Global Air Quality: Assessing the Impact of Pollution on Health and the Environment

Hongyun Liu[1] and Yuehua Bai[2,*]

[1]*School of Marixism, Jilin Agricultural Science and Technology College, Jilin 132000, Jilin, China*
[2]*School of Management, Wuhan Donghu University, Wuhan 430212, Hubei, China*
E-mail: liuhongyun888@jlnku.edu.cn; baiyuehua2024@163.com
[*]*Corresponding Author*

Received 12 April 2025; Accepted 06 January 2026

Abstract

The article proposes improvements to Machine Learning (ML) algorithms, specifically Stochastic Gradient Descent (SGD) and K K-nearest neighbour Classification (KNNC), with two improved optimization techniques: Arithmetic Optimization Algorithm (AOA) and Manta Ray Foraging Optimization (MRFO). The two nature-inspired optimization algorithms are used to tune the hyperparameters of SGD and KNNC toward the goals of maximizing prediction precision and computational cost reduction. SGD, being one of the popular algorithms used for loss function minimization in ML, is typically susceptible to careful fine-tuning of its parameters to prevent low convergence, among other problems, as well as overfitting. Likewise, while KNNC is cherished for convenience alongside performance in a majority of applications, well-tuned parameter values can substantially enhance its classification accuracy. AOA and MRFO, inspired by nature and the behavior of animals, present novel concepts for hyperparameter space exploration with better optimizing in a style than typical styles. Experience confirms that SGD and KNNC models work quite effectively about efficiency gains on various diversified datasets based on these optimization approaches. The study

Strategic Planning for Energy and the Environment, Vol. 45_1, 263–290.
doi: 10.13052/spee1048-5236.45110
© 2026 River Publishers

highlights the value of applying bio-inspired algorithms in ML processes that offer a flexible framework to address tough classification problems in diverse fields. SGD was the superior model since it showed superior accuracy, with the capability to handle enormous as well as intricate datasets with great precision, and also recovered complex patterns at lesser error rates. On the contrary, although KNNC was very accurate for small and localized datasets, it did not perform well with large and complex ones, thereby being the weaker model herein. All these findings validate the strength as well as effectiveness of SGD in handling different as well as sophisticated datasets, especially in air pollution prediction.

Keywords: global air pollution, stochastic gradient descent, k nearest neighbour classification, arithmetic optimization algorithm, manta ray foraging optimization.

1 Introduction

A complex environmental issue, air pollution is caused by a variety of chemical, physical, and biological processes that change the natural composition of the atmosphere [1]. The growing urbanization, industrialization, and consumption of fossil fuels have brought air pollution into the forefront as a major environmental concern, especially for the world in development. Recent global events have also influenced pollution levels significantly. For instance, during the COVID-19 pandemic, strict confinement measures led to a temporary but notable reduction in urban air pollution levels [2]. This highlights how human activity patterns directly affect air quality trends and reinforces the importance of accurate and adaptable predictive models. Aside from its detrimental impact on the environment and ecosystems, air pollution has a direct impact on the physical well-being of people and the quality of their lives, resulting in increased rates of disease and mortality. In attempting to avoid this problem, constant monitoring, assessment, planning, and speedy action are needed [3–5]. The most common tropospheric pollutants include particulate matter (PM), carbon monoxide (CO), Nitrogen Oxides (NOx), formaldehyde (HCHO), Ozone (O_3), Carbon Dioxide (CO_2), Sulfur dioxide (SO_2), as well as Carbon monoxide (CO). These pollutants pose a variety of risks to human beings [6, 7]. These days, energy consumption as well as its impact are inevitable in human life. Man-made air pollution may be caused by a variety of factors such as aerosol containers, automobiles, aircraft, kerosene, coal, as well as burning of straws. Numerous dangerous pollutants

are released into the environment on a daily basis, such as CO, CO_2, PM, NO_2, SO_2, O_3, NH_3, Pb, among others [8–11].

Animals, plants, and people are all impacted by the chemicals and particles that cause air pollution. Humans are susceptible to a wide range of dangerous illnesses brought on by air pollution, including lung cancer, heart disease, pneumonia, and bronchitis [12–14]. Other modern environmental problems including acid rain, global warming, smog, impaired visibility, aerosol production, climate change, and early mortality are all caused by poor air quality. Scientists have realized that air pollution has the potential to negatively impact historical sites [15, 16]. Greenhouse gas emissions are caused by several sources, such as factory and power plant emissions, vehicle emissions, and agricultural exhausts. Greenhouse gases have a negative impact on the climate, which in turn has an impact on plant development. The literature has used a variety of models, including statistical, deterministic, physical, and machine learning (ML) models. The conventional methods, which rely on statistics and probability, are less effective and very complicated. It has been demonstrated that the ML-based Air Quality Index (AQI) prediction models are more dependable and consistent [17, 18]. Data collection was made simple and accurate by sophisticated technology and sensors. Because there are so many environmental data points, making accurate and trustworthy predictions requires thorough analysis, which only ML algorithms are capable of handling. The significance of supervised ML methods for real-world environmental protection problems was covered [19, 20]. Prior to using data visualization techniques to explore hidden patterns and trends and get deeper insights, the dataset is preprocessed and cleansed. Few academics have used the correlation coefficient's essential properties with ML models, but this work does [21, 22]. contrasted twenty distinct literary works about the pollutants examined, the ML techniques used, and each work's performance. The authors discovered that several studies used meteorological information to better correctly forecast pollution levels, including temperature, wind speed, and humidity. They discovered that the boosting and neural network (NN) models performed better than the other top ML techniques [23, 24].

This study introduces a novel approach by applying two recent nature-inspired optimization algorithms (Arithmetic Optimization Algorithm (AOA) and Manta Ray Foraging Optimization (MRFO)) to fine-tune the hyperparameters of Stochastic Gradient Descent (SGD) and K-Nearest Neighbors Classification (KNNC). Unlike existing works that often overlook parameter optimization or focus on complex deep-learning models, this research provides an efficient and interpretable framework for air pollution prediction.

It is the first to employ both AOA and MRFO in this context, offering a comparative analysis that highlights performance differences across data complexities and model types. This contribution addresses a gap in the literature and offers a flexible strategy applicable to similar classification problems in other domains.

1.1 Related Works

Using a deep recurrent neural network (DRNN) reinforced by a unique autoencoder-based pretraining technique, Ong et al. [25] employed the network primarily for time-series prediction. Furthermore, the sparsity of the system was taken into consideration while selecting the sensors inside the DRNN, without lowering the prediction accuracy. This strategy produced more accurate findings than the subpar performance obtained with the noise reduction approach when it came to the forecast of air pollution, specifically for PM2.5 particulate matter concentration.

Four measures of the root-mean-square error (RMSE), precision (P), recall (R), and F measure were utilized to measure the performance. Even though it's built on the LSTM approach, the present work is similar because it uses the same technology (that of RNN). An STDL-based approach to predict air quality was given by Li et al. [26]. Applying a greedy layer-wise strategy to train the stacked autoencoder (SAE) code, intrinsic characteristics of the air quality were obtained. The given model showed temporal consistency during all seasons and was capable of predicting the air quality at each station simultaneously compared to the traditional time-series predicting models. Additionally, a comparison was given to the support vector regression (SVR), autoregression moving average (ARMA), and spatiotemporal artificial neural network (STANN) models. Three performance measures were applied to measure the performance of the model: mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE). This paper is similar in the sense that it employs the same method (RNN) to predict the air quality indices, yet it also addresses the issue of a large number of data points and utilizes an LSTM strategy to enhance network performance.

