DOI: 10.21120/LE/19/2/3

*Acta Geographica Debrecina*
*Landscape & Environment series 19 (2) 2025. 33-53*

# A DEEP ANALYTICS FOR PREDICTION AND FORECASTING OF AIR QUALITY IN CHENNAI

**SASIKALA S[1]*, SHALINI R[2], CHINNAPPARAJ D[3]**

[1, 2, 3]Department of Computer Science IDE, University of Madras, Chennai-5, India
*Email: sasikalarams@gmail.com

Received 3 January 2024, accepted in revised form 10 September 2024


**Abstract**
Air pollution is a global crisis with profound implications for public health and environmental sustainability. In addressing this issue in Chennai, Tamil Nadu, a novel Hadoop-based real-time air pollution prediction system is proposed. This research offers accurate air quality information for specific Chennai regions, aiding decisions and mitigating pollution risks through big data analytics and deep learning for air quality prediction. To expedite air quality prediction, a vast air pollution dataset is rigorously analyzed using a Hadoop-based analytics model. Specific locations in Chennai, including Perungudi, Royapuram, Manali, Alandur, Arumbakkam, Kodungaiyur, and Velachery, are assessed for up-to-date air quality evaluations. The core of the research revolves around implementing deep learning models—Long Short-Term Memory, Convolutional Neural Network, and a hybrid Long Short-Term Memory-Convolutional Neural Network model. These models are trained to forecast AQI for selected areas over the next five years, with the hybrid model emerging as the standout performer, achieving 99% of accuracy rate and mean absolute error, mean square error, root mean square error rates of 7.95, 101.71, 9.65. This high accuracy and low error rates empowers policymakers and environmental agencies to make informed decisions, fostering healthier living conditions in Chennai. The integration of big data analytics and deep learning, promises improved air quality management in urban areas globally, addressing similar environmental challenges and enhancing overall quality of life.

***Keywords:*** air pollution, prediction, forecasting, air quality index, map reduce, deep learning, hybrid model

## 1. Introduction

Air pollution is a consequential environment complication that is attracting increasing consideration globally (Lelieveld et al., 2020). The urban air pollution poses a significant threat to human health (Susanto et al., 2020). The unconstrained discharge of air pollutants is gradually triggering air pollution; human health is severely affected by incessant exposure to polluted air. World Health Organization (WHO) declares ambient air pollution which is typically emitted by industries, power plants, households, vegetation burning and automobiles has the adversative effect on human health, (World Health Organization, 2020). WHO also estimated that elevating levels of pollutant have played a crucial role in causing life threatening disease. Air pollution is assessed through the Air Quality Index (AQI) (en.wikipedia.org/wiki/Air_quality_index),

which quantifies daily pollutant impacts on a scale from 0 to 500. Air quality is influenced by a location's history, and pollutants like particulate matter (PM), nitrogen dioxide (NO2), sulphur dioxide (SO2), carbon monoxide (CO) contribute to levels ranging from Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. Prolonged exposure to high pollution levels poses significant health risks, often linked to a country's rapid economic development. To combat this, research has focused on improving air quality analysis and prediction, shifting towards big data and Artificial Intelligence (AI) techniques for increased accuracy and effectiveness.

Air pollution is a pressing concern, leading to extensive research on air quality and AQI forecasting to combat it effectively. Traditional methods have limitations like low accuracy and complex calculations. Recent advancements have introduced alternatives using big data and AI techniques, with Deep Learning (DL) playing a significant role. Accurate AQI prediction is crucial for public health protection. Research categorizes air quality prediction into three traditional classes: numerical, statistical, and AI forecasting methods. Numerical methods lack accuracy due to incomplete theory, while statistical methods offer limited accuracy in capturing nonlinear patterns. This evolving field aims to harness technology to improve air quality prediction and positively impact human health.

AI-based methods have emerged as a promising alternative to linear models for air quality prediction and forecasting. Research has shown that Artificial Neural Network (ANN) (Seyedeh et al., 2021) outperform linear models because air quality data exhibits clear nonlinear patterns. This work focuses on AI techniques capable of analyzing non-linear data, especially in time series prediction and forecasting. Predicting air quality involves handling extensive input data, which poses a significant challenge in the matter of data analysis and processing rate. The task is to forecast pollutant concentrations based on real-time sensor data and historical trends. Efficient solutions are required for handling large dataset issues, including storing the data and stream processing. While existing AI methods support air quality prediction, they face challenges related to computational and time complexity due to the need for retraining with newly gathered training samples, demanding substantial memory and high-power processors.

The Distributed computing (Hajewski et al., 2020) has emerged as a present-day elucidation to address the substantial memory and computational power demand for training extensive datasets like air pollution. It optimizes processing time by distributing the workload among multiple processors. Hadoop, a scalable framework for big data storage and analysis, incorporates the concept of MapReduce to process massive volumes of data in parallel by storing data in clusters. Recent research has successfully integrated Map Reduce based distributed machine learning algorithms, enhancing accuracy and reducing processing time (Zhai et al., 2014). Simultaneously, DL, a promising machine learning (ML) approach, has gained significant attention in academia and industry, finding successful applications in various fields such as image-classification task, natural language processing, forecasting, object detection, and AI (Liang et al., 2020).

The DL algorithms extract data features in a layered manner, identifying structural patterns without prior knowledge, making them effective for air quality predictions. The study utilizes a Hadoop-based hybrid DL approach, incorporating Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) and Long Short-Term Memory_Convolutional Neural Network (LSTM-CNN), to forecast five years of air quality data with reduced time complexity, addressing the significant concern of AQI prediction and forecasting using big data and AI.

Air quality prediction is advancing with the integration of AI technologies to reduce time complexity, transitioning from traditional

ML methods to advanced DL and big data techniques like LSTM, CNN, Recurrent Neural Network (RNN), and Hadoop Map Reduce. These advanced methods enhance accuracy, particularly for large and nonlinear datasets, improving air quality forecasting. In a study by (Kothandaraman et al., 2022) ML models were applied for predicting PM2.5 and forecast air quality levels. They collected meteorological and PM2.5 data from 2014 to 2019, performed preprocessing to handle null values, extracted meteorological features, and applied various ML methods. The results showed mean absolute error (MAE) of 8.27 and 9.23, root mean-square error (RMSE) of 13.85 and 10.59, mean average percentage error (MAPE) of 0.40 and 0.45, for different models. This research demonstrated effective PM2.5 prediction and air quality forecasting with good performance. (Ghufran et al., 2022) has presented a work on air pollution forecasting with deep learning model called LSTM. The LSTM model finds best hyperparameters by incorporating metaheuristic algorithm called Genetic Algorithm (GA). This hyperparameter tuned model predicts the level of air pollution for the next day using a group of pollutants namely PM10, PM2.5, CO and NOX. The proposed optimized model shows more accurate result of 9.6 RMSE and 19.16 MAE. The training of model utilizes historical dataset from kaggle website which contains time series data in hours for a group of stations in India from 2017 to 2020, containing readings of gases such as PM2.5, PM10, CO, NOX, O3, SO2, NO.

(Jeya et al., 2020) has proposed recurrent neural network called Bi-LSTM to forecast the PM 2.5 pollutant. The work contains dataset gathered through UCI machine learning repository with hourly PM 2.5 data of the US embassy, recorded for Beijing city. The collected dataset is preprocessed by normal distribution with a mean of 0 and a standard deviation 1. Then, Bi-LSTM model has been applied with batch size of 16, epoch as 20 and dropout of 0.2. The hidden layer of Bi-LSTM produces the forward and reverse data which results in effective performance of analyzing time series data of PM 2.5 pollutant. The performance evaluation of the model is resulted with values of 7.53, 9.86, 0.1664 for MAE, RMSE and SMAPE.

(Bekkar et al., 2021) has proposed a hybrid CNN-LSTM forecasting model for AQI. The work has collected dataset from UCI repository which contains pollutant components like PM2.5, PM10, SO2, NO2, CO, O3 and meteorological features such as dew point, temperature, atmospheric pressure, combined wind direction, cumulated wind speed, cumulated hours of snow and rain. The dataset consists of data of adjacent stations in different time periods from March 1st to February 28th, 2017. The dataset is preprocessed by filling the missing values with spline linear interpolation method. Then, the meteorological feature called wind direction containing categorical data is encoded, which helps in determining the concentration of pollutant PM2.5. Followed by, the values of PM2.5 is normalized with min-max method to improve prediction accuracy. And, the feature selection process applies Pearson correlation which selects the pollutant and meteorological feature that act crucial for PM2.5 concentration forecasting. The spatial and temporal characteristics of selected features are extracted by convolution and LSTM layers which helps in achieving better result of 6.742 MAE, 12.921 RMSE and 0.989 $R^2$.

