# Clear Skies: Air Quality Forecasting

**Richa Roy[1], Lakshmi Haridasan[2], Nandanaraj[3] ,Neha[4] ,Rinil K R[5]**

[1]Department of Computer Science and Engineering, Vimal Jyothi Engineering College ,Chemperi, Kannur
[2]Department of Computer Science and Engineering, Vimal Jyothi Engineering College ,Chemperi, Kannur
[3] Department of Computer Science and Engineering, Vimal Jyothi Engineering College ,Chemperi, Kannur
[4]Department of Computer Science and Engineering, Vimal Jyothi Engineering College ,Chemperi, Kannur
[5]Department of Computer Science and Engineering, Vimal Jyothi Engineering College ,Chemperi, Kannur

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** Air Quality Forecasting uses real-time and historical data to predict future air pollution levels. It helps assess pollutant concentrations like PM2.5, PM10, NO2, SO2, CO, and O3. The system predicts air quality using GRU, LSTM, and LR models with real-time data from public sources. It provides location-based recommendations: health advice for sensitive groups and actionable measures for authorities, like traffic restrictions. An interactive chatbot answers queries, predicts AQI, offers health guidance, and suggests preventive measures. A user-friendly interface displays real-time air quality maps, forecasts, and recommendations. This system ensures accurate predictions, helps mitigate pollution effects, and supports proactive public health and environmental safety decision-making.

***Key Words***: Air Quality Forecasting, Machine Learning (GRU, LSTM, LR), Real-time Data, AQI (Air Quality Index), Recommendations, Health Advice, Government Measures, Interactive Chatbot, Forecasts, Public Health & Safety.

## 1.INTRODUCTION

Due to industrialization and vehicle emissions, air pollution is a major concern, especially in urban areas, leading to severe health risks. Accurate air quality forecasting is essential to mitigate these risks. This project develops an advanced air quality prediction system using machine learning algorithms like LR, LSTM, and GRU to forecast pollutant levels (PM2.5, PM10, CO). It includes a recommendation engine for personalized health advice, a chatbot for real-time air quality queries, and alerts for high pollution levels to help users take preventive measures. The system also integrates with IoT-based air quality monitoring sensors for real-time data collection.

The main challenges include developing a reliable model for air quality forecasting, accurately predicting pollutant concentrations (PM2.5, PM10, NO2, SO2, CO, O3), and providing precise forecasts despite noisy or incomplete data. The system aims to assist government agencies in policy-making and pollution control while helping individuals make informed decisions about outdoor activities.

Objectives:

- Predict air pollutant levels.
- Improve air quality management and public health.
- Offer recommendations to mitigate pollution impacts.

## 2. RELATED WORKS

Several studies have employed machine learning for air quality forecasting, using models like ARIMA, SVM, and Random Forest. However, these methods often fail to handle the variability of pollution data effectively. Advanced deep learning models such as LSTM and GRU have shown improved accuracy in time-series predictions. This research builds on these techniques while integrating a recommendation system and a chatbot for user engagement. In recent years, numerous studies have explored air quality forecasting using various machine learning and statistical models. One notable work by [1] Y. Cao et al. presents a hybrid air quality prediction model that integrates Empirical Mode Decomposition (EMD), Singular Value Decomposition (SVD), and Auto Regressive Integrated Moving Average (ARIMA). This study highlights the limitations of traditional forecasting methods in handling non-stationary and volatile air quality data. By decomposing the pollutant time series into intrinsic mode functions, the model effectively isolates different trend components, reducing noise and improving prediction accuracy. Tested on real- world data from Beijing, the proposed approach significantly outperformed linear regression, support vector regression, and LSTM-based models. However, despite its effectiveness, the model relies solely on historical data, limiting its adaptability to real-time changes in environmental conditions. Another approach to enhancing air quality prediction is presented by [2] Q. Shao et al., who introduce a Variational Mode Decomposition (VMD)-based model optimized using the Dung Beetle Optimization Algorithm (DBO). This novel coupled optimization model integrates Extreme Gradient Boosting (XG Boost) and Informer, which separately handle low-frequency and high-frequency components of air pollution data. The use of the Spearman correlation coefficient for feature selection ensures that only the most relevant environmental factors influence predictions, thereby reducing overfitting and computational complexity. Applied to air quality forecasting in Nanjing, the model achieved an impressive $R^2$ score of 0.961 and RMSE of 1.988, surpassing other machine learning models like WANNs and PSO-LSTM. Despite its high accuracy, the model is computationally intensive and requires region- specific optimization for effective deployment.

