Indian Journal of Engineering & Materials Sciences
Vol. 31, August 2024, pp. 589-600
DOI: 10.56042/ijems.v31i4.2132

Design and synthesis of time series forecasting and deep learning prediction model for air quality index prediction in Indian cities

Dhanalakshmi Subramanian* & Poongothai Marimuthu
Department of Electronics and Communication Engineering,
Coimbatore Institute of Technology, Civil Aerodrome post, Coimbatore 641 014, India

Received 31 May 2023; Accepted: 12 June 2024

Modeling air quality by considering the complexities of randomness in pollutant concentrations and meteorological parameters to forecast real-time Air Quality Index helps mitigate public health risks. The uncertainty of prediction exists due to the high-dimensional nature of predictor variables, making the design of an early warning system highly critical and challenging. With the aim of ensuring accurate AQI forecasts, a statistical time series forecasting model, namely Vector Auto Regression (VAR), and an artificial neural network-based Long Short-Term Memory (LSTM) are integrated to form the Fusion Forecasting Model (FFM), referred to as PrediCasting. The proposed PrediCasting FFM (PCFFM) has been tested with air pollutant and meteorological data collected from the Central Pollution Control Board website for three major Indian cities: Noida, Hyderabad, and Vishakhapatnam. This work provides a detailed analysis of forecasting Air Quality Index by considering the correlation between all factors of air pollutants and meteorological parameters. Results demonstrate that, on average, the proposed PCFFM model has reduced the Root Mean Square Error value by 13.18% and 29.07% for 7-days-ahead and 14-days-ahead forecasts, respectively. Compared to existing models, 7-days-ahead and 14-days-ahead forecasts reduce Mean Absolute Percentage Error (MaPE) on average by 0.187 and 0.222, respectively.

Keywords: Long short-term memory, Mean absolute percentage error, Predicasting fusion forecasting model, Root mean square error, Vector auto regression, Air Quality Index

1 Introduction

Pollution levels in India are the second highest in the world. As measured by the Air Quality Life Index (AQLI), average Indian life expectancy is reduced by 6.3 years due to the persisting air pollution[9]. Thus, in India, it becomes absolutely essential in evaluating and forecasting the air quality which paves way for mitigating and controlling the pollution[4]. The non-linear time series data can be forecasted with high accuracy using leading models and algorithms. Time series forecasting has traditionally been dominated by statistical forecasting models like smoothing based model, moving average process, Holt's trend corrected and Autoregressive model. Deep learning neural networks have also been deployed in recent years for learning complex mappings of a time series data[5]. Some deep learning neural networks include Shift Invariant, Multilayer perceptrons (MLP), advanced recurrent neural networks and hybrid models.

In this paper, PrediCasting Fusion Forecasting Model (PCFFM) is proposed which is the blend of time series forecasting based Vector Auto Regression model and artificial neural network based Long Short-Term Memory to forecast and analyze AQI of three major Indian Cities namely Noida, Hyderabad and Vishakhapatnam. In this proposed PCFFM model, a neural network-based prediction model is synthesized with the existing air quality index forecasting system to improve its performance. Many studies and researches are carried out in forecasting Air Quality Index using a hybrid model that could detect outliers, correct them, and use a heuristic intelligent optimization algorithm to minimize the impact of outliers. Several researchers also contributed to developing a hybrid model to predict Air Quality Index factors. Research Paper[13] focuses on hybrid model which integrates Secondary Decomposition, Sample Entropy, Long Short-Term Memory, Bat algorithm and Least square versions of support-vector machines for the prediction of AQI.

Research paper[11] uses three weight models to find weights to combine RBFNN, EEMD-RBFNN and EEMD-RBFNN-ARIMA in the specific weight ratio to create a Combination Forecasting Model which yields 35.32 and 10.43 for RMSE and MaPE values respectively. Research paper[5] proposes OF-CEEM-

*Corresponding author (Email: dhanalakshmi.s@cit.edu.in)

CF3 model to forecast AQI. According to the results of the three datasets, there exists a deprivation in accuracy with 3.75%, 3.33% and 2.32%.

Research paper[4] proposes ICEEMDAN-ICA-ELM algorithm to model NO2 forecasting system. On an average authors achieve Root Mean Square Error value of 4.49 for modelling NO2 and accuracy is reduced by 4.4%, 7.8%, 6.7% and 12.9% in modelling a forecast system for pollutants namely CO, NO_2, SO_2 and PM_{10} respectively Research paper[10] proposes ICEEMDAN-OS-ELM to study the pollutant concentration of VIS, PM_{10} and $PM_{2.5}$. The proposed model shows a mean absolute error of 0.70, 2.90 and 0.01. The author records a minimal degradation in accuracy when compared to existing algorithms.

Research paper[12] proposes a CEEMDVMD-DE-ELM model for estimating Air Quality Indexes. This research focuses on reducing Root mean square error, Mean absolute percentage error and Mean absolute error for the prediction of AQI. The proposed model shows degradation in its performance in reducing MAE by 68.87%, RMSE by 67.53% and MaPE by 64.91%.

The proposed PCFFM model mainly contains four modules (i) data preprocessing module (ii) data analysis module (iii) proposed PCFFM module and (iv) performance evaluation module. In the data pre-processing module, each predictor variable column of the raw data are tamed to detect for any null values and imputed with the corresponding mean value. The processed data is then applied to a data mining algorithm (DMA) to extract AQI values from the larger set of raw data. In the data analysis module, the data are then normalized to understand the correlation between each and every predictor variable. The probability of receiving varieties of air is studied, to understand the real time pollution episodes of the cities. Augmented Dickey Fuller test is used to determine the stationarity of the time series data. The proposed PCFFM module partitions multivariate data into train and test sets. Using the train sets, the proposed PCFFM model forecasts the 7-day-a-head and 14-day-a-head pollutant concentrations which lead to the estimation of AQI in the future. The proposed PCFFM model is evaluated with Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MaPE) for each of the three Indian cities in the performance evaluation module. Experimental results reveal that the proposed model performed better than existing models in terms of accuracy and minimal error.

2 Materials and Methods

The proposed PrediCasting Fusion Forecasting Model (PCFFM) mainly has four modules namely: Data Preprocessing module, Data Analysis module, Proposed PCFFM Module and performance evaluation module. The detailed steps for forecasting are outlined in this section.

2.1 Data Pre-processing Module

In this research, Mean Replacement method is employed to treat the lost data values. In MR method, the mean value of each predictor variable column is calculated and it is been replaced with the missing values and the data is been groomed for further processing.

2.2 Data mining algorithm

The Data Mining Algorithm (DMA) is developed with the guidelines provided by the Indian government. In India around 8 pollutants are identified and recognized as the major pollutants which affect the Air Quality and thus these pollutants are applied to foresee the Air Quality Index for the input datasets. The Indian Government defined 6 AQI categories to address the level of air quality in each region of the country, as shown in Table 1[9].

Equation 1 is the formula to calculate the AQI using the observed pollutant concentration for all the three major Indian cities[9].

$$AQI = \frac{AQI_{high} - AQI_{low}}{PC_{high} - PC_{low}} (PC - PC_{low}) + AQI\ to\ the\ Suffix\ low_{ion} \quad \ldots (1)$$

Where AQI → Air Quality Index, PC → pollutant concentration, PC_{low} → concentration break point that

Table 1 — Pollutants and AQI group[9]

AQI group (Range)	PM_{10} (1 day)	$PM_{2.5}$ (1 day)	NO_2(1 day)	O_3(8hr)	CO(8hr)	SO_2(1 day)	NH_3(1 day)	Pb(1 day)
Good (0→50)	0→50	0→30	0→40	0→50	0→1	0→40	0→200	0→0.5
Satisfactory (51→100)	51→100	31→60	41→80	51→100	1.1→2.0	41→80	201→400	0.5→1.0
Moderate (101→200)	101→250	61→90	81→180	101→168	2.1→10	81→380	401→800	1.1→2.0
Poor (201→300)	251→350	91→120	181→280	169→208	10→17	381→800	801→1200	2.1→3.0
Very Poor (301→400)	351→430	121→250	281→400	209→748	17→34	801→1600	1200→1800	3.1→3.5
Severe (401→500)	>430	>250	>400	>748	>34	>1600	>1800	>3.5

is ≤PC. PC_{high} → concentration breakpoint that is ≥PC. AQI_{low} → index breakpoint corresponding to PC_{low} and AQI_{high} → index breakpoint corresponding to PC_{high}[9]. Air pollutant and meteorological data for this work is obtained for every hour and thus there will be 24 observations for a single day.

