Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

AI Based Air Pollution Forecasting for PM2.5 Pollutants

Khushboo Mahawar[1], Dr. Poongodi C[2*], Dr. Rekha V[3]

[1, 2, 3] Department of AI and Data Science Engineering, School of Engineering and Technology Christ University, Bangalore, Karnataka, India

Abstract:- Accurate air quality forecasting is critical to reducing health risks associated with poor air quality, and it also aids in urban planning. However, many locations do not have sufficient sensing infrastructure to generate reliable data and are reliant on fragmented datasets that are often noisy. [8] PM2.5 is an important pollutant in the air that we measure, and we use various forecasting methods to determine its level. Both time-series analysis and deep-learning techniques have been used to create models to forecast PM2.5 concentrations. We trained LSTM Networks, ANN-LSTM hybrids and Prophets on daily PM2.5 data. The Bi-LSTM and LSTM Networks provided superior accuracy and performance when compared with all other models. This paper describes the creation of a unified, data-driven air-quality forecasting framework that can generate short-term forecasts based on historical air pollution records in areas where there is no air quality monitoring system in place. This framework utilises robust preprocessing, outlier detection, median/mode imputation, temporal feature extraction, a systematic exploratory analysis of the pollutant dynamics, and the development of seven different models of air quality forecasting using Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, XGBoost, Artificial Neural Networks, and LSTM time-series modelling. Experimental evaluation demonstrates that the nonlinear and deep learning models outperform the linear models by a significant margin, [10] with ANN achieving an approximately 93% accuracy rate and LSTM producing effective forecasts with the use of temporal dependencies. In contrast to sensor-only systems, the air quality forecasting framework described herein provides a cost-effective, scalable, and practical alternative to generating real-time alerts, decision-making support, and research insight from data-sparse environments, and is well-positioned for further integration with live application-programming interfaces, meteorological data, and hybrid technologies to continually improve predictive reliability.

Keywords: Air pollution, machine learning, deep learning, time-series forecasting, LSTM, artificial neural networks, environmental monitoring, regression models, data preprocessing, predictive modeling.

1. Introduction

Pollution within the atmosphere has become a major environmental concern—impactful on the health of humans and the balance of the environment and sustainability of our cities. Yearly increase of factories, vehicle traffic, and an overall increase in the number of people displaced from rural areas has resulted in major increases of pollutant emissions from factories, vehicles and through population density. Extended periods of exposure, to extremely high levels of airborne pollution, are strongly linked to respiratory disease and cardiovascular disease, decreased quality of life and increased costs to our healthcare systems. Therefore, understanding the current condition of air quality and being able to predict air quality trends is critical for managing our environment and developing plans for public health.

Historically, air quality has been monitored primarily using fixed monitoring systems requiring fixed sensors placed at the fixed locations continuously collecting, measuring and recording the concentrations of the pollutants present in the air. The use of fixed monitoring stations is capable of providing data with a high level of reliability. The cost to establish and maintain a monitoring station is high, limiting their use to major urban centres. [10][23]A majority of locations around the world lack either complete or timely access to the information regarding the

levels of pollution they are exposed to on a daily basis. Additionally, monitoring systems have focused primarily on providing real-time observations of air quality and limited ability to provide accurate forecasting of pollution levels. These limitations indicate the potential requirements for developing a data-driven predictive modelling approach, which could enhance existing monitoring infrastructure.

The development of machine learning and deep learning models has resulted in providing significant new modelling capabilities in developing models for environmental parameters, capturing the non-linear relationships between environmental variables; developing temporal dependencies; Capturing the non-linear relationships between variables and building relationships over time.

The aim of this study will be to overcome these obstacles through identifying the best imputation techniques for the treatment of missing air pollutant data as well as to compare the performance of different types of forecasting models when predicting PM2.5 concentrations in India. Data will be used from published records of ITO, India and collected through the Central Pollution Control Board (CPCB).[8][3] A variety of imputation techniques including linear regression, forward?backward fill, Fourier?KNN, linear interpolations and other statistical approaches, have been assessed during the imputation process. After imputation, a variety of forecasting methods will be evaluated to see how well they performed at predicting PM2.5 levels, these include Long Short Term Memory (LSTM).

Thanks to advances in machine learning and deep learning technology there are now more ways than ever to model complex environmental processes (e.g. weather, natural disasters, etc.). Recent developments have provided the ability to accurately represent complex, non-linear relationships and learn from empirical data (historical datasets) whilst also being able to determine any temporal dependencies or meaningful information contained within those datasets.[11] Time-series models are exceptionally well suited to provide accurate forecasts of air pollution with LSTMs providing a means of modeling both the sequential and seasonal nature of air pollution. However, before a forecast can be made with high degrees of accuracy, there are many types of inconsistencies, noise and missing values that are often found within environmental datasets that must be taken into account to ensure accurate predictions.

In this study we propose a forecasting framework whereby we seek to integrate rigorous data preprocessing; exploratory analysis and comparative analyses of a selection of machine learning/deep learning models in an effort to create a reliable way to produce short-term air pollution forecasts in those areas where there is little to no access to real-time monitoring data to assist with future planning and/or decision making related to air quality and the environment.

2. Related Work

Air Quality Prediction has been studied extensively because air quality can impact community health and the environment. Historical research focused primarily on using Statistical Models and Deterministic Models (i.e., ARIMA, Linear Regression) to determine air pollutant concentration levels. The use of these techniques to predict air quality was very basic; due to their assumptions of linearity, they could not effectively model the complex, non-linear nature of many air pollutants.

As more data becomes readily available and computing technology continues to advance, Machine Learning (ML) techniques are now being used to analyze air quality. Examples of the types of ML methods include Decision Trees, Support Vector Machines and Random Forest Regressors for air quality predictions based on the non-linear relationships between various environmental factors and air quality indicators. Several studies indicate ML ensemble methods have demonstrated improved accuracy over single learners by reducing over-fitting and by more effectively modelling feature interactions.[1][3][10] However, it is important to note that most of these ML methods rely on completely filled datasets and have reduced accuracy when faced with either missing or erroneous data (i.e., real-world environmental data).

The use of deep learning models for air quality prediction is currently getting attention. The use of Artificial Neural Networks has shown to be able to predict complex non-linear characteristics of pollutant behaviour well. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, are frequently used for

making time-series predictions of air pollution because of their capacity to capture temporal and long-term relationships of the data. These models have provided good performance for both short-term and medium-term prediction tasks, particularly in circumstances where the diurnal and seasonal variations are important considerations.

Even though there have been a number of studies to date showing that deep learning methods are useful for forecasting air quality, a number of current limitations still exist within the scientific literature.[4][5] The dependence on either real-time sensor networks or external meteorological APIs in many of these studies limits their usefulness to areas that do not have the same types of infrastructure. Some studies that have studied air quality prediction have examined only one modelling approach (and not made any comparisons), which prevents an easy identification of which models produce the best results depending upon the types of data being used and the types of information it provides. In addition, there has been very little research to date on coming up with robust pre-processing strategies for dealing with missing values, inconsistencies and outliers, before the training of these models.

3. Literature Review

The impact of air pollution on human health, climate change and urban development makes predicting air pollution an important area of study. Early air pollution studies used only traditional statistical methodologies (i.e., linear regression and autoregressive models) to predict pollutant concentrations.[1][4][9] The traditional statistical approaches gave researchers an easily interpretable method for estimating the concentration of pollutants but had limited predictive ability because they were unable to fully account for the non-linear relationships and temporal complexity found in air quality data. Because air pollution levels are the result of many interrelated factors, statistical-based air pollution models often result in significant forecasting errors under rapidly changing environmental conditions.

As researchers became increasingly aware of the unique aspects of air quality, they began utilizing machine learning algorithms to improve air quality forecasting.[3][8] The two most notable types of machine learning algorithms for air quality forecasting are Support Vector Regression (SVR), Decision Trees and Random Forests due to their inherent ability to model non-linear relationships and to work with multi-dimensional feature spaces. The majority of research using machine learning applies the ensemble approach of combining multiple machine learning models into one model. The ensemble approach reports greater accuracy than individual model predictions.[8] A disadvantage of this approach, however, is that it assumes that only complete and accurate datasets are available when used to predict air quality. Currently, missing and erroneous data from environmental monitoring systems continue to pose significant obstacles to accurate air quality forecasts.

The advancement of pollution forecasting research from DL methods; with the ability to automatically learn features from large volume sets (or datasets) has helped create new advances within how pollution is forecasted.[10] The utilization of artificial neural networks has successfully enabled reproducible models based off of complex interactions between multiple pollutants; while recurrent structured neural networks (i.e. Long short Term Memory) have been very successfully applied to time-based forecasts.[31] The use of LSTM neural networks allows researchers to capture temporal, seasonal, and long range (longer than one year) dependencies associated with previous patterns of pollution data.