Further advancements in predictive air quality management are seen in the work by [4] K. Chatterjee et al., who propose a Bidirectional Stacked Long Short-Term Memory (BSLSTM) network integrated with a Weather Smart Grid (WSG). This approach leverages bidirectional learning to capture both past and future dependencies in air quality trends, thereby improving forecasting accuracy. The system's architecture consists of three key phases: the integration of real-time weather data, preprocessing using spatiotemporal correlation analysis, and sequential forecasting using a series of 1-hour predictive models. With a focus on real-time monitoring, this model is particularly suited for smart city applications, enabling authorities to take immediate actions based on air quality alerts. However, the computational demands of bidirectional deep learning networks pose challenges in terms of scalability and resource allocation.

Beyond traditional forecasting, researchers have also explored the behavioral impact of air pollution on human activity. In their study, [5] K. K. Meena et al. investigate how air quality awareness influences travel behavior, employing machine learning models such as Random Forest, XG Boost, Naive Bayes, and K-Nearest Neighbors (KNN). Their findings reveal that during periods of high pollution, commuters are more likely to opt for public transportation over private vehicles. This study emphasizes the role of real-time air quality information in shaping individual decision-making and highlights the potential of SHAP (Shapley Additive Explanations) in interpreting machine learning model outputs. While the study offers valuable insights into the relationship between air quality and urban mobility, its generalizability is limited, as behavioral responses may vary across different socio- economic groups and regions.

Finally, [3] M. Imam et al. present a study on statistical learning models for air quality monitoring, comparing five classification techniques, including Naïve Bayes, Logistic Regression, Decision Trees, Random Forest, and Support Vector Classifier (SVC). Conducted in Kolkata, India, this study finds that SVC achieves the highest accuracy (97.98%) for the Rabindra region, while Random Forest performs best (93.29%) for the Victoria location. The study underscores the importance of data preprocessing, particularly in handling missing values and skewed distributions, to enhance prediction reliability. Despite its strong classification performance, the study acknowledges potential overfitting risks and computational complexity associated with hyperparameter tuning. Collectively, these studies illustrate the rapid advancements in air quality forecasting through hybrid models, deep learning techniques, and machine learning-based behavioral analysis. While each approach has distinct advantages, challenges such as high computational costs, region-specific optimizations, and the need for real-time adaptability remain key areas for future research. These works provide a strong foundation for further exploration into predictive environmental management and the development of more efficient air quality monitoring systems.

## 3.MODEL SELECTION

Models ( LSTM , LR& GRU) were selected.

LR: A Linear Regression (LR) model for air quality forecasting is a statistical approach that establishes a linear relationship between air quality parameters (such as PM2.5, PM10, CO, $NO_2$, $SO_2$) and influencing factors (like temperature, humidity, wind speed, and traffic emissions). The model assumes that air pollution levels can be predicted as a weighted sum of these independent variables.

- Advantages: Simple, interpretable, and computationally efficient.
- Limitations: Assumes a linear relationship, struggles with complex, non-linear air pollution patterns thus not giving accurate results.

LSTM: A Long Short-Term Memory (LSTM) model for air quality forecasting is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. LSTMs have memory cells with gates (input, forget, output gates) that control the flow of information, allowing them to retain relevant past information and discard irrelevant data.