1.2 Objective of the study

The main goal of this research is to improve upon ML models, namely Stochastic Gradient Descent (SGD) and K Nearest Neighbor Classification

(KNNC), through the implementation of sophisticated bio-inspired algorithms. Through the utilization of the Arithmetic Optimization Algorithm (AOA) and Manta Ray Foraging Optimization (MRFO) algorithms to optimize these models' hyperparameters, the aim is improving accuracy in prediction as well as computational effectiveness. The research tackles issues such as low convergence as well as overfitting. The work further examines these methods in the context of air pollution prediction, showcasing their ability to improve model dependability in environmental sensing. The aim is to apply bio-inspired methods to ML procedures for complex classification issues.

2 Dataset

Air pollution refers to the contamination of the environment by harmful chemical, physical, and biological substances, primarily from household appliances, vehicles, factories, and natural phenomena such as forest fires. The major pollutants involve nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), and particulate matter (PM2.5). These substances produce worrying health outcomes, including respiratory disorders, cardiovascular issues, and death, particularly for children, the elderly, and those with underlying ailments. The dataset offers geolocations for these substances as well as their effect on the quality of the air.

Country: The nation's name

City: The city's name

AQI Value: The city's total AQI

AQI Category: The city's overall classification

CO AQI Value: Carbon Monoxide Air Quality Index (AQI) value for the city.

CO AQI Category: The city's AQI for carbon monoxide

Ozone AQI Value: The city's Ozone Air Quality Index

Ozone AQI Category: The city's AQI-rated ozone

NO2 AQI Value: The city has a NO2 rating on the Nitrogen Dioxide Air Quality Index.

NO2 AQI Category: The NO2 AQI category applies to the city's nitrogen dioxide levels.

PM2.5 AQI Value: PM2.5, or particulate matter with a diameter of 2.5 micrometers or less, is measured by the city's AQI.

PM2.5 AQI Category: Particulate matter in the city having a diameter of 2.5 micrometers or less is classified as PM2.5 in the AQI.

Figure 1 displays marginal histogram plots that illustrate the relationships and correlations between input features and output variables. These plots reveal how variables interact and influence each other, providing insights into their underlying dynamics. For instance, the variable Country shows the highest frequency within the interval of 160, compared to the lower frequency observed in the interval of 120. In contrast, the City variable demonstrates relatively uniform frequencies across the range of 0 to 20,000.

3 Methodology

3.1 Stochastic Gradient Descent (SGD)

Take into consideration the cost function of

$$\begin{gathered}\min \sum_{j:(i,j)\in\Omega} loss(a_{ij}, x_i^T y_j) + \frac{\lambda}{2}\|X\|_F^2 + \frac{\lambda}{2}\|Y\|_F^2 \\ s \times t \times x_I^T y_j \epsilon\{1,-1\},\ \forall_i, \forall_j\end{gathered} \tag{1}$$

About x_i and y_i, its complete (sub)gradients are as follows:

$$G_{x_i} = \sum_{j:(i,j)\in\Omega} \frac{\partial loss(a_{ij}, x_i^T y_j)}{\partial x_i} + \lambda x_i \tag{2}$$

$$G_{y_i} = \sum_{i:(i,j)\in\Omega} \frac{\partial loss(a_{ij}, x_i^T y_j)}{\partial x_i} + \lambda y_i, \tag{3}$$

correspondingly. Note the decoupling of G_{x_i} (or G_{y_i}) across i (or j). Therefore, one easy way to solve (3) is to update all x_i and y_i using a gradient descent method that uses the entire gradients G_{x_i} and G_{y_i}, then project the products $x_i^T\ y_j$ onto the binary set $\{-1,\ 1\}$. To sum up the partial gradients, $\partial loss/\partial x_i$ and $\partial loss/\partial y_j$, must be evaluated for each iteration of the entire gradient descent procedure. When $i : (i,j) \in \Omega$ and $j : (i,j) \in \Omega$ have large cardinalities, this becomes time-consuming.

To address this problem, the stochastic gradient descent method, which substitutes stochastic gradients for the complete gradients, is recommended

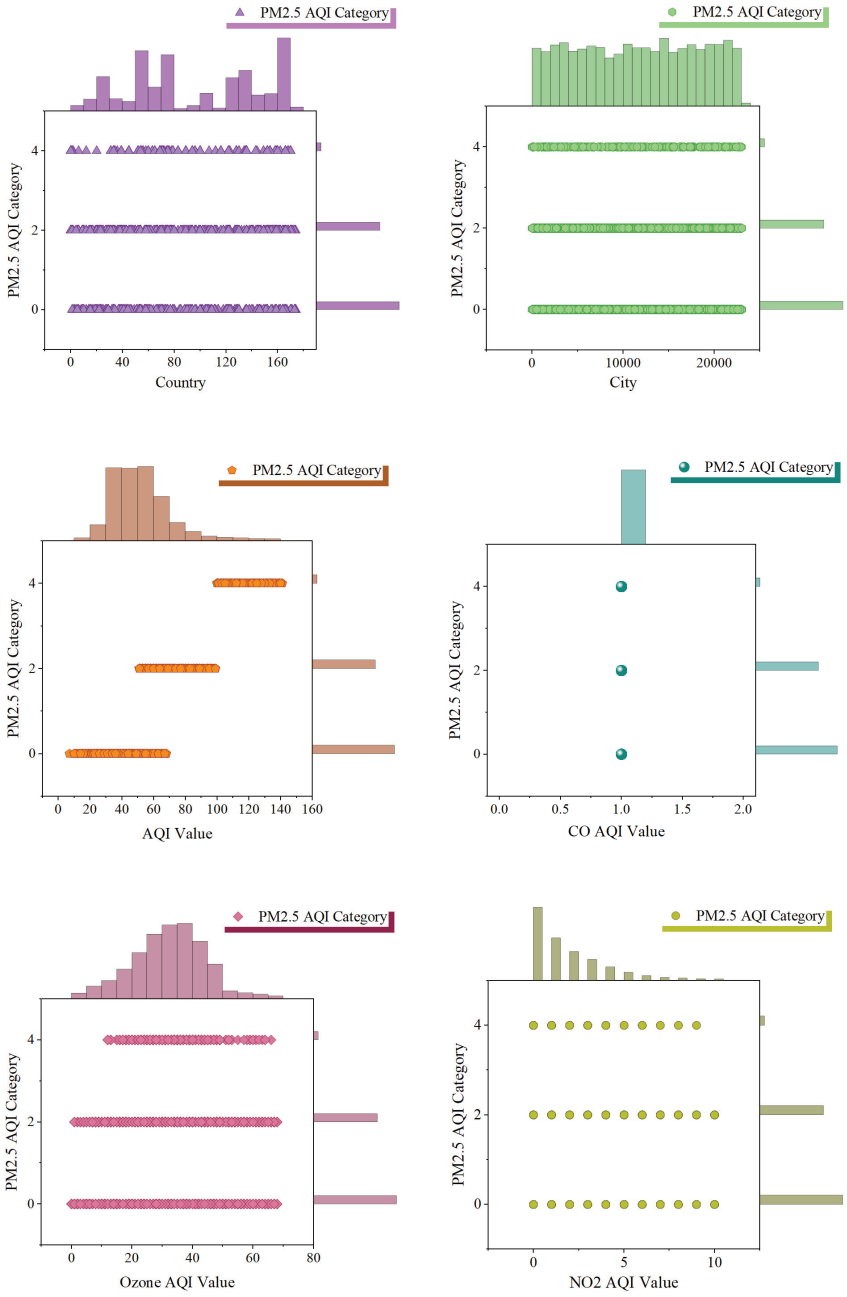

Figure 1 Marginal Histograms plots are used to show the correlation between the input and outputs.

by stochastic approximation and ML theories [27, 28]. Using their stochastic gradients, only a randomly selected pair of x_i and y_j are changed at each iteration:

$$g_{x_i} = \frac{\partial loss(a_{ij}, x_i^T y_j)}{\partial y_j} + \lambda x_i \tag{4}$$

$$g_{y_i} = \frac{\partial loss(a_{ij}, x_i^T y_j)}{\partial y_j} + \lambda y_i \tag{5}$$

The approximations of the complete gradients G_{x_i} and G_{y_i}, respectively, are represented by the stochastic gradients g_{x_i} and g_{y_i}.

