# Enhanced air quality index classification: leveraging genetic algorithm and SMOTE for accurate assessments in Indian cities

**Komal Kumar Napa[1], Ayodeji Olalekan Salau[2,3], Angati Kalyan Kumar[4], Sepiribo Lucky Braide[5], Aitizaz Ali[6], Ting Tin Tin[7,8]**

[1]Department of Artificial Intelligence and Data Science, Saveetha Engineering College, Chennai, India
[2]Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria
[3]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India
[4]Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science (MITS), Deemed to be University, Madanapalle, India
[5]Department of Electrical and Electronics Engineering, Rivers State University, Port Harcourt, Nigeria
[6]Strategic Research Institute, Asia Pacific University of Technology and Innovation (APU), Kuala Lumpur, Malaysia
[7]Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia
[8]School of Information Technology, UNITAR International University, Petaling Jaya, Malaysia

**Article Info**

***Article history:***

Received Jan 31, 2025
Revised May 22, 2026
Accepted Jun 19, 2026

***Keywords:***

Air pollution
Air quality index
Extreme machine learning
Genetic algorithm
Health impact

**ABSTRACT**

The escalating air pollution levels in Indian metropolitan regions necessitate robust predictive systems for air quality assessment. This study presents an advanced air quality index (AQI) forecasting model leveraging the radial basis function (RBF) kernel-based extreme learning machine (ELM) optimized using a genetic algorithm (GA). The proposed model is evaluated on real-time AQI datasets from four major Indian cities: Vishakhapatnam, Delhi, Hyderabad, and Patna. Initial experiments without class balancing yielded prediction accuracies of 83.9%, 88.3%, 87.0%, and 86.9% respectively. To address the class imbalance and enhance predictive performance, the synthetic minority oversampling technique (SMOTE) was applied. Post-balancing, the model achieved significantly improved accuracies of 93.9%, 94.7%, 92.3%, and 96.2% across the respective cities. These results underscore the effectiveness of integrating SMOTE with RBF-ELM for AQI prediction and demonstrate the critical role of data balancing in improving model generalizability. The proposed approach offers a promising solution for urban air quality monitoring and can assist policymakers in formulating timely interventions to mitigate health risks associated with air pollution.



***Corresponding Author:***

Ting Tin Tin
School of Information Technology, UNITAR International University
Petaling Jaya, Malaysia
Email: ting.tintin@researcher.unitar.my

## 1. INTRODUCTION

Air quality degradation poses a major threat to public health, particularly in rapidly urbanizing regions such as India, where rising vehicular emissions and industrial activities have intensified pollution levels and increased the incidence of respiratory and cardiovascular diseases [1]–[3]. The air quality index (AQI) provides a standardized measure of pollution severity based on key pollutants such as PM2.5, PM10, $NO_2$, $SO_2$, CO, $O_3$, $NH_3$, and Benzene, enabling effective public communication and regulatory action [4], [5]. With advances in sensor technologies and real-time monitoring, large-scale environmental datasets have become widely available, prompting extensive research into AQI forecasting using regression models, deep learning (DL) approaches, ensemble methods, optimization-driven architectures, and internet of things (IoT)-enabled systems [6], [7]. However, despite progress in model design, AQI prediction remains hindered by severe class imbalance, where minority pollution categories are underrepresented, leading to biased predictions and reduced generalizability. Although synthetic minority oversampling technique (SMOTE) has proven beneficial in balancing datasets across various machine learning (ML) tasks [8], its application in AQI forecasting remains limited, and few studies examine hybrid optimization frameworks. A clear research gap exists in developing a robust system that jointly addresses data imbalance and model optimization for improved fairness and predictive accuracy.

To bridge this gap, this study proposes a novel AQI prediction framework that integrates SMOTE-based balancing with a genetic algorithm (GA)–optimized radial based function extreme learning machine (RBF-ELM), tailored for four major Indian cities. The contributions include: i) a hybrid GA–ELM model capable of capturing nonlinear urban pollution patterns; ii) the use of SMOTE to enhance class-wise fairness and reduce predictive bias; and iii) a comprehensive evaluation comparing imbalanced and balanced datasets using accuracy, precision, recall, and visual performance metrics [9]. This work provides practical significance by enabling more equitable and reliable AQI classification, supporting informed policymaking, facilitating early warning systems, and contributing to urban public health management.

The remainder of this paper is structured as follows: section 2 reviews related work in air quality prediction and ML approaches. Section 3 outlines the system architecture and dataset description. Section 4 details the methodology including SMOTE, RBF-ELM, and GA integration. Section 5 presents the experimental results and analysis. Section 6 concludes the paper and suggests future research directions.

## 2. RELATED WORKS

Numerous studies have explored the use of ML and artificial intelligence (AI) for AQI prediction across diverse geographic contexts. Rao *et al.* [10] developed a multimodal imputation-based stacked ensemble model that combined multiple regression and classification learners to improve AQI prediction in Indian cities. Expanding on this, their later work proposed a multimodal autoencoder framework integrating data imputation and DL for AQI classification and prediction tasks, demonstrating strong performance in heterogeneous data conditions [11].

Natarajan *et al.* [12] designed an optimized ML model for AQI forecasting, incorporating feature selection and hyperparameter tuning tailored for major Indian cities. Their approach emphasized robust performance across different pollutant metrics. Similarly, Swamynathan *et al.* [13] adopted lightweight ML methods to build scalable AQI prediction models suitable for smart city deployment within IoT ecosystems. Mubeen *et al.* [14] integrated optimization strategies with DL to forecast AQI more accurately, leveraging historical data and meteorological inputs. In a regional context, Islam [15] conducted an AI-driven assessment of air pollution in Dhaka, Bangladesh, and proposed integrated ML techniques for both forecasting and policy-making. Megavarnan and Venkatachalam [16] introduced the POA-DT method, a pollutant-occurrence-aware decision tree model specifically designed for Indian urban datasets, and demonstrated improved accuracy and interpretability. Mim *et al.* [17] presented a comparative analysis of various ML algorithms under different hyperparameter tuning conditions, revealing significant differences in AQI prediction effectiveness. Shetty *et al.* [18] developed a hybrid AI model for both air quality forecasting and environmental mitigation strategy planning, offering a holistic view of pollution assessment.

