PANDIT DEENDAYAL ENERGY UNIVERSITY SCHOOL OF

TECHNOLOGY

Course: **Artificial Intelligence Lab**

“Air Quality Analysis and Ozone Prediction using ML and DL Techniques,”

PROJECT REPORT

B.Tech. (Computer Science and Engineering)

Semester 6

Submitted By:

Sarthak Kapaliya

20BCP072

Abstract

This project aimed to forecast ozone concentration using deep learning and traditional machine learning models. Two deep learning models, LSTM and bidirectional LSTM, were developed, and their performances were compared with two traditional machine learning models, MLR and Random Forest Regressor. The LSTM model achieved a mean squared error (MSE) of 0.00993, root mean squared error (RMSE) of 0.09965, and R2 value of 0.63289. The bidirectional LSTM model achieved an MSE of 0.00965, RMSE of 0.09824, and R2 value of 0.64320, outperforming the LSTM model. The MLR achieved an MSE of 259.99898, RMSE of 16.12448, and R2 value of 0.27389, while the Random Forest Regressor achieved an MSE of 158.15804, RMSE of 12.57609, and R2 value of 0.55830. 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.

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 of 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 because 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.

Several technologies can be used to figure out how much ozone is in the air. There are two main ways to figure out how much pollution is in the air: numerical methods and data-driven methods. The Goddard Earth Observing System with Chemistry, the Weather Research and Forecasting Model with Chemistry, and the Community Multistage Air Quality Model are all used as numerical models. This model allows for clear and strict logic and a strong ability to explain. This also made it hard to make long-term predictions because the cost of computing was high and it was hard to get enough data. By taking the valuable information from a vast quantity of data about the target variable, data-driven models are a potential technique to create precise predictions about the target variable. Data-driven models may be divided into shallow machine learning models and deep learning models, according to several earlier research. 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 as technology improves, new ways are being made to improve ozone monitoring and forecasting. For example, AI and machine learning algorithms are used to make predictions about ozone concentrations more accurate. These technologies can look at a lot of data from many different sources, like ground-based monitoring stations, satellites, and weather models, to make ozone concentration maps that are more accurate and complete. This can help figure out where there is a lot of ozone pollution and guide policy decisions that aim to cut ozone emissions. [7]

Machine learning is extensively used as an empirical method to forecast the ozone concentration. Throughout the time, various ensemble methods have been used. Support vector machines and random forests are some models that have been used in research. Modelling the ozone's fluctuations and making accurate forecasts are two of the most crucial duties for the researchers. There are both deterministic and statistical (black box) models. A lot of physical and chemical interactions between predictor variables must be taken into account during the relatively complex process of using partial differential equations to create a deterministic model to forecast ozone concentrations in a specific area. This process also necessitates a lot of accurate input data (such as emissions, meteorology, 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.

Artificial neural networks are also used to predict pollutants like particulate matter, sulphur dioxide, etc. This paper compares three predictive models: the autoregressive-moving average with exogenous inputs (ARMAX), multilayer perceptron, and the finite impulse response (FIR) neural network. The goal is to figure out the hourly ozone concentration 24 hours in advance. They looked at the highest levels of ozone in the summer between 1996 and 1999 at three Spanish monitoring sites in cities and towns. The MLP neural networks performed better than the linear ARMAX models, which performed better than the dynamic FIR neural networks, based on the five performance criteria that were used. [10]

In this paper, we suggest analysing Indian air quality data using ozone concentration data obtained from a 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 kernel of fixed size. The small changes have resulted in 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 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 correlation coefficient of 0.7. the Root mean square error was 12.01 for entire 2017 year.[14]