As stated in approach 1, the link prediction problem in [29] has been resolved using the stochastic gradient descent approach. Gradient noise, or the difference between the whole gradient and the stochastic gradient, might prevent the iteration from converging if the non-negative step size η isn't set to decrease.

Algorithm 1 stochastic gradient descent (SGD)

Require: training set $\{a_{ij}, (i,j)\ \epsilon\Omega\}$, initial values X and Y While not converged do

Uniformly randomly choose $(i,j)\epsilon\Omega$, and pick a step size η

Calculate stochastic gradients g_{x_i} and g_{y_i} as in (4) and (5)

Update $x_i \longleftarrow x_i - \eta g_{x_i}$ and $y_j \longleftarrow y_j - \eta g_{y_i}$

end while

For all $(i,j) \notin \Omega$, project $x_i^T\ y_j$ onto $\{-1, 1\}$

Completing a large-scale adjacency matrix takes time, even though the stochastic gradient descent technique is a computationally light method of addressing the link prediction issue. It takes a few minutes for a signed social network with 100,000 people to complete the adjacency matrix of ten billion entries using stochastic gradient descent, as the numerical experiments will demonstrate. As a result, the optimization process must be accelerated for applications that require a quick response. Asynchronous distributed stochastic gradient descent implementations will be provided below, utilizing multi-thread computation to drastically shorten execution time.

3.2 *K* Nearest Neighbor Classification (KNNC)

Since it works well for classification tasks and is easy to apply to datasets with many characteristics, the K-Nearest Neighbors technique was used. The use

of KNN was based on its non-parametric nature, which provides the ability to accommodate the underlying patterns in the dataset without assuming a specific distribution. The performance of the model was considered during the utilization of the KNN approach to identify the number of neighbors (k) that is most suitable. Equipped with the Euclidean distance formula, the sample distance was determined [30, 31]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

Where n is the number of qualities and d is the distance between two points, x and y.

5-fold cross-validation was used to ensure the generalizability and reliability of the model. This involved the dataset being split into five individual subsets, each of which was used as a test set one at a time, with the rest of the data used for training. This was done five times to ensure each point of data was used for training as well as for proving the model's performance on theoretical data [32, 33]. Aside from minimizing overfitting, cross-validation provided a better estimation of the model's performance on theoretical data. This method is expressed numerically as:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} M_i \tag{7}$$

If the cross-validation score is $CV_{(n)}$, the number of folds is n and the performance measure for fold i is M_i.

3.3 Arithmetic Optimization Algorithm (AOA)

AOA is a swarm intelligence optimization method that is both simple and effective. This section displays the AOA operating process.

Setting up the population matrix is the initial stage in executing AOA. The job of finding values inside the designated search range is accomplished by using the population matrix.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & \cdots & \ldots & x_{m,n} \end{bmatrix} \tag{8}$$

If $\{x_{1,1}, x_{1,2}, \ldots, x_{1,n}\}$ is an individual from the first population and X is the initial population; as a result, there are m individuals in this population. The individual's dimension is denoted by n.

The user will be able to sort and compute individual fitness values based on their preferences. The optimal individual" is the person whose fitness value is closest to the user's needed value. By the MOA, each person is categorized as either an explorer or an explorer. During the allocation procedure, a random number, $r1$, is chosen between $[0, 1]$. $Once\ r1 < MOA(t)$, the individual moves into the exploration phase; if not, it moves into the exploitation phase. Equation (9) shows the calculation of the MOA.

$$MOA(t) = MOA_{\min} + t \times \left(\frac{MOA_{\max} - MOA_{\min}}{T_{\max}} \right) \tag{9}$$

where t denotes the iteration number that is now in use, $T_{\max}$ denotes the algorithm's final iteration number, and $MOA_{\max}$ and $MOA_{\min}$ stand for the MOA's maximum and minimum values, respectively.

The update function is as follows: if the person moves into the exploring phase:

$$x_{i,j}^{t+1} = \begin{cases} x_{best,j}^{t} \div (MOP + \varepsilon) \times \omega_j \ r_2 \leq 0.5 \\ x_{best,j}^{t} \times MOP \times \omega_j \ r_2 > 0.5 \end{cases} \tag{10}$$

$$\omega_j = (ub_j - lb_j) \times \mu + lb_j \tag{11}$$

where ub_j and lb_j stand for the bounds of the j-dimensional value of an individual; ε is the minimum constant to keep the denominator from being 0; μ is the optimization process control constant, with a value of 0.499; and MOP is given by Equation (12).

$$MOP(t) = 1 - \frac{t^{1/\alpha}}{T_{\max}^{1/\alpha}} \tag{12}$$

where α is the sensitivity coefficient, which is typically fixed at 12.

The update function is: if the person reaches the exploitation phase:

$$x_{i,j}^{t+1} = \begin{cases} x_{best,j}^{t} - MOP \times \omega_j \ r_3 \leq 0.5 \\ x_{best,j}^{t} + MOP \times \omega_j \ r_3 > 0.5 \end{cases} \tag{13}$$

where r_3 is a random value from $[0, 1]$ that is used to determine an individual's update function during the exploitation phase.

3.4 Manta Ray Foraging Optimization (MRFO)

This section outlines the fundamental phases of MRFO. MRFO operates by modeling the three foraging tactics used by manta rays: chain, cyclone, and somersault foraging. In the search space, MRFO creates beginning populations at random, just as other swarm-based metaheuristic algorithms do. After that, the three previously described ways to update it. Below are the corresponding mathematical models for each of these three foraging techniques.

A. CHAIN FORAGING

The manta rays join their heads and tails in a line to create a foraging chain. According to MRFO, increasing the abundance of plankton the manta rays' preferred food is the best course of action. The other individuals in the foraging chain move not just toward food but also toward those who are in front of them, whereas the first person merely travels in the direction of food. The following is a description of the chain foraging mathematical model.

$$x_i^{t+1} = \begin{cases} x_i^t + r_1 \times (x_{best}^t - x_i^t) \\ +\alpha \times (x_{best}^t - x_i^t), \quad i = 1 \\ x_i^t + r_2 \times (x_{i-1}^t - x_i^t) \\ +\alpha \times (x_{best}^t - x_i^t), \quad i = 2, 3, \ldots NP \end{cases} \tag{14}$$

$$\alpha = 2 \times r_3 \times \sqrt{|\log(r_4)|} \tag{15}$$

where the i^{th} individual's location at generation t is denoted by x_i^t. The random vectors $r_i \in [0, 1], i = 1, 2, 3, 4$ are uniformly distributed. It is the optimum individual plankton that has the maximum concentration. Population size is represented by NP. has α weight coefficient of 0.

