Environmental Pollution 315 (2022) 120404

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Air Quality Index prediction using an effective hybrid deep learning model☆

Nairita Sarkar, Rajan Gupta, Pankaj Kumar Keserwani *, Mahesh Chandra Govil

Computer Science and Engineering Department, National Institute of Technology Sikkim, South Sikkim, Ravangla, Sikkim, India

ARTICLE INFO

Keywords:
Air quality index
Long short term memory
Gated recurrent unit
Linear regression
K-nearest neighbor
Support vector machine

ABSTRACT

Environmentalism has become an intrinsic part of everyday life. One of the greatest challenge to the environment's long-term existence is the air pollution. Delhi, the capital of India, has experienced decreasing of air quality for several years. The poor air quality has a significant impact on the lives of individuals. Air Quality Index (AQI) prediction can help to its beneficiaries in taking safeguards about their health before moving to any polluted area. In this study, a variety of data forecasting approaches is evaluated to predict the AQI value for Particulate Matter ($PM_{2.5}$) μm at a particular area of Delhi and several error-prone strategies such as R-Squared (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) methods are catalogued. In the proposed approach two deep learning models like Long-Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are combined to predict the AQI of the environment. Several stand alone machine learning (ML) and deep learning (DL) models such as LSTM, Linear-Regression (LR), GRU, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are also trained on the same dataset to compare their performances with the proposed hybrid (LSTM-GRU) model and it is found that the proposed hybrid model shows supremacy in the performance with the MAE value 36.11 and R^2 value 0.84.

1. Introduction

Air pollution is a significant environmental issue in different parts of the world. When dangerous or excessive amounts of substances, such as gases, particles, and biological molecules, are emitted into Earth's atmosphere, air pollution occurs. Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), Carbon Dioxide (CO_2), Ozone (O_3), Sulfur Dioxide (SO_2) are the fine Particulate Matters as well as pollutants that causes air pollution (Zhu et al., 2017). On the other hand, air pollution, as much as any other health hazard, is responsible for a large proportion of early deaths in India (Glencross et al., 2020), (Pozzer et al., 2020). Growth in population density, expansion of industry and motor vehicles and increase in annual average temperature are the foremost influencing factors of air pollution. Additionally, the process of urbanisation has a detrimental effect on air quality.

Because of the harmful effects that pollution has on human health, the public is highly concerned about the future trend of air quality (Zou et al., 2019). Air pollution causes several diseases such as asthma, weakens lung function, increases cardiopulmonary illnesses (Yuan and Liu, 2014) and also escalates mortality rates. Numerous recommendations are there to shield the population from severe pollution, for example: i) public participation in outdoor activities should be reduced, ii) patients with severe diseases like heart disease, respiratory disease, pregnant women, children should stay in home more times. It is necessary to anticipate air pollution for both warning about and reducing pollution in people's daily lives.

The Air Quality Index, often known as the AQI, is a statistical measurement used to determine the quality of breathing air in our surroundings that consolidates the concentrations of various contaminants into a single numerical form (Zhu et al., 2017). It is used to investigate the long-term health effects of air pollution on a person's health as it explores the relationship among human health and air quality. The new ambient air quality standards (GB3095-2012), which covers six pollutants including Ozone (O_3), Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), and $PM_{2.5}$, PM_{10} (particulate matter with a diameter less than or equal to 10 μm), are used to calculate the AQI (Zhu et al., 2017). The most commonly used air quality evaluation indexes are criteria air pollutants and comprehensive indices (Zhu et al., 2018).

☆ This paper has been recommended for acceptance by Pavlos Kassomenos.

* Corresponding author.
E-mail addresses: phcs220002@nitsikkim.ac.in (N. Sarkar), b180030@nitsikkim.ac.in (R. Gupta), pankajkeserwani.cse@nitsikkim.ac.in (P.K. Keserwani), director@nitsikkim.ac.in (M.C. Govil).

https://doi.org/10.1016/j.envpol.2022.120404
Received 30 June 2022; Received in revised form 27 September 2022; Accepted 6 October 2022
Available online 11 October 2022

There has been a considerable deterioration in the quality of air in several cities of India throughout the years (Singh and Chauhan, 2020). It is very much critical to monitor overall air quality within the region and provide warnings when necessary. Whenever it comes to interpreting air pollution, raw data is difficult to comprehend. To address this issue, a quantitative measurement - AQI is being developed by the research community.

AQI prediction methods for air pollutants may be classified into three categories: i) ML based methods, ii) DL based methods and iii) Hybrid methods.

ML based method is a type of computer program that can identify patterns based on past data. A multi-source ML method was implemented by D. Iskandaryan et al. (2020) to evaluate the AQI scores of large metro polish neighborhoods. They used three significant sample sets (MNR-Air-HCM, SEPHLA-Medieval 2019 and MNR-HCM) to evaluate random forest performance. The researchers discovered that air quality stability has a powerful link between increasing air purity and may also assist to stabilise the environment. Van et al. (2022) employed a strategy for handling data on air pollution and increasing the AQI forecasting process using various ML based models like: Decision-Tree (DT), Random Forest (RF), and Extreme Gradient Boost (XGBoost) and the results are analysed through MAE, RMSE, and R^2 methods. XGBoost beated in foreseeing the AQI values than the other algorithms with RMSE = 0.03684. Mahalingam et al. (2019) discussed the difficulty of forecasting AQI so that pollution can be controlled before it worsens. To resolve it they utilized two ML Algorithms: Neural Network (NN) and Support Vector Machine (SVM) on the data taken from Central Pollution Control Board (CPCB). Liu et al. (2019) used Support Vector Regressor (SVR) and Random Forest Regressor (RFR) to predict Beijing's AQI and the concentration of Nitrogen Dioxides (NO_2). The regression models' performance was evaluated using RMSE and R^2 and SVR-based model predicted the value of AQI better than the RFR-based model with R^2 = 0.9766 and RMSE = 7.666. Pant et al. (2022) made an accurate prediction of the air quality in Dehradun, Uttarakhand by employing supervised machine-learning methods. The test results demonstrated that the decision-tree had a higher degree of accuracy, with a precision of 98.63%. It was interesting to see that the logistic regression had the highest precision of 91.78% for the expectation of air quality. Whereas, Castelli et al. (2020) examined the contaminant and particle levels and estimated the air quality score using Support Vector Regression (SVR) and Radial-basis-function (RBF) and it was proved that SVR was giving the most accurate and trustworthy predictions with accuracy rate was 94.1%. Gocheva-Iliev et al. (Gocheva-Ilieva et al., 2014) suggested the prediction of AQI using the Linear-Regression Algorithm and Gradient-Boost Algorithm. The Naive Forecast method was also used, but the Gradient Boost Algorithm was more accurate with 96% accuracy. This project helped to find the main pollutant that is causing pollution. Campbell-Lendrum et al. (Campbell-Lendrum and Prüss-Ustün, 2019) presented the use ML to predict $PM_{2.5}$ μm level from weather data in a non metropolitan city at a high elevation (Quito, Ecuador). Auto-regression was used to forecast $PM_{2.5}$ μm levels based on past $PM_{2.5}$ μm levels. SVM algorithms, Decision Tree (DT) were able to predict $PM_{2.5}$ μm with accuracy score of about 89%. Bhalgat et al. (2019) presented an overview of many investigations done on air quality prediction using SVM, Decision Tree (DT), and Effective Algorithms on a static dataset. Due to use of static dataset, there was no specific parameter on which the quality of air is dependent that may cause differ in forecast in real-world settings. To solve this challenge, a web scraper might be used to retrieve real-time datasets for training models, enhancing accuracy and offering better results. To determine how well AQI prediction works, Nigam et al. (2015), came up with a method that combined protocols for processing air pollution data and effective machine learning methods. To find the best model for AQI expectations, the MAE, RMSE, and R^2 measurements were used to compare three computation models: DT, RF, and XGBoost. The suggested methods were tested using two different public datasets that were collected from different parts of India. XGBoost method gave better result than others to predict the AQI. So, XGBoost was chosen again for a low-cost device to estimate the AQI at fixed-site assessment units. Similarly, Janarthanan et al. (2021) used two ML algorithms Support Vector Machines and Neural Networks, to predict the AQI.

The previous paragraph explains the detail of ML based approaches used by the research community to predict the AQI value, this paragraph represents a brief review on DL based methods utilized for AQI prediction. Due to its supremacy in terms of accuracy while learn with massive amounts of data, deep learning is becoming increasingly popular. In Sigamani and Venkatesan (2022), Sigaman et al. demonstrated the usage of features like $PM_{2.5}$ μm, PM_{10} μm, and other gaseous components used to predict AQI. The Auto-Regressive Integrated Moving Average (ARIMA) model, Auto Regression model, and Linear Regression model were employed as deep learning algorithms. On the other hand, Zhou et al. (2019) build an AQI prediction model based on Convolutional Neural Network (CNN), an attention mechanism, and GRU in their work. CNN was used to extract features from the input data and Gated Recurrent Unit (GRU) was used to predict the AQI. Weather and air quality data from January 1, 2017 at 00:00 to September 30, 2020 at 23:00 in the Chinese city of Shijiazhuang were collected and were applied to GRU prediction model to prove the better accuracy and it was compared with the performance of other models and the results of the proposed model showed the best overall performance with MAE = 6.099281, MSE = 90.781522, EVS = 0.972560, R^2 = 0.972495.

