Research Article

Iyad Abu Doush*, Khalid Sultan, Ahmad Alsaber, Dhari Alkandari, and Afsah Abdullah

Enhanced Jaya optimization for improving multilayer perceptron neural network in urban air quality prediction

https://doi.org/10.1515/jisys-2023-0310
received December 15, 2023; accepted April 07, 2024

Abstract: The multilayer perceptron (MLP) neural network is a widely adopted feedforward neural network (FNN) utilized for classification and prediction tasks. The effectiveness of MLP greatly hinges on the judicious selection of its weights and biases. Traditionally, gradient-based techniques have been employed to tune these parameters during the learning process. However, such methods are prone to slow convergence and getting trapped in local optima. Predicting urban air quality is of utmost importance to mitigate air pollution in cities and enhance the well-being of residents. The air quality index (AQI) serves as a quantitative tool for assessing the air quality. To address the issue of slow convergence and limited search space exploration, we incorporate an opposite-learning method into the Jaya optimization algorithm called EOL-Jaya-MLP. This innovation allows for more effective exploration of the search space. Our experimentation is conducted using a comprehensive 3-year dataset collected from five air quality monitoring stations. Furthermore, we introduce an external archive strategy, termed EOL-Archive-Jaya, which guides the evolution of the algorithm toward more promising search regions. This strategy saves the best solutions obtained during the optimization process for later use, enhancing the algorithm's performance. To evaluate the efficacy of the proposed EOL-Jaya-MLP and EOL-Archive-Jaya, we compare them against the original Jaya algorithm and six other popular machine learning techniques. Impressively, the EOL-Jaya-MLP consistently outperforms all other methods in accurately predicting AQI levels. The MLP model's adaptability to dynamic urban air quality patterns is achieved by selecting appropriate values for weights and biases. This leads to efficacy of our proposed approaches in achieving superior prediction accuracy, robustness, and adaptability to dynamic environmental conditions. In conclusion, our study shows the superiority of the EOL-Jaya-MLP over traditional methods and other machine learning techniques in predicting AQI levels, offering a robust solution for urban air quality prediction. The incorporation of the EOL-Archive-Jaya strategy further enhances the algorithm's effectiveness, ensuring a more efficient exploration of the search space.

Keywords: multilayer perceptron neural network, air quality prediction, training neural network, archive technique, Jaya algorithm

*** Corresponding author: Iyad Abu Doush,** The College of Engineering and Applied Sciences, American University of Kuwait, Salmiya, Block 2, P.O. Box 3323, Safat 13034, Kuwait; Computer Science Department, Yarmouk University, Irbid, 21163, Jordan, e-mail: idoush@auk.edu.kw
Khalid Sultan: The College of Engineering and Applied Sciences, American University of Kuwait, Salmiya, Block 2, P.O. Box 3323, Safat 13034, Kuwait, e-mail: ksultan@auk.edu.kw
Ahmad Alsaber: Department of Management, College of Business and Economics, American University of Kuwait, Salmiya, Safat 13034, Kuwait, e-mail: aalsaber@auk.edu.kw
Dhari Alkandari: Department of Earth and Environmental Sciences, Kuwait University (Shadadiya Campus), Al-Shadadiya, P.O. Box 12422, Kuwait University City, 13071 Kuwait City, Kuwait, e-mail: dhary.alkandary@ku.edu.kw
Afsah Abdullah: The Office of Research and Grants, American University of Kuwait, Salmiya, Safat 13034, Kuwait, e-mail: aabdullah1@auk.edu.kw

1 Introduction

The tendency toward increased global air pollution has grown clearer with the spread of high-intensity human engineering activities [1–3]. Extreme weather is a frequent occurrence that has contributed to numerous disasters, and air pollution poses a serious threat to human health [4–7]. Population mortality has dramatically increased as a result of high $PM_{2.5}$ and PM_{10} concentrations [8–10]. In 2016, outdoor air pollution caused about 4.09 million deaths worldwide, according to the global burden of disease (GBD) report [11,12].

Large portions of the Gulf Cooperation Council (GCC) region fall under the dry and semi-arid categories, and when the correct wind conditions are present, these regions serve as a front for saltated sand and dust that fuel sandstorms. Dust and sandstorms cause land degradation in addition to the direct effects they have on particulate matter (PM) concentrations, confirmed to have negative health effects [13]. They also cause sand to encroach on important structures, topsoil removal, loss of visibility, soil fertility loss, crop failure, and damage [14]. In addition, the effectiveness of solar cells in capturing energy is decreased by the development of thin crusts of mud and/or carbonate coatings. The loss of animals and flora and the decline of the vegetation on the land constitute land degradation. Because it harms natural resources, biodiversity is a major concern for the GCC desert ecosystem. For instance, Kuwait had a 2.78-fold increase in dust deposition from 109.4 t/km^2 between 1974 and 1980 to 392 t/km^2 between 2011 and 2017 [15,16]. The Arabian Gulf's northern region, which includes Kuwait, experiences more dusty days than its southern region, with a mean total annual number of 255 days [15,17–19].

A recent study using several methods, such as the positive matrix factorization model, concentration rose plots, and backward path profiles, demonstrated that a significant portion of Kuwait's PM originates from nearby continents or states [20]. According to the same study, Kuwait's yearly average PM levels – particles with an aerodynamic diameter smaller than 10 μm and $PM_{2.5}$ levels – particles with an aerodynamic diameter less than 2.5 μm – were 130 and 53 μg/m^3, respectively. Sand dust was identified as Kuwait's primary source of $PM_{2.5}$, accounting for 54% of the country's total, followed by oil combustion (18%), the petrochemical sector (12%), with traffic (11%), also, more than 50% of the analyzed $PM_{2.5}$, according to the same study, had sources other than Kuwait.