From a review standpoint, Malhotra *et al.* [19] provided a systematic evaluation of AI-based AQI prediction models, highlighting common challenges such as data imbalance, missing values, and computational overhead. In response to these challenges, Sawah *et al.* [20] proposed a hybrid binary particle swarm optimization (BPSO) with butterfly optimization algorithm (BWAO) for feature selection and hyperparameter optimization, significantly boosting AQI prediction accuracy. Health impact studies have also emerged. Rao *et al.* [21] proposed a novel hybrid framework that links air quality data with lung health assessment using classification models. Meanwhile, Sanjit and Suchitra [22] applied regression-based techniques to uncover latent AQI patterns and promote sustainable urban air management. Patel *et al.* [23] focused on particulate matter (PM2.5 and PM10) prediction, comparing both ML and DL approaches. Their results highlighted the importance of pollutant-specific models. Agbehadji and Obagbuwa [24] reviewed spatiotemporal prediction models using ML and DL techniques, discussing real-world deployment challenges and evaluation benchmarks. Singh *et al.* [25] implemented a random forest-based approach, optimizing it for better prediction of urban air quality, emphasizing model robustness and scalability.

While these studies highlight the diversity of methods used for AQI prediction, many do not adequately address the challenge of imbalanced datasets, which can bias predictions toward dominant AQI classes. To bridge this gap, the present study introduces a novel RBF kernel-based ELM optimized using GA, in conjunction with the SMOTE, to improve classification fairness and accuracy across all AQI categories.

## 3. SYSTEM OVERVIEW

### 3.1. Dataset

This study employs air quality data obtained from the central pollution control board (CPCB), India, comprising both hourly and daily measurements of various atmospheric pollutants, including the AQI, for the period 2015 to 2020 [26]. The initial dataset consists of 29,532 entries and 16 features, representing monitoring stations distributed across 26 cities nationwide. For this analysis, data from four metropolitan regions, Vishakhapatnam, Delhi, Hyderabad, and Patna were selected to ensure regional diversity and urban relevance. The dataset incorporates critical air quality indicators such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, and Toluene, along with AQI and the associated categorical label 'AQI_Bucket,' which indicates the qualitative air quality level. An excerpt of the dataset for Visakhapatnam is presented in Table 1.

Table 1. Sample dataset of Visakhapatnam city

| Index | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 | O3 | Benzene | Toluene | AQI | AQI_Bucket |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 28075 | 2016-07-07 | 35.33 | 84.31 | 16.95 | 44.35 | 34.22 | 12.06 | 0.84 | 0.57 | 25.24 | 4.87 | 9.37 | 81.0 | Satisfactory |
| 28076 | 2016-07-08 | 40.62 | 92.73 | 28.48 | 52.89 | 47.01 | 11.91 | 1.07 | 0.57 | 16.24 | 6.46 | 12.2 | 112.0 | Moderate |
| 28077 | 2016-07-09 | 39.24 | 93.36 | 22.33 | 45.37 | 39.47 | 12.71 | 1.00 | 0.78 | 18.38 | 5.51 | 9.55 | 81.0 | Satisfactory |
| 28078 | 2016-07-10 | 26.64 | 69.66 | 11.26 | 36.51 | 26.20 | 12.62 | 0.88 | 0.97 | 27.98 | 4.12 | 8.15 | 81.0 | Satisfactory |
| 28079 | 2016-07-11 | 30.04 | 69.69 | 13.78 | 39.01 | 29.02 | 8.92 | 0.92 | 0.92 | 24.73 | 3.87 | 7.32 | 79.0 | Satisfactory |

### 3.2. Analysis of chosen cities and overview of methods

The selected cities, Visakhapatnam, Delhi, Hyderabad, and Patna were chosen for their environmental significance and diverse geographic and industrial profiles, enabling meaningful comparative analysis of regional air quality variations. To prepare the dataset, the 'Xylene' attribute with no valid entries was removed, and all null values were eliminated to ensure data integrity. SMOTE was applied to the training set to address class imbalance by generating synthetic minority samples, improving model generalization. The predictive framework employs an RBF-based ELM-RBF, chosen for its fast training and nonlinear learning capability. Hyperparameters were optimized using a GA implemented through DEAP, with accuracy guiding fitness selection across evolving populations. Model performance evaluated using accuracy, precision, and recall on both imbalanced and SMOTE-balanced datasets demonstrates the robustness and effectiveness of the proposed approach for air quality prediction.

## 4. METHOD

The proposed air quality prediction framework Figure 1 is designed for four major Indian cities Vishakhapatnam, Delhi, Hyderabad, and Patna and follows a structured, reproducible methodology. The dataset, obtained from the CPCB, consists of 29,532 records and 16 attributes from 2015–2020. Preprocessing includes removing attributes with no valid entries (such as Xylene), dropping rows with missing values, and imputing remaining numerical gaps using column means. Zero values, considered non-physical for certain pollutants, are replaced with mean values to prevent distortions. AQI_Bucket labels are encoded numerically, and Standard scaler is applied to normalize all features.

To address severe class imbalance, SMOTE is employed with k=5 neighbour's, applied only after a stratified 70:30 train–test split. This approach balances the training data while preserving the real-world distribution in the test set, thereby avoiding data leakage. The predictive model uses an ELM with an RBF-ELM. Its hyperparameters, including kernel width, are optimized using a GA implemented in DEAP. The GA begins with a population of 100 individuals, each encoding 12 ELM-RBF parameters, and evolves over 50 generations using one-point crossover and Gaussian mutation. Fitness is measured using classification accuracy, and the algorithm tracks key statistics across generations. A convergence threshold of 0.6, determined through exploratory trials, is used for early stopping to balance computational efficiency and optimization quality.

