# AIR QUALITY PREDICTION BASED BIG DATA ANALYTICS FOR SUSTAINABLE ENVIRONMENT USING DEEP NEURAL MATRIX AND HAVERSINE RATIO

[1]Sathish Kumar Sekar MCA.,

Lecturer, School of International Education, East China University of Technology, Nanchang, Jianxi, China.

[2]Dr Pushpa Vinu Amalraj M.Sc., M.Phil., Ph.D.,

Lecturer, School of Software, East China University of Technology, Nanchang, Jianxi, China.

**Abstract**

With the swift evolutionary speed of modern cities in recent years, urban computing, specifically spatiotemporal analysis, and prediction, has become an important research topic with the growing nature of data (i.e., Big Data). There has been a tendency to use deep learning techniques to address urban air quality challenges. Moreover, a general support vector machine and extreme gradient boosting have proposed spatiotemporal data. However, these methods primarily are constructed on sparsely distributed monitoring cities handling trivial data, focusing less on accuracy and outliers. To efficiently address the proposed air quality dataset with densely deployed monitoring cities based on Big Data analysis, we offer an Air Quality Prediction (AQP) method based on a deep neural network. This method is called Deep Neural Network-based Matrix Completion and Haversine Spatiotemporal (DNN-MCHS) for AQP. The DNN-MCHS method has split into three layers. They are the input layer, hidden layers and output layer. First, the PM2.5 data of five Chinese cities provides input to the input layer. Next, three hidden layers do employ for AQP. They are matrix Completion-based pre-processing performed in the first hidden layer; in this work, the performance evaluation metrics used for the root means squared error (RMSE), air quality prediction time, air quality prediction accuracy, and to verify the efficiency of the proposed methods are utilising the PM2.5 atmospheric dataset. The experimental results show that our DNN-MCHS method performs air quality analysis significantly compared with the state-of-the-art techniques.

**Keywords**: Big Data Analysis, Deep Learning, Air Quality Prediction, Spatiotemporal, Matrix Completion, Haversine Ratio, Air Quality Index

## 1. Introduction

A critical matter for decagons has interpreted the Urban air quality due to the rising distress for human health. Vulnerability to indigent-quality air may result in several allergic retorts and, on a par, give rise to respiratory and circulatory diseases. These diseases can even bring about immeasurable economic squandering due to the insistence on medical treatments and decreased productivity. Hence, the blooming objective for the government is to consider the air quality prediction accuracy and for researchers to bestow pertinent information to the public. However, upon comparison with air quality monitoring, the air quality forecasting methods are completed to be more intricate, and the prediction methods usually differ from country to country.

Spatio-temporal prediction using a support vector machine was proposed in [1] to predict undisclosed space and time air quality. Data were acquired from Geographic Information System (GIS) and included a time series. Initially, the temporal prediction was executed in the reference stations, followed by the air quality index (AQI) is predicted to infer future AQI of undisclosed locations. Validation inferred high accuracy for temporal prediction. This study demonstrated the prediction framework in northern Taipei, enhancing accuracy with minimum error.

Because air quality prediction measurements may subdue missing data aspects, in this work, the aid of low-rank matrix completion is to handle the Matrix Completion-based Pre-processing modes. This filling of missing entries removes the noise and reduces the prediction time, therefore corroborating the objective.

With the increase in the complexity of air quality prediction, in [2], the Integrated Dual Long Short-Term Memory (LSTM) has been proposed. Furthermore, the Seq2Seq (Sequence to Sequence) technique was employed to demonstrate single-factor prediction that can acquire the air quality predicted value. Here, the time-series data in the forecasting process has contemplated the air quality. Following this, the LSTM multi-factor prediction modes utilise the attention mechanism.

The controlling elements of air quality, like the neighbouring station and weather data, were also included in this model. Finally, through XGBoosting (eXtreme Gradient Boosting) tree, to increase the prediction accuracy of air quality of the data with the minimum error was combined into the two models. However, immense and complicated

Big data collection necessitates several levels of exploration, therefore, compromising both accuracy and error. Therefore, a deep learning architecture involving the Haversine Ratio of Variances-based Feature Selection model and the spatiotemporal matrix is designed separately, precisely predicting air quality through AQI, reducing the root mean square error rate in a significant manner.

### 1.1 Contributions

The main contributions of this paper are brief as follows:

- We inspect the Air Quality Prediction for a sustainable environment as an augmentation to the conventional quality prediction problem to maximise the accuracy and minimise time and error rate under a dynamic network.
- To propose Matrix Completion-based Preprocessing model in the first hidden layer, which reduces the time consumed in air quality prediction by first handling the missing data and then models based on low rank to obtain pre-processed air quality data by employing observed and low-rank matrix functions.
- To design Haversine Ratio of Variances-based Feature Selection model in the second hidden layer with pre-processed air quality data selects significant features, reducing the error rate.
- To model spatiotemporal data extraction based on haversine function that ensures robust prediction, therefore ensuring accuracy to a greater extent.
- We evaluate the performance of the air quality Chinese dataset. The experiment results confirm our theoretical findings and show that our proposed Deep Neural Network-based Matrix Completion and Haversine Spatiotemporal (DNN-MCHS) for AQP achieve better performance in terms of air pollution prediction accuracy, time, and error rate.

### 1.2 Organisation

The manuscript continues as follows. Section 2 presents a substantial reference list for related work in prediction and control regarding air quality. The results of this exploration unfold for the novelty of the proposal of this work. Section 3 describes the proposed Deep Neural Network-based Matrix Completion and Haversine Spatiotemporal (DNN-MCHS) method for AQP deployed in Python. Section 4 provides the experimental setup, followed by a discussion with a table and graphical representation. Finally, Section 5 includes the conclusion.

## 2. Related works

With the swift evolution of the world economy and the advancement of industrialisation and urbanisation, several cities globally are suffering from air pollution. Issues concerning air pollution are becoming more discernable, menacing human life and sustainable environment earnestly. Moreover, as PM2.5 is the principal component of air pollution, the growing PM2.5 concentration will also directly influence human health. Hence, PM2.5 concentration predictions in the preliminary stage have critical social aspects that play a significant part in controlling air pollution to smoothen a sustainable environment.

In [3], a deep learning technique for predicting air quality using the Temporal Sliding Long Short-Term Memory Extended method (TS-LSTME). The TS-LSTME process combined the optimal time lag to comprehend sliding prediction via multi-layer bidirectional Long Short-Term Memory (LSTM), ensuring significant air quality accuracy. However, the Air Quality Index (AQI) is the primary tool for measuring prediction rates to describe air quality. In [4], the Back Propagation (BP) neural network was first optimised based on enhanced Particle Swarm Optimization (PSO) to predict AQI. With the deployment of improved PSO and the learning factor, a playing means for fast convergence has ensured the global search potentiality.

Although employing machine learning techniques in predicting air quality for urban concerned in substantial victory, accurate prediction is still a demanding issue to be addressed. In [5], low-cost air quality sensors have been used for 393 deployed air quality monitoring stations to acquire a large-scale dataset. The observed data was initially converted from irregularly distributed stations into regular pollution maps. Further processing has applied the advanced deep convolutional networks.

Multivariate Time series data play significant parts in our day-to-day chores. Using these multivariate time series data for prediction has been a research topic for several academicians. In [6], a novel multivariate time series prediction was designed based on a multi-attention generative adversarial network. This multivariate time series prediction method is split into three distinct parts. Initially, the encoder stage divides into two parts, where the input-attention and self-attention encoded the exogenous sequence into latent space. Next, the decoder stage extracted temporal patterns employing the temporal-convolution-attention module. Finally, weight clipping made long-term predictions, resulting in improved multivariate time series prediction. However, another method using multi-criteria technology was proposed in [7] to reduce the consequences of air pollution.

PM2.5 is a significant index for estimating and managing the degree of air pollution and hence has received a big deal of attention over the past few years. Researchers have also identified that vulnerability to pollutants like PM2.5 shoots up the likelihood of cardiovascular and respiratory diseases. China is venturing into sustainable environmental development; however, it still encounters acute pollution due to the extended air pollutants.