2.3 Data Analysis module

In data analysis module, Data normalization is done using linear scaling method. In this research, a number of meteorological variables, including temperature, humidity, wind direction, wind speed, and pressure, are considered to influence the Air Quality Index. Therefore, it is imperative to define whether there is a correlation between pollutants, meteorological variables, and the Air Quality Index. Detailed correlation analysis improves the proposed PCFFM's forecasting performance. The possibilities of receiving the six distinct categories of air are determined using Kernel Density Estimation. The Augmented Dickey Fuller (ADF) test is used to examine the stationary of the series.

In India the health impacts and its breakpoints of the eight pollutants are given in Table 2[9].

2.4 Existing models

In this section, Vector Auto Regression (VAR) based statistical time series forecasting model is described. In continuation with Long Short-Term Memory (LSTM) is elaborated on how the prediction of Air Quality Index (AQI) is performed on the preprocessed time series data.

2.5 Vector Auto Regression (VAR) Model

In this research work, almost 9 air pollutants like PM_{10}, $PM_{2.5}$, NO, NO_2, NO_x, NH_3, SO_2, CO, Ozone and 5 meteorological variables like Temperature, humidity, Wind direction, Wind speed and pressures are considered for the three cities. Thus, the chosen algorithm must have the capability to capture the relationship between multiple parameters as they change over time. VAR algorithm is optimized to clutch all the 14 input variables to forecast 7 and 14-days-a-head future pollutant concentration and meteorological parameters values[1].

Table 2 — Air Quality Index and its health breakpoints[9]

Good (0→50)	Minimum Impact	Poor (201→300)	Breathing discomfort on prolonged exposure
Satisfactory (51→100)	Minor breathing discomfort for sensitive group	Very Poor (301→400)	Respiratory illness on prolonged exposure
Moderate (101→200)	Breathing discomfort for children and older adults and those with lung, heart disease.	Severe (>401)	Respiratory issues even on healthy people

$$i_t = \sigma(\omega_i[h_{t-1},(Y_{1,t+f_d} + Y_{2,t+f_d} + Y_{3,t+f_d} + Y_{4,t+f_d} + Y_{5,t+f_d} + Y_{6,t+f_d} + Y_{7,t+f_d} + Y_{8,t+f_d} + Y_{9,t+f_d} + Y_{10,t+f_d} + Y_{11,t+f_d} + Y_{12,t+f_d} + Y_{13,t+f_d} + Y_{14,t+f_d})] + b_i)$$

$$\widetilde{c}_t = \tanh(\omega_c[h_{t-1},(Y_{1,t+f_d} + Y_{2,t+f_d} + Y_{3,t+f_d} + Y_{4,t+f_d} + Y_{5,t+f_d} + Y_{6,t+f_d} + Y_{7,t+f_d} + Y_{8,t+f_d} + Y_{9,t+f_d} + Y_{10,t+f_d} + Y_{11,t+f_d} + Y_{12,t+f_d} + Y_{13,t+f_d} + Y_{14,t+f_d})] + b_c)$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t \quad \text{... (2)}$$

$$f_t = \sigma(\omega_f[h_{t-1},(Y_{1,t+f_d} + Y_{2,t+f_d} + Y_{3,t+f_d} + Y_{4,t+f_d} + Y_{5,t+f_d} + Y_{6,t+f_d} + Y_{7,t+f_d} + Y_{8,t+f_d} + Y_{9,t+f_d} + Y_{10,t+f_d} + Y_{11,t+f_d} + Y_{12,t+f_d} + Y_{13,t+f_d} + Y_{14,t+f_d})] + b_f)$$

$$o_t = \sigma(\omega_0[h_{t-1},(Y_{1,t+f_d} + Y_{2,t+f_d} + Y_{3,t+f_d} + Y_{4,t+f_d} + Y_{5,t+f_d} + Y_{6,t+f_d} + Y_{7,t+f_d} + Y_{8,t+f_d} + Y_{9,t+f_d} + Y_{10,t+f_d} + Y_{11,t+f_d} + Y_{12,t+f_d} + Y_{13,t+f_d} + Y_{14,t+f_d})] + b_0) \quad \text{.... (3)}$$

$$h_t = o_t * \tanh(c^t)$$

2.6 Long Short-Term Memory (LSTM) Network

In this research, Long Short-Term Memory is optimized based on the number of variables and dataset and it is used to estimate the future values of Air Quality Index. This method is considered to be superior as it results in an extremely high performance output[7].

2.7 Proposed Predicasting Fusion Forecasting Model (PCFFM)

In this research work, statistical forecasting based VAR model is embedded inside the Recurrent Neural Network based LSTM prediction method as a Fusion Forecasting Model. Figure 1 shows the proposed PCFFM model block diagram.

The expression for input gate, cell state and the candidate for cell state are in the amalgamated version of VAR element implanted in the LSTM feature.

Inputs to the proposed PCFFM are optimized to clutch all the 14 predictor variables. In this research, the duration of forecast is considered for 7-days-a-head and 14-days-a-head and the proposed algorithm performs better for both the forecasting pattern and for all the three datasets.

The forget gate expression for the proposed PCFFM model is mentioned in equation 3. In the proposed PCFFM, the input content to the forget gate is the modified version of LSTM forget gate which incorporates the feature of VAR with the duration of forecast added to the current time stamp.

Based on the relevance and importance of data, weights are assigned using tanh and sigmoid function. The expression for output gate is represented in Equation 3.

