HYAQP: A Hybrid Meta-Heuristic Optimization Model for Air Quality Prediction Using Unsupervised Machine Learning Paradigms

Praveena Vasudevan
Department of Electronics Communication and Engineering
SRM Institute of Science and Technology, India
praveenaecev@gmail.com

Chitra Ekambaram
Department of Electronics Communication and Engineering
SRM Institute of Science and Technology, India
chitrae@srmist.edu.in

Abstract: *In the current decade, presence of air pollution leads towards serious health conditions, including respiratory ailments, cardiovascular disorders, and lung cancer, which impacts both the lifespan and overall well-being of individuals. Moreover, the air pollution has the potential to have detrimental effects on the environment, resulting in destruction to ecosystems, decreased agricultural output and so on. Emission control systems, better fuels, stronger laws, energy efficiency improvements, and renewable energy promotion are helping industries reduce air pollution. To monitor the quality of air methods such as Particulate Matter (PM2.5), PM10, Ozone (O3), and other meteorological indicators exists. These measures pose real-time air quality however it fails to predict air quality. Predicting air quality helps reduce air pollution by enabling timely interventions and preventive measures to mitigate pollution peaks. Thus, in this research a hybrid version of optimization algorithm namely Hybrid Air Quality Prediction system (HYAQP) which is a combination of k-means clustering algorithm and meta-heuristic algorithm Sine Cosine Algorithm (SCA) is proposed. The HYAQP holds SCA integrated with k-means algorithm to find optimal cluster centroid for grouping the air data into three clusters good, poor, and moderate quality. Then the cluster which is nearer to the test instance is found and the instances present in those clusters are passed to K-Nearest Neighbor Regressor (K-NNR). Comparing HYAQP on mean absolute error it outperforms 62.9% than Multiple Linear Regression (MLR), 58.5% than Support Vector Regression (SVR), 45.5% than Vanilla-Long Short-Term Memory (Vanilla-LSTM), 44.4% than Sparrow Search Algorithm based-LSTM (SSA-LSTM) and 53.8% than K-Nearest Neighbor (KNN).*

Keywords: *Meta-heuristic, unsupervised machine learning, hybrid optimization, optimized k-means, air quality prediction.*

Received March 12, 2024; accepted September 5, 2024
https://doi.org/10.34028/iajit/21/5/15

1. Introduction

The increase in the development of metropolitan cities with the aid of improving economy leads to increase pollution in the air, water, and noise. Air pollution is a serious threat that causes various diseases and even causes death for living beings. With the growth of modern industries, smokes emitted from vehicles, and the cutting down of trees are the major sources of air pollution. Human life has a direct impact on air pollution because of the suspended particles and pollutants present in the air, and thus air pollution draws severe attention nowadays. Various acts had been enacted from time to time to prevent air pollution among which the Clean Air Act (CAA) states that all-major sources of air pollution including the usage of mobile devices and cell towers must adhere to air quality standards. Mobile devices include all the moving vehicles like cars, buses, motorcycles, etc., [40]. Though various programs had been enacted to prevent air pollution, the death rate crosses nearly 120,000 in countries like India due to Air pollution. Also, air pollution affects the economic rate of the country by nearly 2 lakh crores [19]. As per the report, the major air pollutant called PM2.5 causes an approximate death rate of 160,000 across major cities in India. The various control measures taken by the government include the use of Compressed Natural Gas (CNG), pollution under control certification for all vehicles driven by petrol and diesel, and spreading the use of Electric vehicles. Despite the measures, air pollution seriously affects the lives of humans.

Thus, in this paper, the importance of predicting air pollution in advance is identified and various machine learning algorithms had been studied [13, 33]. Intending to predict the quality of air, the forecasting of air pollution is the only potential solution. Such forecasting can be done through machine learning algorithms.

Hybrid Air Quality Prediction system (HYAQP) forecast the air quality using a novel algorithm designed by the combination of sine cosine optimization, k-means algorithm and K-Nearest Neighbor Regressor (K-NNR). With historical meteorological parameters such as temperature, pressure, wind speed, wind direction, humidity, etc., a model has been built to predict the air pollution to plan for effective measures to control it. The proposed HYAQP initially segregates the given data into three clusters viz. good quality, moderate quality

and polluted air using k-means algorithm. This process will help the proposed prediction model to classify the new data to which category it falls into. Then K-NN regression had been applied over the centroids of three clusters and K-neighbors from the instances in the nearby clusters are found and the average value is taken for measuring the quality of air. In this process, the random centroids in the initial phase, faces a valid reason since the initial centroids holds the chance of deviation from accuracy in prediction that then turns into wrong classification of clustered groups. Hence, to fix appropriate centroids in the initial levels of k-means algorithm, Sine Cosine Algorithm (SCA) is imposed. Metaheuristic algorithm intends to find the best optimal solution among the possible feasible solution [14, 35]. Some of the predominant metaheuristic algorithms are Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), Ant Colony Optimization (ACO), SCA etc., in this article, for finding the optimal cluster centroid, SCA algorithm is used due to the nature of sine and cosine function. Also, from the literature the SCA algorithm has been used for various purposes and to the best of author's knowledge, there are no other methodology exists on integrating SCA with the k-means algorithm to identify and select the most relevant centroids from the dataset, which are then clustered using k-means to group similar data points based on these features. After clustering, each cluster's centroids serve as representative data points that are subsequently utilized by a K-NN regressor. The K-NN regressor predicts air quality by finding the closest clusters' centroids and averaging their corresponding air quality values to make accurate predictions for new, unseen data points.

The key contributions of this research work includes:

- Integration of SCA algorithm on k-means to determine the optimal clusters enhancing the accuracy of the clustering process and ensuring that the k-means algorithm operates on the most informative data points.
- The hybridization of K-NN regressor with the above technique after clustering allows the model to make predictions based on the most similar groups of data points, improving the precision of Air Quality Predictions (AQP) by leveraging localized patterns within the data.
- The combination of SCA and k-means reduces initial deviations from random centroids and complexity of the dataset, making the subsequent K-NN regression more computationally efficient and scalable for large datasets.

This article is organized as follows: Section 2 depicts the various related works present in measuring the quality of air. Section 3 depicts the design of AQP system using improved k-means based on SCA and K-NNR. Section 4 depicts the experimental results stating the influence of the proposed algorithm. Finally, section 5 concludes with future work.

2. Related Work

This section comprises of state-of-the-art methods that are existing to predict the quality of air and the impact of meta heuristic algorithms on various domains.

2.1. Air Quality Measurement and Factors

Air quality measurements are critical for understanding the state of the environment and the potential health impacts on the population. The key points about air quality measurements and the factors influencing them are listed in this section.

2.1.1. Air Quality Measurements

PM2.5 and PM10: PM2.5 refers to Particulate Matter (PM) with a diameter of 2.5 micrometers or smaller. These particles can penetrate deeply into the lungs and even enter the bloodstream, causing various health issues. PM10 refers to PM with a diameter of 10 micrometers or smaller, which can cause respiratory problems and other health issues [44]. Ozone (O3): Ground-level O3 is a harmful air pollutant formed when pollutants emitted by cars, power plants, and other sources react chemically in the presence of sunlight [16]. Carbon Monoxide (CO): CO is a colorless, odorless gas resulting from the incomplete combustion of fossil fuels. It can prevent oxygen from entering the body's tissues and is especially harmful to people with cardiovascular disease [7].

2.1.2. Air Quality Monitoring and Regulation

Air Quality Index (AQI): the AQI is a standardized index used by governments to communicate the level of air pollution to the public. It typically includes measurements of PM, O3, and CO [15]. Regulatory standards: Various countries have regulatory bodies that set and enforce air quality standards, such as the Environmental Protection Agency (EPA) in the United States and the European Environment Agency (EEA) in Europe [17].

2.2. State of the Art Methods on Air Quality Prediction

Precise forecasting of air quality is an essential element of meteorological services, and the ongoing monitoring of air quality is critical for evaluating the paths of pollutant emissions and variations in air quality. Improving the precision of air quality forecasting has the capacity to expedite the attainment of goals related to clean air and carbon neutrality. Hence, it is essential to effectively use the auxiliary function of machine learning models in air quality forecasting.