Hybrid-Single Particle Lagrangian Integrated Trajectory (HYSPLIT) [8] perform the backward trajectory, which provides the mechanism for those mobile source particles from Afghanistan. Also, the study analysed the correlation between particulate matter and concentration particles for air pollution forecasting. In [9], the chi-square test was combined with LSTM to evaluate the determining components of air quality. This test ensured the AQI prediction.

In [10], China National Environmental Monitoring Centre employed a semi-supervised method for predicting PM2.5 concentrations. Empirical Mode Decomposition (EMD) and Bidirectional Long Short-Term Memory (BiLSTM) neural networks were combined to decompose the data, extract the frequency, and remove the amplitude features. This method laid a strong foundation for enhancing short-term trend predictions, specifically for abrupt changes, therefore, ensuring accuracy and reducing the error rate to a greater extent. However, they do not address the separation between long and short term, in [11] proposing an investigation of PM10 and PM2.5 concentrations in Italy employing a random forest model. With Aerosol Optical Depth (AOD) data utilising estimates from atmospheric ensemble models, to be extensively suitable for both long term and short-term health effects find prediction quality.

A detailed comparison of machine learning and land use regression in [12] ensures ambient air pollution investigation. However, urban air pollution predictions and controls for rapid response include measures. Parameterised, non-intrusive reduced order model was designed [13] over a large region in China. With this non-intrusive model, the CPU cost was reduced up to five orders of magnitude and ensured rational accuracy.

A new method utilising Multiple Feature Clustering and Neural Networks was designed in [14] to forecast hourly PM2.5 Concentrations in China. In addition, another machine learning method was proposed in [15] using support vector regression to predict the air quality index in California. Also, applying the radial basis function for prediction accuracy was improved.

In [16], based on the Long Short-Term Memory (LSTM) and MultiVerse Optimization (MVO) model for prediction and analysing of air pollution from Combined Cycle Power Plants (CCPP) design and propose a novel

hybrid intelligent method. Here, the CCPP employs to predict the amount of produced NO2 and SO2 in LSTM. Next, to achieve lower forecasting error and high accuracy, the MVO algorithm optimises LSTM parameters.

In [17] investigates an in-depth analysis of air quality prediction between 2005 and 2020. However, do not obtain the spatiotemporal correlations. Graph Convolution Network (GCN) and Convolutional LSTM were proposed in [18], therefore ensuring spatiotemporal deep predictive accuracy to a greater extent. In addition, deep learning mechanisms were investigated in [19] [20] for air pollution prediction.

Motivated by the above state-of-the-art materials and methods proposed by numerous research personalities in air quality prediction and control, this work designs the Deep Neural Network-based Matrix Completion and Haversine Spatiotemporal (DNN-MCHS) methods for AQP. The following section provides a detailed description of the proposed DNN-MCHS methods for air quality prediction.

## 3. Methodology

A deep learning algorithm comprises of hierarchical framework possessing several layers. Each layer constitutes a complicated and non-linear information processing unit. With the rapid evolution of computation patterns, deep learning algorithms have evolved. Moreover, their complexity has enhanced the solutions for the general issues, specifically in predicting air quality. In this work, analysing of the environment for Big Data regarding air pollution with PM2.5 Data from five Chinese cities explores the efficiency of deep learning. Figure-1 shows the block diagram of the proposed Deep Neural Network-based Matrix Completion and Haversine Spatiotemporal (DNN-MCHS) method for AQP.

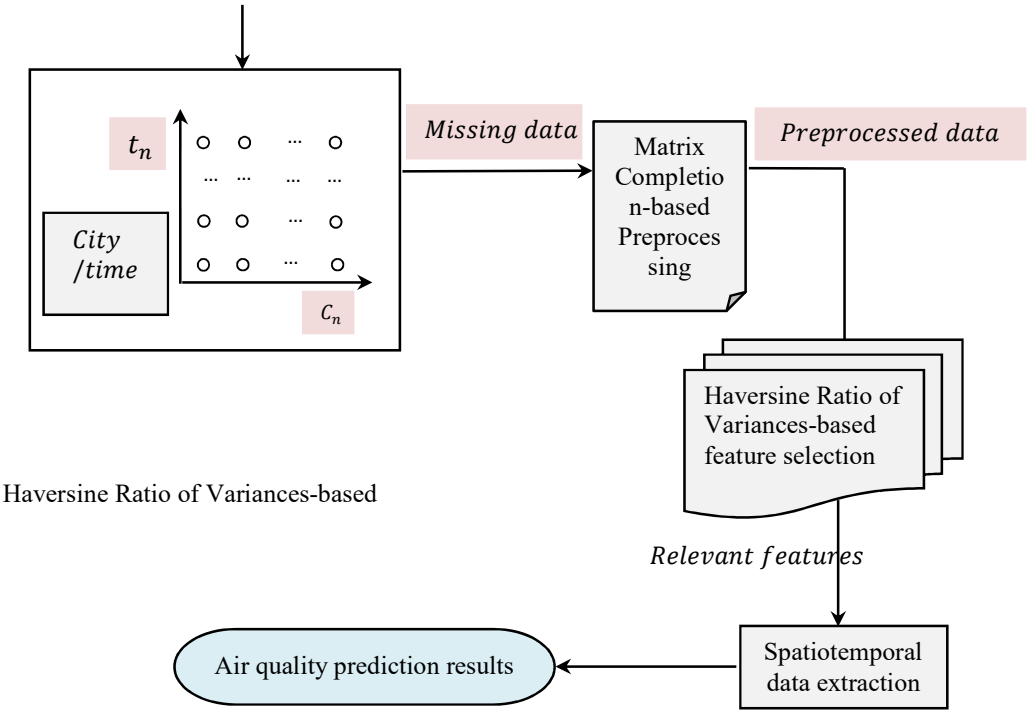
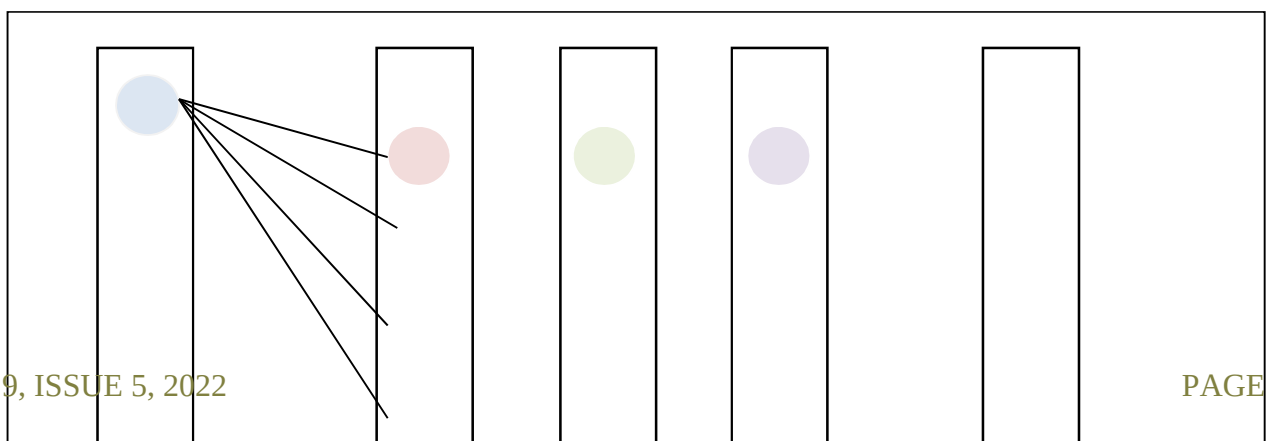



**Figure-1 Block diagram of the DNN-MCHS method**

As shown in the above figure, the proposed DNN-MCHS method for AQP uses matrix Completion-based pre-processing (first hidden layer) and Haversine Ratio of Variances-based Feature Selection (second hidden layer) models. Next, the third hidden layer performs the spatiotemporal data extraction. The proposed work finally introduces a Deep Neural Matrix and Haversine Ratio-based AQP algorithm. The five Chinese cities' benchmark data on particulate matters at different periods has based on the proposed deep learning technique. Figure-2, given below, shows the deep learning architecture for the proposed DNN-MCHS method.



