dew point Temperature (degree Celsius) |
| RAIN | Precipitation (mm) |
| wd | Wind direction |
| WSPM | Wind speed (m/s) |
| station | Name of the air-quality monitoring site |

### B. Dataset Preprocessing – Missing data and scatter plots

The summary statistics of the entire dataset is given in Table III. These results (i.e. Table III) will be used as a base for comparison with the different imputation results.

TABLE III. ATTRIBUTE STATISTICS BEFORE IMPUTATION

| Concentration | Minimum | Median | Mean | Maximum | NA's |
|---|---|---|---|---|---|
| PM2.5 | 3.00 | 58.00 | 82.77 | 898.00 | 925 |
| PM10 | 2.00 | 87.00 | 110.10 | 984.00 | 718 |
| $SO_2$ | 0.29 | 9.00 | 17.38 | 341.00 | 935 |
| $NO_2$ | 2.00 | 53.00 | 59.31 | 290.00 | 1023 |
| CO | 100.00 | 900.00 | 1263.00 | 10000.00 | 1776 |
| $O_3$ | 0.22 | 42.00 | 56.36 | 423.00 | 1719 |

An examination of the NA's (number of missing values) column of Table III shows that each attribute has considerable amount of missing values. Learning from the data with this amount of missing values may lead to poor accuracy for prediction and forecasting. So, it is necessary to predict the missing values for better forecasting. We used three different imputation methods – regression (mean/median/mode), kNN, and MICE (Multivariate Imputation by Chained Equations). Imputation is replacing the missing values with some estimated ones [15], [16]. For the sake of space, we will not show the full results but summary of the imputation methods.

#### 1) Summary of imputation methods

Regression imputation is replacing the missing values with the mean/median/mode and it is a crude way of treating missing values. Using this approach, the median value from this imputation gets deviated more compared to the statistics before imputation (i.e. Table III). This means that since there are a number of missing values in the dataset, imputation with mean alters other metrics median drastically, so the statistics of the entire dataset is getting changed which is not satisfactory. Neighbour-based imputation uses k-Nearest Neighbours approach to impute missing values. Using k-Nearest Neighbours, the median value deviated from the original data (i.e. Table III) is less compared to regression method. We use k = 10 and it is important to note here that the best value of k is chosen after twenty iterations with kNN, k = 1 to 20.

A third imputation method is MICE (Multivariate Imputation by Chained Equations). It uses a slightly uncommon way of implementing the imputation in two steps, multiple models with different imputations are built; and final complete imputation is generated and returned [16]. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations.

TABLE IV. ATTRIBUTE STATISTICS AFTER MICE IMPUTATION

| Concentration | Minimum | Median | Mean | Maximum |
|---|---|---|---|---|
| PM2.5 | 3.00 | 58.00 | 82.45 | 898.00 |
| PM10 | 2.00 | 86.00 | 109.80 | 984.00 |
| $SO_2$ | 0.29 | 9.00 | 17.38 | 341.00 |
| $NO_2$ | 2.00 | 54.00 | 59.38 | 290.00 |
| CO | 100.00 | 900.00 | 1266.00 | 10000.00 |
| $O_3$ | 0.22 | 41.00 | 55.22 | 423.00 |

From Table IV, the statistics of the dataset after the MICE imputation is similar to the original data statistics (Table III) before imputation. There is not much deviation in the median and mean value (see Tables III and IV) for each attribute predicted. It can be concluded that imputation using MICE provides the best result among all three imputation methods. After imputation of missing values, we pre-processed the dataset to daily (24-hourly) data instead of hourly data for all the air pollutants concentration. Table V gives the results of MICE imputation for Aotizhongxin region. The rest of this paper is then based on Aotizhongxin region dataset only.

TABLE V. ATTRIBUTE STATISTICS FOR AOTIZHONGXIN

| Concentration | Minimum | Median | Mean | Maximum |
|---|---|---|---|---|
| PM2.5 | 5.50 | 63.21 | 82.45 | 512.29 |
| PM10 | 7.75 | 91.92 | 109.81 | 545.88 |
| SO2 | 2.00 | 10.38 | 17.38 | 134.08 |
| NO2 | 6.83 | 54.54 | 59.38 | 173.38 |
| CO | 158.30 | 941.70 | 1266.30 | 7487.50 |
| O3 | 1.00 | 81.62 | 94.25 | 343.38 |

We give in Fig. 1, examples of observed distribution of concentration for three pollutants. Fig. 1a represents scatter plot for the Aotizhongxin region PM2.5 concentration vs Date. It can be observed that the PM2.5 concentration value starts increasing day by day and after attaining a maximum value it decreases again to low value. The maximum PM2.5 concentration values are mostly obtained in the winter months of Beijing in China - December to February and minimum values in the summer months from June to August. The minimum and maximum PM2.5 concentration values are approximately around 10 ug/m$^3$ and 500 ug/m$^3$ in the 5 years period.