A more recent study revealed that the production of power and the desalination of water accounted for 23 and 14%, respectively, of the anthropogenic $PM_{2.5}$ and PM_{10} emissions. The yearly mean of $PM_{2.5}$ concentration for urban areas in Kuwait consistently exceed the WHO suggested limit, which is 35 μg/m^3 [21]. Another intriguing finding was that the difference in yearly mean concentrations between rural and urban areas remained almost constant at 8 μg/m^3.

Modern human activities inevitably involve energy usage and its consequences. Kerosene, coal, and straw burning are only a few human-caused bases of air pollution, along with emissions from businesses, vehicles, and aerosol cans. Day by day, a variety of detrimental pollutants, including NO_2, CO, NH_3, PM, CO_2, O_3, Pb, and SO_2, are discharged into the atmosphere.

The substances and particles that constitute air pollution affect the health of people, animals, and even plants. Heart disease, bronchitis, and pneumonia are just a few of the severe illnesses that can be contracted by people as a result of exposure to air pollution. Poor air quality is the root cause of smog, acid rain, aerosol formation, vision impairment, global warming, early death, and other contemporary environmental issues. Researchers have found that air pollution has the propensity to destroy historical monuments [22–24]. The atmospheric emissions from factories and power stations, agriculture discharges, and other sources are to blame for the rise in greenhouse gases. The climate is detrimentally impacted by greenhouse gases, which in turn affects how quickly plants develop [25].

Plant–soil interactions are also affected by emissions of greenhouse gases and inorganic carbons [26]. Climate change has a significant impact on agricultural output as well as human and animal health [27]. Commercial loss is another side effect that is sometimes worth taking into account. The air quality index (AQI), a measurement variable, has a direct bearing on public health. Higher levels of AQI denote a riskier exposure for the masses. Consequently, the yearning to obtain accurate anticipation for AQI drove investigators to monitor, model, and analyze the air quality. Based on rising automotive and manufacturing activities,

assessing and predicting AQI, especially in metropolitan areas, have become an essential and difficult undertaking. Since the concentration of the deadliest pollutant, $PM_{2.5}$, is shown to be multiplied in developing countries, the majority of air quality studies and research efforts focus on these nations [28].

Many models, including statistical, physical, deterministic, and machine learning (ML) approaches, have been employed to predict AQI in the literature. The conventional methods based on statistics and probability are exceedingly intricate and ineffective. It has been proven that ML-based AQI prediction models are more dependable and consistent. Data collection was made simple and accurate by modern technologies and sensors. ML algorithms are capable of handling the rigorous analysis needed to make accurate and trustworthy predictions from such vast environmental data. The significance of ML techniques in dealing with the challenges related to environment protection was extensively explored by Al-Jamimi et al. [29].

In fact, ML and statistical forecasting are the two prominent tools for predicting AQI. To map the association between the goal data and the time-series historical data, statistical forecasting techniques create data-driven mathematical models. These techniques can offer precise and timely predictions with a straightforward mathematical basis. A popular statistical forecasting method that is typically used for short-term forecasting is the auto-regressive integrated moving average (ARIMA). A recent survey of traditional statistical models for AQI forecasting highlights that ARIMA models are better at identifying trends and making forecasts with the lowest root-mean-square error (RMSE) when equated to other statistical models [30]. However, the bulk of statistical-based models do not take atmospheric variables and circumstances into consideration when forecasting future data; instead, they simply use the previously recorded data. In addition, statistical models, unlike ML models, call for computationally intensive data pre-processing, particularly when dealing with historical data's discontinuity [31].

Nevertheless, once combined with environmental applications, ML algorithms, with their demonstrated dominance and effectiveness in many predicting issues, might be quite alluring to academics. For instance, Wang et al. [32] used a radial basis neural network (NN) to estimate SO_2 levels and came to the conclusion that the outcomes could be useful for AQI forecasting in the future. In a similar vein, Cai et al. [33] demonstrated that in terms of predicting various pollutant concentrations, a feedforward NN outperformed multilinear regression. For predicting the pollutant levels that are needed to compute the AQI (per hour) in California, the work proposed by Castelli et al. [34] used the support vector machine (SVM), which is another reliable ML approach. The classification accuracy of the models was 94.1% for the projected air quality. A novel method of forecasting the AQI directly using ensemble learning was suggested Ganesh et al. [35]. In this method, the projected AQI amounts from five separate regression and ML models were additionally processed and then introduced to ensemble models to improve forecasting accuracy.

The current study examines 3 years' worth of data on air pollution collected by five environmental monitoring stations in the state of Kuwait covering the period from January 1, 2018, and December 31, 2021. Data visualization techniques are then used to improve insights and look into undiscovered patterns and trends after the dataset has been cleansed and prepossessed. This study uses the JAYA algorithm model to forecast the AQI in Kuwait.

The JAYA algorithm, developed by Roa in 2016, stands out as an exceptionally efficient metaheuristic approach. Inspired by the concept of "survival of the fittest," this algorithm enables solutions within the population to converge toward the most optimal solution while avoiding inferior ones [36].

One of the key advantages of the JAYA algorithm lies in its simplicity and ease of implementation, as it requires no algorithm-specific parameters. Its conceptual and computational straightforwardness has also contributed to its popularity among researchers [37,38]. Researchers have employed the JAYA algorithm to tackle various optimization problems, ranging from truss structures and parameter estimation to facial emotion recognition, micro-channel heat sinks, optimal power flow, job-shop rescheduling, knapsack problems, feature selection, and shell-and-tube heat exchangers [39–47].