Where,

$$
\begin{aligned}
Y_{1,t+f_d} = {} & \alpha_1 + \beta_{11,1}Y_{1,t+f_d-1} + \beta_{12,1}Y_{2,t+f_d-1} + \beta_{13,1}Y_{3,t+f_d-1} + \beta_{14,1}Y_{4,t+f_d-1} + \beta_{15,1}Y_{5,t+f_d-1} + \\
& \beta_{16,1}Y_{6,t+f_d-1} + \beta_{17,1}Y_{7,t+f_d-1} + \beta_{18,1}Y_{8,t+f_d-1} + \beta_{19,1}Y_{9,t+f_d-1} + \beta_{110,1}Y_{10,t+f_d-1} + \beta_{111,1}Y_{11,t+f_d-1} + \\
& \beta_{112,1}Y_{12,t+f_d-1} + \beta_{113,1}Y_{13,t+f_d-1} + \beta_{114,1}Y_{14,t+f_d-1} + \beta_{11,2}Y_{1,t+f_d-2} + \beta_{12,2}Y_{2,t+f_d-2} + \\
& \beta_{13,2}Y_{3,t+f_d-2} + \beta_{14,2}Y_{4,t+f_d-2} + \beta_{15,2}Y_{5,t+f_d-2} + \beta_{16,2}Y_{6,t+f_d-2} + \beta_{17,2}Y_{7,t+f_d-2} + \\
& \beta_{18,2}Y_{8,t+f_d-2} + \beta_{19,2}Y_{9,t+f_d-2} + \beta_{110,2}Y_{10,t+f_d-2} + \beta_{111,2}Y_{11,t+f_d-2} + \beta_{112,2}Y_{12,t+f_d-2} + \beta_{113,2}Y_{13,t+f_d-2} + \\
& \beta_{114,2}Y_{14,t+f_d-2} + \beta_{11,3}Y_{1,t+f_d-3} + \beta_{12,3}Y_{2,t+f_d-3} + \beta_{13,3}Y_{3,t+f_d-3} + \beta_{14,3}Y_{4,t+f_d-3} + \\
& \beta_{15,3}Y_{5,t+f_d-3} + \beta_{16,3}Y_{6,t+f_d-3} + \beta_{17,3}Y_{7,t+f_d-3} + \beta_{18,3}Y_{8,t+f_d-3} + \beta_{19,3}Y_{9,t+f_d-3} + \\
& \beta_{110,3}Y_{10,t-3} + \beta_{111,3}Y_{11,t-3} + \beta_{112,3}Y_{12,t-3} + \beta_{113,3}Y_{13,t-3} + \beta_{114,3}Y_{14,t-3} + \beta_{11,4}Y_{1,t-4} + \\
& \beta_{12,4}Y_{2,t+f_d-4} + \beta_{13,4}Y_{3,t+f_d-4} + \beta_{14,4}Y_{4,t+f_d-4} + \beta_{15,4}Y_{5,t+f_d-4} + \beta_{16,4}Y_{6,t+f_d-4} + \\
& \beta_{17,4}Y_{7,t+f_d-4} + \beta_{18,4}Y_{8,t+f_d-4} + \beta_{19,4}Y_{9,t+f_d-4} + \beta_{110,4}Y_{10,t+f_d-4} + \beta_{111,4}Y_{11,t+f_d-4} + \\
& \beta_{112,4}Y_{12,t+f_d-4} + \beta_{113,4}Y_{13,t+f_d-4} + \beta_{114,4}Y_{14,t+f_d-4} + \beta_{11,5}Y_{1,t+f_d-5} + \beta_{12,5}Y_{2,t+f_d-5} + \beta_{13,5}Y_{3,t+f_d-5} + \\
& \beta_{14,5}Y_{4,t+f_d-5} + \beta_{15,5}Y_{5,t+f_d-5} + \beta_{16,5}Y_{6,t+f_d-5} + \beta_{17,5}Y_{7,t+f_d-5} + \beta_{18,5}Y_{8,t+f_d-5} + \\
& \beta_{19,5}Y_{9,t+f_d-5} + \beta_{110,5}Y_{10,t+f_d-5} + \beta_{111,5}Y_{11,t+f_d-5} + \beta_{112,5}Y_{12,t+f_d-5} + \beta_{113,5}Y_{13,t+f_d-5} + \\
& \beta_{114,5}Y_{14,t+f_d-5} + \beta_{11,6}Y_{1,t+f_d-6} + \beta_{12,6}Y_{2,t+f_d-6} + \beta_{13,6}Y_{3,t+f_d-6} + \beta_{14,6}Y_{4,t+f_d-6} + \beta_{15,6}Y_{5,t+f_d-6} + \\
& \beta_{16,6}Y_{6,t+f_d-6} + \beta_{17,6}Y_{7,t+f_d-6} + \beta_{18,6}Y_{8,t+f_d-6} + \beta_{19,6}Y_{9,t+f_d-6} + \beta_{110,6}Y_{10,t+f_d-6} + \\
& \beta_{111,6}Y_{11,t+f_d-6} + \beta_{112,6}Y_{12,t+f_d-6} + \beta_{113,6}Y_{13,t+f_d-6} + \beta_{114,6}Y_{14,t+f_d-6} + \beta_{11,7}Y_{1,t+f_d-7} + \\
& \beta_{12,7}Y_{2,t+f_d-7} + \beta_{13,7}Y_{3,t+f_d-7} + \beta_{14,7}Y_{4,t+f_d-7} + \beta_{15,7}Y_{5,t+f_d-7} + \beta_{16,7}Y_{6,t+f_d-7} + \beta_{17,7}Y_{7,t+f_d-7} + \\
& \beta_{18,7}Y_{8,t+f_d-7} + \beta_{19,7}Y_{9,t+f_d-7} + \beta_{110,7}Y_{10,t+f_d-7} + \beta_{111,7}Y_{11,t+f_d-7} + \beta_{112,7}Y_{12,t+f_d-7} + \\
& \beta_{113,7}Y_{13,t+f_d-7} + \beta_{114,7}Y_{14,t+f_d-7} + \beta_{11,8}Y_{1,t+f_d-8} + \beta_{12,8}Y_{2,t+f_d-8} + \beta_{13,8}Y_{3,t+f_d-8} +
\end{aligned}
$$

$$\beta_{14,8}Y_{4,t+f_d-8} + \beta_{15,8}Y_{5,t+f_d-8} + \beta_{16,8}Y_{6,t+f_d-8} + \beta_{17,8}Y_{7,t+f_d-8} + \beta_{18,8}Y_{8,t+f_d-8} + \beta_{19,8}Y_{9,t+f_d-8} +$$
$$\beta_{110,8}Y_{10,t+f_d-8} + \beta_{111,8}Y_{11,t+f_d-8} + \beta_{112,8}Y_{12,t+f_d-8} + \beta_{113,8}Y_{13,t+f_d-8} + \beta_{114,8}Y_{14,t+f_d-8} +$$
$$\beta_{11,9}Y_{1,t+f_d-9} + \beta_{12,9}Y_{2,t+f_d-9} + \beta_{13,9}Y_{3,t+f_d-9} + \beta_{14,9}Y_{4,t+f_d-9} + \beta_{15,9}Y_{5,t+f_d-9} +$$
$$\beta_{16,9}Y_{6,t+f_d-9} + \beta_{17,9}Y_{7,t+f_d-9} + \beta_{18,9}Y_{8,t+f_d-9} + \beta_{19,9}Y_{9,t+f_d-9} + \beta_{110,9}Y_{10,t+f_d-9} + \beta_{111,9}Y_{11,t+f_d-9} +$$
$$\beta_{112,9}Y_{12,t+f_d-9} + \beta_{113,9}Y_{13,t+f_d-9} + \beta_{114,9}Y_{14,t+f_d-9} + \beta_{11,10}Y_{1,t+f_d-10} +$$
$$\beta_{12,10}Y_{2,t+f_d-10} + \beta_{13,10}Y_{3,t+f_d-10} + \beta_{14,10}Y_{4,t+f_d-10} + \beta_{15,10}Y_{5,t+f_d-10} +$$
$$\beta_{16,10}Y_{6,t+f_d-10} + \beta_{17,10}Y_{7,t+f_d-10} + \beta_{18,10}Y_{8,t+f_d-10} + \beta_{19,10}Y_{9,t+f_d-10} + \beta_{110,10}Y_{10,t+f_d-10} +$$
$$\beta_{111,10}Y_{11,t+f_d-10} + \beta_{112,10}Y_{12,t+f_d-10} + \beta_{113,10}Y_{13,t+f_d-10} + \beta_{114,10}Y_{14,t+f_d-10} +$$
$$\beta_{11,11}Y_{1,t+f_d-11} + \beta_{12,11}Y_{2,t+f_d-11} + \beta_{13,11}Y_{3,t+f_d-11} + \beta_{14,11}Y_{4,t+f_d-11} + \beta_{15,11}Y_{5,t+f_d-11} +$$
$$\beta_{16,11}Y_{6,t+f_d-11} + \beta_{17,11}Y_{7,t+f_d-11} + \beta_{18,11}Y_{8,t+f_d-11} + \beta_{19,11}Y_{9,t+f_d-11} + \beta_{110,11}Y_{10,t+f_d-11} +$$
$$\beta_{111,11}Y_{11,t+f_d-11} + \beta_{112,11}Y_{12,t+f_d-11} + \beta_{113,11}Y_{13,t+f_d-11} + \beta_{114,11}Y_{14,t+f_d-11} +$$
$$\beta_{11,12}Y_{1,t+f_d-12} + \beta_{12,12}Y_{2,t+f_d-12} + \beta_{13,12}Y_{3,t+f_d-12} + \beta_{14,12}Y_{4,t+f_d-12} +$$
$$\beta_{15,12}Y_{5,t+f_d-12} + \beta_{16,12}Y_{6,t+f_d-12} + \beta_{17,12}Y_{7,t+f_d-12} + \beta_{18,12}Y_{8,t+f_d-12} + \beta_{19,12}Y_{9,t+f_d-12} +$$
$$\beta_{110,12}Y_{10,t+f_d-12} + \beta_{111,12}Y_{11,t+f_d-12} + \beta_{112,12}Y_{12,t+f_d-12} + \beta_{113,12}Y_{13,t+f_d-12} + \beta_{114,12}Y_{14,t+f_d-12} +$$
$$\beta_{11,13}Y_{1,t+f_d-13} + \beta_{12,13}Y_{2,t+f_d-13} + \beta_{13,13}Y_{3,t+f_d-13} + \beta_{14,13}Y_{4,t+f_d-13} + \beta_{15,13}Y_{5,t+f_d-13} +$$
$$\beta_{16,13}Y_{6,t+f_d-13} + \beta_{17,13}Y_{7,t+f_d-13} + \beta_{18,13}Y_{8,t+f_d-13} + \beta_{19,13}Y_{9,t+f_d-13} + \beta_{110,13}Y_{10,t+f_d-13} +$$
$$\beta_{111,13}Y_{11,t+f_d-13} + \beta_{112,13}Y_{12,t+f_d-13} + \beta_{113,13}Y_{13,t+f_d-13} + \beta_{114,13}Y_{14,t+f_d-13} + \varepsilon_{1,t+f_d} \quad \text{... (4)}$$