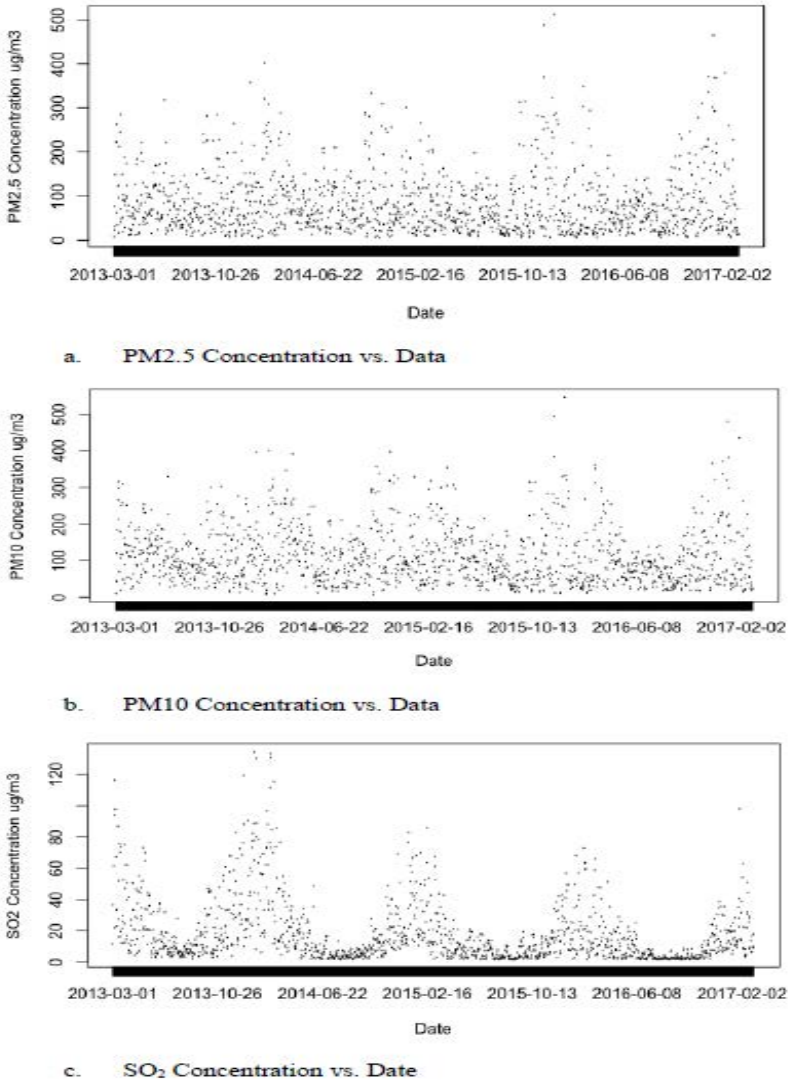


Fig. 1. PM2.5, PM10, and SO2 Concentrations vs. Date

In Fig. 1b, it is observed that the PM10 concentration values approximately ranges from 10 ug/m$^3$ to 500 ug/m$^3$ similar to PM2.5. Also, the maximum and minimum values are obtained often in the winter months and summer months respectively similar to PM2.5. The scatter plot of $SO_2$ concentration vs Date is shown in the Fig. 1c. The $SO_2$ concentration values are in the range from 5 ug/m$^3$ to 120 ug/m$^3$. The maximum and minimum values of $SO_2$ concentration (Fig. 1c) are more distinguishable as compared to the PM2.5 and PM10 concentrations. But similar to PM2.5 and PM10, the maximum and minimum concentration values are predominantly in the winter and summer months.

Similarly, the $NO_2$ concentration values (not shown in the Fig.) ranges from 5 ug/m$^3$ to 170 ug/m$^3$. Like the three previous concentration plots the maximum and minimum values are frequently obtained in the winter and summer months. But the $NO_2$ concentrations have more scattered set of values compared to the three previous pollutants. The CO concentration values are very high as compared to other pollutants. The values are roughly in the range of 1500 ug/m$^3$ to 8000 ug/m$^3$. The CO concentration also has comparable pattern of values in each year. The $O_3$ concentration values are bounded by 1 ug/m$^3$ to 350 ug/m$^3$. The change in concentration values with respect to date follows the same pattern (i.e. winter to summer) as the other pollutants. Note that we did not show the scatter plots for $NO_2$, CO, and $O_3$ because of space.

From the results of the scatter plots, the seasonal pattern of the different pollutants can be observed. This gives an understanding of the dataset pattern. There are more pollutants in winter compared to summer season. These information (i.e. the scatter plots and the imputations) will be used when interpreting the prediction and forecasting results in sections IV and V.

## IV. INITIAL EXPERIMENTS AND RESULTS

Although the primary aim of this paper is to present the novel compositional learning architecture we started by first conducting some baseline experiments with three different machine learning algorithms. The purpose is to use the results of these experiments as comparisons with the results of the disentangled and compositional learning approach in section V.

The experimental environment consists of *Intel Core i7-10700F processors, 16 CPUs, 64 cores, GeForce RTX 3060, and CentOS Linux 8 with 16G RAM, and 2 Terabytes storage*. Note that the experiments and the results in this section are based Aotizhongxin region dataset only. The three machine learning algorithms initially examined are Linear Regression, Random Forest and Random Committee. The choice of these three algorithms is based on the fact that they have been used in previous research works on the dataset. In particular, linear regression is very popular with timeseries forecasting. The dataset is split into 60% training and 40% test data samples for the prediction [22]. The models are evaluated using two metrics Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Then, the AQI value is calculated from the forecasted values of PM2.5, PM10, $SO_2$, $NO_2$, CO and $O_3$ concentrations.

### A. *Linear Regression model*

Fig. 2 presents the prediction results for linear regression model based on 60% training and 40% testing. The results shown is the test data.

The 1-step ahead predictions plot is obtained by taking every prediction of the test data. The evaluation metrics MAE and RMSE for the predictions (test data) are very high in the order of $10^6$. Thus, very high value of evaluation metrics makes Linear Regression not particularly suitable for predicting the entire air quality pollutants.

