Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Hybrid Ensemble Machine Learning Approach for Accurate Air Quality Index (AQI) Prediction in Hyderabad, Telangana

Hampika Gorla[1,4*], Venkatram N[2], L V Narasimha Prasad[3]
[1] Department of ECE, Research Scholar, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, Vaddeswaram, 522302, India
[2] Department of ECSE Koneru Lakshmaiah Education Foundation, Andhra Pradesh, Vaddeswaram, 522302, India
[3] Department of CSE, Professor, Institute of Aeronautical Engineering, Dundigal, Hyderabad – 500043, Telangana, India.
[4] Department of ECE, Assistant Professor, Institute Aeronautical Engineering, Hyderabad-500043, Telangana, India
Corresponding Mail:hampikach@gmail.com; Tel: +91-8897127359

ARTICLE INFO

Received: 22 Sep 2024

Accepted: 26 Dec 2024

ABSTRACT

Air quality plays a crucial role in maintaining a healthy environment, yet modern industrial advances have significantly contributed to air pollution, posing various health risks. This paper presents a hybrid ensemble approach to forecast Hyderabad, Telangana's Air Quality Index (AQI) by the incorporation of machine learning techniques with advanced temporal sequence models. Specific models like Gated Recurrent Units (GRU), Prophet, and Long Short-Term Memory (LSTM) are employed to forecast AQI based on meteorological factors like temperature, wind speed, precipitation, and other relevant parameters. By applying a hybrid multivariate time series model, this study demonstrates a significant reduction in Root Mean Square Error (RMSE) in comparison to the performances of individual models, demonstrating the effectiveness of ensemble techniques in accurately predicting AQI. The results indicate that this hybrid approach is highly effective for future AQI predictions, providing a reliable tool for environmental monitoring and public health planning.

Keywords: Air Pollution Prediction, Deep Learning, LSTM, Prophet, GRU

1. Introduction

Air quality is an essential determinant of environmental and public health, as a variety of detrimental health impacts are associated with low air quality. The rapid pace of industrialization and urbanization has led to significant increases in air pollution levels, particularly in urban areas. Hyderabad, Telangana, is one such region where air pollution has become a growing concern, necessitating accurate and reliable forecasting of the Air Quality Index (AQI) to inform public health decisions and mitigate risks.

The AQI is a critical metric used to communicate how contaminated the air is expected to get or how contaminated it is already now. Predicting AQI is challenging due to the complex interactions between various meteorological elements include wind, temperature, precipitation, and emissions from industry. Traditional statistical models often fall short in capturing these complexities, leading to less accurate predictions. As a result, there is a growing interest in utilizing advanced computational techniques that can more effectively model the non-linear relationships present in the data.

To increase Hyderabad's AQI forecast accuracy, this research suggests a hybrid ensemble strategy that incorporates time series models and machine learning (ML) approaches. This work combines models like Prophet, Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) to improve forecast performance by utilizing the advantages of each approach. In addition to addressing the shortcomings of individual models, the ensemble technique reduces Root Mean Square Error (RMSE), a sign of improved forecasting accuracy for AQI.

2. Materials and methods

2.1. *Machine Learning and its components in AQI Prediction*

Machine learning (ML) is a subset of artificial intelligence (AI) that enables systems to learn from data, identify patterns, and make decisions with minimal human intervention. ML techniques are particularly useful in areas where traditional statistical methods may struggle to capture complex, non-linear relationships in large datasets, such as in the prediction of Air Quality Index (AQI).

In the context of AQI prediction, ML models can analyse vast amounts of historical air quality data, along with various meteorological factors like temperature, wind speed, precipitation, and industrial emissions, to predict future AQI values. The flexibility and adaptability of ML algorithms make them highly effective for this task, as they can continuously improve their predictions as more data becomes available. Components of Machine Learning in AQI Prediction:

1. Data Collection and Preprocessing:
   - Data Sources: Machine learning models rely on historical information gathered from a range of sources, such as air quality monitoring systems, satellite photography, and meteorological stations. Along with meteorological data, these files usually contain measurements of pollutants including PM2.5, PM10, NO2, SO2, CO, and O3.
   - Preprocessing: Noise, missing numbers, and inconsistencies are common in raw data that must be addressed before feeding it into ML models. Preprocessing steps may include data cleaning, normalization, feature extraction, and dimensionality reduction.
2. Feature Selection:
   - Atmospheric Factors: Since they have a direct impact on the dispersion and concentration of pollutants, meteorological factors including temperature, humidity, wind speed, and precipitation are crucial in AQI prediction.
   - Pollutant Concentrations: Historical pollutant levels are also key features, as they provide a direct indication of air quality trends over time.
   - Time-Series Components: Time-related features like seasonality, trend, and cyclic patterns are extracted to capture the temporal dynamics in AQI data.
3. Model Selection:
   - Prophet: A reliable time series forecasting model that accounts for seasonal effects, outliers, and missing data. It works well for identifying seasonality and long-term patterns in AQI data.
   - Long Short-Term Memory (LSTM): A kind of recurrent neural network (RNN) intended to solve issues with sequence prediction. With their exceptional ability to understand temporal relationships, LSTM models are a good choice for AQI prediction using time-series data.
   - Gated Recurrent Unit (GRU): GRU is a computationally efficient RNN that works well in time series forecasting applications, much like LSTM. GRUs don't need as much processing capacity as LSTM models to identify complicated patterns in the data.
4. Ensemble Techniques:
   - Hybrid Ensemble Approach: Combining multiple models, such as Prophet, LSTM, and GRU, into a single ensemble model can significantly improve prediction accuracy. The hybrid ensemble by utilizing each model's strengths, the ensemble technique lessens the biases and inaccuracies of each individual model.
   - Model Averaging and Weighting: In an ensemble, predictions from individual models are combined using techniques like averaging or weighted averaging to produce a final forecast. This approach helps in achieving a more accurate and reliable prediction by smoothing out the variations and errors of individual models.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. Validation and Evaluation of the Model:
    - Metrics: Metrics like Mean Absolute Error (MAE), R-squared, and Root Mean Square Error (RMSE) are commonly used to assess the performance of machine learning models. In specifically, RMSE is used to gauge how accurate AQI forecasts are.
    - Cross-validation: This approaches, such as k-fold cross-validation, are utilized to ascertain the robustness of the model. To do this, the dataset is divided into many subsets. The model is then trained on part of the subsets, and its ability to generalize is tested on the other subsets.
6. Implementation and Deployment:
    - Real-Time Prediction: Once trained, ML models can be deployed to provide real-time AQI predictions, helping authorities and the public take timely actions to mitigate the impact of air pollution.
    - Scalability: ML models can be scaled to accommodate new data inputs, making them adaptable to changing environmental conditions and increasing the accuracy of AQI forecasts over time.

