scientific reports

OPEN

Improving air quality prediction using hybrid BPSO with BWAO for feature selection and hyperparameters optimization

Mohamed S. Sawah[1]✉, Hela Elmannai[2], Alaa A. El-Bary[3], Kh. Lotfy[4,5] & Osama E. Sheta[4]

Air pollution poses a significant threat to public health and environmental sustainability, necessitating accurate predictive models for effective air quality management. This study uses machine learning techniques to forecast air quality through utilizing the annual AQI dataset obtained from the U.S. Environmental Protection Agency (EPA). Feature selection (FS) was conducted using Binary version of Grey Wolf Optimizer (BGWO), Particle Swarm Optimization (BPSO), Whale Optimization Algorithm (BWAO), and a novel hybrid BPSO-BWAO approach to identify the most relevant features for AQI prediction. Among the feature selection methods, BPSO achieved the best Mean Squared Error (MSE) score of 53.56, but with high variance, while BWAO demonstrated lower variance and consistent results. The hybrid BPSO-BWAO method emerged as the optimal solution, achieving an MSE of 53.93 with improved stability and feature set balance, selecting key features such as 'Days with AQI,' 'Median AQI,' 'Days CO,' 'Days NO2,' 'Days PM2.5,' 'Good_Days_Percent,' and 'Unhealthy_Days_Percent.' Machine learning models, including Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and Linear Regression (LR), were evaluated before and after feature selection. The Random Forest model achieved the best performance after feature selection with an MSE of 53.93, R^2 of 0.9710, and reduced fitted time. Further optimization using novel hybrid PSO-WAO enhanced RF performance, achieving an improved MSE of 51.82 and R^2 of 0.9821, demonstrating the efficacy of hyperparameter tuning. The study concludes that feature selection and hyperparameter optimization significantly improve model accuracy and computational efficiency, offering a robust framework for air quality forecasting.

Keywords Air quality prediction, Air quality index (AQI), Hybrid optimization, Feature selection, BPSO-BWAO-RF

Air quality (AQ) is a growing global concern due to its far-reaching effects on human health, the environment, and economic stability[1]. Human activities like industrial emissions, vehicle pollution, deforestation, and urban expansion are the main culprits behind declining air quality. Natural events such as volcanic eruptions, dust storms, and wildfires also contribute to unpredictable fluctuations in pollution levels[2]. According to the World Health Organization (WHO), more than 90% of the world's population breathes air that fails to meet recommended safety standards, leading to millions of premature deaths each year[3]. This dire situation highlights the urgent need for reliable air quality prediction systems to combat pollution and support sustainable development efforts. Managing and understanding air quality is essential for safeguarding public health and protecting the environment[4].

While traditional methods of monitoring air quality through networks of ground-based sensors have provided valuable insights, they often face limitations such as restricted spatial coverage and delayed real-time availability. These challenges underscore the need for more advanced approaches, making air quality prediction a vital tool in the fight against pollution. Air quality prediction combines scientific and computational techniques to forecast

[1]Department of Information Systems, Al Alson Higher Institute, Cairo, Egypt. [2]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. [3]Arab Academy for Science, Technology and Maritime Transport, P.O. Box 1029, Alexandria, Egypt. [4]Department of Mathematics, Faculty of Science, Zagazig University, P.O. Box 44519, Zagazig, Egypt. [5]Department of Mathematics, Faculty of Science, Taibah University, Madinah, Saudi Arabia. ✉email: me1900@fayoum.edu.eg

future pollution levels[5]. By integrating historical data, real-time monitoring, meteorological information, and insights into pollution sources and transport mechanisms, predictive models can anticipate changes in air quality[6]. These forecasts empower timely interventions, enabling authorities to issue public health advisories, implement pollution control measures, and make informed urban planning decisions[7].

Recent developments have significantly reshaped the field of air quality prediction. High-resolution data from ground-based stations, satellites, and mobile sensors now provide a strong foundation for creating and validating models. Advances in meteorological modeling deliver precise forecasts of weather conditions such as wind speed, temperature, humidity, and precipitation which are crucial for understanding how pollutants disperse and transform. Additionally, the rise of advanced statistical and machine learning techniques, including time series analysis, regression models, and artificial neural networks, has made it possible to develop highly accurate and dependable predictive models[8].

Because air pollution dynamics are highly complex, effective prediction requires a multi-faceted approach. A variety of modeling techniques are available, each with their own strengths and weaknesses. Statistical models, which rely on historical data and statistical patterns, are computationally efficient but may struggle to represent complex, non-linear relationships. Chemical transport models (CTMs), which simulate the physical and chemical processes governing pollution's emission, transport, transformation, and deposition, provide detailed insights but are computationally intensive. Hybrid models, blending statistical and CTM methodologies, aim to leverage the advantages of both approaches[9]. Meanwhile, machine learning models, particularly deep learning algorithms, excel at identifying intricate patterns and non-linear interactions in air quality data, often outperforming traditional methods in predictive accuracy[10].

The choice of an appropriate prediction model depends on factors such as the specific pollutants being studied, the availability of data, the desired prediction timeframe, and the computational resources at hand. For short-term predictions addressing immediate health concerns, statistical and machine learning models are favored for their efficiency[11]. Conversely, for long-term predictions aimed at identifying trends and informing policy, CTMs and hybrid models may be more suitable. Accuracy and reliability are important for air quality predictions to be effective. Evaluating and validating models involve comparing predictions with observed data and using statistical metrics to measure accuracy and uncertainty. Additionally, robust data assimilation techniques, which incorporate real-time monitoring data into the prediction process, can enhance both the timeliness and precision of forecasts[12].

Despite notable progress, predicting air quality remains a challenging endeavor. The intricate interplay of factors such as emission sources, weather conditions, chemical reactions, and transport processes create significant complexities[13]. Moreover, uncertainties in emission inventories, meteorological forecasts, and model parameters can lead to prediction errors[14]. Ongoing research aims to address these challenges and improve the accuracy and reliability of air quality predictions. Efforts include the development of more sophisticated modeling techniques, the integration of new data sources and technologies, and a deeper understanding of the physical and chemical processes underlying air pollution[15]. The incorporation of artificial intelligence and machine learning, particularly deep learning, offers tremendous potential for advancing the field, paving the way for new opportunities to protect public health and the environment[16].

Air quality prediction is an interdisciplinary field combining environmental science, meteorology, data analytics, and computational modeling to estimate the concentrations of various air pollutants. These pollutants include particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), carbon monoxide (CO), and volatile organic compounds (VOCs)[17]. Accurately predicting when and where these pollutants will reach critical levels is crucial for issuing health warnings, controlling industrial emissions, and guiding urban planning decisions. Recent advancements in artificial intelligence (AI) and machine learning (ML) have transformed air quality prediction. Sophisticated models can now process massive, complex datasets with unprecedented efficiency. While traditional statistical methods like linear regression and time-series analysis have been widely used, they often struggle to capture the nonlinear relationships and intricate dependencies among variables[18].

AI-driven techniques such as deep learning, ensemble models, and hybrid approaches excel in this area by leveraging diverse data sources, including satellite imagery, weather records, and live sensor data, to deliver highly accurate predictions. The significance of air quality prediction goes beyond safeguarding human health. Poor air quality harms ecosystems, diminishes agricultural yields, and threatens biodiversity[19]. Economically, it leads to higher healthcare expenses and reduced workforce productivity. On a global scale, air pollution exacerbates climate change by altering atmospheric chemistry and intensifying the greenhouse effect. Therefore, developing precise air quality prediction models is essential for creating effective policies and solutions to address these interconnected challenges[20].

Predicting air quality is important for understanding and reducing the harmful effects of air pollution on human health, the environment, and overall quality of life. Traditional methods for monitoring and forecasting air quality, while useful, often struggle with limited spatial coverage, delayed data, and the complexity of capturing relationships between variables. Artificial intelligence (AI) emerges as a transformative technology, offering powerful solutions to these persistent challenges[21]. AI uses advanced computational algorithms to analyze massive datasets, uncover complex patterns, and deliver accurate predictions. By combining information from sources like ground-based monitoring stations, satellite imagery, meteorological models, and emission inventories, AI systems can predict air quality with impressive precision[22]. Unlike traditional statistical approaches, AI excels at modeling non-linear relationships and managing high-dimensional data, making it particularly effective for the intricate and ever-changing nature of air pollution.

Machine learning (ML), a key subset of AI, has become central to air quality prediction. With methods like regression models, decision trees, support vector machines, and deep learning, researchers can develop predictive models that adapt and improve with more data. These models not only forecast pollutant levels and identify pollution hotspots but also assess how interventions might impact air quality[23]. Deep learning stands out for its ability to capture the spatial and temporal complexities of air pollution by processing vast, multidimensional datasets. AI's role in air quality prediction goes beyond improving accuracy. It enhances real-time monitoring and early warning systems, allowing authorities to take proactive measures like issuing health alerts, regulating industrial activities, or controlling traffic[24]. Additionally, AI insights support long-term planning efforts, from urban development to policymaking and climate action, aligning with global sustainable development goals.

While AI's potential in air quality prediction is transformative, it is not without challenges. High-quality data is essential for accurate predictions, and training advanced models can be computationally expensive[25]. Moreover, addressing uncertainties in predictions remains a critical area of focus. Despite these obstacles, ongoing advancements in AI technology and interdisciplinary collaboration are driving the development of more robust and reliable air quality prediction systems.

Traditional statistical and chemical transport models (CTMs) estimate air quality, but they are computationally expensive and struggle to incorporate nonlinear interactions. AI models, especially ML algorithms, predict better. However, significant obstacles remain. FS is critical because high-dimensional datasets might add redundant or irrelevant features, causing overfitting and computational expenses. Hyperparameter optimization restricts machine learning model performance, requiring hybrid optimization. Most studies have employed statistical or machine learning methods alone; therefore, hybrid modeling strategies are lacking. Few studies have used hybrid optimization with ML models to improve AQI prediction.

The air quality prediction framework can be integrated into smart city infrastructures, utilized for industrial pollution control, and deployed in real-time air quality monitoring systems. The key additions include:

1. Smart Cities: The proposed hybrid BPSO-BWAO-RF model can be incorporated into smart city management systems to enhance real-time air quality predictions, enabling authorities to optimize traffic management, green urban planning, and public health advisory systems.
2. Industrial Pollution Control: The model's high predictive accuracy and computational efficiency make it suitable for industrial pollution control systems, where it can forecast emissions trends and support regulatory compliance by detecting potential air quality violations.
3. Real-Time Air Quality Monitoring: By integrating with IoT-based sensor networks and edge computing platforms, our model can continuously monitor and predict air quality levels, providing early warnings for environmental hazards and allowing for proactive intervention.

