Bio-Inspired Optimization-Driven Air Quality Prediction Using Machine Learning

OPEN ACCESS

Volume: 13

Special Issue: 3

Month: February

Year: 2026

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Citation:
Devi, V., et al. "Bio-Inspired Optimization-Driven Air Quality Prediction Using Machine Learning." *Shanlax International Journal of Arts, Science and Humanities,* vol. 13, no. 3, 2026, pp. 107–16.

DOI:
https://doiorg/10.34293/sijash.v13iS3-i1-Feb.10268

Dr. V. Devi
Professor, PG & Research Department of Computer Science
Thiruthangal Nadar College, India

Ms. M. Divya
Research Scholar, PG & Research Department of Computer Science
Thiruthangal Nadar College, India

Dr. A. Ambeth Raja
Associate Professor, PG & Research Department of Computer Science
Thiruthangal Nadar College, India

D. R. Divesh Kumar
Student, Department of Civil Engineering
Chennai Institute of Technology, Chennai

Abstract

Precise air quality forecasting is vital to environmental management and health safety of people but traditional machine learning solutions tend to experience reduced performance as a result of large-dimensional input features and inefficient hyper-parameter settings. To overcome these shortcomings, this paper suggests a new bioinspired optimization-based machine learning framework where both feature selection and hyper-parameter optimization are done through a single co-optimization. The model presents a multiphase hybrid search methodology combining Whale Optimization Algorithm, which is a global exploration method, with the use of the Gay Wolf Optimizer, which is an adaptive convergence method, and the Particle Swarm Optimization, which is a fine-grained exploitation method, to facilitate efficient search of complex solution spaces. In order to develop Air Quality Index predictions models, real-world air quality data in the form of multivariate pollutant concentrations and meteorological variables were used. The optimized learning parameters were found to be very effective in the errors of forecasting and the complexity of the model. The experimental findings show reductions of errors by over 35 percent compared to traditional machine learning methods and a steady improvement in performance as compared to single-optimizer methods. The XG-Boost optimization model performed best on the predictive performance with an R 2 of 0.951 using a smaller magnitude of features (up to 50% reduction in features) which shows the efficiency of the proposed dimensionality-aware optimization framework. The results suggest that bio-inspired co-optimization could be used to provide reliable, scalable, and computationally efficient air quality prediction systems to be applied in real-time environmental monitoring and smart cities.

Keywords: Air Quality, Prediction, Bio Inspired Algorithm, Optimization

Introduction

Air pollution is one of the most critical environmental problems of the twenty-first century which is highly dangerous to human health, ecological balance and economic efficiency. Rapid urbanization, industrial growth, automobile emissions, and energy use have greatly increased the concentration of such harmful pollutants as particulate matter (PM_2.ancac and PM 1 0) and nitrogen dioxide (NO 2), sulfur dioxide (SO 2), carbon monoxide (CO), and ozone (O 3) in the air. The long-term exposure to such contaminants has been associated with respiratory infections, heart diseases, neuronal disorders, and higher mortality rates especially when used in the crowded urban areas[1-4]. This, therefore, has made constant surveillance and proper forecasting of the quality of air in the atmosphere a key priority to the environmental departments, health care organizations, and policy makers to ensure timely measures are adopted and they develop sustainable city planning plans.

The standard air quality forecasting methods have been largely based on deterministic atmospheric dispersion models and classical statistical methods such as autoregressive integrated moving average (ARIMA), linear regression and chemical transport models. Although these techniques might be useful in understanding the behavior of pollutants, they tend to be ineffective with nonlinear behavior, excessive variability and more intricate interactions among meteorological conditions and the sources of emissions[5-8]. Additionally, these models are usually large, need a lot of domain-specific knowledge, emission inventories, and computationally expensive, and they are not scalable or can be used in real time. Due to the increasing complexity of urban settings and the volume of data, and the need to record concealed patterns in very large scale environmental data, there is a rising demand on smart data-oriented approaches that can be utilized to extract patterns in extensive environmental information[9-12].