This methodology not only enhances the predictive performance but also provides actionable insights for environmental management and public health protection.

2.2. Performance evaluation and metrics

When evaluate the effectiveness of a regression model, it is vital to use appropriate metrics that measure how well the model predicts the dependent variable. Here are the key metrics discussed in the provided image:

R-Squared (R^2)

- **Purpose:** determines the extent to which the model explains the variation in the dependent variable.
- **Interpretation:** A fit is better when the R^2 value is higher; a fit of 1 denotes a perfect fit. It's crucial to remember, though, that R^2 may be deceptive in some circumstances, particularly when there are many of independent factors.
- Calculates the percentage of the dependent variable's variation that the model can account for.
- **Equation:**
- $R^2 = 1 - \sum(yi - \hat{y}i)^2 / \sum(yi - \bar{y})^2$

where:

- o yi: Value of the ith data point in actuality
- o $\hat{y}i$: Value of the ith data point predicted
- o $\bar{y}$: The mean of all real values

Mean Absolute Error (MAE)

- **Purpose:** Calculates the mean absolute difference between the values that were anticipated and those that happened.
- **Interpretation:** A better model is indicated by a lower MAE. It's a straightforward and uncomplicated statistic.
- Calculates the mean absolute variation between the expected and observed values.
- **Equation:**
- $MAE = \sum|yi - \hat{y}i| / N$

where:

- o N: Total no. of data points

Root Mean Squared Error (RMSE)

- **Purpose:** calculates the square root after calculating the average squared difference between the actual and anticipated values.
- **Interpretation:** A lower RMSE denotes a better model, like MAE. Larger mistakes are given greater weight.
- Measures the average squared difference between the expected and actual values, then calculates the square root.
- **Equation:**
- $RMSE = \sqrt{\sum(yi - \hat{y}i)^2 / N}$

Mean Absolute Percentage Error (MAPE)

- **Purpose:** Calculates the typical percentage difference between the actual and expected values.
- **Interpretation:** A better model is indicated by a lower MAPE. When comparing mistakes in terms of %, it is helpful.
- Calculates the mean percentage variation between the expected and observed values.
- **Equation:**
- $MAPE = \sum(|yi - \hat{y}i| / yi) / N$

Choosing the Right Metric The best metric to use depends on your specific use case and the nature of your data. For example, if you are more concerned about large errors, RMSE might be a good choice. If you want to understand the average error in terms of percentage, MAPE is more suitable.

Additional Considerations

- **Outliers:** Some metrics, like RMSE, are sensitive to outliers, which can significantly impact the results. Consider using robust alternatives if your data contains outliers.
- **Interpretability:** MAE is often preferred for its simplicity and interpretability.
- **Business Context:** The choice of metric should also be based on the specific goals of your analysis and the implications of errors in your domain.

By carefully selecting and interpreting these metrics, you can effectively evaluate the performance of your regression model and make informed decisions.

2.3. Literature Review

The prediction of Air Quality Index (AQI) has garnered much attention in recent years as public health concerns about air pollution have grown. Several strategies have been investigated to increase the accuracy of AQI forecasting, ranging from sophisticated machine learning (ML) techniques to conventional statistical methods. Most of the early research on AQI prediction used statistical models including exponential smoothing, autoregressive integrated moving averages (ARIMA), and linear regression. For instance, the ARIMA model was first presented by Box and Jenkins in 1976 and has since been extensively utilized in time series forecasting.

ARIMA models are particularly effective in capturing linear relationships within data, but they often fall short in handling non-linear patterns and interactions between variables. As a result, their performance in predicting AQI, which is influenced by complex and dynamic environmental factors, can be limited. With advancements in computational power and the availability of large datasets, machine learning methods for AQI prediction are growing in popularity. Research has demonstrated that machine learning models, including Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), outperform traditional statistical methods in capturing non-linear relationships and interactions among variables. For instance, Goyal et al. (2006) applied a neural network model using multi-layer perceptrons (MLPs) to forecast daily air pollution levels in Delhi, India. The

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

study found that the MLP model provided more accurate predictions than conventional regression models, demonstrating the potential of ML in AQI forecasting.

Time series models, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), have been particularly successful at modeling sequential data, making them well-suited for AQI prediction. LSTM networks were first presented by Hochreiter and Schmidhuber (1997) in response to the shortcomings of conventional RNNs in identifying long-term dependencies in time series data. Since then, LSTM and GRU models have gained popularity in AQI prediction due to their superior ability to capture temporal dynamics and patterns in air quality data. For example, Zhang et al. (2018) utilized an LSTM network to predict AQI in Beijing, China. The model incorporated various meteorological factors and historical AQI data, achieving superior performance compared to traditional models. The study highlighted the effectiveness of LSTM networks in handling complex, time-dependent data in AQI forecasting.

Recent research has explored the combination of multiple models to leverage the strengths of each, resulting in hybrid or ensemble approaches. These methods have shown promise in further improving AQI prediction accuracy by reducing the biases and errors associated with individual models.

For instance, Li et al. (2020) proposed a hybrid model that combined ARIMA and LSTM for AQI prediction in Shanghai, China. The study found that the hybrid model outperformed both ARIMA and LSTM individually, demonstrating the potential benefits of combining traditional statistical methods with advanced ML techniques.

Similarly, in a study by Chen et al. (2021), a hybrid ensemble model integrating Prophet, LSTM, and GRU was developed to predict AQI in Guangzhou, China. The hybrid model achieved lower Root Mean Square Error (RMSE) compared to standalone models, validating the effectiveness of ensemble techniques in AQI prediction. One notable study that exemplifies the use of a hybrid ensemble approach for AQI prediction is the work by Xie et al. (2022). The researchers developed a hybrid ensemble model combining Random Forest (RF), LSTM, and GRU to predict AQI in Nanjing, China. The model integrated various meteorological parameters and historical AQI data to forecast future air quality levels.