Despite its merits, the JAYA algorithm does suffer from certain shortcomings, notably related to premature convergence. When reaching the equilibrium state, it may become trapped in local optima [48]. To address this issue, researchers have proposed modified versions of the JAYA algorithm, such as the self-adaptive, multi-population, chaotic, and performance-guided approaches [38,48–50]. By continually refining and enhancing

the JAYA algorithm and its variants, researchers aim to harness its potential to solve optimization problems with increased efficiency and accuracy.

In this article, elite opposition-based learning (EOL), which is a transformative enhancement, is integrated into the Jaya algorithm focusing on the elite solutions, or best solutions for the sake of revolutionizing solution exploration. The new proposed algorithm is called EOL-Jaya-MLP, which enables the algorithm to simultaneously explore promising elite solutions while also probing the antitheses of these elite points. By striking a harmonious balance between exploitation and exploration, EOL empowers the Jaya algorithm to efficiently and effectively converge toward near-optimal solutions, mitigating premature convergence and enhancing its applicability to complex, real-world optimization challenges. It significantly improves solution precision and speed, as it targets exploration efforts around the elite solutions, steering clear of local optima.

Researchers have introduced archive methods as a powerful approach to enhance population diversity during evolution and preserve potential optima. One notable method, proposed by Lacroix et al. [51], involves collecting the best-known solutions within the evolving population into a single archive. This archive serves as an index to identify weak regions in the search space for exploration, which are then stored in another collection. Throughout evolution, both archives are continuously updated, ensuring a balanced exploration–exploitation trade-off.

Another approach, presented by Zhang et al. and Kundu et al. [52,53], utilizes an archive method for subpopulations. This facilitates regeneration to form an initial population of solutions, promoting diversity and avoiding premature convergence. In the work of Wang et al. [54], an archive method is employed to store stagnant solutions. These solutions are periodically reinitialized along with their neighboring individuals whose fitness is lower, fostering rejuvenation within the population.

Moreover, Sheng et al. and Turky et al. [55,56] have successfully utilized archives to address dynamic optimization problems. By continuously updating the archives, these methods aim to improve population evolution and retain the best potential solutions for subsequent cycles. Empirical findings strongly support the efficacy of archive methods, showing their superiority over other approaches in reliably locating multiple optimal solutions in the problem search space. Consequently, researchers have also leveraged archives to tackle multiobjective optimization problems, showing their versatility and effectiveness in various contexts.

To enhance the exploration capabilities of EOL-Jaya-MLP, an archive strategy is implemented to store the best results, inspired by the methods introduced in the literature [51,54,57]. This modification, named EOL-Archive-Jaya, empowers the algorithm to search more effectively in promising regions. EOL-Archive-Jaya is specifically employed to optimize the training of a single hidden layer NN for enhanced performance in predicting the AQI. The proposed algorithm is evaluated using a 3-year dataset collected from five air quality monitoring stations, comparing its performance against the original Jaya algorithm and three other well-known swarm intelligence algorithms (bat algorithm, particle swarm optimization (PSO), and Moth-flame optimization) as well as six classical ML techniques. The results demonstrate the effectiveness of EOL-Archive-Jaya in obtaining superior solutions and efficiently training the NN, signifying its potential as a valuable tool in optimization and ML domains.

The remaining parts of the present article are organized as follows: in Section 2, we discuss the related work; Section 3 provides an explanation of the multilayer perceptron (MLP); our novel algorithms, EOL-Jaya-MLP and EOL-Archive-Jaya, are presented in Section 4; we present experimental results and discussions in Section 5, thoroughly analyzing the performance of the proposed algorithms; finally, in Section 6, we conclude the article by summarizing key findings and discussing potential future developments. This structured approach enables readers to seamlessly navigate through the article, gaining a comprehensive understanding of our contributions and their implications.

2 Related works

The power of NNs, including MLP NNs and radial basis function networks, in conjunction with optimization algorithms like the Chimp optimization algorithm (ChOA) have been harnessed by AI research community to tackle various problems such as detecting and identifying marine life in underwater settings [58–66]. The

proposed techniques have significantly elevated the efficacy of underwater image detection, and facilitated the development of proficient sonar systems. Moreover, the integration of evolutionary NNs and fuzzy system concepts has led to the refinement of classification strategies for acoustic targets [67].

This section reviews the relevant works that involve the use of NNs and optimization algorithms to predict air quality. Lu et al. [68] introduce an approach to enhance the performance of the back-propagation (BP) NN in predicting the AQI. Their method involves leveraging an advanced PSO algorithm, resulting in improved optimization capabilities. The researchers utilize the enhanced PSO algorithm to optimize not only the different strategies related to the inertia weight but also the learning factor. This optimization process ensures that the algorithm exhibits high global search capability from the initial stages and facilitates fast convergence toward the optimal solution. To prevent the particles from getting trapped in local optima, the authors propose a solution based on an adaptive mutation during the search phase. This algorithm effectively avoids the particles from being stuck in suboptimal solutions. Through in-depth analysis and thorough comparison of empirical results, the researchers provide compelling evidence that the BP NN optimized using the improved PSO algorithm achieves superior accuracy in predicting the AQI.