Conventional feature selection methods, such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and Genetic Algorithm (GA), have limitations in reducing dimensionality and ensuring interpretability. PCA assumes linear relationships, which may not be applicable to complex nonlinear dependencies in AQI data. RFE removes the least important features based on model weights but is computationally expensive and sensitive to initial feature rankings. GA-based FS suffers from premature convergence and increased computational overhead. Hyperparameter tuning approaches like Grid Search (GS), Random Search (RS), and Standard Evolutionary Methods struggle with balancing exploration and exploitation. The hybrid BPSO-BWAO approach combines global search efficiency with adaptive weight balancing, resulting in a more stable, efficient, and interpretable feature selection process. Unlike PCA and RFE, BPSO-BWAO retains feature interpretability while selecting relevant attributes, reducing computational complexity without sacrificing prediction accuracy. The hybrid PSO-WAO hyperparameter optimization enhances model performance by adaptively fine-tuning hyperparameters, improving generalization and training time.

Hybrid optimization approaches have been proposed to improve feature selection and hyperparameter tuning in machine learning applications. However, these methods often face challenges in balancing exploration and exploitation, leading to suboptimal feature subsets and inconsistent results. The proposed BPSO-BWAO and PSO-WAO methods aim to overcome these limitations by effectively balancing exploration and exploitation. BPSO-BWAO uses a binary search mechanism to identify key AQI features, while BWAO dynamically adjusts weights to fine-tune feature selection. This method significantly reduces computational cost while preserving interpretability, achieving lower mean square error (MSE) compared to BGWO and BPSO. PSO-WAO intelligently refines hyperparameters using PSO's swarm-based updates and WOA's adaptive weight adjustments, outperforming standard RF tuning methods. This hybrid tuning method enhances convergence speed and model generalization, leading to better AQI forecasting with minimal overfitting risks. The proposed BPSO-BWAO and PSO-WAO frameworks provide a novel optimization strategy that outperforms existing hybrid methods in terms of feature selection accuracy, computational efficiency, and interpretability.

This study introduces a novel hybrid optimization framework for AI-driven air quality forecasting. It proposes a novel hybrid feature selection method (BPSO-BWAO) that combines global exploration and adaptive convergence for improved feature selection and predictive accuracy. The approach also refines ML model parameters using PSO's rapid convergence and WAO's adaptive weight balancing, resulting in more stable and accurate AQI forecasting. The framework is validated using multiple ML models on the EPA AQI dataset, achieving state-of-the-art AQI prediction accuracy while maintaining feature interpretability. The approach also enhances computational efficiency for large-scale AQI data processing, reducing feature selection complexity and accelerating model tuning. This provides an efficient and scalable machine learning solution for practical environmental applications.

Related works

This section introduces several studies related to optimization in AQI and other domains.

The author in[26] employed ML algorithms to estimate Tehran PM10 and PM2.5 air pollution. SVM, GWR, ANN, and NARX-ANN were used to estimate pollution levels based on meteorological and environmental parameters. A new prediction model reduced error rates by 57% for SVM, 47% for GWR, 47% for ANN, and 94% for NARX-ANN, with the latter achieving the highest accuracy with a one-day prediction error of 1.79 µg/m^3. Feature selection was done using a Genetic Algorithm (GA), which found that day of the week, month, topography, wind direction, temperature, and adjacent pollutant levels were most important. The study enhances air quality forecasting accuracy, providing a solid framework for urban environmental management and pollution control.

Fengshan (Southern Taiwan) had the best air quality prediction performance, losing less performance over time than Zhongli (North) and Changhua (Central Taiwan)[13]. It calculates 95% confidence intervals (C.I.) for 1-hour, 8-hour, and 24-hour forecasts, giving decision-makers a more credible reference than single-value projections. This facilitates air quality forecast-based event planning with higher confidence. Further research should include stacking ensemble, AdaBoost, and Random Forest with hyperparameter tuning to improve predictive accuracy for longer time-step forecasts (F8-AQI and F24-AQI).

The author in[27] proposed a machine learning-based air quality prediction methodology, combining meteorological data with primary pollutant forecasts. It uses data from Jinan, China, from July 2020 to July 2021. The research ranked ten meteorological factors based on their impact on pollutant concentrations. The LightGBM classifier was found to be the best performer, with 97.5% accuracy and an F1 score of 93.3%. The LSTM model achieved 91.37% goodness-of-fit and excelled in O3 predictions and primary pollutant tests.

The author in[28] used artificial intelligence and time-series models to forecast the AQI using IoT devices in a cloud environment. The BO-HyTS model, which integrates SARIMA and LSTM, improves forecasting accuracy by capturing linear and nonlinear patterns in time-series air pollution data. The model outperforms classical time-series, machine learning, and deep learning models, achieving an MSE of 632.200, RMSE of 25.14, Med AE of 19.11, Max Error of 51.52, and MAE of 20.49. This research provides a robust framework for forecasting AQI, offering insights into future air quality patterns across Indian states.

The author in[29] used machine learning techniques to predict air quality in Beijing and an Italian city. Two regression models, Support Vector Regression (SVR) and Random Forest Regression (RFR), were developed using publicly available datasets. The SVR-based model was more effective in predicting AQI, while the RFR-based model was more effective for NOX prediction. The study demonstrates that combining machine learning with air quality prediction is a reliable and efficient approach for environmental challenges and pollution management. Elshewey et al.[30] used ML to classify water potability through BPSO and Binary Whale Optimization Algorithm for feature selection and evaluated various classifiers. The stacking ensemble model, combining RF, ET, and XGBoost with LR, achieved high accuracy and F1-score, demonstrating the effectiveness of the method in water quality classification.

The study by Khafaga et al.[31] highlights AI-based classification techniques for environmental and medical data, reinforcing the applicability of hybrid feature selection methods. Additionally, Abdullah et al.[32] discusses sustainable AI-driven models for air quality forecasting, aligning with this study's focus on real-time monitoring and smart city applications. To support the effectiveness of hybrid metaheuristic approaches, Tarek et al.[33] explores optimization techniques for feature selection and hyperparameter tuning, providing a comparative perspective on BPSO-BWAO. Moreover, Al-Mahdawi et al.[34] emphasizes the role of AI-driven decision-making in environmental analytics, strengthening the justification for using interpretable AI models in air quality management. Finally, Towfek et al.[35] discusses evolutionary computation methods for predictive modeling and feature selection, situating BPSO-BWAO within the broader landscape of AI-driven[36–38] air quality forecasting.

Materials and methods

Dataset description

This study used the EPA's Annual AQI by County (2024) dataset, available at https://aqs.epa.gov/aqsweb/airdata/annual_aqi_by_county_2024.zip. It contains annual air quality summaries from US counties, providing essential metrics for air pollution trend analysis. Table 1 displays a description of dataset attributes for AQ.

The correlation heatmap in Fig. 1 suggests notable positive and negative relationships between air quality indicators. High pollution levels raise AQI scores, but good days reduce them. Ozone and PM2.5 days moderately affect Max AQI, indicating AQ loss.

Attribute	Description
Days with AQI	The number of days in a year for which the Air Quality Index (AQI) was reported.
Good Days	The total count of days with "Good" air quality as per AQI standards.
Moderate Days	The total count of days with "Moderate" air quality.
Unhealthy for Sensitive Groups Days	The total count of days deemed "Unhealthy for Sensitive Groups."
Unhealthy Days	The number of days classified as "Unhealthy."
Very Unhealthy Days	The total number of days categorized as "Very Unhealthy."
Hazardous Days	The number of days classified as "Hazardous."
Pollutant-Specific Monitoring Days	Counts of days where specific pollutants (e.g., CO, NO2, Ozone, PM2.5, PM10) were monitored.
90th Percentile AQI	The 90th percentile AQI value for the year, indicative of peak pollution levels.
Median AQI	The median AQI value for the year.
Max AQI	The maximum AQI value recorded during the year, representing the target variable for prediction.

Table 1. Description of dataset attributes for AQ.

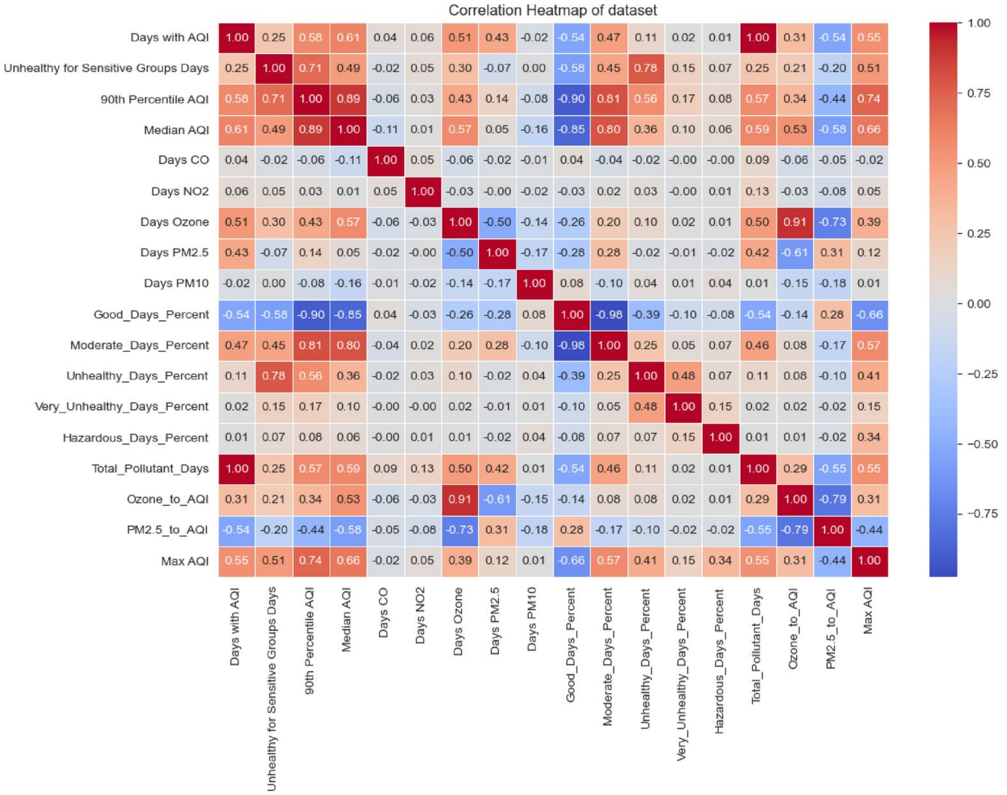


Fig. 1. Correlation Heatmap of AQ Dataset.