Kabesok Ko.et al. has worked on a dataset about ozone concentration. The dataset used here contain 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 paper, Uwe Schlink et al. provides report onn comaparatice analysis of models in which 15 different statistical techniques for ozone forecasting were applied to ten dataset representing different meteorological and emission conditions throughout Europe. They also attempt to compare the performance of the statistical techniques. The relative evaluation of forecasting performance (benchmarking) produced 1340 yearly time series of daily predictions and the results are described in terms of predefined performance indices. Through analysing associations between the performance indices, we found that the success index is of outstanding significance. The 8-h average ozone concentration forecast accuracy was found to be superior to the 1-h mean ozone concentration forecast, which makes the former very significant for operational forecasting. The best forecasts were achieved for sites located in rural and suburban areas in Central Europe unaffected by extreme emissions (e.g. from industries) Results show that a technique might be good but bad. Since they can manage nonlinear linkages and adapt to site-specific variables, neural network and generalised additive models are the best compromise for most scenarios. Nonlinear modelling of univariate ozone time-series dynamics was unprofitable. [16]

Public authorities prefer predictive models to give ground-level ozone (O3) alerts so concerned residents and industrial groups may avoid exposure and minimise dangerous emissions. The class imbalance problem—in some field data, O3 contaminated days are much fewer than non-polluted days—will reduce model performance on minority class days. SVM air quality prediction results are promising, however this study presents a cost-sensitive classification strategy for the traditional support vector classification model (S-SVC) to analyse class imbalance. S-SVC is CS-SVC. CS-SVC is less likely to miss O3 polluted days than S-SVC and SVR, although it performs somewhat worse on non-polluted days for two Hong Kong air monitoring locations. The conventional SVM is sensitive for our unbalanced air quality dataset, hence the cost-sensitive realisation is needed. CS-SVC, with better performance on O3 polluted days, is suggested for public health because false negatives on polluted days are far more damaging than false positives on non-polluted days.[17]

Paper Title	Method	Data	Performance Metric	Main Finding
Brian S. et al. [11]	Hybrid combination of Recurrent Neural Network and Long short Term Memory models	Kuwait, Persian Gulf	Mean Absolute Error (MAE)	Accurately predict air pollution concentration while only monitoring key factors for real-time analysis.
Bilge Ozbay et al. [12]	Multiple Linear Regression (MLR) and Principal Component Analysis (PCA)	Troposphere Ozone, Annual and Seasonal Trend	R2 Score	MLR model with PCA for reducing variables had high R2 score (0.90, 0.92, 0.85) for different seasonal trends.
Ning Jin et al. [13]	Nested LSTM layer with multiple task multiple channels and Discrete Stationary Wavelet Transform	Beijing dataset from 12 observing station for UCI ML repository	R2 Score	Proposed model outperformed other models for ozone with an R2 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	CNN-based model with five convolutional layers and ReLU activation function achieved an IOA of 0.87, correlation coefficient of 0.7, and 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 representing different meteorological and emission conditions throughout Europe	Success Index	Neural network and generalized additive models are the best compromise for most scenarios, while nonlinear modeling of univariate ozone time-series dynamics was unprofitable.

Table 1 Literature Review for Ozone Concentration prediciton

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**, generalised 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