In their research, Seng et al. [69] present an deep learning-based approach to predict air quality. Their method involves the development of a comprehensive prediction model based on long short-term memory (LSTM). This model incorporates multioutput and multiindex of supervised learning (MMSL) techniques. To create their prediction model, the researchers integrate particle concentration data from the current monitoring station with data from neighboring stations, as well as meteorological and gaseous pollutant data from the same period. The LSTM is employed for training the model, enabling the prediction of pollution indicators. To evaluate the model's performance, the researchers utilize a dataset collected from 35 air monitoring stations in Beijing area for a period of 2 whole years, namely, 2016 and 2017.

The study conducted by Cordova et al. [70] focuses on analyzing air quality in Peru, specifically in Metropolitan Lima. They employed the LSTM networks to assess and predict air pollution levels. The researchers gathered data from five monitoring stations and evaluated their model using two different schemes. First, they used the hold-out (HO) method, and second, they used the blocked-nested cross-validation (BNCV). The experimental results yielded promising outcomes, particularly regarding PM_{10} concentrations during periods of low pollution. Moreover, the LSTM network combined with BNCV exhibited superior predictability performance for periods characterized by high contamination levels. Focusing on the Amman area (Jordan), Aljanabi et al. [71] aimed to predict ozone concentration, specifically at the ground level. The authors developed a model that utilizes a combination of meteorological and seasonal variables from the previous day to make accurate predictions. To enhance the predictive performance, the researchers compared multiple algorithms: MLP NN, support vector regression (SVR), decision tree regression (DTR), and extreme gradient boosting (XGBoost). In addition, they investigated the impact of applying different smoothing filters to the time-series data, including moving average, Holt-Winters smoothing, and Savitzky-Golay filters. The findings revealed that the MLP algorithm outperformed the other algorithms in predicting ozone concentration. Furthermore, the authors implemented feature selection techniques to reduce the prediction time. The experimental results demonstrated a significant reduction of 91% in prediction time, along with a decrease in the number of features required for prediction. Specifically, the selected features included ozone, humidity, and temperature values from the previous day.

In addition, Al-Rashed et al. [72] focus also on the air quality in Kuwait. The researchers investigated the variations in air pollutant trends and employed two analytical methods: (1) exponentially weighted moving average (EWMA) scheme and (2) the cumulative sum (CUSUM) technique. The EWMA technique was utilized to calculate the average mean, while the CUSUM method was applied to extract shifts from this average mean. The objective of their study was to investigate the trends of major pollutants in three specific areas in Kuwait over a 5-year period. Data for the study were collected from three monitoring stations located in Ali Subah Al-Salem area, Al-Mutla district, and Al-Mansouriya area. The results indicated that the average trends for carbon monoxide (CO) and non-methane hydrocarbon (NMHC) in all three study areas exceeded the standard concentration levels established in Kuwait and the World Health Organization (WHO) guidelines. In the study by Zhao et al. [73], the researchers employ a deep learning approach to forecast air quality classification (AQC) in three distinct industrial cities within the United States. They utilize a recurrent neural network (RNN), a

popular deep learning technique, to construct their prediction model. Through extensive experimentation, the authors demonstrate that their proposed model exhibits superior performance compared to SVM and random forest models. This highlights the effectiveness and superiority of the RNN-based approach in predicting AQC for the given industrial cities.

Ebrahimi-Khusfi et al. [74] shift their focus to another aspect of air pollution, namely, dust concentration (DC). They introduce an approach to investigate the uncertainty and interpretability of the adaptive neuro-fuzzy inference system (ANFIS), a robust model utilized for dust event prediction. To further enhance their analysis, the authors quantify the uncertainty of the ANFIS model in which they use uncertainty estimation techniques that, in turn, utilize local errors and clustering tools. They also employ a model-agnostic interpretation approach to improve the interpretability of the ANFIS model. In addition, the authors incorporate the bat optimization algorithm (BAT) to enhance the performance of the ANFIS model in prediction. The results obtained from their study demonstrate the effectiveness of the ANFIS+BAT model in predicting DC. In comparison to the ANFIS model, the ANFIS+BAT model exhibits an improvement of 10 and 16% in correlation coefficient for predicting DC during the cold and warm months, respectively. Moreover, the uncertainty analysis reveals a reduced prediction interval when the BAT algorithm is integrated compared to the original ANFIS model.

Maleki et al. [75] devises an artificial neural network (ANN) model with the aim of predicting hourly standard air pollutant concentrations, AQI, and air quality health index. Through an analysis of air pollution data collected in Ahvaz, Iran, spanning from August 2009 to August 2010, they assess the performance of their model by obtaining the predicted correlation coefficient of 0.87 and a RMSE of 59.9. The results highlight the potential of ANNs in predicting air quality, thus enabling proactive measures to prevent potential health impacts. By utilizing this ANN model, policymakers and relevant stakeholders can make informed decisions to mitigate the adverse effects of air pollution on public health.

Yang et al. [76] introduce a hybrid optimization framework aiming to enhance air quality assessment. The primary objective of their model revolves around accurately forecasting the concentration of $PM_{2.5}$, a crucial air pollutant. The framework combines the power of the least-square support vector machine (LSSVM) with the optimization capabilities of the JAYA optimization algorithm, resulting in what is known as JAYA-LSSVM. One notable aspect of the JAYA-LSSVM approach is its comprehensive consideration of various air pollutants, including PM_{10}, SO_2, NO_2, CO, and O_3. In addition, the model takes into account crucial climate conditions such as wind patterns, temperature, sunlight intensity, and humidity. By incorporating these factors, the researchers aim to provide a more holistic and accurate understanding of air quality dynamics. We summarize the most relevant recent research work in Table 1.

