Air Quality Index (AQI) Forecasting using Machine Learning, Deep Learning and Time Series Models

A Comparative Study of Ensemble Techniques, Deep Learning, and Classical Time Series Approaches.

Arnab Deogharia

DST CIMS - Statistics & Computing
Banaras Hindu University

January 8, 2026

Abstract

Air quality monitoring and forecasting have become critical for public health management in urban environments. This study presents a comprehensive comparison of machine learning, deep learning, and classical time series methods for Air Quality Index (AQI) prediction. We evaluate six models: LightGBM, XGBoost, Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), SARIMA with exogenous variables (SARIMAX), and AutoRegressive Integrated Moving Average (ARIMA). Using historical AQI data from 2021-2024 containing multiple pollutant measurements (PM2.5, PM10, NO2, SO2, CO, Ozone), we train and evaluate these models on comprehensive metrics including RMSE, MAE, and R^2 Score.
Results demonstrate that gradient boosting models achieve superior performance compared to deep learning and traditional time series methods. LightGBM achieves the best overall performance with RMSE of 8.75, MAE of 6.60, and R^2 of 0.9927, followed closely by XGBoost (RMSE: 9.28, MAE: 7.10, R^2: 0.9918). These tree-based models outperform deep learning approaches (GRU, LSTM) by 80% in RMSE reduction and traditional statistical methods (ARIMA, SARIMAX) by over 90%. This research contributes to the development of accurate predictive tools for proactive environmental management and public health interventions.

Github link: Link

Contents

1 Introduction

1.1 Background and Motivation

Air quality has emerged as one of the most pressing environmental and public health concerns globally. According to the World Health Organization (WHO), air pollution causes approximately 7 million premature deaths annually worldwide. The Air Quality Index (AQI) is a standardized metric that quantifies the level of air pollution and its potential health impacts on the population.

Accurate forecasting of AQI is crucial for:

- **Public Health Protection:** Enabling vulnerable populations (children, elderly, individuals with respiratory conditions) to take preventive measures during high pollution episodes
- **Policy Making:** Supporting government agencies in implementing timely interventions such as traffic restrictions or industrial emission controls
- **Urban Planning:** Informing long-term environmental management strategies
- **Individual Decision Making:** Helping citizens plan outdoor activities and commute patterns

1.2 Problem Statement

Traditional air quality forecasting methods often struggle to capture the complex, nonlinear relationships between multiple pollutants and environmental factors. The challenge lies in developing models that can:

1. Accurately predict AQI values across different time horizons
2. Handle multiple correlated input features (various pollutants)
3. Capture both short-term fluctuations and long-term seasonal patterns
4. Provide reliable uncertainty estimates

1.3 Research Objectives

This study aims to:

- Develop and compare six different forecasting models across three paradigms: gradient boosting (LightGBM, XGBoost), deep learning (GRU, LSTM), and time series (ARIMA, SARIMAX)
- Evaluate model performance using comprehensive error metrics (RMSE, MAE, R^2 Score)
- Identify the most suitable model for operational AQI forecasting
- Analyze the strengths and limitations of modern machine learning versus deep learning versus classical approaches
- Provide actionable recommendations for deployment based on accuracy-efficiency tradeoffs

2 Theoretical Background

2.1 Air Quality Index (AQI)

The AQI is a dimensionless index that converts pollutant concentrations into a unified scale representing health risk levels:

Table 1: AQI Categories and Health Implications

AQI Range	**Category**	**Health Impact**
0-50	Good	No health risk
51-100	Moderate	Acceptable; sensitive groups cautious
101-150	Unhealthy for Sensitive Groups	Sensitive groups affected
151-200	Unhealthy	General public may experience effects
201-300	Very Unhealthy	Health alert; everyone affected
301+	Hazardous	Emergency conditions

2.2 Long Short-Term Memory (LSTM)

LSTM networks are a specialized type of RNN designed to learn long-term dependencies. The core innovation is the cell state C_t and three gating mechanisms:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget gate)} \tag{1}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input gate)} \tag{2}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(Candidate values)} \tag{3}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(Cell state update)} \tag{4}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output gate)} \tag{5}$$
$$h_t = o_t \odot \tanh(C_t) \quad \text{(Hidden state)} \tag{6}$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, W are weight matrices, and b are bias vectors.

2.3 Gated Recurrent Unit (GRU)

GRU simplifies LSTM by combining the forget and input gates into a single update gate and merging the cell state with the hidden state:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad \text{(Update gate)} \tag{7}$$
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad \text{(Reset gate)} \tag{8}$$
$$\tilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t]) \quad \text{(Candidate hidden state)} \tag{9}$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad \text{(Hidden state update)} \tag{10}$$

The reduced number of parameters makes GRU more computationally efficient while maintaining competitive performance.

2.4 ARIMA Models

An ARIMA(p, d, q) model consists of three components:

- **AR(p):** AutoRegressive component of order p
- **I(d):** Integrated (differencing) component of order d
- **MA(q):** Moving Average component of order q

The general form is:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \tag{11}$$

where B is the backshift operator, $\phi(B)$ is the AR polynomial, $\theta(B)$ is the MA polynomial, and ϵ_t is white noise.