After convergence, the best individual is used as a transformation matrix for feature weighting, and the optimized feature set is used to train the final RBF-ELM model. Performance is evaluated using accuracy, precision, and recall on both imbalanced and SMOTE-balanced datasets.

### 4.1. Data preprocessing

The dataset used in this study, obtained from the CPCB, includes the observation date, city name, pollutant concentrations (PM2.5, PM10, NO, $NO_2$, $NO_x$, $NH_3$, CO, $SO_2$, $O_3$, Benzene, and Toluene), the AQI, and its categorical label AQI_Bucket, which classifies air quality into six levels. Although the dataset spans many Indian cities, this study focuses on Visakhapatnam, Delhi, Hyderabad, and Patna to ensure analytical clarity and capture diverse pollution patterns.

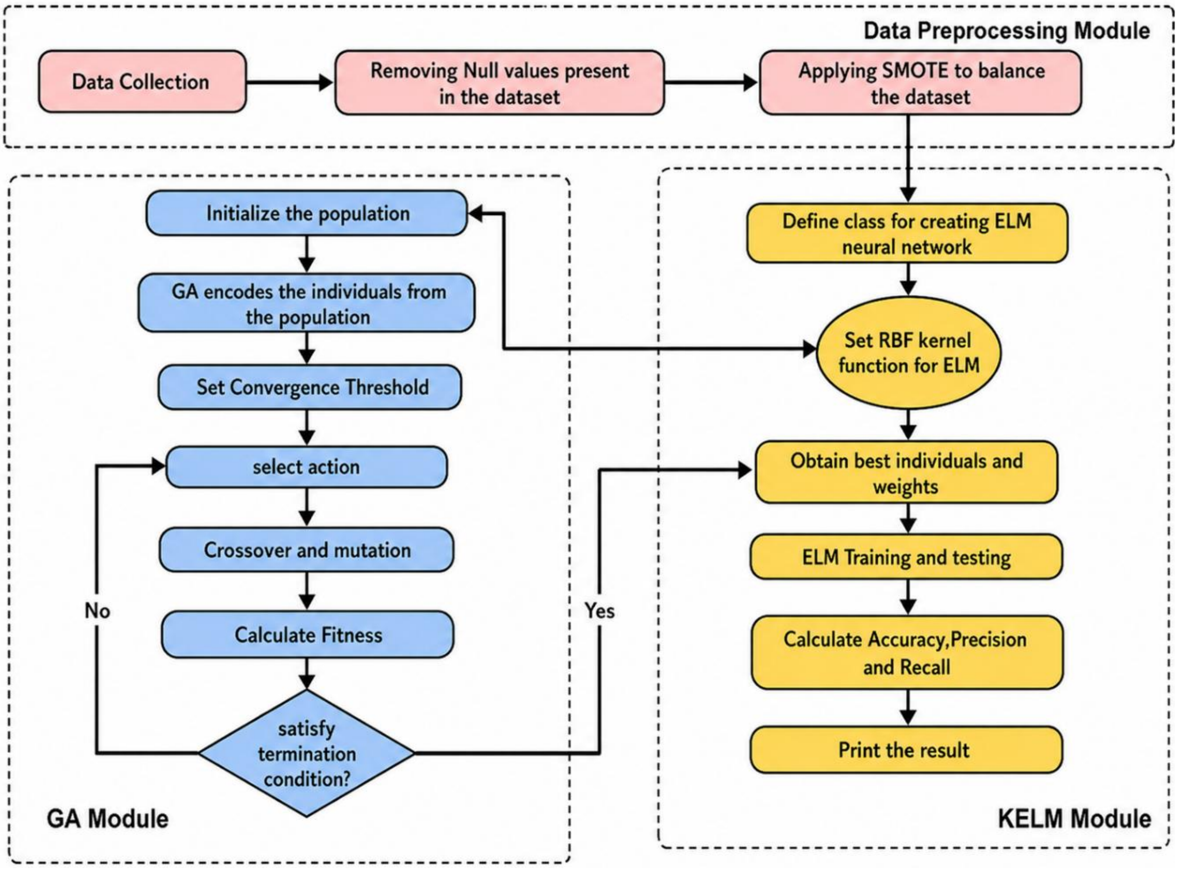


Figure 1. Proposed method

Data preprocessing involves removing records with missing values to maintain data quality. To correct severe class imbalance in AQI categories, the SMOTE is applied, generating synthetic samples for underrepresented classes. Standard scaler is used to normalize all features, ensuring stable model training. The predictive framework employs an ELM with an RBF kernel, optimized using a GA. This comprehensive preprocessing pipeline ensures a reliable foundation for accurate AQI prediction.

#### 4.1.1. Synthetic minority oversampling technique

SMOTE is a powerful technique designed to address class imbalance in ML datasets, particularly prevalent in scenarios where one class significantly outnumbers the others. It focuses on the minority class, generating synthetic instances to rebalance the distribution and enhance model performance. The method was introduced by Chawla *et al.* [8].

Let $X$ be represent the original feature space, $C$ be the minority class, and $x_i$ denote a sample form $C$. For a given minority instance $x_i$, let $x_{zi}$ be one of its k nearest neighbors. The synthetic instance $x_{new}$ is generated using (1):

$$x_{new}=x_i+\lambda.(x_{zi}-x_i) \quad (1)$$

where λ is a random value between 0 and 1.