(Nguyen et al., 2024) has proposed an optimized deep learning model ACNN-QPSO-LSTM for AQI prediction in seoul's urban region. The dataset collected from seoul air data contains hourly concentrations of six air quality components O3, CO, NO2, SO2, PM10 and PM2.5 of 25 district stations. The preprocessing of dataset comprises standardizing the timestamp data, replacing the missing values and negative values with nan. Then, encoder-decoder framework is applied by improving LSTM network where quantum particle swarm optimization fine tunes the LSTM network parameters which reduces redundancy and saves time in capturing irregular patterns. Additionally,

the local and global dependence are captured through Attention based CNN, and ARIMA model predicts linear component values before yielding the non-linear component values through proposed model. The predicted values of linear and non-linear components are synthesized to generate output through XGBoost Regressor. The performance of the model has been proved with the lowest MAE, MSE and competitive R2 values.

(Chang et al., 2020) has introduced an Aggregated Long-Short-Term-Memory (ALSTM) method for predicting the hazardous pollutant PM2.5. The model utilizes dataset features to forecast PM2.5 over 8 hours and includes a wider sub-neural network to learn from different feature sources. ALSTM was compared to Gradient Boost-Tree Regression (GBTR), Support Vector Machine Regression (SVR), and LSTM. ALSTM outperformed these models with lower MAE and RMSE, making it operative in predicting PM2.5 values for both near and distant future time points. (Belavadi et al., 2020) used a Long Short-Term Memory with Recurrent Neural Network (LSTM-RNN) to forecast air quality data gathered from Bengaluru by constructing the model with three layers. The model's RMSE was between 30-40 ppm in Bengaluru and 0-5 ppm in Amaravati, indicating high temporal variance. Improving the model complexity and incorporating weather factors could enhance its performance. (Arora et al., 2022) used a history-dependent RNN with fractional derivatives to predict AQI, collecting data from multiple cities for 2015-2020. Their model achieved a low MAPE of 3.22%, but an automated optimization-based approach is needed for derivative order selection. (Qiao et al., 2019) proposed a model using wavelet transform (WT), SAE, and LSTM for PM2.5 predictions. They collected data from six Chinese air quality study cities and statistically analyzed it. The stack encoder-long short-term memory (SAE-LSTM) model, with optimized wavelet layers, performed well for PM2.5 but may not be suitable for other pollutant time series without adjustments.

(Qadeer et al., 2020) analyzed various ML models for predicting PM2.5 and found LSTM to outperform others. They used datasets containing meteorological parameters and pollutants, and employed DL methods for PM2.5 predictions, with LSTM showing the best results. Future work may involve predicting other correlated pollutants like NO2, O3, and PM10 using ML models. (Janarthanan et al., 2021) proposed a fused DL approach for air quality prediction in Chennai using air data of three monitoring stations. The data covers various pollutants and environmental variables. Pre-processing and feature extraction were performed, and the SVM-LSTM model outperformed existing models with a low RMSE and a high R value. (Zhang et al., 2022) has proposed a fusion of CNN and LSTM methods for air quality forecasting. They addressed missing data with interpolation, used residual CNN for spatial features, and LSTM for temporal features. The deep air model outperformed baseline models, achieving the lowest error rate of 27.1 for forecasting PM2.5, but struggled with irregular O3 concentration changes, leading to lower accuracy.

(Ren et al., 2023) has proposed urban air environmental control policies. They emphasize the importance of legislation, clarified government roles, and public participation. Their recommendations include increased financial support, enhanced legal documents, strict source control standards, and promoting technological innovation, along with the need for a regional monitoring platform. The recent high air pollution levels in Polish cities highlight the need for policy changes and integrated urban development for local air quality management. Urban planners and designers play a vital role in prioritizing air quality as a component of sustainable city development, addressing sources like traffic and industrial activities (Gulia et al., 2015). Solutions include green infrastructure, sustainable transport, building design, air quality sensors, and intelligent prediction

systems. Current AQI prediction methods use ML and DL, with Hadoop for efficiency, but further improvements can be made with hybrid methods. The proposed work combines DL with Hadoop Map Reduce for AQI prediction, labeling, and forecasting.

This proposed work has been structured as follows, "Methodology and Analysis" section explains detailed methodology of proposed model for AQI value prediction, AQI level labeling, forecasting the AQI values of future five years, "Results" and "Discussion" sections presents the implementation and describes the experimental results of the proposed models whereas, "Conclusion" section covers inference of the proposed work.

## 2. Methodology and Analysis

The proposed Hadoop based hybrid DL architecture aims to optimize AQI prediction and forecasting process by applying the real-time data collection, pre-processing and normalization, then feature selection with map reduce which is depicted in Figure 1. Further, prediction of AQI values, labelling of AQI levels and forecasting of AQI values for future five years is performed with DL models like LSTM, CNN, and LSTM-CNN. Finally, the execution of the proposed DL models has been evaluated using accuracy and error rate.

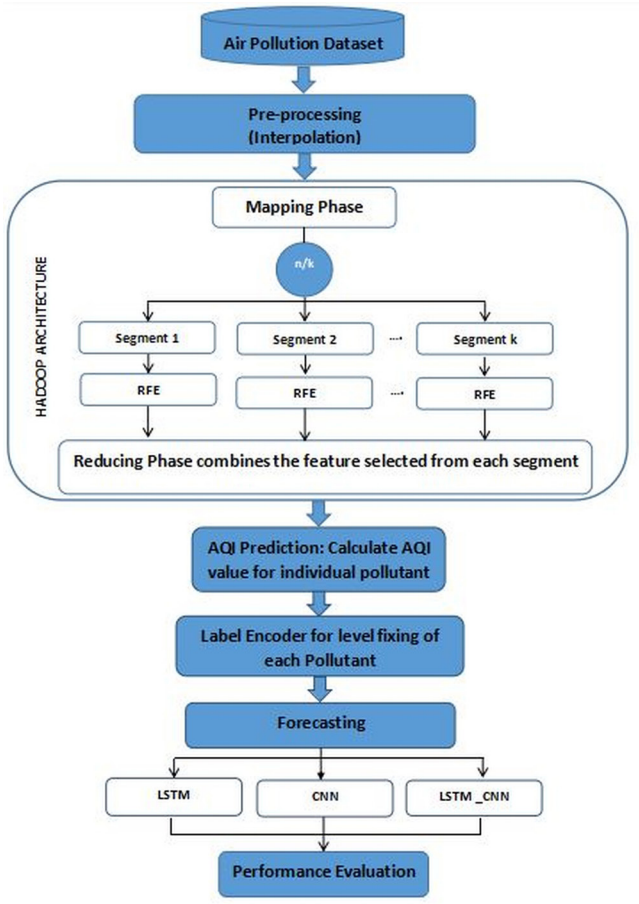



Fig. 1. Hadoop Map Reduce based Deep-Hybrid Air Quality Prediction and Forecasting

## Data collection

The input dataset was gathered from the Central Pollution Control Board (CPCB) website which monitors various pollutants in many cities across India through 883 monitoring stations in 379 cities or towns in 28 states and 7 union territories of the country. The observation of pollutants is carried out for 24 hours with an occurrence of twice a week, to have 104 annotations in a year [cpcb.nic.in]. The average air pollutant concentrations over 8 hours and 24 hours are updated and accessible through a Data Application Programming Interface (DAPI).