Recently, machine learning (ML) has become a strong paradigm in the air quality prediction field as it can logically model nonlinear relationships, dynamic data distributions, and operate on high-dimensional inputs. Other supervised learning algorithms like Support Vector Machines, Random Forests, Gradient Boosting Machines, Artificial Neural Networks and Extreme Gradient Boosting have shown higher predictive accuracy than the traditional statistical methods across a wide range of air pollution forecasting problems[13-15]. These models have the potential to use historic levels of pollutants and meteorological variations like temperature, humidity, speed of the wind, direction of the wind and atmospheric pressure to produce the correct short-term and long-term forecasts of air quality. In spite of these, the performance of machine learning models is greatly influenced by the quality of input features, selection of models and best hyper-parameter settings.

Among the major difficulties in air quality prediction, high-dimensional and noisy datasets should be noted. The environmental monitoring systems tend to aggregate a large number of correlated variables at different scale of time and space, which creates redundancy, multicollinearity, and higher level of complexity. Noisy or weakly informative variables may ruin the performance of the model, enhance the risk of overfitting, and decrease the interpretability. Also, machine learning is generally sensitive to hyper-parameter tuning to perform the best generalization. The grid-based search strategies or manual tuning are computationally heavy and usually not efficient to search the solution space especially in complex ensemble and nonlinear models.

Bioinspired optimization algorithms have been of major interest in overcoming these drawbacks and have become powerful tools in the selection of features, hyper-parameter optimization, and model improvement in predictive analytics. They are based on nature and evolution-like processes and behaviors in biology, including swarm intelligence, evolution, and foraging behavior in animals. Some of the popular bio-inspired methods are Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), Grey Wolf Optimizer (GWO), Genetic Algorithms (GA), Ant

Colony Optimization (ACO), and Firefly Algorithm (FA). The stochastic and population-based search algorithms of them make the effective search of the large and nonlinear solution space possible, which allows them to escape local optima and reach globally optimal solutions.

The rest of this paper will be structured in the following manner. Part 2 is a literature review on machine learning-based air quality forecasting and optimization-based modeling methods. Section 3 introduces the suggested bio-inspired optimization scheme, such as the data preprocessing, feature selection, model training, and optimization process. In section 4, the results of the experiments and performance comparisons are discussed. Section 5 summarizes the research and gives the future research directions.

Literature Review

The air quality prediction has received serious research concern because of the growing effect of air pollution on the environment and human health. Initial studies were mostly based on statistical and deterministic atmospheric models, but the dynamics of pollutants and their nonlinear nature have stimulated the transition to machine learning and deep learning techniques. Cabaneros et al. made extensive review of artificial neural network modeling to ambient air pollution prediction, indicating that they are more effective in modelling the nonlinear pollutant relationships with meteorological variables than the conventional regression methods [1]. Although they have a better accuracy, feature redundancy, computational complexity, and generalization related challenges were detected.

As the deep learning technology has developed, the recurrent and hybrid pollution forecasting architectures have been more and more explored. Xing et al. created deep learning models to forecast air quality changes to emission changes, and they were strong models of predicting complex atmospheric interactions based on large datasets [2]. Their result confirmed that deep learning is appropriate in environmental systems but pointed to the sensitivity of hyper-parameter optimization. In a similar manner, He and Guo presented comparative analysis of deep learning models to predict monthly PM2.5, and LSTM-based networks demonstrated higher effectiveness in predicting the time-varying pollution patterns [3]. Nevertheless, a high degree of overfitting and feature dependency was still of concern.

Hybrid models with optimisation methods are a new trend and a growing trend in modelling. According to Wang et al., a hybrid water quality prediction model was developed using Variational Mode Decomposition and LSTM that were optimized with an Improved Grasshopper Optimization Algorithm, and highly improved accuracy was realized [4]. Though concerned with water quality, their results are very strong about the efficiency of optimization-enhanced learning models in environmental forecasting.

