A Comprehensive Study of Ozone Concentration Forecasting Using Deep Neural Networks for Indian Air Quality Data

Abstract

This project aimed to forecast ozone concentration using deep learning and traditional machine learning models over New Delhi, India for 2023. We employ two deep learning models, Long Short Term Memory (LSTM) and Bidirectional LSTM, and their performances were compared with two traditional machine learning models, Multi Linear Regression MLR and Random Forest Regressor. These models were used to perform multivariate time series forecasting of ozone concentration using each day for the entire year data. Several predictors were used from the data like atmospheric temperature, and PM.25, NO2 concentrations. The model results were analyzed for 5 different monitoring stations in New Delhi for the 6 years from 2018 to 2023. However, CNN is used for capturing daily styles as well as yearly shifts of ozone, it particularly had prediction results for some regions. Air quality Parameters can be used to understand the air quality and meteorological conditions that may affect the concentration of ozone in the atmosphere. Through analyzing these trends, one can make predictions about the future concentration of ozone and develop strategies to mitigate its negative effects on human health and the environment. Models are validated in Beijing China monitoring data to evaluate performance on variable datasets. Through this research work, Recurrent Neural Models like LSTM and Bi-LSTM were found to be more effective in forecasting time series data. Regression metrics such as Mean Square Error, Root Mean Square Error,a nd R2 Score were used for estimation. The results indicated that the deep learning models outperformed the traditional machine learning models, indicating their suitability for forecasting ozone concentration. Graphs were also added to visualize the performances of these models on observed and predicted values.

Introduction

The Earth's atmosphere is the best of all the atmospheres in our solar systems. It is made up of many layers. Since technology has been getting better, people have been moving to cities, which has made pollution worse. There are many signs of pollution, such as ozone, nitrogen dioxide, PM2.5, and sodium dioxide. Ozone is a big part of the pollution that is always in the air in cities. [1] Ozone is a greenhouse gas and a pollutant in the air in cities. It has very bad effects on both climate change and people's health. In the past few years, a lot has been done to lower surface ozone levels by putting in place strict measures to control ozone precursors' emissions. Carbon emissions around the world are also linked to ozone. In terms of controlling emissions, methane, carbon monoxide, and volatile organic compounds, which are precursors to ozone, have a lot to do with ozone and how to control them. [2] Ozone has been found to be a major oxidant, and it is a part of photochemical smog, which is one of the main pollutants that lowers the quality of the air.

Ozone plays a unique role in absorbing certain wavelengths of incoming solar ultraviolet light. One reason ozone is a serious environmental problem is that it is not directly emitted into the air, which makes it hard to predict and control. [3] All life is shielded from the sun's damaging radiation by the ozone layer, but human activities have worn down this barrier. Less UV protection from the ozone layer results from this decrease in ozone concentration. This constant decrease causes higher risks for skin cancer and cataract rates. The combustion of fossil fuels resulted in higher concentrations of trace gases like nitrous oxide and carbon monoxide. As a result of the buildup, spread, and transformation of these air pollutants, the quality of the atmosphere has decreased. [4] Climate change and air pollution are both on the rise, causing environmental conditions to deteriorate. When temperatures increase, climate change leads to a weakening of the ozone layer.

Recent patterns and distribution of field studies have shown that there is an increase in mortality rate during the summer smog due to high ground-level ozone concentration. By the use of strict emission control measures for Ozone precursors, a substantial effort has been made to lower tropospheric ozone concentrations. A monitoring station has been built in accordance with ozone concentration forecasts to predict higher geographical distribution change that aids in ozone reduction. Also, it is crucial to execute accurate regional estimates for ozone concentration in order to reduce greenhouse gas emissions and ensure public health.

To determine the amount of ozone in the air, a variety of methods can be utilized. Numerical and data-driven approaches are the two basic methodologies to estimate air pollution levels. Numerical models include the Community Multistage Air Quality Model, the Weather Research and Forecasting Model with Chemistry, and the Goddard Earth Observing System with Chemistry. This approach supports a strong capacity for explanation and clear, rigid logic. Because of the high cost of computation and the difficulty in gathering sufficient data, it was also difficult to make long-term projections. Data-driven models are a potential method to make accurate predictions about the target variable by extracting useful information from large amounts of data about the variable. According to several earlier research, data-driven models may be divided into shallow machine-learning models and deep-learning models. The quantity of ozone in the air in an hour may be predicted using machine learning in just a few simple steps. The static Multi-Layer Perceptron model.

However, new approaches are being developed to enhance ozone monitoring and forecasting as technology advances. For instance, machine learning and AI are used to improve the accuracy of forecasts concerning ozone concentrations. These tools can analyze a large amount of data from several sources, including satellites, meteorological models, and ground-based monitoring stations, to create more precise and comprehensive ozone concentration maps. This can assist identify areas with high ozone pollution levels and direct the creation of policies to reduce ozone emissions. [7]

The empirical approach of machine learning is often used to anticipate ozone concentration. Numerous ensemble techniques have been employed over the years. Some models that have been employed in the study include random forests and support vector machines. One of the most important tasks for the researchers is to model the variations in ozone and to produce

precise predictions. There are both statistical (black box) and deterministic models. When utilizing partial differential equations to build a deterministic model to estimate ozone concentrations in a particular location, a variety of physical and chemical interactions between predictor variables must be taken into consideration. A lot of precise input data are also required for this procedure, including statistics on emissions, weather, and land cover. These are the key reasons why creating and maintaining deterministic models is expensive. [8] Several studies have improved the ozone concentration forecast system during the last ten years.

Contaminants like sulphur dioxide and particulate matter are among the contaminants that artificial neural networks are used to anticipate. Three prediction models are compared in this study: the multilayer perceptron, the finite impulse response (FIR) neural network, and the autoregressive-moving average with exogenous inputs (ARMAX). Calculating the hourly ozone concentration 24 hours in advance is the aim. At three Spanish monitoring sites in cities and towns, they looked at the peak ozone concentrations over the summer between 1996 and 1999. Based on the five performance criteria, the MLP neural networks outperformed the linear ARMAX models, which outperformed the dynamic FIR neural networks. [10]

In this paper, we suggest analyzing Indian air quality data using ozone concentration data obtained from many monitoring site. This includes forecasting ozone concentration for data on the quality of the air in India using several machine learning and deep learning techniques. All forecasting models may be compared, and the results can be used to make decisions in the future. For the purpose of predicting ozone, the suggested models will learn from regional patterns and long-term spatiotemporal distributions of air quality data. Our goal is to assess the state of ozone in the vicinity of a reputable monitoring station and the relationships between other pollutants and ozone concentration.