The hybrid methods mainly includes the combination of more than one ML or DL models. Chen et al. (Chen et al., Hong) used hybrid model to estimate the contamination, climate, and compounds was collected from the Weather Research and Forecasting-Chem (WRF-Chem) model. The partial mutual information (PMI) based input variable selection (IVS) hybrid model, also known as PMI-IVS model has the superior outcome than the current models from a technical standpoint with Calibration: 0.56, Validation: 0.52. On the other hand, Alireza et al. (2021) makes a use of Hybrid Single Degradation (HSD) and Hybridization Two Phase Decomposition (HTPD) model to make a prediction about the air quality record of Orumiyeh, Turkey, one day in advance. According to the results, the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise-Environmental Liabilities Management (CEEMDAN-ELM) model and the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise-General Regression Neural Network (CEEMDAN-GRNN) model both were successful in accurately forecasting AQI series information by employing HSD models. The HTPD shows the supremacy in accuracy and the value of R^2 = 0.98, RMSE = 4.13 and MAE = 2.99 in the training phase and the value of R^2 = 0.74, RMSE = 5.45 and MAE = 3.87 in the testing phase. An RNN-LSTM model proposed by Partheeban et al. (Partheeban) that can predict the amount of pollution in the air at any given hour and applied in a certain part of Delhi to figure out the AQI. The LSTM model got its information from time sequences of four weather parameters and the pollution levels. Similarly, Wu et al. (Wu and Lin, 2019) proposed a novel optimal-hybrid model in their study for AQI forecasting based on sample entropy (SE), secondary decomposition (SD), least squares support vector machine which is optimized by Bat-algorithm (BA-LSSVM) and LSTM neural network. The proposed SD-SE-LSTM-BA-LSSVM model achieved the supremacy in accuracy while comparing with the other hybrid models. The calculated error rates for the proposed model are: MAE = 6.6885, RMSE = 8.8920, MAPE = 0.0877 for the city Beijing and calculated MAE, RMSE, MAPE values for Guilin are 3.8036, 4.4396, 0.0880 respectively. For more improvement of the accuracy of AQI G Li et al. (2022) introduced an innovative hybrid AQI prediction model named CEEMDAN-mvMDE-BVMD-RSO-KELM by employing complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) for AQI decomposition, multivariate multiscale dispersion entropy (mvMDE) for calculating the complexity of every decomposed element, variational mode decomposition (VMD) optimized by Bald Eagle Search (BES) algorithm i.e., BVMD for selecting the decomposition

level and penalty factor and kernel extreme learning machine optimized by Rat Swarm Optimizer (RSO) algorithm (RSO-KELM) for anticipating the intrinsic mode function (IMF) components which are obtained from the AQI decomposition.

Table 1 summarises some of the mentioned works performed and evaluated to predict the Air Quality Index (AQI).

It is evident from above literature survey that a limited research are expressed for AQI prediction on various datasets. Now it is becoming very difficult to monitor and to predict the AQI, especially for urban areas due to industrial developments and increasing transportation. In order to predict AQI in the urban areas this work first uses samples gathered from the Central Pollution Control Board (CPCB). After that a hybrid model called hybrid LSTM-GRU model is being developed which uses R2, MAE, and RMSE approaches to assess the performance of the developed model, and finally compares the performance of the proposed integrated LSTM-GRU model with some developed ML and DL based models as well as other recent existing approaches. In the proposed approach a step-by-step workflow is being presented for the transformation of raw data into an integrated hybrid feature space to enhance the prediction of Air Quality Index of a certain region of Delhi with minimal error. For conducting the experiments raw data are taken from the Central Pollution Control Board's official website and the several pre-processing and integration processes such as data imputation, data aggregation, and data normalization are being conducted. Hereafter, rich features are selected as a next from the pre-processed dataset by discarding features with less feature importance score. The selected feature vector is fed as the input to the input layer of the ML and DL models. For modifying these time series data, numerous ML and DL models such as: linear regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) networks are deployed and further standalone GRU and LSTM models are combined together to get better result for AQI forecasting. The proposed hybrid LSTM-GRU model shows the supremacy in performance with minimal error rate. The significant contributions of this work are provided below:

Table 1
Summarizing of calibrated works to predict the AQI.

Ref.	Method	YOP	Result(s)
Iskandaryan et al. (2020)	SVM, RF, EGB, LightGBM and CatBoost	2020	RMSE = 9.3953
(Chen et al., Hong)	PMI-IVS and PEK-based ML	2021	Calibration: 0.56, Validation: 0.52
Li et al. (2019)	ML algorithms	2019	AR-Kalman model: MAE = 2.87, MAPE = 3.18, RMSE = 3.25, MSE = IMM Model: MAE = 0.54, MAPE = 0.55, RMSE = 0.95, MSE = 5.23
Van et al. (2022)	DT, RF, and GB	2022	RMSE = 0.03684
Mahalingam et al. (2019)	ANN, SVM	2019	RMSE = 17.21,RMSE = 17.88, RMSE = 13.91
Liu et al. (2019)	SVR, RFR	2019	SVR r = 0.9887, R^2 = 0.9766 RMSE = 7.666, RFR r = 0.9823, R2 = 0.9633, RMSE = 9.602
Srivastava et al. (2018)	LR, SDG, RF, DT, SVM, CNN, GB,	2018	R^2 = 0.65646, 0.65922, 0.67, 0.62, 0.69275, 0.68478, 0.69647, 0.69275
Pant et al. (2022)	DT, LR	2022	DT = 98.63% LR = 91.78%
Castelli et al. (2020)	SVR-RBF, PCA, SVR-RBF	2020	94.1%
Kleine Deters et al. (2017)	Multiple Linear Regressive model	2017	MSE = 15.0
Alireza et al. (2021)	HSD, HTPD, CEEMDAN-ELM, CEEMDAN-GRNN	2020	R^2 = 0.74 RMSE = 5.45 MAE = 3.87
Londhe (2021)	Multiple regression, RMSE, KNN	2021	RMSE = 52.34, R^2 = 0.89, MAE = 24.79
Gocheva-Ilieva et al. (2014)	LR and GB	2014	Accuracy = 96%
Campbell-Lendrum and Prüss-Ustün (2019)	SVM and DT	2019	Accuracy = 89%
Bhalgat et al. (2019)	SVM, DT	2019	MSE = 166.358
Sigamani and Venkatesan (2022)	MLR, ANN, SVM	2022	R^2 = 0.923, 0.931, 0.953
Liu et al. (2018)	RF, SVM,DT, GB,	2018	RMSE = 9.8
Janarthanan et al. (2021)	SVM,NN	2021	RMSE = 10.995, R^2 = 0.570
Zhou et al. (2019)	CNN, GRU	2021	MAE = 6.099281, MSE = 90.781522, EVS = 0.972560, R^2 = 0.972495
(Partheeban)	RNN-LSTM model	2021	MAE = 1 R^2 = 0.89

- We have accumulated the air quality data from a particular city in India and built the hybrid Long-Short-Term Memory-Gated Recurrent Unit (LSTM-GRU) model to predict the AQI value by taking into account the pollutant content in the air as well as many meteorological characteristics, e.g., temperature, wind speed, and relative humidity, amongst others.
- Permutation feature importance method is adopted to find out the feature importance score of each feature of the dataset and the feature with least importance score is discarded.
- Various standalone ML and DL models: LR, KNN, SVM, LSTM, GRU as well as the Hybrid LSTM-GRU model are utilized to receive the chosen feature vector as input and the effectiveness of all the models has been evaluated.
- Efficiency of each of the model is analysed on the basis of RMSE, MAE and R^2 values and the proposed hybrid model shows the best performance w.r.t the MAE and R^2 values.

1.1. Organization of the paper

The remaining parts of the paper are divided into the following sections: materials and methods associated with this study are outlined in Section 2. Section 3 shows the results and the implementation planning. The conclusive comparison of performance is presented in Section 4. The paper is brought to an end in conclusion presented in Section 5, which also discusses the future work direction.

2. Materials and methods

The step by step process flow for developing an AQI prediction model is presented in this section. All involved steps or modules required for implementing the proposed model are exhibited in Fig. 1. The flow diagram consists of several modules: reading of AQI data, data preprocessing, dataset splitting, then utilize the ML and DL models to predict AQI and finally choose the best model on the basis best performance. All the associated steps are explained subsequently.

2.1. Data collection

The process of collecting and evaluating information is referred to as data collection. In this study two sets of data are used. Data is collected from official portal of the Indian government, the Central Pollution Control Board. The dataset includes pollutant concentration and climate data collected every hour for seven years (from March 2015 to December 2021). The dataset contains 2482 number of samples of weather data from Delhi taken every day basis. In the dataset, there are 14 climatic factors, like temperature, wind-speed, humidity, and so on. The Central Pollution Control Board, where data for 18 states, 102 cities,

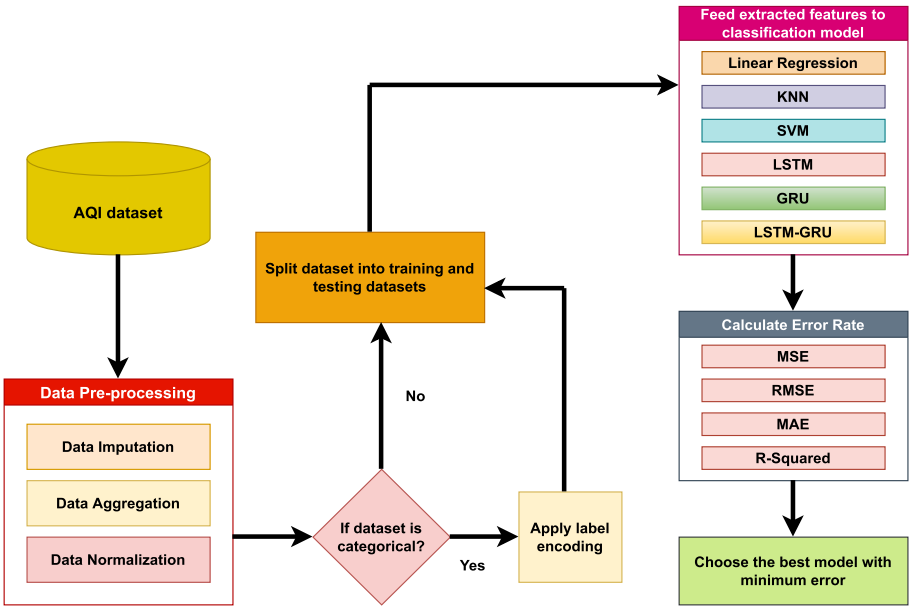

Fig. 1. Proposed process flow of developing an AQI prediction model.

and 170 stations is publicly accessible, is consulted in order to collect the information used to determine the concentration of pollutants and the meteorological parameters. The station considered for data collection is Netaji Subhas University of Technology (NSUT) which was formerly known as Netaji Subhas Institute of Technology (NSIT), situated in Sector 3 of Dwarka. In comparison to the data from other stations, those data from NSUT are quite dense, with some non-assigned values, totalling anywhere from a few days to a few weeks on an annual basis. As a result, data spanning seven years is used in order to accomplish the goal. The chronology for the data begins on March 23, 2015 and continues through December 31, 2021. The number of data points were collected hourly then converted to daily by taking the hourly average value.