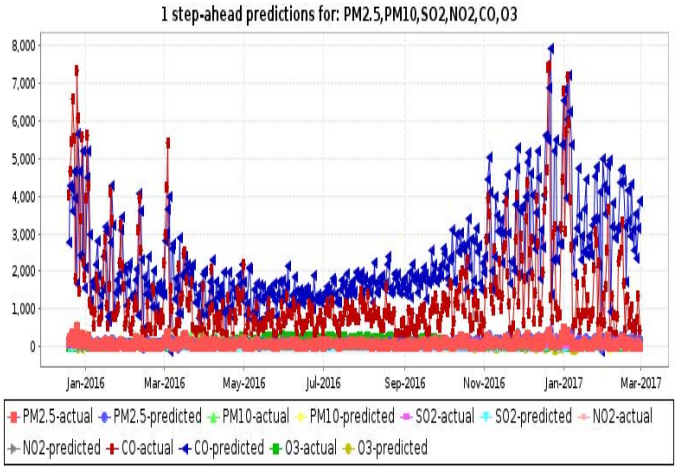


Fig. 2 Linear Regression Test Data Predictions for Aotizhongxin

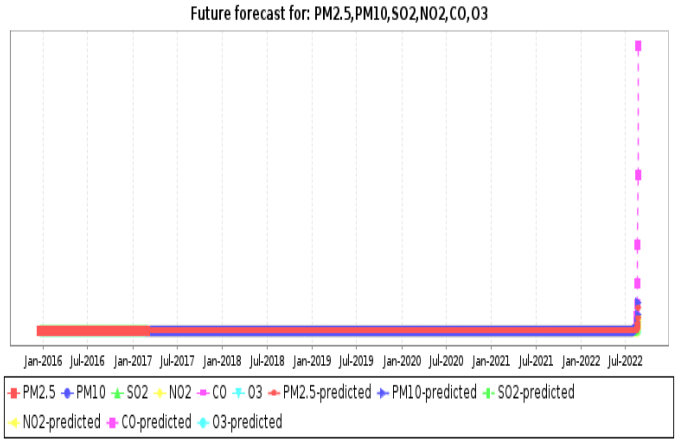


Fig. 3 Linear Regression Future Forecast for Aotizhongxin

In addition, using Linear Regression for future forecast (Fig. 3) shows that the forecasts remain flat for a period (up till after July 2022) and suddenly increases vertically without dropping. As the forecast time unit increases the concentrations of air pollutants suddenly becomes large and both the MAE and RMSE values for future forecast are in the order of $10^{85}$ which is not a possible value in the real life. As a result, linear regression model may not be the best algorithm for long term forecasting.

### B. *Random Forest*

The 1-step ahead predictions and the future forecast plots obtained from Random Forest model on the test data are shown in the Figs. 4 and 5 respectively.

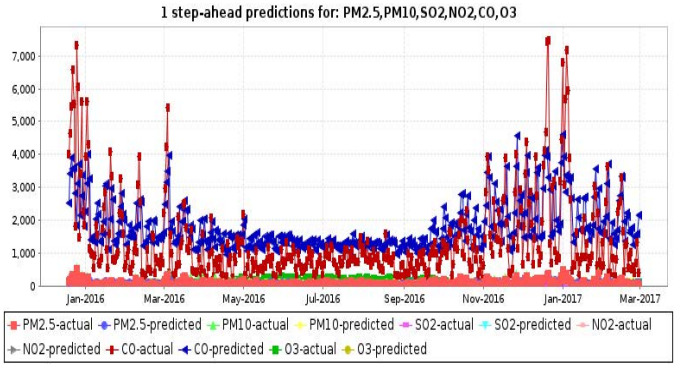


Fig. 4 Random Forest Test Data Predictions for Aotizhongxin

These results (i.e. Figs. 4 and 5) suggest that both the predictions and the future forecasts are better as compared to the Linear Regression (Figs. 2 and 3). Unlike the Linear Regression, the concentration of air pollutants obtained are acceptable, not having very high values. Also, the future forecast values obtained are less than the maximum concentration value determined in the statistics of the dataset (Table V). Since the CO pollutant concentration have very high values compared to other pollutants, the changes can easily be noticed. The analysis of the performance metrics (RMSE and MAE) will be discussed later (in Table VI).

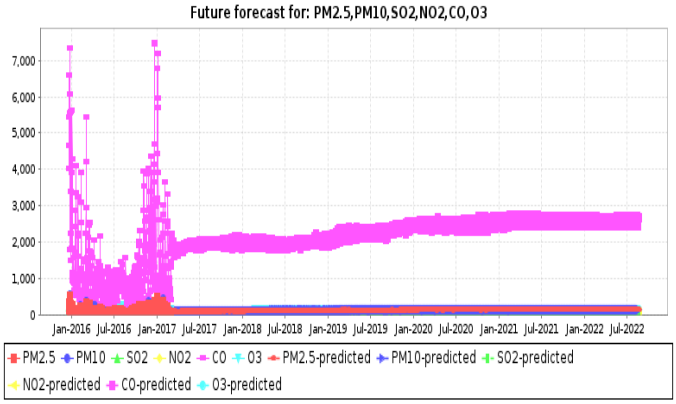


Fig. 5 Random Forest Future Forecast for Aotizhongxin

### C. *Random Committee*

The 1-step ahead predictions and the future forecast plots obtained from Random Committee model on test data shown in the Figs. 6 and 7 respectively. The results indicate that Random Committee model is better than Linear Regression model for the Aotizhongxin region dataset. The concentrations of air pollutants obtained are also comparably similar to that of Random Forest model barring the CO concentration. Forecast values are excellent during the initial forecast period of time around 365 days having less error but as the forecast unit increases the error metrics kept on aggregating.

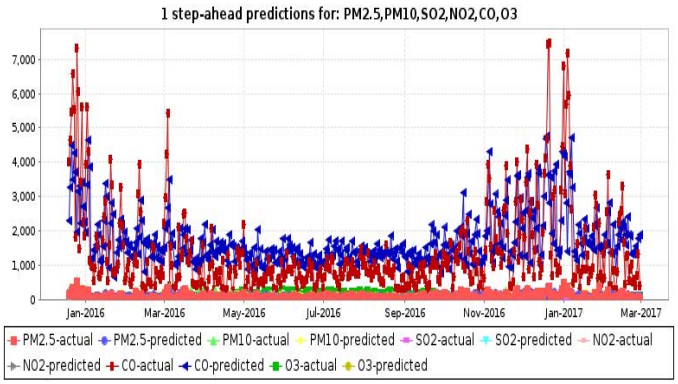


Fig. 6 Random Committee Test Data Predictions for Aotizhongxin

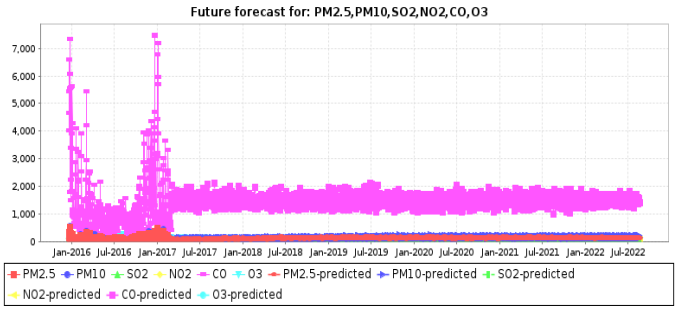


Fig. 7 Random Committee Future Forecast for Aotizhongxin

The predicted values of the test data are almost identical to the actual test data except for the CO air pollutant concentration. Regardless of the difference in values for CO concentration, the values predicted are feasible and the calculation of AQI can be done.