DeepAirNet was designed using deep learning for forecasting air pollution. Recurrent Neural Network

(RNN) when subjected to AirNet data predicted the particulate Matter 10 which was considered as major pollutant and the designed model forecasted air pollution for every hour across Chinese cities [4]. With the intend to predict the Respirable Suspended Particulate Matter (RSPM), Sulfur dioxide (SO2) and Nitrogen dioxide (NO2) for the city Lucknow in India, three neural networks model were designed. The neural network architecture used were multi-layer perceptron network, Radial Basis Function Neural Network (RBFNN) and Generalized RNN (GRNN). Meteorological data including air temperature, relative humidity, wind speed and air quality data including concentration of Suspended Particulate Matter (SPM), NO2, SO2 collected during the period 2005 to 2009 [37]. multi task learning was formulated using the meteorological data gathered for the past day for 24 hours. Parameter reducing normalization was designed to overcome the drawbacks of conventional regression algorithms [47]. Recurrent Network Model (RNM), change point detection model with RNM, Sequential Network Construction Model and Self-Organizing Feature Maps (SOFM) were designed for forecasting air quality. As the air quality data collected through various sensors have high level of noise, SOFM obtained high level of prediction than other forecasting models [5]. LightGBM model was designed by integrating gradient boosted decision trees and XGboost algorithm. As the air quality data collected over real time have noise, LightGBM with its ability to learn parallel and its capability of learning the data at good rate of accuracy, it improved the accuracy of forecasting the quality. Also, the memory requirement of LightGBM is very less, as ii used histogram-based segmentation [46].

The major pollutants for deteriorating air quality is O3 and PM10. Statistical approach such as feed-forward neural network was designed. Pruned Neural Network (PNN) based on parameter-parsimonious technique was used to remove the repeated information from fully connected neural network. Apart from PNN, a Local Linear (LL) algorithm were used to predict the concentration of O3 and PM10 [11]. Deep Spatial-Temporal Ensemble (STE) model was designed which includes Ensemble technique, discovering spatial correlation and temporal predictor. Temporal Predictor was built using Long Short-Term Memory (LSTM) networks. The designed STE model was evaluated on real data collected from 35 monitoring stations situated in Beijing, China [42]. LSTM network was used to predict the air quality from the data collected using various IoT devices across smart cities [25]. A study was made on predicting air quality using various techniques available in machine learning including artificial neural network, GA, decision tree, deep belief network, least square support vector machine. The identified research issues in this study were how quality the data is when collected using IoT devices and monitoring the quality of air dynamically according to meteorological data [22].

Hybrid forecasting model was developed for predicting concentration of pollutants. Back propagation neural network together with fuzzy set theory and analytic hierarchy process were used to predict the quality of air. The designed forecasting model classifies the air quality as good, moderate, lightly polluted, heavily polluted and severely polluted for the data gathered in Chengdu and Hangzhou from China [43]. Neuro fuzzy technique was developed to predict the quality of air. Fuzzy clustering had been done with the membership functions such as mean and variance. Fuzzy rules were extracted from fuzzy clusters. Genetic PSO based neural network was constructed to train the network for optimal prediction of air quality [11, 28]. Transfer learning integrated with Bi-Directional Long Short Term Memory Model (BLSTM) was proposed for predicting the pollutants that affect the air quality. The designed BLSTM was evaluated for the data collected from Guanddong, China [30].

Liao *et al*. [27] proposed AQP model using the hybrid mechanism which is unsupervised learning and pattern learning. The unsupervised pre-training method is helpful for long-term temporal pattern learning in this proposed model. Sui and Han [39] a graph based convolutional network for predicting the quality of air. The multi angle view and the multiple task oriented spatiotemporal information from the graph based network are the prime factors for predicting the air quality. Chen *et al*. [10] proposed a transmit neural network for predicting the quality of air at regional level. The proposed model used spatiotemporal and hierarchical information of the region to predict the quality of air using deep learning model. In the year 2024, Simsek *et al*. [36] proposed an event detection system along with AQP system using the decentralized, fog-assisted system. In the year 2024, authors Chen *et al*. [9] proposed a spatio-temporal assisted neural network for predicting the quality of air. The proposed model used multistage graph and special information to predict the quality of air using deep learning model.

2.3. Impact of Sine Cosine Algorithm

SCA was a population-based optimization algorithm modelled using sine and cosine functions for solving unimodal, multi-modal and composite function [31]. SCA based K-NN was used to predict the phishing attacks. The designed SCAK-NN was compared with decision tree, naïve bayes algorithm in terms of parameters such as accuracy, f-measure, true positive rate, false positive rate and mean absolute error. SCA integrated with K-NN was used for optimal detection of intrusion in wireless sensor network. Polymorphic mutation and compact mechanism were integrated with conventional SCA with the goal to minimize loss, time and space [34]. SCA k-means was used for clustering cloud resources with the aid for optimal resource discovery was designed. The fitness function of designed

SCA k-means relies on Intra-cluster similarity and Inter-cluster similarity. The designed SCA k-means was optimal enough than conventional k-means [23].

Modified Sine Cosine Algorithm (m-SCA) was designed to get rid of problems such as stagnation in local optima by integrating self-adaptiveness to find global optimal solution [20]. Automatic clustering was done through atom search optimization and SCA. The fitness considered was to minimize compact-separated index to find optimal quality of clusters. SCA was used as intensification operator to improve Dunn and silhoutte index [1]. Initial centroids play a prominent role in clustering. SCA-Fuzzy Possibilistic C-Ordered Means (SCA-FPCOM) was designed that integrates fuzzy C-Means with SCA. The outliers caused by initial random centroids are solved using SCA-FPCOM [26]. To improve the convergence rate of SCA, adaptive and modified SCA for clustering has been designed which promotes both diversification and intensification [6].

Zhang *et al*. [45] proposed a method for evaluating the emission in enterprises that induces power using the machine learning method named as support vector machine that are improved with least square.

From the literature survey, it is observed that there needs an optimal mechanism for predicting the air quality. Also, it is inferred that hybridization of SCA algorithm with traditional machine learning algorithms can be better strategy for effective prediction of air quality.

3. Proposed Air Quality Predictor System (AQPS)

Figure 1 shows the schematic architecture of the proposed AQP system. The proposed improved k-means based on SCA HYAQP has been used for finding optimal cluster centroid for k-means, thereby the clusters formed using k-means are optimal enough to create the groups in the air quality dataset. The K-Nearest Neighbor (K-NN) based regressor is then activated, which intends to find the optimal set of instances from the clusters by computing the distance between the cluster centroid with the new test instance. Then, the instances present in the nearest cluster is subjected to find the nearest neighbors and the average value is taken. The proposed AQPS includes data collector which intends to collect the data across various sensors. The raw data collected using sensors has to be preprocessed before subjecting for analyzing. The preprocessing carried out using min-max normalization where for each attribute the minimum value and maximum value has been chosen. All the other values are normalized according to minimum and maximum value between 0 to 1. The preprocessed data is fed to the HYAQP which finds the optimal cluster centroid and optimal clusters. The optimal clusters are returned to the K-NNR which finds the K-NNs of the new test instance.

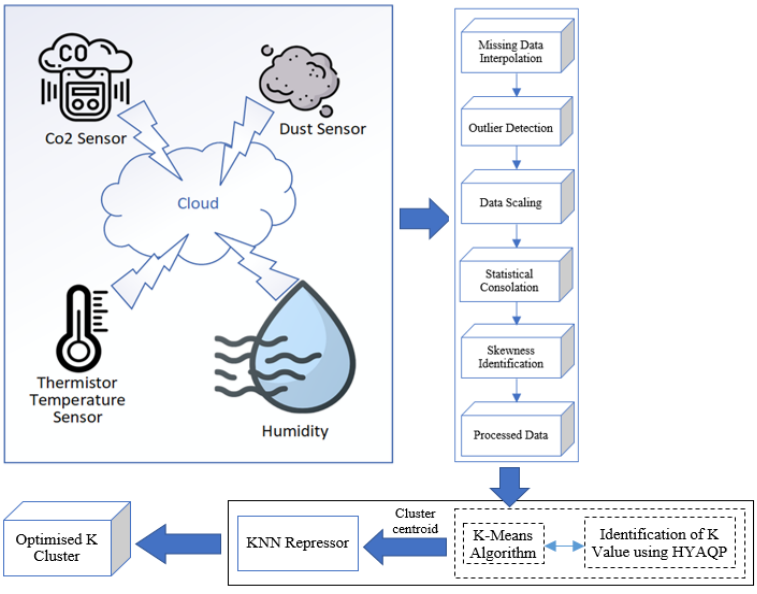


Figure 1. Proposed AQPS.