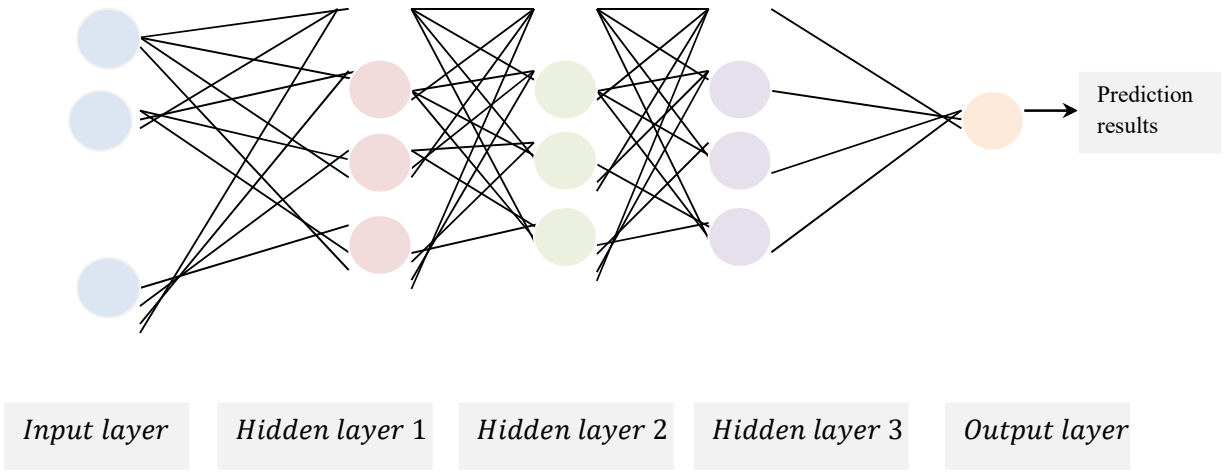



**Figure 2 Deep learning for DNN-MCHS**

As shown in the above architecture for the DNN-MCHS method, five layers are there, one input layer, three hidden layers and one output layer. In this section, giving to the characteristics of the PM2.5 data of five Chinese cities, the construction idea of the air quality data analysis has directed graph first provided and then introduces our deep learning prediction model. Finally, shown in Figure-1, the sliding window method is employed in splitting the data, where '$X-axis$' represents the Directed Graph PM2.5 Data composed of all monitoring cities at time t (i.e., '$Y-axis$').

### 3.1 Directed Graph PM2.5 Data Construction model

Let us consider the spatial distribution of '$C$' monitoring cities at a specific time instance '$t$' as a directed graph '$DG=(V,E,A)$' where '$V$' refers to the monitoring cities (i.e., Beijing, Chengdu, Guangzhou, Shanghai and Shenyang), '$E$' forming the edge set (i.e., 15 features) and '$A \in R^{C*C}$ It Represents the weighted adjacency matrix of the directed graph. Each monitoring city on the directed graph detects the value of PM2.5 concentration at the same sampling frequency.

### 3.2 Input layer

The number of pollution sources across the monitoring cities heavily influences the PM2.5 concentration. In other words, the unbalanced distribution of pollution sources in space obtained from environmental authorities and companies affects the association between the cities. To be considered have the pollution sources obtained around the cities from the PM2.5 data of five Chinese cities. Let a time series be '$F(1), F(2), \ldots F(n)$'. Moreover, the predicted values of the '$t+1$' time series possess the following arrangement.

$$F(t+1) = \alpha_1 F(t) + \alpha_2 F(t-1) + \alpha_3 F(t-2) + \cdots + \alpha_n F(t-q+1) + \varepsilon\,(t+1) \qquad (1)$$

From the above equation (1), '$\alpha_1$' represents the air quality monitoring parameters or features in consideration, with '$q$' denoting the highest order number of air quality monitoring parameters and '$\varepsilon\,(t)$' representing the random noise. Table-1 given below lists the data acquired from the input dataset '$DS$' and provides it as input to the input layer.

**Table-1 PM 2.5 Data Content**

| S. No | Feature | Description |
|---|---|---|
| **1** | No | Row number |
| **2** | Year | Year of data |
| **3** | Month | Month of data |
| **4** | Day | Day of data |
| **5** | Hour | Hour of data |
| **6** | Season | Season of data |
| **7** | PM | PM 2.5 Concentration |
| **8** | DEWP | Dew Point |
| **9** | TEMP | Temperature |
| **10** | HUMI | Humidity |
| **11** | PRES | Pressure |
| **12** | Cbwd | Combined Wind Direction |
| **13** | lws | Cumulated Wind Speed |
| **14** | Precipitation | Hourly precipitation |

| | | |
|---|---|---|
| **15** | lprec | Cumulated precipitation |

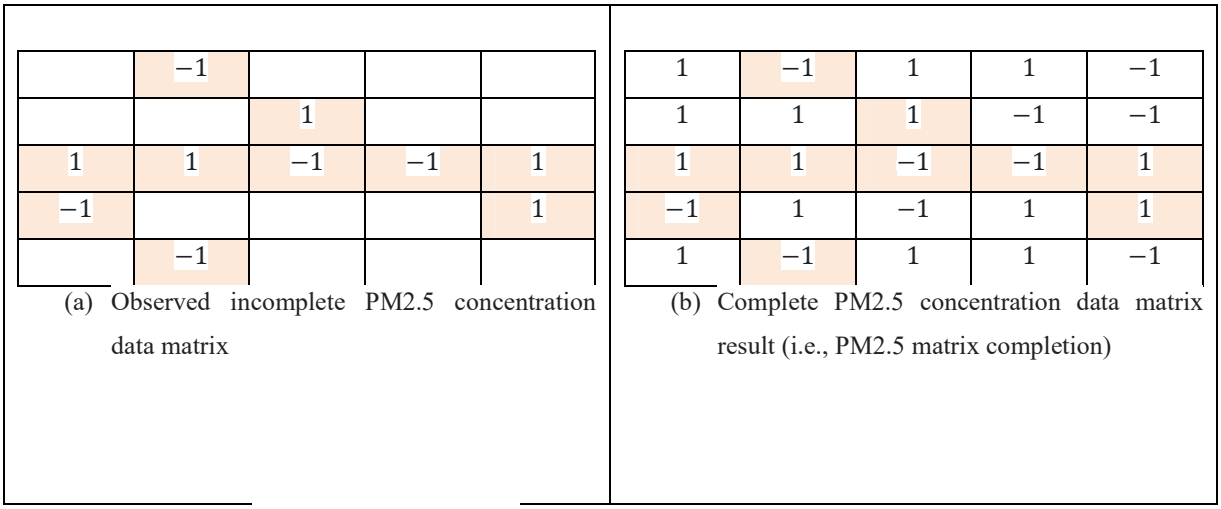
With the above input data, to carry out the next pre-processing task. Air quality reports from environmental authorities and companies have frequently included the PM2.5 reading. To be more specific, PM2.5 refers to atmospheric Particulate Matter (PM) that comprises less than 2.5 micrometres and is a measure of pollution has hance utilised. The period for this PM2.5readings is between Jan 1st, 2010, to Dec 31st, 2015. Table-1, given below, provides the content in the PM2.5 dataset.

### 3.3 Matrix Completion-based Preprocessing model

Data pre-processing is considered a significant step in the deep learning process as it influences the generalisation potentiality of the learning algorithm. As the proposed method uses deep neural networks, numerous layers are present in the execution, including hidden layers. The objective here remains to simplify the overall process by reducing the processing time and the number of attributes. In the proposed method missing data, using Matrix Completion-based Preprocessing model in the first hidden layer or hidden layer 1 handles imputation.

In the PM 2.5 dataset, most missing data are denoted as NA (Not Applicable) and as the percentage of missing data in the employed dataset is more petite, using Matrix Completion as a substitute to get rid of those data performed by replacing missing data in our work. The matrix Completion technique is applied in our work to handle missing data for different concentrations. Matrix completion here refers to filling missing data of a relatively distinguished matrix.