In Visakhapatnam, the AQI bucket distribution was highly imbalanced, with most samples falling under the “Moderate” and “Satisfactory” categories, while “Very Poor,” “Poor,” and “Good” had limited representation and “Severe” was completely absent Figure 2. Such imbalance can bias classification models toward majority classes. To address this, SMOTE was applied to generate synthetic samples for minority categories through interpolation, resulting in a balanced dataset. After oversampling, each AQI bucket Severe, Very Poor, Poor, Satisfactory, Moderate, and Good contained 368 samples Figure 3. This balancing step significantly improved model fairness, robustness, and overall predictive performance.

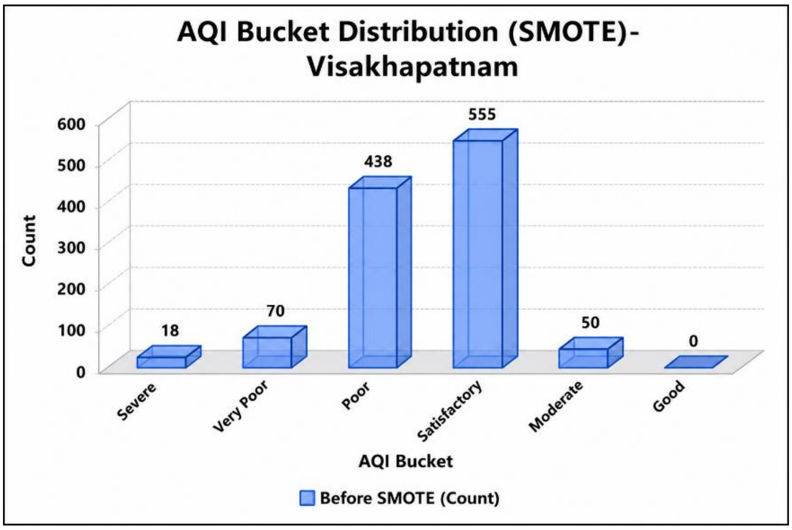


Figure 2. Imbalanced data for Visakhapatnam city

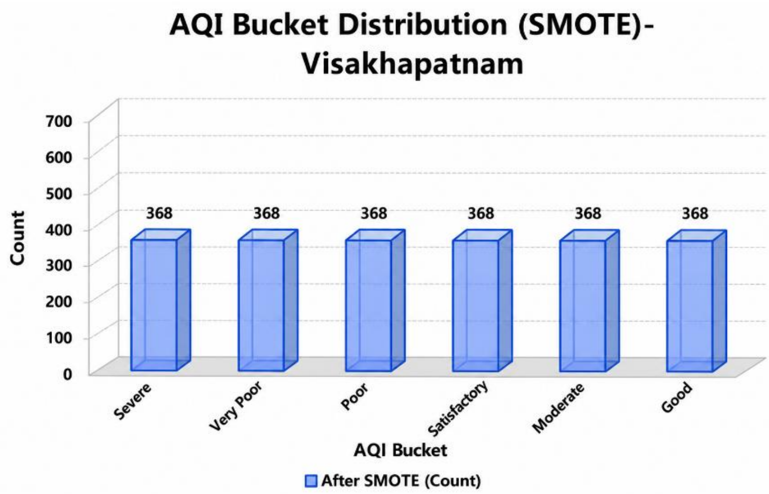


Figure 3. Balanced by SMOTE

#### 4.1.2. Feature scaling using standard scaler

Standard scaling is a crucial preprocessing step in ML, particularly when features exhibit varying scales. The method centers the data by subtracting the mean and scales it by dividing by the standard deviation. This normalization ensures that all features contribute equally to the model preventing biases introduced by variables with larger scales [27]. The standardization process is mathematically expressed as (2):

$$z = \frac{x - mean}{standard\ deviation} \quad (2)$$

where x represents the original feature value, meandenotes the mean of the feature, and standard deviationis the standard deviation [28].

### 4.2. Predictive modelling

#### 4.2.1. Extreme machine learning

The ELM is a fast and efficient ML algorithm designed for training single-layer feedforward neural networks. Introduced by Huang *et al.* [29], ELM differs from traditional gradient-based methods by using randomly assigned input weights and fixed hidden-layer parameters, eliminating iterative optimization. The hidden neurons act as random feature extractors, projecting input data into a high-dimensional space. Output weights are computed analytically using a linear system solution, which significantly accelerates training. This structure enables ELM to achieve strong generalization performance while maintaining exceptional computational efficiency.

The output of the hidden layer can be represented as (3):

$$H=\sigma(W_{in}\,.\,X+b_{in}) \tag{3}$$

where $H$ is the hidden layer output, $W_{in}$ is the input weight matrix, $X$ is the input data, and $b_{in}$ is the bias vector. The activation function σ introduces non-linearity.

The output weights are determined by solving the linear system:

$$W_{out}=H^{+}.Y \tag{4}$$

#### 4.2.2. Radial basis function kernel-based extreme learning machine

The RBF Kernel-ELM is a powerful ML algorithm that combines the efficiency of ELM with the nonlinear mapping capabilities of the RBF kernel. ELM, is known for its simplicity and fast learning speed. It involves a single-hidden layer feedforward neural network (SLFN) with randomly generated input weights and biases. The RBF kernel enhances the ELM by introducing a nonlinear transformation, allowing the model to capture complex relationships within the data. The RBF-ELM employs RBF, which are mathematical functions that depend on the distance from a central point or prototype. These functions facilitate the mapping of input features into a high-dimensional space, enabling the model to better handle intricate patterns in the data.

The mathematical representation of the RBF-ELM can be expressed as (5):

$$f(x) = \sum_{i=0}^{N} \beta_{i.}\, \emptyset(\left||x - c_i|\right|) \tag{5}$$

Here, f(x) represents the output of the RBF-ELM for input x, βi denotes the output weights, ci represents the center of the RBF, and (||x–ci||) is the RBF activation. The key advantage of the RBF-ELM lies in its ability to efficiently approximate complex functions with a reduced risk of overfitting. The RBF kernel introduces a nonlinear transformation, allowing the model to capture intricate relationships in the data without compromising computational efficiency. Siouda *et al.* [30] have demonstrated the effectiveness of the RBF-ELM in various applications, including pattern recognition, regression, and classification tasks. The integration of the RBF kernel with ELM enhances the model's capacity to handle diverse datasets and improve predictive performance.