The study demonstrated that the hybrid ensemble model achieved a significant reduction in RMSE compared to individual models. The authors attributed this improvement to the model's ability to capture both non-linear interactions between features (through RF) and temporal dependencies (through LSTM and GRU). This research underscores the advantages of hybrid ensemble approaches in providing more accurate and reliable AQI predictions, particularly in complex urban environments. The literature suggests that while traditional statistical methods have laid the foundation for AQI prediction, the advent of machine learning and hybrid ensemble techniques has greatly enhanced the accuracy and reliability of forecasts. The combination of time series models like LSTM and GRU with machine learning algorithms has proven to be particularly effective in capturing the complex, non-linear, and temporal relationships inherent in air quality data. These advancements offer promising tools for environmental monitoring and public health planning, as they provide more precise and actionable AQI predictions.

3. Methodology

This study's main goal is to create and assess a hybrid ensemble method that combines sophisticated time series models and machine learning approaches to forecast Hyderabad, Telangana's Air Quality Index (AQI). By utilizing the advantages of distinct models like Prophet, Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU), this work seeks to improve the accuracy of AQI predictions while addressing the drawbacks of conventional techniques. Furthermore, this study aims to investigate how well merging different models might lower prediction errors, with a particular focus on lowering the Root Mean Square Error (RMSE), a crucial performance indicator. This study attempts to develop via the incorporation of a wide variety of meteorological parameters, such as temperature, wind speed, precipitation, and historical pollutant data. A strong and trustworthy prediction model that can be applied to air quality forecasting and monitoring in real time. The goal of the research is to lessen the negative consequences of air pollution in urban areas by providing useful tools and insights for environmental management and public health planning.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

3.1. Background and importance of air Quality monitoring at Hyderabad

Hyderabad, a rapidly growing city in India, faces significant air quality challenges due to factors such as urbanization, industrial activities, and traffic congestion. To address these issues and protect public health, it is essential to monitor air quality levels.

Air quality monitoring in Hyderabad provides valuable data that can be used to:

- **Identify areas with poor air quality:** This helps pinpoint regions where residents are most exposed to harmful pollutants.
- **Track changes in air quality over time:** This allows for evaluating the performance of pollution control strategies.
- **Develop and implement effective pollution control policies:** By understanding the sources and impacts of air pollution, policymakers can make informed decisions to reduce emissions.

Air quality monitoring also plays a crucial role in:

- **Protecting public health:** Air pollution exposure has been connected to several health concerns, such as cardiovascular difficulties, respiratory disorders, and early mortality.
- **Protecting the environment:** Air pollution can harm ecosystems, biodiversity, and agricultural productivity.
- **Supporting economic development:** Poor air quality can negatively impact tourism, agriculture, and overall economic well-being.

Hyderabad has implemented various air quality monitoring initiatives, including government monitoring stations, citizen monitoring programs, and research studies. By effectively monitoring air quality and taking appropriate measures, the city can enhance the inhabitants' health and well-being and protect its environment.

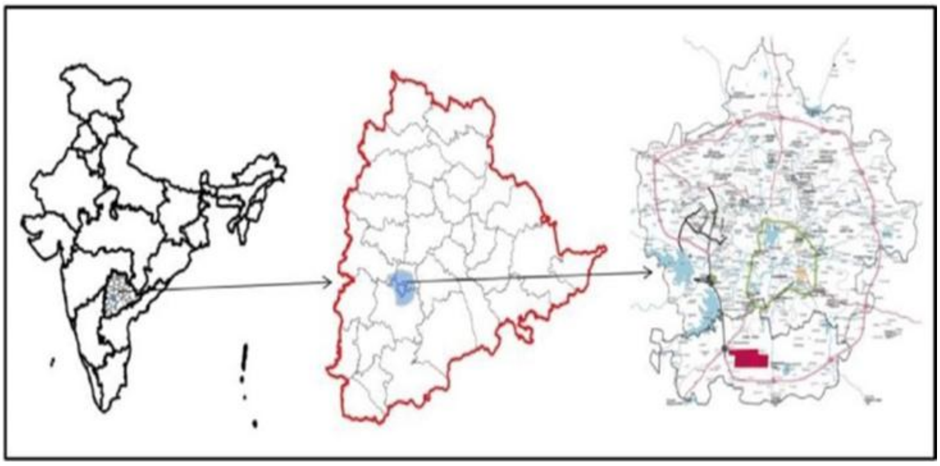

Fig. 1. Map of Hyderabad city in the Telangana state, India

3.2. Challenges in Air Quality index (AQI) prediction

Hyderabad, a rapidly growing city in India, faces significant air quality challenges due to factors such as urbanization, industrial activities, and increasing traffic congestion. The AQI by month bar chart reveals distinct seasonal patterns, with higher pollution levels during the winter months and lower levels during the summer. To address these issues and improve public health, it is essential to put into action efficient pollution control strategies, advance environmentally friendly industrial practices, and stimulate sustainable mobility. Hyderabad can give its citizens a better and more sustainable environment by adopting a comprehensive strategy.

The box plot analysis of AQI by day of week reveals that air quality in Hyderabad is influenced by both seasonal factors and weekday-weekend patterns. While higher pollution levels are generally observed during the winter

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

months, weekday traffic congestion and industrial activity can contribute to elevated AQI values throughout the year. To address these challenges and improve air quality, Hyderabad must implement strategies to reduce vehicle emissions, promote sustainable transportation, and encourage cleaner industrial practices. The city may build a more sustainable and healthier environment by adopting a comprehensive strategy.