Table 2 depicts a list of considered air pollutants: Carbon Monoxide (CO), Nitrogen Oxide (NO), Nitrogen dioxide (NO_2), Nitrogen Oxides (NO_x), Sulfur dioxide (SO_2) and Ozone (O_3). CO an uncolored, unperfumed and tasteless gas which is generated by burning fuels. It is extremely poisonous and a bit less dense than air. CO is measured by the unit mg/m^3 which tells the amount of CO in milligram per unit cubic meter. NO, NO_2 and NO_x belongs to the family of extremely reactive, toxic gases which are also produced by flickering fuels in high temperature. The quantity of these gases in air is measured by the unit microgram per cubic meter of air i.e. $\mu g/m^3$. Similarly, SO_2 is also a poisonous gas that gives off the odour of burnt matches and measured by $\mu g/m^3$. O_3 is a highly reactive gas comprised of three oxygen atoms and belongs to both the natural and artificial substances. It has many bad impacts on human health like chest pain, coughing, lung infections etc. O_3 is measured by $\mu g/m^3$.

Table 3 presents meteorological parameters: temperature, wind speed, relative humidity and solar radiation which have a great impact on air pollution. The brief idea about each of the meteorological parameter is enlisted below:

Table 2
List of pollutants considered.

Sl. No.	Parameters	Units
1	NO (Nitrogen Oxide)	$\mu g/m^3$
2	NO_2(Nitrogen dioxide)	$\mu g/m^3$
3	NO_x(Nitrogen Oxides)	$\mu g/m^3$
4	SO_2(Sulfur dioxide)	$\mu g/m^3$
5	CO(Carbon Monoxide)	mg/m^3
6	O_3(Ozone)	$\mu g/m^3$
7	$PM_{2.5}$(Particulate Matter 2.5 μm)	$\mu g/m^3$

Table 3
Meteorological parameters.

Sl. No.	Parameters	Units
1	Temperature	°C
2	WS (Wind Speed)	m/s
3	RH (Relative Humidity)	%
4	SR (Solar Radiation)	W/m^2

- **Temperature:** A physical quantity that indicates climatic conditions, whether it is cold or hot, is referred to as temperature. High temperature influences the motion of air which makes some consequences on air pollution. It is measured by degree Celsius (°C) or degree Fahrenheit (°F).
- **Wind Speed:** In meteorology, wind speed is a fundamental atmospheric parameter brought on by air shifting from a state of high to low pressure, usually happens as a result of temperature variations. Wind speed is basically a rate in which air is passing through an area. It is measured by meter per second (m/s).
- **Humidity:** The amount of water or moisture that is present in the air as water vapour is referred to as humidity. The quantity of hazardous or dangerous substances in the air increases with high humidity. Humidity is measured in percentage (%).
- **Solar Radiation:** Solar radiation is the electromagnetic radiation released by sun. This radiation makes some effects in temperature which in turns is related with air pollution. Solar radiation is measures by Watt per square meter (W/m^2).

The detail information of statistical measurement i.e. the count, mean, standard deviation etc. of the used dataset is illustrated in Table 4. "Count" calculates the total number of non-missing values for each corresponding feature. "Mean" is the average of all the observations used. The term "standard deviation" refers to a measurement of the dispersion of data from the mean.

2.2. Data pre-processing

A dataset contain important information as well as noises, outliers, and null values. These noises, outliers, and null values are affecting the

Table 4
Information of pollutants and meteorological parameters.

Index	NO	NO_2	NO_x	SO_2	CO	O_3	Temp	RH	WS	SR	$PM_{2.5}$
Count	2337.0	2338.0	2341.0	2365.0	2271.0	2346.0	2370.0	2267.0	2221.0	2370.0	2311.0
Mean	18.61	30.23	30.62	10.52	124.21	32.80	25.97	57.08	0.91	115.06	107.85
Std	20.54	14.48	21.05	7.28	253.52	20.12	7.25	18.17	0.42	76.44	73.13
Min	0.21	0.0	0.0	0.3	0.0	0.8	5.36	8.79	0.05	1.5	8.29
25%	6.49	20.02	16.92	5.88	0.59	20.14	20.312	43.95	0.63	69.65	56.875
50%	10.93	27.52	24.1	8.73	0.91	28.805	27.79	58.7	0.86	101.84	92.08
75%	21.91	37.635	37.96	12.82	2.29	40.67	31.375	70.74	1.14	139.63	137.03
Max	170.37	148.58	171.52	80.05	992.44	240.85	60.57	101.79	3.16	790.99	982.69

forecasting procedure in a negative manner which leads to less accurate prediction. Because of this, it is necessary to know how to deal with these noises, outliers, and null values in right way.

The collected data is containing a number of missing values as well as extreme values that vary over other data observations. This is an indication of measurement variability, errors in the experiment, or uniqueness. This problem is handled by giving the outliers null values to represent their position. The noise that is introduced into the dataset as a result of the missing data, has a detrimental impact on the competence of the model. The linear interpolation is used to fill in the blanks left by the missing data. It is a technique for generating new data points within a certain range of an existing discrete collection of data points that have already been determined. Other varieties of possible configurations of interpolation are linear, bi-linear, piece wise, polynomial, spline, cubic, and bi-cubic. Simply said, linear interpolation is the process of estimating a missing value by joining together dots in a straight line in ascending order in an efficient way (Huang, 2021). Due to its efficiency this study utilizes linear interpolation to replace missing values. The total dataset evaluated, includes 27,236 samples for every contaminant and meteorological parameter. Data pre-processing acts in accordance with the following steps:

2.2.1. Data imputation

The technique of substituting alternative values for missing data is known as data imputation (Zhang). When substituting a data point, the occurrences of missing values in the input parameters are thrown out. An impute function based on mean value technique is used to execute interpolation in order to make an approximation of the missing data in the case of the target object, which is the pollutant. Fig. 2 depicts the graphical representation of the heat map of the missing values. The x-axis of this figure is representing the air pollutants and considered meteorological parameters on the other side y -axis is representing the count of the missing values.

The measurement of the missing and non-missing values is given in Fig. 3. The blue part is representing the non-missing entries whereas the orange part says about the missing values.

The actual AQI value is represented graphically before applying the data normalization from the year 2015–2021 in Fig. 4:

2.2.2. Data aggregation

Data aggregation is the process of storing and representing data in a summary format (Clark and Avery, 1976). The data can come from more

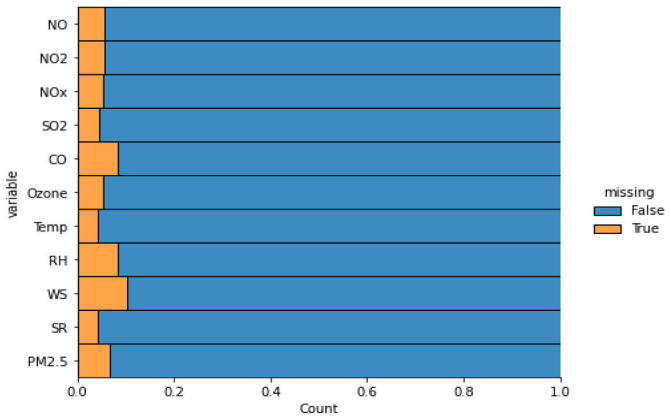

Fig. 3. Graphical representation of missing and non-missing values.

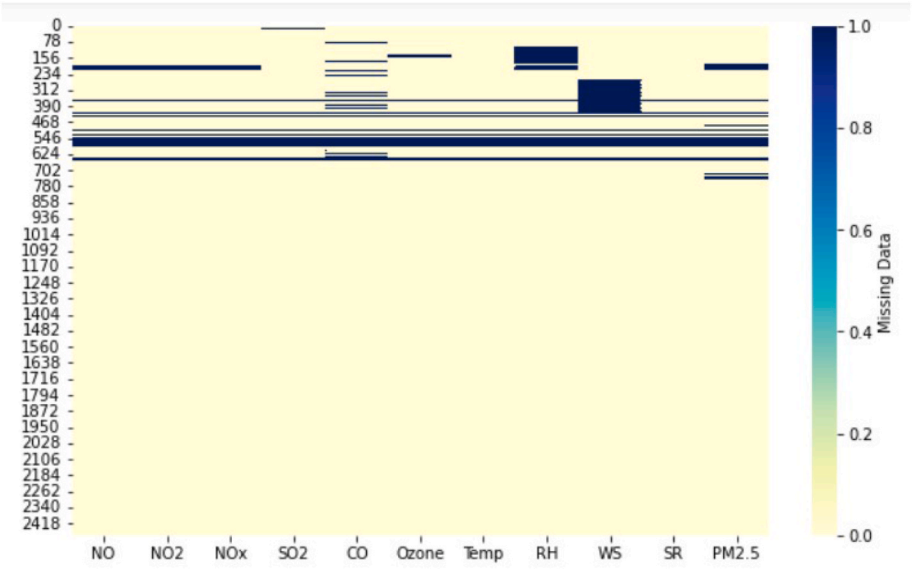

Fig. 2. Heat map of missing values.