3.1. HYQAP

Algorithm (1) demonstrates that the proposed improved k-means algorithm using SCA aims to identify the optimal cluster centroids, thereby facilitating the computation of optimal clusters.

Algorithm 1: k-Means Algorithm (A, D).

Input: Initial Cluster Head points (D), Data Points (A)

while (termination condition satisfied) do
k←size(A)
for each i∈N do
$Z_i \leftarrow argmin_k ||A_i - D_i||$
end for
for each i∈k do
$A_i \leftarrow MEAN(\{D_i : Z_i = k\})$
end for
end while
Output: Cluster Head points (Z)

SCA is accelerated based on mathematical functions such as sine and cosine trigonometric functions. The SCA algorithm is populated with agents where the function intends to move the agents toward the best or outwards from it. The SCA algorithm is a population-based metaheuristic algorithm where initially, the agents are initialized with random values. Each agent in the population will have three dimensions viz. The first dimension represents the good quality of air, the second dimension represents the medium quality air and the third dimension represents the poor-quality air. At each iteration, the algorithm intends to find the globally optimal solution. When the algorithm is made to run for a single iteration, the algorithm may produce local optimal solution. Thus, at each iteration, the HYAQP does exploration which favors to provide global search i.e., diversification. Also, the algorithm undergoes exploitation, which promotes local search (i.e., intensification). Global search takes place by introducing high level of randomness with the agent's position, so that the agent randomly moves in abrupt direction. The term local search intends to promote randomness very slowly as opposed to diversification. Also, the use of sine and cosine function favors to switch between exploitation and exploration to a great extent. At each iteration, the best agent's position is present in the variable A_{Best}. All the other agents tend to update their position according to the best agent. Also, at each iteration the agents are evaluated with the fitness. The fitness of the agent is the linear combination of the objective of the agent and the weight associated with it. The representation of fitness is given in Equation (1).

$$f(A) = -w_1 * \nabla + w_2 * \Phi + w_3 * \psi \tag{1}$$

The w_1, w_2, and w_3 represents the weight associated with the objectives Entropy, Purity and Dunn index. The fitness function considered is a maximum function. The negative sign for the entropy indicates that it has to be minimized. Entropy $\nabla(\Omega_i)$ is a measure of the amount of disorderliness [38] in a cluster represented in Equation (2).

$$\nabla(\Omega_i) = \sum_{j=1}^{|CL|} p * l \tag{2}$$

where p represents the probability that the instance $\overrightarrow{X_\iota}$ in cluster Ω_i have class label CL_i shown in Equation (3). And l represents the logarithm of the probability represented in Equation (4). Entropy for all clusters $\nabla(\Omega_i)$ is computed as the sum of the product of entropy of individual cluster $\nabla(\Omega_i)$ and ratio of the number of instances in each cluster $|\Omega_i|$ to the size of dataset $|X|$ represented in Equation (5).

$$\mathrm{p} = \mathrm{P}\left(\overrightarrow{\mathrm{X_i}} \in \Omega_\mathrm{i}\right)_{\mathrm{CL_i}} \tag{3}$$

$$\mathrm{l} = \log_2\left(\mathrm{P}\left(\overrightarrow{\mathrm{X_i}} \in \Omega_\mathrm{i}\right)_{\mathrm{CL_i}}\right) \tag{4}$$

$$\nabla(\Omega) = \sum_{\Omega_i \in \Omega} \nabla(\Omega_i) * \frac{|\Omega_i|}{|\overrightarrow{X_i}|} \tag{5}$$

Purity of clusters $\Phi(\Omega_i)$ is a measure of the extent of how much the instances in a particular class CL_i belongs to cluster Ω_i [12] represented in Equation (6).

$$\Phi(\Omega_i) = \frac{1}{|X|} * \sum_{i=1}^{|\Omega|} \underset{j}{Max}\left(\Omega_i \cap CL_j\right) \tag{6}$$

where $\underset{j}{Max}(\Omega_i \cap CL_j)$ represents the maximum number of instances in the cluster Ω_i have class label CL_i.

Dunn index, which is a metric used to evaluate the cluster based on the partitioned data [38]. Let Ω_i and Ω_j be two clusters for the dataset X. The diameter of the cluster Ω_i is given as the maximum distance between any two instances $\overrightarrow{X_\iota}$, $\overrightarrow{X_j}$ represented in Equation (7). The diameter of the cluster is otherwise known as intra-cluster distance as shown in Equation (8).

$$\partial(\Omega_i) = \underset{\vec{X}_i, \vec{X}_j \in \Omega_i}{Max}\left(d\left(\vec{X}_i, \vec{X}_j\right)\right) \tag{7}$$

$$\partial(\Omega_i) = \Delta(\Omega_i) \tag{8}$$

$d\left(\overrightarrow{X_\iota}, \overrightarrow{X_j}\right)$, represents the distance between two instances $\overrightarrow{X_\iota}$ and $\overrightarrow{X_j}$. Dunn index ψ is represented in Equation (9).

$$\psi = \frac{\underset{1 \le j \le |\Omega|}{Min}\ \delta\left(\Omega_i, \Omega_j\right)}{\underset{1 \le k \le |\Omega|}{Max}\left(\partial(\Omega_i)\right)} \tag{9}$$

Once the fitness is evaluated, the agent's position will be updated. The position of the agent is updated using the sine function, cos function and best agent also. The computation of position of the agent is represented in Equations (10) and (11).

$$A_i^t \leftarrow A_i^{t-1} + r_1 * sin(r_2) + \left|r_3 * A_{Best}^t - A_i^{t-1}\right| \quad if\ r_4 < 0.5 \tag{10}$$

$$A_i^t \leftarrow A_i^{t-1} + r_1 * cos(r_2) + \left|r_3 * A_{Best}^t - A_i^{t-1}\right| \quad if\ r_4 \ge 0.5 \tag{11}$$

The variable r_1 is used to control between exploration and exploitation and it is computed using Equation (12).

$$r_1 \leftarrow a - \mathrm{t} * \frac{a}{Max\_Iter} \tag{12}$$

The variable t represents the current iteration. *Max_Iter* represents the maximum iteration taken for convergence. The variable a is initialized to 2. And gradually decreased from 2 to 0. The random variable r_2 decides whether the agent is moving towards the best agent or outwards from it. The random variable r_3 assigns weight to the best agent. If the value of r_3 is greater than 1, then the highest weight is assigned to the best agent else the weight is assigned to the agent's previous position. r_4 is used to switch between exploitation and exploration for better convergence. The process of computing the fitness and updating the agent's position is done for maximum number of iterations till convergence. Once the algorithm converged, it gives the optimal cluster centroid which is then used by K-NNR. The working of HYAQP is shown

in Algorithm (2). The diagrammatic representation of AQP using HYAQP is shown in Figure 2.

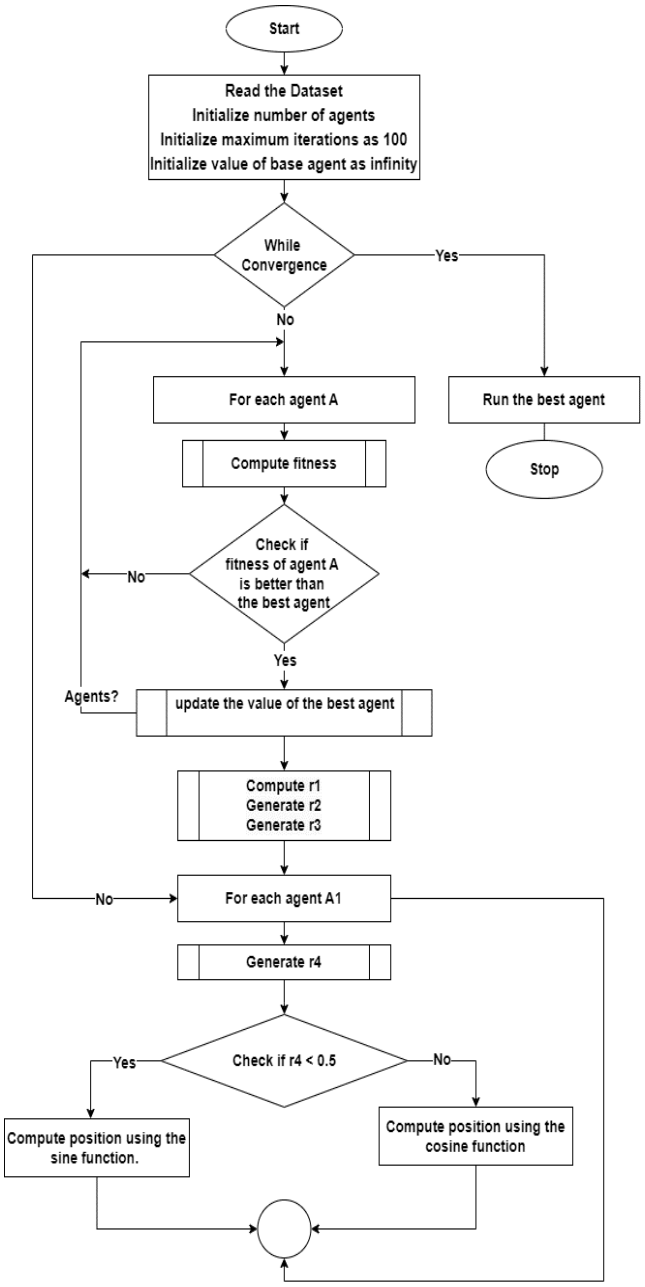


Figure 2. HYAQP workflow.