Let us consider the air pollution monitoring for five Chinese cities, given a time '$t$' period matrix for each entry. '$(i, j)$' represents the historical PM2.5 attentiveness data of '$i$' continuous time steps to forecast the PM2.5 attentiveness of '$j$' continuous time steps in the future; if the forecast is made and is else missing, we would like to predict the remaining entries to make suitable recommendations to the general public to protect in opposition to utmost circumstances by cautioning the people and initiate action accordingly. The figure below shows the sample arrangement of PM2.5 data using the Matrix Completion-based Preprocessing model.



| | | | | |
|---|---|---|---|---|
| | −1 | | | |
| | | 1 | | |
| 1 | 1 | −1 | −1 | 1 |
| −1 | | | | 1 |
| | −1 | | | |

(a) Observed incomplete PM2.5 concentration data matrix

| | | | | |
|---|---|---|---|---|
| 1 | −1 | 1 | 1 | −1 |
| 1 | 1 | 1 | −1 | −1 |
| 1 | 1 | −1 | −1 | 1 |
| −1 | 1 | −1 | 1 | 1 |
| 1 | −1 | 1 | 1 | −1 |

(b) Complete PM2.5 concentration data matrix result (i.e., PM2.5 matrix completion)

**Figure 3 Sample PM2.5 Matrix Completion-based Preprocessing**

We formulate the observed matrix for air quality monitoring parameters below with the above arrangement (i.e., as provided in the input layer).

$$M_{ij} = \begin{bmatrix} F_{11} & F_{12} & \dots & F_{1n} \\ F_{21} & F_{22} & \dots & F_{2n} \\ \dots & \dots & \dots & \dots \\ F_{m1} & F_{m2} & \dots & F_{mn} \end{bmatrix} \quad (2)$$

From the above equation (2), '$M_{ij}$' denotes the observed matrix for the corresponding features. '$F_{mn}$', with '$i = 1,2, \dots, m$' representing the number of features and '$j = 1,2, \dots, n$' denoting the number of test samples, respectively. Then, w the aid of the observed matrix, filling the missing entries has modelled low-ran matrix completion.

$$Min\ Rank\ (P_{ij}) \quad (3)$$

$$Subject\ to\ P_{ij} = M_{ij}\ \forall\ i, j \in E \quad (4)$$

From the above formulation, as in (3) and (4), the objective remains to identify the lowest rank matrix. '$P_{ij}$' which matches the observed matrix. '$M_{ij}$' for all entries in the set '$E$' respectively.

### 3.4 Haversine Ratio of Variances-based Feature Selection model

Upon successful feature engineering by handling missing data for different concentrations, feature selection from the overall 15 features is made in the second hidden layer or hidden layer 2. With feature selection, the dimensionality of data is reduced and removes the existence of irrelevancy. However, the objective of the work

remains to predict air pollutants with the aid of numerous meteorological attributes. The meteorological environment can influence air pollutant concentration via aggregated reactions (i.e., between factors such as temperature, humidity, pressure, combined wind direction, combined wind speed, hourly precipitation, and cumulated precipitation). Therefore, this work retains all features of meteorological conditions in the dataset.

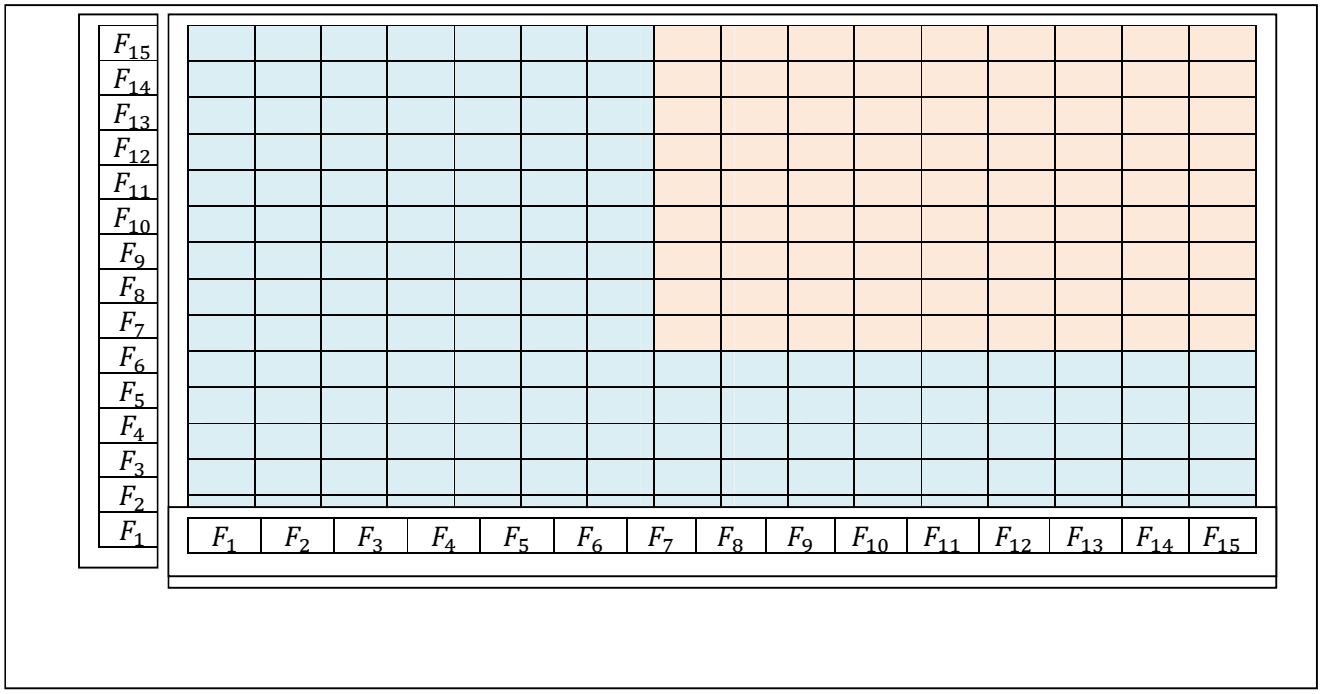



**Figure 4 Ratio of Variances Map of Particulate Matter**

All the significant features have utilised the Haversine Ratio of Variance-based Feature Selection model. To exhibit the existence of irrelevancy between elements has generated a Ratio of Variance Map of Particulate Matter. The figure shows the sample Ratio of Variances Map of Particulate Matter of fifteen different features. After careful consideration, feature selection has eliminated the less significant features that exhibit a high ratio of variances (variances between independent elements).

From the above figure, two coloured regions are observed (i.e., orange and blue). The orange-coloured areas denote the low ratio of variances and are considered relevant features. On the other hand, the blue-coloured regions represent a high percentage of conflict and the considered non-relevant features. Therefore, to be more highly correlated has said to the near thins than distant things. In other words, to be more specific, the influence of the adjacent city on the monitoring city is said to be greater than that of the remote town.

To illustrate the spatiality of the PM2.5 concentration series, using the Haversine function based on the latitude and longitude of each city has evaluated the distance between any two cities. Then the ratio of variances for the corresponding PM2.5 concentration series is evolved. Mathematically this is stated as given below.

$$Haversin\left(\frac{d}{Rad}\right) = Haversin\left(Lat_i - Lat_j\right) + Cos\left(Lat_j\right) * Haversin\left(\Delta\alpha\right) \quad (5)$$

$$Haversin\left(\theta\right) = Sin^2\left(\frac{\theta}{2}\right) + Cos\left(Lat_j\right) \quad (6)$$

From the above equations (5) and (6), the '$Haversin\ (.)$' function denotes the distance between cities, '$Rad$' referring to the radius of the earth, '$Lat_i$' and '$Lat_j$', the latitude between two cities and '$\Delta\alpha$' referring to the distance between two points for the corresponding cities, respectively. The given below has formulated the variance ratio with the estimated distance mathematically (i.e., between class or between the cities and within the category or cities).

$$F(\alpha) = \frac{S_B^2(\alpha)}{S_W^2(\alpha)} \quad (7)$$

From the above equation (7), the ratio of variances '$F(\alpha)$' are obtained based on the sample variance between cities '$S_B^2(\alpha)$' and sample variance within cities '$S_W^2(\alpha)$' respectively.

$$S_B^2(\alpha) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{F_{ij}(\alpha)}{N} \quad (8)$$

$$S_W^2(\alpha) = \sum_{i=1}^{m}\sum_{j=1}^{n}F_{ij}(\alpha) - \left[\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{F_{ij}(\alpha)}{N}\right] \quad (9)$$

From the above equations (8) and (9), '$F_{ij}(\alpha)$' represents the '$\alpha$' feature in the '$j-th$' sample in the '$i-th$' group concerning the number of samples '$N$'. Finally, the significant features selected (PM, DEWP, TEMP, HUMI, Pres, Cbwd, lws, precipitation and lprec) has given below.