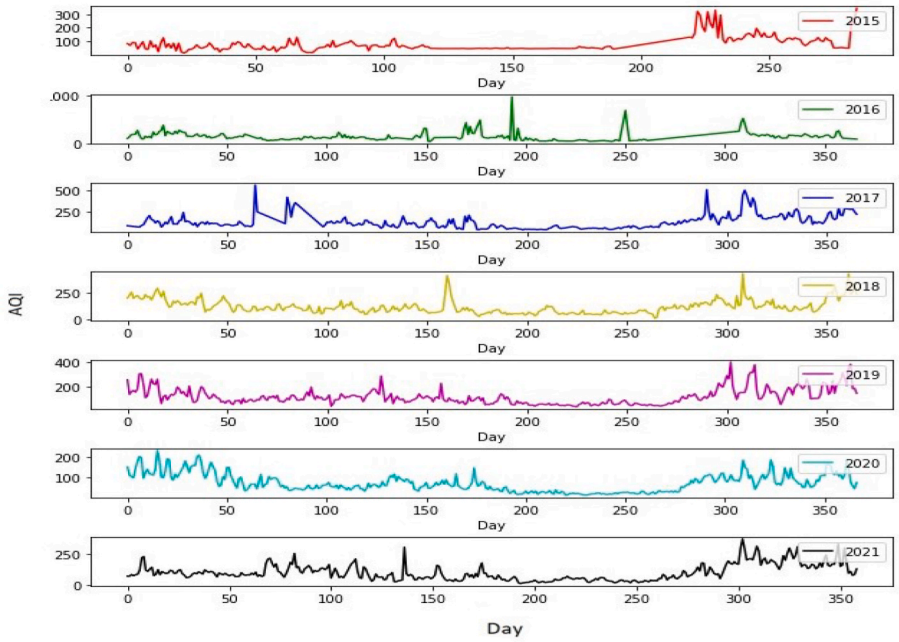

Fig. 4. AQI daily data for the years 2015–2021.

than one sources, and all of these sources can be put together in a data analysis description. This is a very important step because the accuracy of insights from data analysis depends a lot on the amount and quality of the data used. To get useful results, it is necessary to collect accurate, high-quality data in a large enough amount. Collecting data can help in everything-from making decisions about financing or business strategy, setting prices, running operations, and coming up with marketing plans etc.

2.2.3. Data normalization

Attributes consisting of several characteristics with different units, must be scaled to a specific range to ensure that each attribute receives the same amount of weightage. This process is known as data normalization (Turky et al., 2021). This prevents a potentially more significant trait from being overshadowed by a less significant one that has a wider range of values. In this case we re-scale all of the characteristics utilising the Z-score normalization or the mean - standard deviation scaling. The dataset has been normalized by the use of the mathematical formula, which can be written as:

$$X_{norm} = \frac{(X - X_{mean})}{X_{std}} \quad (1)$$

where, X_{norm} = value of attribute obtained after normalization.

X_{mean} = mean of the population and
X_{std} = standard deviation of the population.

AQI value is obtained after testing with normalized dataset which uses a minmaxscalar of range 0–1. The information shown in Table 4 makes it clear that the ranges of mean and standard deviation of our data are not the same. The gradient can fluctuate, and it may take a very long time to get either a local or a global minima since the ranges of possible values for the many attributes are not the same. Therefore, using Min-Max Normalization, the data are normalized on a scale in between 0 and 1 in order to tackle the issue of model learning. This ensures that the values of the various characteristics are not comparable to one another, which enables gradients to converge at a faster rate.

2.2.4. Feature selection

In the construction of a predictive model, one of the most important steps is feature selection (Agrawal and Sharma, 2022), which is the process of minimising the number of input variables or attributes by removing the irrelevant features. It is necessary to limit the number of parameters in order to enhance the performance of the model as well as to reduce the computational cost of modelling. An essential function that helps to reduce the number of input features in predictive modelling is "feature importance score" that provides information and insight of the working dataset. The term "feature importance" relates to the methods for scoring each input attribute for a certain model; the scores merely indicate the "importance" of each feature. A higher score indicates that the particular characteristic will have more impact on the model being used to forecast the target attribute. Here, the feature importance score is calculated based on permutation feature importance method. All features are arranged in ascending/descending order of its in permuted feature importance values.

Permutation feature importance is a technique for model inspection used for any of the fitted non linear estimator on a tabular form of data. The permutation feature importance refers to reduction of a model score for single feature values is shuffled randomly (Breiman, 2001). The process for calculation of permutation feature importance breaks the propinquity among the feature and the target in the tabular dataset. In this way it a technique to determine the dependency of the model on the feature. This technique can be used a number of times with various permutations of the feature of the dataset.

The algorithm to compute the permutation feature importance for each feature of a tabular dataset is given below as:

Algorithm 1
Algorithm to get Permutation Feature Importance of a feature

Require: Predictive Model: m, Dataset: D with J rows and P columns
1: Compute score 's' of the model 'm' on D (it is accuracy for a classifier and R^2 for a regressor)
2: **for** each feature p in D) **do**
3: **for** each repetition j in 1, ….,J **do**
4: Shuffle column 'p' randomly to generate a corrupted data $D_{j,p}$ for column 'p'
5: Compute score $s_{j,p}$ of 'm' on $D^{'}_{j,p}$.

(*continued on next page*)

Algorithm 1 (*continued*)

6: **end for**

7: Compute the feature importance i_p for feature f_p given as $i_p = s - \frac{1}{J}\sum_{j=1}^{J} s_{j,p}$ (2)

8: **end for**

The feature importance score of each attribute of developed dataset is depicted in Fig. 5. As $PM_{2.5}$ is one of the most prime ingredients in AQI calculation, that's why $PM_{2.5}$ must be included as a rich feature in the feature vector. So, in Fig. 5 the feature importance score of the remaining components excluding $PM_{2.5}$ is shown.

2.3. Machine learning and deep learning models

Various Machine Learning models, Deep Learning models and hybrid model are developed and obtained results from these developed models are compared mutually to get a statistical idea about the overall performance comparisons. The basic and brief idea about the used ML and DL models are given below:

2.3.1. Linear regression

In terms of data analytics (Ghani et al., 2019), Linear-Regression is the simplest and most often used ML approach. In linear regression, there is only one independent variable, and the relationship between the independent variable (x) and the dependent variable (y) is linear (Yu and Yao, 2017), (Pal and Bharati, 2019). The line can be modelled is mathematically expressed as follows (Bonaccorso, 2017):

Consider the dataset of real value input vector is given by X and $X = [\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots\ldots, \overrightarrow{x_m}]$ and $\overrightarrow{x_i} \in \mathbb{R}^m$, where $\mathbb{R}$ is the set of real numbers.

Each of the input vector is mapped with a real valued output i.e. y_i. Let us consider, Y represents the set of real valued output mapped with every real valued input vector i.e. $Y = [y_1, y_2, \ldots.., y_m]$, where $y_i \in \mathbb{R}$.

In linear regression it is tried to draw a line that best fits the points based on the given data.

$$y = a_0 + \sum_{i=1}^{m} a_i x_i \tag{3}$$

The best fit straight line is the red line shown in Fig. 6.

The objective of the linear regression method is to determine the optimal a_0 and a_i values. Where, a_0 is estimated intercept and a_i is the slope of the straight line.

2.3.2. K-Nearest Neighbor

A supervised machine learning technique called the K-Nearest Neighbor or KNN relies on the idea that how closely a point's value resembles the values of its neighbors. Therefore, it operates by selecting

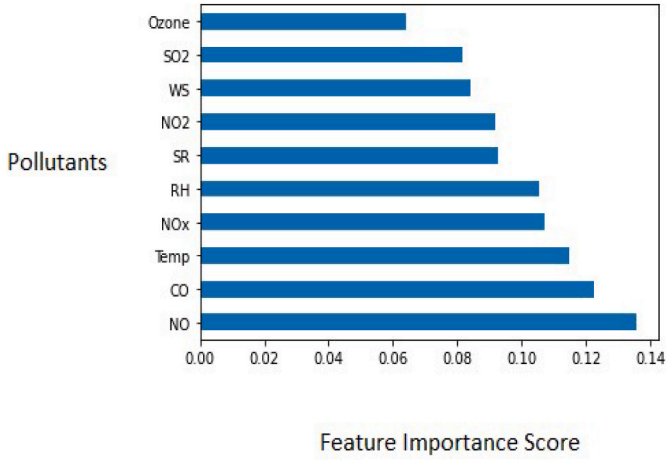

Fig. 5. Feature importance score.

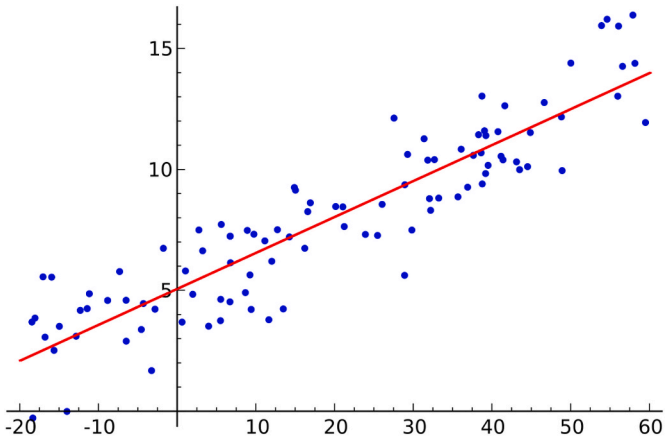

Fig. 6. Linear regression.

the K neighbor value that is nearest to the site of interest and selecting the class that appears the most frequently (Zhang et al., 2017a). The Euclidean distance (d) between two locations (a_1, a_2) and (b_1,b_2) may be calculated as follows (Zhang et al., 2017b):

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \tag{4}$$

For n-dimensional space, d can be obtained as:

$$d = \sum_{i=0}^{n} \sqrt{(a_i - b_i)^2} \tag{5}$$

The graphical representation of KNN is shown in Fig. 7.

2.3.3. Support vector machine

A nonlinear mapping φ_j (x) is used in SVM to translate data events x into a multidimensional feature space F, which then allows a linear regression model to be fitted. Again, a kernel function determines how input data is mapped further into new feature space. An interesting property of SVM is its approach to modelling error minimization, rather than focusing just on the apparent training errors (Fan et al., 2020).

In this case, the training dataset T is shown as:

$$T = (x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n) \tag{6}$$

where $x \in X \subset \mathbb{R}^n$ are the inputs for training and $y \in Y \subset \mathbb{R}^n$ are the anticipated results of training.