Algorithm 2: Hybrid Air Quality Prediction.

Input: $D \leftarrow \{I_1, I_2, \ldots, I_n\}$;
N-Number of Agents; Maximum Iteration (Max_Iter), Objective function (f), initial iteration (t=1)

Begin
// Population initialization
  for each Agent A_i *do*
    for each Agent A_i *do*
      $A_{i,j} \leftarrow rand(0,1)$
    end for
  end for
// Fitness Computation
  for each Agent A_i *do*
    $F(A_i) \leftarrow f(A_i)$
  end for
  while (t≤Max_Iter) do
    for each Agent A_i *do*
      $F(A_i) \leftarrow f(A_i)$
      if $(value_{A_{Best}}) < F(A_i))$ *then*
        $A_{Best} \leftarrow A_i$
      end if
    end for
    for each Agent A_i *do*
      Compute r_1 *using Equation (12)*
      Generate r_2, r_3, r_4

      if $(r_4 < 0.5)$ *then*
        Compute Position using Equation (10)
      else
        Compute Position using Equation (11)
      end if
    end for
  end while
  Return A_{Best}
End
Output: A_{Best}

3.2. Clustering based on Centroid Distance

The K-NNR rather than acting of all the instances to find K-NNs works on optimal cluster centroid given by HYAQP. The distance between the optimal cluster centroid and the new instance is computed. The centroid with minimum distance is considered and, in that K-NNs are found and the average value of the K-neighbors are computed and is the predicted air quality for the new instance. The working of K-NNR is shown in Algorithm (2). If the K-NNR is fed with all the instances present in the dataset, then the time for computation is high. Also, when finding the nearest neighbors of the new instance, the problem of finding the neighbors is critical, when the K-NN is subjected to too large dataset for estimating the optimal value of air quality. To avoid this, originally the instances where clustered into three groups based on the purity of air and contamination. Having clustered the instances using HYAQP, the new instance is tested with three cluster centroids. The cluster centroid which is having minimum distance is chosen and the instances present in that particular cluster centroid is given as input to K-NNR. By passing only a group of instances, the computational complexity is significantly reduced. Also, since the instances are found using HYAQP, the chosen K- Neighbors are optimal enough for the estimation of particulate matter. The nearest instances are found by using Euclidean distance represented in Equation (13).

$$Dist(I_{new}, clus_i) \leftarrow \sqrt{\sum_{j=1}^{num\_d} (I_{new} - clus_i)^2} \qquad (13)$$

4. Experimental Analysis

The proposed algorithm has been evaluated on 6 different datasets [2, 3, 8, 21, 41] and listed in Table 1.

The features include data, time, average concentration of CO, PT08.S1(tin oxide), concentration

of nonmetallic hydrocarbons, benzene concentration, PTO8.S2 (titania), concentration of NOx, PTO8.S2(tungsten oxide), NO2 concentration, PTO8.S4 (tungsten oxide), PT08.S5 (indium oxide), temperature, relative humidity and AH Absolute humidity. The proposed work has been implemented using python, with system configuration of Intel ® core™ i7, 1.80 GHZ and 16 GB RAM. Algorithms taken into comparison of the proposed work includes Multiple Linear Regression (MLR), Support Vector Regression (SVR), K-NN, Vanilla LSTM [18] and Sparrow Search Algorithm based LSTM (SSA-LSTM) [29]. The metrics taken into account for comparison are mean absolute error, root mean square error, purity, Dunn index, coefficient of determination, Mean Bias Error (MBE) and complexity. The parameters used for HYAQP are as follows: *Number of Agents (N)*: 100; Maximum Iteration *(Max_Iter)*: 100; Runs: 10.

Table 1. Datasets used for evaluation.

Si. No.	Dataset	#Instances	#Attributes	Ref.
1	Italian city	9357	14	[21]
2	UK-AIR (cambridge city)	6199	7	[3]
3	UK-AIR (wicken fen)	6844	13	[3]
4	Cities in India	6236	15	[8]
5	London	3169	16	[2]
6	US-pollution data	7301	28	[41]

4.1. Comparison of Mean Absolute Error

Mean absolute error is defined as the ration of sum of the difference between the observed value and actual value to the total number of instances which is represented in Equation (14). Figure 3 represents the comparison of mean absolute error. The mean absolute error of HYAQP is 62.9% reduced than MLR, 58.5% reduced than SVR, 45.5% reduced than Vanilla LSTM, 44.4% reduced than SSA-LSTM and 53.8% reduced than KNN. The reason behind is that proposed mechanism used two level of processing data. In the first level, clustering is done through optimal cluster centroid given by SCA and in the second level, the nearest group of instances is given to K-NNR (K=3) which optimally finds the 3 nearest neighbors and estimate the amount of benzene in the air to determine air quality.

$$MAE \leftarrow \frac{\sum_{i=1}^{N}(|y_i - \hat{y}_i|)}{N} \quad (14)$$

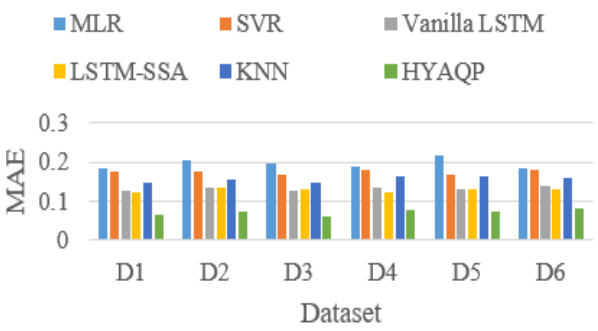


Figure 3. Comparison of mean absolute error.

4.2. Comparison of Root Mean Square Error

Root mean square error is defined as the root of the ratio of sum of squared difference between actual values and predicted to total number of instances which is represented in Equation (15). Figure 4 represents the comparison of root mean square error by various algorithms. It is evident from the Figure 4 that RMSE of HYAQP is minimum than other existing algorithms. The mean absolute error of HYAQP is 51% reduced than MLR, 43.1% reduced than SVR, 29.7% reduced than Vanilla LSTM, 22.2% reduced than SSA-LSTM and 31.8% reduced than KNN. This shows that HYAQP good in forecasting the pollution, in particular the estimation of Benzene. As MLR is biased with slope and intercept, SVR is biased with the hyperplane, HYAQP optimally finds the 3 neighbors of the new instance and takes the average of the values to find the estimation of benzene.

$$RMSE \leftarrow \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (15)$$

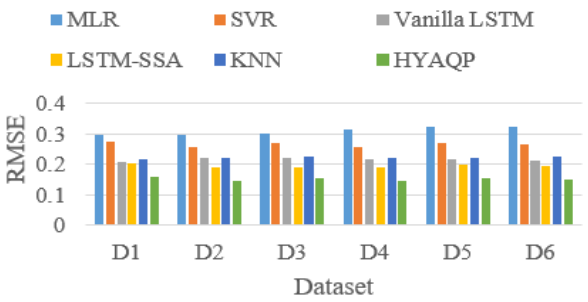


Figure 4. Comparison of root mean square error.