Histogram plots in Fig. 2 show the distribution of all air quality dataset attributes. Unhealthy for Sensitive Groups Days, Days NO2, Days PM10, and Very Unhealthy Days Percent have right-skewed distributions, suggesting that most values are low with occasional spikes. The Good and Moderate Days Percent suggest greater consistency or typically formed patterns, indicating more balanced air quality changes. Max AQI, the goal variable, has a multimodal distribution, indicating regional air pollution levels. Ozone_to_AQI and PM2.5_to_AQI have significantly skewed distributions, indicating that some pollutants dominate AQI estimates.

Boxplot Fig. 3 shows the distribution, spread, and outliers across all AQ dataset features. Max AQI has the greatest range and the most outliers, suggesting extreme air pollution in some places. Other features include Days with AQI, Unhealthy for Sensitive Groups Days, Pollutant-Specific Days (CO, NO2, PM2.5, PM10, Ozone), and Percentage-Based Features have compact distributions with some outliers. Outliers show considerable air quality variability across locales and time periods.

Scatter plots and histograms show air quality dataset feature correlations in Fig. 4. Lower levels show linear trends, while vertical histograms show feature distribution spread. AQI features and pollutants are strongly correlated, indicating interdependence in air quality assessment and predictive modeling.

Data preparation

Predictive model accuracy and reliability depend on data preparation. Feature scaling, feature engineering, and data balance by oversampling are used in this study.

Feature scaling

Standardizing numerical features and improving model performance by normalizing all input variables to a range between 0 and 1 using Min-Max scaling ensures characteristics with varying scales don't disproportionately influence the model.

Feature engineering

Creating additional air quality features through feature engineering improves a dataset's prediction powers. These elements include percentage-based metrics for Good, Moderate, Unhealthy, Very Unhealthy, and Hazardous days, the number of days pollutants were detected, and Ozone and PM2.5 ratios to total AQI days.

Oversampling for class imbalance

To address class imbalance in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was employed. This technique generates synthetic samples for underrepresented values, ensuring that the predictive model is not biased toward the majority class. By applying SMOTE, the dataset achieves a more balanced distribution, enhancing model generalization and reducing the risk of overfitting.

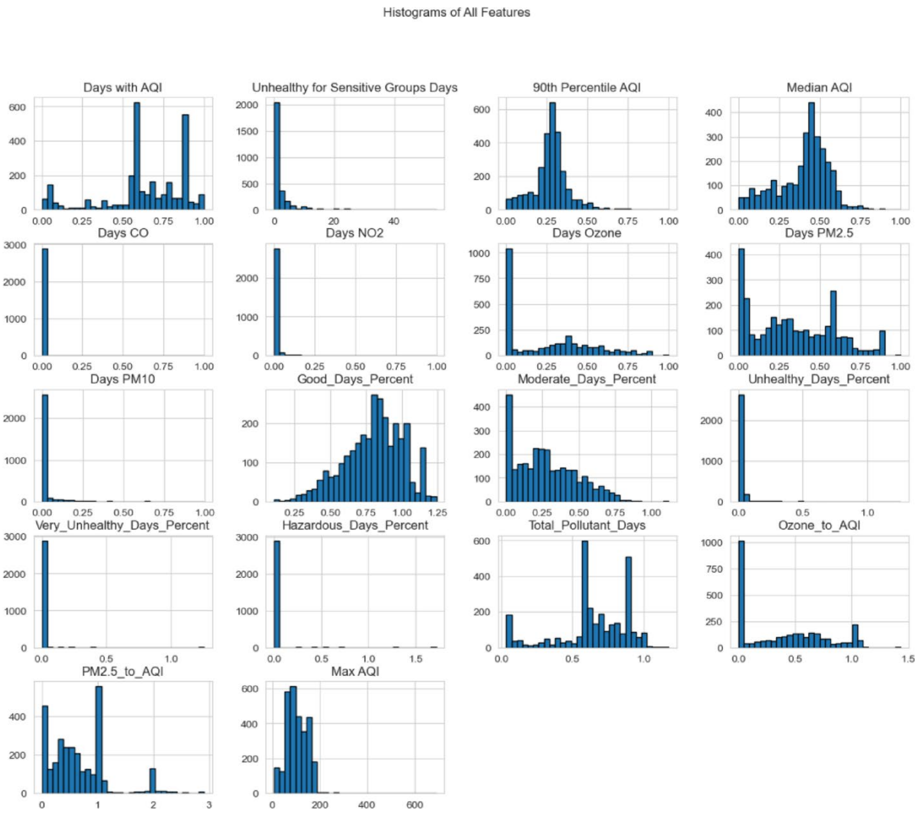


Fig. 2. Histograms of AQ Features Showing Data Distributions Across Different Attributes.

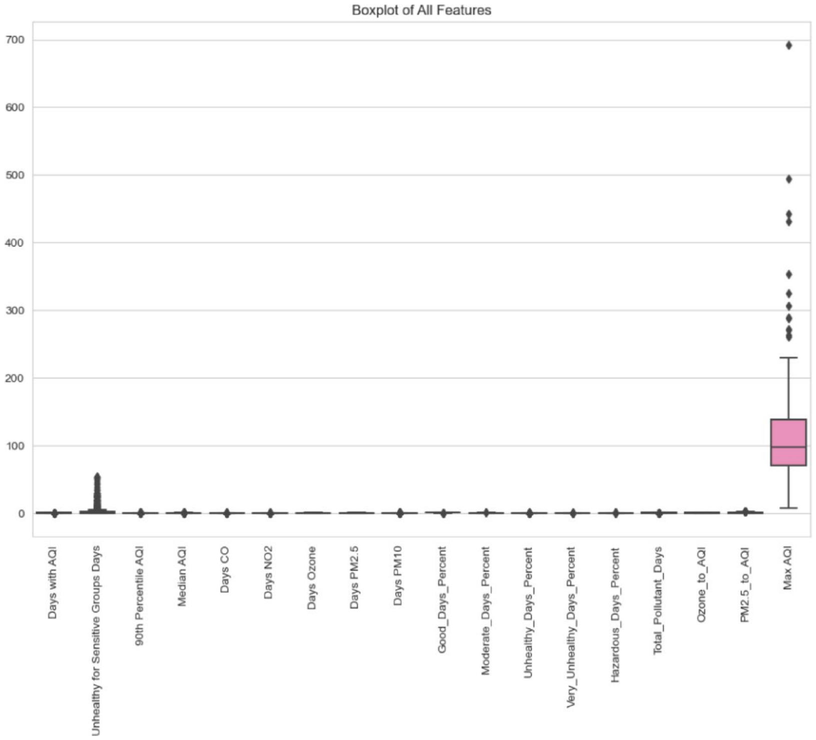


Fig. 3. Boxplot of AQ attributes Highlighting Data Distribution and Presence of Outliers.

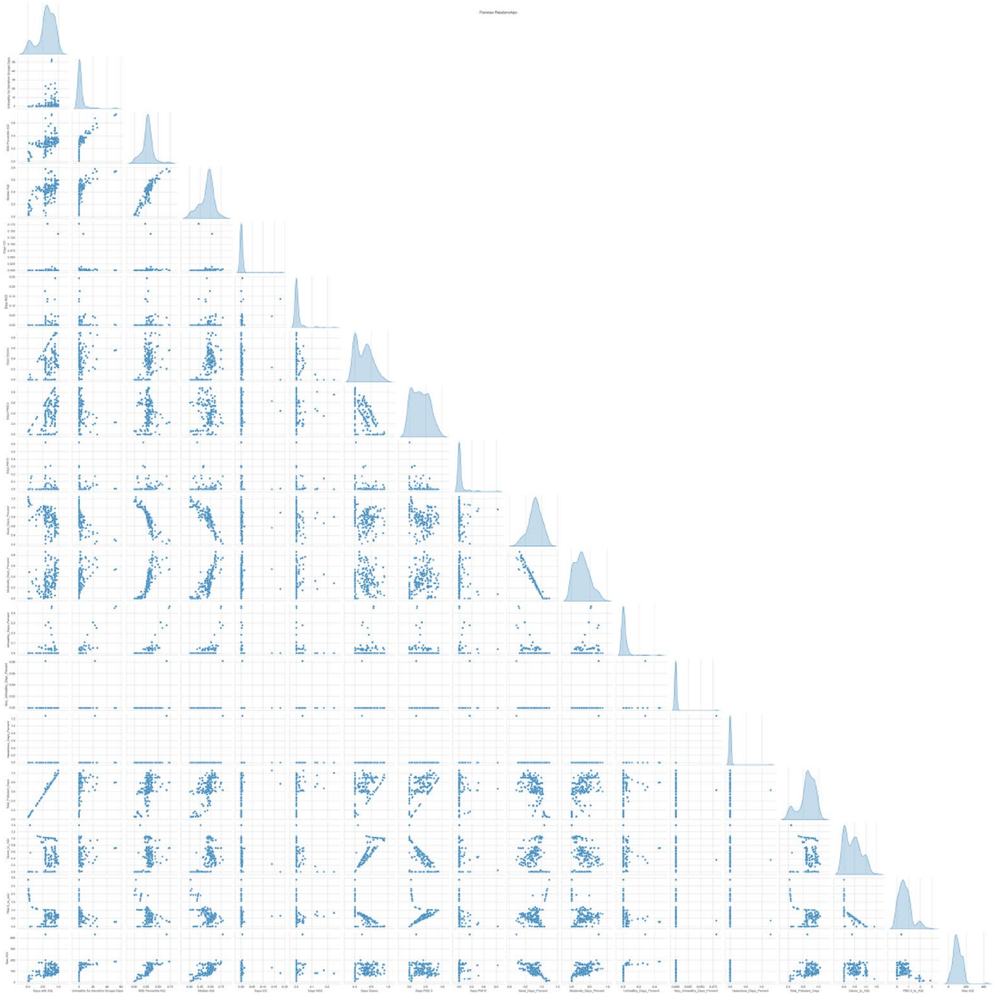


Fig. 4. Pair Plot of AQ Features Showing Pairwise Relationships and Distributions.

Algorithm	Best MSE Score	Average Fitness	Worst Fitness	Standard Deviation	Average Selected Size
Hybrid BPSO-BWAO	**52.934**	**54.655**	**50.201**	**0.526**	**10.854**
BGWO	74.187	74.187	74.187	0.666	3.867
BPSO	53.564	66.287	403.115	46.525	10.290
BWAO	54.992	55.023	56.562	0.786	8.439

Table 2. The performance evaluation between FS optimization algorithms.