3 MLP

The feedforward neural network (FNN) is a powerful type of ANN that emulates the interconnectedness of neurons in the human brain to model knowledge. It comprises three interconnected layers: input, hidden, and output. The input layer receives the features from the data, while the hidden layers process this information to recognize the desired output patterns based on the input features. Finally, the output layer identifies the predicted classes corresponding to the given set of inputs [79].

A specific class of FNN is the MLP, where data move in a unidirectional manner, passing through the network from the input layer to the output layer. The parameters of an MLP include the input features, the weights (w), and the biases (b). The MLP's outcome is computed in three distinct phases:

- **Initializing weights**: The weighted sum score is assigned for each input in the MLP network. The weighted sum is computed using equation (1).

$$S_j = \sum_{i=1}^{n} (w_{ij}. X_i) - \beta_j, \quad j = 1, 2, \ldots, h. \tag{1}$$

Table 1: Summary of related work

Study	Method	Data resolution	Pollutants	Location of data collection
[68]	NN with PSO-BP	Daily	PM 2.5, PM 10, O_3, NO_2, CO, SO_2	China AQ monitoring and analysis platform
[69]	LSTM with MMSL	Hourly	$PM_{2.5}$, CO, NO_2, O_3, SO_2	Beijing, China
[70]	LSTM with HO and BNCV	Hourly	PM_{10}	Metropolitan Lima - Peru
[71]	MLP, SVR, DTR, and XGBoost algorithms	Daily	Ozone (ppb)	Amman, Jordan
[77]	Multicriteria decision-making process	Hourly	O_3 in NO_2 and NOx	City of Wroclaw
[78]	Deep learning armed by a RNN with LSTM	Hourly	O_3 concentration	Kuwait
[72]	weighted moving average (EWMA) and CUSUM analyzing methods	Daily	CO, NO_2, SO_2, PM_{10}, H2S, NMHC	Kuwait
[73]	RNN of deep learning	Daily	CO, NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10}	three industrial cities in USA
[74]	Adaptive neuro-fuzzy inference system (ANFIS)	Hourly	Air temperature, precipitation (P), wind speed (WS), relative humidity (RH)	semi-arid regions of Iran
[75]	ANN	Hourly	O_3, NO_2, PM_{10}, $PM_{2.5}$, SO_2, and CO	Ahvaz, Iran
[76]	Hybrid optimization prediction involving: (1) complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), (2) variational mode decomposition, and LSSVM	Daily	PM_{10}, SO_2, NO_2, CO, and O_3	Xi'an and Shenyang

In an MLP, it is important to note the following definitions: First, n represents the total number of inputs present in the MLP architecture. Each input, denoted as X_i, is connected to a hidden neuron j through a weight vector, w_{ij}, which governs the influence of that specific input on the activation of the hidden neuron. In addition, the hidden neuron j is equipped with a bias term, β_j, which plays a significant role in determining the neuron's overall activation process.

- **Activation function computation**: The Sigmoid activation function is commonly used to process the weighted vector output. After that, the computed output vector is transferred to the next layer. As the Sigmoid output value is between 0 and 1, it is not used in our proposed framework. The Leaky ReLU activation function is used instead [80]. The calculation is presented in equation (2).

$$\text{Leaky ReLU} = \max(\alpha x, x). \tag{2}$$

Note that the value of hyperparameter α is between 0.01 and 0.1.

- **Computing the outcome**: The outcome of the last layer is calculated as shown in equation (3):

$$\hat{y}_k = \sum_{i=1}^{m} w_{kj} f_i + b_k. \tag{3}$$

It is worth noting that w_{jk} refers to the weight associated with the connection from the hidden neuron j to the output neuron k in the NN. This weight value influences how the hidden neuron's activation contributes to the final output neuron's result. Moreover, the output neuron k is accompanied by a bias term denoted as b_k, which plays a critical role in adjusting the output neuron's activation threshold.

Finally, the optimal set of weights and biases are used in the MLP as shown in equations (1) and (3) to achieve better classification accuracy.

4 Methodology

This section presents the main contributions of this work. The MLP is enhanced by proposing an EOL Jaya optimization algorithm (EOL-Jaya-MLP) for predicting air pollution in Kuwait. The utilization of EOL is to improve the algorithm exploration power by introducing a new set of solutions generated using the current best solutions. After that, the Enhanced Archive-Jaya (EOL-Archive-Jaya) technique is proposed to enhance the performance of the MLP NN. As previously discussed, the key variables in MLP NNs are the weights and biases. EOL-Archive-Jaya optimizes these parameters by selecting MLPs that achieve the highest classification accuracy.

The groundbreaking feature of EOL-Archive-Jaya lies in its integration of an external archive, designed to preserve the best solutions found during the optimization process for utilization in future iterations. In each iteration, EOL-Archive-Jaya leverages the MLP to evaluate the current solutions, with the weights and biases serving as input vectors. The archive rate (A_r) determines the proportion of the finest solutions from the population, encompassing the most optimal weights and biases for the fittest MLP. These elite solutions are then stored in the external archive and subsequently incorporated into the initial population for the next run. This intelligent approach enables EOL-Archive-Jaya to capitalize on the best solutions acquired thus far throughout the algorithm's evolution, leading to enhanced performance and convergence.

The iterative process, involving the use of the external archive, is repeated in every run. In the first run, since there is no feedback from historical runs, the population is randomly constructed. However, once the archive is available after the first run, it comes into play, significantly influencing subsequent iterations. By leveraging the collective knowledge captured in the external archive, EOL-Archive-Jaya optimizes the selection of weights and biases, driving the MLP toward improved classification accuracy. This dynamic and iterative approach ensures that the algorithm efficiently explores the solution space, leading to enhanced performance in training the MLP NNs.