4.3. Comparison of Purity

Next level of comparison is done on measuring the purity of the cluster created using PSO [24], GA [32], SSA and Grey Wolf Optimizer (GWO). Table 2 shows the purity obtained by various algorithms taken for consideration. It is observed that the proposed HYAQP achieves a higher level of purity than other existing algorithms. The purity of clusters created using HYAQP is 16.6% higher than clusters created using Inverse Kinematics-Particle Swarm Optimization (IKPSO). The proposed HYAQP has a balanced exploration and exploitation which is the prime factor for high rate of purity. The purity of Inverse Kinematics-Genetic Algorithm (IKGA) is very minimum i.e., 18.4% reduced purity when comparing to HYAQP. The reason behind minimum purity in IKGA is that the cross over and mutation probability played a critical role in doing exploration and exploitation. As a result, the solution obtained by IKGA is not optimal enough to create optimal clusters. Similarly, on comparing Inverse Kinematics Sparrow Search Algorithm (IKSSA) and Inverse Kinematics-Grey Wolf Optimizer (IKGWO), HYAQP improved its performance with 10.3% and 3.77% respectively.

Table 2. Comparison of purity.

	D1	D2	D3	D4	D5	D6
IKGA	0.76	0.76	0.77	0.73	0.74	0.77
IKPSO	0.77	0.77	0.77	0.77	0.77	0.75
IKSSA	0.81	0.82	0.83	0.82	0.87	0.83
IKGWO	0.91	0.89	0.87	0.90	0.90	0.86
HYAQP	0.91	0.95	0.92	0.91	0.93	0.93

4.4. Comparison of Dunn Index

Dunn index is another important metric taken into account for evaluating the clustering. Dunn index should be high for good clustering algorithm. Table 3 represents the Dunn index obtained for various clustering algorithms. As expected, HYAQP ranks highest in maximizing the Dunn index compared to the other algorithms considered. The HYAQP improves Dunn index by 16% than IKPSO and 21% than IKGA. The reason behind the improvement of Dunn index in proposed HYAQP is that the clusters created are optimal enough. Also, SCA algorithm tends to switch between exploration and exploitation with random probability. Also, the algorithm uses mathematical functions such as sine and cosine functions for doing intensification and diversification, the centroids returned by HYAQP is optimal enough for clustering the instances into three groups of viz. good quality, moderate quality and poor quality of air. Similarly on comparing IKSSA and IKGWO, HYAQP improved its performance with 10.3% and 5.1% respectively.

Table 3. Comparison of Dunn index.

	D1	D2	D3	D4	D5	D6
IKGA	0.761	0.759	0.720	0.731	0.729	0.726
IKPSO	0.765	0.786	0.756	0.774	0.786	0.795
IKSSA	0.863	0.827	0.868	0.818	0.818	0.809
IKGWO	0.904	0.870	0.889	0.863	0.874	0.875
HYAQP	0.949	0.941	0.921	0.945	0.901	0.950

4.5. Comparison of Coefficient of Determination

The coefficient of determination of the proposed HYAQP had been compared with other algorithms like MLR and SVR which is represented in Figure 5. The coefficient of determination represented using R^2 is calculated using Equation (16). The proposed HYAQP achieves 4.20% greater coefficient of determination than MLR. Also, SVR achieves 12.567% less coefficient of correlation than the proposed HYAQP.

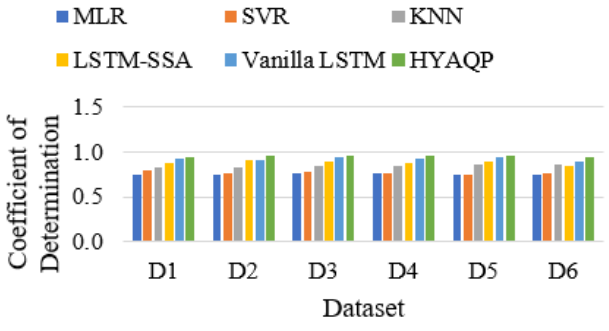


Figure 5. Comparison of coefficient of determination.

Where N represents number of instances, σ_x represents the standard deviation for the input instances, σ_y represents the standard deviation of the predictor variable, $\bar{x}$ and $\bar{y}$ represents the mean of the input instances and predictor variable.

$$R^2 = \frac{1}{N}\frac{\sum_{j=1}^{N}(y_j - \bar{y})(x_j - \bar{x})^2}{\sigma_y \sigma_x} \quad (16)$$

4.6. Comparison of Mean Bias Error

The MBE determines the average prediction error. For a good predictor model, the MBE must be low. Equation (17) represents the calculation of MBE. From the Figure 6 it is evident that, the proposed HYAQP achieves 40.38% and 54.389% reduction of MBE than MLR and SVR respectively.

$$MBE = \frac{\sum_{j=1}^{N}(y_j - x_j)}{N} \quad (17)$$

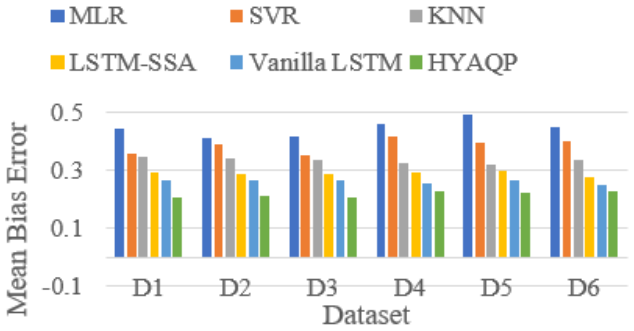


Figure 6. Comparison of MBE.