The dataset contains 15 years (2009-2003) data of AQI values covering three different areas such as industrial, commercial and residential with seven specific stations at Chennai namely Alandur Depot, Arumbakkam, Kodungaiyur, Manali, Perungudi, Royapuram, Velachery. And the dataset comprises 46418 rows with 24 columns as shown in Figure 2. The investigative variables collected from the seven stations include pollutant features and meteorological parameters as represented in Figure 3, where the pollutant features are Particulate Matter 10, Particulate Matter 2.5 (PM 2.5), Nitric Oxide (NO), Nitric Dioxide (NO2), Nitric x-oxide (NOX), Ammonia (NH3), Sulphur Dioxide (SO2), Carbon Monoxide (CO), Ozone (O3), Benzene and Toluene. The observation of meteorological parameters comprises Wind Speed (WS), Wind Direction (WD), Relative Humidity (RH), Temperature, Atmospheric Pressure (AP), Solar Radiation (SR), Radio Frequency (RF). The size of the collected dataset is 9.9 MB MB as represented in Figure 3.

## Data Pre-processing

The CPCB air pollution dataset is a repository of huge information collected from various locations in different time series, which contains missing values and huge noise in the few features due to instrument failure. The analysis of missing values in the dataset for all variables results below 89% where the pollutant features has 64% and meteorological parameters has 89% as represented in Table 1. To alleviate the loss of information, the imputation process has been done using interpolation method. Interpolation ensures continuity in the dataset, preventing issues arising from missing or noisy data. The linear interpolation is the mathematical method as given Eq. (1), which is applied to descend value between two points having an approved value. In AQI prediction and forecasting work, the linear interpolation can be described as a method

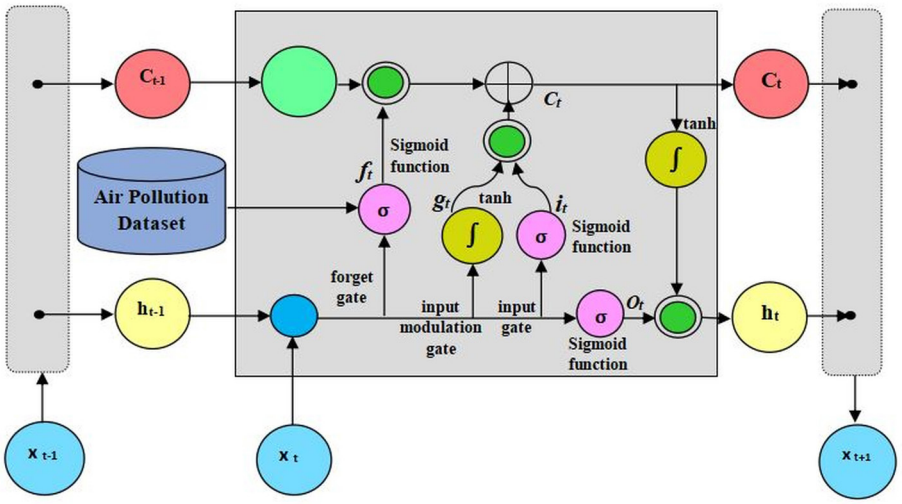



Fig. 2. Sample of Air Quality Dataset in Chennai from 2009-2023

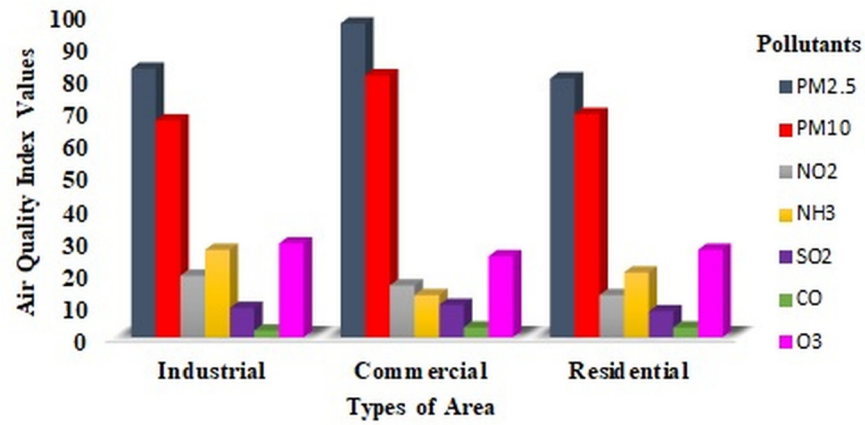



Fig. 3. Details of the Air Quality Dataset used in the Stud

Table 1. Summary of Air Quality Dataset Sample Size

| Variables | Number of Samples | Non-Missing Values | Missing Values | Percentage of Missing Values |
|---|---|---|---|---|
| PM 2.5 | 46418 | 27979 | 18439 | 39.72% |
| PM 10 | 46418 | 23981 | 22437 | 48.34% |
| NO | 46418 | 31395 | 15023 | 32.36% |
| $NO_2$ | 46418 | 31477 | 14941 | 32.19% |
| $NO_x$ | 46418 | 30709 | 15709 | 33.84% |
| $NH_3$ | 46418 | 27126 | 19292 | 41.56% |
| $SO_2$ | 46418 | 29912 | 16506 | 35.56% |
| CO | 46418 | 18386 | 28032 | 60.39% |
| $O_3$ | 46418 | 16821 | 29597 | 63.76% |
| Benzene | 46418 | 4950 | 41468 | 89.34% |
| Toluene | 46418 | 4885 | 41533 | 89.48% |
| WS | 46418 | 28976 | 17442 | 37.58% |
| WD | 46418 | 29392 | 17026 | 36.68% |
| RH | 46418 | 30483 | 15935 | 34.33% |
| Temperature | 46418 | 12336 | 34082 | 73.42% |
| AP | 46418 | 10596 | 35822 | 77.17% |
| SR | 46418 | 17792 | 28626 | 61.67% |
| RF | 46418 | 20190 | 26228 | 56.50% |

Algorithm 1

| Algorithm 1: Recursive Feature Elimination |
|---|
| Input: Dataset D={$d_1$,$d_2$ ,... $d_n$}<br>Output: Feature Rank R={$r_{d1}$,$r_{d2}$ ,... $r_{dn}$} |
| Step 1: Built a classifier by the training dataset from D<br>Step 2: Observe the performance of the classifier<br>Step 3: For each feature $d_i$ in D, **D'=D-$d_i$**<br>Step 4: Train the classifier with **D'**.<br>Step 5: Compute the classification accuracy<br>5.1. Find out the accuracy loss of classifier due to elimination of **$d_i$**.<br>Step 6: Compile the loss profile of features {$d_1$,$d_2$ ,... $d_n$}<br>Step 7: Compute features rank R={$r_{d1}$,$r_{d2}$ ,... $r_{dn}$} from the loss profile. |

of resembling the pollutant value of a given function at a given set of discrete pollutant values.

$$Y = Y1 + \frac{(Y2-Y1)}{(X2-X1)} * (X - X1) \qquad (1)$$

Where, the interpolation value is denoted by y, the independent variable is denoted by x, values of the function at one point are denoted by x1, y1, values of the function at another point are denoted by x2, y2.

**Feature Selection**

The accuracy of the AQI prediction and forecasting process can be improved through selection of the optimal features subset. The real time CPCB data on air pollution provides spatio-temporal dataset based on multiple parameters such as pollutants, meteorology, location etc. Since, the researchers face the issue of handling large dataset in AQI prediction and forecasting process, the research exertion currently utilizes the feature selection concept based on hadoop map reduce in order to reduce the time complexity.