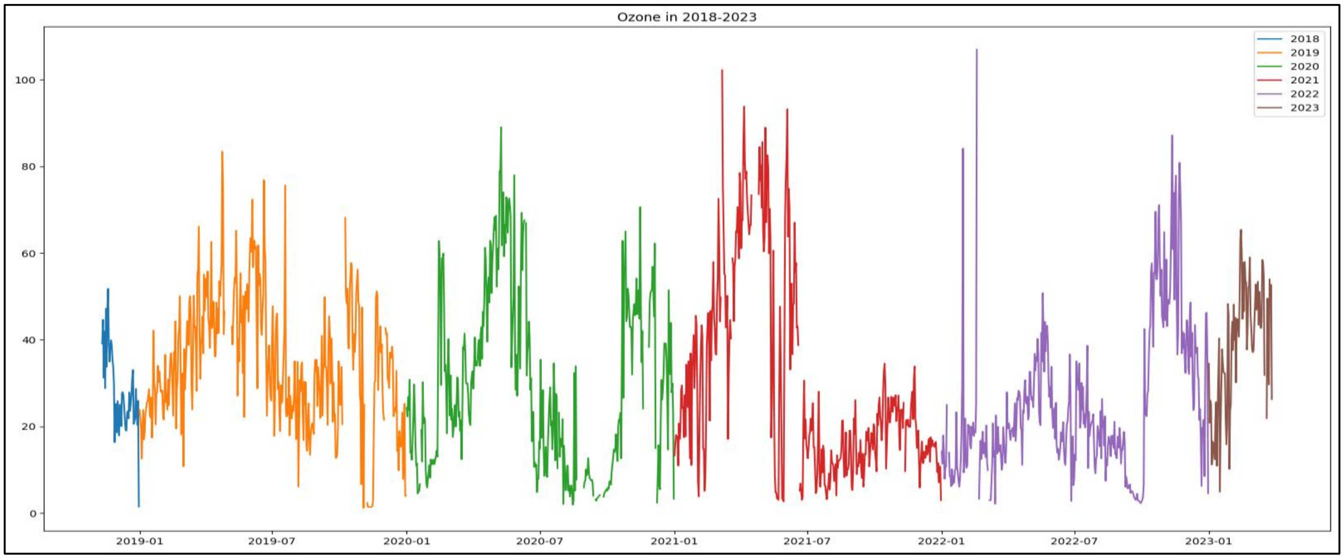
Data Collection:

The Central Pollution Control Board (CPCB) in Delhi, India, gathered the data for this research from numerous stations there. Stations like Alipur, AnandVihar, AshokVihar, AyaNagar, and Bawana Delhi - DPCC were considered for this research. PM2.5, PM10, NO, NO2, NOx, NH3, SO2, CO, Ozone, Benzene, Toluene, Eth-Benzene, MP-Xylene, RH, WD, SR, BP, AT, TOT-RF, RF, and Xylene were the variables measured at these sites. The information was gathered on average every 24 hours throughout a six-year period, from 2017 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.

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.