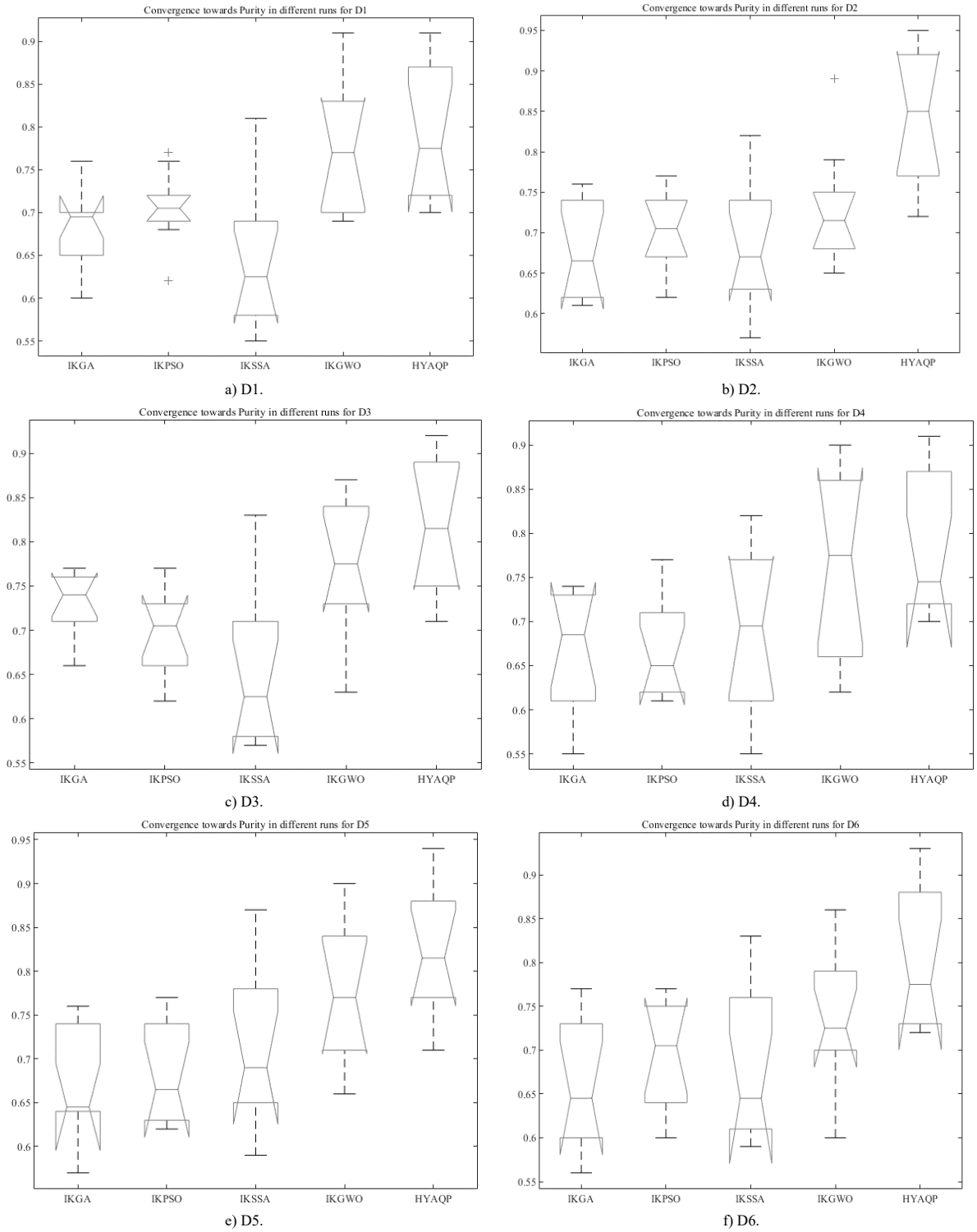
4.7. Comparison of Computational Time

Next level of comparison is done to measure the time complexity of proposed HYAQP with conventional K-NNR. The conventional K-NNR, which is a lazy learning algorithm, intends to find the K neighbors by computing the distance with all the instances present in the dataset. Thus, the time complexity is directly proportional to the number of instances present in the dataset. In the case of HYAQP, where the clusters are formed using optimized k-means and the nearest cluster is passed to the K-NNR. Thus, time taken to process the test instance is minimum in HYAQP.

4.8. Diversity Analysis

Figure 7 shows the overall convergence analysis with respect to purity on different datasets. From the figures it is evident that the proposed model is having a significant convergence towards the better purity value in all the runs. In particular, the convergence analysis w.r.t. purity for D1 dataset Figure 7-a) shows that the proposed model is balanced between the worst and the best results whereas the IKPSO is biased in worst case results in many runs. The convergence analysis w.r.t. purity for D2 dataset Figure 7-b) shows that the proposed model is more towards the best results in most

of the runs whereas the IKKSA is biased in worst case results in many runs. The convergence analysis w.r.t. purity for D3 dataset Figure 7-c) shows that the proposed model is more towards the best results in most of the runs and similarly the existing methods are also in line with the proposed model. The convergence analysis w.r.t. purity for D4, D5, D6 dataset Figure 7-d), (e) and (f) shows that the IKGWO is more towards the best results in most of the runs. However, the proposed model is more towards worst solutions in many runs.



a) D1. b) D2. c) D3. d) D4. e) D5. f) D6.

Figure 7. Convergence analysis w.r.t. purity on different datasets.

4.9. Statistical Analysis

The results of purity are interpreted with Two-Way ANOVA test and Post Hoc tests results are discussed in this section.

Univariate ANOVA is used to establish the association between control factors, and a solitary numerical dependent variable. In this system, the control factors are datasets and the algorithms and the dependent variable is Purity. Table 4 shows the analysis of algorithms on different dataset its mean and standard deviation (Std. Dev.).

Table 4. Analysis of algorithms on different dataset its mean and standard deviation.

Algorithms	Dataset (D1)		Dataset (D2)		Dataset (D3)		Dataset (D4)		Dataset (5)		Dataset (D6)		Total	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
IKGA	.6520	.06563	.6770	.05794	.6800	.07379	.6770	.05056	.6820	.06746	.6560	.08235	.6707	.06530
IKGWO	.7590	.09632	.7740	.10058	.7530	.08970	.7420	.10644	.7460	.10793	.7680	.08456	.7570	.09441
IKPSO	.6690	.06540	.6850	.06346	.6950	.06671	.6750	.05968	.7110	.03900	.7000	.04922	.6892	.05747
HYAQP	.8150	.08017	.8310	.08478	.8060	.07321	.7960	.08222	.8400	.08206	.8490	.06262	.8228	.07687
IKSSA	.6890	.08020	.7150	.07075	.7230	.08551	.6790	.07651	.6870	.10350	.6940	.09009	.6978	.08294
Total	.7168	.09724	.7364	.09445	.7314	.08762	.7138	.08884	.7332	.09946	.7334	.09977	.7275	.09430

Table 5. Levene's test of equality of error variances.

Levene's test of equality of error variances					
		Levene statistic	df1	df2	Sig.
Purity	Based on Mean	1.440	29	270	.073
	Based on Median	1.229	29	270	.201
	Based on Median and with adjusted df	1.229	29	217.825	.205
	Based on trimmed mean	1.426	29	270	.078

Levene's test, introduced by Levene in 1960, is used to determine whether k samples exhibit equal variances. Homogeneity of variance refers to equal variances across samples. It tests the null hypothesis that the error variance of the dependent variable is equal across groups. In this model, Purity is indulged for the comparison of homogeneity and then results are tabulated in Table 5 The design model of conducting the Levene's test includes the design as the sum of intercept, dataset, algorithms, dataset with algorithms.

Table 6. White test for heteroskedasticity.

White test for heteroskedasticity		
Chi-square	df	Sig.
47.421	29	.017

Table 6 shows the white test for heteroskedasticity where it lists the chi-quare test values, degree of freedom and sigma factor. The dependent variable chosen for comparison is purity. This model test the null hypothesis that the variance of the errors does not depend on the values of the independent variables.

Table 7. Tests of between-subjects effects.

Dependent variable: Purity					
Source	**Type 3 sum of squares**	**df**	**Mean square**	**F**	**Sig.**	**Partial eta squared**
Corrected model	.997[a]	29	.034	5.587	.000	.375
Intercept	158.777	1	158.777	25797.026	.000	.990
Dataset	.023	5	.005	.754	.584	.014
Algorithms	.932	4	.233	37.868	.000	.359
Dataset * algorithms	.042	20	.002	.339	.997	.024
Error	1.662	270	.006			
Total	161.436	300				
Corrected total	2.659	299				
R Squared= 375 (Adjusted R Squared= 308)						

All model terms in the between-subjects effects tests are statistically significant with significance values below 0.05. in this respective concern the sigma values of Algorithms corrected model and intercept are less than 0.05 which indicates that the concerns are statistically significant to each other with respect to purity. Table 7 holds the comparison between subject effects.

Table 8. Pairwise comparison of algorithms.

(I) Algorithms	(J) Algorithms	Mean difference (I-J)	Sig.	95% Confidence interval for difference	
				Lower bound	Upper bound
IKGA	IKGWO	-.086*	.000	-.127	-.046
	IKPSO	-.018	1.000	-.059	.022
	HYAQP	-.152*	.000	-.193	-.112
	IKSSA	-.027	.589	-.068	.013
IKGWO	IKGA	.086*	.000	.046	.127
	IKPSO	.068*	.000	.027	.108
	HYAQP	-.066*	.000	-.106	-.025
	IKSSA	.059*	.000	.019	.100
IKPSO	IKGA	.018	1.000	-.022	.059
	IKGWO	-.068*	.000	-.108	-.027
	HYAQP	-.134*	.000	-.174	-.093
	IKSSA	-.009	1.000	-.049	.032
HYAQP	IKGA	.152*	.000	.112	.193
	IKGWO	.066*	.000	.025	.106
	IKPSO	.134*	.000	.093	.174
	IKSSA	.125*	.000	.084	.166
IKSSA	IKGA	.027	.589	-.013	.068
	IKGWO	-.059*	.000	-.100	-.019
	IKPSO	.009	1.000	-.032	.049
	HYAQP	-.125*	.000	-.166	-.084

Table 8 shows the pairwise comparison of algorithms of each model with every other model in all aspects such as mean difference standard error, significance and confidence interval. Adjustment on multiple comparisons are taken care by Bonferroni. On comparing the results the mean difference is significant at level 0.05.

In Table 9. the F score tests the effect of Algorithms.

This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Table 10 shows the comparison of algorithms for every datasets on 95% confidence intervals.

Table 9. Univariate tests.

Dependent variable: Purity						
	Sum of squares	df	Mean square	F	Sig.	Partial eta squared
Contrast	.932	4	.233	37.868	.000	.359
Error	1.662	270	.006			

Table 10. Comparison of algorithms for every datasets on 95% confidence intervals.