Literature Review

Brian S. et. al. has trained a deep learning model which is a hybrid combination of Recurrent Neural Network and Long short Term Memory models. The author had done forecasting the prediction for 8hr of Ozone concentration. During the Training of model, Data had been collected from Kuwait, small country located in the Persian Gulf. The Differential optical absorption spectroscopy analyzers were used to collect data for tropospheric Ozone. They have reduced the dimension of data with the help of PCA and correlation filters. The Mean Absolute Error was used as a regression metric. This allowed air station to accurately predict air pollution concentration while only monitoring key factors for real time analysis.[11]

Bilge Ozbay et. al. have applied multivariate statistical methods in predicting ozone (O3) measurements at the ground level of the troposphere as the function of pollution and atmospheric considerations. Various correlation between ozone and other pollutants were investigated using statistical correlation methods like bivariate and Pearson correlation methods. The paper contain two main techniques. One of the way to use (MLR) Multiple Linear Regression for prediction Ozone concentration while this also focused on PCA for reducing number of variables for prediction. The Model made for both annual and seasonal trend for forecasting the pollutant to test

the model efficiency in different condition. Annual and Seasonal Trends were evaluated individually in MLR models and calculated R2 values were found as 0.90, 0.92, 0.85 respectively. The Warming condition got the highest R2 Score 0.92.[12]

Ning Jin et. al. proposed a novel model in the log short term memory model. The model is an enhanced using nested LSTM layer and using multiple task multiple channels for forecasting AQI data. The data was pre-processed using Discrete Stationary Wavelet Transform that decompose the real data into multiple sub-signals including eliminate lower frequency and diminish higher frequency components. Many models like SVR, LSTM, NLSTM and MLP are used to compare the efficiency of proposed model. They have done prediction for six air pollutants from which the results for ozone was 0.99 of R2 score. They have used Beijing dataset from 12 observing station for UCI ML repository [13].

Ebrahim et al. have a one of its first kind of Convolutional Neural Network based regression model that is used to predict the real time hourly concentration in Seoul, South Korea for 2017. This paper emphasis on how we can also use convolutional layers for regression as CNN captures temporary variations of the input data by convolution through time series using the kernel of fixed size. The small changes have resulted in a better correlation that in term gives beneficial results for the prediction system. The model is made of five convolutional layers with continuation by a fully connected layer and then an output layer. The ReLU activation function is implemented to normalize input data. The Index of agreement and Correlation coefficient, MAE, and RMSE are used for evaluation. The IOA achieved was 0.87 and a correlation coefficient of 0.7. the Root mean square error was 12.01 for the entire 2017 year.[14]

Kabesok Ko. et al.haves worked on a dataset about ozone concentration. The dataset used here contains planetary boundary information. National Ambient Air Quality Standards include surface ozone as one of six hazardous air pollutants that can affect humans and the environment. Machine learning has boosted data-driven forecasting methods. Data-driven ozone forecasting models now include PBL height. Ozone predictions include PBL since it affects ozone concentrations. This article examines how PBL height affects surface ozone projections. MLP and bidirectional LSTM surface ozone forecasting models are shown. IOA, MAE, and RMSE are used to assess two ozone forecasting models. This shows that anticipated PBL height can enhance data-driven surface ozone concentration prediction models.[15]

In this study, Uwe Schlink et al. report on a comparative comparison of models, in which 10 datasets covering various meteorological and emission circumstances across Europe were subjected to 15 different statistical methodologies for ozone forecasting. They also make an effort to compare the effectiveness of the statistical methods. 1340 yearlong time series of daily forecasts were constructed for the comparative evaluation of forecasting performance (benchmarking), and the findings are reported in terms of established performance metrics. We discovered that the success index is of exceptional relevance through the analysis of correlations between the performance metrics. It was discovered that the accuracy of the 8-h average ozone concentration prediction was better than the 1-h mean ozone concentration forecast, making the former extremely important for operational forecasting. The sites in Central Europe's rural and suburban zones free from intense emissions (such as those from businesses) had the best estimates. Results demonstrate

that a method may be both beneficial and terrible. For the majority of situations, neural networks and generalized additive models are the best compromises since they can handle nonlinear connections and adjust to site-specific factors. It was not profitable to simulate the dynamics of the univariate ozone time series nonlinearly. [16]

In order to prevent exposure and reduce hazardous emissions, public authorities prefer predictive algorithms to issue ground-level ozone (O3) notifications to concerned citizens and industrial organizations. On days with a minority class, model performance will be affected by the class imbalance issue (in some field data, O3-contaminated days are much less than non-polluted days). Although this work proposes a cost-sensitive classification technique for the conventional support vector classification model (S-SVC) to investigate class imbalance, the SVM air quality prediction findings are optimistic. S-SVC equals CS-SVC. Although it performs somewhat worse on non-contaminated days for two Hong Kong air monitoring sites, CS-SVC is less likely to miss O3-polluted days than S-SVC and SVR. Our uneven air quality dataset makes the traditional SVM sensitive, necessitating the use of cost-sensitive realization. Because false negatives on polluted days are far more harmful than false positives on non-contaminated days, CS-SVC, which performs better on O3-polluted days, is recommended for public health.[17]

Paper Title	**Method**	**Data**	**Performance Metric**	**Main Finding**
Brian S. et al. [11]	hybridization of long short-term memory and recurrent neural network models	Kuwait, Persian Gulf	Mean Absolute Error (MAE)	Accurately predict air pollution values while o checking key factors for real-time analysis.
Bilge Ozbay et al. [12]	(MLR) and (PCA)	Troposphere Ozone, Annual and Seasonal Trend	R2 Score	MLR model with PCA for reducing variables had a high R2 score (0.90, 0.92, 0.85) for different seasonal trends.
Ning Jin et al. [13]	Nested LSTM layer with multiple tasks multiple channels and Discrete Stationary Wavelet Transform	Beijing dataset from 12 observing stations for the UCI ML repository	R2 Score	The proposed model outperformed other models for ozone with a coefficient of determination score of 0.99.

Ebrahim et al. [14]	Convolutional Neural Network (CNN)	Seoul, South Korea, 2017	Index of Agreement (IOA), Correlation Coefficient, MAE, and RMSE	A CNN-based model with five convolutional layers and a ReLU activation function achieved an IOA of 0.87, a correlation coefficient of 0.7, and an RMSE of 12.01 for the entire 2017 year.
Kabesok Ko. et al. [15]	Machine learning with PBL height	Ozone concentration	IOA, MAE, and RMSE	Anticipated PBL height can enhance data-driven surface ozone concentration prediction models.
Uwe Schlink et al. [16]	15 different statistical techniques	Ten datasets that reflect various European meteorological and emission conditions	Success Index	The optimum compromise in the majority of cases is between neural networks and generalized additive models, while nonlinear modeling of unilabiate ozone time-series dynamics was not beneficial.