#### 4.2.3. Genetic algorithm

In the proposed model, the GA is used to optimize parameters of the RBF-based ELM for improved AQI prediction [31]. Each GA individual encodes an ELM parameter set, and evolutionary operations selection, crossover, and mutation iteratively refine the population toward optimal configurations. This enables the model to capture complex nonlinear pollution patterns more effectively than fixed-parameter approaches. A convergence threshold of 0.6 serves as an early-stopping criterion, preventing unnecessary computation and limiting overfitting. By balancing exploration and exploitation, the GA enhances model robustness, reduces computational effort, and contributes to a more accurate and efficient AQI prediction framework [32], [33].

## 5. EXPREIMENTAL ANALYSIS AND FINDINGS

The proposed air quality prediction model was evaluated using standard performance metricsaccuracy, precision, and recallto provide a comprehensive analysis of its predictive capabilities across four urban regions: Vishakhapatnam, Delhi, Hyderabad, and Patna. In the initial phase, the model was trained and tested on imbalanced datasets without applying any balancing methods, resulting in baseline classification accuracies of 83.90% for Vishakhapatnam, 88.31% for Delhi, 87.08% for Hyderabad, and 86.91% for Patna.

In this study, the hyperparameters of the RBF kernel, including kernel width, were not fixed but optimized using a GA. The GA systematically tuned key parameters such as kernel width, convergence threshold, and population size to achieve optimal performance, ensuring that the RBF-ELM configuration was adapted to the specific characteristics of the dataset, thereby enhancing predictive accuracy. To address class imbalance, SMOTE was applied, with particular care taken to preserve the class distribution during the train–test split. Specifically, SMOTE was used exclusively on the training set after the initial split, ensuring that the test set retained its original real-world class proportions. This approach prevented data leakage and allowed for an unbiased assessment of the model's ability to generalize to naturally imbalanced distributions while benefiting from balanced data during training.

The experiments were conducted in Python 3.10 using Jupyter Notebook on an Intel i7, 16 GB RAM system. An RBF-ELM model optimized via a GA (population 100, 50 generations, threshold 0.6) was used. SMOTE-balanced, stratified 70:30 data splits enabled performance evaluation on both imbalanced and balanced datasets. Table 2 depicts the findings of the study. Figure 4 represents the performance metrics before and after applying SMOTE for four cities.

Table 2. Performance metrics comparison for air quality prediction

| City | Accuracy (before SMOTE) (%) | Accuracy (after SMOTE) (%) | Precision (before SMOTE) (%) | Precision (after SMOTE) (%) | Recall (before SMOTE) (%) | Recall (after SMOTE) (%) |
|---|---|---|---|---|---|---|
| Vishakhapatnam | 83.90 | 93.87 | 83.90 | 93.87 | 83.90 | 93.87 |
| Delhi | 88.31 | 94.85 | 88.31 | 94.85 | 88.31 | 94.85 |
| Hyderabad | 87.08 | 92.55 | 87.08 | 92.55 | 87.08 | 92.55 |
| Patna | 86.91 | 96.25 | 86.91 | 96.25 | 86.91 | 96.25 |

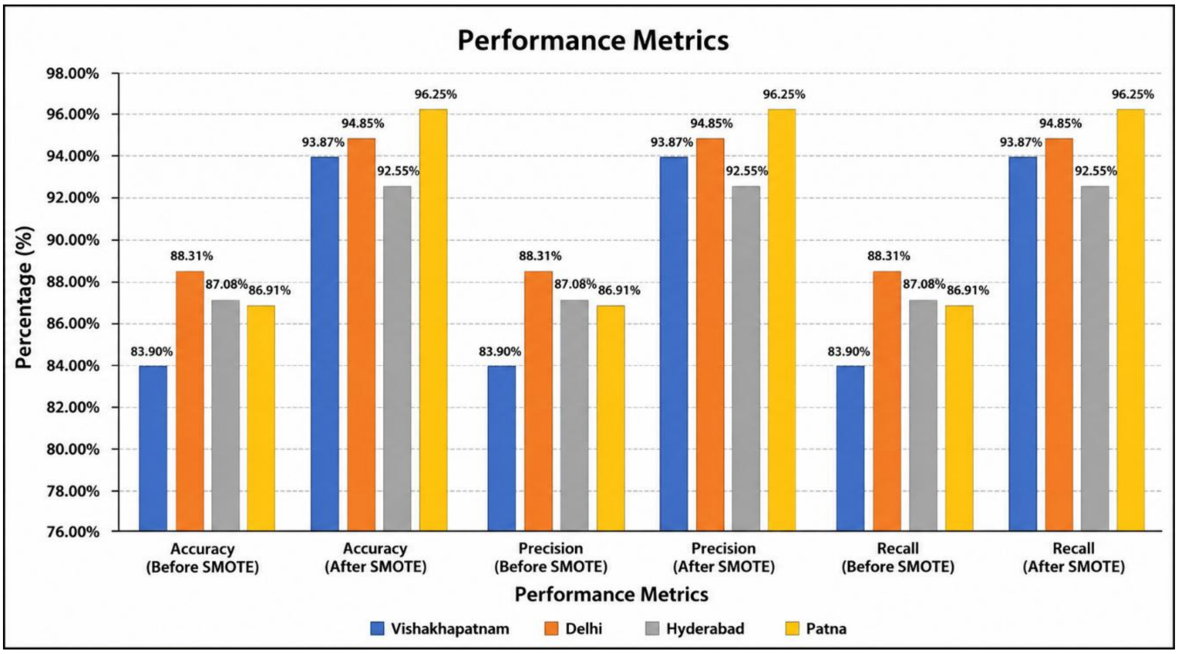


Figure 4. Performance metrics before and after SMOTE

A substantial improvement in model performance was observed after applying the SMOT to address class imbalance in the AQI datasets. Before balancing, the model achieved only moderate accuracy across the four cities. After introducing SMOTE, accuracy increased markedly to 93.87% in Vishakhapatnam, 94.85% in Delhi, 92.55% in Hyderabad, and 96.25% in Patna, demonstrating that balancing the dataset enables better generalization and reduces bias toward majority AQI classes.