More recent work has also emphasised the critical role of feature extraction/data pre-processing steps in enhancing the accuracy of prediction results.[1][8][15] Studies have demonstrated that one must effectively deal with missing values, normalise input data, and extract temporal features before training any model to effectively achieve reliable predictions.[10] Furthermore, despite the need to incorporate effective exploratory analysis and comparison techniques when evaluating models, only a few studies have created complete modelling frameworks that incorporate all these considerations.

This study presents a comprehensive framework which integrates historical pollution datasets with systematic pre-processing and exploratory analyses so as to compare machine learning vs deep learning solutions on their predictive reliability within environments where data is limited). [12][14][15] Thus, This work seeks to create

more scalable, as well as viable applications of air quality forecasting through use of improved analytical methodologies, than has been previously established.

4. Dataset Description

This research study uses a data set that contains air pollution information compiled at many different air quality monitor sites. The data includes information about pollution concentrations taken in many cities and states during a particular time. This allows for both a spatial and temporal analysis of the pollution trends.[8] The individual records of air pollution at a given site and time are contained in the data set with information on how much of each pollutant type is present.

Numerical attributes (that is, attributes that are real numbers) and categorical attributes (that is, attributes that consist of words or phrases) comprise the data set. Numerical attributes contain the concentrations of each pollutant measured, as well as the geographical coordinates for the monitors and summary statistics about each concentration at that location. Categorical attributes provide information on where pollution was measured (that is, city, state, and pollutant type), and include timestamps for when each measurement was taken. These timestamps will ultimately be broken down into year, month, day, and hour so that they can be used to do time-series analysis and forecasting.

The dataset possesses a diverse range of characteristics, including inconsistencies in reported pollutant levels and variations in pollutant levels by geographic location, owing to difficulties in collecting actual/real-world data. The variation in these characteristics makes the data particularly useful for developing robust preprocessing methodologies and predictive models that can work effectively within the framework of realistic conditions.

Table 1. Required Libraries and techniques

Library	Purpose in This Project
NumPy	Needed for efficient numerical operations, especially for ML, LSTM, CNN calculations.
Pandas	Dataset loading, handling missing values, grouping, aggregations, EDA, preprocessing.
Matplotlib	For base-level visualizations like histograms, line plots of pollution trends.
Seaborn	For correlation heatmaps, boxplots, pairplots, pollutant distribution analysis.
Warnings	Removes clutter so we focus only on meaningful outputs.

Pollutant Description:

Key air pollutants have been identified that have a significant impact on human health and environmental quality due to their physical and chemical properties, as well as how they affect how air pollution is formed and distributed. Each of the pollutants listed below has its own physical and chemical characteristics and adds to the different ways that we understand how air pollution occurs.

Particulate Matter (PM2.5)

Particulate matter with a diameter of 2.5 micrometers or less is defined as PM2.5. These tiny particles can be inhaled, enter deep into the lungs, and to some extent, directly into the bloodstream. PM2.5 has the potential to adversely affect respiratory and cardiovascular health. It is one of the most important constituents used to assess the quality of air in an urban area and is strongly affected by emissions from vehicles, industries, and burning fuels.

Particulate Matter (PM10)

Particulate matter with a diameter up to 10 micron is called PM10. Although the diameter is larger than that of PM2.5, it still has the potential to adversely affect respiratory health by affecting primarily the upper airways. PM10 can originate from road dust and from construction sites as well as from natural sources like dust storms.

Nitric Oxide (NO2):

The principal source of nitric oxide is automobile emissions and industrial combustion processes. It acts as a precursor for other pollutants, namely ozone, and particulate matter. Chronic exposure to NO2 can cause respiratory problems, including decreased lung function, and irritation of the respiratory tract.

Sulfide (SO2):

Sulfide is primarily generated from fossil fuel combustion for energy generation and industrial activity, contributing to acid rain formation and causing respiratory issues among sensitive populations.

CO (Carbon Monoxide):

Carbon Monooxide (CO) takes on the state of a gas that has neither colour nor smell when it is produced by the incomplete burning of carbonated fuels. When too much CO accumulates in the body it limits the amount of O2 being delivered to the bloodstream, which causes dizziness, tiredness, and even death.

02 Ground-Level Ozone (Ozone O_3):

Ground-Level Ozone (Ozone O_3) is an example of a secondary pollutant that happens during the process of combining nitrogen oxides and volatile organic chemicals using the energy from sunshine. Ground-Level Ozone (Ozone O_3) has a negative impact on human health, as it causes people to have increased asthma attacks, decreased lung performance and plant life, and much more.

Target Variable

The study's main objective will be to determine an average concentration of these pollutants, which we will refer to as pollutant_avg, and this will serve as the main outcome variable for both predictives using regression and those predictive methods based on time.

Characteristics of the Dataset

The dataset contains temporal, seasonal, and regional variability and shows the difference in how pollutants are released based on Geographic Location, Seasonality, Urban Activity Patterns, and Environmental Conditions and as such provides an excellent database for exploratory analysis, predictive regression modelling, and forecasting using Deep Learning Methods.

5. Data Preprocessing and Feature Engineering

Environmental air quality datasets typically contain large amounts of missing values that are due to a variety of factors, including sensor failure, inconsistent data collection (e.g., missing days or times), and transmission errors (i.e., improper transmission during switching to different networks).[8][9][24] These factors significantly affect the quality of input data, which, in turn, severely impact the reliability of the resulting predictive model. For this reason, extensive pre-processing of these datasets was conducted prior to establishing the predictive model. The goal of this stage of the project was to improve data quality and to retain important patterns in the data necessary for accurate air pollution forecasting.

An initial review of the dataset indicated that there were a number of missing values throughout the dataset, both in numerical and categorical attribute types. If we had removed all records that contained any missing attributes, a significant percentage of the available information would have been lost;[15][18][21] therefore, we applied specific imputation strategies based on the characteristics and distributions of each attribute. For the numerical pollutant concentrations, we used primarily the median value as an imputation strategy, as many air pollution datasets are highly skewed and have occasional outliers due to sudden emissions. The use of the median provides a more stable estimate of central tendency that is less sensitive to outliers than the mean. The mean approach to imputing values for certain numerical values with somewhat symmetrical distributions was only used selectively in order to maintain the long-term trends in the data without introducing bias to the predictive model.