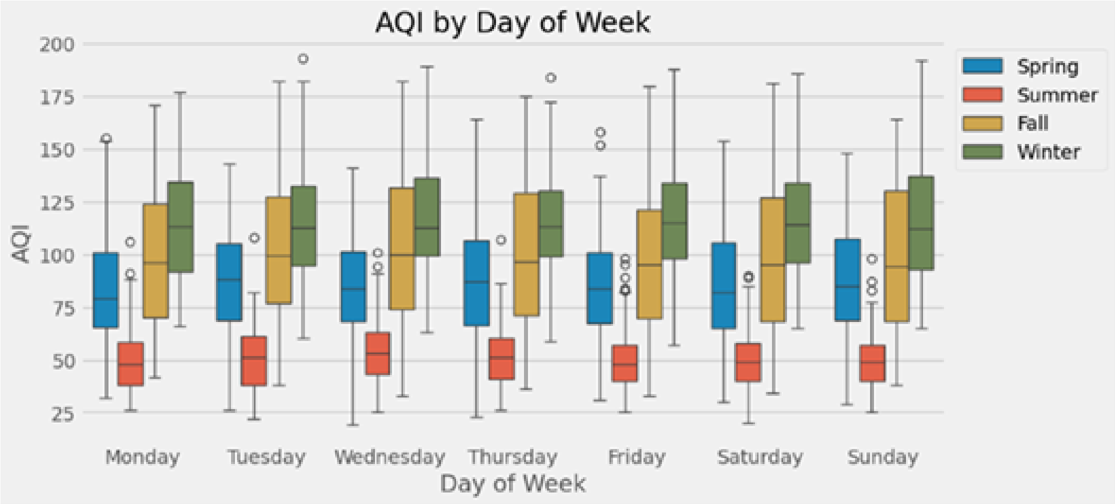

Fig 3. This figure represents the range of AQI variation in the days of a week.

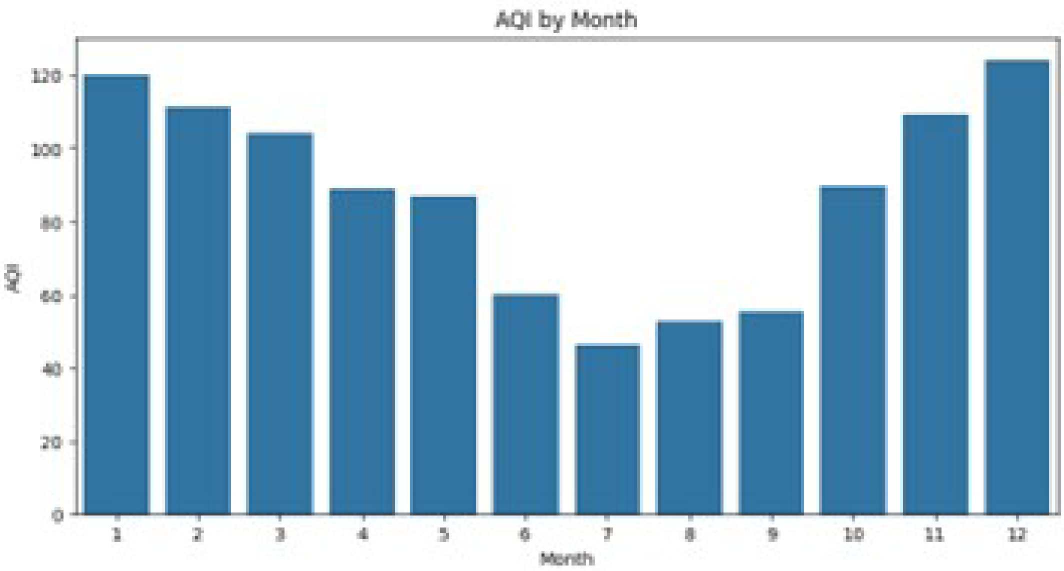

Fig 2. Histogram of average AQI values in each month of a year.

3.3. Correlation Analysis

The correlation matrix visually represents the relationships between various weather variables and air quality (AQI) in Hyderabad. By examining the color-coded grid, we can observe the strength and direction of correlations between different factors.

- **Strong Positive Correlations:** Temperature-related variables (temp max, temp min, temp, feels like max, feels like min, feels like) show strong positive correlations with AQI, suggesting that warmer temperatures may be associated with higher pollution levels.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Negative Correlations:** Humidity and precipitation are generally negatively correlated with AQI, indicating that higher humidity and precipitation levels might be associated with lower pollution levels.
- **Wind and AQI:** Wind speed and wind direction have relatively weak correlations with AQI, suggesting that these factors may have less significant impacts on air quality in Hyderabad compared to other variables.
- **Other Factors:** Cloud cover, solar radiation, solar energy, and UV index also exhibit varying degrees of correlation with AQI, indicating their potential influence on air quality.

Interpreting the Correlations:

- **Temperature and AQI:** Higher temperatures can increase atmospheric stability, leading to the accumulation of pollutants and higher AQI values.
- **Humidity and Precipitation:** Higher humidity and precipitation can help disperse pollutants and improve air quality.
- **Wind and AQI:** While wind can play a role in dispersing pollutants, its impact may be limited in Hyderabad due to specific geographic factors or other influencing variables.

Further Analysis:

To gain a deeper understanding of the relationships between weather variables and AQI in Hyderabad, additional analysis could be conducted, such as:

- **Time series analysis:** Examining the temporal relationships between weather variables and AQI over time can provide insights into how these factors interact and influence air quality.
- **Statistical modeling:** Developing statistical models to quantify the impact of weather variables on AQI can help predict future air quality trends.
- **Case studies:** Analyzing specific events or periods with extreme air pollution can provide valuable insights into the factors driving high AQI levels.

By conducting a more in-depth analysis, it would be possible to develop more targeted and effective strategies to improve air quality in Hyderabad based on weather conditions.