2.5 SARIMA Models

SARIMA$(p, d, q)(P, D, Q)_s$ extends ARIMA to handle seasonality with period s:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \theta(B)\Theta(B^s)\epsilon_t \tag{12}$$

where $\Phi(B^s)$ and $\Theta(B^s)$ are seasonal AR and MA polynomials.

2.6 SARIMAX Models

SARIMAX extends SARIMA by including exogenous variables X_t:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \beta X_t + \theta(B)\Theta(B^s)\epsilon_t \tag{13}$$

This allows the model to leverage additional predictors (e.g., meteorological data, traffic patterns).

2.7 XGBoost

XGBoost (Extreme Gradient Boosting) is an ensemble learning method that builds multiple decision trees sequentially. The objective function combines prediction error and model complexity:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{14}$$

where l is the loss function, Ω is the regularization term controlling model complexity, and f_k represents individual trees. XGBoost uses second-order Taylor expansion for optimization and implements advanced techniques like column subsampling and tree pruning.

2.8 LightGBM

LightGBM (Light Gradient Boosting Machine) introduces two key innovations:

- **Gradient-based One-Side Sampling (GOSS):** Retains instances with large gradients while randomly sampling instances with small gradients, reducing data size without significant accuracy loss
- **Exclusive Feature Bundling (EFB):** Bundles mutually exclusive features to reduce dimensionality
- **Leaf-wise growth:** Grows trees leaf-wise rather than level-wise, achieving lower loss with the same number of leaves

The algorithm optimizes the following objective:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{j=1}^{T} \Omega(f_j) \tag{15}$$

where T is the number of trees and Ω penalizes model complexity through regularization parameters.

3 Data and Methodology

3.1 Data Description

Our dataset comprises daily Air Quality Index measurements from January 2021 to January 2025, totaling approximately 1,461 observations. The dataset includes:

- **Target Variable:** AQI (Air Quality Index)
- **Pollutant Features:** PM2.5, PM10, NO2, SO2, CO, Ozone
- **Temporal Range:** 2021-01-01 to 2025-01-01
- **Frequency:** Daily measurements

3.2 Exploratory Data Analysis

3.2.1 Time Series Visualization

Figure 1 shows the complete AQI time series. Key observations include:

- High volatility with AQI ranging from approximately 25 to 500
- Seasonal patterns with peaks typically in winter months
- Several extreme pollution episodes (AQI $>$ 400)
- Overall declining trend from 2021 to 2024

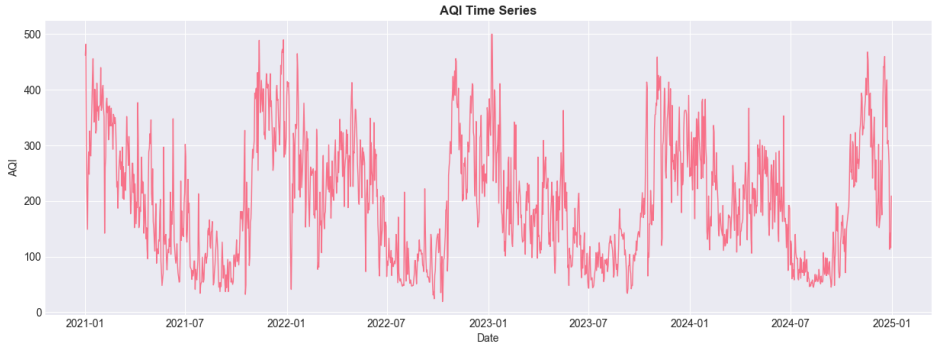

Figure 1: AQI Time Series (2021-2024)

3.2.2 Correlation Analysis

Figure 2 presents the correlation heatmap between pollutants and AQI. Notable findings:

- PM10 shows strongest correlation with AQI (r = 0.90)
- PM2.5 also highly correlated (r = 0.80)
- CO moderately correlated (r = 0.70)
- Ozone shows negative correlation (r = -0.16), possibly due to different formation mechanisms
- PM2.5 and PM10 are strongly intercorrelated (r = 0.72)