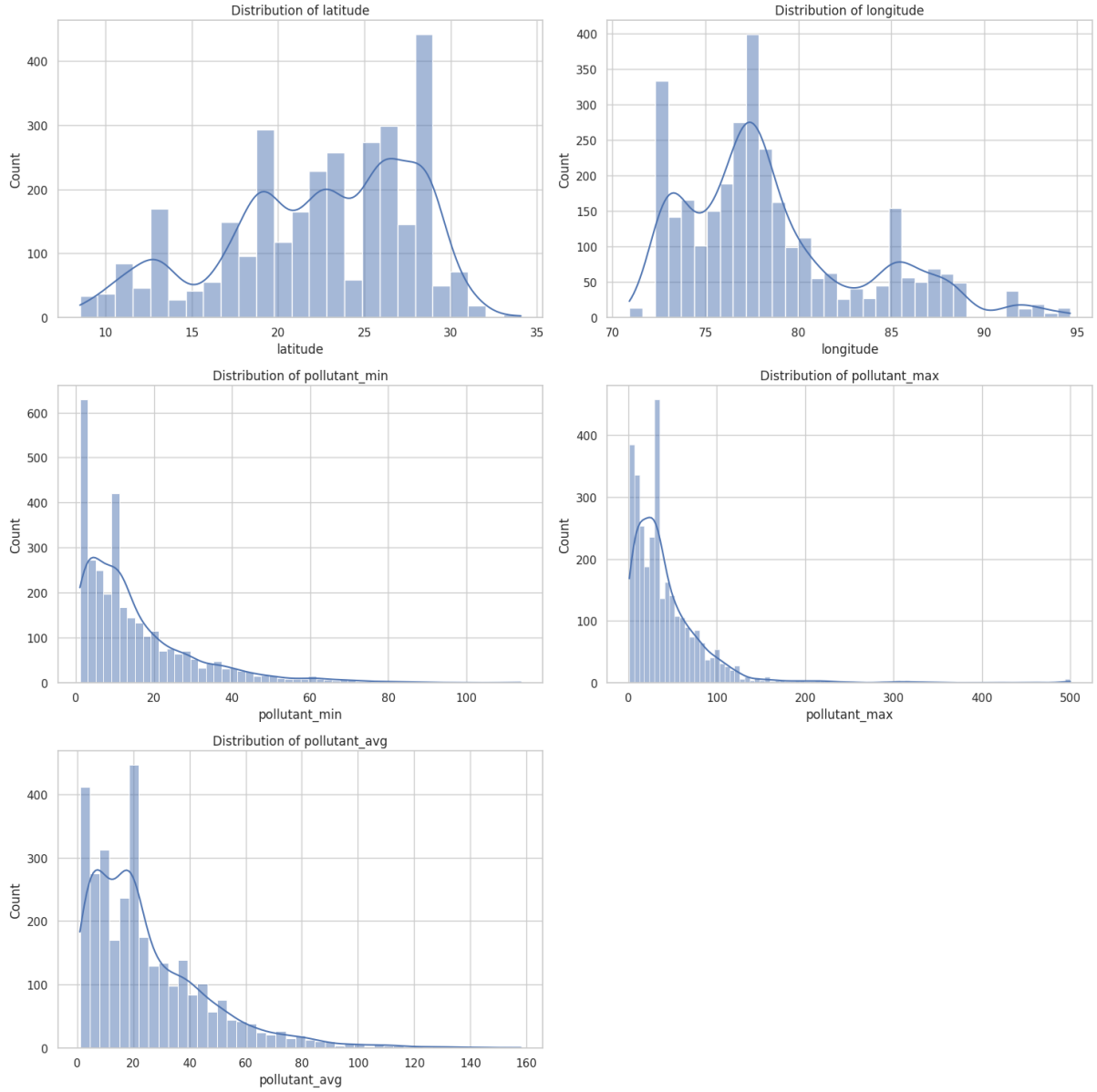

Figure 1. Shows the numeric variables behave individually in the dataset

Imputation of categorical variables includes location and pollutant categories which were completed by entering the most common (mode) value in each column. Doing so provides semantic alignment and avoids the introduction of artificial or ambiguous category types that hinder effective model learning. Through the combination of these two imputation methods, both types of data were treated equally and the total amount of distortion to the existing data was kept to a minimum.

The timestamp was converted into standardised date/time format to facilitate temporal analysis. Then new temporal features were created from this field including year, month, day, and time to allow the model to learn how to represent seasonal changes, daily traffic patterns, and hourly changes in pollution levels. The extraction of temporal features is essential for time series models because pollution is often strongly influenced by time.

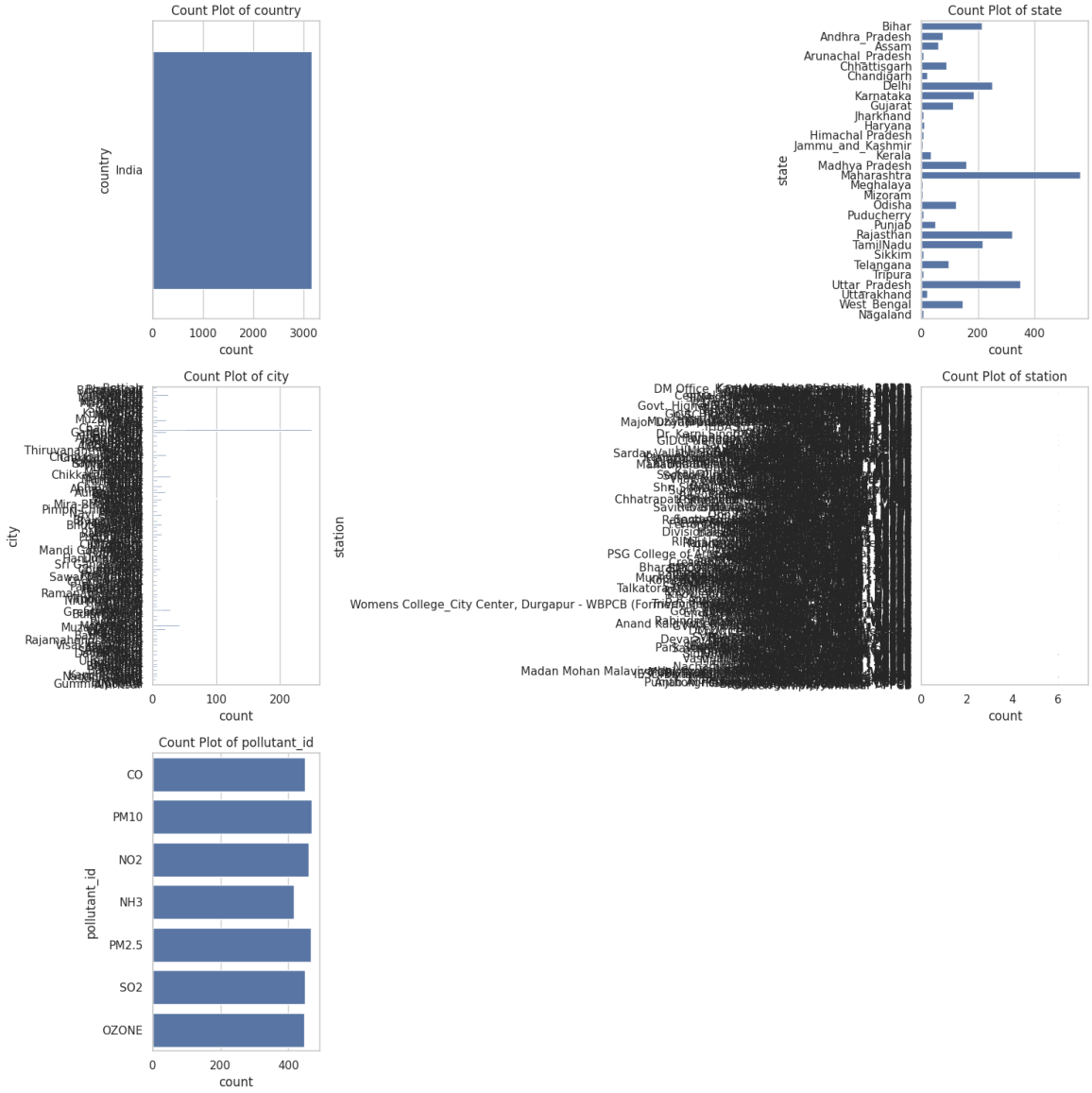

Figure 2. Shows the frequency of each category in a column

Data cleaning that was performed on this dataset also consisted of removing records with missing or invalid target values as well as correcting any data that did not contain consistent data types. In order for numerical features to be trained properly on certain algorithms (namely gradient descent and neural network algorithms), those numerical features were rescaled. Finally, a subset of relevant features was selected based on both domain expertise and exploration analytics.[32] The average concentration of pollutants for each location was used as the target variable. Through all the above processes, a consistent foundation was established for all the model-building and evaluation processes described in subsequent sections of this document.

6. Exploratory Data Analysis

The initial stages of EDA help to further support and understand the degree to which the pollutants produced an overall dataset, as well as the potential for predictive modeling. Ultimately, EDA helps to validate the quality of data collected and provides a comprehensive understanding of how pollutants interact and behave with respect to each other as well as assist in the selection of features for predictive modeling.[8][10][16] EDA was conducted using statistical and visual techniques to explore each of the pollutant variables individually along with exploring the correlation between pollutant variables, ensuring that all of the modeling decisions made going forward are supported by empirical evidence.

The distribution of the numeric variables produced clear differences between the concentrations of the pollutants. The PM2.5 and PM10 concentration variables are both right-skewed distributions, indicating that there are periods of time in which the pollutant concentrations are higher than what would normally be expected and many times these are due to an extreme pollution event. The spikes in concentrations are consistent with real-life occurrences such as heavy traffic, industrial emissions, and seasonality. On the other hand, the distribution of the remaining numeric pollutant variables produced relatively consistent distributions, exhibiting relatively uniform behavior over the course of the study, thus validating the medians of the pollutant variables to be used in place of the means. Additionally, these distributions reinforced the requirement for the use of robust modeling techniques that account for skewed data.

Through bivariate analysis, we learned how pollutants' concentrations relate to each variable. The scatter & box plots provided examples of both positive & negative correlation,[17][19] depending upon the pollutant being analyzed and its context. Certain pollutants displayed a tendency to be higher at certain times of the year and/or day (e.g., pollutants from cars during rush hours), while the pollution levels at different locations across different states affected by industries and transportation also have different levels of pollution than each other based on environmental factors.