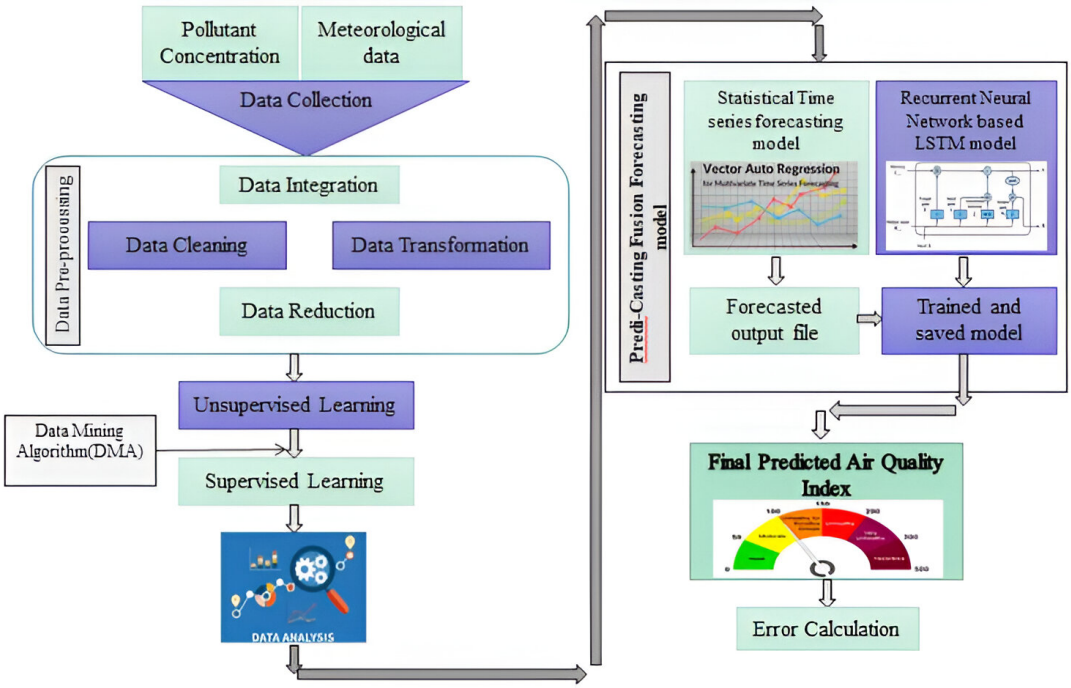

Fig. 1 — Proposed PCFFM model.

Where from equation 2 till 4, f_t→forget gate, i_t→ input gate, o_t→output gate, σ→function of sigmoid, w_x→ weight of gate neuron, h_{t-1}→previous LSTM block output(at time stamp), x_t→current timestamp input, b_x→respective gates(x) biases, c_t→ cell state(memory) at timestamp(t), $\tilde{c}_t$→ cell state candidate at timestamp(t)[14] and f_d is the duration of forecast which is taken as 7-days-a-head and 14-days-a-head. The proposed PCFFM yield the better forecasting results for the multistep prediction with minimal errors for all the three cities.

2.8 Performance evaluation module

The proposed model is evaluated using two performance metrics Root Mean Square Error (RMSE) and the Mean absolute Percentage Error (MaPE) is mentioned in equation 5[6].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{A_i - F_i}{A_i}\right| X100 \quad \text{... (5)}$$

Where,

F_i→forecast values, A_i→observed actual values and N→Sample size

3 Results and Discussion

In this research, A PrediCasting Fusion Forecasting Model (PCFFM) using data preprocessing and Data Mining algorithms was constructed to forecast the Air Quality Index. The experimental process and the analysis of the forecasting results are discussed below.

3.1 Data set description

Based on the close study, this research collects data from three major Indian cities like Noida, Vishakhapatnam and Hyderabad to carry on case studies. The Noida is a satellite city of Delhi and is a part of National Capital Region of India (NCR). Hyderabad is the fourth most populous city and is the capital of Telangana and Vishakhapatnam is a largest and most populous coastal city situated in shore of Bay of Bengal. The air quality in these two cities is comparatively better when compared to Noida.

Due to the diversity and characteristics of these three cities, air pollutant data like $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, Ozone, NH_3, NO, NO_x and meteorological variables like Pressure, Relative humidity, Temperature, Wind direction and Wind speed are collected for all the three cities from Central Pollution Control Board (CPCB), India from 1st January 2018 to 31st December 2021.

For the estimation of the proposed PCFFM model, this research focuses on 7-days-a-head and 14-days-a-head forecasting. Observations of these three datasets are partitioned into training and validation sets.

3.2 Dataset Preprocessing

In this research, Mean Replacement method is employed to treat the lost data values. The preprocessed data are then fed as an input to the Data Mining Algorithm (DMA) to calculate the Air Quality Index for all the three datasets. Thus the unsupervised learning problem has been converted into supervised learning problem. The calculated AQI values for Noida are shown in Fig. 2. In the similar way, AQI values are calculated and plotted for Hyderabad and Vishakhapatnam. In this chapter for simplicity, we have used "Dataset 1" for Noida, "Dataset 2" for Hyderabad and "Dataset 3" for Vishakhapatnam.

In data analysis phase, the pollutant concentration data, meteorological data and AQI data are normalized to determine the correlation between each predictor variable and also to study the AQI by understanding the pollution episodes for the cities under study. The statistical table with details of the data collected for the three cities is mentioned in Table 3, 4 and 5.