A nonlinear function representing the SVM is given by equation (7):

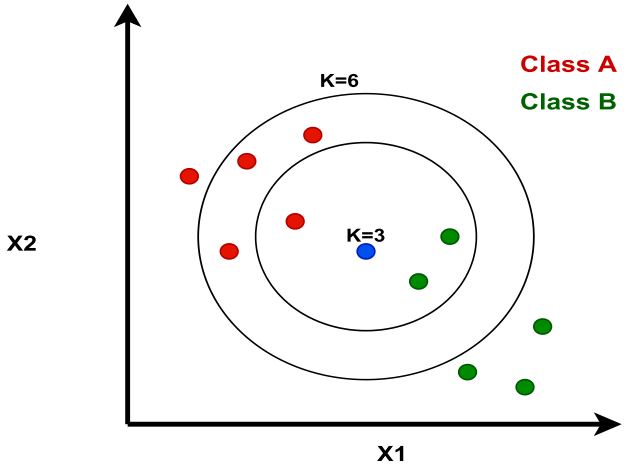

Fig. 7. K-nearest neighbor.

$$f(x) = w^T \varphi(x_i) + b \tag{7}$$

where, w denotes the vector of weights, bias is denoted by b, x is the input space, $\varphi(x_i)$ is the high-dimensional feature space. The aim is to develop a function f(x) that has the least possible deviation $\in$ from the goals y_i while fitting the training dataset T. Three training factors determine SVM's effectiveness and generalizability, these are:

- C (the regularization parameter)
- The kernel functions
- $\in$ (the insensitive zone)

Fig. 8 shows the graphical presentation of SVM.

2.3.4. Long Short-Term Memory

In the LSTM model, there are three different types of gates: input gates, forget gates, and output gates (Lei et al., 2019; Alvi et al., 2022). The output from the input modulator gate is fed into memory cell, which is responsible for the collection of any new information that is received from the outside world. The forget gate provides the following generation with instructions that which data are to be preserved and which data are to be thrown away. This is how it determines which delays are optimal for the data series that is provided. The output gate receives the outcomes of the computations in order to process them (Danihelka et al., 2016).

The structure of an LSTM cell is shown in Fig. 9:

A mathematical representation of the cells may be found as follows (Dey and Salem, 2017):

Input gate is represented as follows:

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{8}$$

Forget gate is expressed by f_t:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{9}$$

Equation (10) depicts the mathematical expression for output gate. Output gate:

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

The Memory cell is represented as follow:

$$c_t = f_t \circ c_{(t-1)} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{11}$$

Shadow state is given by h_t:

$$h_t = o_t \circ \sigma_h(c_t) \tag{12}$$

initial values $c_o = 0$ and $h_o = 0$ and the operator $\circ$ designates product in an element-wise manner.

The time stamp is represented by the subscript t.

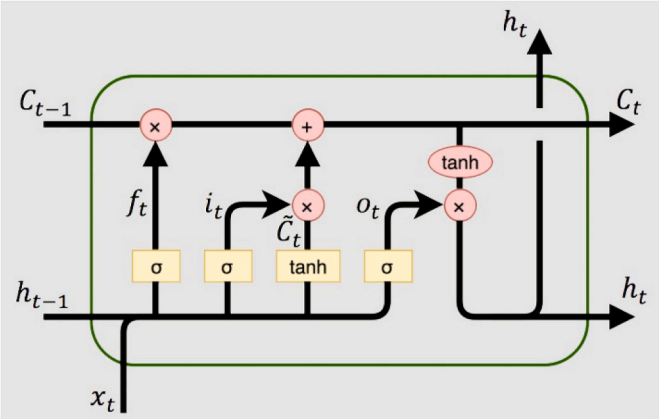

Fig. 9. Long short term memory cell.

$x_t \in R^d$: Vector to be used as input to LSTM unit.
$f_t \in R^h$: Activation vector used for Forget gate of LSTM unit.
$i_t \in R^h$: Activation vector used for Input gate of LSTM unit.
$o_t \in R^h$: Activation vector used for Output gate of LSTM unit.
$h_t \in R^h$: Output vector of LSTM unit.
$c_t \in R^h$: Vector to be used for cell state.
$W \in R^{(h*d)}$, $U \in R^{(h*h)}$, $b \in R^{(h)}$: Weight matrices and bias vector parameters that are trainable where superscripts $\in h$ and $\in d$ refer to the number of input features and number of hidden units respectively.
σ_g: Sigmoid function
σ_c: Hyperbolic tangent function
σ_h: Hyperbolic tangent function or identity function

A completely connected hidden layer is used to pass through the output from LSTM layers. The output layer produces the pollutant concentration at time (k+1), which is represented by $C^p_{(k+1)}$.

2.3.5. Gated Recurrent Unit

In addition to maintain the LSTM's properties, GRU significantly simplifies the structure of the algorithm. In the GRU model, there are numerous levels of interconnection (Tang et al., 2016) (Alvi et al., 2021). GRU's duplicate module structure is simpler than LSTM but both have recurrent neural networks with duplicate modules (Chen et al., 2019). Fig. 10 shows the GRU cell which is containing only two gates, the update gate (Z_t) and the reset gate (r_t).

A mathematical representation of the cell may be found as (Dey and Salem, 2017):

$$r_t = \sigma(W_r.[h_{t-1}, x_t]) \tag{13}$$

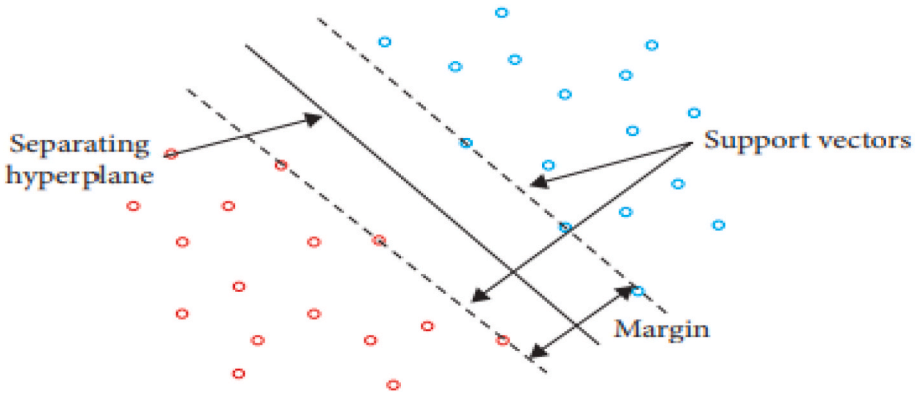

Fig. 8. Support vector machine.

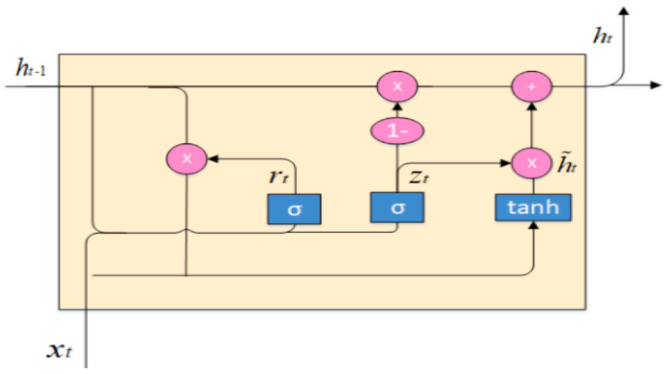

Fig. 10. Gated recurrent unit cell.

$$z_t = \sigma(W_z.[h_{t-1}, x_t]) \tag{14}$$

$$h_t = \tanh(W_h.[r_t * h_{t-1}, x_t]) \tag{15}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \tag{16}$$

$$y_t = \sigma(W_o.h_t) \tag{17}$$

Here, σ is the logarithmic sigmoid. The input value x and the preceding hidden state value h_t are both represented in these expressions. Learned weight matrices are represented by: W_r, W_z and W_h.

2.4. Performance evaluation metric

Performance evaluation metric is nothing but the benchmark for measuring the efficiency and efficacy of the ML and DL models. Model construction cannot be started until a comprehensive and in-depth evaluation has been completed. So, performance evaluation is one of the most important step for building a model and there are numerous metrics that can evaluate the performance of the ML and DL models. The performance evaluation metrics utilized in this work are enlisted below:

- **R-squared (R^2):** R-squared (R^2) (Rights and Sterba, 2021) is a measure of the amount of variation in the result that can be accounted by the factors.
- **Mean Squared Error (MSE):** In mathematics, the Mean Squared Error (MSE) (Azeez and Adewoye) is the same thing as the root square of the mean squared error and the MSE is the lowest absolute disagreement between observed actual output values and predicted values by the model. The average squared difference between the actual output and the predicted output gives the MSE values which is represented by the following mathematical formula.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(P_{ACT(i)} - Q_{PRED(i)}\right)^2 \tag{18}$$

- **Root Mean Squared Error (RMSE):** Using the Root Mean Squared Error (RMSE) (Karunasingha, 2022), we may determine how accurate a model's prediction are in light of actual data. The RMSE is calculated by taking the square root of the MSE. The RMSE is the minimum absolute discrepancy between these two sets of values. Mathematically RMSE is represented as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_{ACT(i)} - Q_{PRED(i)}\right)^2} \tag{19}$$

- **Mean Absolute Error (MAE):** Mean Absolute Error (MAE) (Brassington, 2017) performs the same thing that RMSE does, which is to quantify the error in the model's prediction. MAE is computed by aggregating the absolute difference between the actual output and the predicted output of each observation over the entire dataset and then dividing the acquired sum by the total number of observations in the dataset. Equation (19) denotes the mathematical expression for MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_{ACT(i)} - Q_{PRED(i)}| \tag{20}$$

Here, $P_{ACT(i)}$ is the actual output of the ith sample, $Q_{PRED(i)}$ is the predicted output of the ith sample, n is the total number of samples. The evaluation of MSE, RMSE and MAE is portrayed by the following example. Table 5 shows the values of actual and predicted output for three observations.