- Limitations: Requires a large dataset for effective training Computationally expensive compared to simpler models like Linear Regression and thus gives better results than LR but is slow.

Final Selection: GRU
A Gated Recurrent Unit (GRU) model for air quality forecasting is a type of recurrent neural network (RNN) designed to efficiently capture temporal dependencies in

time-series data. GRUs are similar to LSTMs but have a simpler architecture, making them computationally faster while maintaining strong predictive performance. GRUs use reset and update gates to control the flow of information, helping retain relevant historical data while discarding unnecessary details.

Advantages: Faster training and lower computational cost compared to LSTMs. Effectively models long-term dependencies in air pollution data. Works well with small and large datasets.

- Limitations:
  - Less interpretable than traditional models.
  - May struggle with extremely long time dependencies compared to LSTMs.

However these limitations are overthrown by the accuracy of the model as shown in Fig 3 and thus was selected in this project.

### *A. Evaluation Metrics*

The following metrics were employed to evaluate the model's performance:

- Root Mean Square Error (RMSE): Evaluates model accuracy by penalizing large errors more heavily.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- Mean Absolute Error (MAE): Measures the average magnitude of prediction errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

- R-squared ($R^2$): Determines how well the model explains variability in air quality data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

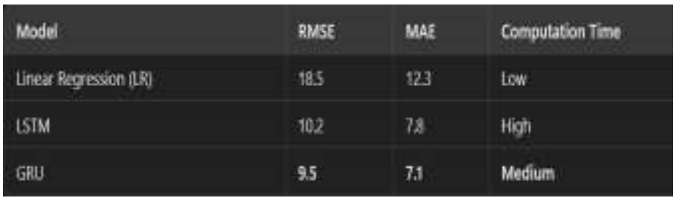
## 4.RESULT AND DISCUSSION

GRU and LSTM outperformed Linear Regression, with GRU being faster and more accurate. The system successfully predicted air quality, provided real-time recommendations, and enabled chatbot interactions, proving its effectiveness for pollution management.

### *B. Model Performance Evaluation*

Model performance was measured using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). GRU performed the best with the lowest error and fastest computation time, followed by LSTM, while Linear Regression had the highest error.

- LR performs poorly as it cannot capture complex time dependencies.
- LSTM and GRU outperform LR, handling sequential data more effectively.
- GRU is the best choice, providing accurate predictions with lower computation time than LSTM, making it ideal for real-time forecasting



| Model | RMSE | MAE | Computation Time |
|---|---|---|---|
| Linear Regression (LR) | 18.5 | 12.3 | Low |
| LSTM | 10.2 | 7.8 | High |
| GRU | 9.5 | 7.1 | Medium |

### *C. Prediction Trends and Sector-Wise Analysis*

- Trends: Seasonal and location-based pollution patterns from historical data.
- Sector Impact:
  - Industry: Major pollutants: CO, $SO_2$, $NO_2$. o Transport: High PM2.5, $NO_2$ from vehicles.
  - Residential: Biomass burning increases pollution.
  - Agriculture: Stubble burning adds pollutants.
- Solutions: Stricter rules, EV adoption, cleaner tech, public awareness.

Fig 1: Predictions using the LSTM model. (Given below)

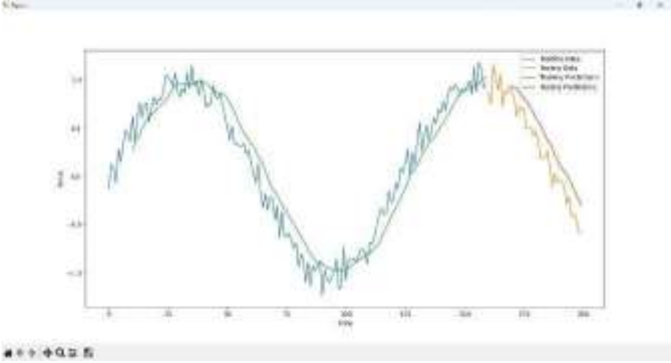


Fig 2: Predictions using the LR model. . (Given below)