Table 1: Literature Review for Ozone Concentration Prediction

Using different approaches, datasets, and performance measures, each paper made ozone concentration predictions. To reduce data dimension and improve forecasting, the majority of papers used statistical techniques and machine learning models. According to a comparison of these studies **Table 1**, generalized additive models and neural networks provide the optimum compromise in the majority of situations. The study also highlights cost-sensitive classification techniques for real-time analysis, PBL height prediction, and class imbalance correction.

Proposed Methodology

Data Collection:

The Central Pollution Control Board (CPCB) in Delhi, India, gathered the data for this research from numerous stations there[22]. Stations like Alipur, AnandVihar, AshokVihar, AyaNagar, and Bawana, Delhi - DPCC were considered for this research as shown in below **Figure 1** on Google Satellite Image. The information was gathered on average every 24 hours throughout a six-year period, from 2018 to March 2023. The CPCB is a renowned agency in charge of keeping tabs on and managing pollution levels throughout India, and the information gathered from its stations is accurate and credible.

(Figure 1: Geographical Location of the Monitoring Stations)

Most of the stations have common meteorological features like PM2.5, PM10, NH3, SO2, CO, Ozone, Benzene, Toluene, Benzene, MP-Xylene, RH, WD, SR, BP, AT, TOT-RF, RF, and Xylene were the variables measured at this site[23]. Following **Table 2** show an in-depth picture of all the parameters. All the concentrations are measured in ug/m^3.

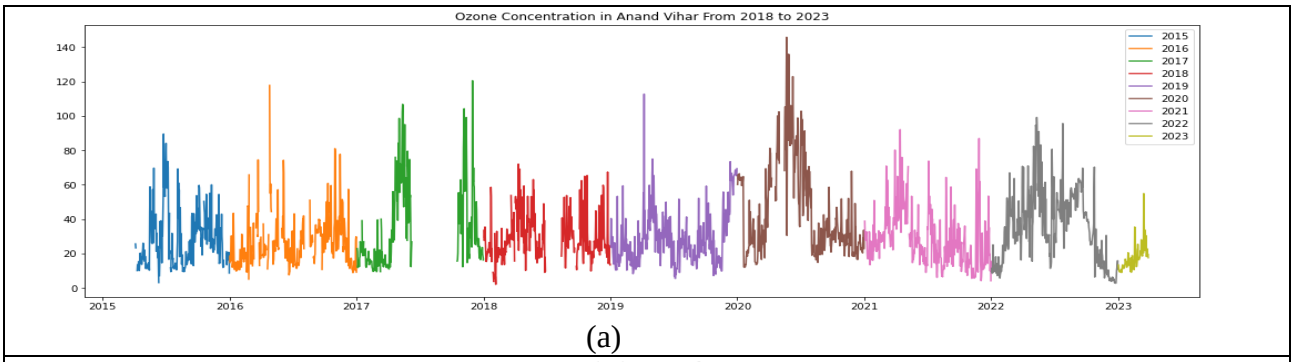
Sno.	Parameter	Description
1.	PM2.5	Particulate Matter 2.5 is a common air pollutant that can have negative health effects when inhaled.
2.	PM10	Particulate matter includes particles less than 10 µm in diameter
3.	NO,NO2,NOx	Nitrogen and its oxides are great cause of respiratory problems, (Nitrogen Oxide, dioxide and mixture of both)
4.	NH3	Ammonia, agricultural air pollutant.
5.	SO2	Sulphur Dioxide a major factor in acid rain and breathing problem
6.	CO	Carbon monoxide is a deadly and incombustible gas
7.	Ozone O3	A gas that is both a natural and man-made air pollutant. It is formed by the reaction of other pollutants in the presence of sunlight and can cause respiratory problems.
8.	Organic Compounds	Many concentration of Benzene, Xylene and Toluene compounds exists that cause neurological and respiratory problems due to pollution.
9.	Physical factors	Relative Humidity(RH), Wind Direction(WD), Solar Radiation(SR), Barometric Pressure(BP), Air Temperature(AT), Total Radiative forcing (TOT-RF)

(**Table 2**: Description of the Air Quality Parameters)

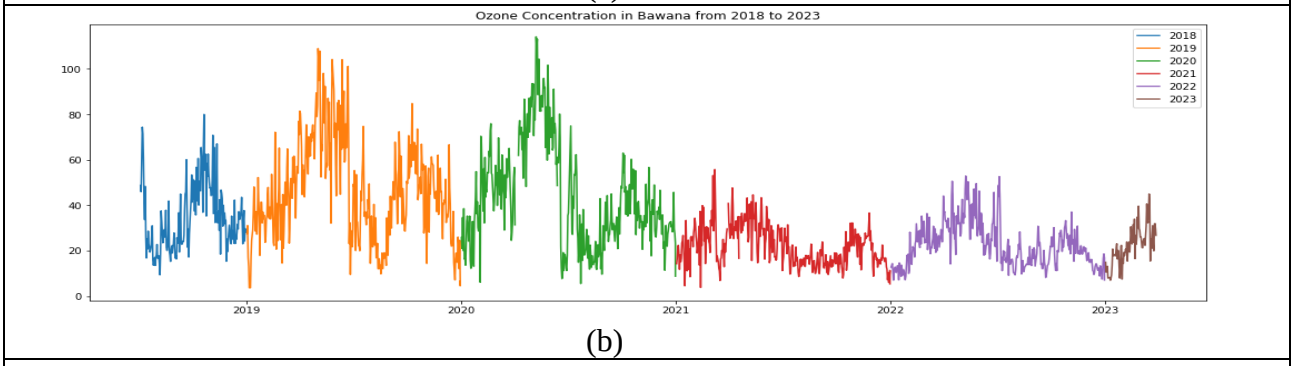
Data Preparation and Visualization

Ozone prediction requires data visualisation and preprocessing. Data visualisation helps highlight patterns and trends in the data that may not be visible from the raw figures. To prepare data for analysis, preprocessing cleans and transforms it. This might involve deleting missing numbers, dealing with outliers, and scaling the data.