3. Implementation and result

The dataset is comprised of air quality data and meteorological data. Based on the empirical research the best results are obtained by partitioning the dataset into 70–80% for training and 20–30% for testing. That's why the dataset in this work is bisected as 80% for training and 20% for testing purposes. Different ML and DL models are trained with the training dataset, tested on testing dataset to evaluate the results for analysing the performance of the models.

Further subsections discuss the experimental setup, implementation and result analysis of the different ML and DL models such as Linear Regression model, K-Nearest Neighbor model, Support Vector Machine model, Long Short Term memory model, Gated Recurrent Unit model and finally the proposed hybrid LSTM-GRU model.

3.1. Experimental setup

The work presented in this article aims to develop a prediction model to predict the AQI value using ML and DL algorithms. The proposed approach is the amalgamation of the LSTM and GRU model. The simulation and modelling of this work were performed on a machine with an Intel Core i7-5500U CPU @ 2.40 GHz 2401 Mhz and 12 GB RAM running Microsoft Windows 10 Professional. All models for AQI forecasting are designed and developed using the Python programming environment created on the mentioned machine.

3.2. Implementation using linear regression model

The graph for AQI prediction with respect to time (day) for the actual data and the predicted data for linear regression model is shown in Fig. 11. The performance assessment of the Linear Regression model is presented using MAE, RMSE and R^2 methods. The obtained MAE, RMSE and R^2 values are 38.86, 57.09 and 0.42 respectively.

3.3. Implementation using K-Nearest Neighbor Model

The graph for AQI prediction with respect to time (day) for the actual data and the predicted data for K Nearest Neighbor Model with 10 neighbors is shown in Fig. 12. The performance assessment of the K-

Table 5
Actual vs. Predicted Output.

	Actual Output	Predicted Output
Observation 1	0.8	0.7
Observation 2	0.8	0.9
Observation 3	1.1	1.2

$$\text{MSE} = \frac{1}{3}[(0.8-0.7)^2 + (0.8-0.9)^2 + (1.1-1.2)^2] = 0.01.$$

$$\text{RMSE} = \sqrt{\frac{1}{3}[(0.8-0.7)^2 + (0.8-0.9)^2 + (1.1-1.2)^2]} = 0.1.$$

$$\text{MSE} = \frac{1}{3}|(0.8-0.7) + (0.8-0.9) + (1.1-1.2)| = 0.1.$$

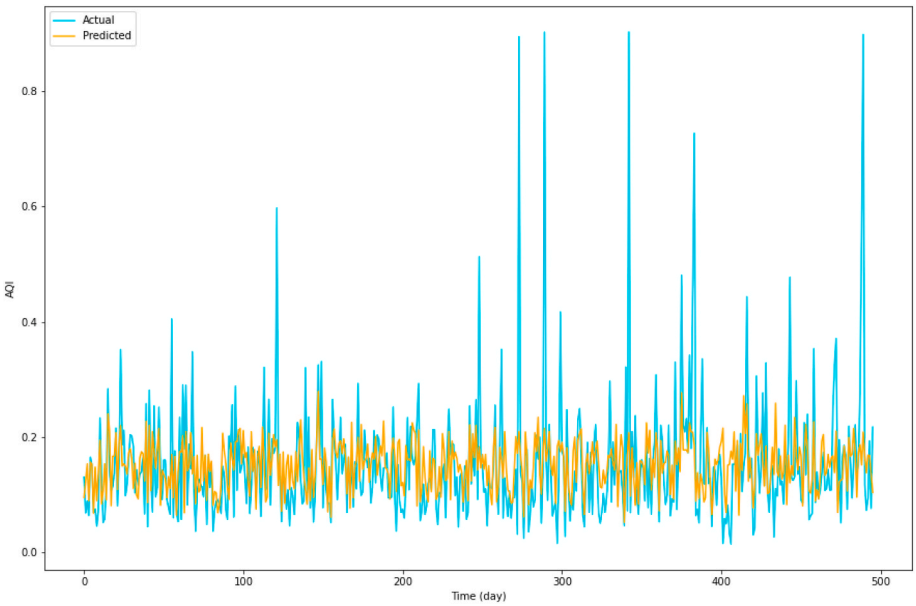

Fig. 11. Actual output Vs Predicted output for Linear Regression Model.

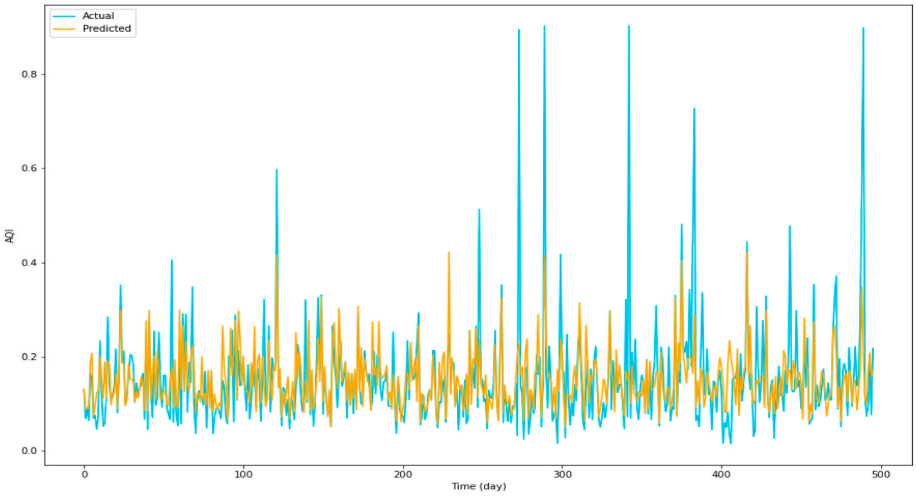

Fig. 12. Actual output Vs Predicted output for K-Nearest neighbor Model.

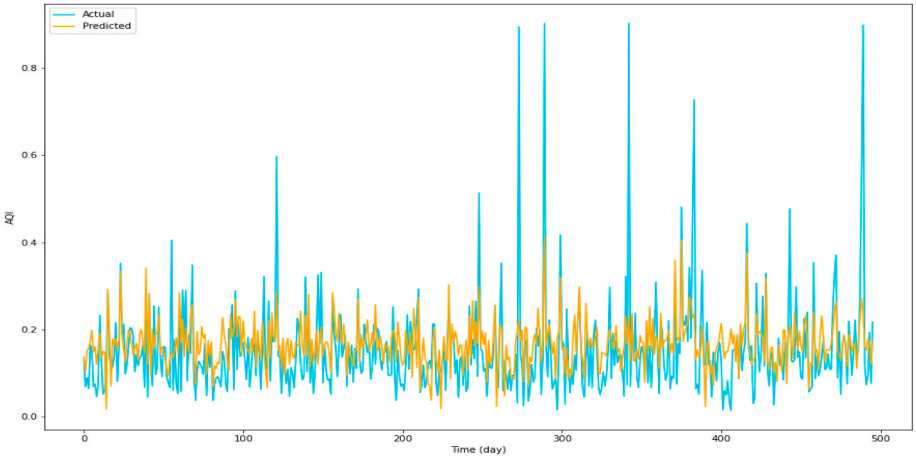

Fig. 13. Actual output Vs Predicted output for Support Vector Machine.

Nearest Neighbor model is presented using MAE, RMSE and R^2 methods. The obtained MAE, RMSE and R^2 values are 37.86, 57.65 and 0.47 respectively.

3.4. Implementation using support vector machine

A Support Vector Machine model has been constructed by utilising the Support Vector Regressor function and the Gaussian Radial Basis function (RBF) kernel parameter to the train dataset and generate predictions on the test dataset. The graph for AQI prediction with respect to time (day) for the actual data and the predicted data for SVM Model is shown in Fig. 13. The performance assessment of the SVM model is presented using MAE, RMSE and R^2 methods. The obtained MAE, RMSE and R^2 values are 49.09, 60.63 and 0.56 respectively.

3.5. Implementation using Long Short Term Memory

It is essential to scale the data before moving on to LSTM model for getting more accurate result. In order to maintain the same range of attribute variances throughout both the train dataset and the test dataset, Python's MinMaxScaler() method has been used. In the LSTM model 10 input feature vectors are fed into the input layer and one hidden layer are used with 50 neurons. The output is given by 1 target neuron using 'Relu' activation function. The number of batch size and epochs taken for the LSTM model are 16 and 10 respectively.

The graph for AQI prediction with respect to time (day) for the actual data and the predicted data for Long Short-Term Memory model on the testing samples is shown in Fig. 14. The performance assessment of LSTM model is presented using MAE, RMSE and R^2 methods. The obtained MAE, RMSE and R^2 values are 58.97, 61.37 and 0.78 respectively.

3.6. Implementation using Gated Recurrent Unit

The Gated Recurrent Unit (GRU) network is designed similar to LSTM for evaluation. It uses Min-Max function to scale data. At first the data are transmitted in a sequence. Then the dropout function has been applied to prevent data from being overfit and also a dense layer consisting of a single unit is utilized to produce a single variable output. The graph for the actual output and the predicted output for Gated Recurrent Unit model for AQI prediction with respect to time (day) on testing samples is shown in Fig. 15. The performance assessment of GRU model is presented using MAE, RMSE and R^2 methods. The obtained MAE, RMSE and R^2 values are 42.66, 49.45 and 0.65 respectively.

3.7. Implementation using LSTM-GRU model

In this work, a two-stage framework for foreseeing data on air quality is developed. When the process of input data preparation is started, first the method searches for any missing values and then uses the median function to replace them. Min-max scalar approaches are used for the datasets that have previously been normalized. Then, the GRU and LSTM deep learning models are integrated into a hybrid model.

To build up the LSTM-GRU integrated model, the data spanning from the seven years prior to the present is taken. In the proposed compound model, the LSTM layer is in charge of gathering the input samples and sending it on to the GRU layer, where a dropout function is used to prevent the data from being overfit. For the LSTM framework 10 input feature vectors are fed to the input nodes of the input layers which are connected with 100 neurons of the hidden layer using the 'tanh' activation function. A dropout with 0.2 is used to avoid the overfitting. Whereas, the GRU framework consists of a hidden layer with 50 neurons along with 'Relu' activation function and 0.2 dropout.