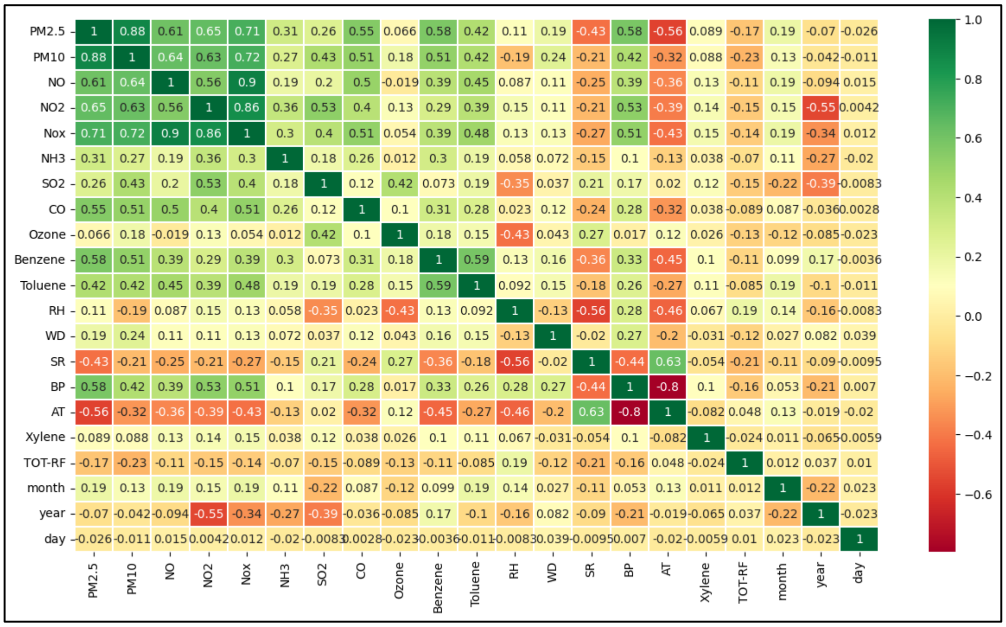


Figure 2 Correlation Matrix of all the air quality Features

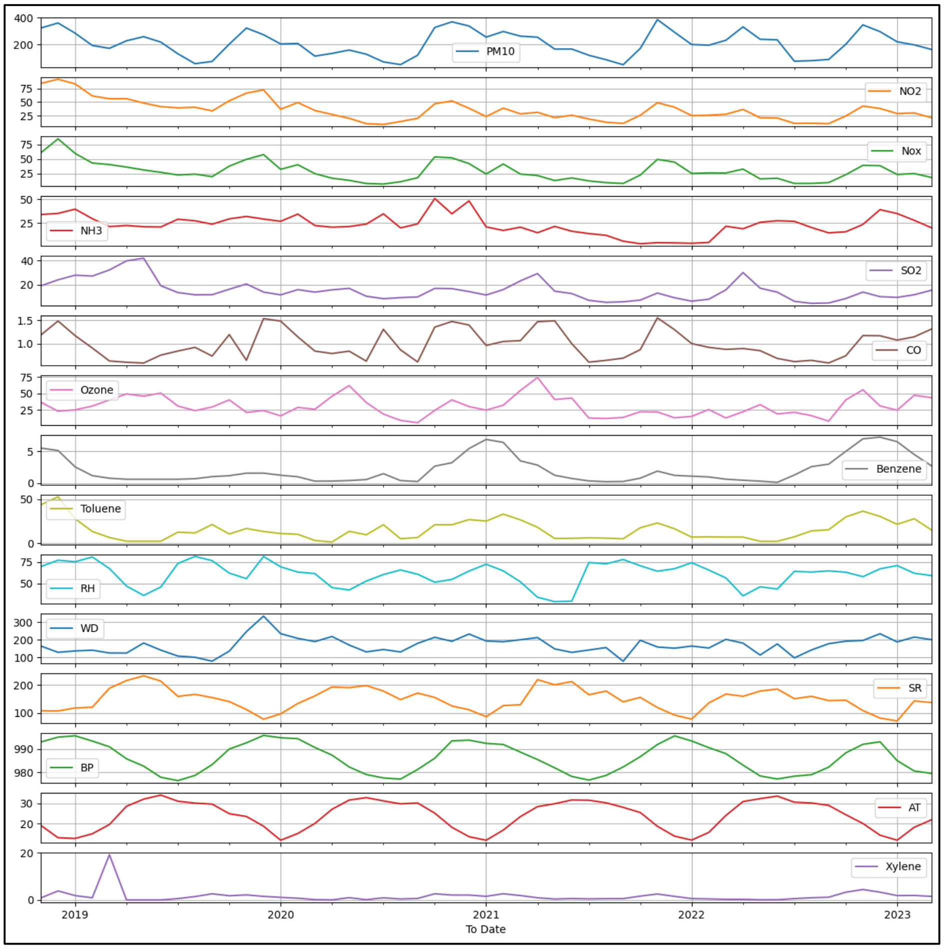
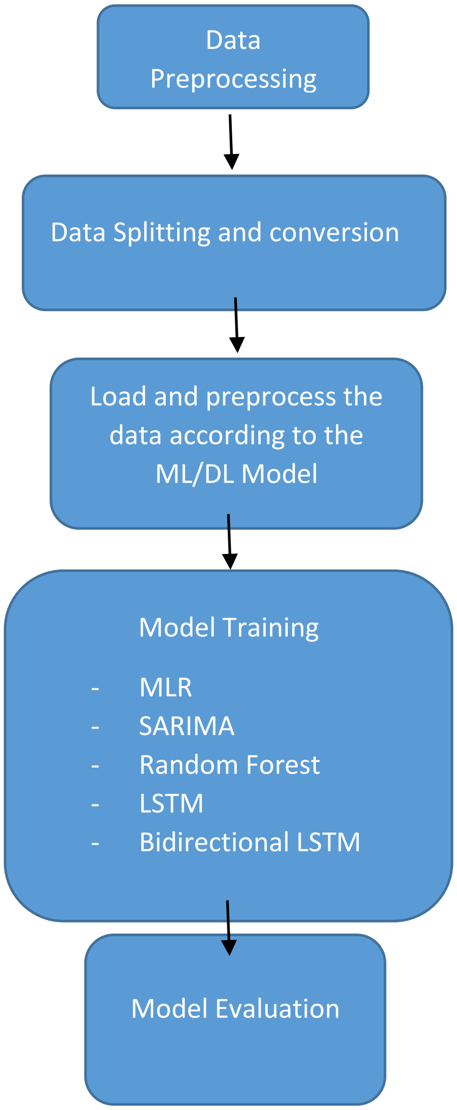