Selection of features has been identified as an important parameter that determines the reliability of prediction. Liu et al. created machine learning-based prediction of air quality class using monitoring data and secondary modeling strategies and proved that irrelevant variables worsen the performance of the model [5]. On the same note, Mazuruse et al. implemented algorithmic optimization methods to develop better air quality prediction models of the Shillong City, India with improved forecasting performance and strength using optimization-based learning frameworks [6].

Recent works have explicitly applied bio-inspired optimization algorithms to concurrently optimizing features as well as hyper-parameters. The framework of a hybrid approach of Binary Particle Swarm Optimization and Binary Whale Optimization Algorithm approach to predicting air quality was proposed by Sawah et al., which demonstrated excellent results in the elimination of redundant features and the optimal tuning of model parameters [7]. Their findings supported that the use of hybrid optimization improves the machine learning generalization dramatically.

Similarly, Sudha et al. also suggested a pyramid dilated and optimal weighted feature selection adaptive residual Bi-LSTM architecture to predict air quality [8]. Their model was able to nature multi-scale temporal dependencies well and showed better predictive accuracy as compared to traditional deep learning methods, which determined the critical role of optimized feature weighting mechanisms.

Hybrid learning frameworks are also supported by the development of sophisticated optimization algorithms. Pan et al. proposed Gannet Optimization Algorithm, a new metaheuristic that has a high level of global search and a quick convergence rate on engineering optimization problems [9]. These current algorithms provide some potential solutions in terms of feature selection and hyper-parameter optimization in air quality prediction.

In other high-dimensional areas, besides the environmental modeling, optimization-driven learning frameworks have performed very well. Bilal et al. used quantum computational strategies into extreme learning machines to perform multi-cancer detection, which has a significant enhancement of predictive accuracy [10]. They find that their findings support the wider relevance of optimization-enhanced machine learning to both noisy and complex datasets properties that air pollution data possesses.

Methodology

The paper presents a single bio-inspired optimization-based learning framework to improve the quality of air quality prediction via joint feature selection and hyperparameter optimization. In contrast to the traditional methods which analyze these tasks separately, the proposed one constructs the air quality prediction as a joint optimization, making it possible to effectively process high-dimensional environmental information and enhance the extrapolation of the model.

The main novelty is the capability to use swarm-based metaheuristic optimization with machine learning predictors to find the most informative pollutant-meteorological variables and best learning parameters automatically within a single adaptive search procedure.

Dataset Utilization

The test analysis will use the UCI Air Quality Dataset which has hourly readings of pollutants as well as weather factors. The purpose of the output is the Air Quality Index (AQI), which is based on the standardized levels of pollutants. Multivariate pollutant indicators and climate factors are the only ones that are retained to evaluate the effectiveness of a high-dimensional optimization.

The main characteristics that will be employed in this study are003A

- PM2.5 concentration (μg/m³)
- PM10 concentration (μg/m³)
- Carbon monoxide (CO) (mg/m³)
- Nitrogen dioxide (NO_2) (μg/m³)
- Sulfur dioxide (SO_2) (μg/m³)
- Ozone (O_3) (μg/m³)
- Temperature (°C)
- Relative humidity (%)
- Wind speed (m/s)
- Atmospheric pressure (hPa)

The dependent variable is the Air Quality Index (AQI) which is calculated considering standardized pollutant concentration levels, as set out by environmental bodies. The dataset has more than 9,000 hourly observations, which is a sufficient temporal variation in model learning.

Problem Formulation

Assume that the data set is represented as:

$$D = \{(X_i, y_i)\}_{=1i}^{N}, \quad X_i = [x_1, x_2, \ldots, x_m]$$

where m denotes pollutant and meteorological features and is AQI.