The plots obtained from the models conclude that Random Forest and Random Committee algorithm perform better for the Aotizhongxin region dataset compared to linear regression. The best algorithm between these two is obtained by comparing the overall MAE and RMSE metrics for the predictions (Table VI).

*Table VI MAE & RMSE values for* Aotizhongxin

| **Features** | **MAE** | | **RMSE** | |
|---|---|---|---|---|
| | Random Forest | Random Committee | Random Forest | Random Committee |
| PM2.5 | 55.405 | 58.651 | 74.539 | 78.134 |
| PM10 | 58.028 | 58.667 | 79.531 | 82.240 |
| $SO_2$ | 11.043 | 11.081 | 14.124 | 14.468 |
| $NO_2$ | 30.135 | 27.489 | 34.306 | 33.276 |
| CO | 1240.897 | 1147.630 | 1428.121 | 1403.337 |
| $O_3$ | 72.740 | 66.895 | 95.493 | 88.635 |

From Table VI, the performance evaluation values of MAE and RMSE are less for the PM2.5, PM10 and $SO_2$ concentrations in the case of Random Forest compared to Random Committee. But for the $NO_2$, CO and $O_3$ concentrations, Random Committee model has slightly better MAE and RMSE performance values compared to Random Forest.

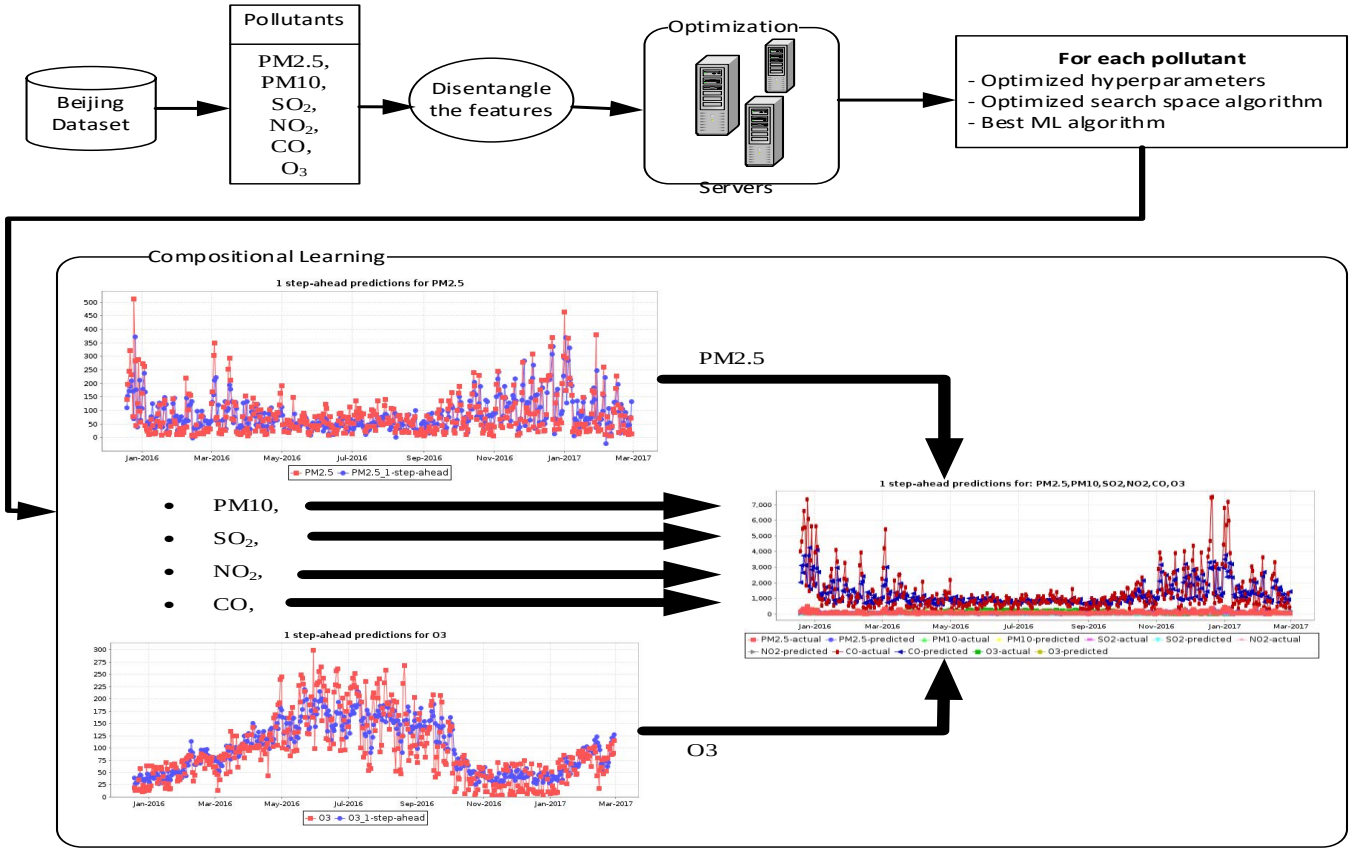


Fig. 8 Commotional Learning Architecture

*Table VII The results of using different algorithms*

| | | Attribute | | Performance | |
|---|---|---|---|---|---|
| **Features** | **Best Classifier** | **Search** | **Evaluator** | **MAE** | **RMSE** |
| PM2.5 | Linear Regression | Best First | Cfs | 39.2378 | 53.9483 |
| PM10 | meta.bagging [MLP] | null | null | 46.2151 | 60.6940 |
| $SO_2$ | Linear Regression | null | null | 08.8364 | 13.6525 |
| $NO_2$ | Additive Regression | null | null | 16.2288 | 21.3743 |
| CO | Random Forest | Best First | Cfs | 252.4395 | 334.1110 |
| $O_3$ | meta.bagging [MLP] | Best First | Cfs | 29.4006 | 39.3750 |

Three conclusions can be drawn from the initial experiments. Firstly, the performance evaluation (i.e. the MAE and RMSE values) of the two machine learning models is different for different features – Random Forest performs better for certain features of the air quality attributes while Random Committer performs better for a different set of the features. Secondly, since the AQI value is mostly affected by the two features PM2.5 and PM10 concentrations, we may conclude that Random Forest model performs slightly better [overall] than the Random Committee for the dataset.