$$FS = Haversin\left(\frac{d}{Rad}\right).Haversin\left(\theta\right).F(\alpha) \quad (10)$$

### 3.5 Spatiotemporal data extraction

With the obtained feature selection, the third hidden layer contains two types of data, namely spatial data and temporal data for each of the selected features utilised to acquire the arbitrary spatiotemporal data variances. The material data and spatial data are obtained in the form of a temporal matrix and spatial matrix separately, as given below.

$$TM = P_{TM}.\sigma\left[(RF)^T X_1.X_2(X_3 RF) + C_{TM}\right] \quad (11)$$

$$SM = P_{SM}.\sigma[RF'Y_1.Y_2(Y_3 RF') + C_{SM}] \quad (12)$$

$$STM = [TM][SM] \quad (13)$$

From the above equations (11) and (12), '$P_{TM}$', '$X_1$', '$X_2$', '$X_3$' and '$C_{TM}$' forms the learning parameters for the temporal matrix of the relevant features '$RF$'. In a similar manner, '$P_{SM}$', '$Y_1$', '$Y_2$', '$Y_3$' and '$C_{SM}$' forms the learning parameters for the spatial matrix of the relevant features '$RF$' respectively.

### 3.6 Output layer

Finally, the output layer consists of temporal and fully connected layers. The temporal layer combines the temporal result output by spatiotemporal matrix value to obtain '$STM$' and then gets the final prediction result via the fully connected layer. Given below has been formulated mathematically.

$$Y(AQI) = (STM)W + b \quad (14)$$

From the above equation (14), the output in the output layer '$Y$' is obtained via spatiotemporal matrix value '$STM$', the learnable parameters, '$W$' and '$b$', respectively. The table given below lists the AQI values and corresponding air quality classification.

**Table-2 AQI value and Air quality level**

| ***S.No*** | ***AQI Value*** | ***Air Quality Level*** |
|---|---|---|
| **1** | $0-50$ | *Excellent* |
| **2** | $51-100$ | *Good* |
| **3** | $101-150$ | *Light Pollution* |
| **4** | $151-200$ | *Moderate Pollution* |
| **5** | $201-300$ | *Heavy Pollution* |
| **6** | $>300$ | *Serious Pollution* |

With the aid of the above AQI value and the air quality level, the results obtained from the output unit in the output layer has measured and provided the prediction results accordingly. The pseudo-code representation of the Deep Neural Matrix and Haversine Ratio-based AQP algorithm is below.

| |
|---|
| **Input**: Dataset '$DS$', Features '$F = F_1, F_2, \ldots, F_n$', data '$D = D_1, D_2, \ldots, D_n$', Sample '$Sample$' |
| **Output**: Precise and accurate air quality prediction for a sustainable environment |

1: **Initialise** '$m$', '$n$', '$t$', number of samples '$N$', latitude between two cities '$Lat_i$' and '$Lat_j$', '$Rad = 6371$', learning parameters. '$P_{TM}$', '$X_1$', '$X_2$', '$X_3$' and '$C_{TM}$' for temporal matrix, learning parameters. '$P_{SM}$', '$Y_1$', '$Y_2$', '$Y_3$' and '$C_{SM}$' for spatial matrix, learnable parameters, '$W$' and '$b$'

2: **Begin**

**//Input layer**

3: **For** each Dataset '$DS$' with Features '$F$' and data '$D$'

4: Formulate air quality monitoring parameters as given in (1)

5: **End for**

**//Hidden layer – 1 [preprocessing]**

6: **For** each Dataset '$DS$' with Features '$F$' and data '$D$'

7: Formulate the observed matrix as given in (2)

8: Formulate low-rank matrix completion as given in (3) and (4)

9: **Return** pre-processed data '$PD$'

10: **End for**

**//Hidden layer – 2 [feature selection]**

11: **For** each Dataset '$DS$' with Features '$F$' and pre-processed data '$PD$'

12: Evaluate the distance between cities as given in (5) and (6)

13: Estimate ratio of variances as given in (7)

14: **Return** features selected '$FS$'

15: **End for**

**//Hidden layer – 3 [temporal and spatial matrix generation]**

16: **For** each Dataset '$DS$' with Features '$F$' and features selected '$FS$'

17: Evaluate the temporal matrix as given in (11)

18: Evaluate spatial matrix as given in (12)

19: Evaluate the spatiotemporal data matrix as given in (13)

20: **Return** spatiotemporal data matrix '$STM$'

21: **End for**

**//Output layer [Air quality index analysis results]**

22: **For** each Dataset '$DS$' with Features '$F$' and features selected '$FS$' and spatiotemporal data matrix '$STM$'

23: Obtain the output '$Y\ (AQI)$'

24: **If** '$AQI\ lies\ between$ 0 & 50'

25: **Then** '$Air\ Quality\ Level = Excellent$'

26: **End if**

27: **If** '$AQI\ lies\ between$ 51 & 100'

28: **Then** '$Air\ Quality\ Level = Good$ '

29: **End if**

30: **If** '$AQI\ lies\ between$ 101 & 150'

31: **Then** '$Air\ Quality\ Level = Light\ pollution$'

32: **End if**

33: **If** '$AQI\ lies\ between$ 151 & 200'

34: **Then** '$Air\ Quality\ Level = Moderate\ pollution$'

35: **End if**

36: **If** '$AQI\ lies\ between$ 201 & 300'

37: **Then** '$Air\ Quality\ Level = Heavy\ pollution$'

38: **End if**

39: **If** '$AQI > 301$'

40: **Then** '$Air\ Quality\ Level = Serious\ pollution$'

41: **End if**

42: **End for**

42: **End**

**Algorithm 1 Deep Neural Matrix and Haversine Ratio-based AQP**

As given in the above Deep Neural Matrix and Haversine Ratio-based AQP, the objective remains to predict the air quality with maximum accuracy and minimum error. A deep neural network-based technique using Matrix Completion for this objective has proposed the Haversine Spatiotemporal function. First, input in the input layer obtains the PM2.5 Data of five Chinese cities. Next, in the first hidden layer, missing data are handled by employing

Matrix Completion-based Preprocessing model. Then, in the second hidden layer, dimensionality reduced features are selected using the Haversine Ratio of Variances-based Feature Selection model. Then, the third hidden layer extracts comparable spatiotemporal data for the relevant chosen features. Finally, output in the output layer through the AQI value provides the air quality prediction results.

## 4. Experimental setup

In this section, the performance of air quality prediction for the sustainable environment is called Deep Neural Network-based Matrix Completion, and Python has performed the Haversine Spatiotemporal (DNN-MCHS). To measure the DNN-MCHS method, air quality data in PM2.5 Data of five Chinese cities (measurements for Shenyang, Chengdu, Beijing, Guangzhou and Shanghai) – https://www.kaggle.com/uciml/pm25-data-for-five-chinese-cities has used. First, images of five Chinese cities has provided.