The optimization objective is to determine:

- Optimal feature subset $F^* \subseteq \{1, \ldots, m\}$
- Optimal hyperparameter vector H*

Such that prediction error is minimized:

$$(F^*, H^*) = argminL(f(X_F, H), y)$$

L is the aggregate loss, and f (.) represents the machine learning model.

Bio-Inspired Co-Optimization Strategy Proposal

The suggested algorithm coded feature selection and hyper-parameter tuning into a single solution vector allowing joint optimization a major breakthrough of available sequential ones.

The candidate solutions are defined as: $S = [b_1, b_2, \ldots, b_m, h_1, h_2, \ldots, h_k]$

where:

- $b_i \in \{0,1\}$ indicates feature inclusion
- h_j represents continuous hyperparameters of the learning model

This hybrid binarycontinuous representation is permitting:

- auto-dimensionality reduction
- learning dynamics that are optimize
- minimized computing load.

Adaptive Bio-Inspired search mechanism:

The framework uses a multi-phase adaptive optimization process as opposed to using one optimizer.

Phase 1: Global Exploration

Whale Optimization Algorithm (WOA) actively utilizes extensive solution spaces in order to prevent early convergence.

Phase 2: Balanced Search

Grey Wolf Optimizer (GWO) narrows down potential regions with the help of hierarchical leadership modeling.

Phase 3: Local Exploitation

The combinations of feature-hyper-parameters are optimized to convergence using the Process of Particle Swarm Optimization (PSO).

This cascading search strategy is a key novelty that:

- accelerates convergence
- improves stability
- avoids local minima

Fitness Function Design (Innovation-Oriented)

In order to have balanced accuracy and generalization, a multiobjective fitness function is proposed:

$$Fitness = w1 \cdot RMSE + w2 \cdot MAE + w3 \cdot \frac{|F|}{m}$$

where:

- first two terms guarantee accuracy of prediction.
- third takes a penalty on overuse of features.

This directly implements tight and understandable models in contrast to older error-only optimization.

Algorithm:1 Bio-Inspired Co-Optimized Air quality Prediction Framework

Input: Dataset D, feature set X, ML model f, population size P, iterations T

Output: Optimized model f*

1. Initialize population randomly
2. Evaluate fitness for each
3. for t = 1 to T do
4. Apply WOA for global exploration
5. Apply GWO for adaptive convergence
6. Apply PSO for fine optimization
7. Update feature-hyperparameter vectors
8. Recompute fitness values
9. end for
10. Select best solution S*
11. Train final ML model using optimized F*, H*

The Benefit of Computational Complexity

As compared to grid or random search methods which have exponential complexity:

$O(n^k)$

the suggested framework works on the basis of:

$O(PxT)$

ensuring that it is scalable to large environmental datasets.

Findings and Performance Review

The given section provides the results of the experiment of the bioinspired co-optimized machine learning framework to predict air quality. The comparison is based on the analysis of prediction quality, model size, convergence speed, and stability in comparison with traditional machine learning models and single-optimization strategies.

Training all input features with manually-set hyper-parameters was used to set reference performance with conventional machine learning models. Table 1 shows the performance of the baseline models.

Table 1 Performance of Baseline Models

Model	RMSE	MAE	R^2
Random Forest	14.62	10.83	0.861
SVR	16.48	12.71	0.823

Gradient Boosting	13.91	10.24	0.878
XGBoost	13.42	9.87	0.885

The findings show that ensemble models are better than SVR because they can obtain the nonlinear interactions of pollutants. Nevertheless, the error values are also rather high, and it reflects the limitations due to redundant features and the lack of an optimal tuning process.

The suggested framework also combined feature subsets and hyperparameters optimization through multi-phase bio-inspired search strategy. The table 2 displays the performance.