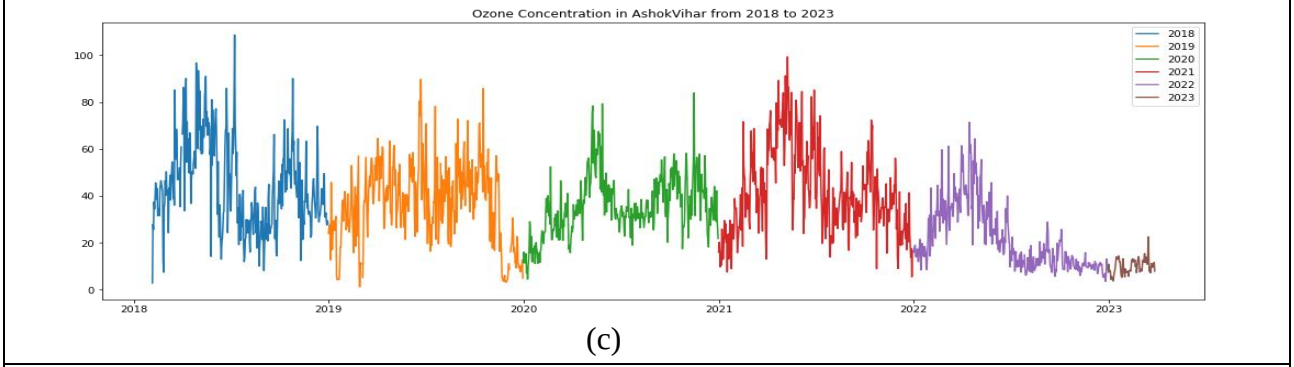
Data visualisation can discover relationships between ozone concentration, temperature, wind speed, and particulate matter levels for ozone forecast. This may help choose features and choose predictive model parameters. Pre-processing is crucial for predictive modelling. Missing values may be imputed using statistical approaches like mean or regression imputation. Visualisation can identify outliers and truncate them. Scaling data may make it simpler to compare the effects of multiple factors on ozone concentration, particularly when the parameters are recorded on different scales. [24]


Ozone Concentration in Anand Vihar From 2018 to 2023
2015
2016
2017
2018
2019
2020
2021
2022
2023

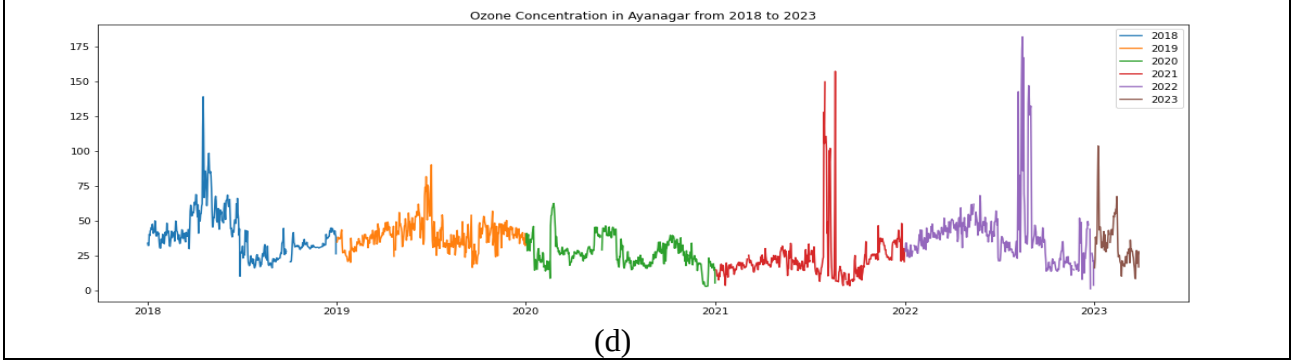
(a)


Ozone Concentration in Bawana from 2018 to 2023
2018
2019
2020
2021
2022
2023

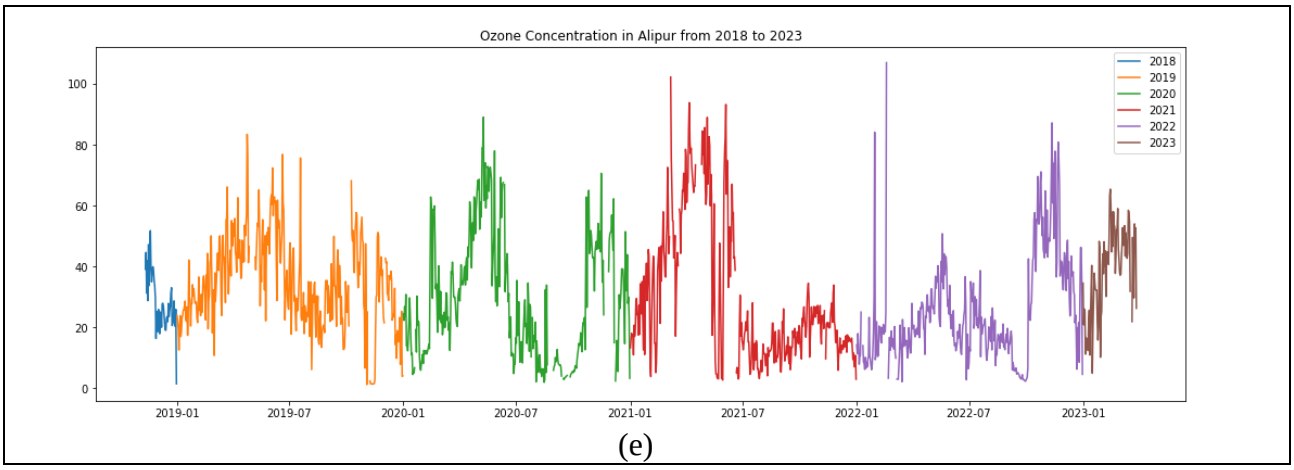
(b)


Ozone Concentration in AshokVihar from 2018 to 2023
2018
2019
2020
2021
2022
2023

(c)


Ozone Concentration in Ayanagar from 2018 to 2023
2018
2019
2020
2021
2022
2023

(d)



(e)

(**Figure 2:** Plotting Ozone concentration for all the available station data year wise

(a) Anand Vihar (b) Bawana (c) Ashok Vihar (d) Aya Nagar (e) Alipur)

In this envision the data is done to see the seasonality and trends in air quality data for ozone concentration. First, analyze air quality data seasonal patterns and trends to anticipate ozone concentration. In **Figure 2**, there has been visualization of the Ozone data according to the different locations. Anand Vihar has the highest amount of data as it was collected from 2015 to 2023. Additionally, there was some missing information in between which was not appropriate for the prediction. In the time period of 2018 to 2021, Anand Vihar had the highest concentration of 140 ug/m^3 in the summer of 2020. Most of the time there was no trend was observed in many of the station data. On the other hand, some seasonal patterns were observed. One of the observation was that in a year Ozone gradually increase with time and reach its maximum concentration then there would be a gradual decrease with time. This leads to roughly a Gaussian distribution curve but for some period of time for a particular location. Recently we observed that all the stations are indicating an increase in ozone concentration as the months passed which can be seen in **Figure 2** further from 2023.

In F**igure 2** (c), Ashok Vihar can be seen to have a regular oscillation in the range of 100 ug/m^3 to 20 ug/m^3 on average. In 2022, Aya Nagar had the highest concentration reaching 175 ug/m^3 but there was no helpful information that can be extracted throughout the year. Overall, these visualizations conclude that there is no fixed pattern that can be useful for predicting ozone concentration in the future. Howsoever, small intervals can be used to predict future concentration like using the past month's data to predict two-three days forecast.