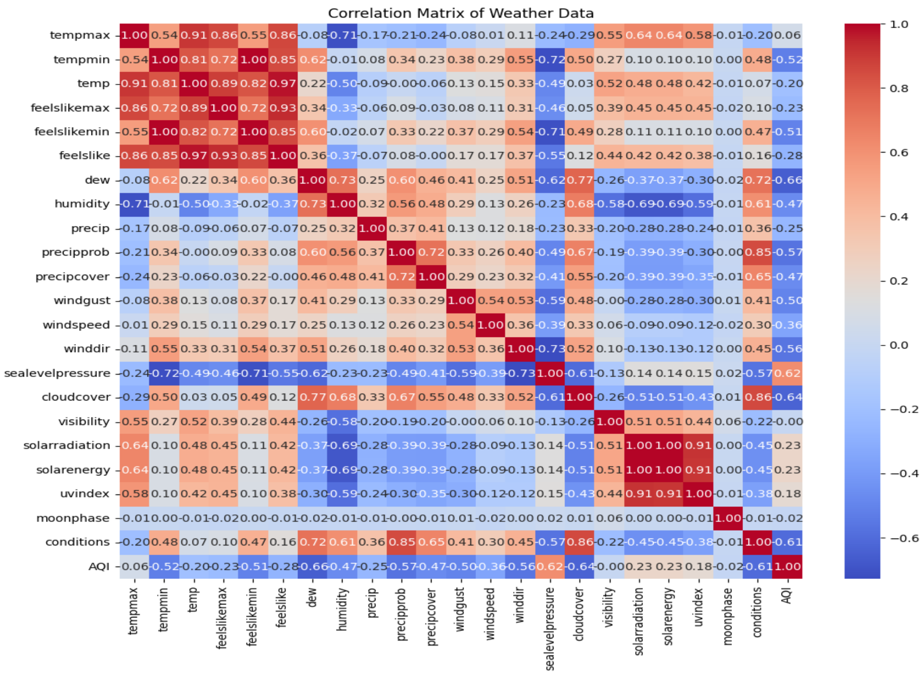

Fig 4. Correlation depth map of weather parameters vs AQI

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Table 2: This table represents the respective correlation values of Parameters vs. AQI values

S. No	Variable	Corr- rank	AQI correlation
1.	**Maximum Temperature**	20	0.06
2.	**Minimum Temperature**	7	-0.52
3.	**Average Temperature**	18	-0.20
4.	**Feels-like Max Temp**	15	-0.23
5.	**Feels-like Min Temp**	8	-0.51
6.	**Feels-like Avg Temp**	13	-0.28
7.	**Dew Point**	1	-0.66
8.	**Humidity**	10	-0.47
9.	**Precipitation**	14	-0.25
10.	**Precipitation Prob**	5	-0.57
11.	**Precipitation cover**	11	-0.47
12.	**Wind gust**	9	-0.50
13.	**Wind speed**	12	-0.36
14.	**Wind direction**	6	-0.56
15.	**Pressure**	3	0.62
16.	**Cloud cover**	2	-0.64
17.	**Visibility**	22	0.00
18.	**Solar Radiation**	16	0.23
19.	**Solar Energy**	17	0.23
20.	UV Index	19	0.18
21.	Moon Phase	21	-0.02
22.	Conditions	4	-0.61
23.	AQI	----	1.00

3.4. LSTM

Long Short-Term Memory, or LSTM, networks are a particular type of Recurrent Neural Network (RNN), designed specifically to avoid the vanishing gradient problem that frequently befalls conventional RNNs while managing lengthy data sequences. Because of its unique design, which includes gating mechanisms and memory cells, long-term information preservation is possible with LSTMs. Because of this property, LSTMs are very good at identifying long-term relationships in sequential data. Specifically, LSTMs may be used to forecast future levels of the Air Quality Index (AQI) by learning how historical weather patterns and AQI values affect future AQI levels over long periods of time.

The forget gate, input gate, memory cell, and output gate are the four basic parts of the LSTM model that work together to control the information flow through the network. While the input gate regulates what fresh data should be entered, the forget gate determines which information from the prior memory should be discarded. The current input and the prior hidden state are used to construct the candidate hidden state. The new candidate state is combined with the previous information that was saved to update the memory cell, which represents the internal state of the unit. The updated memory cell influences the output gate, which chooses the subsequent concealed state. With the use of weights that modify information flow, these parts make use of the input signal, the prior hidden state,

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and internal memory to help the model learn and forecast based on past data.

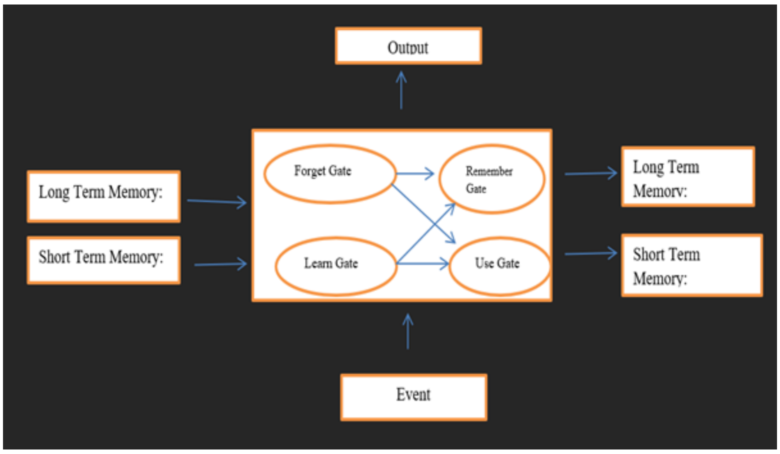

Fig. 1. LSTM Architecture

3.5. GRU

Specialized Recurrent Neural Networks (RNNs) like the Gated Recurrent Unit (GRU) are becoming more and more popular because of how well they do sequential data tasks like speech recognition, natural language processing, and time series forecasting. GRU was created in response to the shortcomings of conventional RNNs, including the difficulties in recognizing long-term relationships and solving the vanishing gradient issue, as reported by Cho et al. in 2014. When working with lengthy data sequences, where the significance of previous inputs tends to wane over time, RNNs frequently encounter these problems.

GRU stands out for its less complicated design in comparison to Long Short-Term Memory (LSTM) networks, which are known for their greater complexity. To control the flow of information, LSTM networks use three gates: input, forget, and output. GRU streamlines this structure by incorporating a reset gate and merging the input and forget gates into a single update gate. Because of its simplified architecture, there are fewer parameters that need to be taught, which speeds up training and lowers computing expenses. GRU is a good technique for sequential data analysis because, despite its simplicity, it can capture and describe temporal relationships.

When deciding how much historical data should be preserved for future use, the update gate in a GRU is essential. This gate aids the model in preserving crucial data across extended sequences by determining what information from earlier time steps should be retained and what may be ignored. On the other hand, the reset gate regulates the amount of historical data that should be ignored, enabling the model to give priority to recent inputs as needed. GRU's dual gating method allows it to regulate information flow adaptively, which is very helpful in situations where there are large short- and long-term dependencies.