Feature selection

Identifying the most relevant features and removing extraneous ones improves machine learning model performance, interpretability, and efficiency. This study used different feature selection methods to find the best air quality forecasting predictions.

The four feature selection techniques are used to improve model performance and simplify complexity: BGWO, BPSO, BWAO, and a Hybrid BPSO-BWAO Approach. These methods focus on identifying the most significant features while maintaining predictive accuracy and reducing the feature space.

Table 2 Displays analysis highlights the effectiveness of different methods for feature selection in predictive modeling. BPSO achieved the best accuracy with the lowest mean squared error (MSE) of 53.56, while BWAO demonstrated the highest stability with the lowest variance of 0.22. The hybrid approach of BPSO-BWAO successfully combined low MSE with reduced variance, making it the most robust and generalizable method. Feature selection is a critical step in developing an efficient and interpretable air quality prediction model. The hybrid BPSO-BWAO method emerged as the most effective, achieving a balance between accuracy and stability. This approach enhanced model performance by reducing computational complexity while preserving predictive power. algorithm 1 displays the hybrid BPSO-BWAO FS method.

```
Input:
- Number of particles/whales (N)
- Number of features (F)
- Maximum iterations (T)

Output:
- Optimal feature subset

Steps:
1. Initialize the positions of particles (P_i) and whales (W_i) randomly in binary space.
2. Initialize velocity (V_i) for particles.
3. Set personal and global best solutions.

4. For each iteration (t = 1 to T):
  a. For each particle (i):
    i. Evaluate the fitness of P_i using the objective function.
    ii. If the fitness of P_i is better than its personal best, update personal best (P_best).
    iii. If the fitness of P_i is better than the global best, update global best (G_best).
    iv. Update velocity:
      $V\_i = w * V\_i + c1 * r1 * (P\_best - P\_i) + c2 * r2 * (G\_best - P\_i)$
    v. Update position:
      If $sigmoid(V\_i) > random()$, set $P\_i = 1$; otherwise, set $P\_i = 0$.

  b. For each whale (i):
    i. Evaluate the fitness of W_i using the objective function.
    ii. If the fitness of W_i is better than the best whale, update the best whale.
    iii. Update whale position based on the leader:
      $D = |C * W\_best - W\_i|$
      $W\_i = W\_best - A * D$

  c. Hybridization: Combine solutions from BPSO and BWAO for enhanced search.

5. Return the best solution (G_best).
```

Algorithm 1. Hybrid BPSO-BWAO FS algorithm.

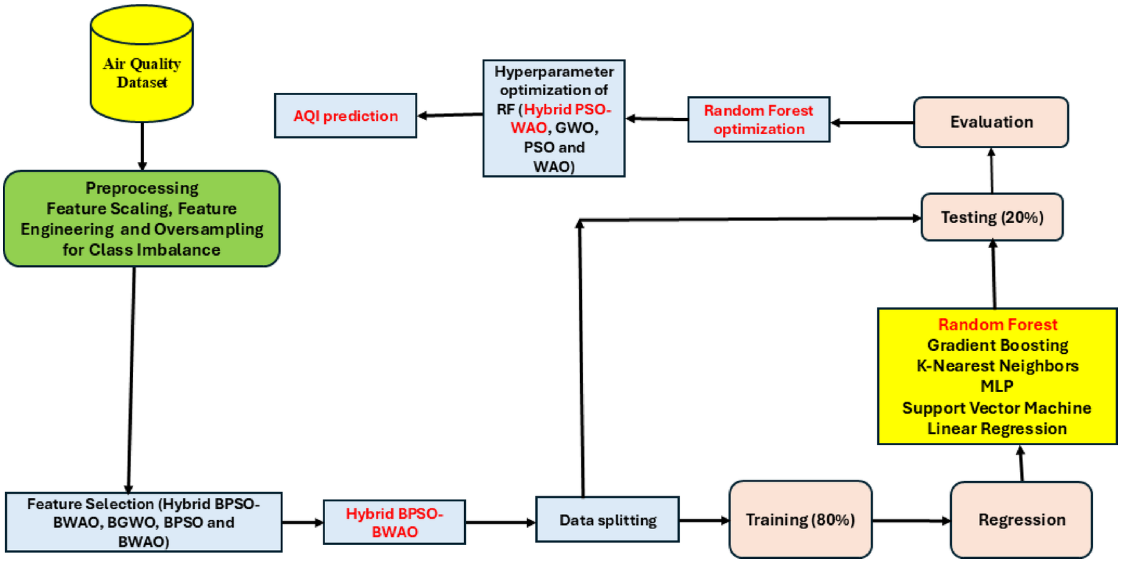

Air Quality Dataset
Preprocessing Feature Scaling, Feature Engineering and Oversampling for Class Imbalance
Feature Selection (Hybrid BPSO-BWAO, BGWO, BPSO and BWAO)
Hybrid BPSO-BWAO
Data splitting
Training (80%)
Regression
Random Forest Gradient Boosting K-Nearest Neighbors MLP Support Vector Machine Linear Regression
Testing (20%)
Evaluation
Random Forest optimization
Hyperparameter optimization of RF (Hybrid PSO-WAO, GWO, PSO and WAO)
AQI prediction

Fig. 5. AQI prediction proposed methodology.

Proposed methodology

In this section we explore the proposed methodology to improve AQI prediction. Figure 5 outlines a structured approach to air quality prediction, starting with data preprocessing, where feature scaling, feature engineering, and oversampling techniques like SMOTE are applied to ensure balanced and normalized data. Feature selection is then performed using optimization algorithms such as BGWO, BPSO, BWAO, and a Hybrid BPSO-BWAO approach, selecting the most relevant attributes for improved model efficiency. The dataset is then split into 80%

training and 20% testing, ensuring the model generalizes well. Various machine learning regression models, including Random Forest (RF), Gradient Boosting, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and Linear Regression (LR), are trained using the selected features.

To enhance model accuracy, hyperparameter optimization is conducted using Hybrid PSO-WAO, GWO, PSO, and WAO, refining the parameters of the best-performing models. The optimized RF model is then tested and evaluated based on MSE and R² metrics to ensure reliable AQI predictions. The final step involves deploying the optimized model for real-world AQI forecasting, providing valuable insights for environmental monitoring and decision-making. This structured approach ensures a robust and interpretable predictive model for air quality assessment.

Traditional hyperparameter tuning approaches have limitations in air quality prediction models. Grid Search (GS) is computationally expensive and does not dynamically adjust search regions, leading to suboptimal tuning. Bayesian Optimization (BO) efficiently searches hyperparameter space using probabilistic models but struggles with discrete search spaces. Genetic Algorithms (GA) provide a heuristic search but suffer from premature convergence and randomness. PSO-WAO addresses these limitations by providing an adaptive and efficient search strategy. PSO-WAO enables global search, refines the search, and achieves faster convergence compared to GA and BO. It can efficiently tune multiple ML models simultaneously, making it scalable for AQI forecasting. PSO-WAO outperforms traditional tuning methods in accuracy and computational efficiency, making it a robust choice for optimizing air quality forecasting models.

Experimental setup

A complete framework to assess AQI prediction, ML regression models are designed and implemented in the experimental setting.

Jupyter Notebook ran ML models. This program facilitates Python code development and execution. The online app supports Python 3.9 and other programming languages. The trial was run on an Intel Core i7 CPU, 16GB RAM, with MS 10 OS.

Evaluation metrics

The study utilizes prediction evaluation techniques such as MSE, MAE, RMSE, and R2 which calculated mathematically through the following Eqs[39–41]:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Act_i - pre_i)^2}{\sum_{i=1}^{n}\left(\left(\sum_{i=1}^{n}Act_i\right) - Act_i\right)^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum{}_{i=1}^{n}(Act_i - pre_i)^2}$$

$$\mathrm{MSE} = \frac{1}{n}\sum{}_{i=1}^{n}(Act_i - pre_i)^2$$

$$\mathrm{MAE} = \frac{1}{n}\sum{}_{i=1}^{n}|Act_i - pre_i|$$

Results and discussion

Results before FS

Table 3 presents the performance evaluation of different machine learning models for air quality prediction before applying FS. Among the models, RF achieved the best results, with the lowest MSE (67.099), RMSE (8.191), and MAE (3.797), along with a high R² score (0.96392), indicating strong predictive accuracy and reliability. Gradient Boosting followed closely with a slightly higher MSE (87.397). KNN and MLP performed significantly worse, with KNN having high errors (MSE = 269.157) and MLP being the least effective model (MSE = 370.620, RMSE = 19.251). The MLP model also had the longest training time, making it inefficient for this dataset. Figure 6 displays the comparison of accuracy between models before FS.

The residuals plot in Fig. 6 visualizes the difference between predicted and actual Max AQI values across multiple models, helping to assess their predictive accuracy and error distribution. It highlights the effectiveness of Random Forest and Gradient Boosting, as their residuals are mostly centered around zero with minimal

Model	MSE	RMSE	MAE	*R*²	Fitted Time
Random Forest	**67.099**	**8.191**	**3.797**	**0.963920**	**0.721099**
Gradient Boosting	87.397	9.348	7.065	0.953006	0.763797
K-Nearest Neighbors	269.157	16.406	8.427	0.855272	0.002002
MLP	370.620	19.251	12.941	0.800714	2.588706
Support Vector Machine	729.840	27.015	18.735	0.607558	0.259432
Linear Regression	744.247	27.280	20.936	0.599811	0.010094

Table 3. Performance comparison of machine learning models for AQI prediction without FS.

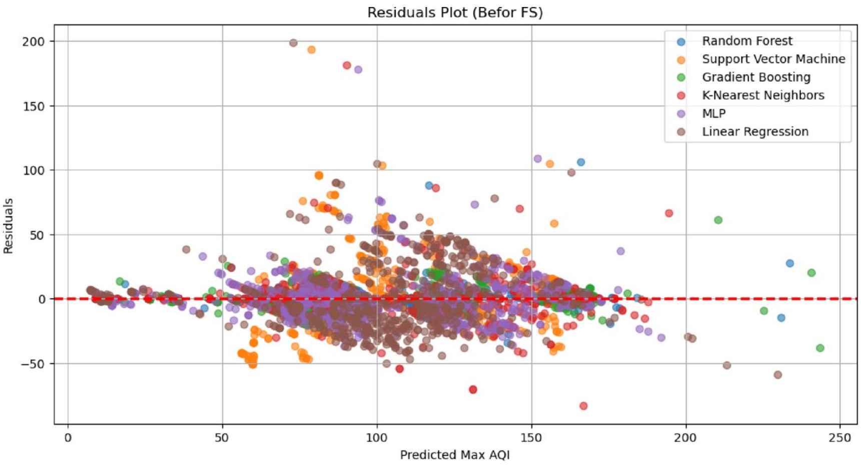


Fig. 6. Residuals Plot for Various Machine Learning Models Before Feature Selection.