B. CYCLONE FORAGING

At the bottom of the ocean, manta rays find plankton, which they use to construct lengthy foraging chains as they spiral toward food. Whale optimization algorithm (WOA) toward food and trailing the people in front of it, this behavior is comparable to WOA. The subsequent equation provides the mathematical description of cyclone foraging.

$$x_i^{t+1} = \begin{cases} x_{best}^t + r_5 \times (x_{best}^t - x_i^t) \\ +\beta \times (x_{best}^t - x_i^t), \quad i = 1 \\ x_{best}^t + r_6 \times (x_{i-1}^t - x_i^t) \\ +\beta \times (x_{best}^t - x_i^t), \quad i = 2, 3, \ldots, NP \end{cases} \tag{16}$$

$$\beta = 2 \times \exp(r_7 \times (iter_{\max} - iter + 1)/iter_{\max}) \times \sin(2\pi r_7) \quad (17)$$

where the random vectors $r_i \in [0,1], i = 5,6$ are uniformly distributed. is the coefficient of weight. A random integer with a uniform distribution is $r_7 \in [0,1]$. The maximum number of iterations is denoted by $iter_{\max}$, whereas the current number of iterations is represented by $iter$.

Food is primarily employed in Eq. (17) as a point of reference for spiral foraging, which helps to fully use the area around food. Furthermore, a randomly generated position inside the search space is employed as a spiral foraging reference site to broaden the search area. This enables everyone to look for locations that are distant from their ideal starting point right now. Because the random spiral foraging mechanism primarily prioritizes exploration, MRFO is able to conduct a thorough worldwide search. The particular mathematical model is explained in the following way.

$$x_{rand} = lb + r_8 \times (ub - lb) \quad (18)$$

$$x_i^{t+1} = \begin{cases} x_{rand} + r_9 \times (x_{best}^t - x_i^t) \\ +\beta \times (x_{best}^t - x_i^t), \quad i = 1 \\ x_{rand} + r_{10} \times (x_{i-1}^t - x_i^t) \\ +\beta \times (x_{best}^t - x_i^t), \quad i = 2,3,\ldots,NP \end{cases} \quad (19)$$

where x_{rand} is an arbitrary point generated at random inside the search space. Random vectors with uniform distributions are $r_i \in [0,1], i = 8,9,10$. The $upper$ and $lower$ boundaries of the search space are denoted by ub and lb, respectively.

C. SOMERSAULT FORAGING

The position of the meal is seen as a pivot point at this stage. Every person rotates around the pivot, therefore seeking a new spot. This is how this phase is represented mathematically.

$$x_i^{t+1} = x_i^t + S \times (r_{11} \times x_{best}^t - r_{12} \times x_i^t), \quad i = 1,2,\ldots,NP \quad (20)$$

where S is the somersault factor and $S = 2$ determines the somersault range of manta rays. In [0, 1], there are two random numbers, r_{11} and r_{12}.

With the regulation of $(iter \ / \ iter_{\max})$ change, MRFO governs the behavior of exploration and exploitation. Food sources are created at random as reference points in the search space, and exploratory behavior is mostly carried out when $(iter \ / \ iter_{\max}) < rand$. The technique may be exploited

more easily since the optimal individual is utilized as a reference point when $(iter \ / \ iter_{\max}) < rand$. Additionally, spiral or chain foraging is chosen at random using a random integer. Then, Somersault's foraging takes place.

3.5 Performance Evaluators

Accuracy, defined by True Positives, True Negatives (correctly predicted negative cases), False Positives (incorrectly predicted as positive), and False Negatives (incorrectly predicted as negative), measures the overall correctness of the model by calculating the ratio of correctly predicted instances (both positive and negative) to the total number of instances. Precision focuses on the correctness of positive predictions by measuring the proportion of true positive predictions among all positive predictions made by the model, with True Positives and False Positives being the relevant metrics, and indicates higher precision when there are fewer false positives. Recall measures how well the model can recognize all positive cases by considering the proportion of true positives over all actual positive instances, based on True Positives and False Negatives, and shows greater recall as a result of fewer false negatives. The F1-score, which averages Precision and Recall, offers a measure that strikes a balance on the trade-off between Precision and Recall.

$$Accuracy : \frac{TP + TN}{TP + FP + FN + TN} \tag{21}$$

$$Precision : \frac{TP}{TP + FP} \tag{22}$$

$$Recall : \frac{TP}{TP + FN} \tag{23}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{24}$$

The F1-measure is the harmonic mean of precision and recall, offering a single score that weighs both dimensions. It is particularly useful in considering both false positives as well as false negatives. The higher score for F1 implies a improved correlation between precision and recall.

4 Result and Discussion

Figure 2 demonstrates the effect of feature selection with different input parameters, i.e., country, city, AQI, CO, Ozone, and NO2 AQI values.

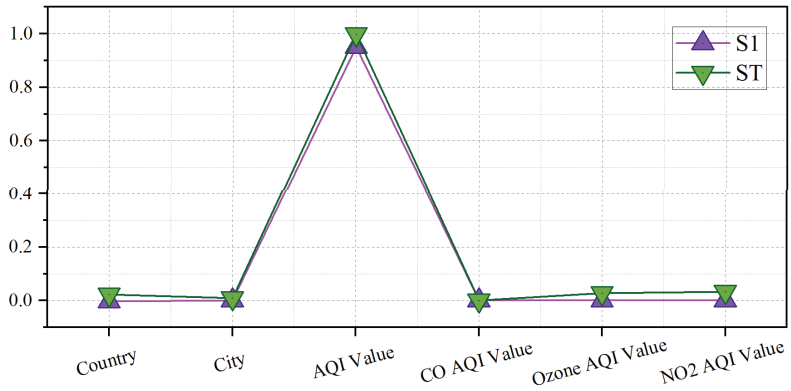

Figure 2 Analysis of feature selection given the input parameters.

The analysis here depicts how feature selection influences model performance regarding increasing accuracy and decreasing complexity. It illustrates that the selection of key features, like SI (Statistical Indicator) and ST (Statistical Test), is essential in improving model performance and also in producing accurate and reliable predictions. The analysis illustrates that AQI values are most frequent at a value of 1 in SI and in ST, indicating their overall dominance on the model. On the other hand, attributes such as country, city, CO, Ozone, and NO2 AQI values are of lower frequencies compared to AQI values. Moreover, ST is more frequent than SI, and this affirmatively indicates its greater relevance when selecting features. This reflects the significance of effective feature selection to enhance model accuracy as well as performance through guaranteeing that the model is effectively calibrated to produce accurate as well as credible predictions.

Figure 3 shows convergence curves, which graphically show the progress of an algorithm's learning process over time or several iterations. The curves reflect the advancement in the model as well as its effectiveness in learning. A well-behaved convergence curve, revealing constant improvement until it peaks, denotes that the model has optimized its performance. Convergence curves are useful for model evaluation and model optimization to ensure they provide the most desirable outputs while minimizing issues. In 200 iterations that follow, the Stochastic Gradient Ascent Optimization (SGAO) algorithm outperforms the Stochastic Gradient Method Optimization (SGMO) with a value of 0.89424 compared to 0.88342 achieved by SGMO. The improvement shows that SGAO is more effective in both learning and optimizing the model under the same number of iterations. Both algorithms show progress, with

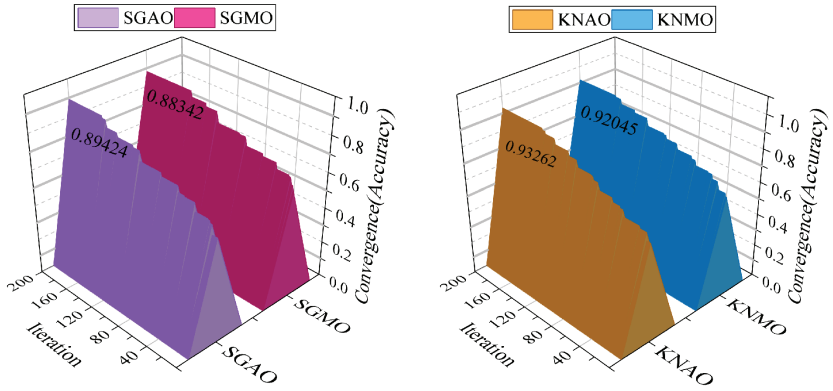

Figure 3 Modeling the convergence curve is the output 3D bars

SGAO's higher value signifying superior convergence as well as possibly more accurate outputs. The minor performance gap reflects SGAO's effectiveness in approaching peak performance faster, thus making it the optimum option for model training as well as evaluation.