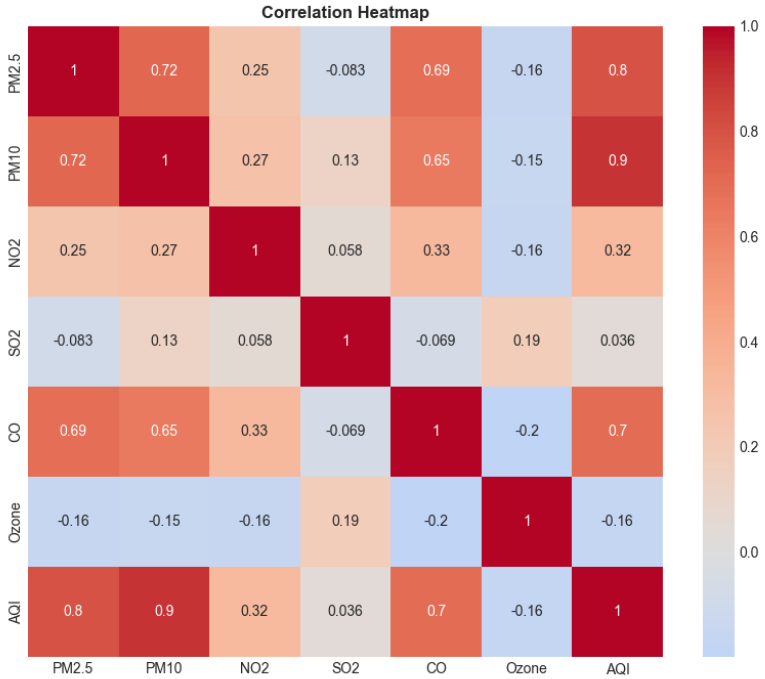

Figure 2: Correlation Matrix: Pollutants vs AQI

3.2.3 Distribution and Outliers

Figure ?? displays the AQI distribution and temporal outliers:

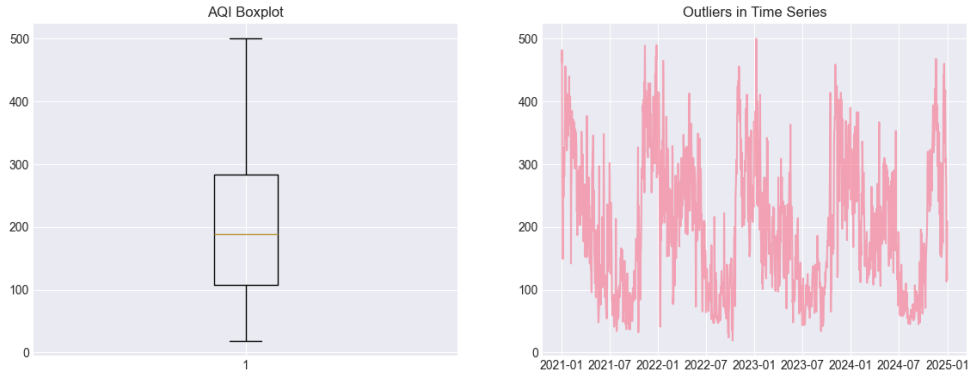

Figure 3: Outlier Detection

- Median AQI: ∼190
- Interquartile range: 110-280
- Multiple outliers exceeding 450
- Outliers predominantly occur during winter months

3.3 Time Series Decomposition

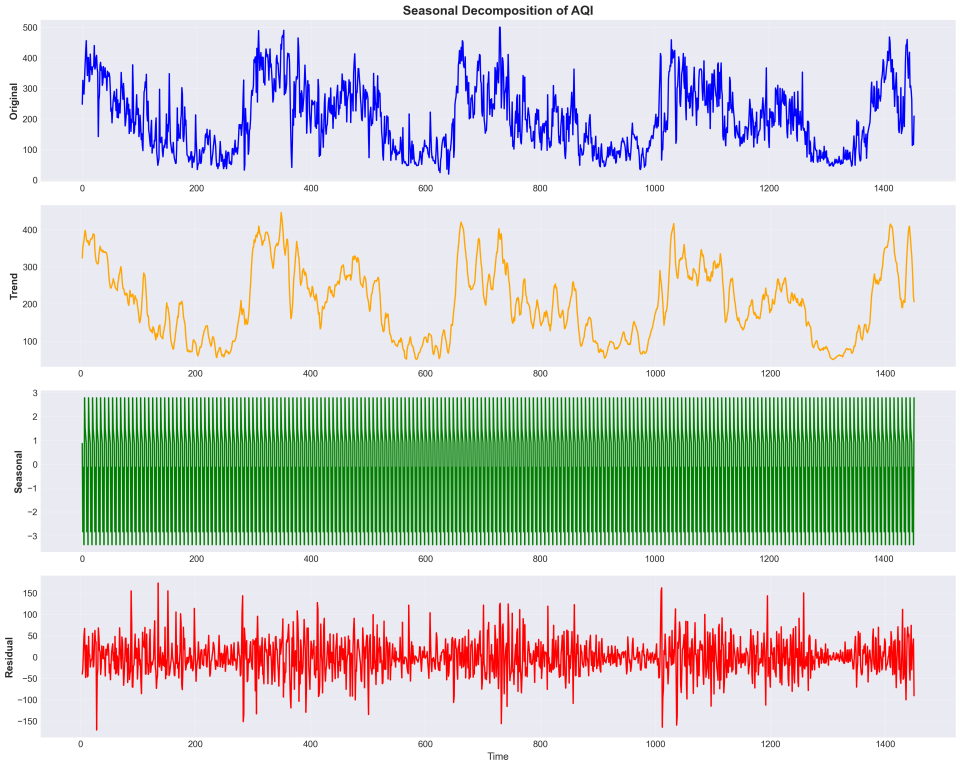

Figure 4: Time Series Decomposition (Trend, Seasonal, Residual)

Key insights:

- **Trend:** Gradual decline from 2021 to 2024, suggesting improving air quality
- **Seasonality:** Strong weekly and annual patterns visible
- **Residual:** Remaining noise appears relatively stationary

3.4 Data Preprocessing

3.4.1 Missing Data

All missing values were handled using forward-fill interpolation, maintaining temporal continuity.