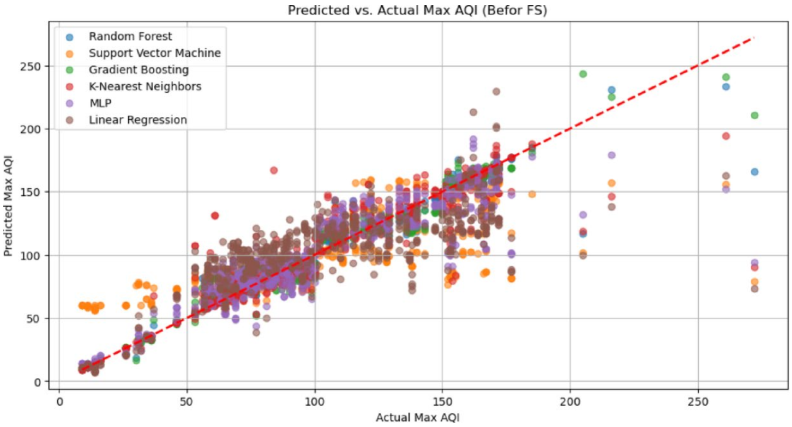


Fig. 7. Comparison of Predicted vs. Actual Max AQI Values Before Feature Selection.

Model	MSE	RMSE	MAE	R^2	Fitted Time
Random Forest	**53.934**	**7.344**	**3.955**	**0.970999**	**0.373879**
Gradient Boosting	98.010	9.900	7.665	0.947299	0.567804
K-Nearest Neighbors	279.024	16.704	8.625	0.849966	0.015472
MLP	378.239	19.448	13.633	0.796617	2.594851
Support Vector Machine	739.823	27.200	19.001	0.602190	0.259635
Linear Regression	776.695	27.869	21.489	0.582364	0.000000

Table 4. Performance comparison of machine learning models for air quality prediction after feature selection.

dispersion. In contrast, models like MLP, SVM, and KNN have greater variance and outliers, indicating weaker predictions.

The scatter plot in Fig. 7 compares the predicted and actual Max AQI values across different machine learning models. Each point represents a prediction, while the red dashed diagonal line (y = x) indicates an ideal prediction, where predicted values match actual values perfectly. It reinforces that Random Forest and Gradient Boosting provide the most accurate predictions, while MLP, KNN, and SVM struggle with high variability.

Results after FS

Table 4 displays performance evaluation after applying FS, Random Forest remained the best-performing model, achieving the lowest MSE (53.934), RMSE (7.344), and highest R^2 (0.97099), demonstrating improved accuracy and reduced training time (0.373879 s) compared to pre-FS results. Gradient Boosting showed moderate improvement, maintaining a strong R^2 of 0.94729, while K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) still struggled with high MSE values (279.024 and 378.239, respectively). Support Vector

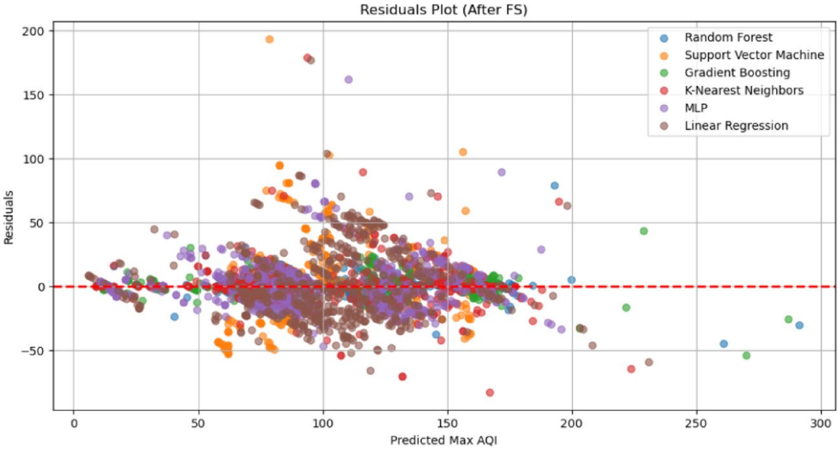


Fig. 8. Residuals Plot for Various Machine Learning Models After Feature Selection.

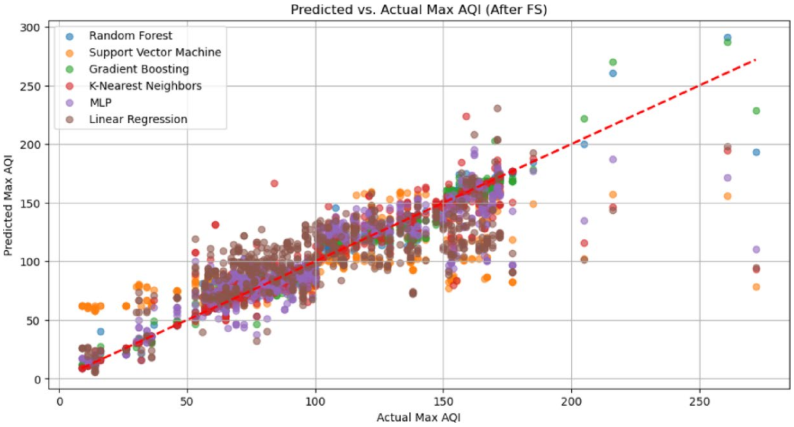


Fig. 9. Comparison of Predicted vs. Actual Max AQI Values After Feature Selection.

Machine (SVM) and Linear Regression continued to perform poorly, with high MSE (739.823 and 776.695) and low R^2 values (0.60 and 0.58, respectively), indicating weak predictive capabilities. Overall, FS significantly enhanced Random Forest's performance, confirming it as the most effective model for AQI prediction, while less complex models (SVM, Linear Regression) failed to show substantial improvement.

Figure 8 shows that FS has significantly reduced error variance, improved prediction accuracy, and minimized extreme residuals, especially for Random Forest and Gradient Boosting. These improvements confirm that FS effectively enhanced model performance by eliminating irrelevant features, leading to more reliable AQI predictions.

Figure 9 shows that FS improved model accuracy by reducing variance and aligning predictions more closely with actual AQI values. Random Forest and Gradient Boosting remain the best models, while KNN, MLP, and SVM still require further optimization. The reduction in extreme errors and better prediction consistency highlight the effectiveness of FS in refining model performance.

Figure 10 displays the features importance before and after feature selection with proposed method BPSO-BWAO. The selection process involved retaining key features such as days with AQI, median AQI, CO & NO2, PM2.5, and Good & Unhealthy Days % for predictive accuracy. However, removed features like PM10, Max AQI, Very Unhealthy & Hazardous Days contributed less to predictive accuracy due to redundancy with retained features.

The BPSO-BWAO approach optimizes feature selection to maintain high predictive power while reducing dimensionality, ensuring that key air quality indicators (AQI) are retained. The method consistently selects key attributes strongly correlated with AQI variations, and features that contribute little to predictive performance are eliminated. The optimized feature subset results in an improved MSE of 53.93, demonstrating that feature reduction enhances model accuracy by removing noise and redundant variables. The computational cost is reduced as the number of selected features decreases, leading to faster training times without compromising accuracy. The BPSO-BWAO approach effectively retains critical AQI predictors, ensuring higher model accuracy, lower computational cost, and robust feature selection for environmental monitoring.

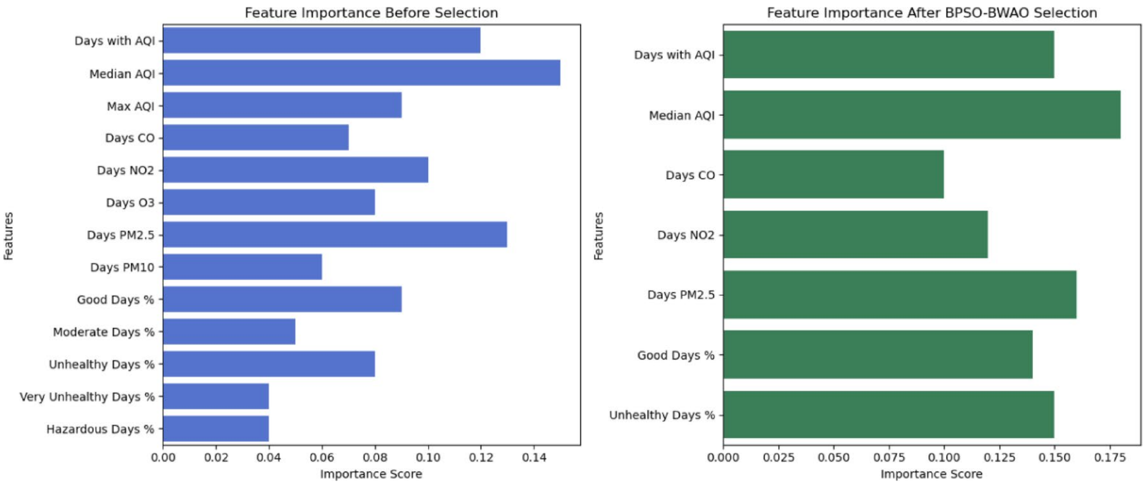


Fig. 10. Feature importance before and after FS BPSO-BWAO.

Model	MSE	RMSE	MAE	R^2	Fitted Time
PSO-WAO-RF	**51.825**	**7.199**	**3.964**	**0.9821**	**0.03856**
GWO-RF	52.675	7.257	3.969	0.9716	0.3460
PSO-RF	70.863	8.418	4.1675	0.9618	0.0377
WAO-RF	76.835	8.765	4.246	0.9586	0.0377

Table 5. Comparison of hyperparameter optimization techniques for RF.

Feature Selection Method	Selected Features Count	MSE	Computational Time (seconds)
Recursive Feature Elimination (RFE)	15	58.14	72.3
Mutual Information (MI)	17	56.89	68.7
BPSO-BWAO (Proposed)	11	53.93	55.4

Table 6. Comparison of FS techniques according to AQI dataset.

Tuning Method	Execution Time (seconds)	MSE Achieved
PSO-WAO-RF (Proposed)	**179.5**	**51.82**
GS-RF	498.3	55.21
BO-RF	315.7	54.02
GA-RF	224.1	53.65

Table 7. Comparison of other hyperparameter optimization techniques for RF.