Figure 4 shows the ROC curves of the most effective hybrid models, demonstrating their performance during classification. The ROC curve is used to assess model effectiveness by graphing the True Positive rate vs. the False Positive rate at different thresholds. The ROC curve for the Good model has high performance, starting at the origin (0, 0) and demonstrating

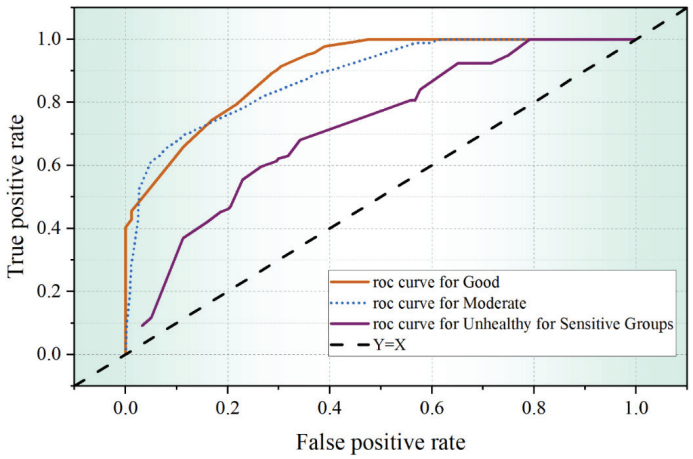

Figure 4 The ROC curves for the performance of the most efficient hybrid models.

a steep increase in the True Positive rate, with a low False Positive rate. It peaks at 0.4 and 1, indicating good performance at all times, with the model's discrimination ability clearly distinguishing classes with minimal missteps. The ROC curve for the Moderate model traces a similar course with slightly less effectiveness. Although it traces the Good model's course, overall effectiveness is a little lower, with less efficiency in reaching the highest True Positive levels. The ROC curve for the Unhealthy for sensitive groups model demonstrates less performance. Although it begins the same and starts making progress, the True Positive rate rises at a slower pace, leading to a low curve. Even at reaching its highest impact, it cannot match the effectiveness of the Good or Moderate models, showcasing its lower classification effectiveness for sensitive populations. Overall, the Good model has less performance followed by the others: the Moderate and then the Unhealthy.

Figure 5 compares measured and predicted air quality values across three categories: Moderate, Good, and Unhealthy for Sensitive Groups. The measured data show a higher count in the "Good" category, indicating generally better air quality. Predictions from various models (SGAO, SGMO, SGDC, KNAO, KNMO, and KNNC) align mostly with the measured values, though some discrepancies exist. In the "Moderate" category, the KNAO model predicts the highest values, while KNNC forecasts the lowest, with SGDC providing intermediate predictions. The "Good" category has fairly uniform predictions concerning measured observations, demonstrating good agreement. The "Unhealthy for Sensitive Groups" category shows substantial differences across models, wherein KNMO and KNNC are predicting smaller values than those observed. It shows that models require fine-tuning for precision in predicting this category and provides areas of improvement.

Figure 6 presents different performance profiles for different air quality categorization models. SGAO performs well in the Unhealthy for Sensitive Groups category with minimal errors but poorly in the Good category. KNAO performs well in the Unhealthy for Sensitive Groups but has higher errors in the remaining categories. SGMO is very good in the Good category, though with a higher general error rate, but performing well on other categories. KNMO performs well on the Good category and is less error-prone on "Unhealthy for Sensitive Groups" but poorer with more error rates on the "Moderate" category. SGDC is excellent, with high precision in the Good and Moderate classes and very marginal errors in the Unhealthy for Sensitive Groups class, and it has good performance. KNNC is also good in the Unhealthy for Sensitive Groups class but with high error in the Good

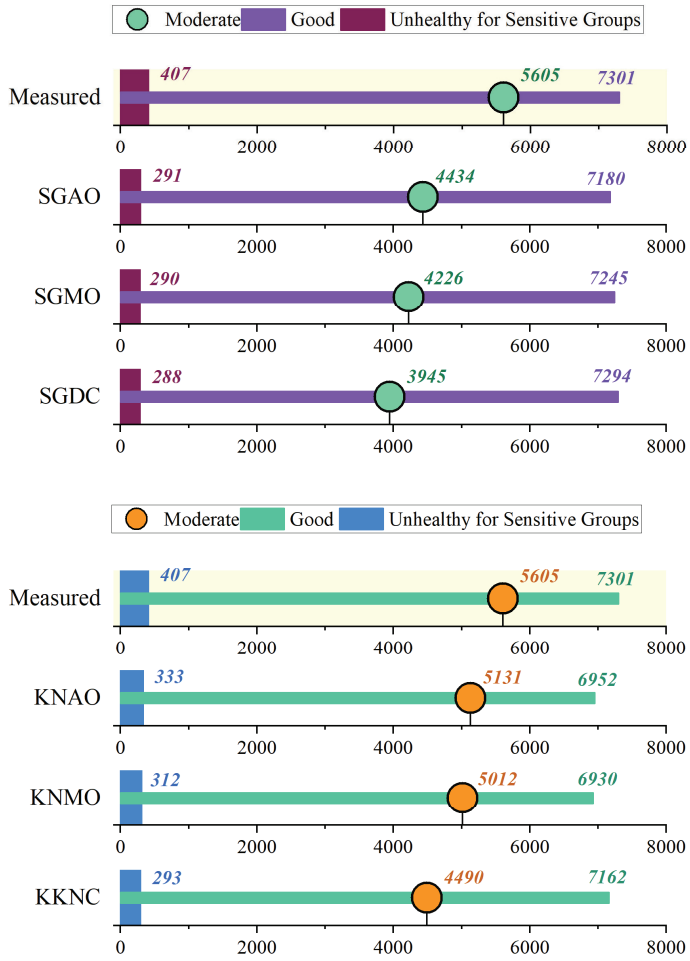

Figure 5 Compare the measured and predicted values.

class. Generally, the comparison indicates both models have strengths and weaknesses in different classes. This outcome is significant for the selection of the most appropriate model according to individual needs of air quality classification and also demonstrates avenues of potential improvement to reach increased overall accuracy and reliability.