The graph for the actual output and the predicted output for LSTM-GRU model for AQI prediction with respect to time (day) on testing samples is shown in Fig. 16.

The performance assessment of LSTM-GRU model is presented using MAE, RMSE and R^2 methods. The obtained MAE, RMSE and R^2 values are 36.11, 52.21 and 0.84 respectively.

3.8. Flow diagram of the implemented integrated LSTM-GRU model

Fig. 17 depicts the work flow diagram of the proposed integrated LSTM-GRU Model.

It is depicted from this section that the proposed integrated LSTM-GRU model has the least MAE value and R^2 value while predicting the AQI based on the collected dataset from CPCB. The subsequent section discusses the comparative study on the basis of the performance metrics of the proposed integrated LSTM-GRU model with the other existing ML and DL models.

4. Comparative analysis of performance of different models

This section discusses the comparison of evaluation metrics of proposed integrated LSTM-GRU model with other ML and DL models. Evaluation is a crucial step for any model implementation and that also allows to identify the optimal model among different models based on the performance result. In this study the evaluation analysis or the comparative analysis of the performance of different models are conducted utilising three metrics. In order to conclude the calculations, the actual value was compared to the anticipated outcomes. In the context of

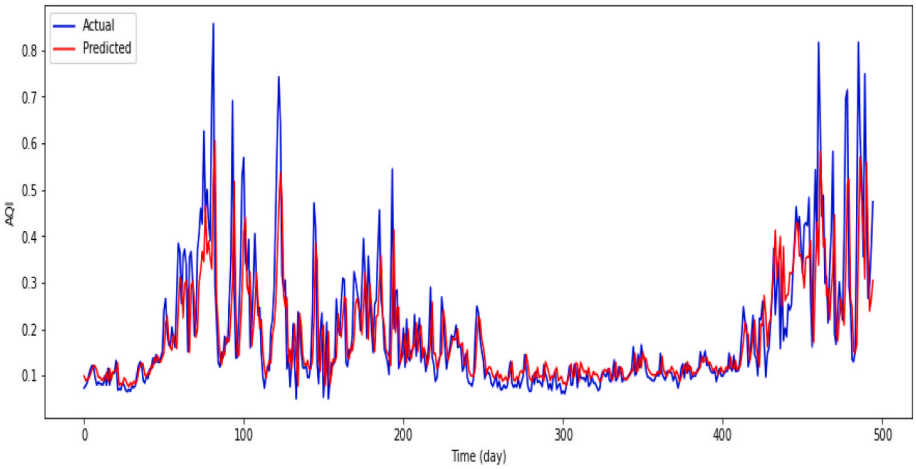

Fig. 14. Actual output Vs Predicted output for Long Short Term Memory.

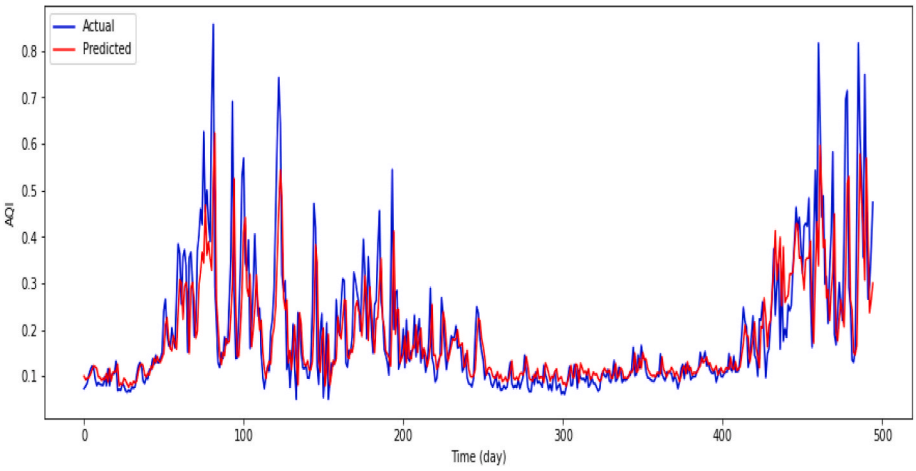

Fig. 15. Actual output Vs Predicted output for Gated Recurrent Unit.

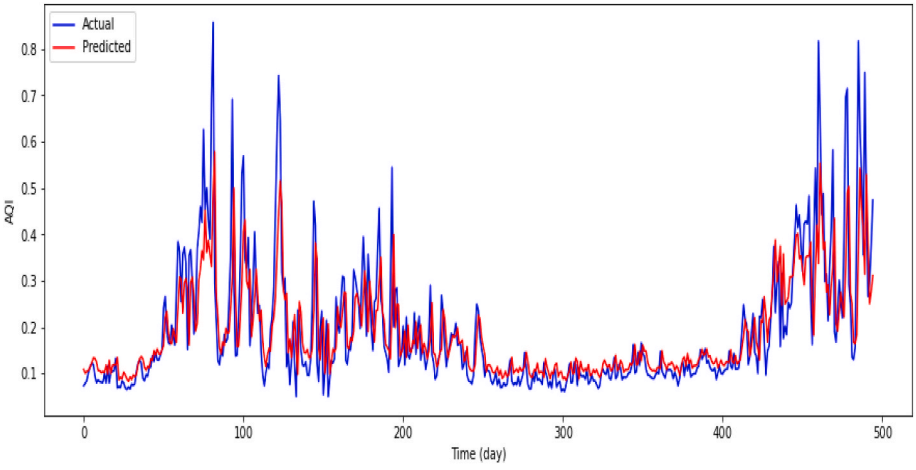

Fig. 16. Actual output Vs Predicted output for LSTM-GRU Model.

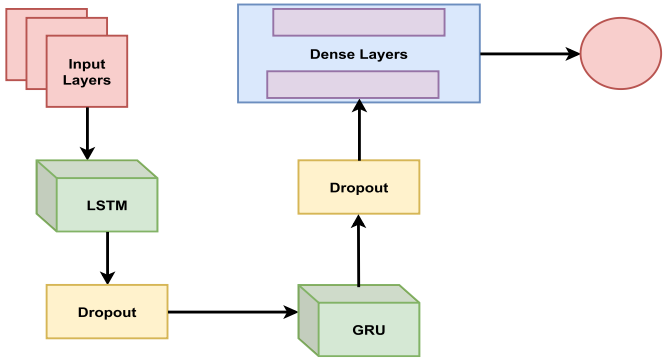

Fig. 17. LSTM-GRU model.

this investigation, RMSE, R^2, and MAE are applied as metrics to assess how well the models are performing.

The optimal result that are obtained after carefully evaluating with various ratios of training-testing data and comparison of the scores of the models as shown in Table 6. The dataset is divided into 80 percent of training data and 20 percent testing data for the purposes of evaluation. This decision is made due to the fact that the optimal result has been obtained after carefully evaluating with different ratios of training-testing data.

Table 7 is representing the metrics wise performance comparison of the different models.

Using the accuracy metric as a point of comparison in Table 7, it has been shown that the MAE 36.11 of the suggested model is lower than that of other models. Hence, it is reasonable to assume that the proposed model is appropriate for projecting data related to air quality. The suggested model, on the other hand, has an R^2 of 0.84, while LSTM has an R^2 of 0.78. This indicates that LSTM-GRU performs better than the LSTM model in this particular scenario. When compared to other models, some measurements provide less favourable findings.

The result of the proposed model is compared with the other existing approaches and the corresponding comparative study is reported in the following table.

5. Conclusion and future scope

A precise and steady AQI forecasting is not only essential for promoting urban public health, but it is also important for maintainable development of environment under the adverse effects of air pollution. This research proposes a hybrid LSTM-GRU model for predicting the AQI in a very polluted city. The performance of the LSTM-GRU model is compared with the other standalone ML and DL models' performance in terms of R^2, RMSE and MAE parameters. The results manifest that the developed hybrid model produces less mistakes than stand-alone models, indicating its superiority. The proposed method can also be extended to predict AQI of other cities jointly for which we must require

Table 6
Performance Comparison based on training validation test split method.

Model	Metrics	60:40	65:35	70:30	75:25	80:20	85:15	90:10	95:05
LR	R^2	0.37	0.38	0.37	0.42	0.40	0.38	0.40	0.42
	MAE	38.87	38.87	38.87	38.86	38.86	38.86	38.86	38.86
	RMSE	61.01	61.20	60.87	57.09	59.22	61.38	56.26	55.01
KNN	R^2	0.42	0.43	0.41	0.47	0.48	0.49	0.52	0.52
	MAE	37.86	37.86	37.86	37.86	37.86	37.86	37.86	37.86
	RMSE	54.90	54.97	55.28	50.65	51.63	52.13	46.66	46.50
SVM	R^2	0.57	0.54	0.54	0.56	0.58	0.58	0.51	0.51
	MAE	49.09	49.10	49.09	49.09	49.09	49.09	49.09	49.10
	RMSE	62.19	64.02	63.64	60.63	61.39	62.59	61.01	60.86
LSTM	R^2	0.73	0.71	0.73	0.78	0.78	0.81	0.79	0.81
	MAE	58.97	58.98	58.98	58.97	58.97	58.97	58.97	58.97
	RMSE	61.37	61.37	61.37	61.37	61.37	61.37	61.37	61.37
GRU	R^2	0.61	0.61	0.63	0.64	0.66	0.64	0.63	0.66
	MAE	42.67	42.67	42.67	42.66	42.66	42.66	42.66	42.66
	RMSE	54.21	54.87	53.52	49.45	50.73	51.25	47.95	46.85
LSTM-GRU	R^2	0.69	0.70	0.73	0.75	0.84	0.72	0.73	0.75
	MAE	36.12	36.11	36.12	36.11	36.11	36.11	36.11	36.11
	RMSE	57.77	57.91	54.33	52.21	52.03	54.43	52.21	51.36

Table 7
Performance comparison.