The third conclusion is a hypothesis: *"Using a modified compositional learning approach with optimized hyperparameters and improved search method will increase the prediction and forecast accuracies in terms of MAE and RMSE [23], [24]."* This hypothesis is proposed as a result of the experimental results obtained from Random Forest and Random Committee models. The subject of next section is to show if this hypothesis can be accepted or rejected.

## V. A MODIFIED DISENTANGLED AND COMPOSITIONAL LEARNING NETWORK

In order to accept or reject the hypothesis proposed at the end of section IV, we design a novel architecture for a modified compositional learning to find the best algorithm, search space, and attribute evaluator for each air quality feature (see Fig. 8).

Fig. 8 left to right and top to bottom shows the layout of our compositional learning network. The dataset is disentangled feature by feature. Disentanglement learning is in general an unsupervised learning technique that can be used to breakdown or disentangle dataset features. In our case, we modified this approach to apply optimization to each feature.

The pollutants are selected for optimization. For each pollutant, the best machine learning algorithm with optimized hyperparameters, and the best search space algorithm for the feature are identified through experiments. The results of this stage of learning is given in Table VII columns 1 to 4, and the 1-step ahead prediction for each feature is given in Fig. 9.

The best attribute evaluator selected for three of the features (Table VII) is *CfsSubsetEval (Cfs)*. Cfs evaluates the worth of attributes subset when considering the individual predictive ability of each feature with the degree of redundancy between them [17], [18], [19]. In general, the subsets that are highly correlated while having inter-correlation are preferred. Other types of attribute evaluators include ChiSquared-, Classifier-, Consistency-, Principal Components, and GainRatio-attribute evaluators [20], [21]. The *null* value indicates a choice of any attribute evaluator or search space.

The best search space selected for the two of the features is *Best First* method (Table VII). Best first searches the space of attribute subsets using greedy hill climbing augmented with a backtracking capacity. This search algorithm (*Best First*) may start with an empty set of attributes and search forward, or start with all the set of attributes and search backward, or start at any point searching in both directions.

The results shown in Table VII indicate that for PM2.5, the best classifier is Linear Regression with *Best first* search method, and *Cfs* for attribute evaluator. PM10 has Multi-layer Perceptron with bagging as the best classifier and any type of search method and attribute evaluator. The best classifier for $SO_2$ is Linear Regression and any type of search method and evaluator. The best algorithm for $NO_2$ is Additive Regression method using any type of search method and evaluator. Random Forest is the best classifier for CO with *Best first* as the best search method and *Cfs* as the best attribute evaluator. Finally, Multi-layer Perceptron is the best algorithm for $O_3$ with *Best first* search method and *Cfs* for attribute evaluator.

The next step is to apply compositional learning (see bottom of Fig 8, left to right) using the output of the disentangled learning with optimization results. Compositional learning is based on the idea that more than one machine learning models are needed to effectively recognize the patterns in a dataset. The performance metrics in term of MAE and RMSE values are given in Table VII columns 5 and 6. The 1-step ahead prediction with compositional learning is shown in Fig. 10.

*Table VIII Comparing compositional learning model with Random Forest and Random Committee*

| **Features** | **MAE** | | | **RMSE** | | | **MAE** | **RMSE** |
|---|---|---|---|---|---|---|---|---|
| | Random Forest (RF) | Random Committee (RC) | Compositional Learning (CL) | Random Forest (RF) | Random Committee (RC) | Compositional Learning (CL) | % improvement of CL over "single" learners | |
| PM2.5 | 55.405 | 58.651 | 39.2378 | 74.539 | 78.134 | 53.9483 | 28.18 | 27.62 |
| PM10 | 58.028 | 58.667 | 46.2151 | 79.531 | 82.240 | 60.6940 | 20.36 | 23.69 |
| $SO_2$ | 11.043 | 11.081 | 08.8364 | 14.124 | 14.468 | 13.6525 | 19.98 | 03.34 |
| $NO_2$ | 30.135 | 27.489 | 16.2288 | 34.306 | 33.276 | 21.3743 | 40.96 | 35.77 |
| CO | 1240.897 | 1147.630 | 252.4395 | 1428.121 | 1403.337 | 334.1110 | 78.00 | 76.19 |
| $O_3$ | 72.740 | 66.895 | 29.4006 | 95.493 | 88.635 | 39.3750 | 56.05 | 55.58 |