Figure 7 shows a comparison of different models on accuracy, precision, recall, and F1-score to understand their performance. The SGDC model starts with good performance with high precision but low recall, which

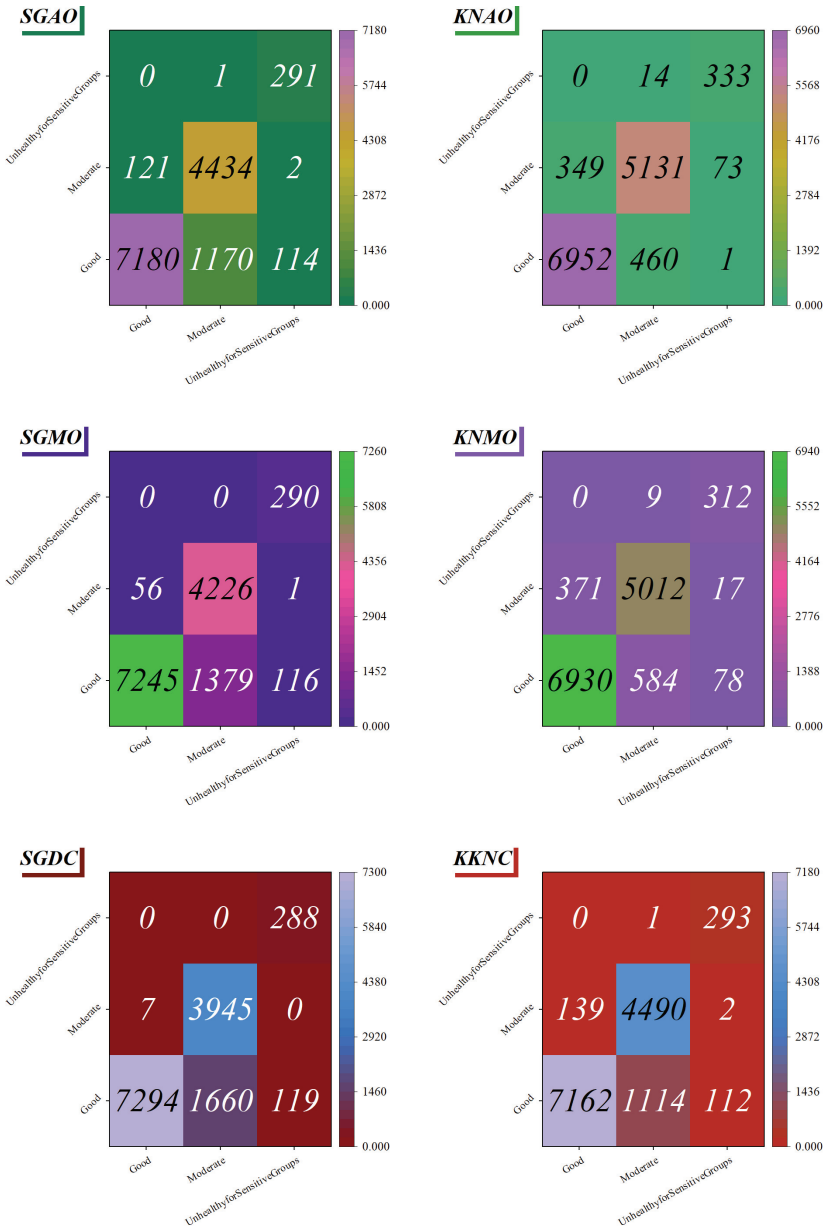

Figure 6 Confusion matrix for the accuracy of each model.

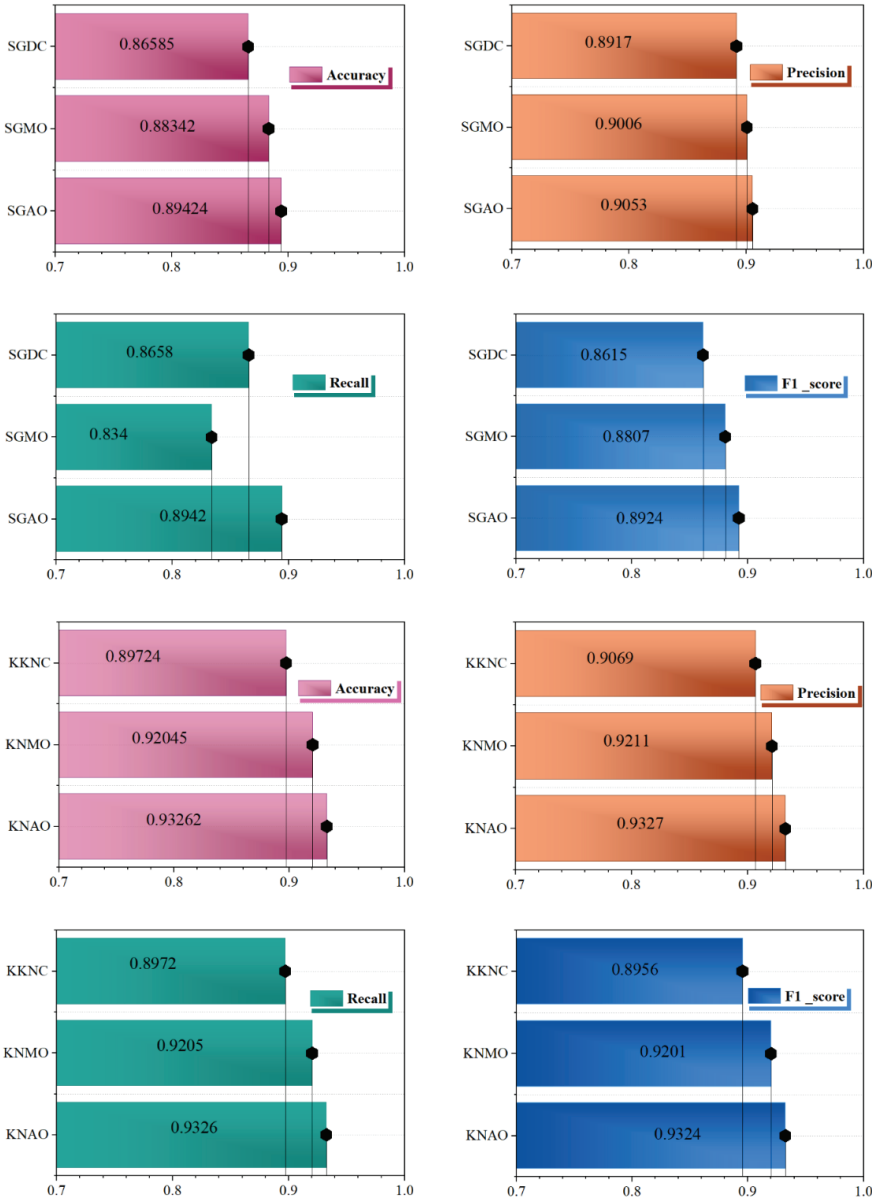

Figure 7 Bar plot used to assess the performance of the models.

translates to a moderately good F1 score. The SGMO model has a minor improvement in accuracy as well as precision over SGDC. Its recall, however, is slightly less, which brings down the overall F1-score such that it is only marginally better than SGDC. The SGAO model takes performance one step higher by improving accuracy, precision, and recall, to get a balanced as well as higher F1 score than SGMO. This shows that SGAO has a better overall performance. The KKNC model continues improving on the previous advancements by SGAO by bringing higher accuracy and precision. The recall also sees a slight improvement, leading to the highest F1-score and demonstrating the KKNC model's performance on all measures. KNMO takes performance to the next level by improving the recall significantly, along with high accuracy and precision. This corresponds to a considerably higher F1 score, indicating a well-balanced model performing well on various evaluation measures. Then the KNAO model is the highest performer, with the highest accuracy, precision, recall, and F1-score. This model is the peak of the optimized models compared here, as it provides the most balanced and better performance on every measure.