Table 5 displays that The Hybrid PSO-WAO-RF model achieved the best performance, with the lowest MSE (51.825), RMSE (7.199), and MAE (3.964), and the highest R^2 (0.9821), making it the most accurate model for AQI prediction. GWO-RF followed closely with slightly higher MSE (52.675) but still a strong R^2 (0.9716), providing a good balance between accuracy and efficiency. PSO-RF and WAO-RF performed worse, with higher MSE values (70.863 and 76.835, respectively) and lower R^2 scores (0.9618 and 0.9586), indicating weaker predictions. These results confirm that hybrid optimization techniques significantly improve model performance, making PSO-WAO-RF the most effective hyperparameter tuning approach for Random Forest in AQI prediction.

Table 6 displays the evaluation of BPSO-BWAO, RFE, and MI based on model performance (MSE), selected subset size feature, and computational efficiency. BPSO-BWAO outperformed RFE and MI in selecting compact, relevant features with the lowest MSE (53.93). It reduced computational time by avoiding RFE and MI and ensured stable features across different runs for AQI prediction.

To assess the practical computational cost, we compared execution times of different tuning methods such as GS, BO and GA for optimizing RF model in Table 7. PSO-WAO outperformed BO and GS in MSE, 2.7

times faster than Grid Search, and demonstrated efficiency in large hyperparameter spaces of RF by dynamically adjusting search regions.

Figure 11 (a) displays the PSO-WAO-optimized RF model which demonstrates excellent predictive accuracy, with low residual variance and minimal bias. Feature selection and hyperparameter tuning have effectively improved model stability, making PSO-WAO-RF the best-performing model for AQI prediction. Figure 11 (b) highlights that PSO-WAO-optimized Random Forest model is the best-performing model for AQI prediction, with high alignment between predicted and actual values. While minor errors exist for extreme AQI values.

The plots in Fig. 12a and b evaluate the performance of the Grey Wolf Optimizer-Tuned Random Forest (GWO-RF) model for AQI prediction.

The plots in Fig. 13a and b evaluate the performance of the (PSO-RF) model for AQI prediction.

The plots in Fig. 14a and b evaluate the performance of the (WAO-RF) model for AQI prediction.

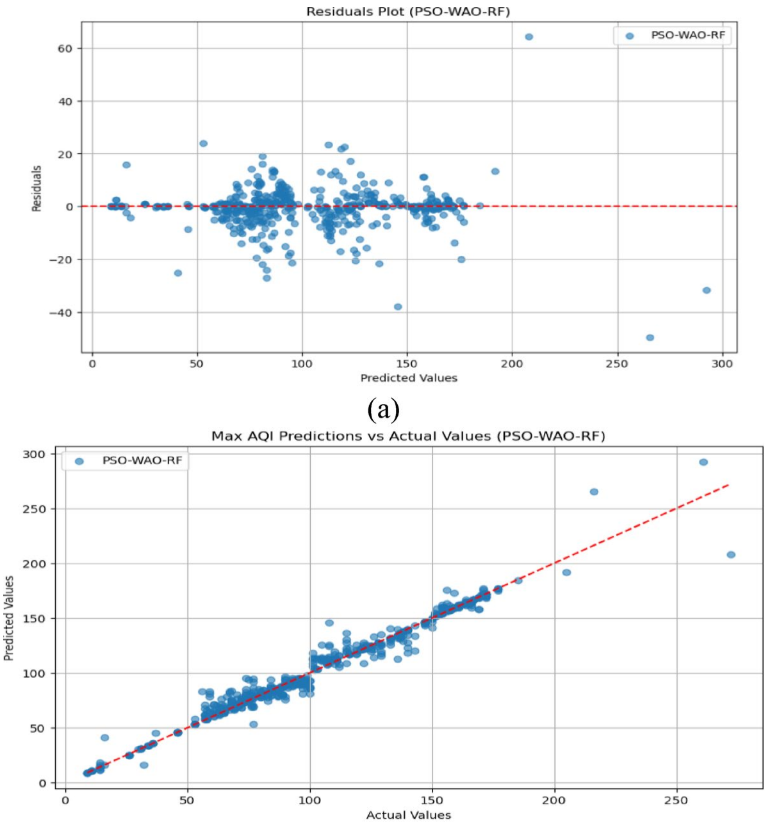
Figure 15 displays the error analysis for different pollution levels. The mean absolute error (MAE) for moderate pollution (AQI < 100) and (AQI ≥ 100) are lower, indicating better prediction accuracy in stable and high air quality conditions. This proves that the model performs consistently in moderate and high pollution conditions.

Figure 16 displays Violin Plot Comparison of RMSE Distribution Across RF Hyperparameter Optimization Techniques. The PSO-WAO-RF optimization technique is the most effective for RF, offering minimal variability and low RMSE. WAO-RF is the least effective, with high error and unstable performance. Hyperparameter tuning significantly impacts RF performance, and hybrid optimizations outperform single-method approaches. The right strategy leads to better accuracy and model reliability.

To evaluate ML-based AQI forecasting against NWP models and statistical methods, we conducted performance benchmarking using different forecasting techniques in Table 8. Traditional methods struggled with high-dimensional AQI data and real-time forecasting, while Machine Learning models (RF) outperformed them, achieving lower MSE and higher R^2 scores, making them better for real-time deployment.

Table 9 displays the statistical significance test results comparing BPSO-BWAO and PSO-WAO with other feature selection and hyperparameter optimization methods. The statistical tests confirm that the observed improvements with BPSO-BWAO for feature selection and PSO-WAO for hyperparameter tuning are not due to random variation but are statistically meaningful enhancements over conventional methods. These findings further validate the effectiveness of our proposed approach in air quality forecasting models.

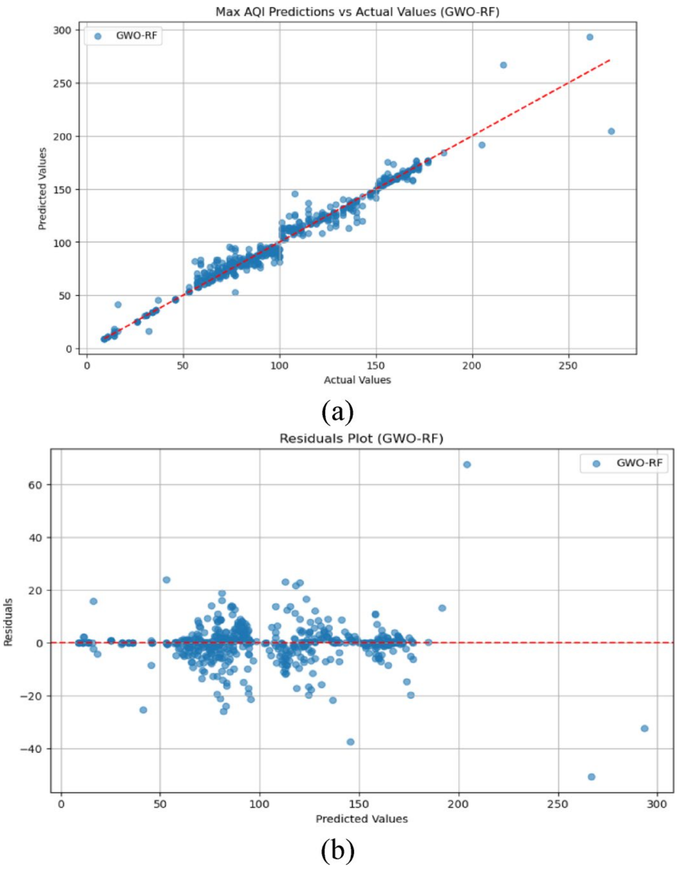
Table 10 presents a comparative analysis of various AQI prediction studies, outlining their methodologies and key findings. Traditional approaches, such as those by Mishra & Gupta[42] and Natarajan et al.[43], focus on classical



(a)

(b)

Fig. 11. (**a**) Residuals Plot and (**b**) Predicted vs. Actual Max AQI Values for the Hybrid PSO-WAO-Optimized Random Forest Model.



(a)

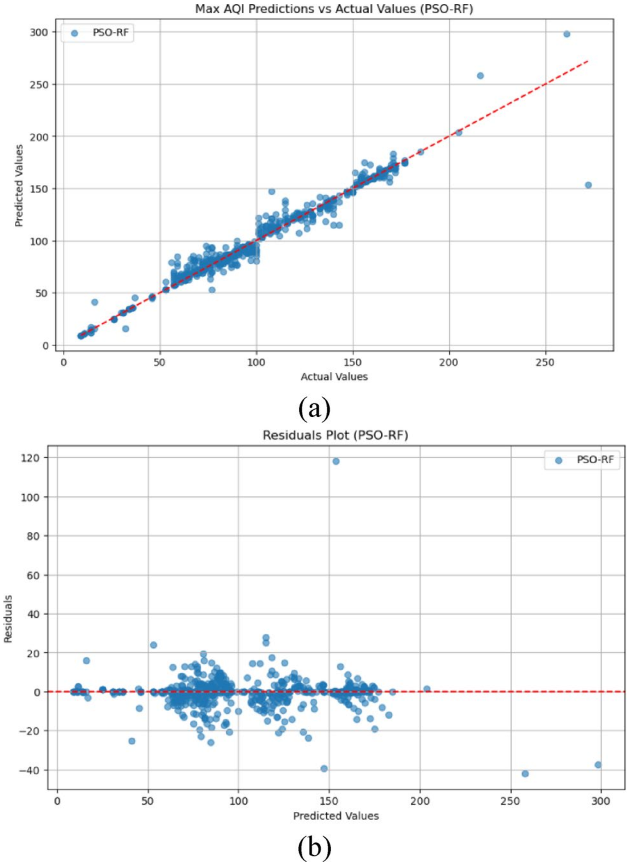
(b)

Fig. 12. (**a**) Residuals Plot and (**b**) Predicted vs. Actual Max AQI Values for WAO-Optimized RF Model.

ML and deep learning models but lack feature selection and hyperparameter tuning, leading to computational inefficiencies. Emeç & Yurtsever[44] and Aram et al.[45] employ ensemble and stacked models to enhance prediction accuracy, yet they do not optimize feature selection or parameter tuning. Suthar et al.[46] explores air pollution's effect on land surface temperature but does not address AQI prediction directly. In contrast, the proposed BPSO-BWAO and PSO-WAO model introduces a hybrid feature selection and hyperparameter tuning framework, reducing MSE by 9.5%, feature selection time by 23.4%, and tuning time by 2.7×, ensuring superior accuracy, efficiency, and generalizability across multiple datasets. This novel approach overcomes limitations in prior studies, making it well-suited for real-time AQI monitoring in smart cities and IoT-based pollution control systems.