Table 2 After Proposed Optimization Framework Performance

Model	Selected Features	RMSE	MAE	R^2
RF + Proposed	6	9.38	6.94	0.936
SVR + Proposed	5	10.11	7.62	0.921
GBM + Proposed	6	8.92	6.58	0.942
XGBoost + Proposed	5	8.41	6.13	0.951

The optimalized models obtained:

- RMSE reduction of 35%-40%
- MAE reduction 38% exceeding
- Significant R2 improvement

It is worth noting that the optimized XGBoost model had the best prediction accuracy and only half of the original features were used.

In order to determine the effectivity of the suggested multiphase hybrid optimizer, the comparison of models optiized with the help of single algorithms was conducted as in table 3.

Table 3 Comparison of Optimization Strategy (XG-Boost Model)

Optimization Method	RMSE	MAE	R^2	Features
PSO Only	9.76	7.42	0.928	7
WOA Only	9.51	7.18	0.932	6
GWO Only	9.32	7.03	0.935	6
Proposed Hybrid	8.41	6.13	0.951	5

With hybrid optimization strategy, it was always better than single-optimizers because it could attain:

- lower error values
- faster convergence
- smaller feature subsets

giving an advantage to adaptive multi-phase search.

The suggested algorithm had a high dimensionality reduction without reducing the accuracy.

Table 4 Feature Reduction Analysis

Original Features	Selected Features	Reduction (%)
10	6	40

10	5	50
10	6	40
10	5	50

It proves that in case of intelligent optimization, compact feature representations are effective to predict pollutants.

The hybrid optimization model was also found to converge quicker than the single algorithms and terminated within 40-50 iterations, whereas PSO and WOA took more than 80 iterations to converge. This fastens convergence saves on training cost and it allows deployment in real time to be possible.

Mean gain made by the suggested framework over the baseline models as in table 6

Table 6 Mean improvement of the proposed framework

Metric	Improvement (%)
RMSE	38.6
MAE	41.2
R^2	8.9

These improvements prove that joint optimization is a significant improvement of learning performance.

The results validate that:

- Duplicated atmospheric characteristics severely deteriorate baseline ML models.
- Compact and highaccuracy predictors are generated using joint featurehyperparameter optimization.
- Premature convergence is prevented in multiphase bioinspired search.
- Ensemble learners are optimized to give the highest AQI forecasting.

The proposed framework has a high accuracy, and its computational complexity is lower than in the recent studies carried out with the focus on optimization.

Conclusion

This work introduced a new bio-inspired optimization-based machine learning architecture to the truthful and effective forecasting of air quality, overcoming essential drawbacks of the traditional data-driven forecasting models in terms of the high-dimensional feature space and sub-optimal hyper-parameter settings. In contrast to classical methods, which conduct the feature selection and model-tuning processes independently, the suggested framework modeled air quality prediction as a co-optimization, that is, one can identify informative pollutant-meteorological features and optimum learning parameters simultaneously in a single adaptive search. The application of a multi-phase hybrid hybridization of Whale Optimization Algorithm, as well as the Grey Wolf Optimizer and Particle Swarm Optimization, has had a major positive effect on the exploration efforts around the globe, the stability of convergence, and the refinement in the local region. This cascading search algorithm was very successful to prevent early convergence and at the same time using rapid optimization of complex solution spaces. The performance of the machine learning models, as exhibited by the experimental results of real-world air quality data, showed significant improvements in performance across the entire set of models considered. Specifically, the best predictor based on XGBoost attained the smallest prediction errors and the largest coefficient of

determination using only a small set of features, which validates the usefulness of the suggested dimensionality-sensitive optimization approach. The quantitative analysis showed that the reduction of errors was over 35 percent compared to the baseline models and gradual improvement in performance compared to the use of single-optimizers. Also, the framework was found to reduce features up to 50 percent or more without loss of prediction, providing better model understandability and scalability. These results underscore the fact that intelligent co-optimization is not only able to improve the accuracy of the forecasts, but also can be used to generate scalable models to be used in real-time systems of environmental monitoring. The proposed method has a better generalization ability than the recent optimization-enhanced air quality prediction studies in terms of adaptive multiphase search and explicit complexity inhibition in the fitness formulation. The framework is versatile and can be applied to other environmental forecasting activities including water quality control, climate anomaly control and pollution source control. Further research will involve when developing spatiotemporal deep learning models the inclusion of the proposed optimization framework to learn the propagation of pollution patterns in the region.