The present study involves prediction of air quality with specific pollutant features such as $PM_{10}$, $PM_{2.5}$, $NO_2$, $O_3$, CO, $SO_2$, $NH_3$ and specific AQI levels namely Good, Satisfactory, Moderate, Poor, Very Poor, and Severe as shown in Table 2. This specific seven pollutant features have varied characteristics

Table 2. Summary of AQI Levels

| AQI Level (Range) | PM10 24-hr | PM2.5 24-hr | NO2 24-hr | O3 8-hr | CO 8-hr | SO2 24-hr | NH3 24-hr |
|---|---|---|---|---|---|---|---|
| Good (0-50) | 0-50 | 0-30 | 0-40 | 0-50 | 0-1.0 | 0-40 | 0-200 |
| Satisfactory (51-100) | 51-100 | 31-60 | 41-80 | 51-100 | 1.1-2.0 | 41-80 | 201-400 |
| Moderate (101-200) | 101-250 | 61-90 | 81-180 | 101-168 | 2.1-10 | 81-380 | 401-800 |
| Poor (201-300) | 251-350 | 91-120 | 181-280 | 169-208 | 10.1-17 | 381-800 | 801-1200 |
| Very Poor (301-400) | 351-430 | 121-250 | 281-400 | 209-748 | 17.1-34 | 801-1600 | 1201-1800 |
| Severe (401-500) | 430+ | 250+ | 400+ | 748+ | 34+ | 1600+ | 1800+ |

*AQI category, pollutants and health breakpoints, according to Central Pollution Control Board

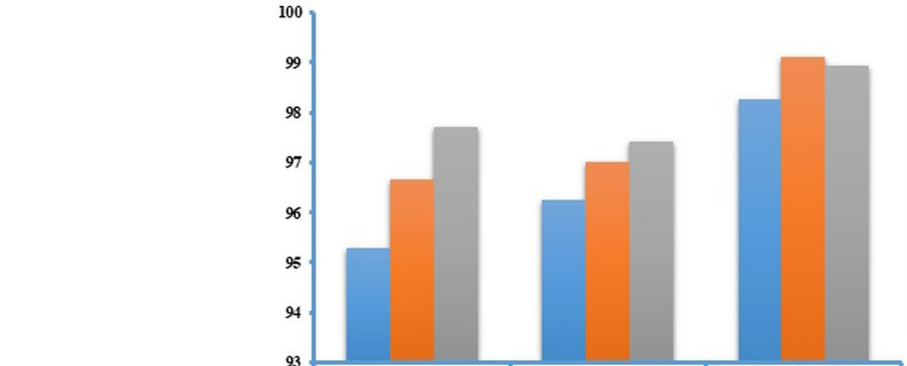
Algorithm 2

| Algorithm 2: Label Encoding |
|---|
| Input: AQI levels - A list of categorical AQI levels (Good, Satisfactory, Moderate, Poor, Very Poor, and Severe)<br>Output: Encoded_ AQI levels - A list of corresponding numerical values (0, 1, 2, 3, 4 and 5) |
| Step 1: Initialize an empty dictionary levels_map to store the mapping of unique AQI levels to numerical values.<br>Step 2: Initialize an empty dictionary reverse_levels_map to store the mapping of numerical values back to AQI.<br>Step 3: Find the set of unique AQI levels in the input list 'AQI levels' and store it in unique_labels.<br>Step 4: Initialize index i as 0.<br>Step 5: For each unique AQI level in unique_labels, do the following:<br>5.1. Add the key-value pair (level, i) to level_map.<br>5.2. Add the key-value pair (i, level) to reverse_level_map.<br>5.3. Increment i by 1.<br>Step 6: Initialize an empty list encoded_ AQI levels.<br>Step 7: For each level in the input list 'AQI levels', do the following:<br>7.1. Look up the numerical value of the AQI level in level_map.<br>7.2. Append the numerical value to encoded_ AQI levels.<br>Step 8: Return encoded_ AQI levels as the output. |

with potential health effects as described in Table 3. so, proposed AQI prediction and forecasting work based on specific air pollutants of varied locations implements recursive feature elimination (RFE) algorithm for feature selection with the Map Reduce (MR) model, the specified air dataset (AirDs) is spilt into number of smaller sets and scattered across the network and for every single division, the RFE (Misra, P. et.al, 2020) algorithm is applied in parallel. The dataset trials are correspondingly distributed and processed in parallel so as to achieve the class balance. In MR, (AirDsi) is mapped into the equivalent mapi task. Throughout the mapping phase, AirDsi involves the RFE.

Algorithm 3

| Algorithm 3: Long Short Term Memory |
|---|
| Input: Predicted AQI values<br>Output: Forecasted AQI values |
| Step 1: Model Initialization<br>Initialize the weights and biases of the LSTM units.<br>Initialize the cell state, c, and the hidden state, h, to zeros.<br>Step 2: Compute the input gate, i_t: $i_t=\text{sigmoid}\ (W_i\cdot[h_{t-1},x\_t\ ]+b_i)$<br>Step 3: Compute the forget gate, f_t: $f_t=\text{sigmoid}\ (Wf\cdot[h_{t-1},x_t\ ]+b_f)$<br>Step 4: Compute the output gate, o_t: $o_t=\text{sigmoid}\ (Wo\cdot[h_{t-1},x_t]+b_o)$<br>Step 5: Compute the candidate cell state, g_t: $gt=\tanh\ (WC\cdot[h_{t-1},x_t]+b_C\ )$<br>Step 6: Compute the new cell state, c_t: $Ct=f_t*\ C_{t-1}+i_t*C_t$<br>Step 7: Compute the new hidden state, h_t: $ht=(o_t*\tanh C_t)$<br>Where $f_t$,$i_t$, and ot represent the outputs generated by forget gate, input gate, and the output gate, respectively. $C_t$ and $g_t$ are the cell states respectively and ht is the output of LSTM network. |




| | Long Short Term Memory | Convolutional Neural Network | Long Short Term Memory - Convolutional Neural Network |
|---|---|---|---|
| | ACCURACY LEVELS | | |
| COMMERCIAL : Alandur, Arumbakkam | 95.3 | 96.25 | 98.25 |
| RESIDENTIAL : Kodungaiyur, Velachery | 96.67 | 97.02 | 99.12 |
| INDUSTRIAL : Perungudi, Royapuram, Manali TNPCB, Manali CPCB | 97.71 | 97.42 | 98.92 |

Fig. 2. Sample of Air Quality Dataset in Chennai from 2009-2023

The RFE algorithm is applied to each partitions, the output of each map function is represented as rfei = ¬ ($rfe_{i1}$, ..., $rfe_{iA}$), where the number of selected features is denoted by 'A'. The reduce phase combines the features selected from each partitions, obtaining a vector 'sp' given in Eq. (2), where spj denotes the jth feature. This is the outcome of the complete feature selection process, which is used for further air quality in AQI prediction process:

$$sp = \{sp_1, \ldots, sp_A\}, sp_j = \frac{1}{n}\sum_{i=1}^{n} rfe_{ij} \qquad (2)$$

where n=number of tasks in the Map Reduce. Commonly, the reduce phase is conceded out by a separate process thus dropping the implementation time in Map Reduce (Florence, A. et.al, 2021). The complete execution is done with a single Map Reduce procedure which abolished the additional disk admittances.

Algorithm 4

Algorithm 4: Convolutional Neural Network

Input: Predicted AQI values
Output: Forecasted AQI values

Step 1: Model Initialization with Hyperparameter Tuning

- Build CNN model with hyperparameters such as the number of layers, drop out, batch size, epoch and learning rate.
- Choose the hyperparameters, h=(Li, Di, Bi, Ei, Lri).
- Let C denote, CNN algorithm with N hyperparameters
- The domain of nth hyperparameter given as Λn= {Ln, Dn, Bn, En, Lrn}, where Λ denotes N-ary operation
- Since, the vector of hyperparameter is initialized, it is denoted as $\lambda \in \Lambda$.
- CNN model represented with hyperparameters is denoted as $C\lambda$.

$$C\lambda = \Lambda n$$

Step 2: Model Training with predicted AQI values and encoded AQI levels.
Step 3: Forecasting of AQI values.

**RFE Algorithm**

The RFE is a wrapper-type feature selection algorithm which works by examining for a subdivision of features, beginning with all features in the training dataset and successfully selecting the relevant features. The proposed RFE with Map Reduce (RFE-MR) method aims at selecting seven specific air pollutants such as $PM_{10}$, $PM_{2.5}$, $NO_2$, $O_3$, CO, $SO_2$ and $NH_3$ of three different areas with seven specific locations of chennai like industrial (Perungudi, Royapuram, Manali), commercial (Alandur, Arumbakkam) and residential (Kodungaiyur, Velachery) which are relevant to the subsequent processes like AQI prediction and labelling of AQI levels, by plummeting the time complexity. The proposed RFE algorithm is outlined in algorithm 1.

**Prediction of AQI Values**

The proposed work performs prediction of AQI values by calculating the AQI values of the seven pollutants covering three different areas with seven specific locations of Chennai as mentioned earlier. The proposed model has predicted the AQI value of each pollutant based on the pollutant quantity by applying AQI standards provided by PCB (Mustakim et al., 2023) as given in Eq. (3).