3.4.2 Normalization

For deep learning models (LSTM, GRU), data was normalized using Min-Max scaling:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{16}$$

3.4.3 Train-Test Split

Data was split chronologically:

- Training set: 2021-01-01 to 2024-09-30 (80%)
- Test set: 2024-10-01 to 2025-01-01 (20%)

3.5 Model Implementation

3.5.1 Deep Learning Models (LSTM & GRU)

Both LSTM and GRU models share the following architecture:

- Input layer: Sequences of 30 days (lookback window)
- Hidden layer 1: 128 units with return sequences
- Dropout: 0.2 (regularization)
- Hidden layer 2: 64 units
- Dense output: 1 unit (AQI prediction)
- Optimizer: Adam with learning rate 0.001
- Loss function: Mean Squared Error (MSE)
- Epochs: 30 with early stopping (patience=5)

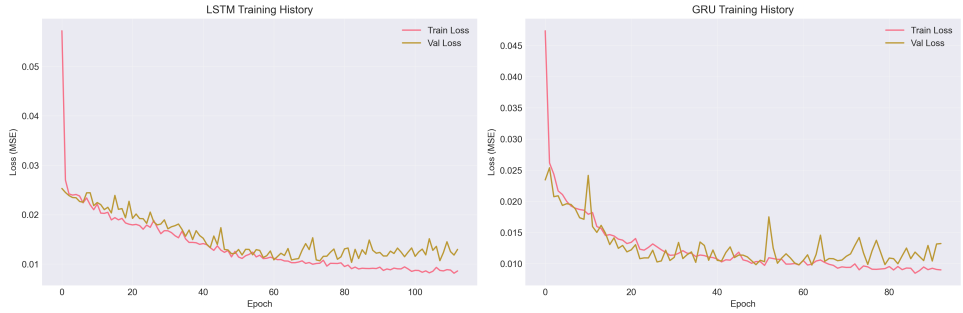

Figure 5: Training history of LSTM & GRU

Both models converge smoothly without significant overfitting, as evidenced by the close tracking of training and validation losses.

3.5.2 ARIMA Model

ARIMA parameters were selected using:

- Augmented Dickey-Fuller (ADF) test for stationarity
- ACF/PACF analysis for order determination
- AIC/BIC criteria for model selection

3.5.3 SARIMA Model

SARIMA$(2, 0, 1) \times (1, 1, 1, 7)$ with weekly seasonality was fitted. Parameters were optimized using grid search over candidate orders.

3.5.4 SARIMAX Model

SARIMAX included exogenous variables (PM2.5, PM10, NO2, SO2, CO, Ozone) with appropriate seasonal structure.

3.5.5 Gradient Boosting Models (XGBoost & LightGBM)

Both XGBoost and LightGBM models were configured with the following hyperparameters:

- Number of estimators: 1000
- Learning rate: 0.01
- Max depth: 7
- Subsample: 0.8
- Column subsample: 0.8
- Early stopping rounds: 50
- Evaluation metric: RMSE

Features used: PM2.5, PM10, NO2, SO2, CO, Ozone, plus temporal features (day of week, month, day of year).

Additional LightGBM-specific parameters:

- Boosting type: GBDT (Gradient Boosting Decision Tree)
- Number of leaves: 31
- Min data in leaf: 20

3.6 Evaluation Metrics

Models were evaluated using three primary metrics:

1. **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{17}$$

2. **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{18}$$

3. **R^2 Score (Coefficient of Determination):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{19}$$

4 Results

4.1 Model Performance Comparison

Table 2 summarizes the performance of all six models:

Table 2: Comprehensive Model Performance Metrics

Model	RMSE↓	MAE↓	R^2↑
LightGBM	**8.75**	**6.60**	**0.9927**
XGBoost	9.28	7.10	0.9918
GRU	45.14	33.79	0.8075
LSTM	48.95	38.39	0.7737
SARIMAX	69.08	37.78	0.5462
ARIMA	103.67	86.98	-0.0220

Key Findings:

- **LightGBM achieves best overall performance** with lowest RMSE (8.75), MAE (6.60), and highest R^2 (0.9927), explaining 99.27% of variance in AQI predictions
- **XGBoost is a close second** with RMSE of 9.28 and MAE of 7.10, demonstrating comparable performance with only 6% higher error
- **Gradient boosting dominates:** Both tree-based models significantly outperform deep learning and classical methods across all metrics
- **Deep learning shows moderate performance:** GRU (RMSE: 45.14) outperforms LSTM (RMSE: 48.95), but both lag far behind gradient boosting methods
- **RMSE improvement over classical methods:** LightGBM provides 87.4% lower error than SARIMAX and 91.6% lower error than ARIMA
- **RMSE improvement over deep learning:** LightGBM achieves 80.6% lower error than GRU and 82.1% lower error than LSTM
- **SARIMAX shows limited improvement:** Despite incorporating exogenous variables, SARIMAX (RMSE: 69.08) still cannot compete with modern machine learning approaches