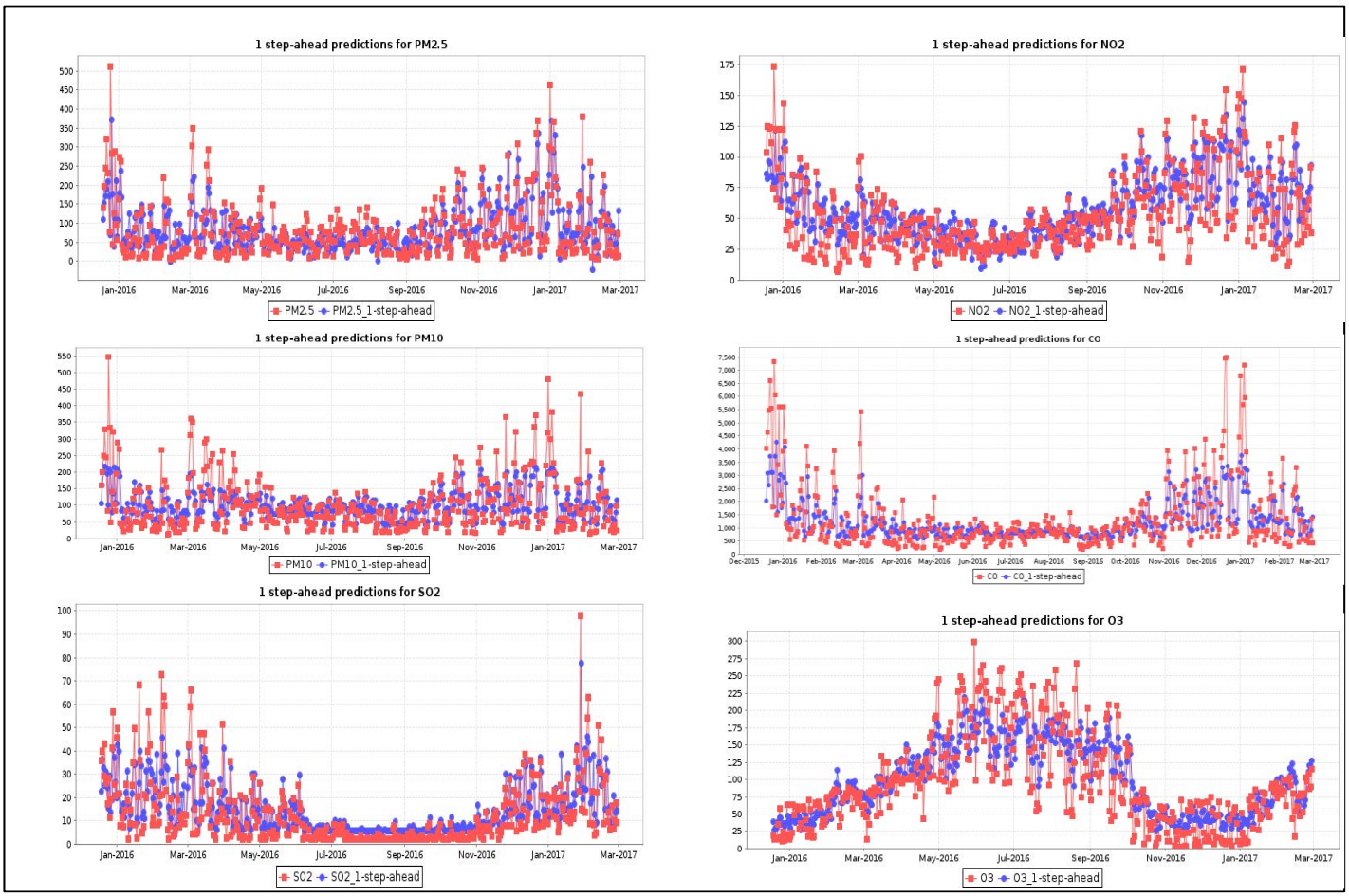


Fig. 9 Disentangled learning for ML with optimized hyperparameters and search space

Table VIII shows the comparisons between the performance metrics in term of MAE and RMSE values for Random Forest (RL), Random Committee (RC), and Compositional Learning (CL). The percentage improvement of CL in the case of MAE values ranges from 19.98% to 78.00%. The improvement in the case of two particles – *CO* and $O_3$ is particularly interesting – 78% for *CO* and 56.05% for $O_3$. Similarly, for RMSE values, the percentage improvement of CL ranges from 3.34% to 76.19%. In fact, the 3.34% for $SO_2$ is more or less an outlier because the next smaller value is 23.69% for *PM10*. Similar results are obtained in the case of *CO* with 76.19% and $O_3$ with 55.58%.

Furthermore, we used 40% of the dataset to test the trained compositional learning model and generate a future forecast between April 2013 and January 2016 (see Fig. 11).

A closer examination of Fig. 11 and comparing it with the scatter plots in Fig. 1 generated for the raw dataset during the pre-processing shows that the prediction patterns in Fig.11 is visually similar to that of the original dataset. This further confirms that using compositional learning with the best search method and attribute evaluator for each attribute is likely to generate a better prediction and forecast accuracy compare to just using a single algorithm learning approach.

Fig. 12 shows future forecast beyond the life of the dataset using the trained compositional learning model. The generated performance metrics in term of MAE and RMSE values are similar to the results shown in Table VIII.