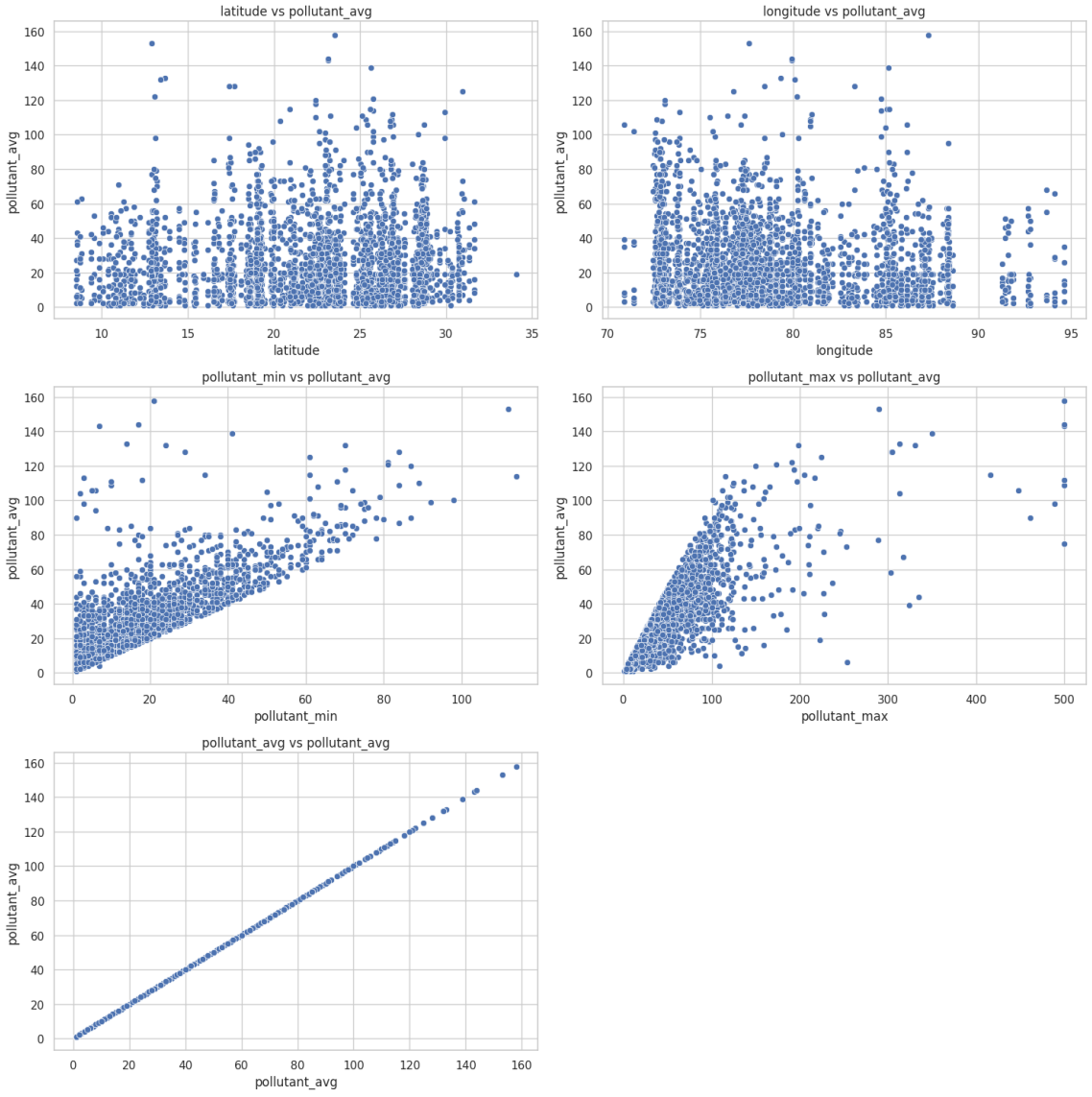

Figure 3. Bivariate analysis (Supports identifying conditions (like high temperature or low wind speed) associated with higher pollution)

Based on these findings we recognize that the level of air pollution is influenced not only by one variable (e.g., a specific time of day), but multiple variables working together.

Multivariate analysis was performed to explore how multiple numeric features can interact together as a whole.[19] Visually, by comparing the different correlated variables, such as minimum and maximum value, on each individual pollutant produced some of the highest positive correlations. While some pollution-related variables had a negative or weak correlation with how they influenced pollution levels. The application of our correlation heat maps helped us identify and eliminate redundant variables and to determine which variables have the greatest impact on the prediction of pollution-related variables.

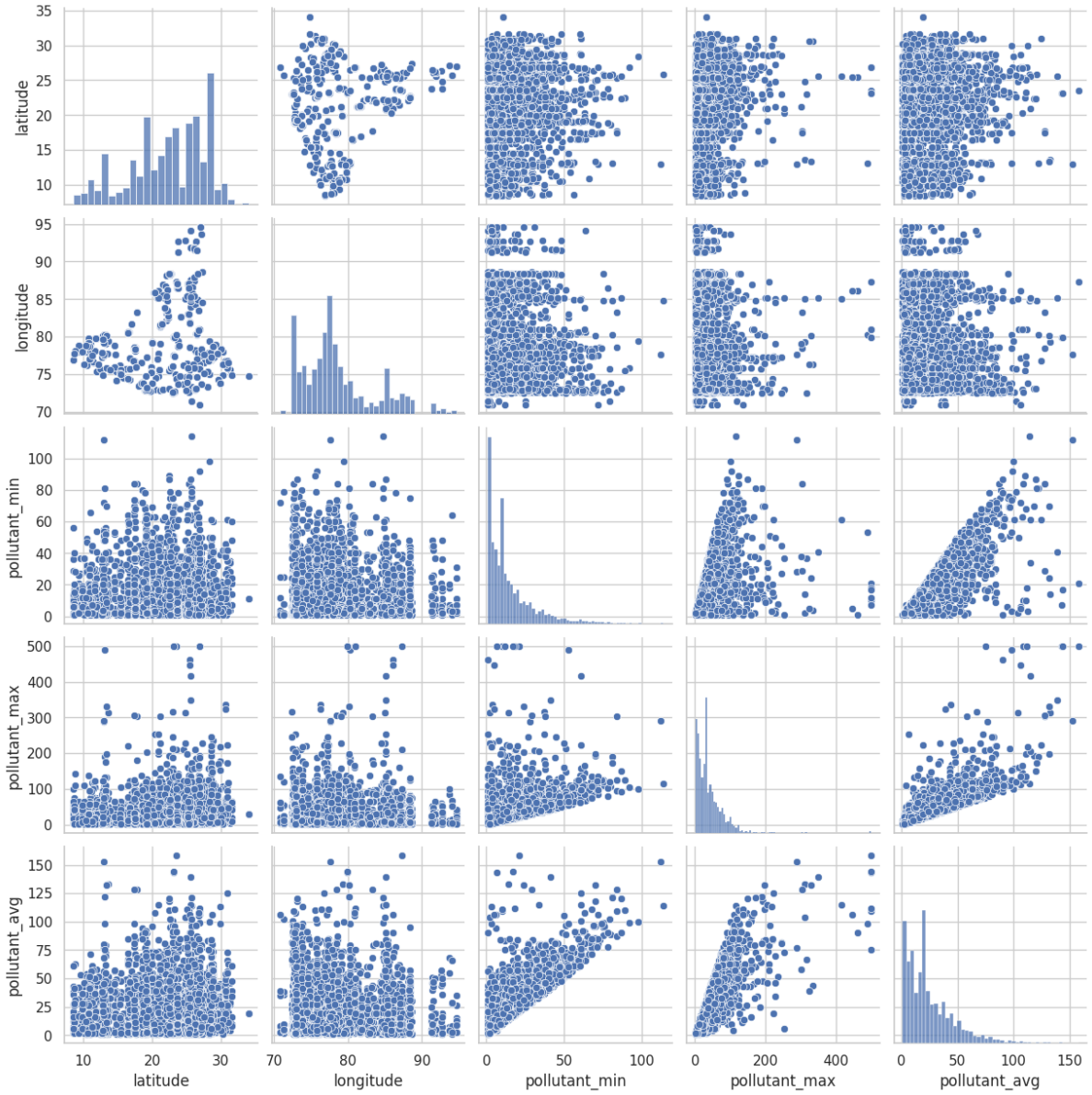

Figure 4. Multivariate analysis (relationships between all numeric variables simultaneously.)

The results of exploratory data analysis indicated that the patterns of air pollution are not linear or uniformly distributed over time, but rather are influenced by both location and seasonality. This provides evidence that the use of ensemble learning and deep learning models for analysis in the next phase of research would be useful in recognising complex interactions in environmental data.

7. Proposed Methodology

To assist with the forecasting of air pollution levels, a hybrid approach was employed, which combined conventional machine learning and deep learning techniques with time-series forecasting to provide an operational workflow to predict air quality. [20][22][29][30] The proposed approach consists of utilizing historical pre-processed datasets, creating features, and training the model, before being evaluated, and identifying future pollutant levels as predicted by the model. In addition, this hybrid methodology is designed to capture the static relationship between the variables and also the dynamic time-series relationships that affect pollutants in the environment.