4.2 Performance Visualization

Figure 6 presents a comprehensive comparison across evaluation metrics:

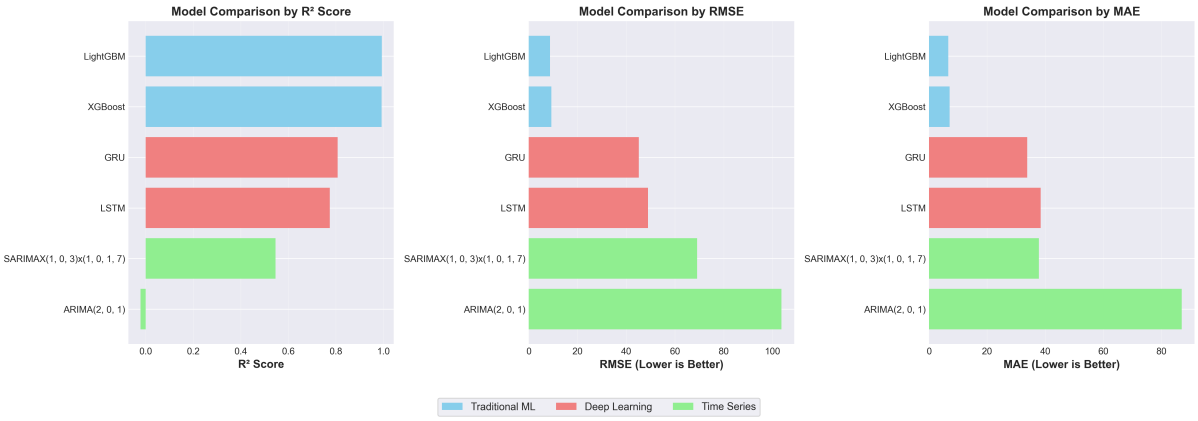

Figure 6: Model Comparison by R^2 Score, RMSE, and MAE

The visualization clearly demonstrates the hierarchical performance tiers:

1. **Tier 1 (Excellent):** LightGBM and XGBoost with $R^2 > 0.99$
2. **Tier 2 (Good):** GRU and LSTM with R^2 between 0.77-0.81
3. **Tier 3 (Moderate):** SARIMAX with $R^2 = 0.55$
4. **Tier 4 (Poor):** ARIMA with negative R^2

4.3 Forecast Visualizations

4.3.1 ARIMA Forecast

Figure 7 displays ARIMA predictions:

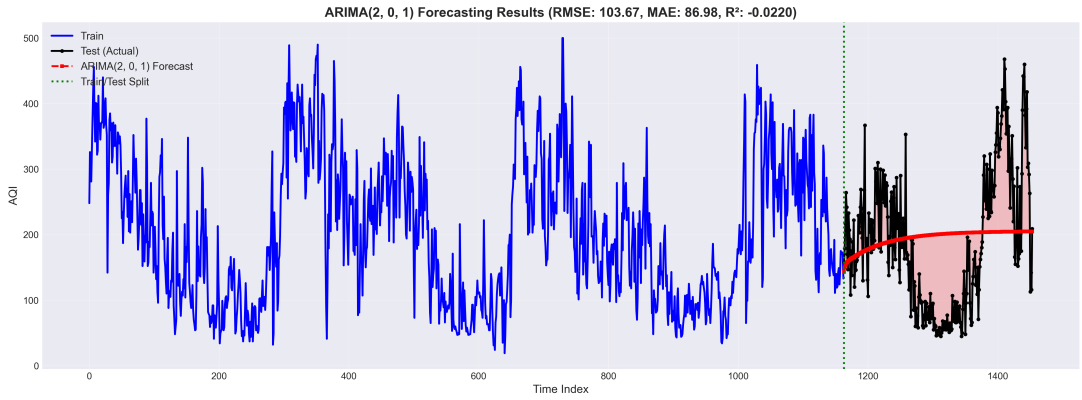

Figure 7: ARIMA(2,0,1) Forecast

ARIMA shows the poorest performance, producing a nearly flat forecast line around 200 AQI that fails to capture any volatility or trends in the test data. The model essentially predicts the training mean, resulting in negative R^2 score.