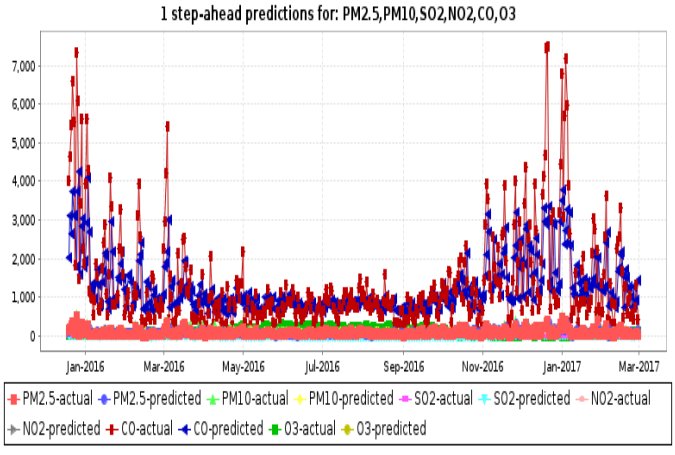


Fig. 10 Aotizhongxin 1-step ahead prediction with compositional learning

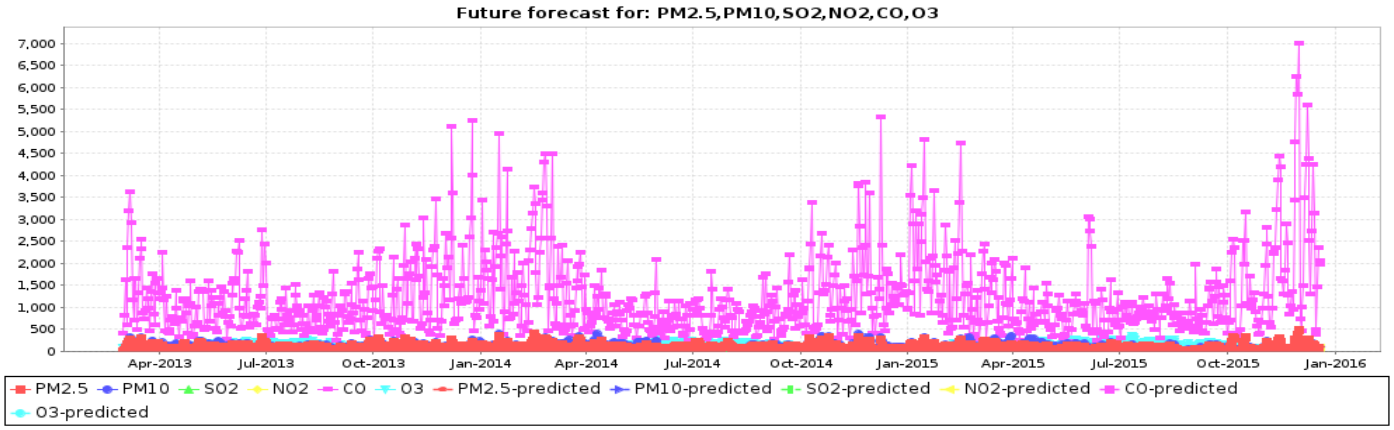


Fig. 11 Aotizhongxin Future forecast using 30% testing

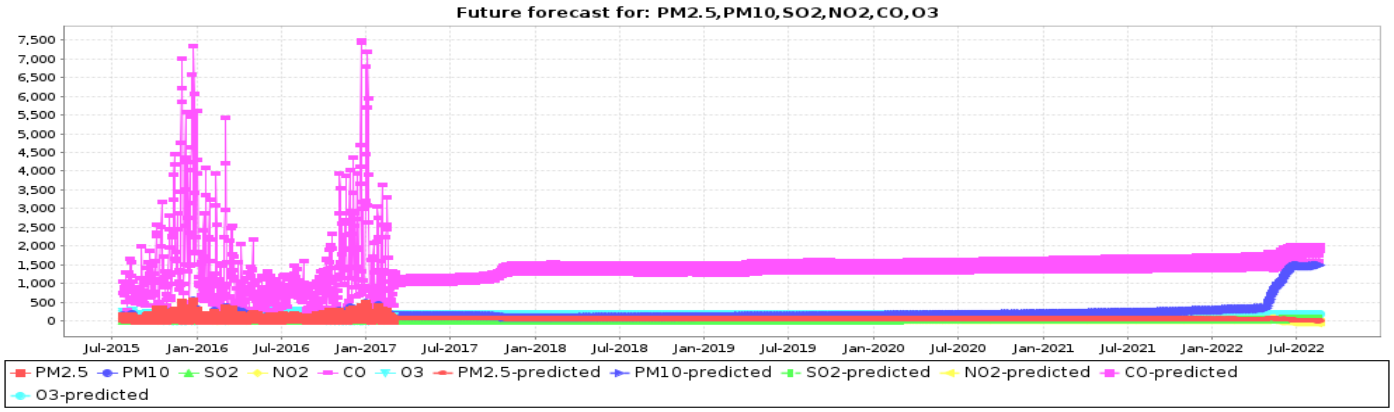


Fig. 12 Aotizhongxin future forecast using compositional learning

## VI. DISCUSSIONS AND CONCLUSIONS

The prediction results (Figs. 2 and 3) of Linear Regression in terms of MAE and RMSE show that using the algorithm is not particularly a good fit for all of the air quality pollutants. The MAE and RMSE values are in the order of $10^6$ for the predictions while that of forecasting are in the order of 1085 for both MAE and RMSE.

The prediction and forecast results (Figs. 4, 5, 6, and 7) generated by both Random Forest and Random Committee indicate a better performance compared to Linear Regression. Comparing both Random Forest and Random Committee in terms of MAE and RMSE values (Table VI) led to the following observations. First, the Random Forest performance evaluation values for MAE and RMSE are better for the PM2.5, PM10 and $SO_2$ concentrations compared to Random Committee. Second, for the $NO_2$, CO and $O_3$ concentrations Random Committee model has slightly better MAE and RMSE performance values compared to Random Forest. These observations show that using different machine learning algorithm for different pollutant feature is likely to produce a better prediction and forecast results compared to using just a single machine learning model for all the attributes.

As a result, a hypothesis was proposed, that *"using a modified compositional learning approach with optimized hyperparameters and improved search method will increase the prediction and forecast accuracies in terms of MAE and RMSE."*

The experimental results obtained (Table VIII) confirmed the hypothesis. Comparing the results of using only Random Forest or Random Committee for all the attributes with using compositional learning, optimized hyperparameters and improved search space shows that in all cases (Table VIII) compositional learning model performs better in terms of MAE and RMSE values.

The conclusion that can be drawn here is that there is no one fit-all machine learning algorithm when it comes to forecasting the future especially dealing with weather related features. So, it is very important to first understand the dataset with respect to the different features and the relationships between these features. Additionally, experiments need to be carried out to find the best way to predict each feature in terms of machine learning model, attribute search method, and attribute space evaluation.

A possible threat to validity in this research is in converting the dataset from hourly to 24-hourly (i.e. daily) instances. It is very difficult to say if the results obtained from the different experiments will be different if the dataset is left as hourly instances. Another source of threat could be with the dataset.

We only focused on one region in terms of the experiments presented in sections IV and V, and we are not sure if the results obtained will be different if the four regions are merged into one big dataset.

In conclusion, we have used all the four regions datasets for the pre-processing in terms of missing data and imputation. However, the training, testing, and forecasting results presented in sections V and VI are based on one region only, that is, Aotizhongin region of Beijing. In the process of our experiments, we forecasted five-year data (mid-2017 to mid-2022) for the six air pollutant features (PM2.5, PM10, $SO_2$, $NO_2$, CO, and $O_3$) using Random Forest and Random Committee algorithms. In addition, using the trained compositional learning model we generated five-year (mid-2017 to mid-2022) for the six features with better performance compared to Random Forest and Random Committee. Unfortunately, we have no way of comparing these generated sets of data with the dataset captured from mid-2017 to date by the Beijing air-quality authority because we have no access to the current dataset.