Data Preparation is another important part of machine learning. In this case, Data was collected according to different years. While training, it was really difficult to accumulate most of the data from Excel sheets into one combined spreadsheet for a particular location. Most of the research time went into the preparation and processing of data. Many Data processing was done so that an appropriate model can be used to predict ozone concentration. After careful observation, It can be seen that predicting Ozone concentration is a Multivariate Time Series forecasting or

regression problem. Overview of the work is shown in **Figure 3** as it depicts the flow of Research work.

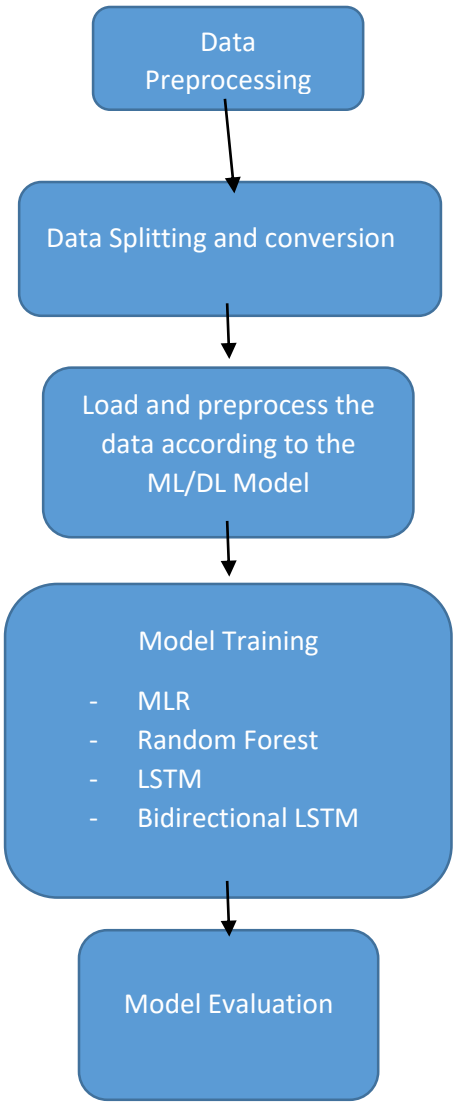


Figure 3: Flowchart of Ozone concentration prediction model

Data Pre-processing Techniques:

The following passage will discuss all the pre-processing that has been done to make data fit for training as shown in flowchart **Fig 3**. Many techniques were used as follows:

1. Indexing Time Series Column: In the available data, there were two date columns (From and To Date) that need to be converted into date time data type. However, the format was not appropriate so we made a customized function to convert the object data into date time data type. The To Date column was used for the indexing of data. This helps us to query data faster by applying simple Numpy SQL-like query and also eases the visualization.

2. Replacing Garbage values: In the real data, the unknown values were indicated using strings like “None” or “-”. This string was replaced with Numpy NAN value so this could tell us the total count of the missing information in the data which needs to be solved.
3. Filling Missing Data: The main issue of data processing is handling missing data can arise from instrument failure or human error and can negatively impact the accuracy of forecasting algorithms. To ensure a fair inter-comparison between different forecasting techniques, it was necessary to systematically replace missing data to create a harmonized and uniform time-series data set. To accomplish this, a mean value of the past observation was used to ensure that it fits well with other data.
4. Removing Correlated Values: Values with Zero correlation or dependencies were removed as they do not help to improve model efficiency. It also takes more training time and makes the model complex. TOT-RF or Organic Compounds like Benzene were having very less correlation. Here Pearson Correlation was used which measures the strength of the linear bond between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.
5. Splitting Data into Train and Test: In our research work, data was collected by 24 hrs durations so it has around 1800 observations for a location. For training purposes, we divided the data in an 80:20 ratio where 80 percent data was used for training and
6. Transformation of Data: Various Machine Learning and Deep Learning Model requires a fixed shape of input so to use these models. As an example, CNN requires 3 dimensional Shape. The main focus was to transform data for LSTM and Bidirectional LSTM. Various methods like Min-Max Standardisation and Normalization were used. For time series forecasting we need to convert the multiple variable data frames into sequences. This is done by splitting a sequence into input-output pairs which is useful for creating input and final target data for time series forecasting. This array of sequences is made using time steps. Time steps mean the number of steps to look back to predict future outcomes and average those previous steps into one.

In this study, data pre-processing makes it easier to load data into a machine-learning model. Here we focus on deep learning models like LSTM and Bi-directional LSTM with a comparison with traditional ML models.

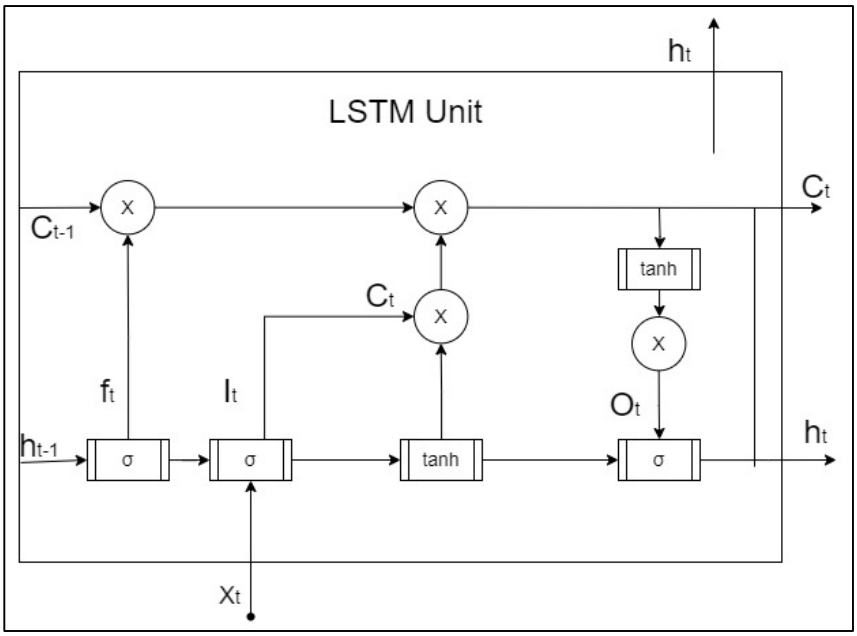
Prediction Models/ Methodology/Model Training:

The following models are used for our research purpose:

- Multi Linear Regression: As a linear model, it is a statistical regression algorithm used to model the relationship between numerous independent variables and dependent variables by fitting a linear equation to observed data. It validates a simultaneous statistical relation between the single continuous outcome Y and the independent variables Xi as shown in **equation 1**. [25]