Algorithm 5

Algorithm 5: Long Short Term Memory – Convolutional Neural Network

Input: Predicted AQI values
Output: Forecasted AQI values

Step 1: Initialize the model architecture.
Step 2: Add a CNN layer with specified hyperparameters such as layers, drop out, batch size, epoch and learning rate.

- N hyperparameters, Λi=(Li, Di, Bi, Ei, Lri)
- A vector of hyperparameter is initialized, it is denoted as $\lambda$
- CNN layer instantiated with hyperparameters is denoted as $C\lambda$.

$$C\lambda = \Lambda i$$

Step 3: Add a pooling layer with the specified pool size (P).
Step 4: Reshape the output of the pooling layer to have the dimensions (new_height, new_width * F).
Step 5: Add a LSTM layer with the specified number of units (U).

$$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_i)$$
$$C_t = \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_c)$$
$$C_t = f_t\, C_{t-1} + i_t * C_t$$
$$o_t = \sigma\,(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$H_t = (o_t * \tanh C_t)$$

Where ft,it, and ot represent the outputs generated by forget gate, input gate, and the output gate, respectively. $W_f$, $W_i$, $W_C$, and $W_o$ are the input weights, respectively. $b_f$,$b_i$,$b_C$, and bo are bias terms and Ht is the output of LSTM network.
Step 6: Add fully connected layers with the specified number of units and 'relu' activation function.
Step 7: Add an output layer with the specified number of classes (C =6, that is, AQI levels) and 'softmax' activation function.
Step 8: Compile the model with the categorical cross-entropy loss function.
Step 9: Train the model on the training data (X_train, y_train) with predicted AQI values by specifying number of epochs, using the specified batch size and validation data (X_val, y_val).
Step 10: Evaluate the model on the test data (X_test, y_test) with AQI values, and then calculate the accuracy.
Step 11: Perform forecasting of AQI values on the test data. Use the trained model to generate predicted AQI values.

Table 3. Brief Description and Health Effect of Pollutant Features

| S.No | Pollutant | Brief Description | Health Effect |
|---|---|---|---|
| 1. | Particulate Matter ($PM_{10}$ and $PM_{2.5}$) | The PM10 are particles with a diameter of 10 micrometres or less, which are tiny enough to enter the lungs. The PM2.5 are particles with a diameter of 2.5 micrometres or less, which are so tiny and intensely harm the lungs and bloodstream. | The Short-term exposure accelerate heart attacks and arrhythmias. The Long-term exposure influence improper pulmonary functioning, and cardiovascular disease. |
| 2. | Nitrogen Dioxide ($NO_2$) | NO2 is a highly reactive gas which is released into the air by motor vehicles, industry, unflued gas-heaters and gas stove tops. The high concentrations of NO2 can be found notably in traffic roads and indoors includes unflued gas-heaters cigarette smoke or canister gas. | Increased susceptibility to frequent asthma attacks and airway inflammation. |
| 3. | Ozone ($O_3$) | O3 is highly reactive form of oxygen gas. Ozone is formed in two layers of the Earth's atmosphere. Ozone in the upper atmosphere defends us by filtering out harming ultraviolet radiation from the sun while, ozone at ground level is harming human health. | Irritation and inflammation of eyes, nose, throat, lower airways, reduced lung function with exacerbation of asthma and chronic bronchitis. |
| 4. | Carbon Monoxide (CO) | CO is a noxious gas without odour, taste or color and it is produced due to burning fuels, including gas, wood, propane or charcoal. | The earlier symptom includes nausea, vomiting, dizziness and severity causes brain injury or death. |
| 5. | Sulphur Dioxide ($SO_2$) | SO2 is highly reactive gas with a pungent putrid. It is produced by burning fuel at power plants. Also decomposition and combustion of organic matter, spray from the sea, and volcanic eruptions release sulphur gases. | Initially narrowing of the airways, leading to wheezing, chest tightness, shortness of breath, then frequent asthma attacks, and aggravated cardiovascular issues. |
| 6. | Ammonia ($NH_3$) | NH3 is a colorless gas with a strong odor which is termed as a primary air pollutant. This gas is deleterious to human health and the environment, and can cause ecological issues such as acidification, and nitrification. | Inhaling large amounts of ammonia can be venomous, and diluted concentrations leading to pulmonary damage and death at very high concentrations. |

$$I_p = \frac{I_{High} - I_{Low}}{BP_{High} - BP_{Low}} (C_p - BP_{Low}) + I_{Low} \qquad (3)$$

Where Ip is the index for pollutant p, Cp is the shortened concentration of pollutant p, BPHigh is the concentration division that is greater than or equal to Cp, BPLow is the concentration division that is less than or equal to Cp, IHigh is the AQI value corresponding to BPHigh, ILow is the AQI value corresponding to BPLow.

**Labelling of AQI Levels**

The proposed work includes the labeling of six AQI levels which is basically done based on the AQI range provided by CPCB air quality standards (app.cpcbccr.com). And, this AQI levels varies depending upon the

different areas like industrial, commercial and residential. The proposed work has applied label encoding algorithm for labeling the list of six categorical AQI levels with the corresponding encoded numerical values (0, 1, 2, 3, 4 and 5). The label encoding algorithm assigns a unique integer values to each categorical AQI level as represented in algorithm 2.

**Forecasting of AQI Values**

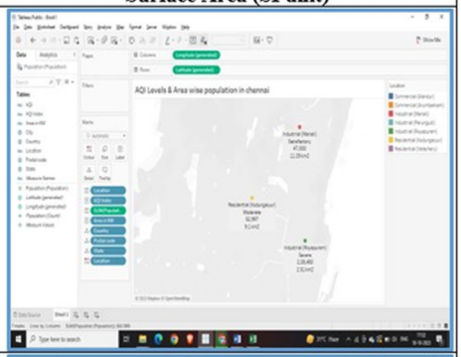
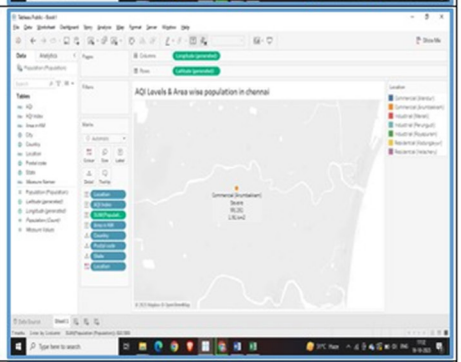
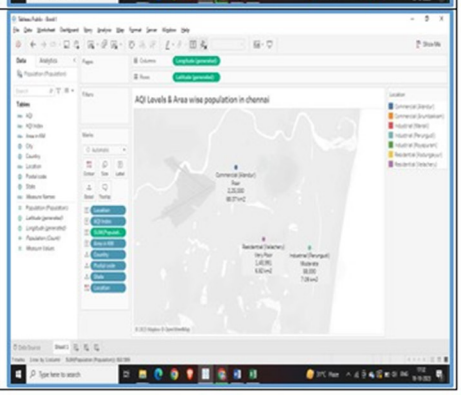
The forecasting of AQI values can significantly prevent air polluting by taking appropriate actions and measures to control the air-pollutant. Based on the results of predicted AQI values and labeling of AQI levels with the varied air pollutants and areas, the AQI values can be forecasted for future years. As mentioned earlier, the process of forecasting AQI values can be advanced by provoking AI models. The right choice of time series data analyzing AI methods can achieve better accuracy on forecasting AQI values. Conferring to the review study of predicting and forecasting the AQI values, the current work applies time series data analyzing DL methods namely LSTM, CNN, and LSTM-CNN to forecast the AQI values efficiently. Training the proposed DL models involves a large dataset of predicted AQI values with corresponding labeled AQI levels to perform the forecasting of AQI values. So, the AQI values are forecasted by including the seven pollutants and six AQI levels covering three different areas with seven specific locations for five future years (2024, 2025, 2026, 2027 and 2028).

LSTM: The LSTM is a type of RNN design befitting for handling sequential data like time series and capture long-term dependencies (Bihter et al., 2022). The main motivation behind LSTM networks is addressing the limitations of traditional RNNs, which struggle to retain information from earlier time steps as it propagates through the network, the LSTM network introduces memory cells and gates that allow the mechanism to capture long-term dependencies and handle varying time warp in the data. Air pollution analytics typically involves analyzing time series data collected from various sensors or monitoring stations. Since, the LSTM models have demonstrated potential in AQI prediction with better accuracy rates in the previous studies. The present study proposed the LSTM method to forecast the AQI values of future five years at varied locations of Chennai based on the historical data of the specific pollutants. The proposed LSTM algorithm is outlined in algorithm 3.