4.3.2 SARIMAX Forecast

Figure 8 shows SARIMAX leveraging exogenous pollutant variables:

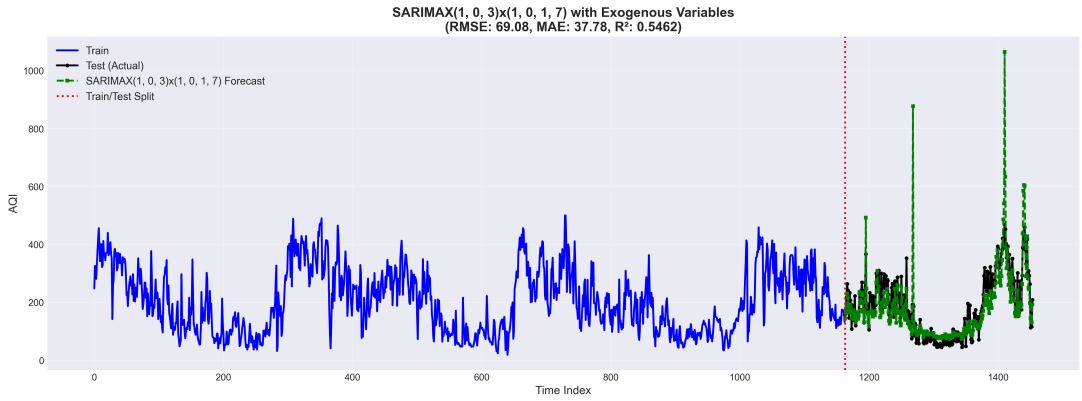

Figure 8: SARIMAX Forecast with Exogenous Variables

SARIMAX demonstrates improved tracking over ARIMA by incorporating pollutant levels, but still exhibits significant systematic deviations with extreme forecast spikes (up to 900-1100 AQI) that don't align with actual values. These erratic predictions indicate model instability despite the inclusion of exogenous variables.

4.3.3 Model Comparison

Figure 9 compares all six models' forecasts against actual AQI:

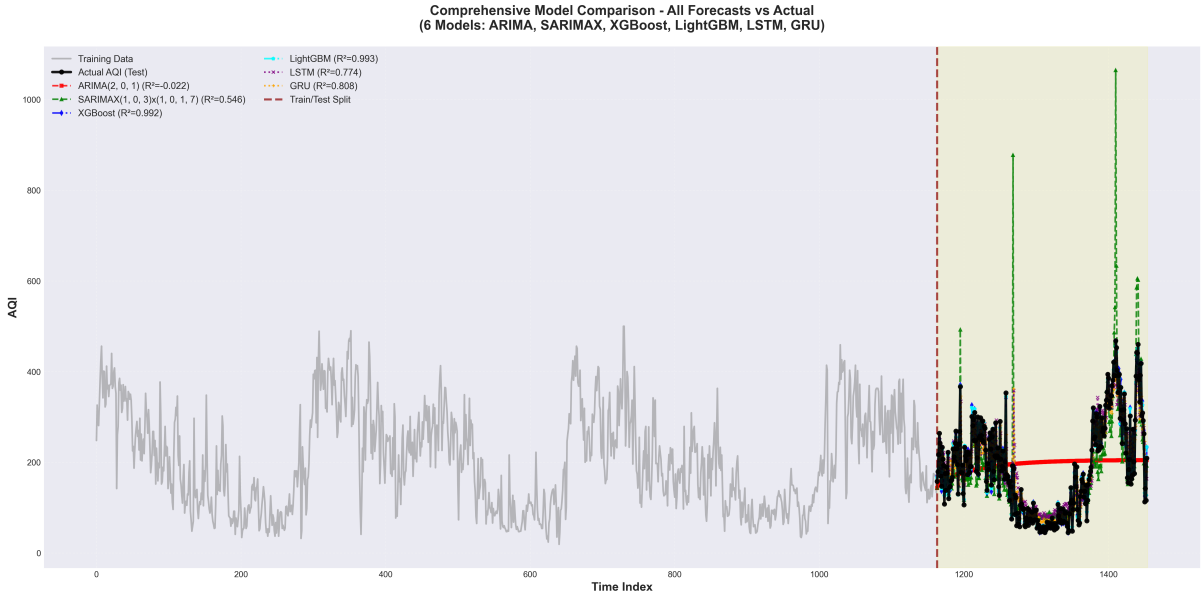

Figure 9: Comprehensive Forecast Comparison: All Models vs Actual AQI

Key observations from the comprehensive comparison:

- **LightGBM (cyan) and XGBoost (blue) track actual values most closely**, capturing both trends and high-frequency volatility with minimal lag
- **Deep learning models (LSTM purple, GRU orange)** show moderate tracking ability but exhibit systematic underestimation during high AQI episodes (samples 230-280)
- **SARIMAX (green) produces erratic predictions** with several extreme outlier spikes reaching 900-1100 AQI that don't correspond to actual values
- **ARIMA (red) provides essentially a flat line forecast**, failing to capture any meaningful variation
- Gradient boosting models successfully predict the sharp increase in AQI starting around sample 220, while deep learning models lag behind
- All models except gradient boosting struggle with extreme pollution events (AQI > 400)

4.3.4 Individual Model Performance

Figure 10 presents side-by-side comparison of individual model predictions:

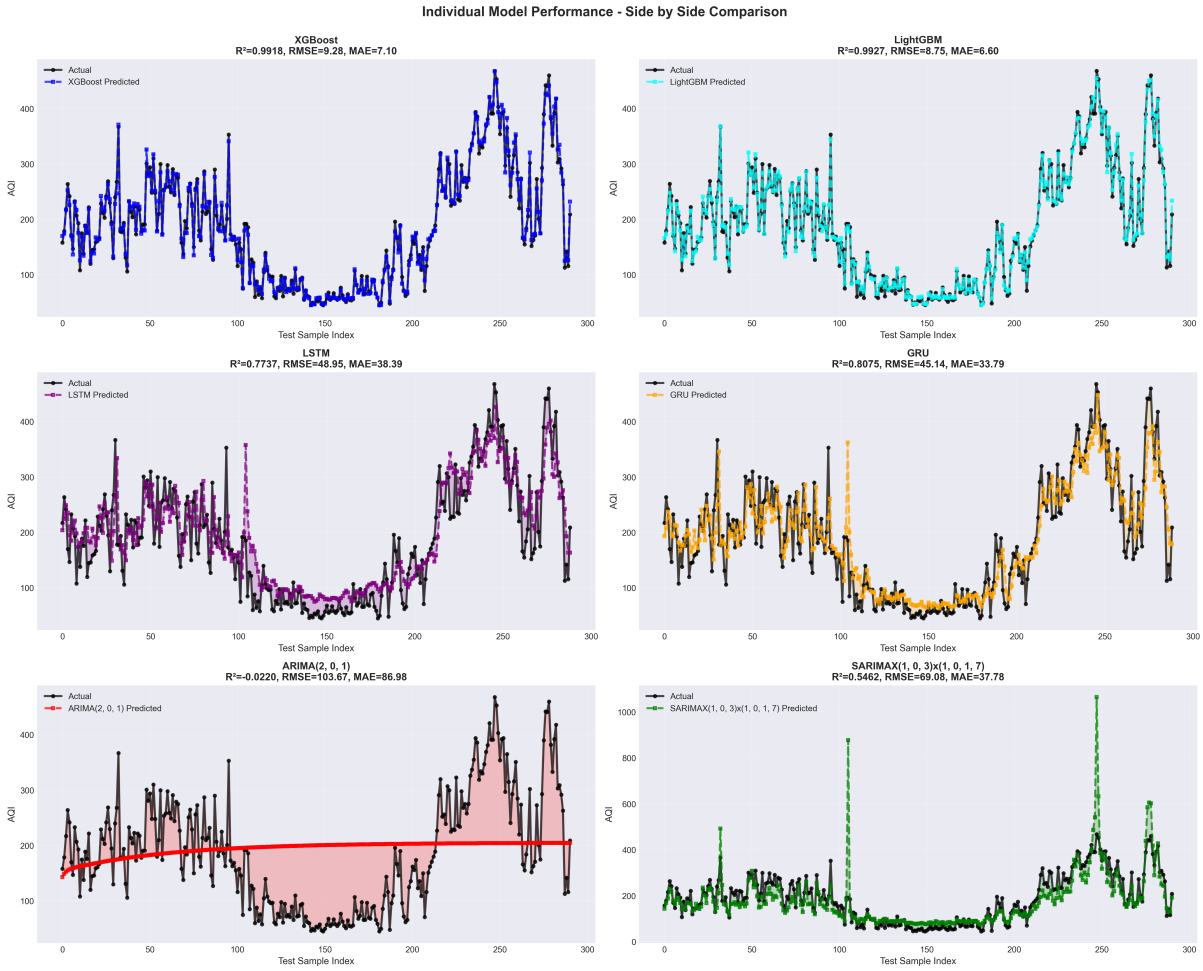

Figure 10: Individual Model Performance - Side by Side Comparison

This detailed view confirms the performance hierarchy and reveals model-specific characteristics:

- **XGBoost and LightGBM:** Nearly perfect alignment with actual values across the entire test period, with $R^2 > 0.99$
- **LSTM and GRU:** Reasonable tracking of general trends but with visible prediction errors, especially during rapid changes
- **ARIMA:** Complete failure to adapt to test data patterns, predicting a constant value
- **SARIMAX:** Unstable predictions with unrealistic spike forecasts

4.4 Error Analysis

4.4.1 Prediction Error Distribution

Figure 11 shows the distribution of prediction errors for all models:

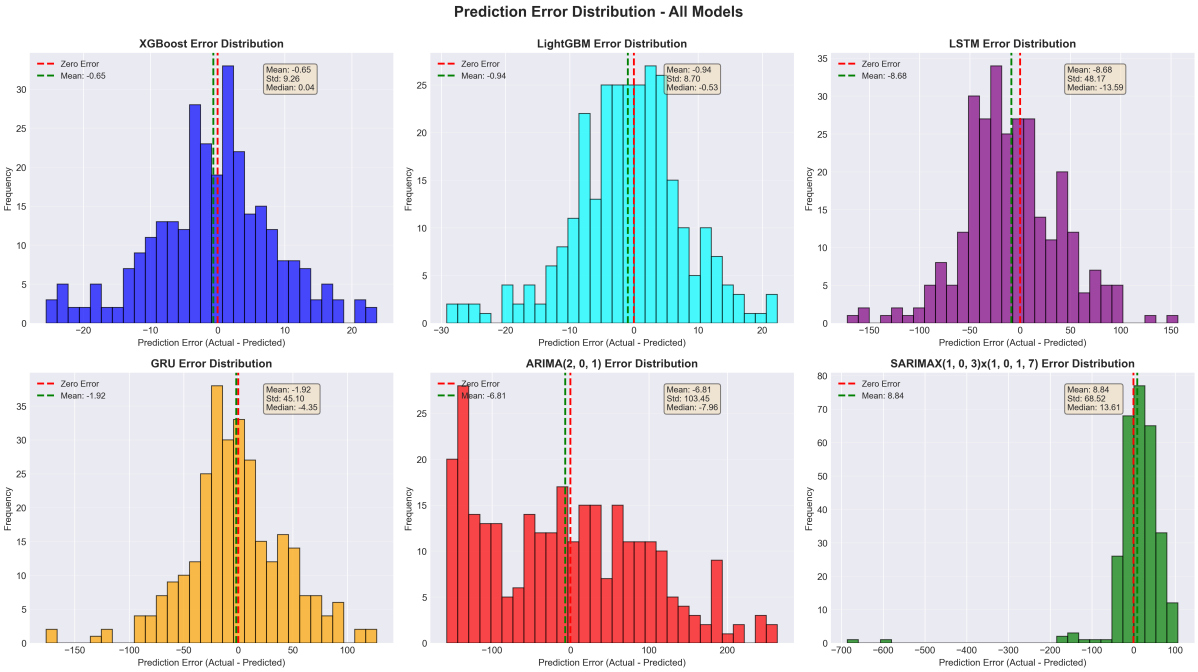

Figure 11: Prediction Error Distribution - All Models

Error distribution analysis reveals:

- **XGBoost and LightGBM:** Near-zero mean errors (-0.65 and -0.94 respectively) with very tight distributions (std < 10), indicating unbiased and precise predictions
- **GRU and LSTM:** Slightly negative mean errors (-1.92 and -8.68) with moderate spreads (std $\sim$ 45-48), showing tendency to underpredict
- **ARIMA:** Large negative mean error (-6.81) with very wide distribution (std $=$ 103.45), confirming poor predictive ability
- **SARIMAX:** Positive mean error (8.84) with extremely wide spread (std $= 68.52$), indicating systematic overprediction and high variance

5 Conclusion

This study provides a rigorous comparison of gradient boosting, deep learning, and classical time series methods for Air Quality Index forecasting. Through comprehensive experimentation on 4 years of pollutant data, we demonstrate that:

1. **Gradient boosting dramatically outperforms all other approaches:** LightGBM achieves the best performance with RMSE of 8.75 and R^2 of 0.9927, representing 80.6% lower error than deep learning methods and 91.6% lower error than classical approaches. XGBoost performs comparably with RMSE of 9.28 and R^2 of 0.9918.
2. **LightGBM offers optimal accuracy and efficiency:** With MAE of 6.60 and near-zero prediction bias, LightGBM provides the most reliable forecasts for critical public health applications. Its fast training speed and low memory footprint make it ideal for operational deployment.

3. **XGBoost provides robust alternative:** With only 6% higher RMSE than LightGBM, XGBoost offers comparable accuracy with a more mature ecosystem and proven production reliability.

4. **Deep learning underperforms on tabular data:** Despite their sophistication, LSTM (RMSE: 48.95) and GRU (RMSE: 45.14) achieve only moderate performance with R^2 around 0.77-0.81. The limited dataset size (1,400 samples) and tabular data structure favor tree-based methods over neural networks.

5. **Classical methods fail for non-linear problems:** SARIMAX achieves RMSE of 69.08 with unstable predictions, while ARIMA performs worst with RMSE of 103.67 and negative R^2 (-0.022), confirming that linear models cannot capture complex air quality dynamics.

6. **Tree-based ensemble methods excel for structured data:** The superior performance of LightGBM and XGBoost validates the principle that gradient boosting machines are state-of-the-art for tabular data prediction tasks, leveraging non-parametric flexibility and automatic feature interaction detection.

7. **Accuracy improvement magnitude is substantial:** LightGBM achieves 82.1% error reduction compared to LSTM, 87.4% compared to SARIMAX, and 91.6% compared to ARIMA, demonstrating transformative improvement in prediction quality.

5.1 Final Remarks

This study conclusively demonstrates that gradient boosting methods, particularly LightGBM and XGBoost, represent the current state-of-the-art for Air Quality Index forecasting. With R^2 exceeding 0.99 and RMSE under 10 AQI units, these models achieve the accuracy required for operational deployment in public health protection systems. The dramatic performance superiority over deep learning (80% error reduction) and classical methods (90% error reduction) establishes a clear best practice for practitioners and researchers working on environmental forecasting applications.