Table 1 offers a detailed comparison of performance measures for SGAO, SGMO, SGDC, KNAO, KNMO, KKNC, SGDC and KKNC measuring accuracy, precision, recall, and F1 score at both the training and testing stages. SGDC shows strong performance at the training stage with 0.9007 accuracy as well as high precision, recall, and F1 measure values of 0.9159, 0.9007, and 0.8987, respectively. At the testing stage, its performance decreases significantly, with accuracy decreasing to 0.7844 alongside lower precision, recall, and F1 measure values of 0.8428, 0.7844, and 0.7686. This dramatic reduction indicates that SGDC has possibly overfitted the training set, thus encountering difficulty in generalization to new, unseen observations.

Conversely, KKNC performs well in both the training and test phases. It has a high training accuracy of 0.9238 along with a precision, recall, and F1

Table 1 Result of presented model for SGDC and KKNC

	Section/Metric							
	Train				Test			
Model	Accuracy	Precision	Recall	F1_score	Accuracy	Precision	Recall	F1_score
SGAO	0.9221	0.9287	0.9221	0.9212	0.8942	0.9053	0.8942	0.8924
SGMO	0.9136	0.9235	0.9136	0.9123	0.8130	0.8521	0.813	0.8034
SGDC	0.9007	0.9159	0.9007	0.8987	0.7844	0.8428	0.7844	0.7686
KNAO	0.9503	0.9503	0.9503	0.9502	0.8913	0.8914	0.8913	0.891
KNMO	0.9417	0.9423	0.9417	0.9415	0.8708	0.8716	0.8708	0.87
KKNC	0.9238	0.9297	0.9238	0.923	0.8353	0.8557	0.8353	0.8301

score of 0.9297, 0.9238, and 0.9230, respectively. KKNC demonstrates effective learning. A slight decrease in performance on the test set is observed, as it has a precision, recall, and F1 score of 0.8557, 0.8353, and 0.8301, with accuracy at 0.8353. The smaller performance difference between stages in the case of KKNC reflects improved generalization as well as stability over different data sets. Overall, SGDC has superior performance during the training stage, yet a very poor performance during the test stage, indicating possible generalization problems. KKNC, however, provides more reliable as well as stable performance in both stages, thus proving a better model for real-time applications. This analysis demonstrates KKNC's improved performance balance as regards effective training as well as generalization ability.

Table 2 illustrates a comparison of SGAO, SGMO, and SGDC performance measures across different grades: Good, Moderate, and Unhealthy for Sensitive Groups. SGAO achieves high precision as well as recall in the Good class, offering strong performance. The precision of SGAO remains unaffected, but the recall is low in the Moderate grade, thus negatively impacting the overall F1 score. In the "Unhealthy for Sensitive Groups" classification, SGAO has good precision as well as low recall, leading to a low F1 score as compared to other grades. SGMO offers strong performance in the Good class with good precision as well as recall. SGMO has good precision in all classes, though recall is low in the classification of Moderate as well as Unhealthy for Sensitive Groups. Notwithstanding this, the precision of SGMO is good, ensuring good classification. SGDC has a strong performance in the Good grade with perfect recall as well as good precision. The Moderate grade

Table 2 Evaluation indexes of the developed models' performance based on grades in SGDC

		Index values		
Model	Grade	precision	recall	F1-score
SGAO	Good	0.8483	0.9834	0.9109
	Moderate	0.973	0.7911	0.8727
	Unhealthy for Sensitive Groups	0.9966	0.715	0.8326
SGMO	Good	0.8289	0.9923	0.9033
	Moderate	0.9867	0.754	0.8548
	Unhealthy for Sensitive Groups	1	0.7125	0.8321
SGDC	Good	0.8039	0.999	0.8909
	Moderate	0.9982	0.7038	0.8256
	Unhealthy for Sensitive Groups	1	0.7076	0.8288

shows very good precision with low recall, which leads to a low F1 score. For the "Unhealthy for Sensitive Groups" class, SGDC has good precision with a little decrease in recall, leading to performance similar to that of SGMO. Overall, though all models show good precision, SGMO and SGDC ensure good performance in different classes, with SGDC ensuring good precision as well as recall.

Table 3 illustrates the performance measures of KNAO, KNMO, and KNNC by three different evaluation grades: Good, Moderate, and Unhealthy for Sensitive Groups. All models demonstrate high competence with performance differences across grades. In the Good grade, all models KNAO, KNMO, and KNNC demonstrate high precision as well as recall, indicating high-quality performance with high F1 scores. KNAO and KNMO perform very well, with KNNC performing well but slightly lower in F1-score compared to the other two. When the grade changes to Moderate, KNAO and KNMO demonstrate good precision, with a decline in recall, leading to a decrease in F1 scores. KNNC demonstrates high precision but suffers a drastic decline in recall, leading to a lower F1 score compared to its performance in the Good grade. In the category of Unhealthy for Sensitive Groups, all models demonstrate high precision, with a decline in recall. KNAO and KNMO demonstrate a high decline in recall, leading to a decline in their F1 scores. KNNC also demonstrates a decline in recall, leading to a similar effect on its F1 score. Overall, although KNAO, KNMO, and KNNC all demonstrate high performance in the Good category, they suffer varying levels of performance decline in the Moderate as well as the Unhealthy for Sensitive Groups categories.

Table 3 Evaluation indexes of the developed models' performance based on grades in KKNC

		Index values		
Model	Grade	precision	recall	F1-score
KNAO	Good	0.9378	0.9522	0.945
	Moderate	0.924	0.9154	0.9197
	Unhealthy for Sensitive Groups	0.9597	0.8182	0.8833
KNMO	Good	0.9128	0.9492	0.9306
	Moderate	0.9281	0.8942	0.9109
	Unhealthy for Sensitive Groups	0.972	0.7666	0.8571
KNNC	Good	0.8538	0.981	0.913
	Moderate	0.9696	0.8011	0.8773
	Unhealthy for Sensitive Groups	0.9966	0.7199	0.8359

5 Conclusion

GAP analysis using Stochastic Gradient Descent (SGD) and K-nearest Neighbors Classification (KNNC) produced significant findings. SGD showed high accuracy and a strong ability to process large and complex datasets, learning intricate patterns with lower error rates and improved model performance. In contrast, KNNC demonstrated excellent precision when applied to smaller, localized data, particularly in identifying specific subsets with high accuracy. The use of optimization algorithms played a central role in these results. Arithmetic Optimization Algorithm (AOA) effectively fine-tuned SGD parameters, leading to higher accuracy, while Manta-Ray Foraging Optimization (MRFO) enhanced KNNC's generalization by reducing classification errors. Beyond performance metrics, this study offers a novel contribution by introducing a comparative framework where two bio-inspired optimization techniques are applied to two distinct ML models for environmental prediction tasks. This dual-optimization approach provides not only a practical improvement in prediction outcomes but also a scalable and interpretable method that can extend to similar classification problems in other domains. The findings reinforce the importance of model selection and parameter tuning in environmental modeling, particularly for applications such as air pollution prediction. Overall, SGD emerged as the more robust model, with a training accuracy of 0.9007 and a test accuracy of 0.7844, while KNNC showed limitations on large and complex datasets despite a higher training accuracy of 0.9238 and a test accuracy of 0.8353.

5.1 Policy Implications

The findings highlight the importance of using optimized machine learning models in environmental monitoring and decision support. Accurate air pollution prediction can help policymakers allocate resources more efficiently, implement timely interventions, and evaluate the impact of existing regulations. The demonstrated superiority of SGD for large-scale data suggests its suitability for nationwide monitoring systems, while KNNC may be useful for localized or community-based assessments. Additionally, the use of nature-inspired optimization techniques shows potential in building cost-effective, adaptable, and interpretable predictive systems that align with sustainable development goals. These insights can assist policymakers in adopting data-driven strategies for air quality management and long-term climate planning.