Precision, which reflects the model's ability to minimize false positives, showed a similar upward trend. Initial precision values ranged from 83.90% to 88.31%, indicating moderate reliability in positive predictions. With SMOTE, precision rose substantially to 93.87%, 94.85%, 92.55%, and 96.25%, showing improved correctness in identifying AQI categories and fewer erroneous classifications. This highlights SMOTE's effectiveness in strengthening prediction confidence. Recall, measuring the model's ability to detect true positive cases, also improved significantly. Originally between 83.90% and 88.31%, recall increased to 93.87%, 94.85%, 92.55%, and 96.25% after SMOTE, confirming enhanced sensitivity to all AQI classes, including minority categories.

Overall, the consistent gains in accuracy, precision, and recall affirm SMOTE's critical role in improving model fairness, minority-class representation, and overall robustness in real-world air quality prediction.

## 6. CONCLUSION

This study proposes a robust air quality prediction framework for four Indian cities Vishakhapatnam, Delhi, Hyderabad, and Patna. By applying the SMOTE, the model effectively addressed class imbalance, leading to significant improvements in predictive performance. The integration of an RBF-based ELM optimized with a GAusing a convergence threshold of 0.6, a population of 100, and 50 generations further enhanced model accuracy and stability. Post-SMOTE, the model achieved strong and consistent results, with accuracy, precision, and recall reaching 93.87% in Vishakhapatnam, 94.85% in Delhi, 92.55% in Hyderabad, and 96.25% in Patna. These improvements highlight the framework's reliability for real-world air quality monitoring and its value in supporting data-driven public health and policy decisions. Future enhancements may include incorporating real-time IoT sensor data for continuous updates and integrating anomaly detection to identify unusual pollution patterns. Additionally, benchmarking the proposed method against transformer-based architectures and deep temporal forecasting models such as TCN and LSTM variants may further improve performance by leveraging their ability to capture long-range temporal dependencies and complex nonlinear trends, aiding sustainable urban planning and public health management.


## ACKNOWLEDGMENTS

The authors are grateful to their respective organizations.

## FUNDING INFORMATION

This paper was funded by INTI International University and UNITAR International University.


## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Komal Kumar Napa | ✓ | ✓ | | ✓ | ✓ | | | ✓ | ✓ | ✓ | | | ✓ | |
| Ayodeji Olalekan Salau | ✓ | ✓ | | | | ✓ | | ✓ | ✓ | ✓ | | | ✓ | |
| Angati Kalyan Kumar | | ✓ | ✓ | | | | ✓ | | | ✓ | ✓ | | | |
| Sepiribo Lucky Braide | ✓ | | ✓ | ✓ | | | | ✓ | | ✓ | ✓ | | | |
| Aitizaz Ali | | ✓ | | ✓ | ✓ | | | | | ✓ | | | | |
| Ting Tin Tin | | | | | ✓ | | ✓ | ✓ | | ✓ | | | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing -**O**riginal Draft
E : Writing - Review &**E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [Komal Kumar Napa], upon reasonable request.

## REFERENCES


[1] Q. A. Chakkaravarthy, "Human survival and environmental pollution," in *Proceedings of the Third International Conference on Environment and Health*, 2003, pp. 66–74.

[2] A. Monoson *et al.*, "Air pollution and respiratory infections: the past, present, and future," *Toxicological Sciences*, vol. 192, no. 1, pp. 3–14, Mar. 2023, doi: 10.1093/toxsci/kfad003.

[3] A. Nakhjiri and A. A. Kakroodi, "Air pollution in industrial clusters: A comprehensive analysis and prediction using multi-source data," *Ecological Informatics*, vol. 80, May 2024, doi: 10.1016/j.ecoinf.2024.102504.

[4] S. A. Horn and P. K. Dasgupta, "The air quality index (AQI) in historical and analytical perspective a tutorial review," *Talanta*, vol. 267, Jan. 2024, doi: 10.1016/j.talanta.2023.125260.

[5] D. R. T. Prajapati, "Iot base solutions for real time air quality monitoring in smart cities," *International Journal of Environmental Sciences*, pp. 826–842, Jul. 2025, doi: 10.64252/smtzhc34.

[6] M. N. A. Ramadan, M. A. H. Ali, S. Y. Khoo, M. Alkhedher, and M. Alherbawi, "Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment," *Ecotoxicology and Environmental Safety*, vol. 283, Sep. 2024, doi: 10.1016/j.ecoenv.2024.116856.

[7] A. Mishra and Y. Gupta, "Comparative analysis of air quality index prediction using deep learning algorithms," *Spatial Information Research*, vol. 32, no. 1, pp. 63–72, Feb. 2024, doi: 10.1007/s41324-023-00541-1.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[9] R. Abdulin, "Visual aesthetics in environmental data visualization: A preference-based study on color palettes, themes, and chart types for displaying air pollution data," Ph.D. dissertation, MID Sweden University, 2025.

[10] R. S. Rao, L. R. Kalabarige, B. Alankar, and A. K. Sahu, "Multimodal imputation-based stacked ensemble for prediction and classification of air quality index in Indian cities," *Computers and Electrical Engineering*, vol. 114, Mar. 2024, doi: 10.1016/j.compeleceng.2024.109098.