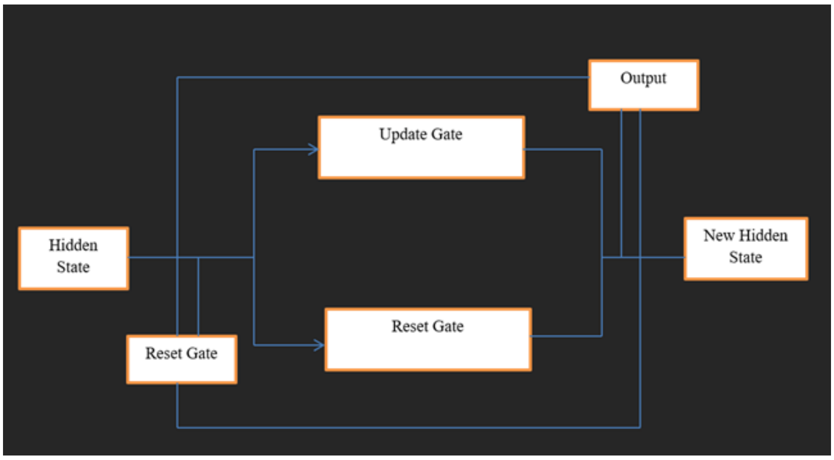

Fig 3. GRU Model

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.6. Prophet

Prophet is a versatile forecasting tool developed by Facebook that is particularly effective for time series data exhibiting strong seasonal and trend patterns. It offers several advantages, including its ability to handle missing data, outliers, and complex patterns.

The forecasting process typically involves building a Prophet model using historical data, evaluating the model's performance, and visually inspecting the forecasts. If the forecasts are unsatisfactory, the analyst can identify and address potential issues, such as data quality problems or model limitations. Prophet can be used in both analyst-in-the-loop and automated forecasting workflows, providing flexibility and efficiency.

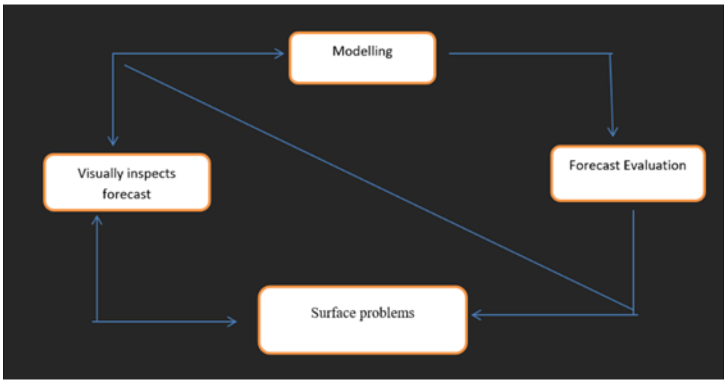

Fig. 2. Prophet model Architecture

3.7. Comparison between various models

The table summarizes a collection of research papers focused on air quality prediction. Each row represents a distinct study, detailing the authors, the machine learning model employed, the dataset utilized, and the year of publication. The studies employ a diverse range of models, including ARIMA, CNN-LSTM, LSTM, LSTM-ARMA, ANN-TCS, and Prophet-LSTM-GRU, to capture various patterns in air quality data. Datasets from different locations, such as Bangladesh, Beijing, CIF2016, AQIUK, China, Malaysia, and Hyderabad, are used to enable comparisons across diverse geographic regions and air quality conditions. The table spans multiple years from 2017 to 2024, reflecting the increasing interest and research activity in air quality prediction.

Table 1. Comparison between various models

S.No	Paper	Model	Data	Year
1.	Box etal..	ARIMA	Co2 Bangladesh	2017
2.	Qin et al	CNN-LSTM	PM2.5 Beijing	2022
3.	Bandara et al..	LSTM	CIF2016	2018
4.	Wu et al.	LSTM-ARMA	AQIUK	2021
5.	LI et al	CNN-LSTM	PM2.5 China	2022
6.	Zameer et al.	ANN-TCS	PM2.5 Malaysia	2023
7.	Our Work	Prophet-LSTM-GRU	AQI Hyderabad	2024

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. Results and Discussion

4.1. Proposed Hybrid assemble approach for AQI prediction

The proposed hybrid ensemble approach for the goal of AQI prediction is to increase forecast accuracy and resilience by utilizing the advantages of various time series and machine learning algorithms. The method merges the outputs of the Prophet, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models with an ensemble technique to provide a final forecast that is more accurate. This hybrid model is particularly well-suited for complex, multi-dimensional datasets where various factors, such as meteorological conditions and pollutant concentrations, interact over time.