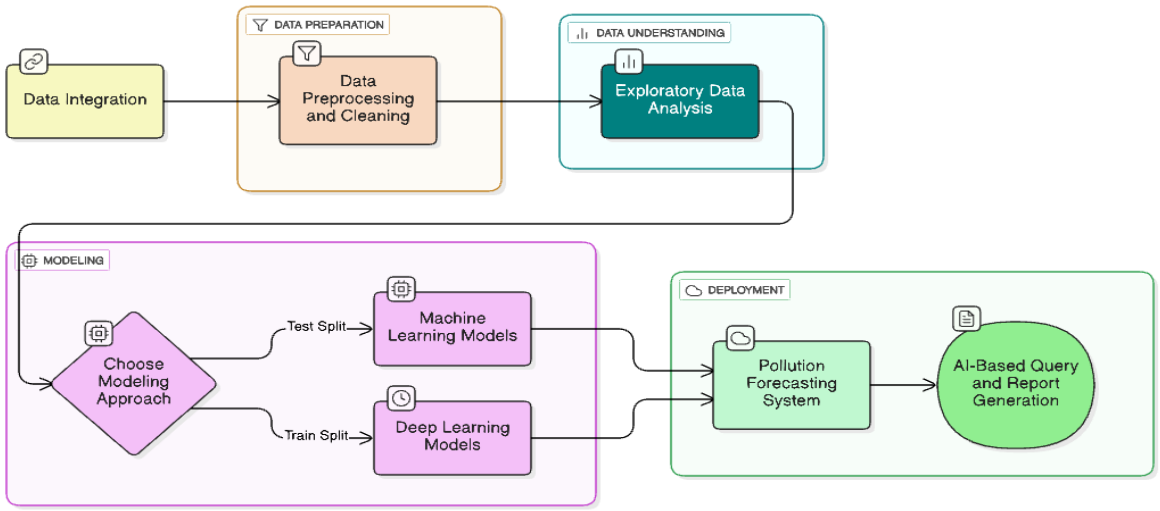

Figure 5. Implementations of Objectives

The problem statement indicates a supervised regression problem, where the prediction is made for the average pollutant level given a set of input features. The database used to represent this is as follows:

$D = \{(xi,yi)\}i=1N,$

where $xi \in Rm$ *represents the input features and* $yi \in R$ *represents the predicted pollutant concentration.*

The objective is to determine $f(\cdot)$ *that reduces the error between yi and yi^, where* $yi^=f(xi)$.

Multiple machine learning models were employed to learn this mapping, including Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, and XGBoost. Linear Regression serves as a baseline model and assumes a linear relationship between features and target, expressed as

$y^=\beta0+\sum j=1m\beta jxj \hat{y} = \beta_0 + \sum_{j=1}^{m} \beta_j x_j y^=\beta0+\sum j=1m\beta jxj.$

While simple and interpretable, this model is limited in capturing nonlinear interactions present in environmental data.

Ensemble-based models such as Random Forest and Gradient Boosting were incorporated to address nonlinear dependencies. Random Forest predicts outputs by aggregating results from multiple decision trees, expressed as

$y^=1T\sum t=1Tht(x) \hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) y^=T1\sum t=1Tht(x),$

where hth_tht represents the prediction of the ttt-th tree. These models improve generalization by reducing variance and learning feature interactions effectively.

To further capture complex nonlinear patterns, Artificial Neural Networks were implemented. ANN models learn hierarchical feature representations through weighted transformations and nonlinear activation functions. The output of a neuron is defined as

z=σ(Wx+b)z = \sigma(Wx + b)z=σ(Wx+b),

where WWW represents weights, bbb is the bias, and σ(·)\sigma(\cdot)σ(·) is a nonlinear activation function. ANN models demonstrated strong predictive capability due to their flexibility and ability to approximate complex functions.

For time-dependent forecasting, a Long Short-Term Memory network was adopted. LSTM is designed to model sequential data by maintaining a memory state that evolves over time. The internal operations of an LSTM cell are governed by gating mechanisms, defined as

ft=σ(Wf[ht−1,xt]+bf)f_t = \sigma(W_f [h_{t-1}, x_t] + b_f)ft=σ(Wf[ht−1,xt]+bf),

it=σ(Wi[ht−1,xt]+bi)i_t = \sigma(W_i [h_{t-1}, x_t] + b_i)it=σ(Wi[ht−1,xt]+bi),

ct=ft⊙ct−1+it⊙tanh⁡(Wc[ht−1,xt]+bc)c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c [h_{t-1}, x_t] + b_c)ct=ft⊙ct−1+it⊙tanh(Wc[ht−1,xt]+bc),

ht=ot⊙tanh⁡(ct)h_t = o_t \odot \tanh(c_t)ht=ot⊙tanh(ct),

where ftf_tft, iti_tit, and oto_tot denote the forget, input, and output gates, respectively. These mechanisms enable the model to retain long-term dependencies and effectively forecast future pollution levels.

The integration of machine learning and deep learning models allows the framework to address both spatial-feature relationships and temporal dynamics. While regression and ensemble models focus on feature-driven prediction, LSTM emphasizes sequential learning for future forecasting. This complementary design ensures robust performance across different prediction scenarios.

The designs of individual models were consistent across all learning methods. Each model was trained and tested on the same preprocessed data and feature set. By using the same data and features for all models, any differences observed in performance could directly relate to the models themselves and not differences in how the data were processed. Having the same data and feature sets makes it possible to compare multiple algorithms that employ different statistical techniques within one predictively-oriented framework.

Additionally, the proposed methodology placed an equal emphasis on understanding the workings of the models and maximizing predictive accuracy. When working with the ensemble models, it was possible to conduct a feature importance analysis that identified the variables that contributed significantly to predicting pollution levels. Knowing which features are most influential in creating predictions will provide useful information concerning how pollution behaves and increase the transparency of environmental decision making, thereby enhancing the practical use of the proposed methodology.

Although past pollution records were used to develop and evaluate the methodology, a key component of the methodology will be its ability to adapt and incorporate additional input data as it becomes available. The current version was built primarily around historical pollution record data; however, the framework can easily accommodate other types of input data (i.e., meteorological inputs and/or real-time sensor data). This adaptive feature allows for the continuous evolution of the architecture of the system.

8. Experimental Setup and Evaluation Metrics

All experiments were carried out in a controlled and reproducible manner to provide the most rigorous and unbiased assessment possible. [8] To do this, the pre-processed dataset was separated into training and testing sets through fixed-split ratio; the majority of data was available for training and the balance of data was set aside for performance evaluation. By limiting access to the testing set during training, the models were able to be assessed on the basis of their ability to generalise to unseen samples.

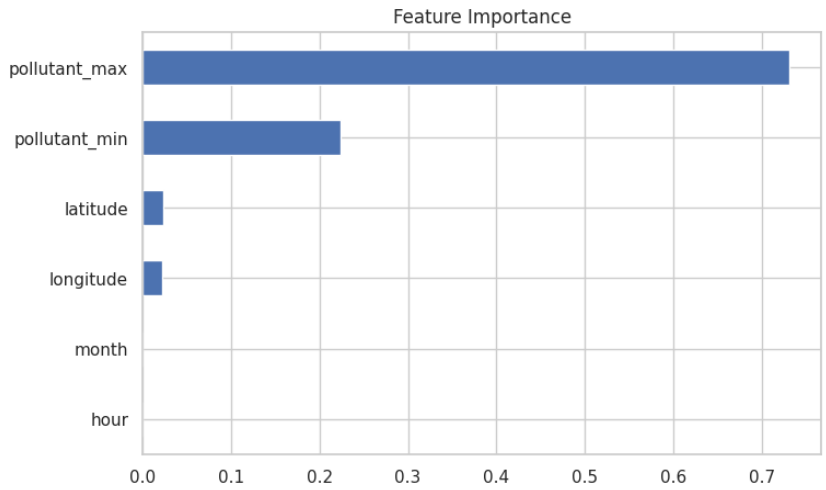

Figure 6. Shows how this feature is Useful for targeting pollution control measures or improving forecasting models

The feature set and target variable for all machine learning models were identical to each other so that a fair comparison could be made. Ensemble model hyperparameters (such as Random Forest, Gradient Boosting and XGBoost) were selected according to common industry best practices and through empirical tuning, in an effort to achieve a balance between model complexity and performance.[17] A linear regression model was used as a baseline against which the performance of nonlinear models could be measured. Deep learning models were constructed with the Artificial Neural Network (ANN) using more than one hidden layer and optimised with gradient descent, while the Long Short Term Memory (LSTM) model was used with sequential time-windowed data to account for the temporal nature of the training data.

Multiple epochs were used to train models wherever possible, and when using Batch Optimization on Neural Networks, Batch Optimization enabled models to stabilize while converging. Careful consideration was taken to minimize the likelihood of overfitting by regularly monitoring the performance of each Machine Learning model throughout the training process and limiting model complexity to the greatest degree possible.[25] The same computing environment was used for all experiments to maintain consistency with respect to runtime characteristics and reporting of model performance results.

The most suitable Regression Metric was chosen for evaluating predictive accuracy. The measure of the ability of a Model to explain the variance within actual pollution levels vs. predicted values was recommended as the primary Metric, meaning that all Machine Learning Models are to be assessed based upon the coefficient of determination (R^2). The definition of R^2 can be found in the following formula:

$$R^2 = 1 - (\Sigma (y_i - \hat{y}_i)^2) / (\Sigma (y_i - \bar{y})^2)$$

(yi)= actual observations of the target variable (yi^)= predictionsfor the target variable (y-)= mean target variable value across all observations (i.e. observations used for training, validating, and testing) R^2 values indicate how well a model's predictions agree with the actual observation(s). High R^2 indicates a Model's ability to predict the target variable accurately.