References

[1] G.-J. Liu, E.-J. Fu, Y.-J. Wang, K.-F. Zhang, B.-P. Han, and C. Arrowsmith, "A framework of environmental modelling and information sharing for urban air pollution control and management," *Journal of China University of Mining and Technology*, vol. 17, no. 2, pp. 172–178, 2007.

[2] A. G. Progiou, I. Sebos, A.-M. Zarogianni, E. M. Tsilibari, A. D. Adamopoulos, and P. Varelidis, "Impact of covid-19 pandemic on air pollution: the case of athens, greece.," *Environmental Engineering & Management Journal (EEMJ)*, vol. 21, no. 5, 2022.

[3] S. Al-Kallas, M. Al-Mutairi, H. Abdel Basset, A. Abdeldym, M. Morsy, and A. Badawy, "Climatological study of ozone over Saudi Arabia," *Atmosphere (Basel)*, vol. 12, no. 10, p. 1275, 2021.

[4] E. Dons *et al.*, "Concern over health effects of air pollution is associated to NO 2 in seven European cities," *Air Qual Atmos Health*, vol. 11, pp. 591–599, 2018.

[5] M. Filonchyk, "Characteristics of the severe March 2021 Gobi Desert dust storm and its impact on air pollution in China," *Chemosphere*, vol. 287, p. 132219, 2022.

[6] D. L. Crouse *et al.*, "Ambient PM2. 5, O3, and NO2 exposures and associations with mortality over 16 years of follow-up in the Canadian Census Health and Environment Cohort (CanCHEC)," *Environ Health Perspect*, vol. 123, no. 11, pp. 1180–1186, 2015.

[7] M. E. Emetere *et al.*, "Indoor Air Pollution: A Review on the Challenges in Third World Countries," *Air, Soil and Water Research*, vol. 17, p. 11786221241239892, 2024.

[8] E. Tsepi, I. Sebos, and G. L. Kyriakopoulos, "Decomposition Analysis of CO 2 Emissions in Greece from 1996 to 2020.," *Strategic Planning for Energy & the Environment*, vol. 43, no. 3, 2024.

[9] V. Bozoudis and I. Sebos, "The carbon footprint of transport activities of the 401 Military General Hospital of Athens," *Environmental Modeling & Assessment*, vol. 26, pp. 155–162, 2021.

[10] J. L. Martín-Ortega, J. Chornet, I. Sebos, S. Akkermans, and M. J. López Blanco, "Enhancing transparency of climate efforts: MITICA's integrated approach to greenhouse gas mitigation," *Sustainability*, vol. 16, no. 10, p. 4219, 2024.

[11] I. Nydrioti, I. Sebos, G. Kitsara, and D. Assimacopoulos, "Effective management of urban water resources under various climate scenarios in semiarid mediterranean areas," *Sci Rep*, vol. 14, no. 1, p. 28666, 2024.

[12] A. Hajat, C. Hsia, and M. S. O'Neill, "Socioeconomic disparities and air pollution exposure: a global review," *Curr Environ Health Rep*, vol. 2, pp. 440–450, 2015.

[13] J. Grigg, "Air pollution and children's respiratory health–gaps in the global evidence," *Clinical & Experimental Allergy*, vol. 41, no. 8, pp. 1072–1075, 2011.

[14] R. Matyssek *et al.*, "Forests under climate change and air pollution: gaps in understanding and future directions for research," *Environmental Pollution*, vol. 160, pp. 57–65, 2012.

[15] R. Vilcassim and G. D. Thurston, "Gaps and future directions in research on health effects of air pollution," *EBioMedicine*, vol. 93, 2023.

[16] R. Turnbull, K. Rogers, A. Martin, M. Rattenbury, and R. Morgan, "Human impacts recorded in chemical and isotopic fingerprints of soils from Dunedin City, New Zealand," *Science of the total environment*, vol. 673, pp. 455–469, 2019.

[17] S. Fahad *et al.*, *Plant growth regulators for climate-smart agriculture*. CRC Press, 2021.

[18] P. Hystad, S. Yusuf, and M. Brauer, "Air pollution health impacts: the knowns and unknowns for reliable global burden calculations," 2020, *Oxford University Press*.

[19] H. A. Al-Jamimi, S. Al-Azani, and T. A. Saleh, "Supervised machine learning techniques in the desulfurization of oil products for environmental protection: A review," *Process Safety and Environmental Protection*, vol. 120, pp. 57–71, 2018.

[20] M. E. Marlier, A. S. Jina, P. L. Kinney, and R. S. DeFries, "Extreme air pollution in global megacities," *Curr Clim Change Rep*, vol. 2, pp. 15–27, 2016.

[21] W. H. Organization, "Ambient air pollution: A global assessment of exposure and burden of disease," 2016.

[22] I. O. Alade, M. A. Abd Rahman, and T. A. Saleh, "Predicting the specific heat capacity of alumina/ethylene glycol nanofluids using support vector regression model optimized with Bayesian algorithm," *Solar Energy*, vol. 183, pp. 74–82, 2019.

[23] K. Pozo, T. Harner, S. C. Lee, F. Wania, D. C. G. Muir, and K. C. Jones, "Seasonally resolved concentrations of persistent organic pollutants in the global atmosphere from the first year of the GAPS study," *Environ Sci Technol*, vol. 43, no. 3, pp. 796–803, 2009.

[24] V. M. Madhuri, G. G. H. Samyama, and S. Kamalapurkar, "Air pollution prediction using machine learning supervised learning approach," *Int. J. Sci. Technol. Res*, vol. 9, no. 4, pp. 118–123, 2020.

[25] B. T. Ong, K. Sugiura, and K. Zettsu, "Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM 2.5," *Neural Comput Appl*, vol. 27, pp. 1553–1566, 2016.

[26] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, pp. 22408–22417, 2016.

[27] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[28] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, Springer, 2010, pp. 177–186.

[29] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon, "Low rank modeling of signed networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 507–515.

[30] T. A. Reist, D. Koo, D. W. Zingg, P. Bochud, P. Castonguay, and D. Leblond, "Cross validation of aerodynamic shape optimization methodologies for aircraft wing-body optimization," *AIAA Journal*, vol. 58, no. 6, pp. 2581–2595, 2020.

[31] K. M. Bain *et al.*, "Cross-validation of three Advanced Clinical Solutions performance validity tests: Examining combinations of measures to maximize classification of invalid performance," *Appl Neuropsychol Adult*, vol. 28, no. 1, pp. 24–34, 2021.

[32] A. Das and T. Basu, "Assessment of peri-urban wetland ecological degradation through importance-performance analysis (IPA): A study on Chatra Wetland, India," *Ecol Indic*, vol. 114, p. 106274, 2020.

[33] K. Nidhul, S. Kumar, A. K. Yadav, and S. Anish, "Enhanced thermo-hydraulic performance in a V-ribbed triangular duct solar air heater: CFD and exergy analysis," *Energy*, vol. 200, p. 117448, 2020.

Biographies

Hongyun Liu was born in SuiHua, Heilongjiang, China, in 1995.,She is a lecturer in JiLin Agricultural Science and Technology University, She received the bachelor's degree from Changchun Normal University, her master's degree from Northeast Electric Power University. Her research interest include Research on Social Governance.

Yuehua Bai was born in Yixian, Hebei, P.R. China, in 1981. She received the Master's degree from Central China Normal University, P.R. China. Now, she works at the School of Management, Donghu College of Wuhan. Her research interest include management science and engineering, regional sustainable development.