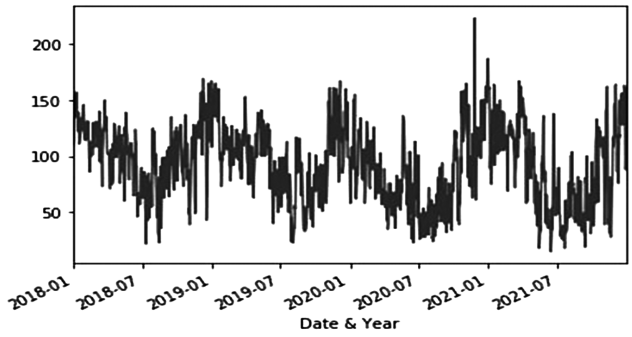

Fig. 2 — AQI Time series distribution for Dataset-1

Table 3 — Dataset-1 Statistics and normalization with year wise mean

Year	Data Count	AQI	AT	BP	CO	NH_3	NO	NO_2	NO_x	Ozone	PM_{10}	$PM_{2.5}$	RH	SO_2	WD	WS
2018	8715	0.3	-0.2	-0.1	0.6	1.1	0.6	0.5	0.7	0.7	0.4	0.27	-0.01	0.30	0.68	0.53
2019	8714	0.1	-0.2	0.1	0.3	0.3	0	0.7	0.2	-0.6	0	0.07	0.20	-0.22	0.04	0.43
2020	8739	-0.3	-0.2	-0.2	-0.2	-0.4	-0.4	-0.6	-0.4	-0.2	-0.4	-0.19	0.34	0.25	-1.10	0.20
2021	8715	-0.1	0.6	0.2	-0.6	-1	-0.3	-0.7	-0.5	0.1	-0.1	-0.15	-0.53	-0.33	0.39	-1.16

The table reveals that around 34,000 data are collected from each city and all the data are normalized to determine the correlation among the pollutants and meteorological variables and the matrix for the same is mentioned in Table 6, 7 and 8 for Dataset-1, 2 and 3 respectively. The correlation matrix reveals that all the pollutant concentration is interlinked and thus there is at least a moderate correlation with the AQI whereas, it is not possible to neglect the meteorological data as well, because most of the meteorological data are in positive correlation with the AQI. Thus, in this research, the proposed

Table 4 — Dataset-2 Statistics and normalization with year wise mean

Year	Data Count	AQI	AT	BP	CO	NH_3	NO	NO_2	NO_x	Ozone	PM_{10}	$PM_{2.5}$	RH	SO_2	WD	WS
2018	8715	0.3	0.5	0.5	0.3	-0.1	-0.1	0	-0.1	0.4	0.3	0.2	-0.3	-0.3	0.1	-0.4
2019	8715	0.2	-0.1	0.1	0.3	0	0	0.3	0.2	0.2	0.2	0.1	0	0	0.2	0.7
2020	8739	-0.3	-0.6	-0.2	-0.5	0.2	0.1	-0.2	-0.1	0.2	-0.3	-0.3	-0.1	0.1	-0.4	-0.2
2021	8725	-0.1	0.2	-0.3	-0.1	-0.1	0	-0.1	0	-0.7	-0.1	-0.1	0.4	0.2	0.1	0

Table 5 — Dataset-3 Statistics and normalization with year wise mean

Year	Data Count	AQI	AT	BP	CO	NH_3	NO	NO_2	NO_x	Ozone	PM_{10}	$PM_{2.5}$	RH	SO_2	WD	WS
2018	8735	0.14	0.23	-0.09	-0.03	-0.14	-0.09	0.20	0.02	0.29	0.17	0.19	-0.39	-0.07	0.65	0.12
2019	8735	0.17	-0.13	0.13	0.43	-0.30	0.00	0.11	0.06	0.11	0.13	0.10	-0.01	0.25	-0.19	0.36
2020	8759	-0.24	-0.37	-0.03	-0.33	-0.46	-0.01	-0.39	-0.21	0.10	-0.25	-0.21	0.14	-0.46	-0.39	-0.16
2021	8737	-0.07	0.27	-0.01	-0.07	0.91	0.10	0.09	0.13	-0.49	-0.04	-0.08	0.27	0.28	-0.07	-0.35

Table 6 — Dataset-1 Correlation matrix-Pollutant and meteorological data

CM_Dataset-1	AQI	AT	BP	CO	NH_3	NO	NO_2	NO_x	Ozone	PM_{10}	$PM_{2.5}$	RH	SO_2	WD	WS
AQI	1.0	-0.4	0.4	0.6	0.3	0.4	0.4	0.5	0.1	0.9	0.9	0.0	0.3	0.2	-0.2
AT	-0.4	1.0	-0.7	-0.5	-0.3	-0.3	-0.4	-0.4	0.1	-0.3	-0.5	-0.5	0.0	0.1	0.0
BP	0.5	-0.7	1.0	0.4	0.1	0.3	0.3	0.4	-0.1	0.4	0.5	0.0	0.1	0.0	-0.3
CO	0.6	-0.5	0.4	1.0	0.5	0.6	0.6	0.7	0.0	0.5	0.6	0.2	0.1	0.1	0.0
NH_3	0.3	-0.3	0.1	0.5	1.0	0.4	0.6	0.5	0.2	0.3	0.3	0.0	0.2	0.2	0.4
NO	0.4	-0.3	0.3	0.6	0.4	1.0	0.6	0.9	0.1	0.4	0.4	0.0	0.2	0.2	-0.1
NO_2	0.4	-0.4	0.3	0.6	0.6	0.6	1.0	0.8	0.0	0.4	0.4	0.1	0.2	0.1	0.1
NOx	0.5	-0.4	0.4	0.7	0.5	0.9	0.8	1.0	0.1	0.5	0.5	0.0	0.2	0.2	0.0
Ozone	0.1	0.1	-0.1	0.0	0.2	0.1	0.0	0.1	1.0	0.1	0.0	-0.2	0.2	0.2	0.1
PM_{10}	0.9	-0.3	0.4	0.5	0.3	0.4	0.5	0.5	0.1	1.0	0.8	-0.2	0.3	0.2	-0.2
$PM_{2.5}$	0.9	-0.5	0.5	0.6	0.3	0.4	0.4	0.5	0.0	0.8	1.0	0.1	0.2	0.1	-0.3
RH	0.0	-0.5	0.1	0.2	0.0	0.0	0.1	0.0	-0.2	-0.2	0.1	1.0	-0.3	-0.2	0.0
SO_2	0.3	0.0	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.3	0.2	-0.3	1.0	0.1	0.0
WD	0.2	0.1	0.0	0.1	0.2	0.2	0.1	0.2	0.2	0.2	0.2	-0.2	0.1	1.0	-0.2
WS	-0.2	0.0	-0.3	0.0	0.4	-0.1	0.1	-0.1	0.1	-0.2	-0.3	0.0	0.0	-0.2	1.0