In using Deep Learning and Time Series Models as a comparison metric for both, the MAE was chosen due to the fact that it offers the greatest Interpretability and robustness.[8][9][20] The MAE offers a “per-mean error” in that it does not exaggerate outlier effects and has been specifically designed for Pollution Data containing occasional spikes. This research provides a uniform, equitable, and reasonable methodology by which to compare Numerous Different Predictive Models. All model results presented reflect the ability to predict accurately and are not biased towards experimental errors/conditions, as all models were trained using the Same Data Set in a manner that provided the Same Evaluation Criteria and Controlled Training Conditions. In the following paragraphs a comprehensive discussion of All Results and their Respectful Implications are discussed.

9. Model Designing

The creation of the model was completed through a staged approach to incremental improvement of accuracy and temporal cognition for air pollution. Rather than going the route of one singular algorithm, a multi-model system was employed, where classic machine learning techniques were first utilized and then sophisticated time-series forecasting supported by deep-learning technology was used thereafter, through a progressive approach; this concept of layered-structured methodology produces assurance, comparability, and enlargement of the possibility of the system in question.

To commence, a selection of five types of machine learning-based regression models was used, focusing upon the production of a base-line premise to predict pollution. These model types were:

1. Linear regression;
2. Random Forest Regressor;
3. Gradient Boosting Regressor;
4. Support Vector Machines (SVR); and
5. XGBoost Regressor.

Each of these models was created using the same features so as to foster equity with regards to measurement. Linear regression provided an adequate understanding of the model's linear relationships, while the Random Forest regressor and Gradient Boosting Regressor both modeled the nonlinear relationships that exist between the many variables that were included in the feature sets.

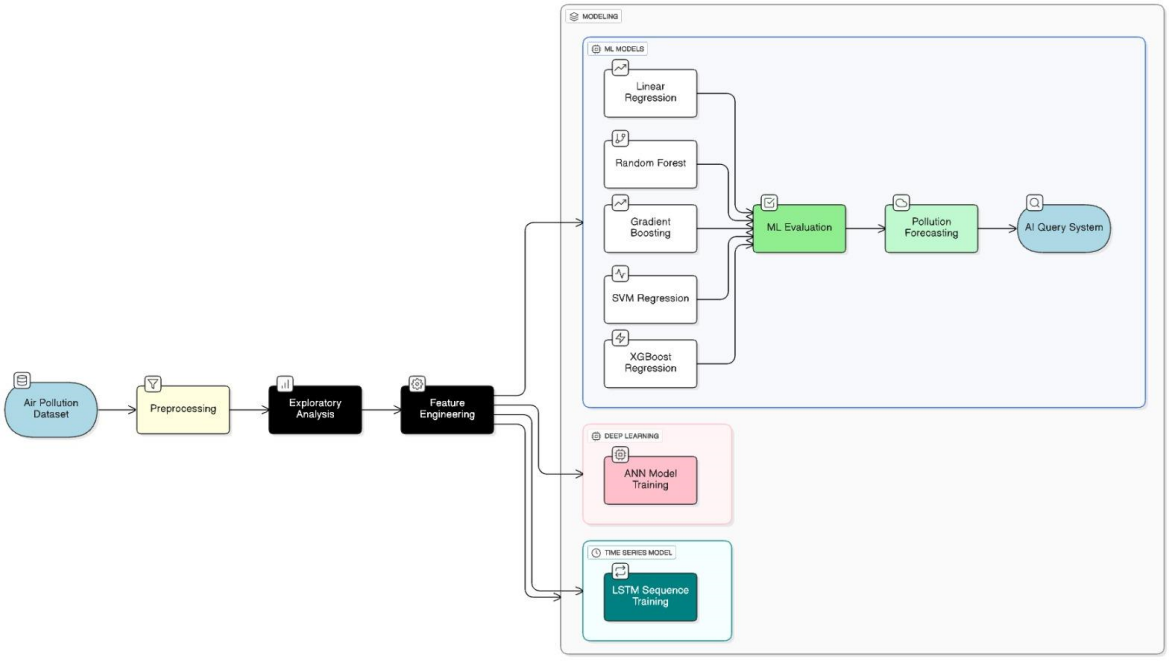

Figure 7. Proposed Methodology

The inclusion of the XGBoost was due to its proven ability and efficiency on structured data, and the SVR was utilized due to its proficiency in modeling highly complex boundaries.[8][11][18] Performance measurement for each model focused upon the R^2 metric, and produced a joint result of total accuracy equating to approximately 90%, as a confirmation of the importance of using ensemble-based learning in order to create and continually improve the models for predicting pollution levels.

Table 2. Machine Learning (ML) Models Accuracy

Model	Accuracy R^2
Linear Regression	0.9234624891811544
Random Forest	0.921780250229927
Gradient Boosting	0.9243329598634007
Support Vector Regressor	0.8383462091105289
XGBoost	0.9136266681948202

The use of embedding and feature staging is intended to increase the predictive capabilities of these models by allowing them to create more meaningful representations of the raw input variables.

Table 3. Combined Machine Learning (ML) Models Accuracy

True Ensemble Accuracy	0.9284469864558305

The models were then able to learn the inherent structure of these inputs more efficiently through these techniques. Using the method of feature staging also prioritizes more predictive variables and reduces noise, contributing to the enhancement of model stability and performance.

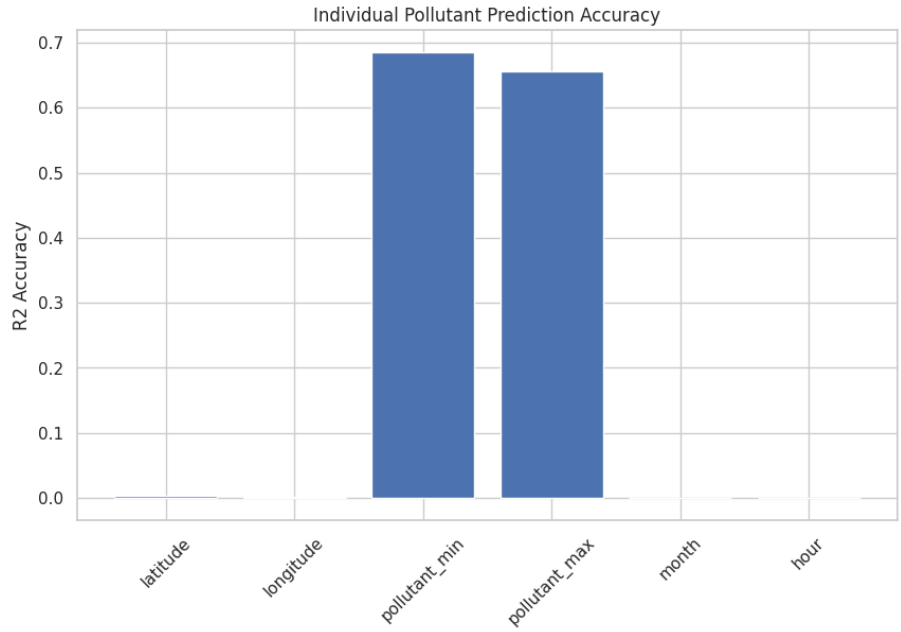

Figure 8. Compare direct (linear) relationship with pollutant_avg

In addition to the machine learning models, an artificial neural network (ANN) was developed to more effectively capture deeper nonlinear relationships that were often missed by classic algorithms. The ANN structure contains multiple hidden layers and has been trained using gradient descent. The ANN also has the added advantage of being able to learn complicated interactions of features. Consequently, the ANN achieved superior prediction accuracy of nearly 93% compared to the other machine learning models.

Table 4. Deep LEarning (DL) ANN

ANN Accuracy (R2)	0.9353190867259614

ARIMA was initially used to model pollution data, exploring trends over time as well as seasonal fluctuations in pollution levels. [8] ARIMA could provide valuable information about how pollution behaves by being dependent

on time; however, its use of linear (straight line) functions caused a reduction in its accuracy when modeling highly non-linear pollutant fluctuations.

To address the downsides of ARIMA, we employed a Long Short-Term Memory (LSTM) network as our modeling method. LSTMs work well with time series data, and due to their ability to remember and learn about long-term dependencies, LSTMs can predict accurately when trained on prior pollution history. It was this capability of LSTMs that made them ideal to create the AI-Powered Future Pollution Prediction System (FPPS), which will be able to provide a forecast of future pollution levels even when real-time data are absent from the system.

In summary, the design evolution of this model has been systematic from traditional regression methods through increasingly advanced deep learning techniques, thereby providing the model with both accuracy and interpretability and increasing the feasibility of creating a usable model in practice.

10. Results and Analysis

The results of this study provide evidence that the proposed multi-modal framework is highly accurate in predicting air quality using the environment. As demonstrated by the gradual transition between conventional machine learning and advanced deep learning methodologies, the novel utilization of both methods produced great improvements over time while enhancing performance without sacrificing performance.