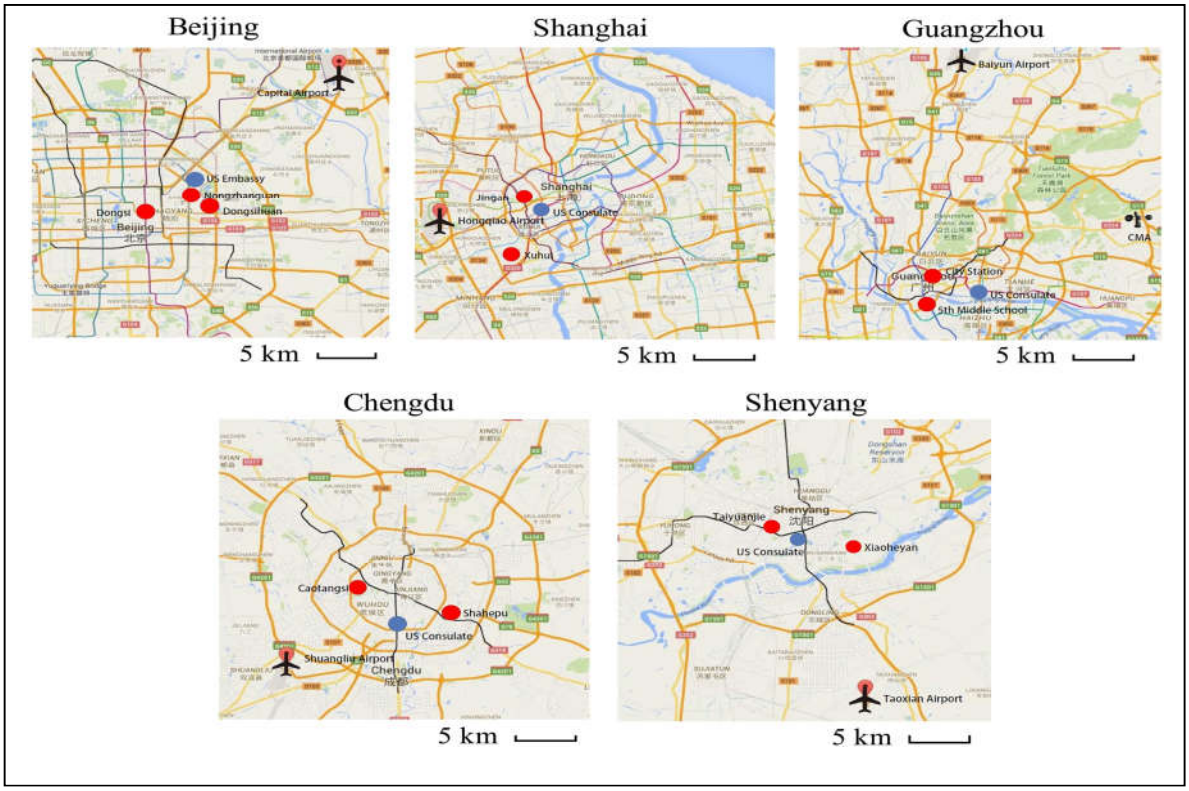



**Figure-5 Map of five Chinese cities**

Experiments on air quality prediction time, accuracy, and root mean square error to distinct air quality samples have followed. Finally, the existing methods have compared Spatio-temporal prediction using a support vector machine [1] and Integrated dual LSTM [2] for a simulation of 10 runs.

### 4.1 Case 1: Air quality prediction time

The first and foremost metric in analysing or predicting the air quality for a sustainable environment for big data is the air quality prediction time. This metric is paramount because the earlier the prediction regarding air quality, the higher the possibility of disease spreading to the public and vice versa. Given below has started the air quality prediction time mathematically.

$$AQP_{time} = \sum_{i=1}^{n} Sample_i * Time\ [AQP] \quad (15)$$

From the above equation (15), the air quality prediction time. '$AQP_{time}$' is measured based on the air quality samples provided as input. '$Sample_i$' and the time consumed in predicting the same '$Time\ [AQP]$'. It has measured in terms of the millisecond (ms). Table-3, given below, shows the performance analysis of the proposed air pollution monitor and controlling method, DNN-MCHS, concerning the air quality prediction and forecasting time. The five Chinese cities' air quality data obtained from PM2.5 have applied the proposed method.

The five Chinese cities' air quality data obtained from PM2.5 has applied this proposed method. In this work, other states of the art methods, including Spatio-temporal prediction using a support vector machine [1] and Integrated dual LSTM [2], has compared the proposed air quality prediction and simulation results.

**Table-3 Tabulation for air quality prediction time using DNN-MCHS, Spatio-temporal prediction using support vector machine [1] and Integrated dual LSTM [2]**

| Air quality samples | Air quality prediction time (ms) | | |
|---|---|---|---|
| | DNN-MCHS | Spatio-temporal prediction using support vector machine | Integrated dual LSTM |
| **5000** | 1750 | 2500 | 3750 |
| **10000** | 2035 | 2815 | 4025 |

| | | | |
|---|---|---|---|
| **15000** | 2415 | 3325 | 4585 |
| **20000** | 2835 | 3585 | 5155 |
| **25000** | 3045 | 4025 | 5845 |
| **30000** | 3355 | 4815 | 6135 |
| **35000** | 4125 | 5845 | 6835 |
| **40000** | 5345 | 6535 | 7535 |
| **45000** | 5825 | 6915 | 8035 |
| **50000** | 6035 | 7325 | 8325 |

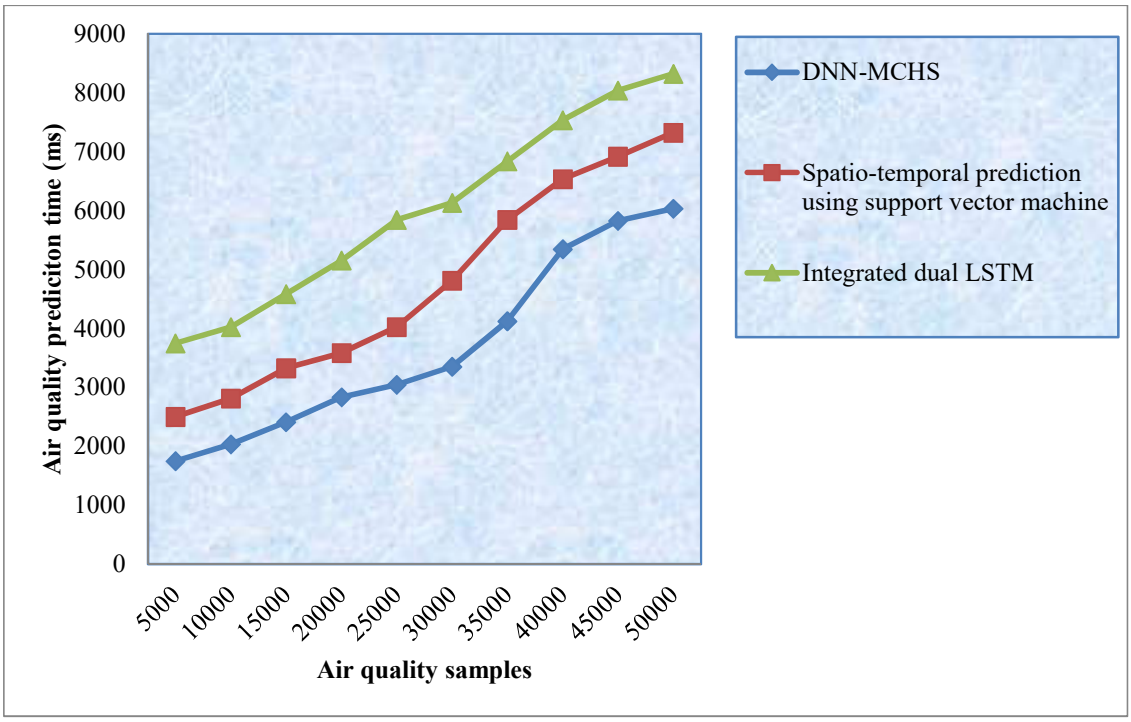



**Figure-6 comparison of air quality prediction time for various air quality samples**

Figure-6 above illustrates the performance metric of air quality prediction time for varying air quality samples obtained from different Chinese cities, i.e., Beijing and Shanghai, in the range of 5000 to 50000 collected at different time intervals. The figure shows that the air quality prediction time increases with the number of air quality samples. Increasing the number of air quality samples increases the number of samples involved in pre-processing, which also

causes an increase in the air quality prediction time. However, with '5000' air quality samples were considered for simulation. The time consumed in air quality prediction for single sample data being '$0.35ms$' using DNN-MCHS, the overall air pollution forecasting time had observed to be '$1750ms$', time consumed in air pollution forecasting being '$0.50ms$' using [1] and '$0.75ms$' using [2], the overall air quality prediction time had observed to be '$2500ms$' and '$3750ms$'. From the result, the air quality prediction time via pre-processing was minimum using DNN-MCHS upon comparison with [1] and [2]. The improvement was due to the application of matrix Completion-based pre-processing. By applying this model, the missing data is handled effectively via low-rank matrix completion for different concentrations. Also, in the observed and lowest rank matrix, two distinct factors were considered based on the respective features irrespective of all samples. With this, the air pollution forecasting time using the DNN-MCHS method was reduced by 24% compared to [1] and 41% compared to [2].