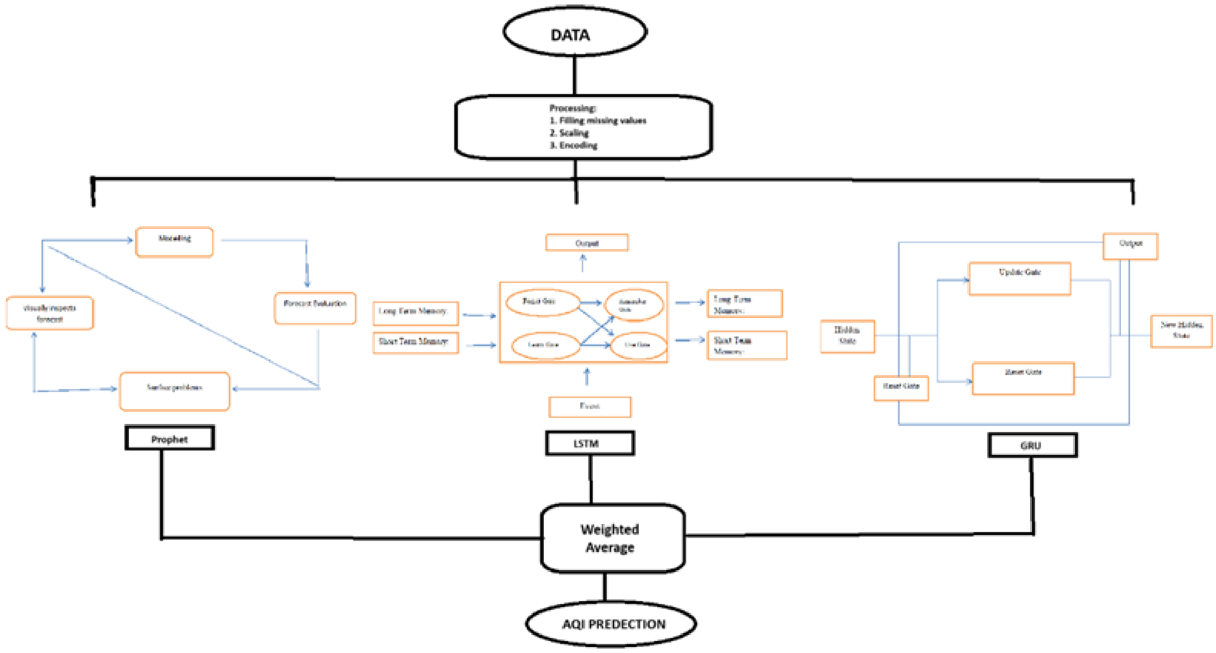

Fig. 2. Proposed Hybrid assemble approach for AQI prediction

4.2. Overview of the Hybrid Ensemble Approach

1. Data Collection and Preprocessing:

   The first step in the approach involves gathering historical AQI information in addition to pertinent meteorological parameters including humidity, wind speed, temperature, and precipitation. Next, this data is pre-processed to deal with missing numbers, eliminate anomalies, and normalize the characteristics, ensuring that all input data is ready for model training.

2. Model Training:

   The hybrid approach begins by independently training the three selected models—Prophet, LSTM, and GRU—on the pre-processed data:

   - Prophet: Seasonal trends and other recurrent patterns in time series data are well captured by this methodology. It is especially helpful in managing significant data gaps and accounting for vacations or other special occasions that may have an impact on air quality.
   - LSTM: Long-term dependencies in the time series data are captured by the LSTM model during training, which is essential for comprehending how historical air quality and weather patterns affect present-day AQI readings.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

- GRU: As a compromise between the Prophet and LSTM models, the GRU model's streamlined design is trained to capture both medium-term trends and short-term perturbations.

3. Ensemble Strategy:

   Once the individual models are trained, the next step involves integrating their outputs into a single, unified prediction. The ensemble strategy typically uses a weighted averaging method, where the weights are assigned based on the performance of each model during cross-validation. This ensures that models with higher accuracy have a greater influence on the final prediction, while those with lower accuracy contribute less.

4. Prediction and Evaluation:

   The outputs of the Prophet, LSTM, and GRU models are combined by the ensemble model to provide the final AQI forecast. Metrics like Mean Absolute Error (MAE), R-squared (R^2), and Root Mean Square Error (RMSE) are then used to assess the performance of this hybrid ensemble model to make sure it performs better than the individual models. To evaluate the robustness and generalizability of the model, cross-validation techniques are frequently utilized.

5. Deployment and Monitoring:

   Once the hybrid model demonstrates satisfactory performance, it is deployed for real-time AQI prediction. The system continuously receives updated meteorological and AQI data, enabling it to provide ongoing forecasts that can be used for environmental monitoring and public health planning.

   The proposed hybrid ensemble approach offers a sophisticated and effective solution for predicting AQI by combining the strengths of different time series and machine learning models. Through careful preprocessing, model training, and ensemble integration, this approach aims to provide highly accurate forecasts that can support better decision-making in environmental management and public health initiatives. The use of an ensemble strategy not only enhances predictive performance but also ensures that the model remains robust and adaptable to different conditions over time.

The line plot illustrates the accuracy with which the Prophet, LSTM, and GRU models forecast AQI. The basic trend and seasonality in the data are quite well captured by all three models. However, being recurrent neural networks, LSTM and GRU are better at identifying transient patterns and fluctuations, whereas Prophet could be more appropriate for identifying broad trends. The optimal model selection is contingent upon the particular needs of the application as well as variables like data quality and hyperparameter adjustment. These parameters may be carefully considered in order to determine which model is best for AQI prediction in Hyderabad.

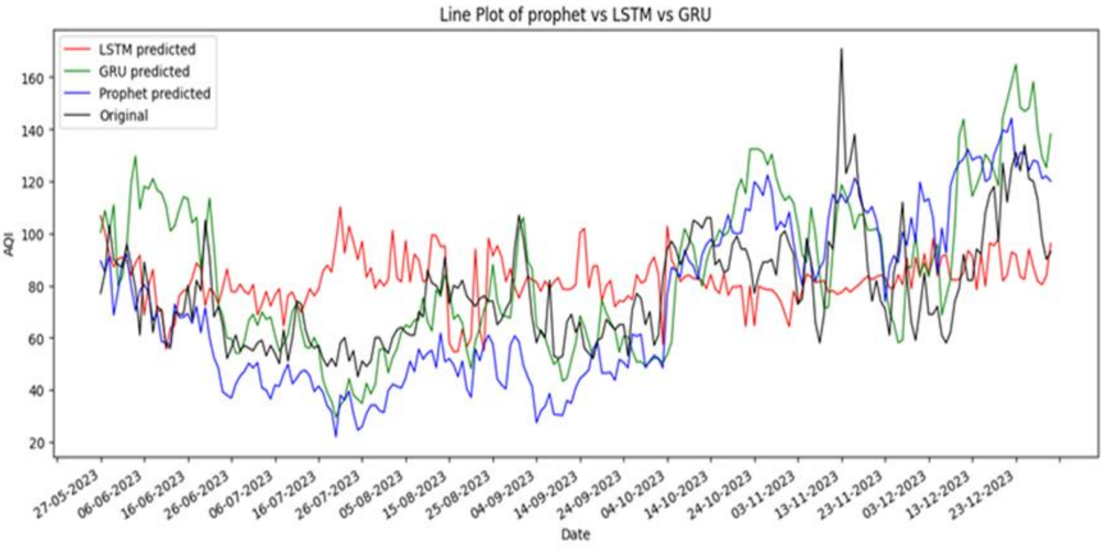

Fig 11: Actual AQI values vs (Prophet, LSTM, GRU) prediction

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

- **Model Performance:** All three models (Prophet, LSTM, and GRU) appear to extract the seasonality and overall trend from the AQI data. However, there are noticeable differences in their specific predictions.
- **LSTM and GRU:** LSTM and GRU, both recurrent neural network architectures, exhibit similar performance in capturing short-term fluctuations and patterns in the AQI data.
- **Prophet:** Prophet, while effective in capturing overall trends and seasonality, may struggle to accurately predict short-term spikes or dips in AQI.
- **Model Comparison:** The application's priorities and particular requirements may influence which model is better for example, if accurate short-term predictions are crucial, LSTM or GRU might be preferred. If capturing overall trends and seasonality is the primary goal, Prophet could be a suitable choice.