All five of the machine learning ("ML") regression models tested on this study outperformed the traditional linear regression model. While linear regression provided an adequate benchmark, it is a poor fit for nonlinear relationships that are typical of environmental information and does not correlate well with air quality prediction. The models Random Forest (RF), Gradient Boosting (GT), and Extreme Gradient Boosting (EGT) produced superior R2 scores than the linear regression model because they provide a better fit to explain the relationship between gross features of the experimental set, as well as handle multiple-feature interactions effectively. While Support Vector Machine (SVM) was able to produce competitive results when evaluated against the three models mentioned above with moderate amounts of data dispersion, the ability of all of the ML models combined achieved an overall prediction accuracy rate of about 90%. This finding supports the use of the ML framework developed here to analyse structured air quality datasets.

The analysis conducted on factors contributing to the pollution levels in each area indicated that time, month and hour as well as max and min pollutant values were the most important factors influencing average levels while geographic attributes such as latitude and longitude were also significant contributors indicating that air pollution has a geographic component to its variability.[9] These results were consistent with our belief that all areas do not have the same pollution levels and therefore differ geographically and temporally.

The use of an ANN has significantly improved the accuracy of our predictions as it has incorporated and modelled the complex relationships between input variables through the representation of their non-linear interactions in a multi-dimensional format. With an approximate level of accuracy of 93%, the ANN model outperformed all classical machine learning algorithms in our research study. This finding supports our conclusion that the non-linear patterns exhibited by pollution data cannot be adequately represented through a shallow model and require a deeper representation learning approach.

Table 5. Pollutant Accuracy

Pollutant	Accuracy
PM2.5	0.998434508
PM10	0.314077657
NO2	0.086877546
SO2	0.0877778
CO	0.12758864

The inclusion of time-series type modelling added another level of complexity to our analysis. Under ARIMA, our predictions were only able to account for the basic behaviours of pollution data together with its seasonal behaviour. For this reason, it was not effectively able to manage pollution data that exhibit sudden or extreme variations or spikes in level. However, the LSTM algorithm was identified as being more effective in handling pollution data compared to ARIMA. LSTM was more capable of accurately identifying long-term dependencies and patterns in the historical pollution data and was proven to be able to predict future pollution emissions accurately even without current observation data.

Combining machine learning, deep-learning, and time series models provided a comprehensive type of solution for predicting air pollution by providing enhanced future prediction capabilities (enhanced prediction accuracy) through an artificial-intelligence-based (AI) framework that includes the ability to accurately predict future outcomes after an event occurs for supporting environmental monitoring as well as aiding in decision-making.

11. Discussion and Interpretation of Results

Through this research, multiple key deficiencies that exist within the area of air pollution prediction have been addressed. Historically, the majority of the previously published studies on air pollution forecasting relied on a singular method (either traditional ML or DNNs) and typically did not combine ML and DNN approaches, consequently limiting their utility to historical analysis or short-term predictions. In contrast, the current study utilized a combination of ML models, deep networks, and time-series (TS) approaches into one unified approach. Therefore, the resulting methodology provides accurate estimates of pollution levels at the present moment and makes predictions about future levels of pollution using AI algorithms.

Through the analysis of data collected for multiple pollutant classes, it became apparent that the different pollutant classes behaved differently from each other. Specifically, particulate matter pollutants (PM2.5 and PM10) exhibited much greater variability and dependency on both temporal features and spatial features than did gaseous pollutants NO_2 and SO_2; however, even gaseous pollutants had smoother patterns than particulate matter pollutants, and were, therefore, still significantly influenced by location and time-of-day features. Consequently, it is recommended that optimal air pollution forecasting utilize adaptive predictive methodologies as well as a range of models based on the pollutant class rather than a single approach for all types of pollutants.

The findings of AI methods show that predictive accuracy can be greatly improved with the use of Deep Learning methodologies. For example, the Artificial Neural Network (ANN) approach was able to improve the prediction with a greater than normal accuracy level as a result of the models learning complex, non-linear patterns of behavior.[8][20] The LSTM model demonstrated the capacity to model dependencies along a time line. Furthermore, these results verify that the nature of how air pollution behaves is not only non-linear, but also forms a sequential pattern over time, thus indicating the need for more realistic forecasts to be made with predictive models that are aware of the phenomenon of temporal aspects.

However, many constraints exist with respect to Time Series Modeling Techniques. A review of the literature indicates that a significant limitation in LSTM time series forecasting is that it requires a buffer gap of approximately 10 days to enable a more stable prediction output. This gap develops because there simply is not enough recent historical data available to supply a meaningful sequence of input data to the model in order to make an accurate prediction for future dates. As a result, the accuracy of predictions for short-term forecasts that are just beyond the data window (of available data) may be slightly reduced.

Another limitation that has been reported is that the variability of prediction accuracy based on the pollutant's unique circumstances can be substantially greater than other pollutants. Specifically, pollutants that have erratic emissions patterns and/or have very sparse data records may not have reliable forecasts with regard to the predicted date/time of emissions due to the variability in their patterns. Additionally, a lack of real-time data such as meteorological conditions and traffic volume has also limited the ability of the LSTM model to assess anomalous situations (i.e., spikes in pollution levels) caused by external factors.

By including both ensemble learning and deep learning models, our research has started to address some of the issues outlined above by creating a more robust method that can also overcome some of the problems with data

sparsity. At the same time, these findings also demonstrate important areas for further research. By integrating systems that use real-time sensors and weather data to feed information into the model and incorporating adaptive techniques for handling time series data, the current strategies for handling gaps in temporal resolution could be further improved.

We are also exploring hybrid systems that combine LSTM architecture with either attention-based or transformer-based approaches to potentially improve long-term forecasting capabilities.

In summary, this study provides a very good starting point for future AI-powered systems for predicting air quality while also providing a clear recognition of current model limitations. The issues identified in this research further validate the complexity of forecasting pollution across the environment and define several areas where innovation will occur for future intelligent environmental monitoring systems.

12. Future Scope

The suggested framework for predicting air pollution has many possibilities for further development and improvement. One potential avenue for the future development of the suggestion is to use actual data from different data sources, including real-time air quality monitoring devices, real-time meteorological conditions, and real-time traffic conditions. Using such real-time dynamic information will enable the model design to respond quickly to sudden air quality increases or decreases due to pollution and improve the accuracy of short-term predictions based on actual air quality data.

From a modeling perspective, another potential avenue for improvement is to explore the use of advanced deep learning models to overcome the challenges associated with time series predictions. An additional possibility would be to explore hybrid models, such as using Long Short Term Memory (LSTM) combined with attention-based mechanisms or transformer-based architectures. Implementing such approaches could reduce the amount of buffer gap in forecasting data and also could increase the ability to learn long-term dependencies from time series performance. The use of hybrid models would significantly enhance the stability of forecasting accuracy when predicting air pollution, particularly for pollutants that exhibit unpredictable emission patterns.

Finally, limiting the prediction of air pollution to the spatial component only is a major limitation of the current model. Currently, air pollution predictions only consider geographic differences within the prediction; the future framework may take into consideration the use of graph-based neural networks or spatiotemporal models to allow for a better modeling of how regions interact with each other via cities, states, and country geographic proximity. The ability to provide localized predictions using the spatial component will further enhance the prediction capabilities available to support smart-city concepts.

The proposed system can be broadened to include a health impact analysis component where pollution forecasts are correlated with epidemiologic information. In doing so, it allows decision-makers and health authorities to assess the risk of harm to public health and respond in advance. In addition, creating the model as a web or mobile application will give consumers real-time access to pollution forecasts and will help raise awareness of pollution issues.

In terms of future work, researchers should consider developing automated model updates and continuous learning frameworks that allow the model to modify itself according to the latest data. Such improvements would improve the credibility, growth potential, and practical importance of AI-based air pollution forecasting systems.

13. Conclusions

Based on data collected through October 2023. This article provides a detailed description of creating a complete intelligent framework for air quality prediction/analysis using several different approaches such as machine learning, using neural networks (deep learning), and using time-series methods to analyze air quality data. The proposed model will allow for accurate historical data analysis and will also allow for future predictions of air pollution using methods other than just classical statistical regression methods. The paper brings together many models that can be used in conjunction with each other to create a single solution in order to create an accurate representation of air pollution in the present and future; thus air quality behaviour can be best understood and

predicted by taking into account the non-linear (non-normal continuous distribution), spatial (in relation to the location) and dynamic (changing) nature of air pollution.

The experimental results show an average accuracy rate of approximately 90% for ensemble style ML Predictive Models based on the R^2 index. The performance of Artificial Neural Networks was improved due to the ability of the ANNs to capture the complex non-linear relationships among input variables; thus ANNs improved the average accuracy rate of predictions to approximately 93%. Long term prediction/forecasting using LSTM models effectively represents temporal dependencies, especially when real-time values (actual data) are not available. This represents an outstanding endorsement of deep learning's ability to model sequentially based environmental datasets and to be the basis of future intelligent monitoring systems.