A LSTM network uses cell state (Ct), hidden state (ht), input gate (it), output gate (ot), forget gate (ft) and input-modulation gate (gt) to process the data sequentially passing the information as it propagates forward. The architecture of LSTM model is shown in Figure 4. The cell state (Ct) is shown as a horizontal line runs through the entire network and has the ability to add or remove the information with the help of gates. The process of the cell-state is to carry the information through the sequence processing and theory information from earlier time steps can be carried all the way through the last time step thus reducing the effect of short-term memory. As the process goes on, the information is added or removed from the cell states to gate states. Gates decide which information is allowed on the cell state. The first gate that is the forget gate is responsible for learning what information is necessary to keep or forget as they contain sigmoid function. The sigmoid function generates values between 0 and 1, describing the portion of each component to be let through. The tanh function generates a new vector, which is added to the state. The new cell state (ct) is updated based upon the outputs generated from the gates. The hidden state (ht) provides the final output.

CNN: The CNN is a category of advanced NN that are being applied for AQI prediction and forecasting in recent days. Specifically, the CNN models are applied for forecasting of the AQI values in an automated manner. The basic architecture of a CNN includes multiple layers of convolutional and pooling operations, followed by fully connected layers that

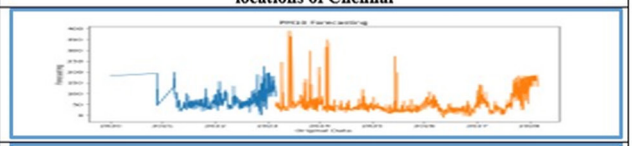
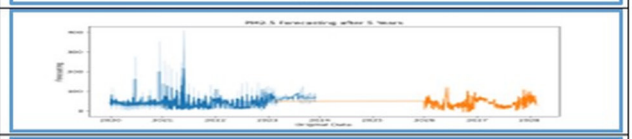
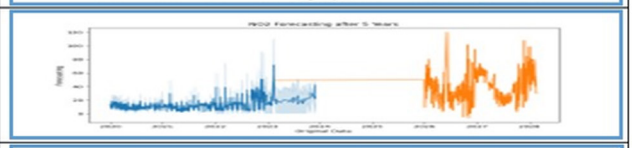
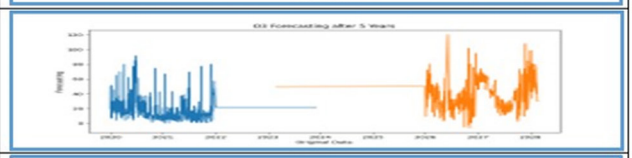
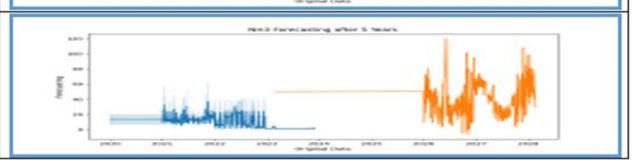
Table 4. Labelling of AQI Levels in Seven Locations of Chennai

| S.No | Locations | AQI Levels and Human Population Count, Surface Area (SI unit) |
|---|---|---|
| 1. | Manali (Industrial)<br>Kodungaiyur<br>(Residential)<br>Royapuram<br>(Industrial) |  |
| 2. | Arumbakkam<br>(Commercial) |  |
| 3. | Alandur (Commercial)<br>Velachery<br>(Residential)<br>Perungudi (Industrial) |  |

perform the classification task. Additionally, in order to improve the efficiency of the classification process the hyperparameter tuning can be done. Selecting appropriate hyperparameters, including the quantity of layers, filter dimensions, and learning rate are crucial factors to contemplate, as they can substantially influence the model's performance.

The existing studies have indicated that tuned CNN models achieve high levels of accuracy in AQI prediction and forecasting, with some surpassing state-of-the-art performance. So, the proposed AQI value forecasting with CNN model takes input and extracts relevant features like predicted AQI values of specific pollutants and varied locations through convolutional and pooling operations. The features are then fed into fully connected layers that perform the forecasting of AQI values for future five years. The outputs are produced in the form of discrete

Table 5. Forecasting of AQI Values in Seven Locations of Chennai

| S. No | Pollut -ants | Forecasting of AQI Values for 5 future Years with 7 Air Pollutants covering 7 locations of Chennai |
|---|---|---|
| 1. | PM 10 |  |
| 2. | PM 2.5 |  |
| 3. | NO2 |  |
| 4. | O3 |  |
| 5. | CO |  |
| 6. | SO2 |  |
| 7. | NH3 |  |

categories, such as forecasting continuous scale of AQI values by representing the previous AQI values. The proposed CNN algorithm is outlined in algorithm 4.

LSTM - CNN: An LSTM-CNN architecture can be effectively used for air pollution data analytics. This combination allows for capturing both the temporal patterns and spatial dependencies present in the data, enabling accurate AQI prediction and forecasting. So, the LSTM-CNN architecture has been applied to air quality data analytics. By combining the strengths of CNN and LSTM, the architecture has effectively captured both the spatial and temporal patterns in trained air pollution data. This can help in predicting future AQI values accurately. The proposed LSTM-CNN algorithm is outlined in algorithm 5.

The proposed LSTM - CNN model initially performs spatial feature extraction, since the air pollution data is influenced by spatial factors. The CNN layer has been employed as a first layer to extract spatial features from the air pollution dataset. Through the convolutional layers the relevant pollutants for AQI forecasting are extracted. The Pooling layers are then used to reduce the spatial

dimensions of the feature maps, while maintaining the relevant pollutants. The second layer has been employed with LSTM to perform temporal dependency modeling. The output of the CNN layer is passed to the LSTM layer. The LSTM layer excels at capturing the temporal dependencies present in the AQI values of pollutants. They analyze the sequence of AQI values of the pollutants across time and learn to model the long-term patterns, such as daily, weekly, or seasonal variations in AQI values. So, the LSTM layers consider the previous AQI value predictions and update their memory cells to retain relevant pollutant information over time.

The fully connected layers has been added following the LSTM layers, to further process the extracted features and map them to perform the forecasting of AQI values. This layer performs additional feature transformations and dimensionality reduction to facilitate the forecasting of AQI values. Finally, the output layer produces the forecasted AQI values of future five years and this output layer is followed by the fully connected layers.

## 3. Results

The proposed research work is implemented in the system of 16 GB RAM and HDD of 1 TB using Python (parallel computing). The air datasets were collected through secondary mode via online repository of CPCB. The dataset has been applied with processes like preprocessing, feature selection, prediction of AQI values (Figure.5), labelling of AQI levels (Table.4), forecasting of AQI values for future five years (Table. 5).

The AQI values of the pollutants varies depending upon the different areas like industrial, commercial and residential, where the emissions of industries, commercial crowds, domestic disposals are the major causes of higher pollutant values. From

Table 6. Default model Hyperparameter Settings

| S. No | Model Type | Model Name | Hyperparameter settings |
|---|---|---|---|
| 1. | **Deep Learning model** | LSTM | • No. of input layer : 1<br>• No. of hidden layers : 4<br>• No. of output layer : 1<br>• No. of Epochs : 64<br>• Batch size : 32<br>• Activation Function : sigmoid |
| | | CNN | • No. of cells in each layer : [70, 70]<br>• Dropout rate : 20%<br>• Activation Function : ReLU<br>• Dense-layer unit : 1<br>• Batch size : 32 |
| 2. | **Hybrid Deep Learning model** | Hybrid CNN - LSTM | 1. Convolution 1D layer:<br>• No. of filters : 32<br>• Kernel : 3<br>• Stride : 1<br>2. Maxpooling layer<br>• Pool size : 3<br>3. LSTM layer<br>• No. of cells : 32<br>• Dropout rate : 30%<br>• Activation Function : ReLU<br>• Dense-layer unit : 1<br>• Batch size : 32 |

Figure.5, it is observed that the commercial areas like Alandur and Arumbakkam has shown higher level of AQI values which is mainly due to vehicular emission. According to a study on air pollution, the vehicle emission is a main source of air pollution at commercial areas (Harrison et al., 2021).