[11] A. Garcia, Y. Saez, I. Harris, X. Huang, and E. Collado, "Advancements in air quality monitoring: A systematic review of IoT-based air quality monitoring and AI technologies," *Artificial Intelligence Review*, vol. 58, no. 9, Jun. 2025, doi: 10.1007/s10462-025-11277-9.

[12] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports*, vol. 14, no. 1, Mar. 2024, doi: 10.1038/s41598-024-54807-1.

[13] S. Swamynathan, N. Sneha, S. P. Ramesh, R. Niranjana, D. D. N. Ponkumar, and R. Saravanakumar, "A machine learning approach for predicting air quality index in smart cities," in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, Dec. 2024, pp. 1609–1615, doi: 10.1109/ICICNIS64247.2024.10823106.

[14] M. Mubeen *et al.*, "Smart prediction and optimization of air quality index with artificial intelligence," *Journal of Environmental Sciences*, vol. 158, pp. 761–775, Dec. 2025, doi: 10.1016/j.jes.2025.02.041.

[15] F. A. S. Islam, "A comprehensive analysis of air pollution in Dhaka City, Bangladesh, and the application of artificial intelligence and machine learning for enhanced management and forecasting," *International Journal of Applied and Natural Sciences*, vol. 3, no. 1, pp. 131–167, May 2025, doi: 10.61424/ijans.v3i1.303.

[16] G. Megavarnan and K. Venkatachalam, "POA-DT: A novel method for predicting air quality in major Indian cities," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1320–1329, Apr. 2025, doi: 10.11591/eei.v14i2.8958.

[17] S. T. Mim, F. Alam, and N. Sharmin, "Air quality index prediction: Comparative study based on various hyper parameter tuning," *Iranian Journal of Science*, vol. 49, no. 6, pp. 1563–1586, Dec. 2025, doi: 10.1007/s40995-025-01842-w.

[18] C. Shetty *et al.*, "A machine learning approach for environmental assessment on air quality and mitigation strategy," *Journal of Engineering*, no. 1, Jan. 2024, doi: 10.1155/2024/2893021.

[19] M. Malhotra, S. Walia, C.-C. Lin, I. K. Aulakh, and S. Agarwal, "A systematic scrutiny of artificial intelligence-based air pollution prediction techniques, challenges, and viable solutions," *Journal of Big Data*, vol. 11, no. 1, Oct. 2024, doi: 10.1186/s40537-024-01002-8.

[20] M. S. Sawah, H. Elmannai, A. A. El-Bary, K. Lotfy, and O. E. Sheta, "Improving air quality prediction using hybrid BPSO with BWAO for feature selection and hyperparameters optimization," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-95983-y.

[21] B. R. Rao, K. S. Chakradhar, D. Ganesh, D. Nataraj, and C. Rohini, "Impact of air quality and its effects on lungs health: A novel hybrid approach for quality assessment," in *2025 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, Apr. 2025, pp. 1–6, doi: 10.1109/RAEEUCCI63961.2025.11048212.

[22] C. R. Sanjit and S. Suchitra, "Unveiling hidden air quality patterns: A data-driven predictive regression analysis for sustainable air management," in *2024 2nd International Conference on Networking and Communications (ICNWC)*, Apr. 2024, pp. 1–10, doi: 10.1109/ICNWC60771.2024.10537472.

[23] P. Patel *et al.*, "A systematic study on PM2.5 and PM10 concentration prediction in air pollution using machine learning and deep learning model," *Environmental Chemistry and Ecotoxicology*, vol. 7, pp. 1401–1415, 2025, doi: 10.1016/j.enceco.2025.07.001.

[24] I. E. Agbehadji and I. C. Obagbuwa, "Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction," *Atmosphere*, vol. 15, no. 11, Nov. 2024, doi: 10.3390/atmos15111352.

[25] S. Singh, M. Kumar, B. K. Verma, and S. Kumar, "Optimizing air pollution prediction with random forest algorithm," *Aerosol Science and Engineering*, vol. 10, no. 2, pp. 208–221, Apr. 2026, doi: 10.1007/s41810-025-00292-6.

[26] Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change, Government of India, "Real-Time Air Quality Data." [Online]. Available: https://cpcb.nic.in/real-time-air-quality-data/. (Accessed Jul. 25, 2025).

[27] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] H. Abdi and L. J. Williams, *Principal component analysis*. Wiley Interdisciplinary Reviews: Computational Statistics, 2010.

[29] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1–

3, pp. 489–501, Dec. 2006, doi: 10.1016/j.neucom.2005.12.126.

[30] R. Siouda, M. Nemissi, and H. Seridi, "Diverse activation functions based-hybrid RBF-ELM neural network for medical classification," *Evolutionary Intelligence*, vol. 17, no. 2, pp. 829–845, Apr. 2024, doi: 10.1007/s12065-022-00758-3.

[31] K. Cornelius, N. K. Kumar, S. Pradhan, P. Patel, and N. Vinay, "An efficient tracking system for air and sound pollution using IoT," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 22–25, doi: 10.1109/ICACCS48705.2020.9074301.

[32] N. K. Kumar, D. Vigneswari, and C. Rogith, "An effective moisture control based modern irrigation system (MIS) with Arduino Nano," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Mar. 2019, pp. 70–72, doi: 10.1109/ICACCS.2019.8728446.

[33] S. S. Pawanekar, J. S. Kallimani, G. Udgirkar, M. R. Rashmi, M. C. Sanitha, and L. H. Pin, "Efficient AQI prediction: A comparative study of artificial neural networks, LSTM, random forest, and gradient boosting techniques," in *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Oct. 2024, pp. 1597–1603, doi: 10.1109/I-SMAC61858.2024.10714872.