Dataset	Algorithms	Mean	95% Confidence interval	
			Lower bound	Upper bound
D1	IKGA	.652	.603	.701
	IKGWO	.759	.710	.808
	IKPSO	.669	.620	.718
	HYAQP	.815	.766	.864
	IKSSA	.689	.640	.738
D2	IKGA	.677	.628	.726
	IKGWO	.774	.725	.823
	IKPSO	.685	.636	.734
	HYAQP	.831	.782	.880
	IKSSA	.715	.666	.764
D3	IKGA	.680	.631	.729
	IKGWO	.753	.704	.802
	IKPSO	.695	.646	.744
	HYAQP	.806	.757	.855
	IKSSA	.723	.674	.772
D4	IKGA	.677	.628	.726
	IKGWO	.742	.693	.791
	IKPSO	.675	.626	.724
	HYAQP	.796	.747	.845
	IKSSA	.679	.630	.728
D5	IKGA	.682	.633	.731
	IKGWO	.746	.697	.795
	IKPSO	.711	.662	.760
	HYAQP	.840	.791	.889
	IKSSA	.687	.638	.736
D6	IKGA	.656	.607	.705
	IKGWO	.768	.719	.817
	IKPSO	.700	.651	.749
	HYAQP	.849	.800	.898

Table 11 shows the post Hoc test R-E-G-W range on purity that shows the order of algorithms as per the subset. Means for groups in homogeneous subsets are displayed. Based on observed means the error term is mean square error is 0.006. Critical values are not monotonic for these data. Substitutions have been made to ensure monotonicity. Type I error is therefore smaller.

Table 11. Post hoc test (Ryan-Einot-Gabriel-Welsch) range on purity.

Algorithms	N	Subset		
		1	2	3
HYAQP	60	.8228		
IKGWO	60		.7570	
IKSSA	60			.6978
IKPSO	60			.6892
IKGA	60			.6707
Sig.		1.000	1.000	.225

4.10. Time Complexity of HYAQP

The computational time of the entire HYAQP is carried out in two different algorithms optimized k-means followed by K-NNR. To compute the entire time complexity of HYAQP, these two method's time complexities will be computed and added together and in the latter case the asymptotic notations will be used to denote the time complexity of the proposed model. (a) On examining the time complexity of HYAQP algorithm's, it is found that it mostly relies on two factors:

1. The randomness of the initialization
2. The use of the Sine approach to update each individual node's location.

Big O notation may describe both as O(M×N), where M is the population and N is the number of dimensions of the problem. (b) The time complexity of the K-Means algorithms is highly dependent on three factors

1. The number of attributes in the dataset (Q).
2. Based on the total number of partitions (P).
3. Total number of iterations (T) and the total time complexity can be computed as O (Q×P×T).

Since both the algorithms are working in a sequential manner (i.e., optimization of the centroid points and then fine tuning of the points using k-means) the overall time complexity can be represented as O(M×N)+O(Q×P×T) and the asymptotic representation is O(Q×P×T).

The template is designed for, but not limited to, six authors. A minimum of one author is required for all journal articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

5. Conclusions

This research work is aimed to measure the air quality by estimating the pollutant level through a novel AQP system. AQP system plugs the advantage of metaheuristic algorithm called SCA integrated with k-means. The cluster centroid obtained from proposed improved k-means based on SCA is given to K-NNR for optimal estimation of pollutant. The novelty of this contribution is the utilization of the meta heuristic algorithm to optimize the prediction accuracy of the pollutant. Experimental analysis has been carried on air quality dataset taken from University of California, Irvine (UCI) repository. The proposed HYAQP is compared with conventional regression algorithms such as MLR, SVR and K-NNR. In all the levels of comparison, HYAQP achieves best because of the integration of SCA which intends to balance between diversification and intensification paving the way for optimal global search and local search thereby resulting in optimal cluster centroid. K-NNR is exposed with the optimal cluster centroid and test instances which not only reduces the complexity but also reduce the mean

absolute error and root mean squared error. Also, the purity of the cluster created using HYAQP has 8.3% and 22.97% greater level of purity than IKPSO and IKGA respectively. By proposed model can be adopted to any country or city just by replacing the dataset. The pollution level prediction claimed by the HYAQP is highly accurate and hence, this model can potentially be used in places where there is a high risk in air pollution level with high uncertainty. The proposal removes the barrier of false prediction by tuning the centroid position. Future work of this contribution is planned to collect data from real world using IoT sensors and also by using Unmanned Aerial Vehicle to locate the source that cause pollution.

References

[1] Abd Elaziz M., Nabil N., Ewees A., and Lu S., "Automatic Data Clustering Based on Hybrid Atom Search Optimization and Sine-Cosine Algorithm," *in Proceedings of the IEEE Congress on Evolutionary Computation*, Wellington, pp. 2315-2322, 2019. DOI:10.1109/CEC.2019.8790361

[2] Air Quality Data from London, https://datahub.io/core/london-air-quality#readme, Last Visited, 2024.

[3] Air Quality Data from UK-AIR https://data.world/datagov-uk/3f4405c6-0aa7-4087-8067-8efe0c364564, Last Visited, 2024.

[4] Athira V., Geetha P., Vinayakumar R., and Soman K., "Deepairnet: Applying Recurrent Networks for Air Quality Prediction," *Procedia Computer Science*, vol. 132, pp. 1394-1403, 2018. https://doi.org/10.1016/j.procs.2018.05.068

[5] Barai S., Dikshit A., and Sharma S., *Soft Computing in Industrial Applications*, Springer, 2007. https://link.springer.com/chapter/10.1007/978-3-540-70706-6_27

[6] Boushaki S., Bendjeghaba O., and Brakta N., "Accelerated Modified Sine Cosine Algorithm for Data Clustering," *in Proceedings of the IEEE 11th Annual Computing and Communication Workshop and Conference*, Nevada, pp. 0715-0720, 2021. DOI:10.1109/CCWC51732.2021.9376122

[7] Centers for Disease Control and Prevention-Carbon Monoxide Poisoning, https://www.cdc.gov/carbon-monoxide/about/?CDC_AAref_Val=https://www.cdc.gov/co/faqs.htm, Last Visited, 2024.

[8] Central Pollution Control Board, https://cpcb.nic.in/, Last Visited, 2024.

[9] Chen X., Hu Y., Dong F., Chen K., and Xia H., "A Multi-Graph Spatial-Temporal Attention Network for Air-Quality Prediction," *Process Safety and Environmental Protection*, vol. 181, pp. 442-451, 2024. https://doi.org/10.1016/j.psep.2023.11.040

[10] Chen X., Xia H., Wu M., Hu Y., and Wang Z., "Spatiotemporal Hierarchical Transmit Neural Network for Regional-Level Air-Quality Prediction," *Knowledge-Based Systems*, vol. 289, pp. 111555, 2024. https://doi.org/10.1016/j.knosys.2024.111555

[11] Corani G., "Air Quality Prediction in Milan: Feed-Forward Neural Networks, Pruned Neural Networks and Lazy Learning," *Ecological Modelling*, vol. 185, no. 2-4, pp. 513-529, 2005. https://doi.org/10.1016/j.ecolmodel.2005.01.008

[12] De Vito S., Massera E., Piga M., Martinotto L., and Di Francia G., "On Field Calibration of an Electronic Nose for Benzene Estimation in an Urban Pollution Monitoring Scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750-757, 2008. https://doi.org/10.1016/j.snb.2007.09.060

[13] Debnath J., Majumder D., and Biswas A., "Air Quality Assessment Using Weighted Interval Type-2Fuzzy Inference System," *Ecological Informatics*, vol. 46, pp. 133-146, 2018. https://doi.org/10.1016/j.ecoinf.2018.06.002

[14] Ding C., Zheng Z., Zheng S., Wang X., Xie X., Wen D., Zhang L., and Zhang Y., "Accurate Air-Quality Prediction Using Genetic-Optimized Gated-Recurrent-Unit Architecture," *Information*, vol. 13, no. 5, pp. 223, 2022. https://www.mdpi.com/2078-2489/13/5/223

[15] Environmental Protection Agency-Air Quality Index Basics, https://www.airnow.gov/aqi/aqi-basics/, Last Visited, 2024.

[16] Environmental Protection Agency-Ground-level Ozone Basics, https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics, Last Visited, 2024.

[17] Environmental Protection Agency-National Ambient Air Quality Standards, https://www.epa.gov/naaqs, Last Visited, 2024.

[18] Fang W., Zhu R., and Lin J., "An Air Quality Prediction Model Based on Improved Vanilla LSTM with Multichannel Input and Multiroute Output," *Expert Systems with Applications*, vol. 211, pp. 118422, 2023. https://doi.org/10.1016/j.eswa.2022.118422

[19] Green Peace, https://www.greenpeace.org/, Last Visited, 2024.