5. Conclusion

The proposed hybrid ensemble approach for Air Quality Index (AQI) prediction combines the strengths of Prophet, models such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) deliver more accurate and robust forecasts. By integrating these models through a weighted ensemble strategy, the approach effectively captures both transient variations and persistent patterns in data on air quality, resulting in reduced error rates and more accurate forecasts.

This model not only provides better performance in terms of reducing metrics like RMSE but also demonstrates robustness and adaptability, making it suitable for real-time AQI forecasting in dynamic and complex environments. This hybrid ensemble system holds significant potential for supporting urban planning, public health initiatives, and environmental policy-making.

There are several potential avenues for expanding and improving this hybrid ensemble approach:

Incorporation of Additional Data Sources: Future work can involve integrating more diverse data sources such as satellite imagery, traffic patterns, and industrial emissions data to further improve the model's accuracy and comprehensiveness in predicting AQI.

Advanced Machine Learning Techniques: The ensemble model's prediction power might be increased by investigating the application of sophisticated machine learning techniques like deep reinforcement learning or attention mechanisms, which would enable it to discover more intricate associations in the data.

Hybrid Ensemble Model Deployment in Real-Time Systems Integrated with Internet of Things (IoT) Devices: Using the hybrid ensemble model in real-time systems integrated with IoT devices, including air quality sensors and weather stations, the model can continuously receive live data, enabling it to make real-time predictions and provide immediate insights for decision-makers.

Transfer Learning for Different Locations: The model can be adapted and fine-tuned for different geographical locations with varying environmental conditions, allowing for more localized and precise AQI predictions. Transfer learning techniques can be employed to re-use the model's knowledge across different regions.

Scenario-Based Forecasting: Future extensions could focus on scenario-based AQI predictions, simulating various environmental changes (e.g., pollution control measures or natural disasters) and providing forecasts under different hypothetical situations, thus assisting policymakers in proactive planning.

Public Awareness and Health Impact Predictions: Integration with health data to forecast not only air quality but also its direct impact on human health can provide actionable insights for the public, offering early warnings for vulnerable populations during periods of poor air quality.

Through these enhancements, the proposed hybrid ensemble approach can continue to evolve, playing a crucial role in predictive modelling for air quality, ensuring healthier living environments, and contributing to more informed policy decisions.

Acknowledgements

We thank Institute of Aeronautical engineering college for providing research facilities to carry out my research work

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and special thanks to KLEF, Vaddeswaram, Guntur,522 502, Andhra Pradesh, India for encouraging scholars by providing guidance and provisions to take results on time. From the bottom of my heart I am thanking, Venkatram N, G. China Satyanarayana and L V Narasimha Prasad for supporting and guiding on my PhD journey.

Author's contributions

Hampika Gorla, Venkatram N and L V Narasimha Prasad contributed to the propose methodology and implementation of the research to the analysis of the designed model results and to writing the manuscript.

Availability of data and materials

The datasets in this research article are available in the following links.

Data: https://airquality.cpcb.gov.in/

https://www.visualcrossing.com/weather-history/Hyderabad%2CIndia

References:

[1] Landrigan, P. J., Fuller, R., Acosta, N. J. R., Adeyi, O., Arnold, R., Basu, N., ... & Zhong, M. (2018). The Lancet Commission on pollution and health. Vol:The Lancet, 391(10119), 462-512.

[2] Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, Estep & Pope, C. A., III. (2017). Estimates and 25- year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907-1918.

[3] Title: Impact of Industrialization, Urbanization, and Energy Consumption on Environmental Degradation: Evidence from India. Authors: Sohail Farooq, Shabana Parveen, Habib Elahi Sahibzada . Citation: Global Economics Review, Volume IV, Issue II, 2019

[4] Ruhul Salim, Shuddhasattwa Rafiq, Sahar Shafiei : ADBI Working Paper 718, 2017 : Urbanization, energy consumption, and pollutant emission in Asian developing economies: An empirical analysis :
https://www.adb.org/sites/default/files/publication/239591/adbi-wp718.pdf

[5] Jonathan I. Levy, Alex L. Apte, Loretta J. Mickley, Dallas Burtraw, Erica Myers, Stephen Pacala, Arlene M. Fiore : The impact of power plants on local air quality in the United States : Atmospheric Environment, Volume 43, Issue 33, 2009, Pages 5062-5070.

[6] Turner, M. C., Krewski, D., Pope, C. A., III, Chen, Y., Gapstur, S. M., Diver, W. R., ... & Burnett, R. T. (2016). Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. American Journal of Respiratory and Critical Care Medicine, 194(3), 289-298.

[7] Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., ... & Hoek, G. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. The Lancet, 383(9919), 785-795.

[8] Box, G. E. P. and Jenkins, G. M.: Time Series Analysis:
Forecasting and Control, Holden-Day, San Francisco, 1970.

[9] Zhang, Y., & Box, G. E. P. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 17(1), 3-22 (2001).

[10] Masood, A.; Ahmad, K. A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. J. Clean. Prod. 2021, 322, 129072.

[11] Qin, Y., Liu, F., & Li, Q. A novel combined prediction model for PM2.5 concentration based on secondary decomposition algorithm and optimized back propagation neural network. Science of the Total Environment, 565, 451-464 (2016).

[12] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Title: Deep learning for time series classification:
A review. Citation: Data Mining and Knowledge Discovery, 33(4), 917-963.

[13] Bandara, K., Bergmeir, C., & Smyl, S. (2020). Title:
Forecasting across Time Series Databases using Recurrent Neural Networks on Groups of Similar Series: A Clustering