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} + \in_i \quad \textbf{(1)}$$

- Random Forest Regression: It is an ensemble algorithm with the potential of performing both regression and classification tasks with the use of multiple decision trees. It has additional support through the bagging method. It is a voting-type mechanism where all the output from the tress is considered for the final output. It is an estimator that employs bootstrap and aggregation to increase predicted accuracy and reduce overfitting.
- Long Short-Term Memory (LSTM): In 1997, Hochreiter and Schmidhuber proposed a recurrent neural network model which was capable of taking the output of the previous moment as the input of the next moment to affect the weights at the next moment without causing vanishing gradient problem. This is the reason for state of the art performance of LSTM. Similar to ANN and RNN, this neural network consists of a series of hidden layers with input and outpuvariablesle. The main differentiator for LSTM is the ability to compose more than one inter-recurrent memory block which forward passes the value into the preserved block and subsequently recovered at a vital rate. [26]
  The architecture of LSTM is explained further with the following **Figure 4**.



(**Figure 4** Architecture of LSTM Unit)

The main components of an LSTM unit are forget gate ft, the output gate Ot, the input gate It, and the cell state node Ct. All the gates work together to regulate the cell state. The amount of information to be dropped is measured by forget gate using a sigmoid-based output layer. Here σ is the sigmoid function. On the other hand, Input gate decides how many new features can move into the cell. It does two function. Firstly, it decides the use of sigmoid to update necessary values and creates new vectors using tanh function. Secondly, It combines the output from forget gate to update the state of the cell. Output gate controls the information that handles the cell state at a particular movement and multiplies the state with the sigmoid output.[27]

- Bidirectional LSTM: It has capabilities of traditional LSTM by including data from both past and future values. It process the input array by beginning to end and end to beginning which helps to effectively extract more complete historical patterns. [28]

These models help us to forecast future value based on historical data. This models can be used by researcher and practitioner to achieve optimal performance on their specific time series data.

Result Analysis

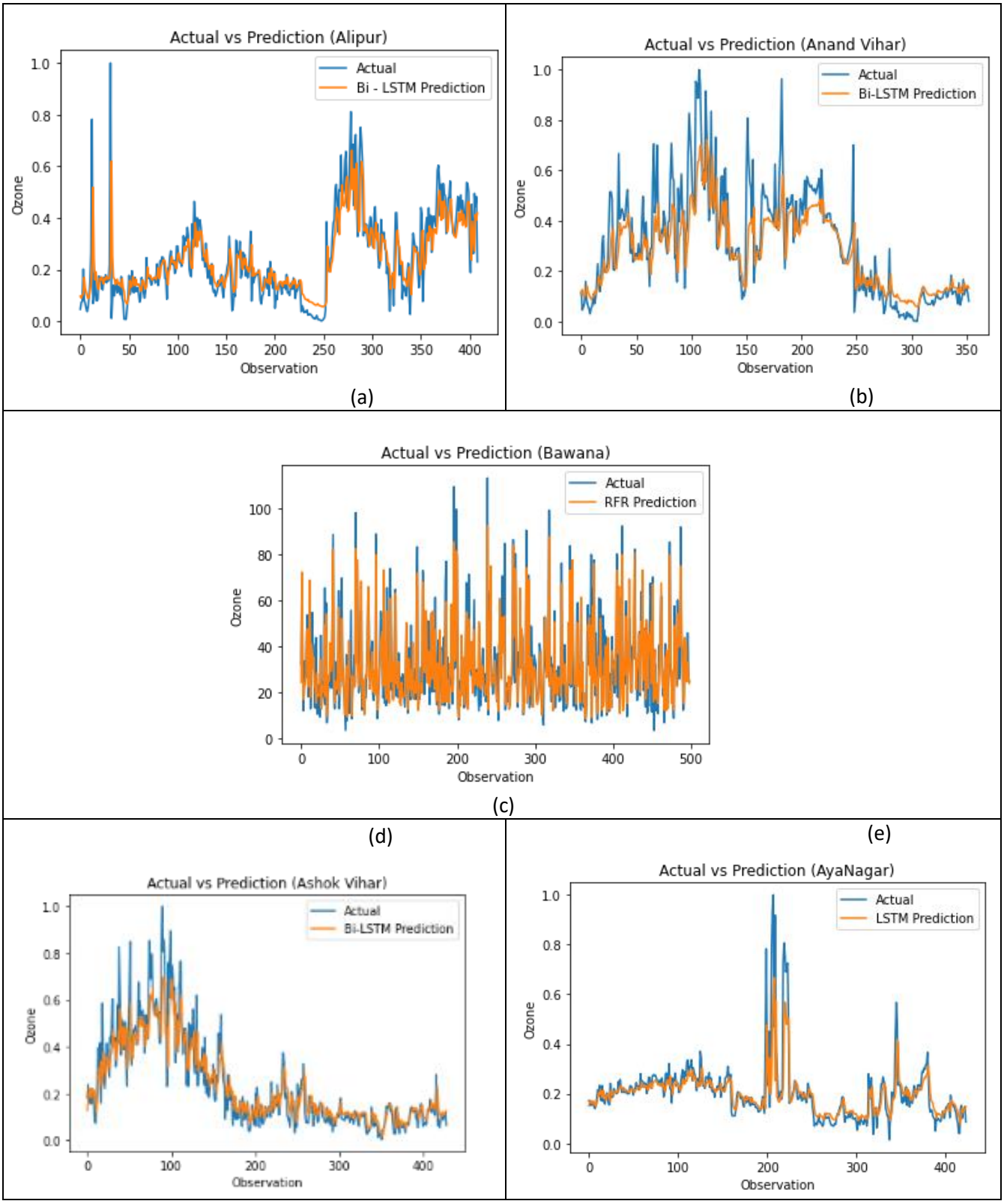
The LSTM model was developed using 64 LSTM units, 2 hidden layers with 32 units each, a 'tanh' activation function and a 20% dropout rate. The bidirectional LSTM model was developed with 64 LSTM units, 50 dense units with 'tanh' activation function and optimized using Adam optimizer with learning rate of 0.001. The some results indicate that the bidirectional LSTM model performed slightly better than the LSTM model. The model performance was measured using regression metrics like Root Mean Square Error, Mean Square Error and R2 Score[32]. All th deep learning model were run for 50 epochs.

The **Table 3** is showing the performance of 4 models with respect to the location. Starting with Alipur data was collected from 2018 to 2023 so had about 1700 data points and was split into train and test for model training. The Best model for this data was Bi directional LSTM. It achieved an R2 score of 0.64 with LSTM as second best model. For different location, data varies with the different parameters range. So the highest efficient result was found in Ashok Vihar Data. The coefficient of determination was 0.79, MSE is 0.00803 and RMSE is 0.08962.