[20] Gupta S. and Deep K., "A Hybrid Self-Adaptive Sine Cosine Algorithm with Opposition Based Learning," *Expert Systems with Applications*, vol. 119, pp. 210-230, 2019. https://doi.org/10.1016/j.eswa.2018.10.050

[21] Jain A., "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010. https://doi.org/10.1016/j.patrec.2009.09.011

[22] Kang G., Gao J., Chiao S., Lu S., and Xie G., "Air Quality Prediction: Big Data and Machine

Learning Approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8-16, 2018. https://www.ijesd.org/show-103-1491-1.html

[23] Kaur M., Kaur R., Singh N., and Dhiman G., "SChoA: A Newly Fusion of Sine and Cosine with Chimp Optimization Algorithm for HLS of Datapaths in Digital Filters and Engineering Applications," *Engineering with Computers*, vol. 38, no. 2, pp. 975-1003, 2022. https://link.springer.com/article/10.1007/s00366-020-01233-2

[24] Kennedy J. and Eberhart R., "Particle Swarm Optimization," *in Proceedings of the International Conference on Neural Networks*, Perth, pp. 1942-1948, 1995. DOI:10.1109/ICNN.1995.488968

[25] Kok I., Simsek M., and Ozdemir S., "A Deep Learning Model for Air Quality Prediction in Smart Cities," *in Proceeding of the IEEE International Conference on Big Data*, Boston, pp. 1983-1990, 2017. DOI:10.1109/BigData.2017.8258144

[26] Kuo R., Lin J., and Nguyen T., "An Application of Sine Cosine Algorithm-based Fuzzy Possibilistic C-ordered Means Algorithm to Cluster Analysis," *Soft Computing*, vol. 25, no. 5, pp. 3469-3484, 2021. https://link.springer.com/article/10.1007/s00500-020-05380-y

[27] Liao H., Yuan L., Wu M., and Chen H., "Air Quality Prediction by Integrating Mechanism Model and Machine Learning Model," *Science of the Total Environment*, vol. 899, pp. 165646, 2023. https://doi.org/10.1016/j.scitotenv.2023.165646

[28] Lin Y., Lee S., Ouyang C., and Wu C., "Air Quality Prediction by Neuro-Fuzzy Modeling Approach," *Applied Soft Computing*, vol. 86, pp. 105898, 2020. https://doi.org/10.1016/j.asoc.2019.105898

[29] Liu X. and Guo H., "Air Quality Indicators and AQI Prediction Coupling Long-Short Term Memory and Sparrow Search Algorithm: A Case Study of Shanghai," *Atmospheric Pollution Research*, vol. 13, no. 10, pp. 101551, 2022. https://doi.org/10.1016/j.apr.2022.101551

[30] Ma J., Cheng J., Lin C., Tan Y., and Zhang J., "Improving Air Quality Prediction Accuracy at Larger Temporal Resolutions Using Deep Learning and Transfer Learning Techniques," *Atmospheric Environment*, vol. 214, pp. 116885, 2019. https://doi.org/10.1016/j.atmosenv.2019.116885

[31] Mirjalili S., "SCA: A Sine Cosine Algorithm for Solving Optimization Problems," *Knowledge-based Systems*, vol. 96, pp. 120-133, 2016. https://doi.org/10.1016/j.knosys.2015.12.022

[32] Mirjalili S., *Evolutionary Algorithms and Neural Networks*, Springer, 2019. https://link.springer.com/chapter/10.1007/978-3-319-93025-1_4

[33] Navares R. and Aznarte J., "Predicting Air Quality with Deep Learning LSTM: Towards Comprehensive Models," *Ecological Informatics*, vol. 55, pp. 101019, 2020. https://doi.org/10.1016/j.ecoinf.2019.101019

[34] Pan J., Fan F., Chu S., Zhao H., and Liu G., "A Lightweight Intelligent Intrusion Detection Model for Wireless Sensor Networks," *Security and Communication Networks*, vol. 2021, no. 1, pp. 1-15, 2021. https://onlinelibrary.wiley.com/doi/10.1155/2021/5540895

[35] Shakil M., Mohammed A., Arul R., Bashir A., and Choi J., "A Novel Dynamic Framework to Detect DDoS in SDN Using Metaheuristic Clustering," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 3, pp. 3622, 2022. https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3622

[36] Simsek M., Kok I., and Ozdemir S., "DeepFogAQ: A Fog-Assisted Decentralized Air Quality Prediction and Event Detection System," *Expert Systems with Applications*, vol. 251, pp. 123920, 2024. https://doi.org/10.1016/j.eswa.2024.123920

[37] Singh K., Gupta S., Kumar A., and Shukla S., "Linear and Nonlinear Modeling Approaches for Urban Air Quality Prediction," *Science of the Total Environment*, vol. 426, pp. 244-255, 2012. https://doi.org/10.1016/j.scitotenv.2012.03.076

[38] Sripada S. and Rao M., "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering," *Indian Journal of Computer Science and Engineering*, vol. 2, no. 3, pp. 343-346, 2011. file:///C:/Users/user/Downloads/BDCC-02-00005-v2.pdf

[39] Sui S. and Han Q., "Multi-View Multi-Task Spatiotemporal Graph Convolutional Network for Air Quality Prediction," *Science of the Total Environment*, vol. 893, pp. 164699, 2023. https://doi.org/10.1016/j.scitotenv.2023.164699

[40] Tillman D., Duong D., and Harding N., *Solid Fuel Blending: Principles, Practices, and Problems*, Elsevier Butterworth-Heinemann, 2012. https://lib.ugent.be/en/catalog/ebk01:2670000000157860

[41] U.S. Pollution Data, https://www.kaggle.com/sogun3/uspollution, Last Visited, 2024.

[42] Wang J. and Song G., "A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction," *Neurocomputing*, vol. 314, pp. 198-206, 2018. https://doi.org/10.1016/j.neucom.2018.06.049

[43] Wang J., Zhang X., Guo Z., and Lu H.,

"Developing an Early-Warning System for Air Quality Prediction and Assessment of Cities in China," *Expert Systems with Applications*, vol. 84, pp. 102-116, 2017. https://doi.org/10.1016/j.eswa.2017.04.059

[44] World Health Organization-Ambient Air Pollution, https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, Last Visited, 2024.

[45] Zhang W., Huang R., and Ye L., "Evaluation of Emission Reduction Performance of Power Enterprises Based on Least Squares Support Vector Machine," *The International Arab Journal of Information Technology*, vol. 21, no. 5, pp. 854-865, 2024. https://doi.org/10.34028/iajit/21/5/7

[46] Zhang Y., Wang Y., Gao M., Ma Q., Zhao J., Zhang R., Wang Q., and Huang L., "A Predictive Data Feature Exploration-based Air Quality Prediction Approach," *IEEE Access*, vol. 7, pp. 30732-30743, 2019. DOI:10.1109/ACCESS.2019.2897754

[47] Zhu D., Cai C., Yang T., and Zhou X., "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," *Big Data and Cognitive Computing*, vol. 2, no. 1, pp. 1-15, 2018. https://www.mdpi.com/2504-2289/2/1/5

Praveena Vasudevan is a Research Scholar from the Department of ECE, SRM Institute of Science and Technology, Deemed to be University, Kattankulatur, Chennai, India. She is currently working as an Assistant Professor in the Department of Computer Science Engineering, SRM Institute of Science and Technology, Deemed to be University, Ramapuram, Chennai, India, since 2010. She has conferred the M.Tech. degree in VLSI Design, from Bharath University, Chennai, India, in 2006. She has published 10 research papers in Journals, and presented 5 papers in Conferences. She has earned 15 years of academic experience from various reputed educational institutions. Her areas of interest are Machine Learning, IoT, and Artificial Intelligence.

Chitra Ekambaram obtained her B.E degree in ECE from Madras University and M.Tech. in VLSI Design from Bharath Institute of Higher Education and Research, Chennai, India. She was awarded Ph.D. Degree in Electronics and Communication Engineering, SRM Institute of Science and Techchnology, Kattankulathur, Chennai,India. She is an Assistant professor in Electronics and Communication Engineering, SRMIST, Kattankulathur, Chennai. She has 20+ years of Teaching Experience in Various Academic Institutions . She has presented and Published research Articles in Various Research Journals. Her research interest includes VLSI Low Power High Speed Design, DSP Structures and VLSI Design Automation, Internet of Things (IoT), Machine Learning.