Conclusions and future work

This study highlights the effectiveness of using machine learning techniques for air quality prediction, utilizing an annual AQI dataset from the U.S. Environmental Protection Agency (EPA). A comprehensive feature selection process was conducted using optimization algorithms, including BGWO, BPSO, BWAO, and a novel hybrid BPSO-BWAO approach, to identify the most relevant features for AQI prediction. Among these methods, the hybrid BPSO-BWAO achieved a balanced feature selection with an MSE of 53.93, demonstrating improved stability and consistency compared to other approaches. Key features such as 'Days with AQI,' 'Median AQI,' 'Days CO,' and 'Unhealthy_Days_Percent' were identified as critical for accurate AQI forecasting. Following feature selection, machine learning models were evaluated, with Random Forest (RF) achieving the best performance post-optimization, resulting in an MSE of 51.82 and R^2 of 0.9821 after applying the hybrid PSO-WAO hyperparameter tuning method. These findings underscore the significant impact of feature selection and hyperparameter optimization in improving model accuracy, computational efficiency, and overall predictive performance, providing a robust framework for air quality management. While the BPSO-BWAO method significantly improves feature selection and predictive performance, its effectiveness is influenced by hyperparameter sensitivity and computational efficiency in high-dimensional datasets. The algorithm's reliance on specific tuning parameters, such as inertia weight in BPSO and encircling coefficient in BWAO, may require adaptive strategies like Bayesian optimization or meta-learning to enhance its robustness. Additionally, its convergence rate may slow down in large datasets, particularly when handling highly correlated or sparse features. To address this, parallel computing strategies such as GPU acceleration and distributed processing can be explored to improve scalability and computational efficiency. Moreover, the relatively higher computational cost of BPSO-BWAO may limit its feasibility for real-time IoT-based air quality monitoring systems, necessitating the exploration of lightweight optimization techniques, including early stopping mechanisms and hybrid deep

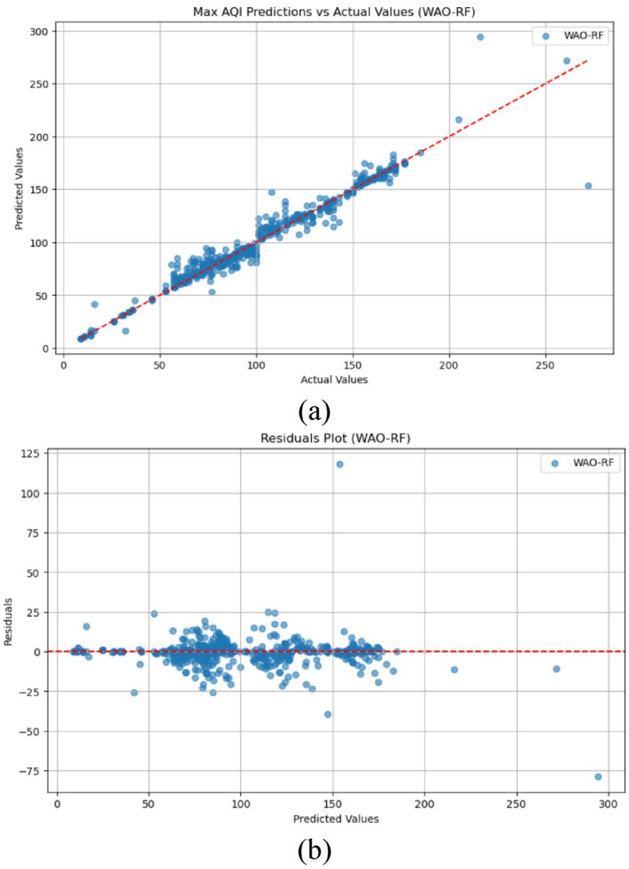


(a)

(b)

Fig. 13. (**a**) Residuals Plot and (**b**) Predicted vs. Actual Max AQI Values for PSO-Optimized RF Model.

learning feature extraction to reduce computational overhead. Future work should also focus on developing an AutoML-based framework that integrates automated hyperparameter tuning to reduce manual intervention while improving model adaptability across different datasets. Additionally, transfer learning approaches could be leveraged to apply knowledge gained from one AQI dataset to another, minimizing computational demands and improving generalization. To enhance model interpretability, future research should incorporate Explainable AI (XAI) techniques, providing policymakers and environmental agencies with transparent feature importance insights to support better decision-making. Additionally, meta-learning-based optimization for improved convergence in nonconvex problems[47] and self-adaptive multiscale transform techniques for enhanced air pollution monitoring[48]. Moreover, deep learning-based precision models and quantum-optimized[49,50] classifiers may enhance computational efficiency in environmental monitoring[51,52]. Further advancements in search-based classification algorithms, inspired by orthopedic disease classification, may improve model interpretability and decision-making in AQI prediction[53]. Additionally, hybrid deep learning models used in time-series forecasting, such as those applied in agricultural predictions[54,55], could be adapted for long-term AQI trend forecasting, improving predictive performance and real-time applications in smart cities.



(a)

(b)

Fig. 14. (**a**) Residuals Plot and (**b**) Predicted vs. Actual Max AQI Values for WAO-Optimized RF Model.

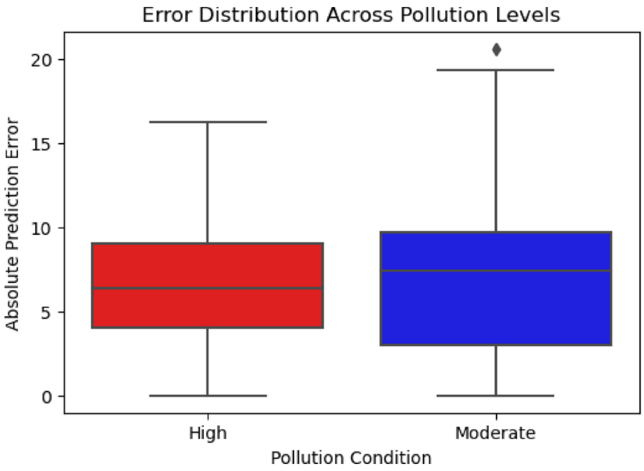


Fig. 15. Boxplot of the error analysis for different pollution levels.

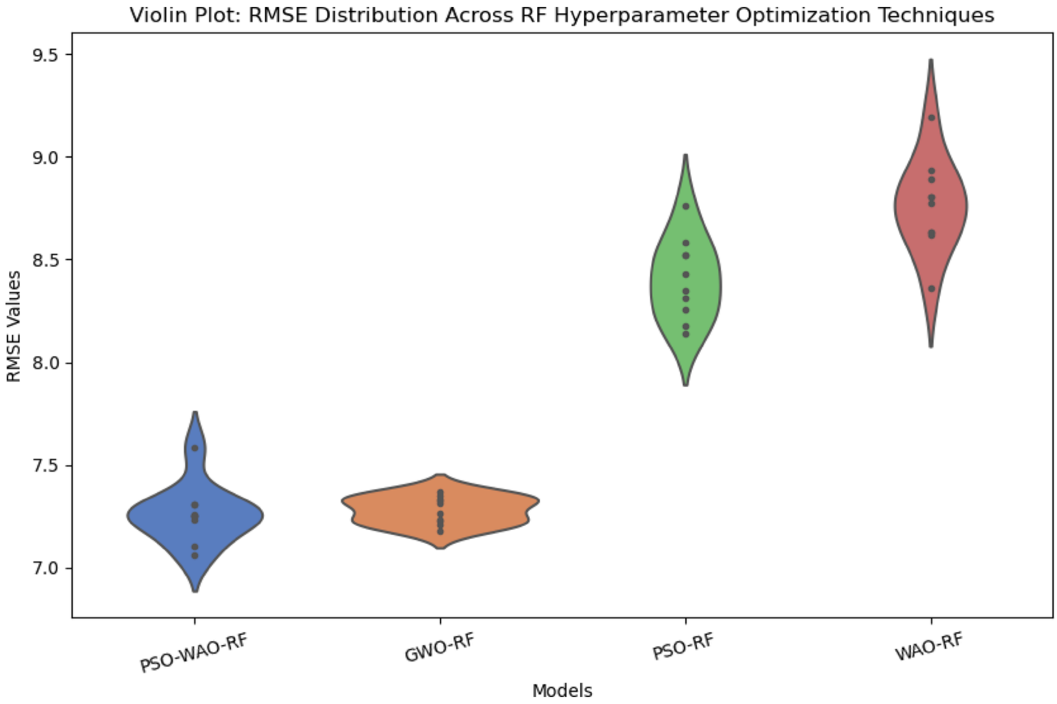


Fig. 16. Violin Plot Comparison of RMSE Distribution Across RF Hyperparameter Optimization Techniques.

Method	MSE	R^2	Computational Time (Seconds)
ARIMA (Statistical Model)	71.42	0.82	12.4
NWP-Based Model	65.73	0.85	300
ML Model (RF, BPSO-BWAO Selected Features)	53.93	0.91	55.4

Table 8. Comparison of ML-based AQI forecasting against NWP models and statistical methods.

	P-Value (Wilcoxon Test)	*P*-Value (t-Test)	Significance
RFE vs. BPSO-BWAO	0.031	0.027	yes
MI vs. BPSO-BWAO	0.042	0.038	
GS vs. PSO-WAO	0.019	0.015	
BO vs. PSO-WAO	0.025	0.021	

Table 9. Statistical significance test results comparing BPSO-BWAO and PSO-WAO with other feature selection and hyperparameter optimization methods.