In future work, the authors will like to research the application of deep machine learning approach for data pre-processing, prediction, and forecasting. Furthermore, a different imputation approach will be applied to the missing data. Finally, each of the four regions will be processed independently comparing the results obtained. In addition, the datasets will be merged into one big dataset and the prediction and forecasting results will be compared to individual region.

## REFERENCES


[1] Shuyi Zhang1, Bin Guo, Anlan Dong, Jing He, Ziping Xu, Song Xi Chen, Cautionary tales on air-quality improvement in Beijing, Proc. R. Soc. A 473: 20170457, http://dx.doi.org/10.1098/rspa.2017.0457, April 2017

[2] Liang X, Zou T, Guo B, Li S, Zhang H, Zhang S, Huang H, Chen SX. 2015 Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proc. R. Soc. A 471, 20150257. (doi:10.1098/rspa.2015.0257)

[3] Beijing Multi-Site Air-Quality Data Set. (2019, Sep 20) Retrieved from https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data# , Accessed March 2020

[4] A. Das, S. A. Ajila, C.-H. Lung, A Comprehensive Analysis of Accuracies of Machine Learning Algorithms for Network Intrusion Detection, IFIP International Federation for Information Processing 2020, Published by Springer Nature Switzerland AG 2020, S. Boumerdassi et al. (Eds.): MLN 2019, LNCS 12081, pp. 40–57, 2020.

[5] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics, The MIT Press, Cambridge, Massachusetts, USA, 2015, ISBN 978-0-262-02944-5

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining – Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann, Elsevier, New York, USA, 2017, ISBN 978-0-12-804291-5

[7] Paul Cohen, Empirical Methods for Artificial Intelligence, MIT Press, ISBN 0-262-03225-2, 1995.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, The MIT Press, Cambridge, Massachusetts, London, England, 2016, ISBN 9780262035613

[9] Python-AQI Package. (2019, Jan 9). Retrieved from https://pypi.org/project/python-aqi/, Accessed April 2020

[10] AQI Basics. Retrieved from https://www.airnow.gov/aqi/aqi-basics/, Accessed April, 2020

[11] Mengmeng Liu and Xin Cao., Beijing PM2.5 Time Series Analysis and Prediction using Regression Models. (2016, May 2) Retrieved from https://www2.isye.gatech.edu/~yxie77/6416_spring16_proj/CSE6416_PM25%20Beijing%20Time%20Series%20Analysis_Report_MengmengLiu_XinCao.pdf, Accessed March 2020

[12] Mingye Yang., A Machine Learning Approach to Evaluate Beijing Air Quality. (2018, June) Retrieved from https://www.math.ucdavis.edu/files/2015/2717/8083/Mingy_Yang_Spring_2018.pdf, Accessed March, 2020

[13] Shengdong Du, Tianrui Li, Yan Yang, and Shi-Jinn Horng, Deep Air Quality Forecasting Using Hybrid Deep Learning Framework, IEEE Transactions on Knowledge and Data Engineering, 2021, Vol. 33:6

[14] Hong Zheng, Haibin Li, Xingjian Lu, and Tong Ruan. A Multiple Kernel Learning Approach for Air Quality Prediction, Hindawi Advances in Meteorology, Volume 2018, Article ID 3506394

[15] James, G. (2015). An introduction to statistical learning: With applications in R.(pp. 303-335) New York: Springer.

[16] Cord, M. (2008)., Machine learning techniques for multimedia: Case studies on organization and retrieval ; 20 tables (pp. 21-49). Berlin: Springer.

[17] Lars Kotthoff, Chris Thornton, Holger H. Hoos, Frank Hutter, and Kevin Leyton-Brown, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA," Journal of Machine Learning Research 18(25), 1 – 5, (2017)

[18] WEKA Select Attributes, available online - https://dataminingntua.files.wordpress.com/2008/04/weka-select-attributes.pdf, Access January 2020.

[19] Lars Kotthoff, Chris Thornton, Holger H. Hoos, Frank Hutter, and Kevin Leyton-Brown, Auto-WEKA: Automatic Model Selection and Hyperparameter Optimization in WEKA, 2019 F. Hutter et al. (eds.), *Automated Machine Learning*, The Springer Series on Challenges in Machine Learning, https://doi.org/10.1007/978-3-030-05318-5_4, pp 81-95

[20] S. Singhal, M. Jena, "A Study on WEKA Tool for Data Pre-processing", International Journal of Innovative Technology and Exploring Engineering, vol. 2, no. 6, pp. 250-253, 2013

[21] R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "WEKA-Experiences with a Java Open-Source Project", Journal of Machine Learning Research, vol. 11, pp. 2533-2541, 2010.

[22] Chen, Sheng & Kan, Guangyuan & Liang, Ke & Li, Jiren & Hong, Yang & Zuo, Depeng & Lei, Tianjie & Xu, Weihong & Zhang, Mengjie & Shi, Wenbin & Chen, Xiaona. (2017). Air Quality Analysis and Forecast for Environment and Public Health Protection: A Case Study in Beijing, China. Transylvanian Review. XXIV. 3575-3593.

[23] M. M. S. Lira, R. R. B. de Aquino, A. A. Ferreira, M. A. Carvalho, O. N. Neto and G. S. M. Santos, "Combining Multiple Artificial Neural Networks Using Random Committee to Decide upon Electrical Disturbance Classification," 2007 International Joint Conference on Neural Networks, Orlando, FL, 2007, pp. 2863-2868.

[24] J. Vanschoren, Meta-Learning, 2019 F. Hutter et al. (eds.), Automated Machine Learning, The Springer Series on Challenges in Machine Learning, https://doi.org/10.1007/978-3-030-05318-5_2 , pp 35-61