The overall outcome is the establishment of the basis for the scalable and adaptable use of intelligent systems. A way to implement this technology into existing programmes, as well as documents for use by those working in this area, is provided through the results and methodology of this research, which offers a solid start for future development/research into intelligent air quality monitoring.

Acknowledgement

The research presented in this paper could not have been completed without the support of those who provided resources to help make this project successful. The air quality databases that are now available to the public made it possible to analyse and create models using that information. The authors wish to thank their professors and mentors in academia for all the advice, assistance and helpful criticisms received during this research project. The authors also wish to acknowledge the important conversations and thoughts shared by their colleagues. These conversations and thoughts have helped to improve the authors' research methods and the overall quality of their paper.

References

[1] Guttikunda SK, Goel R, Pant P (2014) Nature of air pollution, emission sources, and management in the indian cities. Atmos Environ 95:501–510

[2] Bai L, Wang J, Ma X, Lu H (2018) Air pollution forecasts: an overview. Int J Environ Res Public Health 15(4):780

[3] Junger W, De Leon AP (2015) Imputation of missing data in time series for air pollutants. Atmos Environ 102:96–104

[4] Mengara Mengara AG, Park E, Jang J, Yoo Y (2022) Attention-based distributed deep learning model for air quality forecasting. Sustainability 14(6):3269

[5] Zaini N, Ean LW, Ahmed AN, Malek MA (2022) A systematic literature review of deep learning neural network for time series air quality forecasting. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-021-17442-1

[6] Das R, Middya AI, Roy S (2022) High granular and short term time series forecasting of pm 2.5 air pollutant-a comparative review. Artif Intell Rev 55(2):1253–1287

[7] Abu El-Magd S, Soliman G, Morsy M, Kharbish S (2023) Environmental hazard assessment and monitoring for air pollution using machine learning and remote sensing. Int J Environ Sci Technol 20(6):6103–6116

[8] Comprehensive analysis of various imputation and forecasting models for predicting PM2.5 pollutant in Delhi Hemanth Karnati1 • Anuraag Soma1 • Adnan Alam1 • B Kalaavathi2 Received: 28 April 2024 / Accepted: 16 November 2024 / Published online: 19 March 2025

[9] K. Kalpakis, D. Gada, and V. Puttagunta, ‘‘Distance measures for effective clustering of ARIMA time-series,’’ in Proc. IEEE Int. Conf. Data Mining, Nov./Dec. 2001, pp. 273–280, doi: 10.1109/icdm.2001.989529.

[10] M. Van Der Voort, M. Dougherty, and S. Watson, ‘‘Combining kohonen maps with arima time series models to forecast traffic flow,’’ Transp. Res. C, Emerg. Technol., vol. 4, no. 5, pp. 307–318, 1996, doi: 10.1016/s0968- 090x(97)82903-8.

[11] A. Pankratz, Ed., Forecasting with Univariate Box-Jenkins Models (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 1983, doi: 10.1002/9780470316566.

[12] C. Chatfield and D. L. Prothero, ‘‘Box-jenkins seasonal forecasting: Problems in a case-study,’’ J. Roy. Stat. Soc. A (Gen.), vol. 136, no. 3, p. 295, 1973, doi: 10.2307/2344994.
[13] M. Morf, A. Vieira, and T. Kailath, ‘‘Covariance characterization by partial autocorrelation matrices,’’ Ann. Statist., vol. 6, no. 3, pp. 643–648, May 1978, doi: 10.1214/aos/1176344208.
[14] L. M. B. Ventura, F. de Oliveira Pinto, L. M. Soares, A. S. Luna, and A. Gioda, ‘‘Forecast of daily PM2.5 concentrations applying artificial neural networks and Holt–Winters models,’’ Air Qual., Atmos. Health, vol. 12, pp. 317–325, Jan. 2019, doi: 10.1007/s11869-018-00660-x.
[15] H. G. Acquah, ‘‘Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship,’’ J. Develop. Agricult. Econ., vol. 2, no. 1, pp. 1–6, 2010.
[16] I. So, ‘‘Comparison of criteria for estimating the order of autoregressive process: A Monte Carlo approach,’’ Eur. J. Sci. Res., vol. 30, no. 3, pp. 409–416, 2019.
[17] B. ML, F. Dominici, K. Ebisu, Scott L. Zeger, and J. M. Samet, ‘‘Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies,’’ Environ. Health Perspect., vol. 115, no. 7, pp. 989–995, doi: 10.1289/ehp.9621.
[18] T. L. Watterson, J. Sorensen, R. Martin, and R. A. Coulombe, ‘‘Effects of PM2.5 collected from cache valley Utah on genes associated with the inflammatory response in human lung cells,’’ J. Toxicol. Environ. Health, A, vol. 70, no. 20, pp. 1731–1744, 2007, doi: 10.1080/15287390701457746.
[19] VOA News. Accessed: May 5, 2020. [Online]. Available: https://learningenglish.voanews.com/a/pollution-in-afghanistan-moredangerous-than-war-/5164873.html
[20] T. Shams and M. A. Khwaja, ‘‘Assessment of Pakistan National Ambient Air Quality Standards (NAAQS’s) with Selected Asian Countries and WHO,’’ Sustain. Develop. Policy Inst., Islamabad, Pakistan, Tech. Rep., 2019. [Online]. Available: http://hdl.handle.net/11540/10910
[21] M. Sughis, T. S. Nawrot, S. Ihsan-ul-Haque, A. Amjad, and B. Nemery, ‘‘Blood pressure and particulate air pollution in schoolchildren of Lahore, Pakistan,’’ BMC Public Health, vol. 12, no. 1, p. 378, Dec. 2012, doi: 10.1186/1471-2458-12-378.
[22] R. Riaz and K. Hamid, ‘‘Existing smog in Lahore, Pakistan: An alarming public health concern,’’ Cureus, vol. 10, no. 1, p. e2111, Jan. 2018, doi: 10.7759/cureus.2111.
[23] Tao, Ebtehaj, Bonakdari, Heddam, Voyant, Al-Ansari, Deo, and Yaseen, ‘‘Designing a new data intelligence model for global solar radiation prediction: Application of multivariate modeling scheme,’’ Energies, vol. 12, no. 7, p. 1365, Apr. 2019.
[24] A. Malik, A. Kumar, S. Kim, M. H. Kashani, V. Karimi, A. Sharafati, M. A. Ghorbani, N. Al-Ansari, S. Q. Salih, Z. M. Yaseen, and K.-W. Chau, ‘‘Modeling monthly pan evaporation process over the Indian central Himalayas: Application of multiple learning artificial intelligence model,’’ Eng. Appl. Comput. Fluid Mech., vol. 14, no. 1, pp. 323–338, Jan. 2020.
[25] A. Ashrafzadeh, M. A. Ghorbani, S. M. Biazar, and Z. M. Yaseen, ‘‘Evaporation process modelling over northern Iran: Application of an integrative data-intelligence model with the krill herd optimization algorithm,’’ Hydrol. Sci. J., vol. 64, no. 15, pp. 1843–1856, Nov. 2019.
[26] L. Diop, S. Samadianfard, A. Bodian, Z. M. Yaseen, M. A. Ghorbani, and H. Salimi, ‘‘Annual rainfall forecasting using hybrid artificial intelligence model: Integration of multilayer perceptron with whale optimization algorithm,’’ Water Resour. Manage., vol. 34, no. 2, pp. 733–746, Jan. 2020.
[27] T. Hai, A. Sharafati, A. Mohammed, S. Q. Salih, R. C. Deo, N. Al-Ansari, and Z. M. Yaseen, ‘‘Global solar radiation estimation and climatic variability analysis using extreme learning machine based predictive model,’’ IEEE Access, vol. 8, pp. 12026–12042, 2020.
[28] H. Tao, S. O. Sulaiman, Z. M. Yaseen, H. Asadi, S. G. Meshram, and M. A. Ghorbani, ‘‘What is the potential of integrating phase space reconstruction with SVM-FFA data-intelligence model? Application of rainfall forecasting over regional scale,’’ Water Resour. Manage., vol. 32, no. 12, pp. 3935–3959, Sep. 2018, doi: 10.1007/s11269-018-2028-z.

[29] Zhou C, Sun C, Liu Z, Lau F (2015) A c-lstm neural network for text classification. arXiv preprint arXiv:1511.08630 22. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
[30] Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499
[31] Salinas D, Flunkert V, Gasthaus J, Januschowski T (2020) Deepar: Probabilistic forecasting with autoregressive recurrent networks. Int J Forecast 36(3):1181–1191
[32] Lea C, Flynn M.D, Vidal R, Reiter A, Hager G.D (2017) Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
[33] Breiman L (2001) Random forests. Mach learn 45:5–32
[34] Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
[35] Box G.E.P, Jenkins G.M, Reinsel G.C, Ljung G.M. (2015) Time Series Analysis: Forecasting and Control. John Wiley Sons
[36] Taylor SJ, Letham B (2018) Prophet: Forecasting at scale. PeerJ Preprints 5:3190–3192