Table 7 — Dataset-2 Correlation matrix-Pollutant and meteorological data

CM_Dataset-2	AQI	AT	BP	CO	NH_3	NO	NO_2	NO_x	Ozone	PM_{10}	$PM_{2.5}$	RH	SO_2	WD	WS
AQI	1.0	0.1	0.2	0.6	-0.1	0.1	0.6	0.4	0.4	1.0	0.9	-0.5	-0.1	-0.4	-0.1
AT	0.1	1.0	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.2	0.0	-0.3	0.0	0.1	0.1
BP	0.2	0.0	1.0	0.2	-0.1	0.0	0.0	-0.1	0.0	0.1	0.2	-0.1	-0.2	0.0	0.1
CO	0.6	0.1	0.2	1.0	-0.1	0.0	0.6	0.3	0.3	0.6	0.6	-0.3	-0.1	-0.2	-0.2
NH_3	-0.1	0.0	-0.1	-0.1	1.0	0.1	0.1	0.1	-0.1	0.0	-0.1	0.2	0.5	0.2	0.0
NO	0.1	0.0	-0.1	0.0	0.1	1.0	0.0	0.8	0.0	-0.1	-0.1	0.1	0.1	0.0	0.0
NO_2	0.6	0.1	0.0	0.6	0.1	0.0	1.0	0.7	0.4	0.6	0.5	-0.4	0.2	-0.3	-0.2
NOx	0.4	0.0	-0.1	0.3	0.1	0.8	0.7	1.0	0.2	0.3	0.3	-0.1	0.2	-0.1	-0.1
Ozone	0.4	0.1	0.1	0.3	-0.1	-0.1	0.4	0.2	1.0	0.4	0.4	-0.4	0.0	-0.3	-0.2
PM_{10}	1.0	0.2	0.1	0.6	0.0	-0.1	0.6	0.4	0.4	1.0	0.9	-0.5	0.0	-0.3	-0.1
$PM_{2.5}$	0.9	0.0	0.2	0.6	-0.2	-0.1	0.5	0.3	0.4	0.9	1.0	-0.4	-0.1	-0.4	-0.2
RH	-0.1	-0.3	-0.1	-0.3	0.2	0.1	-0.4	-0.1	-0.4	-0.5	-0.4	1.0	0.3	0.4	0.1
SO_2	-0.1	0.0	-0.2	-0.1	0.5	0.1	0.2	0.2	0.1	0.0	-0.2	0.3	1.0	0.2	0.0
WD	-0.4	0.1	0.0	-0.2	0.2	0.0	-0.3	-0.2	-0.3	-0.4	-0.4	0.4	0.2	1.0	0.2
WS	-0.1	0.1	0.1	-0.2	0.0	0.0	-0.2	-0.1	-0.2	-0.1	-0.2	0.1	0.0	0.2	1.0

Table 8 — Dataset-3 Correlation matrix-Pollutant and meteorological data

CM_Dataset-3	AQI	AT	BP	CO	NH_3	NO	NO_2	NO_x	Ozone	PM_{10}	$PM_{2.5}$	RH	SO_2	WD	WS
AQI	1.0	-0.4	0.4	0.6	0.1	0.4	0.6	0.6	0.5	0.9	0.9	-0.3	0.2	-0.2	-0.4
AT	-0.4	1.0	-0.5	-0.4	0.1	-0.3	-0.1	-0.3	-0.5	-0.3	-0.5	0.0	-0.1	0.4	0.3
BP	0.4	-0.5	1.0	0.3	0.1	0.1	0.1	0.1	0.5	0.3	0.4	-0.2	0.1	-0.5	-0.3
CO	0.6	-0.4	0.3	1.0	0.0	0.5	0.5	0.6	0.3	0.6	0.6	-0.1	0.2	-0.1	-0.3
NH_3	0.1	0.1	0.1	0.1	1.0	0.0	0.1	0.1	-0.1	0.2	0.1	0.0	0.1	0.0	-0.1
NO	0.4	-0.3	0.1	0.5	0.0	1.0	0.3	0.9	0.0	0.5	0.4	-0.1	0.2	0.0	-0.4
NO_2	0.6	-0.1	0.1	0.5	0.1	0.3	1.0	0.7	0.1	0.6	0.5	-0.2	0.3	0.2	-0.3
NOx	0.6	-0.3	0.1	0.6	0.1	0.9	0.7	1.0	0.0	0.7	0.5	-0.2	0.3	0.1	-0.4
Ozone	0.5	-0.5	0.5	0.3	-0.1	0.0	0.1	0.0	1.0	0.4	0.6	-0.4	0.0	-0.5	-0.2
PM_{10}	0.9	-0.3	0.3	0.6	0.2	0.5	0.6	0.7	0.4	1.0	0.9	-0.4	0.2	-0.1	-0.4
$PM_{2.5}$	1.0	-0.5	0.4	0.6	0.1	0.4	0.5	0.5	0.6	0.9	1.0	-0.3	0.2	-0.3	-0.4
RH	-0.3	0.0	-0.2	-0.1	-0.1	-0.1	-0.2	-0.2	-0.4	-0.4	-0.3	1.0	-0.1	0.1	0.0
SO_2	0.3	-0.1	0.1	0.2	0.1	0.2	0.3	0.3	0.0	0.2	0.2	-0.1	1.0	0.1	-0.1
WD	-0.2	0.5	-0.5	-0.1	0.0	0.0	0.2	0.1	-0.5	-0.1	-0.3	0.1	0.1	1.0	0.3
WS	-0.4	0.3	-0.3	-0.3	-0.1	-0.4	-0.3	-0.4	-0.2	-0.4	-0.4	0.0	-0.1	0.3	1.0

AQI Category (Range)	Dataset-1	Dataset-2	Dataset-3
Good (0-50)	4%	10%	11%
Satisfactory (51-100)	7%	47%	28%
Moderate (101-200)	25%	40%	52%
Poor (201-300)	31%	0%	5%
Severe (301-400)	20%	0%	0%
Hazardous (401+)	9%	0%	0%

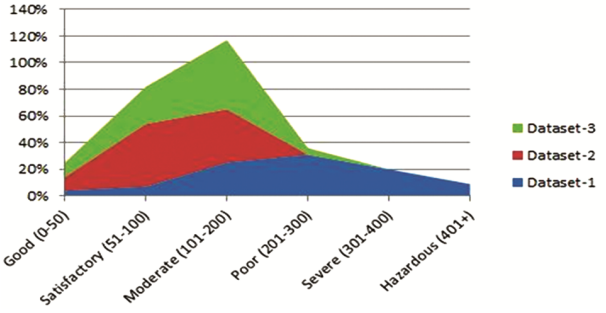

Fig. 3 — Probability statistics for Air Quality Health Index

model uses all 14 predictor variables including AT, BP, CO, NH_3, NO_2, NO_x, NO, Ozone, $PM_{2.5}$, PM_{10}, SO_2, RH, WD and WS for more accurate prediction of Air Quality Index.

The Air Quality Health Index (AQHI) is obtained by finding the probability of receiving Good, Satisfactory, Moderate, Poor, Severe, Hazardous air for all the three Indian Cities for the stipulated time period from January 2018 to December 2021 and it is shown in Fig. 3.

The results reveals that Noida has a very bad air quality as more than 60% of the data falls under the AQI categories with poor, severe and hazardous air quality circumstances, which is an alarming signal for the residents of Noida. Although Hyderabad is the metropolis of Telangana and centre of major technology industry, the Air quality in this city is better when compared to Vishakhapatnam. Vishakhapatnam is a port city and industrial center, the air quality is moderate. The proposed PCFFM model is very versatile which suits 3 unique data patterns and produces results with extremely better accuracy. Finally, all the three dataset underwent stationarity test and the results reveals that all data obeys stationarity property. The sample result of stationarity test conducted for Dataset-1 is mentioned in Table 9.

Table 9 — Stationarity test-Dataset-1

	Pollutant	ADH
0	AQI	Stationarity based on ADH
1	PM_{10}	Stationarity based on ADH
2	$PM_{2.5}$	Stationarity based on ADH
3	CO	Stationarity based on ADH
4	NH_3	Stationarity based on ADH
5	NO	Stationarity based on ADH
6	NO_2	Stationarity based on ADH
7	NO_x	Stationarity based on ADH
8	SO_2	Stationarity based on ADH

3.3 Performance comparison of Predi Casting Fusion Forecasting Model based on Root Mean Square Error and Mean Absolute Percentage Error

In this section, five existing models and PCFFM are compared and analyzed for forecasting results. This research compares the proposed algorithm with the five existing models across three datasets with two different time horizons. A forecasting performance comparison for 7-days-a-head and 14-days-a-head time periods is shown in Fig. 4 (a to f). The proposed PCFFM model was clearly found to be superior than other existing models like KF-AR[2], MA-OWA[3],