The steps of the proposed method are demonstrated in Figure 1. The proposed EOLJaya-MLP and EOL-Archive-Jaya will improve the NN performance by choosing the most promising values of weights and biases vectors, thus providing a better prediction for air pollution.

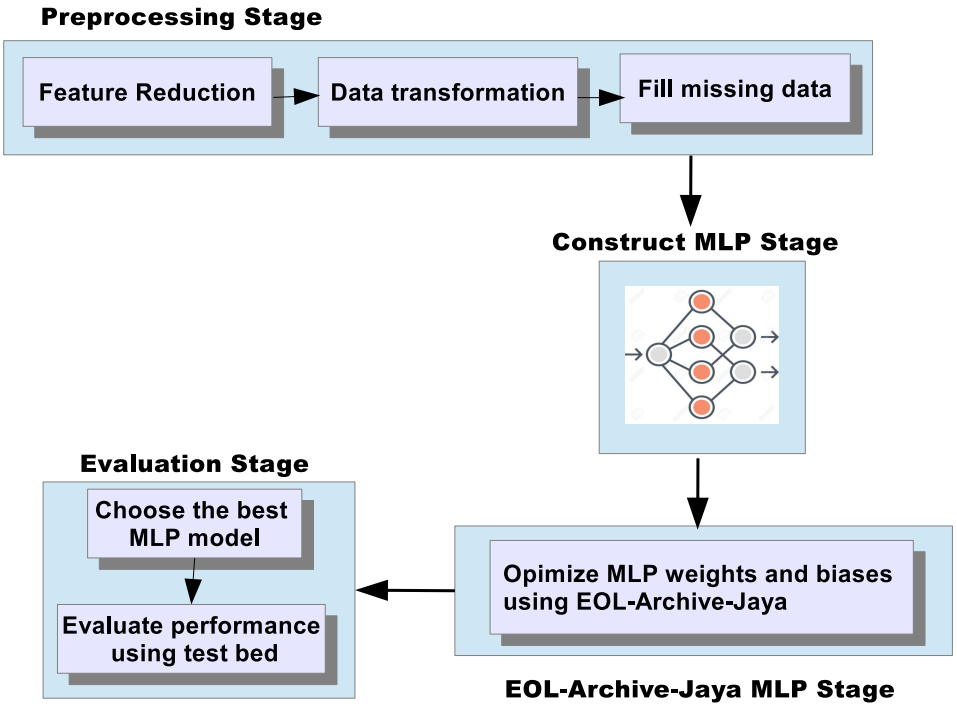

Figure 1: The stages of the proposed method [81].

4.1 Jaya optimization algorithm with EOL and external archive for MLP

JAYA algorithm is a metaheuristic algorithm inspired by the nature evolutionary concept "survival of the fittest" [36]. JAYA algorithm tries to amplify the number of fittest candidate solutions and to avoid the worst solutions. The strength of JAYA algorithm comes from its simplicity, ease of implementation, and no algorithm-specific parameters [82]. The procedural steps of the Jaya optimization algorithm with EOL and external archive for MLP are demonstrated next:

Step 1: **Define the external archive**. The external archive, denoted as ARCH, is represented as a matrix with dimensions $K \times N$, as shown in equation (4). Here, N corresponds to the total number of weights and biases present in the solution vector. The value of K is determined by selecting the top-performing MLPs after each training session for inclusion in the archive. The specific value of K is determined by the equation (5), where A_r refers to the archive rate. The parameter A_r is an important component of the EOL-Archive-Jaya algorithm and is carefully chosen through a preliminary experiment.

It is worth noting that the ARCH matrix starts as an empty set during the first run and will be subsequently updated with the selected MLPs after the initial training session. This approach allows for the preservation and storage of the best-performing solutions for further use and analysis in the algorithm.

$$\boldsymbol{ARCH} = \begin{bmatrix} w_1^1 & \cdots & w_n^1 & b_1^1 & \cdots & b_m^1 \\ w_1^2 & \cdots & w_n^2 & b_1^2 & \cdots & b_m^2 \\ \vdots & \vdots & \cdots & \vdots & & \\ w_1^K & \cdots & w_n^K & b_1^K & \cdots & b_m^K \end{bmatrix}. \tag{4}$$

$$K = HIS \times A_r. \tag{5}$$

The parameter A_r holds significant importance as it influences the extraction of the best solutions from the previous run. Specifically, A_r represents the archive rate, which corresponds to the ratio of MLPs

selected from the training set to be included in the next version of the archive. The entire population size, denoted as HIS, determines the number of solutions, each comprising a vector containing weights and biases for individual MLPs.

Importantly, the size of the archive, denoted as ARCH, is dependent on the population size, and this relationship is defined by the archive rate (A_r). Specifically, the archive size is calculated as a proportion of the population size. As a result, the ARCH matrix adapts to changes in the population size based on the specified archive rate.

Furthermore, it's worth mentioning that the construction of the ARCH matrix occurs only at the beginning of the execution, and it undergoes continuous updates after each run. This approach ensures that the best-performing solutions are preserved and incorporated into the evolving archive as the algorithm progresses.