## BIOGRAPHIES OF AUTHORS

**Komal Kumar Napa** is currently working as an Assistant Professor (SG) in the Department of Artificial Intelligence and Data Science at Saveetha Engineering College, Chennai, India. His research interests include machine learning, data mining, and cloud computing. He can be contacted at email: komalkumarnapa@gmail.com.

**Dr. Ayodeji Olalekan Salau** is an Associate Professor in the Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria (2022-present). He serves as a reviewer for numerous reputable international journals, and is Co-Chair and technical program committee member of various international conferences. He has authored 4 Academic books and has over 300 international research articles in major reputable international Journals, book series, and conferences. His research interests include research in the fields of computer vision, image processing, signal processing, artificial intelligence, and power systems engineering. In 2019, his paper was awarded the best paper of the year in Cogent Engineering. According to the SciVal (SCOPUS- Elsevier) analysis/ranking of academics in 2026, he is among the top researchers in Nigeria (ranked number 6 out of the top 500). He was ranked among the World's Top 2% Scientists in the field of Artificial Intelligence by Scopus Elsevier/Stanford University in 2025, 2024, and 2023. He received the Award of 2nd Position in the category of Best AI researcher in Higher Education (Universities Category) in Nigeria on the 2025 AI awareness day, on 17th February, 2025, given by the National Association of AI Practitioners (NAAIP) in conjunction with the National Universities Commission (NUC) and TETFund. More recently, he received the Best Researcher Award at the prestigious International UILA Awards (IUILA 2025) which took place in conjunction with the 6th International Conference on Data Analytics & Management (ICDAM-2025), organised by London Metropolitan University, London, UK. He was also awarded the 2025 Outstanding Associate Editor of Frontiers in Bioinformatics (Computational Bioimaging section). He can be contacted at email: ayodejisalau98@gmail.com.

**Mr. Angati Kalyan Kumar** is currently working as Assistant Professor in the Department of Computer Science and Engineering (Data Science) at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research interests include ML, and data mining (DM). He can be contacted at email: kalyanangatikumar@gmail.com.

**Dr. Sepiribo Lucky Braide** is an Associate Professor in the Department of Electrical Electronics Engineering, Rivers State University, Port Harcourt, Nigeria. He is the past Head of Department, Electrical Engineering from 2018 to 2021, Fellow of the Nigeria Institute of Electrical Electronics Engineers (FNIEEE) and currently the Post Graduate Coordinator in the Department, from 2021 till date. He is a member of several professional bodies/organizations among many includes; Institute of Electrical Electronics Engineers (IEEE), Nigeria Society of Engineers (MNSE), Council for the Regulation of Engineering in Nigeria (COREN), Nigeria Institute of Electrical Electronics Engineers (MNIEEE), International Association of Engineers (MIAENG), he had the following degrees in Electrical Engineering profession which includes; Bachelor of Technology (B.Tech.), Master of Technology (M.Tech.) and Doctor of Philosophy (Ph.D.). He has attended several conferences these includes: ICEPT, NSE, COREN, IEEE, WCECS (San Francisco, USA Oct., 2019), WCE (London UK, July, 2019). His strong pragmatic confrontation to analyses of finding solutions to challenging engineering problems made him compete in many highly reputable international journals, most recently in 2018, International Journal Engineering and Science Invention (IJESI) classified one of his articles as Best Paper Award - 2018 certified and titled: "A Mathematical Model of Double Exponential Wave Shape (Impulse Generator) for Power Sub-station using Laplace Transform". He has received so many honors and awards both local and international. He can be contacted at email: sepiribo.braide@ust.edu.ng.

**Dr. Aitizaz Ali** is a highly respected academic and professional known for his extensive contributions to the fields of artificial intelligence (AI), cybersecurity, and blockchain technology. He holds a distinguished position as a Senior Member of the IEEE, reflecting his significant impact on the global engineering and technology communities. Over the course of his career, he has authored and co-authored more than 100 research papers, establishing himself as a prolific researcher in his domains. His research interests span across several emerging technologies, and his papers have been published in prestigious journals and conferences. His expertise in AI, cybersecurity, and blockchain has earned him recognition as a thought leader in these areas, and his contributions continue to shape the future of these technologies. As a Senior Member of IEEE, he has not only contributed to research but has also been actively involved in various IEEE initiatives, where he shares his knowledge and expertise with the next generation of engineers and researchers. His work and leadership continue to inspire colleagues and students alike, and his commitment to advancing technology remains unwavering. He can be contacted at email: aitizazengr@gmail.com.

**Dr. Ting Tin Tin** received her B.Sc. and Ph.D. in Computer Sciences from University of Science, Malaysia. She joined Gemalto as telecommunication software engineer in Singapore before joined academic industry after her Ph.D. graduation. She has more than 12 years of lecturing, supervising projects, and research. She received her professional certification in project management from PMI and data analytics from SAS. Currently, she is attached to INTI International University responsible primary in research and postgraduate supervision. She is a passionate researcher in her publication collaborate with national and international researchers i.e., Singapore, Bangladesh, India, Nigeria, Pakistan, United Arab Emirates, and Saudi Arabia. Her research interests including big data analytics, information systems engineering, educational data mining, psycho-academic research, and software engineering. She started being active in her research since 2022, increasing her google scholar H-index from 1 to 4 (2023). Currently she is a freelancer serves as freelance lecturer in INTI International University, Monash University, Tunku Abdul Rahman University of Management and Technology, and Methodist College Kuala Lumpur. Her responsibility including Ph.D. students' supervision; Master's programme course leader; Course leader; delivery, coordination and development of learning materials for undergraduates, setting of examination papers and marking schemes, conduct lecture, tutorial, practical laboratory, mid-term test; open day faculty consultant; final year project supervisor and moderator; research publication. She can be contacted at email: ting.tintin@researcher.unitar.my and tintin.ting@newinti.edu.my.