Study	Methodology	Results
Proposed Work (BPSO-BWAO + PSO-WAO)	**Hybrid Feature Selection (BPSO-BWAO) to remove redundant features + Hyperparameter Optimization (PSO-WAO) for adaptive tuning. Models: RF, XGBoost, Extra Trees, SVM, DT, Deep Learning. Tested across EPA & Global City-Level AQI datasets.**	**R^2 of 0.9821, 9.5% MSE reduction, feature selection time reduced by 23.4%, hyperparameter tuning time reduced by 2.7×. Ensures better accuracy, computational efficiency, and generalization across different regions.**
Mishra & Gupta[42]	Compared Deep Learning (LSTM) with ML models (ARIMA, DT, KNN, XGBoost, GB, Adaptive Boosting, Huber Regressor, Dummy Regressor) using daily and hourly AQI data.	LSTM outperformed ARIMA for hourly data (RMSE: 44.65 vs. 69.65), but ARIMA was better for daily data (RMSE: 97.88 vs. 143.07).
Natarajan et al.[43]	Proposed Decision Tree optimized with Grey Wolf Optimization (GWO) for AQI prediction in major Indian cities. Compared with KNN, RF, and SVR.	Best accuracy: 97.68% (Visakhapatnam), 97.66% (Hyderabad), 94.48% (Kolkata), 88.98% (New Delhi).
Emeç & Yurtsever[44]	Developed a stacking ensemble model combining MLP, SVR, and RF for PM2.5 prediction. Used AQI data from Beijing & Istanbul.	Stacking model outperformed individual models (MAE: 6.67, RMSE: 8.80, R^2 = 0.91). Demonstrates the advantage of ensemble methods in PM2.5 forecasting.
Suthar et al.[45]	Used ANN, RF, SVR, and MLR to predict Land Surface Temperature (LST) based on AQI and meteorological factors in Bengaluru, India.	ANN achieved R^2 = 0.92 (summer), R^2 = 0.95 (winter), outperforming RF, SVR, and MLR.
Aram et al.[46]	Compared RF, GB, LASSO, and Stacked Regressor for AQI prediction, and KNN, SVM, DT, MLP, RF, and Stacked Classifier for AQG (Air Quality Grade) classification.	Stacked Regressor outperformed RF, GB, and LASSO (R^2 = 0.973, RMSE = 7.568, MAE = 4.596).

Table 10. Comparative analysis of AQI prediction proposed model compared to previous studies.

Data availability

This study used the EPA's Annual AQI by County (2024) dataset, available at https://aqs.epa.gov/aqsweb/airdata/annual_aqi_by_county_2024.zip. It contains annual air quality summaries from US counties, providing essential metrics for air pollution trend analysis.

Received: 14 February 2025; Accepted: 25 March 2025
Published online: 16 April 2025

References

1. Kang, G. K. et al. Air quality prediction: big data and machine learning approaches. *Int. J. Environ. Sci. Dev.* **9**(1), 8–16 (2018).
2. Athira, V. et al. DeepAirNet: applying recurrent networks for air quality prediction. *Procedia Comput. Sci.* **132**, 1394–1403 (2018).
3. World Health Organization. *Air quality guidelines for Europe* (WHO, 2020).
4. Madan, T. et al. Air Quality Prediction Using Machine Learning Algorithms–A Review. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, 2020, pp. 140–145.
5. Singh, K. P. et al. Linear and nonlinear modeling approaches for urban air quality prediction. *Sci. Total Environ.* **426**, 244–255 (2012).
6. Yi, X. et al. Deep distributed fusion network for air quality prediction. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 965–973.
7. Iskandaryan, D. et al. Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. *Appl. Sci.* **10**(7), 2401 (2020).
8. Mao, W. et al. Modeling air quality prediction using a deep learning approach: method optimization and evaluation. *Sustain. Cities Soc.* **65**, 102567 (2021).
9. Kök, I. et al. A Deep learning model for air quality prediction in smart cities. 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 1983–1990.
10. Zhang, Y. et al. A predictive data feature exploration-based air quality prediction approach. *IEEE Access.*. **7**, 30732–30743 (2019).
11. Yang, Z. & Wang, J. A new air quality monitoring and early warning system: air quality assessment and air pollutant concentration prediction. *Environ. Res.* **158**, 105–117 (2017).
12. Wang, J. et al. A deep spatial-temporal ensemble model for air quality prediction. *Neurocomputing.* **314**, 198–206 (2018).
13. Liang, Y. et al. Machine learning-based prediction of air quality. *Appl. Sci.* **10**(24), 9151 (2020).
14. Russo, A. et al. Air quality prediction using optimal neural networks with stochastic variables. *Atmos. Environ.* **79**, 822–830 (2013).
15. Ma, J. et al. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmos. Environ.* **214**, 116885 (2019).
16. Seng, D. et al. Spatiotemporal prediction of air quality based on LSTM Neural network. *Alex. Eng. J.* **60**(2), 2021–2032 (2021).
17. Zhu, D. et al. A machine learning approach for air quality prediction: model regularization and optimization. *Big data cogn. comput.* **2**(1), 5 (2018).
18. Li, X. et al. Deep learning architecture for air quality predictions. *Environ. Sci. Pollut. Res.* **23**, 22408–22417 (2016).
19. Soh, P. et al. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. *IEEE Access.* **6**, 38186–38199 (2018).
20. Bhalgat, P. et al. Air quality prediction using machine learning algorithms. *Int. J. Comput. Appl. Technol. Res.*. **8**(9), 367–370 (2019).
21. Chen, H. et al. Air quality prediction based on integrated dual lstm model. *IEEE Access.* **9**, 93285–93297 (2021).
22. Wang, J. et al. Developing an early-warning system for air quality prediction and assessment of cities in china. *Expert Syst Appl.* **84**, 102–116 (2017).
23. Lin, Y. et al. Air quality prediction by neuro-fuzzy modeling approach. *Appl. Soft Comput.* **86**, 105898 (2020).
24. Wang, J. et al. Air quality prediction using CT-LSTM. *Neu Comput. Appl.* **33**, 4779–4792 (2021).
25. Huang, Y. et al. Air quality prediction using improved pso-bp neural network. *IEEE Access.* **8**, 99346–99353 (2020).
26. Delavar, M. R. et al. A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. *ISPRS Int. J. Geo-Inf.* **8**(2), 99 (2019).
27. Liu, Q. et al. Air quality class prediction using machine learning methods based on monitoring data and secondary modeling. *Atmosphere.* **15**(5), 553 (2024).
28. Ansari, M. & Alam, M. An Intelligent IoT-Cloud-based air pollution forecasting model using univariate time-series analysis. *Arab. J. Sci. Eng.* **49**(3), 3135–3162 (2023).
29. Liu, H. et al. Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Appl. Sci.* **9**(19), 4069 (2019).
30. Elshewey, A. M., et al. Water Potability Classification Based on Hybrid Stacked Model and Feature Selection. *Environmental Science and Pollution* pp. 1–17 (2025).

31. Khafaga, D. S. et al. An al-biruni earth radius optimization-based deep convolutional neural network for classifying monkeypox disease. *Diagnostics*. **12**(11), 2892 (2022).
32. Abdullah, S. M. et al. Optimizing traffic flow in smart cities: soft gru-based recurrent neural networks for enhanced congestion prediction using deep learning. *Sustainability*. **15**(7), 5949 (2023).
33. Tarek, Z. et al. Wind power prediction based on machine learning and deep learning models. *Comput. Mater. Contin.* **74**(1), 715–732 (2022).
34. Al-Mahdawi, H. et al. Solving the inverse initial value problem for the heat conductivity equation by using the picard method. *J. Artif. Intell. Metaheuristics*. **2**(2), 46–55 (2022).
35. Towfek, S. K. & Elkanzi, M. A review on the role of machine learning in predicting the spread of infectious diseases. *Metaheuristic Optimiz. Rev.* **2**(1), 14–27 (2024).
36. Güven, A. F. et al. Comprehensive optimization of pid controller parameters for dc motor speed management using a modified jellyfish search algorithm. *Optim. Control Appl. Methods*. **46**(1), 320–342 (2025).
37. Güven, A. F. et al. Optimization of a hybrid microgrid for a small hotel using renewable energy and ev charging with a quadratic interpolation beluga whale algorithm. *Neural Comput. Appl.* **37**(5), 3973–4008 (2024).
38. Güven, A. F. et al. Multi-objective optimization and sustainable design: a performance comparison of metaheuristic algorithms used for on-grid and off-grid hybrid energy systems. *Neural Comput. Appl.* **36**(13), 7559–7594 (2024).
39. Alkhasawneh, M. S. Hybrid cascade forward neural network with elman neural network for disease prediction. *Arab. J. Sci. Eng.* **44**, 9209–9220 (2019).
40. Fouad, Y. et al. Adaptive visual sentiment prediction model based on event concepts and object detection techniques in social media. *Int. J. Adv. Comput. Sci. Appl.* https://doi.org/10.14569/IJACSA.2023.0140728 (2023).
41. El-Kenawy, E. M. et al. Optimizing HCV disease prediction in Egypt: The hyOPTGB framework. *Diagnostics* **13**(22), 3439 (2023).
42. Alkhammash, E. H. et al. Optimized multivariate adaptive regression splines for predicting crude oil demand in Saudi arabia. *Discrete Dyn. Nature Soc.* **2022**(1), 8412895 (2022).
43. Alzakari, S. A. et al. "Early detection of Potato Disease using an enhanced convolutional neural network-long short-term memory Deep Learning Model." Potato Research, 2024, pp. 1–19.
44. Assiri, S. A. et al. Application of machine learning to predict COVID-19 spread via an optimized BPSO model. *Biomimetics* **8**(6), 457 (2023).
45. Radwan, M. et al. Optimized deep learning for potato blight detection using the waterwheel plant algorithm and sine cosine algorithm. *Potato Res.* pp. 1–25 (2024).
46. Alhussan, A. et al. EEG-based optimization of eye state classification using modified-BER metaheuristic algorithm. *Sci. Rep.* **14**(1), 24489 (2024).
47. Aljuaydi, F. et al. A deep learning prediction model to predict sustainable development in Saudi Arabia. *Appl. Math. Inf. Sci.* **18**(6), 1345–1366 (2024).
48. Elkenawy, E. M. et al. Greylag goose optimization and multilayer perceptron for enhancing lung cancer classification. *Sci. Rep.* **14**(1), 23784 (2024).
49. El-Rashidy, N. et al. Multitask multilayer-prediction model for predicting mechanical ventilation and the associated mortality rate. *Neural Comput. Appl.* **37**(3), 1321–1343 (2025).
50. Bilal, A. et al. Quantum computational infusion in extreme learning machines for early multi-cancer detection. *J. Big Data*. **12**(1), 1–48 (2025).
51. Khafaga, D. et al. Enhancing heart disease classification based on greylag goose optimization algorithm and long short-term memory. *Sci. Rep.* **15**(1), 1277 (2025).
52. Bilal, A. et al. A quantum-optimized approach for breast cancer detection using SqueezeNet-SVM. *Sci. Rep.* **15**(1), 3254 (2025).
53. Elshewey, A. M. et al. Orthopedic disease classification based on breadth-first search algorithm. *Sci. Rep.* **14**(1), 23368 (2024).
54. Eed, M. et al. Potato Consumption forecasting based on a hybrid stacked deep learning model. *Potato Res*. pp. 1–25 (2024).
55. Abdelhamid, A. A. et al. Potato harvesting prediction using an improved resnet-59 model. *Potato Res*. pp. 1–20 (2024).

Acknowledgements

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R747), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

All authors are equally contributed.

Funding

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R747), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.