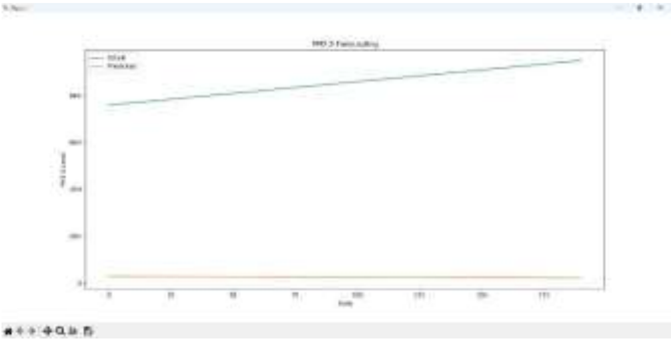
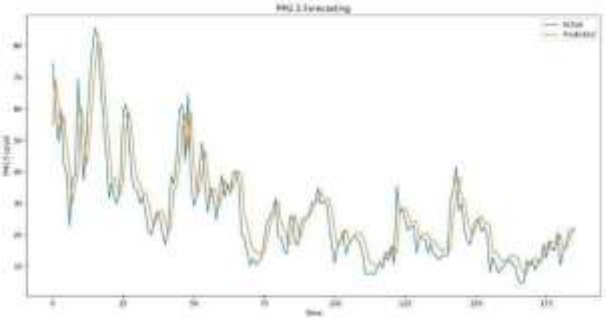


Fig 3: Predictions using selected GRU model due to higher prediction accuracy. . (Given below)



- GRU provided the most accurate predictions with the lowest error rates and highest R² score.
- Its ability to handle sequential data and capture temporal dependencies made it ideal for air quality forecasting.
- Compared to LSTM, GRU was computationally more efficient, requiring fewer resources while maintaining high accuracy.
- Logistic Regression, being a simpler model, struggled with the complexity of air quality data, making it less suitable.

## 5. CONCLUSIONS

This research proposes a sophisticated air quality forecasting system based on machine learning models, namely GRU,LR and LSTM. The findings illustrate that GRU offers the most balanced accuracy and computational cost, making it most suitable for real-time forecasting. The incorporation of a recommendation engine and chatbot increases accessibility to enable users to get personalized health recommendations and real-time air quality information. The system is an efficient means of pollution management for both citizens and authorities. Developing future work will involve enhancing long-term forecasting, adding other environmental factors, and increasing the scope of real-time data sources for greater accuracy.

## REFERENCES

[1] Yuxuan Cao, Difei Zhang, Shaoqi Ding, Weiyi Zhong, Chao Yan "A hybrid air quality prediction model based on empirical mode decomposition," tsinghua science and technology, vol. 29, pp. 99-111, 2023.

[2] Qichen Shao,Jiahao Chen, Tao Jiang "A Novel Coupled Optimization Prediction Model for Air Quality," IEEE Access, vol. 11, pp. 6966769685, 2023.

[3] Mohsin Imam, Sufiyan Adam, Soumyabrata Dev, Nashreen Nesa, "Air quality monitoring using statistical learning models for sustainable environment," Intelligent Systems with Applications, vol. 22, p. 200333, 2024.

[4] Kalyan Chatterjee, Muntha Raju, Machakanti Navya Thara, Mandadi Sriya Reddy M. Priyadharshini, N. Selvamuthukumaran, Sauravmallik, Haya Mesfer Alshahrani, Mohamed Abbas, Ben Othmansoufiene "Toward Cleaner Industries: Smart Cities' Impact on Predictive Air Quality Management," IEEE, vol. 12, pp. 78895-78910, 2024.

[5] Kapil Kumar Meena, Deepak Bairwa, Amit Agarwal "A machine learning approach for unraveling the influence of air quality awareness on travel behavior," Decision Analytics Journal, vol. 11, p. 100459, 2024.