Step 1: **Initialization parameters of JAYA algorithm and MLP parameters**. The MLP NN can be modeled as an optimization problem using a one-dimensional vector with a set of weights and biases to be tuned to increase the fitness of the MLP on the input features. The number of weights and biases can be computed using equation (6):

$$h = 2 \times F + 1, \tag{6}$$

$$n1 = N \times h + h \times o, \tag{7}$$

$$b = h + o. \tag{8}$$

The hidden layer contains a number of neurons represented by h, which plays a crucial role in the network's capacity to learn complex patterns. The dataset comprises a total of F features, where each feature serves as input to the network's neurons. The number of weights in the MLP is represented by g, and these weights are essential in determining the strength of connections between neurons. The output layer generates the final predictions, and the number of outputs from the MLP NN is indicated by o. In addition, the biases, which contribute to the overall flexibility and generalization ability of the network, are denoted by b.

The MLP is used to predict the output for each instance in the dataset, and the prediction accuracy is measured using mean squared error (MSE). The MSE is a common metric that calculates the difference between the MLP output (prediction) and the actual data. The MSE equation is demonstrated in equation (9). Note that y is the actual value, $\hat{y}$ is the predicted value, and k is the number of training samples.

$$\text{MSE} = \frac{1}{k}\sum_{i=1}^{k}(y - \hat{y})^2. \tag{9}$$

The predicted output, denoted as $\hat{y}$, is obtained by feeding the current weights and biases into the MLP and determining the corresponding output for each data input. To assess the performance of the MLP, which is characterized by its specific set of weights and biases, a comparison is made between the MLP's output and the actual output for the given data.
Generally speaking, the classification problem can be presented as follows:

$$\min_{x} f(\boldsymbol{x}) \quad \boldsymbol{x} \in [\boldsymbol{lb}, \boldsymbol{ub}]. \tag{10}$$

The value $f(\boldsymbol{x})$ represents the MSE calculated for a specific case, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$. In the context of the MLP, the vector $\boldsymbol{x}$ encompasses both the weights (n) and biases (m), given by $\boldsymbol{x} = (w_1, w_2, \ldots, w_n, b_1, b_2, \ldots, b_m)$. Here, x_i denotes the decision variable indexed by i, and there are a total of N decision variables in each individual, with $N = n + m$. The weights and biases for the MLP are confined to an interval $[lb_i, ub_i]$, where lb_i and ub_i represent the lower and upper limits of the variable x_i respectively. This constraint ensures that the weight and bias values stay within reasonable bounds, facilitating effective learning and preventing extreme parameter values.

The two parameters of the JAYA algorithm are: (i) The population size (i.e., P), and (ii) the number of iterations (i.e., *MaxIter*).

Step 2: **Constructing the initial population**. In this step, a set of solutions P is randomly generated. Each solution represents the weights and biases as input for the MLP. Each solution is a vector $\boldsymbol{x} = (w_1, w_2, \ldots, w_n, b_1, b_2, \ldots, b_m)$.

$$\boldsymbol{Population} = \begin{bmatrix} w_1^1 & \cdots & w_n^1 & b_1^1 & \cdots & b_m^1 \\ w_1^2 & \cdots & w_n^2 & b_1^2 & \cdots & b_m^2 \\ \vdots & \vdots & \cdots & \vdots & & \\ w_1^P & \cdots & w_n^P & b_1^P & \cdots & b_m^P \end{bmatrix}. \tag{11}$$

Note that each row represents a solution $\boldsymbol{x}^j$ which is a set of weights and biases. The solution is developed using the following equation: $x_i^j = lb_i + (ub_i - lb_i) \times U(0,1), \forall i = 1, 2, \ldots, P$. The cost function is computed using MSE as presented in equation (9). For simplicity, the term x_i^j is used next to refer to the variable i (weight or bias) of a solution vector j.

It is important to highlight that following the initial run, a portion of solutions, denoted as K, is directly copied from the archive ARCH, while the remaining solutions are randomly generated to complete the population.

Step 3: **Elite opposition-based learning**. Elite opposition-based learning (EOL) is a special type of learning proposed by Sihwail et al. [83]. The technique insight is obtained from developing a new set of solutions that is the opposite of elite (best) solutions with the hope to come closer to the global optimal. EOL is used to enhance the learning of different optimization algorithms such as harris hawks [83], cuckoo search [84], and grasshopper optimization [85].

Let $\vec{X} = \langle x_1, x_2, \ldots, x_m \rangle$ be the elite candidate solution that has m decision variables. EOL can be applied to the candidate solutions by calculating the elite opposite-based solution $\vec{X}^\circ$ using equation (12).

$$\vec{X}^\circ = \langle x_1^\circ, x_2^\circ, \ldots, x_m^\circ \rangle, \quad \text{where } x_i^\circ = \delta(da_i + db_i) - x_i. \tag{12}$$

Note that $\delta \in (0, 1)$ and it controls the amount of opposition, and da_i and db_i are the dynamic boundaries which are computed as follows:

$$da_i = \min(x_i), \; db_i = \max(x_i). \tag{13}$$

A correction mechanism is used if the opposite decision variable x_i° is outside $[LB_i, UB_i]$ using the equation:

$$x_i^\circ = \text{rand}(LB_i, UB_i), \textit{ if } x_i^\circ < LB_i \textit{ or } x_i^\circ > UB_i. \tag{14}$$

Note that $\text{rand}(LB_i, UB_i)$ is a random number between LB_i and UB_i. This technique is used to improve the set of generated weights and biases in the population.

Step 4: **Selecting the best and the worst solutions**. The best candidate solution and the worst candidate solution from the population are selected (i.e., $\boldsymbol{x}^{\text{best}}$ and $\boldsymbol{x}^{\text{worst}}$). Note that $\boldsymbol{x}^{\text{best}}$ is the solution (i.e., set of weights and biases) with the minimum MSE, while $\boldsymbol{x}^{\text{worst}}$ is the solution with the maximum MSE.