Model	R^2	MAE	RMSE
Proposed Model	0.84	36.11	52.03
LR	0.40	38.86	59.22
KNN	0.48	37.86	51.63
SVM	0.58	49.09	61.39
LSTM	0.78	58.97	61.37
GRU	0.66	42.66	50.73

Table 8
Comparison of the proposed LSTM-GRU model with other existing approaches.

Sl. No.	Ref.	Classifier	R^2-value
1	Srivastava et al. (2018)	LR, SDG, RF, DT, SVM, CNN, GB,	R^2 = 0.65646, 0.65922, 0.67, 0.62, 0.69275,0.68478, 0.69647, 0.69275
2	Alireza et al. (2021)	HSD, HTPD, CEEMDAN-ELM, CEEMDAN-GRNN	R^2 = 0.74
3	Proposed Approach	LSTM-GRU	R^2 = 0.84

high performance machine along with GPU. In future a large dataset with many features may be created by taking data from different areas for enhancing the experiments with different feature creation and feature selection techniques and classification techniques such as ensemble technique, fuzzy logic techniques etc. with different evaluation metrics such as accuracy, precision, recall etc. where model hyper-parameter optimization can be explored and exploited.

Author statement

Nairita Sarkar: Writing – original draft preparation Conceptualization. Rajan Gupta: Methodology, Software and Data curation. Dr.Pankaj Kumar Keserwani: Supervision. Prof. Mahesh Chandra Govil: Visualization Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data source is cited.

References

Agrawal, S., Sharma, D.K., 2022. Feature extraction and selection techniques for time series data classification: a comparative analysis. In: 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, pp. 860–865.

Alireza, R., Jamil, A., Tzanis, C.G., 2021. Air quality data series estimation based on machine learning approaches for urban environments. Air Q. Atmos. Health 14 (2), 191–201.

Alvi, A.M., Siuly, S., Wang, H., 2021. Developing a deep learning based approach for anomalies detection from eeg data. In: International Conference on Web Information Systems Engineering. Springer, pp. 591–602.

O. I. Azeez, K. B. Adewoye, Mean Square Error in Ml Estimation of Two-Level Time Series Models.

Alvi, A.M., Siuly, S., Wang, H., 2022. A long short-term memory based framework for early detection of mild cognitive impairment from eeg signals', in. IEEE Trans. Emerg. Topics Comput. Intell. https://doi.org/10.1109/TETCI.2022.3186180.

Bhalgat, P., Bhoite, S., Pitare, S., 2019. Air quality prediction using machine learning algorithms. Int. J. Comput. Appl. Technol. Res. 8 (9), 367–390.

Bonaccorso, G., 2017. Machine Learning Algorithms. Packt Publishing Ltd.

Brassington, G., 2017. Mean absolute error and root mean square error: which is the better metric for assessing model performance?. In: EGU General Assembly Conference Abstracts, p. 3574.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Campbell-Lendrum, D., Prüss-Ustün, A., 2019. Climate change, air pollution and noncommunicable diseases. Bull. World Health Organ. 97 (2), 160.

Castelli, M., Clemente, F.M., Popovič, A., Silva, S., Vanneschi, L., 2020. A Machine Learning Approach to Predict Air Quality in california. Complexity.

Chen, J., Jing, H., Chang, Y., Liu, Q., 2019. Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process. Reliab. Eng. Syst. Saf. 185, 372–382.

S. Chen, G. Kan, J. Li, K. Liang, Y. Hong, Investigating China's urban air quality using big data, information theory, and machine learning., Pol. J. Environ. Stud. 27 (2).

Clark, W.A., Avery, K.L., 1976. The effects of data aggregation in statistical analysis. Geogr. Anal. 8 (4), 428–438.

Danihelka, I., Wayne, G., Uria, B., Kalchbrenner, N., Graves, A., 2016. Associative long short-term memory. In: International Conference on Machine Learning. PMLR, pp. 1986–1994.

Dey, R., Salem, F.M., 2017. Gate-variants of gated recurrent unit (gru) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, pp. 1597–1600.

Fan, J., Wu, L., Ma, X., Zhou, H., Zhang, F., 2020. Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renew. Energy 145, 2034–2045.

Ghani, N.A., Hamid, S., Hashem, I.A.T., Ahmed, E., 2019. Social media big data analytics: a survey. Comput. Hum. Behav. 101, 417–428.

Glencross, D.A., Ho, T.-R., Camina, N., Hawrylowicz, C.M., Pfeffer, P.E., 2020. Air pollution and its effects on the immune system. Free Radic. Biol. Med. 151, 56–68.

Gocheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S., Boyadzhiev, D.T., 2014. Time series analysis and forecasting for air pollution in small urban area: an sarima and factor analysis approach. Stoch. Environ. Res. Risk Assess. 28 (4), 1045–1060.

Huang, G., 2021. Missing data filling method based on linear interpolation and lightgbm. In: Journal of Physics: Conference Series, vol. 1754. IOP Publishing, 012187.

Iskandaryan, D., Ramos, F., Trilles, S., 2020. Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Appl. Sci. 10 (7), 2401.
Janarthanan, R., Partheeban, P., Somasundaram, K., Elamparithi, P.N., 2021. A deep learning approach for prediction of air quality index in a metropolitan city. Sustain. Cities Soc. 67, 102720.
Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? use their ratio as well. Inf. Sci. 585, 609–629.
Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y., 2017. Modeling pm2. 5 urban pollution using machine learning and selected meteorological parameters. J. Electr. Comput. Eng.
Lei, J., Liu, C., Jiang, D., 2019. Fault diagnosis of wind turbine based on long short-term memory networks. Renew. Energy 133, 422–432.
Li, J., Li, X., Wang, K., 2019. Atmospheric pm2. 5 concentration prediction based on time series and interactive multiple model approach. Adv. Meteorol.
Li, G., Tang, Y., Yang, H., 2022. A new hybrid prediction model of air quality index based on secondary decomposition and improved kernel extreme learning machine. Chemosphere 305, 135348.
Liu, T., Lau, A.K., Sandbrink, K., Fung, J.C., 2018. Time series forecasting of air quality based on regional numerical modeling in Hong Kong. J. Geophys. Res. Atmos. 123 (8), 4175–4196.
Liu, H., Li, Q., Yu, D., Gu, Y., 2019. Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl. Sci. 9 (19), 4069.
Londhe, M., 2021. Data mining and machine learning approach for air quality index prediction. Int. J. Eng. Appl. Phys. 1 (2), 136–153.
Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., Kedam, G., 2019. A machine learning model for air quality prediction for smart cities. In: 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET). IEEE, pp. 452–457.
Nigam, S., Rao, B., Kumar, N., Mhaisalkar, V., 2015. Air quality index-a comparative study for assessing the status of air quality. Res. J. Eng. Technol. 6 (2), 267–274.
Pal, M., Bharati, P., 2019. Introduction to correlation and linear regression analysis. In: Applications of Regression Techniques. Springer, pp. 1–18.
Pant, A., Sharma, S., Bansal, M., Narang, M., 2022. Comparative analysis of supervised machine learning techniques for aqi prediction. In: 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA). IEEE, pp. 1–4.
P. Partheeban, Application of lstm models in predicting particulate matter (pm2. 5) levels for urban area, J. Eng. Res..
Pozzer, A., Dominici, F., Haines, A., Witt, C., Münzel, T., Lelieveld, J., 2020. Regional and global contributions of air pollution to risk of death from covid-19. Cardiovasc. Res. 116 (14), 2247–2253.
Rights, J.D., Sterba, S.K., 2021. R-Squared Measures for Multilevel Models with Three or More Levels. Multivariate Behavioral Research, pp. 1–28.
Sigamani, S., Venkatesan, R., 2022. Air quality index prediction with influence of meteorological parameters using machine learning model for iot application. Arabian J. Geosci. 15 (4), 1–12.
Singh, R.P., Chauhan, A., 2020. Impact of lockdown on air quality in India during covid-19 pandemic. Air Qual. Atmos. Health 13 (8), 921–928.
Srivastava, C., Singh, S., Singh, A.P., 2018. Estimation of air pollution in Delhi using machine learning techniques. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, pp. 304–309.
Tang, Y., Huang, Y., Wu, Z., Meng, H., Xu, M., Cai, L., 2016. Question detection from acoustic features using recurrent neural network with gated recurrent unit. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6125–6129.
Turky, S.N., Al-Jumaili, A.S.A., Hasoun, R.K., 2021. Deep learning based on different methods for text summary: a survey. J. Al-Qadisiyah Comput. Sci. Math. 13 (1). Page–26.
Van, N., Van Thanh, P., Tran, D., Tran, D.-T., 2022. A new model of air quality prediction using lightweight machine learning. Int. J. Environ. Sci. Technol. 1–12.
Wu, Q., Lin, H., 2019. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci. Total Environ. 683, 808–821.
Yu, C., Yao, W., 2017. Robust linear regression: a review and comparison. Commun. Stat. Simulat. Comput. 46 (8), 6261–6282.
Yuan, Y., Liu, M., 2014. Discussion on the difference between air quality index (aqi) and air pollution index (api) j. Guangzhou Chem. Ind. 42, 164–166.
Z. Zhang, Missing data imputation: focusing on single imputation, Ann. Transl. Med. 4 (1).
Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2017a. Efficient knn classification with different numbers of nearest neighbors. IEEE Transact. Neural Networks Learn. Syst. 29 (5), 1774–1785.
Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D., 2017b. Learning k for knn classification. ACM Trans. Intell. Syst. Technol. (TIST) 8 (3), 1–19.
Zhou, X., Xu, J., Zeng, P., Meng, X., 2019. Air pollutant concentration prediction based on gru method. In: Journal of Physics: Conference Series, vol. 1168. IOP Publishing, 032058.
Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., Che, J., 2017. Daily air quality index forecasting with hybrid models: a case in China. Environ. Pollut. 231, 1232–1244.
Zhu, J., Wu, P., Chen, H., Zhou, L., Tao, Z., 2018. A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. Int. J. Environ. Res. Publ. Health 15 (9), 1941.
Zou, B., You, J., Lin, Y., Duan, X., Zhao, X., Fang, X., Campen, M.J., Li, S., 2019. Air pollution intervention and life-saving effect in China. Environ. Int. 125, 529–541.