From Table 4, it is experimentally proven that, the industrial areas like Manali, Perungudi and Royapuram has population of 47,000, 58,000 and 2, 28, 450 covering the area of 11.29 $km^2$, 7.09 $km^2$ and 2.51 $km^2$. The residential areas like Kodungaiyur and Velachery has population of 52,997 and 1, 43, 991 covering the area of 9.1 $km^2$ and 6.82 $km^2$. The commercial areas like Alandur and Arumbakkam has population of 2, 25,000 and 55,151 covering the surface area of 68.07 $km^2$ and 1.91 $km^2$. Among these three areas, the locations Royapuram, Velachery, and Arumbakkam covers a small surface area of 2.51 $km^2$, 6.82 km2, 1.91 $km^2$ with exceeded population count of 2,28,450, 1,43,991 and 55,151. Here, the air quality is found to be very poor and severe because of higher population. Then, the locations like Perungudi and Kodungaiyur covers a medium surface area of 7.09 km2 and 9.1 km2 with the population count of 58,000 and 52,997 here, the air quality seems to be moderate because these locations have a compromising population count. Whereas the Alandur location with the exceeded population of 2, 25,000 resulting poor air quality because it covers a compromising surface area of 68.07 $km^2$. Finally, the Manali location resulting satisfactory air quality because of the lesser population count of 47,000 with large surface area coverage of 11.29 $km^2$. Even though the population count is compromising due to poor air quality people's health are affected.

The forecasting of AQI value is an operative way of shielding public health providing that caution against hazardous air pollutants. The proposed work concentrates on the forecasting of AQI values based on the seven pollutants, six AQI levels, covering three different areas with seven specific locations of Chennai for five future years which is represented in Table 5. The experiential results come about the air quality of the future five years seems poor compared to the present air quality. The forecasting results that impend meagre air quality can deliver early notice and cautionary to individuals and communities to support them limit exposure and lessen air polluting lifestyle.

**Performance Analysis of LSTM, CNN, and LSTM – CNN Models**

The effectuation of the proposed CNN, LSTM, LSTM – CNN models forecasting AQI of future five years are depicted in Figure.6 and Figure.7 for where error rate and accuracy metrics are used for analysis. In ML, computing a model's performance in one-off number is substantial, regardless of during training, cross-validation, or monitoring after deployment. From the previous studies, the error rate and accuracy are found to be the proper grading rule that is instinctive to understand and accordant with some of the most common statistical assumptions. The frequent statistical methods for finding the error rate between predicted AQI values and forecasting AQI values are MSE, RMSE, MAE, and MAPE, loss value. Hence, the reliability of the proposed LSTM, CNN and LSTM - CNN models are evaluated with error rates like MAE, MSE, and RMSE then efficiency of the model has been proved with Accuracy (Acc) rate. Here, the accuracy and error rates takes into account the probabilities or uncertainty of prediction and forecasting of AQI value based on how much the predicted AQI values varies from the forecasted AQI values. This gives us a more nuanced view into how well the model is performing.

Mean Absolute Error: The MAE is a measure of the middling size of the errors in a collection of AQI predictions as given in Eq. (4). It is measured as the middling total difference between the predicted AQI and the forecasted AQI.

$$\text{Mean Absolute Error} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \qquad (4)$$

Where, $y_i$ is predicted value, $x_i$ is actual value, n total number of data points.

Mean square error: The MSE calculates the average squared loss per forecasting of AQI value over the whole dataset as given in Eq. (5). To calculate MSE, sum up all the squared losses for individual samples and then divide by the number of samples.

$$\text{Mean Square Error} = \frac{1}{N}\sum_{(x,y)\in D}\left(y - \text{prediction}(x)\right)^2 \quad (5)$$

Where (x, y) is an example in which, X is the set of features that the model uses to make predictions and Y is the example's label, Prediction (x) is a function of the weights and bias in combination with the set of features x, D is a data set containing many labeled examples, which are (x, y) pairs, N is the number of examples in D.

Root Mean Square Error: The RMSE is the standard deviation of the prediction errors which is a commonly used measure for evaluating the quality of predictions. In AQI prediction and forecasting, the RMSE is computed by calculating the difference between the residual (predicted AQI values and forecasted AQI values) for each data point. Then, compute the norm of residual for each data point, compute the mean of residuals and take the square root of that mean as given in Eq. (6).

$$\text{Root Mean Square Error} = \sqrt{\overline{(f-o)^2}} \quad (6)$$

Where, f is the ground truth or predicted AQI values, o is observed or forecasted AQI values.

Accuracy: To compute the accuracy in AQI prediction and forecasting, accuracy is simply the fraction of correct AQI value forecasting as given in Eq. (7). The current work carries accuracy evaluation by computing the summation of correctly forecasted value divided by summation of whole forecasted values.

$$\text{Accuracy} = \frac{\text{Correct Forecasting Results}}{\text{All Forecasting Results}} \quad (7)$$

**Comparison of Error rates and Accuracy**

This section compares outcomes of AQI value forecasting for future five years using LSTM – CNN, LSTM and CNN models which is depicted in Figure.6 and Figure.7. It is observed that, the LSTM - CNN model performances in the accuracy aspect are 98.25, 99.12, and 98.92 which are found to be higher than those of the LSTM and CNN models proposed. And, the LSTM – CNN has MSE of 101.81, 102.5 and 100.82 which are reduced than those of the LSTM and CNN models thereby implying the efficacy of the Hybrid LSTM – CNN. And respectively.

The results show that by comparing error rates and accuracy values, the proposed LSTM-CNN model with its spatio-temporal balancing feature outperforms the LSTM and CNN models in forecasting of AQI values for future five years. The high accuracy made on AQI prediction and minimum error rates of AQI values forecasting in different locations of Chennai using three different models are represented in Figure.6 which proves that the DL models can globally support the system for prediction and forecasting of AQI values.

**Comparison of Actual and Proposed Results of AQI in Chennai**

The results include the consolidated report of the AQI value prediction covering three varied areas with seven specific locations of Chennai including seven pollutants for 15 years, which is depicted in Figure. 8. From the figure, it can be seen that the PM2.5 is highly polluting the Chennai locations. As per the current status of Tamil Nadu Pollution Control Board's (TNPCB) website report, at Chennai, the range of $PM_{2.5}$ is seven or eight times more than the WHO guideline limit (iqair.com).

The result of AQI value forecasting for future five years which is depicted in Figure.9 shows that the AQI levels will be higher in the future. Because, the pollutant called $PM_{2.5}$ is highly polluting in Chennai and

other pollutant values were also relatively higher so the results conclude that the future AQI values may be higher. Hence, the precautionary measures of CPCB could be enhanced in order to avoid the future hazards due to pollutants called $PM_{10}$, $PM_{2.5}$, NO2, O3, CO, $SO_2$ and $NH_3$.

## 4. Discussion

The implementation of the proposed DL models for predicting and forecasting of AQI values has been developed using deep neural network (DNN) Application Programming Interface, Keras with Tensorflow back end. The proposed work has tried three altered parameter tuning to design three DL models by tuning varied parameters, such as number of neurons, number of layers, optimizing function, and learning rate, attaining the best DL model which not only executes well on the train data but also achieving prediction results on the unobserved test data. The Hyperparameter (epochs, batch size, and dropout) settings for LSTM, CNN and LSTM - CNN were assigned same for comparison. During model construction process, dropout has been introduced to prevent overfitting in neural networks. These three DL models were trained using with the default parameter settings as presented in Table 6. At first the assembly of LSTM network has been done followed by the number of neurons in CNN model, dropout rate, and other parameters in each layer from top to bottoms are given. Finally, the parameter configured for each layer from top to bottom for LSTM – CNN model are given.

The proposed DL models has been loaded with the predicted AQI values and labeled AQI levels during the training process, in order to perform the forecasting of AQI values for five future years. The spatial and temporal constraints of varied locations give different prediction results, therefore average evaluation values are calculated for three areas like Industrial, Commercial, and Residential. The experimental results show that the hybrid LSTM – CNN model performs better than LSTM and CNN models with lower error rates. So, it is better to forecast the AQI values through LSTM – CNN than LSTM and CNN. Also, the LSTM – CNN model learns and forecast the AQI values in varied areas with good accuracy. Hence, the experimental graphs of the proposed DL models for AQI prediction and forecasting proves that the results of the proposed work can support the air quality prediction system in identifying the pollution hotspots, understanding the impact of AQI on air quality, and decision-making for pollution control measures.