Figure 2: Monthly Representation of different features of data collected from Monitoring Station

In this we have visualize to see the seaonality and trends in air quality data forecasting for ozone concentration. First, analyse air quality data seasonal patterns and trends to anticipate ozone concentration. Line graphs, scatter plots, and histograms may help. We can identify seasonal, weather, or other patterns by graphing pollutant and meteorological variable concentrations versus time. In summer, sunshine and heat may speed up chemical processes that generate ozone, which may increase its concentration. Due to emissions, laws, and other causes, concentration may gradually rise or fall. We can create more accurate forecasting models by studying these patterns and trends. Cleaning, normalisation, and feature engineering may be needed to prepare data for analysis and modelling.


Data Preprocessing

Data Splitting and conversion

Load and preprocess the data according to the ML/DL Model

Model Training

- MLR
- SARIMA
- Random Forest
- LSTM
- Bidirectional LSTM

Model Evaluation

Figure 3: Flowchart of Prediction of Ozone concentration prediction model

This paragraph discusses the issue of missing data in large empirical data sets and the approach taken to address this issue in the context of ozone concentration forecasting. 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 hybrid model was used where short gaps in the data were filled using linear interpolation, which is a relatively simple mathematical method for estimating missing values by drawing a straight line between the two known data points. Longer gaps in the data were filled using the more advanced method of self-organizing maps (SOM), which is a type of artificial neural network that can identify and predict patterns in the data To determine the length of gap that could be replaced using linear interpolation, an index reflecting the persistence of a variable was used. This index can be thought of as a measure of how much a given variable changes over time. If a variable is highly persistent, it means that it changes relatively little over time, making it easier to estimate missing values using linear interpolation. By contrast, variables with low persistence are more difficult to estimate using linear interpolation and may require the use of more advanced techniques such as SOM.

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 model achieved a mean squared error (MSE) of 0.00993, root mean squared error (RMSE) of 0.09965, and R2 value of 0.63289. 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 model achieved an MSE of 0.00965, RMSE of 0.09824, and R2 value of 0.64320. The results indicate that the bidirectional LSTM model performed slightly better than the LSTM model. As shown in **Table 2.**

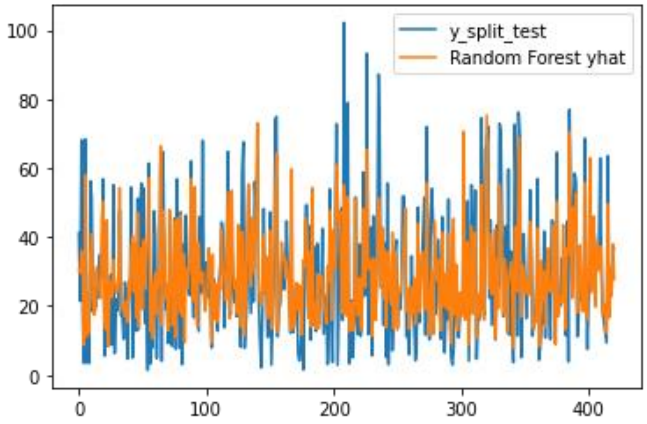
In comparison to traditional machine learning models, the MLR achieved an MSE of 259.99898, RMSE of 16.12448, and R2 value of 0.27389. The Random Forest Regressor achieved an MSE of 158.15804, RMSE of 12.57609, and R2 value of 0.55830. The deep learning models outperformed the traditional machine learning models, indicating the suitability of deep learning models for forecasting ozone concentration.