Additionally, Machine learning model also perform good in Bawana Region as it has 0.75 R2 score. LSTM and Bi-LSTM had quite low MSE and RMASE but reasonable R2 score. Random Forest outperforms MLR. In the Anand vihar region, deep learning model performed better than MLR and RFR. Aya Nagar had poor results for MLR model as it has higher MSE and RMSE. Here LSTM model achieved moderate results.

Overall, across different location, the LSTM and Bi-LSTM models outperformed MLR and RFR models in comparision. In addition to lower MSE and RMSE values, it indicated the model has high accuracy predicting ozone concentration. MLR and RFR generally perform less accurately as they have high loss value and lower R2 score. This conclude the state of the art mechanism of neural network based models, specifically LSTM and Bi-LSTM, for the given time series analysis.

We also included graphs to visually represent the performance of each model. These graphs provide a clear comparison of the predicted values versus the actual values, allowing for a more in-depth analysis of the model's performance which can be seen in below. Only the best models graphs has been displayed as a purpose for the visualization as shown in **Figure 5**.



(Figure 5: Prediction vs Actual visualization of the best models as per the location: (a) Bi-LSTM Alipur (b) Bi-LSTM Anand vihar (c) Random Forest Regressor Bawana (d) Bi-LSTM Ashok Vihar (e) LSTM Aya Nagar)

S No.	Location	Model	Model Metrics		
1.	Alipur		**MSE**	**RMSE**	**R2**
		LSTM	0.00993	0.09965	0.63289
		Bi-LSTM	0.00965	0.09824	0.64320
		MLR	259.99898	16.12448	0.27389
		RFR	158.15804	12.57609	0.55830
2.	Ashok Vihar		**MSE**	**RMSE**	**R2**
		LSTM	0.00919	0.09585	0.76501
		Bi- LSTM	0.00803	0.08962	0.79455
		MLR	124.85358	11.17379	0.57166
		RFR	74.13415	8.61012	0.74566
3.	Anand Vihar		**MSE**	**RMSE**	**R2**
		LSTM	0.01542	0.12418	0.66485
		Bi-LSTM	0.01392	0.11800	0.70037
		MLR	239.294	15.46915	0.26656
		RFR	158.15804	12.57609	0.55830
4.	Aya Nagar		**MSE**	**RMSE**	**R2**
		LSTM	0.00590	0.07683	0.62152
		Bi- LSTM	0.00624	0.07898	0.60004
		MLR	303.85658	17.43148	- 0.03498
		RFR	137.96394	11.74581	0.53007
5.	Bawana		**MSE**	**RMSE**	**R2**
		LSTM	0.01538	0.12401	0.62847
		Bi-LSTM	0.01531	0.12372	0.63017
		MLR	173.86771	13.18589	0.55108
		RFR	94.06914	9.69892	0.75712

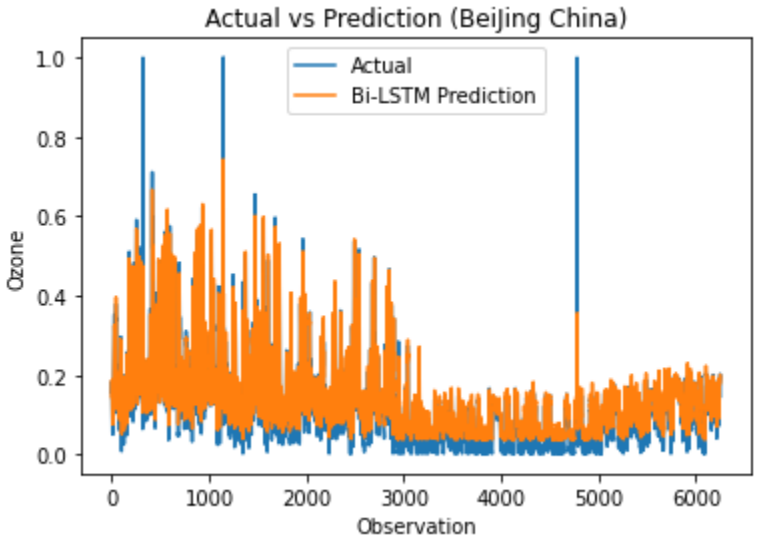
(**Table 3 :** Performance of all four models measured using regression metrics as per locations)

From **Table 3** it is clearly observed that the Deep learning model are more successfully forecasting Ozone Data. Our approach was also validated on the famous open source time series data. The UCI repository data of the Beijing Monitoring station[29-31] were taken into account to validate our deep learning models. This method was used to validate our model which had a good result as shown in **Table 4**

Model	**Model Metrics**		
	MSE	**RMSE**	**R2**
LSTM	0.00157	0.03962	0.86892
Bi-LSTM	0.00133	0.03654	0.88850
MLR	1352.40516	36.77506	0.53304
RFR	576.40043	24.00834	0.80098

Table 4: Performance of our proposed on Beijing China Dataset for Air Quality

The **table 4** display the performance of different models for forecasting Ozone Concentration on the Beijing china Dataset, specifically focusing on the metrics of MSE, RMSE and coefficient of determination. Among the evaluated models, LSTM and Bidirectional LSTM had achieved great efficiency compared to MLR and RFR. Both neural network accomplished lower MSE and RMSE score, indicating the ability to make more accurate prediction about Ozone. Higher R2 score implies the good fitting of model on the air quality data and extracting most amount of variance in the Ozone concentration. **Figure 6** shows the performance of BI-LSTM model with observed values and predicted values of ozone to predict the 20 percent of test data.



(**Figure 6**: Bi-LSTM performance on Beijing China dataset)

Conclusion

In this paper, Deep learning models were applied on real time collected air quality data from various station of New Delhi, India. The models were compared with Machine learning models and also with Beijing China Dataset from UCI repository. LSTM models were trained with proper training time of 50 epochs and were cross validated with testing data. The aim of this study was to develop a prediction model for Ozone forecasting based on data provided by an air quality stations which collected data in a time interval of a day or 24 hrs. Due to 24hrs, there were less data available for training the model on Indian data as they have approximately 1800 – 2500 values only. One reason of getting moderate R2 score for deep learning models is less available data. This can be witness on Beijing China data. The China data is collect every 1hr, resulting into large amount of data. The Ozone concentration for China data is Dense and has about 30000 data points for time period 2017 to 2020 data. Another factor affecting the model accuracy can be missing values. As it was real time data, it had many missing information and many zero correlated model. Though all the hurdles, model achieve good accuracy for few of the location like Ashok Vihar. So far Bi-LSTM model is best for predicting O3 concentration as it will help organization to formulate better schemes and decision making for the environment.