## 5. Conclusion

The improvement of air quality index prediction and forecasting is useful for society since the AQI is an obligatory entity to predict the air quality for taking necessary efforts and measures to control the pollutants. With the advancement of ML, DL and big data technologies, the real time air quality prediction is desirable for future air quality predictions. So, the present study proposes Hadoop-map reduce based DL models to predict AQI values by demanding resolutions like big air pollution dataset, less memory usage and low time intricacy. The prediction of AQI values and labeling of AQI level has been done with air pollutant parameters and forecasting of AQI value has been done by using DL approaches. The experimental results of forecasting of AQI value for future five years imply that the performance of proposed hybrid DL structure (LSTM – CNN) has modified the problems of the proposed LSTM and CNN models to some extent. Compared with LSTM and CNN, the LSTM – CNN model with its spatio-temporal convoking feature, has forecasted the AQI values for future five years in varied areas of Chennai more accurately. The LSTM has achieved 97% accuracy with average MAE, MSE, RMSE error rates of 7.85, 102.38, 10.29. The CNN has achieved 97% accuracy with average MAE, MSE, RMSE error rates of 8.76, 102.67, 9.65. The hybrid LSTM – CNN model has attained better accuracy of 99% and

average MAE, MSE, RMSE error rates of 7.95, 101.71, 9.65. As a result, the proposed Hybrid LSTM – CNN model has been considered effective for AQI prediction and forecasting since it has achieved higher accuracy and reduced error values.

The proposed work has identified three areas covering seven specific locations of chennai as areas of concern for air quality prediction and forecasting. These regions encompass a diverse landscape, ranging from industrial zones to commercial hubs and residential neighborhoods, all of which contribute to the complex tapestry of pollutants that taint the city's air. To address this pressing issue, a holistic and concerted effort is needed, one that combines policy interventions, technological advancements, community engagement, and public awareness initiatives. This includes promoting public transportation accessibility and affordability, incentivizing solar and wind energy adoption, increasing green spaces, enforcing updated industrial emissions standards, providing financial incentives for eco-friendly choices, transitioning to electric vehicles, establishing air quality index systems with alert mechanisms, issuing health advisories during poor air quality episodes, and considering the impact of population growth and tree planting. Innovative technologies like oxy-combustion and liquid trees can also play a role. Regional cooperation, ongoing research and community involvement are essential for effective air pollution reduction.

## *Acknowlegdements*

*We show gratitude to Central Pollution Control Board for providing the air dataset for the years 2005 – 2023. The present study is encouraged and financed by RUSA 2.0 – Research Innovation and Quality Improvement. And, the Authors hereby states that there is no conflict of interest.*

## 6. References

Air quality index - Wikipedia

Arora, S. – Sawaran Singh, N. S. – Singh, D. – Rakesh Shrivastava, R. – Mathur, T. – Tiwari, K. – Agarwal, S. (2022). Air Quality Prediction Using the Fractional Gradient-Based Recurrent Neural Network. Comput Intell Neurosci, 1-14.

Bekkar, A. – Hssina, B. – Douzi, S. (2021): Air-pollution prediction in smart city, deep learning approach. Journal of Big Data, 8 (161): 1-21.

Belavadi, S. V. – Rajagopal, S. – Ranjani, R. – Mohan, R. V. (2020): Air quality forecasting using LSTM RNN and wireless sensor networks. Procedia Computer Science, 170: 241-248.

Bihter, D. – Ömer, O. D. – Nighot, S. – Suat, T. (2022): Prediction of air pollutants for air quality using deep learning methods in a metropolitan city. Urban Climate, 46: 101291.

Chang, Y. S. – Chiao, H. T. – Abimannan, S. – Huang, Y. P. – Tsai, Y. T. – Lin, K. M. (2020): An LSTM-based aggregated model for air pollution forecasting. Atmos Poll Res, 11(8): 1451-1463.

CPCB | Central Pollution Control Board

CPCB: https://app.cpcbccr.com/ccr_docs/FINAL-REPORT_AQI_.pdf

Florence, A. – Jha, P. – Nighot, S. – Mudaliyar, V. (2021): Air quality analysis and visualization using big data approach. Journal of Engineering and Sciences, 3(1): 14.

Ghufran Isam Drewil. – Riyadh Jabbar Al-Bahadili. (2022): Air pollution prediction using LSTM deep learning and metaheuristics algorithms. Measurement: Sensors, 24: 1-7.

Hajewski, J. – Oliveira, S. (2020): Distributed SmSVM Ensemble Learning. Recent Advances in Big Data and Deep Learning, Springer, 1: 7-16.

Harrison, Roy M. – Tuan Van Vu. – Hanan Jafar. – Zongbo Shi. (2021): More mileage in reducing urban air pollution from road traffic. Environment International, 149: 1-8.

IQAIR: Chennai Air Quality Index (AQI) and India Air Pollution | IQAir

Janarthanan, R. – Partheeban, P. – Somasundaram, K. – Elamparithi, P. N. (2021): A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities. Soc, 67:102720.

Jeya, S. – Sankari, L. (2020): Air Pollution Prediction by Deep Learning Model. 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE: 736-741.

Kothandaraman, D. – Praveena, N. – Varadarajkumar, K. – Madhav Rao, B. – Dhabliya, D. – Satla, S. – Abera, W. (2022): Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology.

Lelieveld, J. – Pozzer, A. – Pöschl, U. – Fnais, M. – Haines, A. – Münzel, T. (2020): Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. Cardiovascular research, 116(11): 1910-1917.

Liang, Y. C. – Maimury, Y. – Chen, A. H. L. – Juarez, J. R. C. (2020): Machine learning-based prediction of air quality. Applied sciences, 10(24): 9151.

Misra, P. – Yadav, A. S. (2020): Improving the classification accuracy using recursive feature elimination with cross-validation. Int. J. Emerg. Technol, 11(3): 659-665.

Mustakim, Nurul. – Ul-Saufie, Ahmad Zia. – Shaziayani, Wan. – Mohamed Noor, Norazian. – Mutalib, Sofianita. (2023): Prediction of Daily Air Pollutants Concentration and Air Pollutant Index Using Machine Learning Approach. Pertanika Journal of Science and Technology, 31(1): 123-135.

Nguyen, A.T. – Pham, D.H. – Oo, B. (2024): Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. Journal of Big Data, 11(71): 1-38.

Ren, C. – Yu, C. W. – Cao, S. J. (2023): Development of urban air environmental control policies and measures. Indoor. Built Env, 32(2): 299-304.

Seyedeh, R. S. – Ali, J. – Saba, K. – Mazaher, M. – Nematollah, K (2021): The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration. Urban Climate, 37: 100837.

Gulia, S – Shiva Nagendra, S.M. – Mukesh Khare – Isha Khanna. (2015): Urban air quality management-A review. Atmospheric Pollution Research, 6(2): 286-304.

Susanto, A. D. (2020): Air pollution and human health. Medical Journal of Indonesia, 29(1): 8-10.

Qadeer, K. – Rehman, W. U. – Sheri, A. M. – Park, I. – Kim, H. K. – Jeon, M. (2020): A long short-term memory (LSTM) network for hourly estimation of PM2. 5 concentration in two cities of South Korea. Appl. Sci, 10(11): 3984.

Qiao, W. – Tian, W. – Tian, Y. – Yang, Q. – Wang, Y. – Zhang, J. (2019): The forecasting of PM2. 5 using a hybrid model based on wavelet transform and an improved deep learning algorithm. IEEE Access, 7: 142814-142825.

World Health Organization (12 January 2020): Global Health Observatory (GHO) Data for Ambient Air Pollution (www.who.int/gho/phe/outdoor_air_pollution/en/).

Zhai, Y. – Ong, Y.S. – Tsang, I. W. (2014): The Emerging Big Dimensionality. IEEE Computational Intelligence Magazine, 9(3): 14-26.

Zhang, Q. – Han, Y. – Li, V. O. – Lam, J. C. (2022): Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE Access, 10: 55818-55841.