### 4.2 Case 2: Air quality prediction accuracy

The second metric of significance is the air quality prediction accuracy. In other words, the method's efficiency can be analysed or validated by measuring the accuracy rate. As a result, better accuracy is more efficient air quality analyses are made, and hence safeguards can be made to the public against hazardous elements. Given below has stated the air quality prediction accuracy mathematically.

$$AQP_{acc} = \sum_{i=1}^{n} \frac{Sample_{ap}}{Sample_i} * 100 \quad (16)$$

From the above equation (16), the air quality prediction accuracy, '$AQP_{acc}$' is made through the samples provided for simulation, '$Sample_i$' and the samples accurately predicted, '$Sample_{ap}$'. It had measured in terms of percentage (%). The traditional methods for air quality prediction and control process utilised air quality data, and Table-4 has tabulated the results

From Table-4, it is evidence that the air quality prediction achieves maximum accuracy using DNN-MCHS upon comparison with [1] and [2].

**Table-4 Tabulation for air quality prediction accuracy using DNN-MCHS, Spatio-temporal prediction using support vector machine [1] and Integrated dual LSTM [2]**

| Air quality samples | Air quality prediction accuracy (%) |
|---|---|

| | DNN-MCHS | Spatio-temporal prediction using support vector machine | Integrated dual LSTM |
|---|---|---|---|
| **5000** | 96.3 | 92.5 | 82.7 |
| **10000** | 95.25 | 91.35 | 81.35 |
| **15000** | 94.35 | 90.25 | 81 |
| **20000** | 93.25 | 88.15 | 80.35 |
| **25000** | 91.45 | 86.35 | 80 |
| **30000** | 90 | 85.4 | 78.35 |
| **35000** | 90.15 | 85.85 | 79 |
| **40000** | 91.35 | 87.35 | 80.25 |
| **45000** | 94.25 | 89.25 | 81 |
| **50000** | 95 | 90 | 83.45 |

Second, figure-7 illustrates the air quality prediction accuracy for 50000 varying air quality samples. The figure shows that the air quality prediction accuracy decreases with the increase in the air quality samples. This is because while handling missing data. During pre-processing, had retained a small portion of the presence of artefacts. Therefore, to be identified a result, variation inaccuracy is also said. However, simulations conducted with '5000' air quality samples '4815' air quality samples were correctly detected using DNN-MCHS, whereas '4625' and '4135' were correctly detected using [1] and [2]. From these analyses, the air quality prediction accuracy was '96.3%', '92.5%' and '82.7%' using DNN-MCHS, [1] and [2], respectively. This result shows that the air quality prediction accuracy using the DNN-MCHS method is comparatively better than [1] and [2]. The reason behind the improvement is the application of the Haversine Ratio of the Variances-based Feature Selection model. By applying this model, air quality samples predicted accurately that form the basis for early and precise air quality monitoring and control is said to be obtained via the Haversine function based on the latitude and longitude of every city for each pre-processed air sample data. Then, the ratio of variances was obtained based on the sample variance between cities and sample

variance within cities for selecting the best feature for air quality prediction. The air quality prediction accuracy using the DNN-MCHS method improved by 5% compared to [1] and 15% [2].

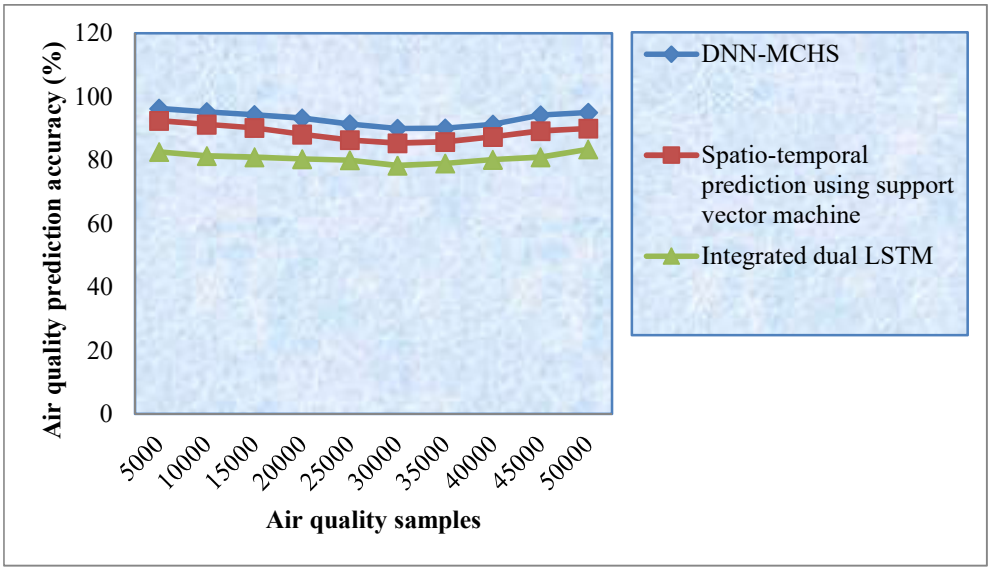



**Figure-7 comparison of air quality prediction accuracy for various air quality samples**

### 4.3 Case 3: Root mean square error

Finally, to assess the effectiveness of the proposed method, root means square error (RMSE) is used as the metric index to evaluate the performance.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \quad (17)$$

From the above equation (17), the root means square error '$RMSE$' is evaluated based on the predicted air pollutant concentration. '$P_i$' observed air pollutant concentration. '$O_i$' concerning the number of test samples '$N$' respectively. Table-5 compares the proposed air quality prediction error rate with other states of the art methods [1] and [2], respectively.

**Table-5 Tabulation for RMSE using DNN-MCHS, Spatio-temporal prediction using support vector machine [1] and Integrated dual LSTM [2]**

| Air quality samples | RMSE | | |
|---|---|---|---|
| | DNN-MCHS | Spatio-temporal prediction using support vector machine | Integrated dual LSTM |
| **5000** | 0.21 | 0.28 | 0.35 |
| **10000** | 0.24 | 0.3 | 0.37 |
| **15000** | 0.27 | 0.31 | 0.39 |
| **20000** | 0.28 | 0.32 | 0.4 |
| **25000** | 0.3 | 0.32 | 0.4 |
| **30000** | 0.31 | 0.34 | 0.41 |
| **35000** | 0.32 | 0.34 | 0.41 |
| **40000** | 0.34 | 0.36 | 0.42 |
| **45000** | 0.34 | 0.37 | 0.42 |
| **50000** | 0.37 | 0.4 | 0.44 |

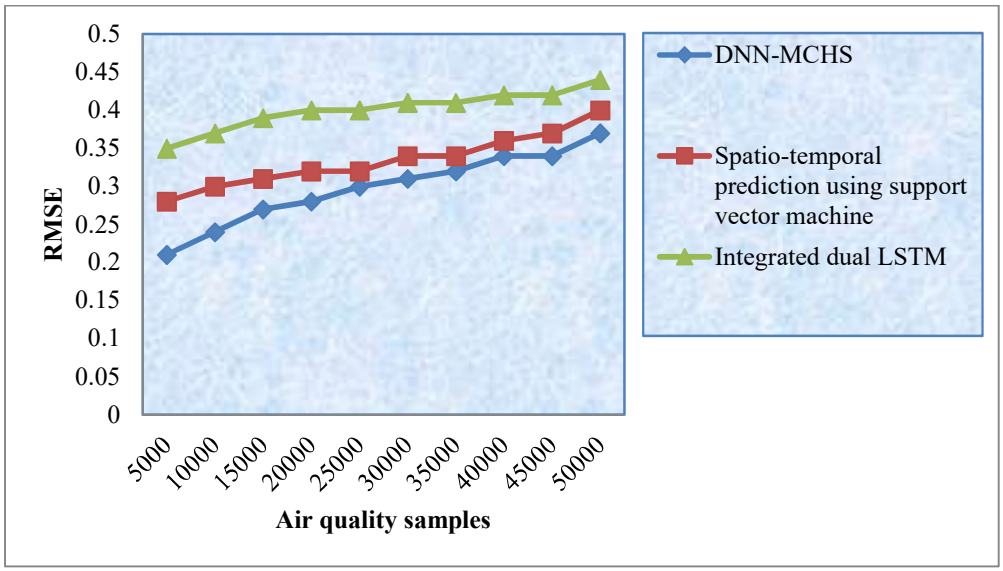


**Figure-8 Comparison of RMSE for various air quality samples**

Finally, figure-8 above illustrates the root mean square error (RMSE) concerning 50000 different air quality samples obtained from five different Chinese cities at different time instances. Using the DNN-MCHS method upon comparison with [1] and [2], found to be comparatively reduced; the error rate increases with the increase in the air quality samples, as the figure shows. The figure shows that the error rate increases with the increase in the air quality samples but was comparatively reduced using the DNN-MCHS method upon comparison with [1] and [2]. However, with simulations performed with 50000 samples, the predicted air pollutant concentration was '95', whereas the observed air pollutant concentration using the DNN-MCHS method was '80' and '75', '70' using [1] and [2], respectively. The overall root means square error rate was 0.21 using the DNN-MCHS method, 0.28 using [1] and 0.35 using [2]. The minimum root means square error rate contribution using the DNN-MCHS method was due to evolving the temporal Matrix and spatial Matrix separately, then combining them to form an overall spatiotemporal matrix. With this, the air quality index was first measured and then, based on learning parameters, multiclass air quality levels have obtained, reducing the root mean square error using the DNN-MCHS method by 11% compared to [1] 17% compared to [2] respectively.