Model	**MSE**	**RMSE**	**R2**
LSTM	0.00993	0.09965	0.63289
Bidirectional LSTM	0.00965	0.09824	0.64320
MLR	259.99898	16.12448	0.27389
Random Forest Regressor	158.15804	12.57609	0.55830

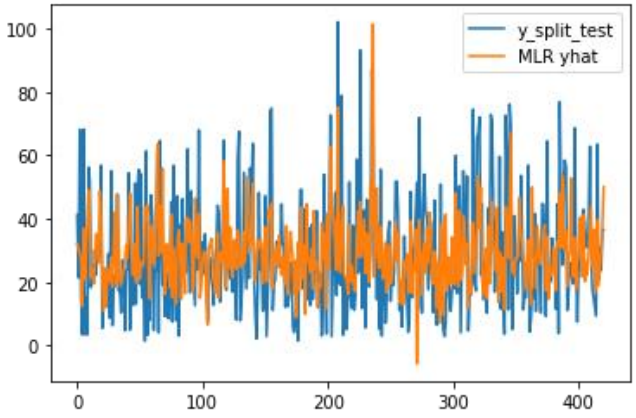
Table 2: Results evaluation of different models

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 accuracy and precision.

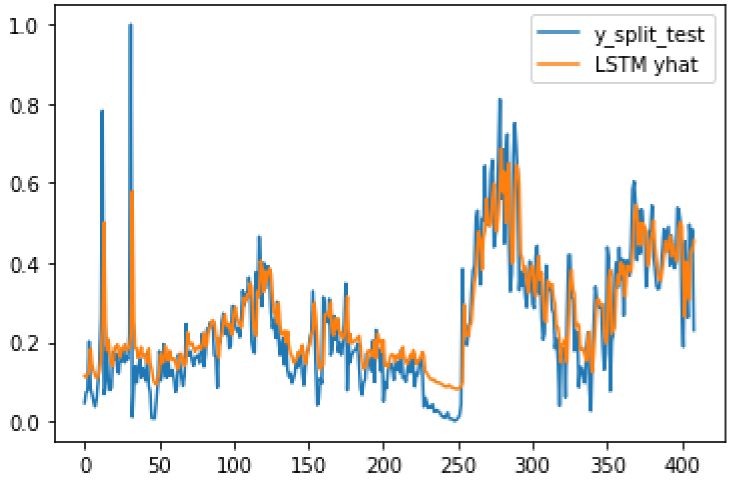
Random Forest



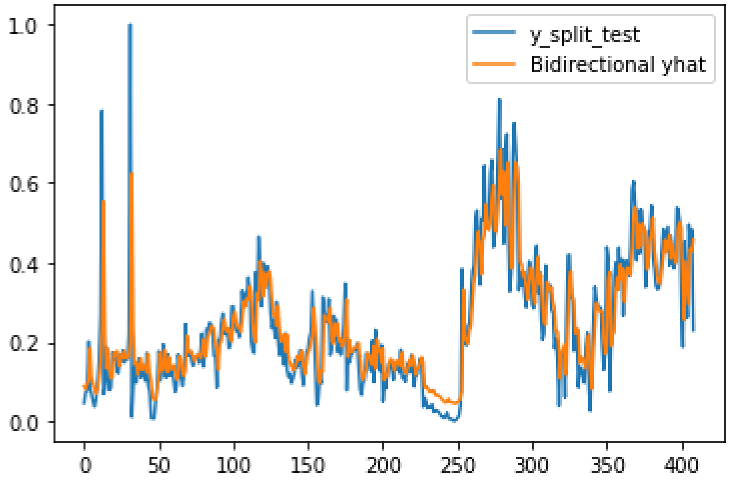
MLR



LSTM



Bidirectional LSTM



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. Eslami, E., Choi, Y., Lops, Y. et al. A real-time hourly ozone prediction system using deep convolutional neural network. Neural Comput & Applic 32, 8783–8797 (2020). https://doi.org/10.1007/s00521-019-04282-x