For future purpose, there is need for large data points for Indian cities so sequential model like LSTM and Bi-LSTM, can perform well and predict the future values more accurately by memorizing the trends effectively. Also more complex neural network like Auto encoders and Nested or Hybrid Models can used for more statistical prediction.

Reference

1. Tree-based ensemble deep learning model for spatiotemporal surface ozone (O3) prediction and interpretation Zhou Zang
2. Regional prediction of ground-level ozone using a hybrid sequence-tosequence deep learning approach
3. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone
4. Seasonal ground level ozone prediction using multiple linear regression (MLR) mode
5. Hybrid deep learning model for ozone concentration prediction: comprehensive evaluation and comparison with various machine and deep learning algorithms
6. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone
7. Long time series ozone prediction in China: A novel dynamic spatiotemporal deep learning approach
8. Hourly ozone prediction for a 24-h horizon using neural networks
9. Balaguer Ballester, E., Camps i Valls, G., Carrasco-Rodriguez, J.L., Soria Olivas, E., del Valle-Tascon, S., 2002. Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks. Ecological Modelling 156, 27–41.
10. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework

11. Brian S. Freeman, Graham Taylor, Bahram Gharabaghi & Jesse Thé (2018) Forecasting air quality time series using deep learning, Journal of the Air & Waste Management Association, 68:8, 866-886, DOI: 10.1080/10962247.2018.1459956
12. Bilge Özbay, Gülşen Aydın Keskin, Şenay Çetin Doğruparmak, Savaş Ayberk,Multivariate methods for ground-level ozone modeling,Atmospheric Research, Volume 102, Issues 1–2, 2011, Pages 57-65, ISSN 0169-8095, https://doi.org/10.1016/j.atmosres.2011.06.005. (https://www.sciencedirect.com/science/article/pii/S0169809511001839)
13. N. Jin, Y. Zeng, K. Yan and Z. Ji, "Multivariate Air Quality Forecasting With Nested Long Short Term Memory Neural Network," in IEEE Transactions on Industrial Informatics, vol. 17, no. 12, pp. 8514-8522, Dec. 2021, doi: 10.1109/TII.2021.3065425.
14. A real-time hourly ozone prediction system using deep convolutional neural network Ebrahim Eslami, Yunsoo Choi*, Yannic Lops, Alqamah Sayeed Department of Earth and Atmospheric Sciences, University of Houston, TX 77004 *corresponding author, ychoi23@central.uh.edu
15. : Ko, K.; Cho, S.; Rao, R.R. Machine-Learning-Based Near-Surface Ozone Forecasting Model with Planetary Boundary Layer Information. Sensors 2022, 22, 7864. https://doi.org/10.3390/ s22207864
16. . Uwe Schlink, Stephen Dorling, Emil Pelikan, Giuseppe Nunnari, Gavin Cawley, Heikki Junninen, Alison Greig, Rob Foxall, Krystof Eben, Tim Chatterton, Jiri Vondracek, Matthias Richter, Michal Dostal, Libero Bertucco, Mikko Kolehmainen, Martin Doyle,A rigorous inter-comparison of ground-level ozone predictions,Atmospheric Environment,Volume 37, Issue 23,2003,Pages 3237-3253,ISSN 1352-2310,
17. https://doi.org/10.1016/S1352-2310(03)00330-3.
18. Wei-Zhen Lu, Dong Wang, Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme, Science of The Total Environment,Volume 395, Issues2–3,2008,Pages109-116,ISSN0048-9697,https://doi.org/10.1016/j.scitotenv.2008.01.035.
19. Multivariate methods for ground-level ozone modeling Bilge Özbay a, ∗, Gülşen Aydın Keskin b, Şenay Çetin Doğruparmak a, Savaş Ayberk aa Department of Environmental Engineering, Kocaeli University, 41380 Kocaeli, Turkey b Department of Industrial Engineering, Kocaeli University, 41380 Kocaeli, Turkey
20. Eberly, L.E. (2007). Multiple Linear Regression. In: Ambrosius, W.T. (eds) Topics in Biostatistics. Methods in Molecular Biology™, vol 404. Humana Press. https://doi.org/10.1007/978-1-59745-530-5_9
21. Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change, Government of India. (n.d.). Retrieved from http://cpcb.nic.in/
22. Gong, C. and Liao, H.: A typical weather pattern for ozone pollution events in North China, Atmos. Chem. Phys., 19, 13725–13740, https://doi.org/10.5194/acp-19-13725-2019, 2019.
23. Zhang, A., Fu, T.-M., Feng, X., Guo, J., Liu, C., Chen, J., et al. (2023). Deep learning-based ensemble forecasts and predictability assessments for surface ozone pollution. *Geophysical Research Letters*, 50, e2022GL102611. https://doi.org/10.1029/2022GL102611

24. E. Debry, V. Mallet,Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10 on the Prev'Air platform,Atmospheric Environment,Volume 91,2014,Pages 71-84,ISSN 1352-2310,https://doi.org/10.1016/j.atmosenv.2014.03.049.
25. Vance, Jesse M., et al. "An empirical MLR for estimating surface layer DIC and a comparative assessment to other gap-filling techniques for ocean carbon time series." Biogeosciences 19.1 (2022): 241-269.
26. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks Ralf C. Staudemeyer, Eric Rothstein Morris https://doi.org/10.48550/arXiv.1909.09586
27. Federico Landi, Lorenzo Baraldi, Marcella Cornia, Rita Cucchiara,Working Memory Connections for LSTM,Neural Networks,Volume 144,2021,Pages 334-341,ISSN 0893-6080,https://doi.org/10.1016/j.neunet.2021.08.030.
28. S. Siami-Namini, N. Tavakoli and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 3285-3292, doi: 10.1109/BigData47090.2019.9005997.
29. Xie, Y., Dai, H., Zhang, Y., Hanaoka, T., & Masui, T. (2017).Health and economic impacts of ozone pollution in China:A provincial level analysis. Health and Economic Impactsof Ozone Pollution in China: A Provincial Level Analysis,130(104881), 1–63. https://doi.org/10.5194/acp-2017-849.
30. Yafouz, A., Ahmed, A. N., Zaini, N., & El-shafie, A. (2021).Ozone concentration forecasting based on artificial intelligence techniques: A systematic review. Water, Air, & Soil Pollution, 232(2). https://doi.org/10.1007/s11270-021-04989-5.
31. Zhou, Y., Chang, F., Chang, L., Kao, I., & Wang, Y. (2019).Explore a deep learning multi-output neural network forregional multi-step-ahead air quality forecasts. Journal ofCleaner Production, 209, 134–145. https://doi.org/10.1016/j.jclepro.2018.10.24
32. Botchkarev, Alexei. "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology." arXiv preprint arXiv:1809.03006 (2018).