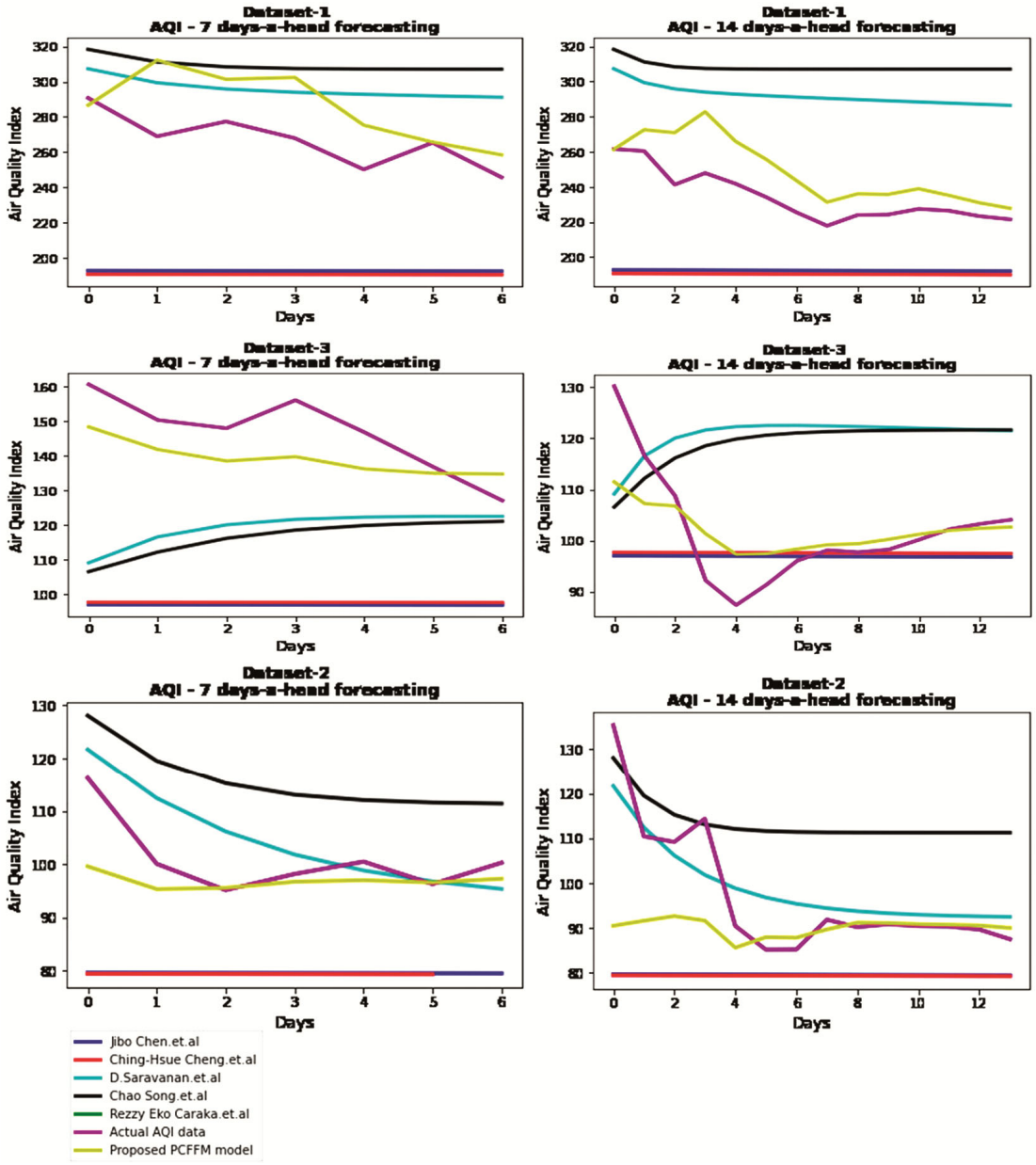

Fig. 4 — (a) 7 days-a-head forecasting- Dataset-1, (b) 14 days-a-head forecasting-Dataset-1, (c) 7 days-a-head forecasting- Dataset-2, (d) 14 days-a-head forecasting- Dataset-2, (e) 7 days-a-head forecasting- Dataset-3, (f) 14 days-a-head forecasting- Dataset-3.

FA-ANN-ARMA[8], EEMD-RBFNN-ARIMA[11], VAR-NN-PSO[1] in terms of its fitting ability which is revealed by its accurate forecasted results closer to the actual value. The PCFFM model has therefore proven to be better in forecasting than the other models. This research observes three diversified and lengthy datasets, showing the model's adaptability and stability for any real-time environment.

In this research work, statistical forecasting model based Vector Auto Regression is embedded inside the Recurrent Neural Network based Long Short-Term Memory prediction method as a Fusion Forecasting Model. Five existing models are compared with the proposed model. In order to demonstrate the predictive ability of the proposed model, the Mean Absolute Percentage Error (MaPE) and Root Mean Square Error (RMSE) forecasting errors are analyzed. Detailed test results can be found in Table 10.

Based on the results, the proposed PCFFM model outperforms the existing five models. The RMSE value of the proposed PCFFM model is 28.7, 4.53 and 12.18 for dataset-1, dataset-2 and dataset-3

Table 10 — Comparison of Proposed PCFFM model with Existing Prediction models

City & Study Site	Designated Predictive model	Reference	Step	RMSE	MaPE
Dataset-1 Noida Sector-125 Greater Noida Expressway	KF-AR	Jibo Chen *et.al*[2]	7	61.39	0.21
			14	58.14	0.21
	MA-OWA	Ching-Hsue Cheng. *et.al*[3]	7	28.9	0.09
			14	45.82	0.18
	FA-ANN-ARMA	D.Saravanan. *et.al*[8]	7	61.09	0.22
			14	74.10	0.31
	EEMD-RBFNN-ARIMA	Chao Song. *et.al*[11]	7	70.72	0.27
			14	86.97	0.37
	VAR-NN-PSO	Rezzy Eko Caraka. *et.al*[1]	7	29.05	1.27
			14	31	0.8
	Proposed PCFFM Model		7	28.7	0.09
			14	24.7	0.1
Dataset-2 Hyderabad Bollaram Industrial Area	KF-AR	Jibo Chen *et.al*[2]	7	36.03	0.29
			14	53.13	0.36
	MA-OWA	Ching-Hsue Cheng. *et.al*[3]	7	18.62	0.12
			14	36.01	0.21
	FA-ANN-ARMA	D.Saravanan. *et.al*[8]	7	14.02	0.10
			14	38.32	0.22
	EEMD-RBFNN-ARIMA	Chao Song. *et.al*[11]	7	13.94	0.03
			14	27.25	0.16
	VAR-NN-PSO	Rezzy Eko Caraka.*et.al*[1]	7	15	0.36
			14	16	0.37
	Proposed PCFFM Model		7	4.53	0.03
			14	10.6	0.07
Dataset-3 Vishakapatnam GVM Corporation	KF-AR	Jibo Chen *et.al*[2]	7	22.03	0.15
			14	54.39	0.28
	MA-OWA	Ching-Hsue Cheng.*et.al*[3]	7	13.47	0.10
			14	38.95	0.20
	FA-ANN-ARMA	D.Saravanan.*et.al*[8]	7	13.72	0.08
			14	35.28	0.16
	EEMD-RBFNN-ARIMA	Chao Song.*et.al*[11]	7	12.91	0.08
			14	35.73	0.16
	VAR-NN-PSO	Rezzy Eko Caraka.*et.al*[1]	7	14	0.4
			14	22	0.4
	Proposed PCFFM Model		7	12.18	0.07
			14	8.08	0.04

respectively for 7-days-a-head time horizon and it is 24.7, 10.6 and 8.08 for 14 days-a-head time horizon. MaPE values of the proposed PCFFM model are 0.09, 0.03, 0.07 for dataset-1, dataset-2 and dataset-3 respectively for 7-days-a-head time horizon and it is 0.1, 0.07, 0.04 for 14-days-a-head time horizon. The application of proposed PrediCasting Fusion Forecasting Model (PCFFM) leads to better forecasting accuracy with minimal error compared to 5 existing models for all the three distinct datasets, thus the proposed model proves its adaptability and versatility for any real time data and environment.