## 5. Conclusion

Deep learning techniques act as a refining service to combat air quality prediction that swiftly increases researchers' and academics' attention. Hypothesising from preceding research work enumerate that there is a necessity to structure an efficient method to provide accurate prediction and control measures to be taken accordingly based on the air pollution via air quality index. To be more specific, there arises a necessity to address the time, accuracy and error involving air quality prediction to prevent hazardous effects on the public. Hence, this work aims to address air quality prediction via deep learning technique to develop a Deep Neural Network-based Matrix Completion and Haversine Spatiotemporal (DNN-MCHS) method that ensures prediction in an accurate and timely manner with a minimum error rate. The deep neural learning framework has designed the entire framework. The entire framework was designed based on the deep neural learning framework. First, a Matrix Completion-based Preprocessing model was designed in the first hidden layer to handle missing data, resulting in pre-processed air samples. Second, the second hidden layer designed the Harversine Ratio of Variance-based Feature Selection model that selects optimal and relevant features. Third, spatiotemporal data are extracted in the third hidden layer to ensure prompt monitoring.

Finally, in the output layer, air quality prediction was made to classify the air quality levels based on AQI. The experimental results show that the DNN-MCHS method can get better results in terms of air quality prediction time, accuracy and error rate using quality data in the Chinese dataset, which fully shows that applying it to increase the prediction accuracy deep and therefore paving the way controlling against hazards caused to the public.

**References**

[1] Chih-Chun Liu, Tzu-Chi Lin, Kuang-Yu Yuan, Pei-Te Chiueh, " Spatio-temporal prediction and factor identification of urban air quality using support vector machine", Urban Climate, Elsevier, Feb 2022 [Spatio-temporal prediction using support vector machine]

[2] Hongqian Chen, Mengxi Guan, Hui Li, "Air Quality Prediction Based on Integrated Dual LSTM Model", IEEE Access, Jul 2021 [Integrated dual LSTM]

[3] Wenjing Mao, Weilin Wang, Limin Jiao, Suli Zhao, Anbao Liu, "Modeling air quality prediction using a deep learning approach: Method optimisation and evaluation", Sustainable Cities and Society, Elsevier, Oct 2020

[4] Yuan Huang, Yuxing Xiang, Ruixiao Zhao, Zhe Cheng, "Air Quality Prediction Using Improved PSO-BP Neural Network", IEEE Access, May 2020

[5] Gao Huang, Chunjiang Ge, Tianyu Xiong, Shiji Song, Le Yang, Baoxian Liu, Wenjun Yin, Cheng Wu, "Large scale air pollution prediction with deep convolutional networks", Information Sciences, Springer, Aug 2021

[6] Xiang Yin, Yanni Han, Hongyu Sun, Zhen Xu, Haibo Yu, Xiaoyu Duan, "Multi-Attention Generative Adversarial Network for Multivariate Time Series Prediction", IEEE Access, Apr 2021

[7] Raquel Espinosa, José Palma, Fernando Jiménez, Joanna Kamińska, Guido Sciavicco, Estrella Lucena-Sánchez, "A time series forecasting based multi-criteria methodology for air quality prediction", Applied Soft Computing, Elsevier, Sep 2021

[8] Uzair Aslam Bhatti, Yuhuan Yan, Mingquan Zhou, Sajid Ali, Aamir Hussain, Huo Qingsong, Zhaoyuan Yu, Linwang Yuan, "Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM**2**:**5**): An SARIMA and Factor Analysis Approach", IEEE Access, Mar 2021

[9] Jingyang Wang, Jiazheng Li, Xiaoxiao Wang, Jue Wang, Min Huang, "Air quality prediction using CT-LSTM", Neural Computing and Applications, Springer, Nov 2020

[10] Luo Zhang, Peng Liu, Lei Zhao, Guizhou Wanga, Wangfeng Zhangd, Jianbo Liua, "Air quality predictions with a semi-supervised bidirectional LSTM neural network", Atmospheric Pollution Research, Elsevier, Sep 2020

[11] Massimo Stafoggia, Tom Bellander, Simone Buccia, Marina Davolia, Kees de Hoogh, Francesca de Donatoa, Claudio Gariazzo, Alexei Lyapustin, Paola Michelozzia, Matteo Renzia, Matteo Scortichinia, Alexandra Shteing, Giovanni Viegih, Itai Kloogg, Joel Schwartzi, "Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model", Environment International, Elsevier, Jan 2019

[12] Xiang Ren, Zhongyuan Mia, Panos G. Georgopoulosa, "Comparison of Machine Learning and Land Use Regression for fine-scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States", Environment International, Elsevier, May 2020

[13] D. Xiao, F. Fanga, J. Zhengc, C.C. Paina,e, I.M. Navond, "Machine learning-based rapid response tools for regional air pollution modeling", Atmospheric Environment, Elsevier, Nov 2018

[14] Hui Liu, Zhihao Long, Zhu Duan, Huipeng Shi, "A New Model Using Multiple Feature Clustering and Neural Networks for Forecasting Hourly PM2.5 Concentrations, and Its Applications in China", Engineering, Elsevier, Jun 2020

[15] Mauro Castelli, Fabiana Martins Clemente, Aleˇs Popovic, Sara Silva, and Leonardo Vanneschi, "A Machine Learning Approach to Predict Air Quality in California", Complexity, Wiley, Aug 2020

[16] Azim Heydari, Meysam Majidi Nezhad, Davide Astiaso Garcia, Farshid Keynia, Livio De Santoli, 'Air pollution forecasting application based on deep learning model and optimisation algorithm", Clean Technologies and Environmental Policy, Springer, Apr 2021

[17] David A. Wood, "Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining", Sustainability Analytics and Modeling, Elsevier, Feb 2022

[18] Pratyush Muthukumar, Emmanuel Cocom, Kabir Nagrecha, Dawn Comer, Irene Burga, Jeremy Taub, Chisato Fukuda Calvert, Jeanne Holm, Mohammad Pourhomayoun, "Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data", Air Quality, Atmosphere and Health, Springer, Nov 2021

[19] Abdellatif Bekkar, Badr Hssina, Samira Douzi2and Khadija Douzi, "Air-pollution prediction in smart city, deep learning approach", Journal of Big Data, Springer, Oct 2021

[20] R. Janarthanan, P. Partheeban, K. Somasundaram, P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city", Sustainable Cities and Society, Elsevier, Jan 2021

**Authors Contribution:**

Two authors prepared this article. Sathish Kumar Sekar performed materials preparation, data collection, coding and analysis. The first draft of the manuscript was written and evaluated by Dr Pushpa Vinu Amalraj. Both the two authors read and approved the final manuscript.

**Declaration of Competing Interest:**

A self-serving stake in the research result will be a promotion in my career.

**Availability of data and material:**

The datasets generated and analysed during the current study are available in the "PM2.5 Data of Five Chinese Cities" repository, https://www.kaggle.com/datasets/uciml/pm25-data-for-five-chinese-cities

**Ethics approval and consent to participant:**

Not Applicable

**Approval for animal experiments:**

Not Applicable.

**Approval for human experiments:**

Not Applicable

**Consent of publication:**

Not Applicable

**Funding:**

Not Applicable