References

1. S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," Environ. Model. Softw., vol. 119, pp. 285–304, 2019.
2. J. Xing et al., "Deep learning for prediction of the air quality response to emission changes," Environ. Sci. Technol., vol. 54, pp. 8589–8600, 2020.
3. Z. He and Q. Guo, "Comparative analysis of multiple deep learning models for forecasting monthly ambient PM2.5 concentrations," Atmosphere, vol. 15, no. 12, p. 1432, 2024.
4. Z. Wang, Q. Wang, and T. Wu, "A novel hybrid model for water quality prediction based on VMD and IGOA optimized LSTM," Front. Environ. Sci. Eng., vol. 17, no. 7, p. 88, 2023.
5. Q. Liu, B. Cui, and Z. Liu, "Air quality class prediction using machine learning methods based on monitoring data and secondary modeling," Atmosphere, vol. 15, no. 5, p. 553, 2024.
6. G. Mazuruse et al., "Algorithmic optimization for efficient air quality prediction models through machine learning: A case study of Shillong City in India," Next Research, vol. 2, no. 2, p. 100346, 2025.
7. M. S. Sawah, H. Elmannai, and A. A. El-Bary, "Improving air quality prediction using hybrid BPSO with BWAO for feature selection and hyperparameters optimization," Sci. Rep., vol. 15, p. 13176, 2025.
8. R. Sudha, A. Damodaran, and G. Manohar, "Enhanced air quality prediction using adaptive residual Bi-LSTM with pyramid dilation and optimal weighted feature selection," Sci. Rep., vol. 15, p. 30428, 2025.
9. J.-S. Pan, L.-G. Zhang, R.-B. Wang, V. Snasel, and S.-C. Chu, "Gannet optimization algorithm: A new metaheuristic algorithm for solving engineering optimization problems," Math. Comput. Simul., vol. 202, pp. 343–373, 2022.
10. A. Bilal et al., "Quantum computational infusion in extreme learning machines for early multi-cancer detection," J. Big Data, vol. 12, no. 1, pp. 1–48, 2025.
11. H. R. Naqvi, M. Datta, G. Mutreja, M. A. Siddiqui, D. F. Naqvi, and A. R. Naqvi, "Improved air quality and associated mortalities in India under COVID-19 lockdown," Environmental Pollution, vol. 268, Art. no. 115691, 2021.
12. T. Le, Y. Wang, L. Liu, J. Yang, Y. L. Yung, G. Li, and J. H. Seinfeld, "Unexpected air pollution with marked emission reductions during the COVID-19 outbreak in China," Science, vol. 369, pp. 702–706, 2020.

13. Z. Wang, Q. Wang, and T. Wu, "A novel hybrid model for water quality prediction based on VMD and IGOA optimized for LSTM," Frontiers of Environmental Science & Engineering, vol. 17, no. 7, p. 88, 2023.
14. A. Jufriansah, A. Khusnani, Y. Pramudya, N. Sya'bania, K. T. Leto, H. Hikmatiar, and S. Saputra, "AI big data system to predict air quality for environmental toxicology monitoring," Journal of Novel Engineering Science and Technology, vol. 2, no. 1, pp. 21–25, 2023, doi: 10.56741/jnest.v2i01.314.
15. S. Zhu, X. Lian, H. Liu, J. Hu, Y. Wang, and J. Che, "Daily air quality index forecasting with hybrid models: A case in China," Environmental Pollution, vol. 231, pp. 1232–1244, 2017, doi: 10.1016/j.envpol.2017.08.069.