Table 11 and 12 gives percentage of improvement in RMSE and MaPE values for 7-days-a-head and 14-days-a-head forecast for all three datasets by implementing the proposed PCFFM model in comparison with the existing models. The results reveal that the proposed PCFFM (DMA-VAR-LSTM) model has reduced the Root Mean Square Error value by 13.18% and 29.07% for 7-days-a-head and 14-days-a-head forecast respectively. The Mean absolute Percentage Error of 7-days-a-head forecast and 14-days-a-head forecast is reduced on an average by 0.19% and 0.22% respectively for three versatile datasets.

An overall performance comparison between the proposed PCFFM model and the existing model is presented in Table 13. The average RMSE improvements of the proposed model are 13.18% and 29.07% for 7-days-a-head and 14-days-a-head forecast. The average MaPE improvements are 0.32% and 0.27% for 7-days and 14-days-a-head forecast.

Table 11 — Comparison of RMSE obtained by the proposed algorithm with the Existing algorithm

Dataset	Designated Predictive model	Improved RMSE for 7-days-a-head forecast (%)	Improved RMSE for 14- days-a-head forecast (%)
	KF-AR	32.69	33.44
	MA-OWA	0.2	21.12
1	FA-ANN-ARMA	32.39	49.4
	EEMD-RBFNN-ARIMA	42.02	62.27
	VAR-NN-PSO	0.35	6.3
	KF-AR	31.5	42.53
	MA-OWA	14.09	25.41
2	FA-ANN-ARMA	9.49	27.72
	EEMD-RBFNN-ARIMA	9.41	16.65
	VAR-NN-PSO	10.47	5.4
	KF-AR	9.85	46.31
	MA-OWA	1.29	30.87
3	FA-ANN-ARMA	1.54	27.2
	EEMD-RBFNN-ARIMA	0.73	27.65
	VAR-NN-PSO	1.82	13.92
Proposed PCFFM Average Improvement (%)		13.18%	29.07%

Table 12 — Comparison of MaPE obtained by the proposed algorithm with the Existing algorithm

Dataset	Designated Predictive model	Improved MaPE for 7-days-a-head forecast in terms of %	Improved MaPE for 14-days-a-head forecast in terms of %
	KF-AR	0.12	0.11
	MA-OWA	0	0.08
1	FA-ANN-ARMA	0.13	0.21
	EEMD-RBFNN-ARIMA	0.18	0.27
	VAR-NN-PSO	1.18	0.7
	KF-AR	0.26	0.29
	MA-OWA	0.09	0.14
2	FA-ANN-ARMA	0.07	0.15
	EEMD-RBFNN-ARIMA	0	0.09
	VAR-NN-PSO	0.33	0.3
	KF-AR	0.08	0.24
	MA-OWA	0.03	0.16
3	FA-ANN-ARMA	0.01	0.12
	EEMD-RBFNN-ARIMA	0.01	0.12
	VAR-NN-PSO	0.33	0.36
Proposed PCFFM Average Improvement (%)		0.19%	0.22%

Table 13 — Comparative study of existing and proposed PCFFM models

Evaluation Parameter	RMSE		MaPE	
Time-horizon	7-days-a-head	14-days-a-head	7-days-a-head	14-days-a-head
Dataset-1	21.53%	34.51%	0.32%	0.27%
Dataset-2	14.99%	23.54%	0.15%	0.19%
Dataset-3	3.04%	29.19%	0.09%	0.2%
Average Improvement	13.18%	29.07%	0.19%	0.22%

4 Conclusion

The development of an effective and highly accurate forecasting model that serves as an early warning system for air quality is imperative. To monitor the quality of air and to analyze the pollution contamination, the PrediCasting Fusion Forecasting model was presented in this paper. In this work, around 14 different predictor variables which include air pollutant concentration and meteorological parameters are collected for three major Indian cities namely Noida, Hyderabad and Vishakhapatnam. The proposed PCFFM model was built in four modules namely data preprocessing module, data analysis module, proposed PCFFM module and performance evaluation module. In data preprocessing module, the Air Quality Index (AQI) for each dataset is determined using Data

Mining Algorithm. The data is analyzed to figure out the real time pollution episodes of the Indian cities and the data is then applied to the PrediCasting Fusion Forecasting Model which blends the concepts of Vector Auto Regression based statistical forecasting model and the RNN based Long Short-Term Memory prediction method. On three different datasets, the developed PCFFM model was applied and evaluated. Five exisitng models namely KF-AR[2], MA-OWA[3], FA-ANN-ARMA[8], EEMD-RBFNN-ARIMA[11], VAR-NN-PSO[1] are studied in this work. This work concentrates on 2 error measurement criteria namely RMSE and MaPE for two different time horizons (7-days-a-head and 14-days-a-head). The experimental results indicated that the proposed PCFFM (DMA-VAR-LSTM) model has reduced the Root Mean Square Error value by 13.18% and 29.07% for 7-days-a-head and 14-days-a-head forecast respectively. The Mean absolute Percentage Error of 7-days ahead forecast and 14 days-a-head forecast is reduced on an average by 0.19% and 0.22% respectively for three versatile datasets. In our future work, seasonal variations are considered to preprocess the dataset and thereby enhancing the prediction accuracy by reducing the error terms.

Acknowledgement

The data for this work was supported by Uttar Pradesh Pollution Control Board(UPPCB), Telangana State Pollution Control Board(TSPCB) and Andhra Pradesh Pollution Control Board(APPCB) from the website of Central Pollution Control Board(CPCB).

References

1 Caraka, Rezzy Eko, Rung Ching Chen, Toni Toharudin, Bens Pardamean, Hasbi Yasin and Shih Hung Wu, *IEEE access*, 7 (2019) 161654.
2 Chen, Jibo, Keyao Chen, Chen Ding, Guizhi Wang, Qi Liu and Xiaodong Liu*, IEEE access*, 8 (2020) 4265.
3 Cheng, Ching-Hsuen and Sue-Fen Huang, *IEEE International Conference on Systems, Man, and Cybernetics* (2009).
4 Li, Chen and Zhijie Zhu, *Sci Total Environ, Elsevier*, 626 (2018) 1421.
5 Li, Hongmin, Jianzhou Wang and Hufang Yang, *Atmos Pollut Res, Elsevier*, 11 (8) (2020) 1258.
6 Liu, Hui, Guangxi Yan, Zhu Duan and Chao Chen, *Appl Soft Comput, Elsevier*, 102 (106957) (2021) 1.
7 RuiYan, JiaqiangLiao, JieYang, WeiSun, MingyueNong and FeipengLi, *Expert Syst Appl, Elsevier*, 169 (114513) (2021) 1.
8 Saravanan D and Santhosh Kumar K, *Mater Today, Elsevier*, 56 (4) (2022) 1809.
9 Shah, Dipsha Paresh and Dr.Piyushkumar Patel, *Environ Chall, Elsevier*, 5 (100356) (2021) 1.
10 Sharma, Ekta, Ravinesh C. Deo, Ramendra Prasad, Alfio V and Parisi, *Sci Total Environ, Elsevier*, 709 (135934) (2020) 1.
11 Song, Chao A and Xiaoshuang Fu, *J Clean Prod, Elsevier*, 261 (121169) (2020) 1.
12 Wang, Deyun, Shuai Wei, Hongyuan Luo, Chenqiang Yue and Olivier Grunder, *Sci Total Environ, Elsevier*, 580 (2017) 719.
13 Wu, Qunli and Huaxing Lin, *Sci Total Environ, Elsevier*, 683 (2019) 808.
14 Wu, Qunli and Huaxing Lin, *SCS*, 50 (101657) (2019) 1.