Step 5: **Improvement process**. The EOL-Jaya-MLP algorithm is used to improve the weights and biases of MLP. The weight or bias (x_i^j) (i.e., $x_i^j = w_j^i$ or $x_i^j = b_j^i$) for the individual $\boldsymbol{x}^j$ stored in the population are stochastically modified at each iteration using the following equation:

$$x_{i,j}^{\text{new}} = x_{i,j} + r_1 \times (x_j^{\text{best}} - |x_{i,j}|) - r_2 \times (x_j^{\text{worst}} - |x_{i,j}|). \tag{15}$$

In the context of candidate solution evaluation, it is important to consider the value of the decision variable j for both the best candidate solution, denoted as $\boldsymbol{x}^{\text{best}}$, represented by x_j^{best}, and the worst candidate solution, $\boldsymbol{x}^{\text{worst}}$, represented by x_j^{worst}. When generating a new solution, $\boldsymbol{x}_i^{\text{new}}$, the old solution at the ith position, $\boldsymbol{xi}$, is utilized, and the updated value of the decision variable j in the old solution is

denoted as xi, j^{new}. In addition, the terms r_1 and r_2 are introduced as two random numbers falling within the range of 0 and 1.

Step 6: **Update process**. In this step, the fitness value $f(x_j^{\text{new}})$ of the new solution (i.e., newly generated weights and biases) x_j^{new} is computed, and it replaces the current solution x_j if better, such as $f(x_j^{\text{new}}) < f(x_j)$.

Step 7: **Stop criterion**. Repeat Steps 3 to Step 6 until the highest number of steps *MaxIter* is achieved. After that, the MLP with the lowest MSE value is tested with the test dataset.

Step 8: **Update the external archive**. After each run, the solutions in the population, represented by vectors of weights and biases for each MLP, are sorted in the ascending order based on their MSE values. Subsequently, the archive (*ARCH*) is cleared, making way for the next iteration. During this process, the best A_r solutions are selected from the sorted population and copied into a new version of the archive (*ARCH*).

Despite the archive being emptied at the start of each training run, its K MLPs remain part of the new population of MLPs to be trained. After the training, the top A_r solutions from this new population are retained to form the latest version of the archive (*ARCH*). Consequently, the archive effectively serves as a store of accumulated knowledge, housing the best-performing solutions across multiple runs.

The detailed steps of the proposed EOLJaya-MLP are presented in Figure 2, and the pseudo-code is illustrated in Algorithm 1 . The steps of the proposed EOL-Archive-Jaya for MLP are presented in Figure 3 and the pseudo-code is illustrated in Algorithm 2.

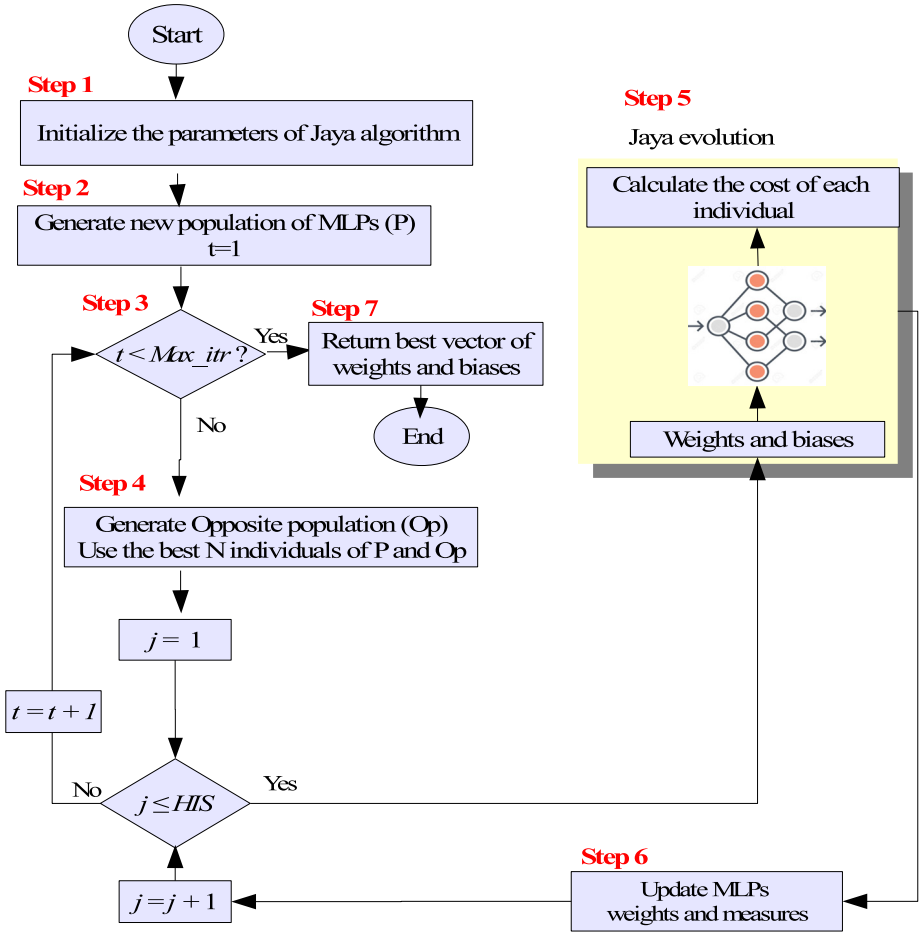

Figure 2: The steps of the proposed EOL-Jaya-MLP.

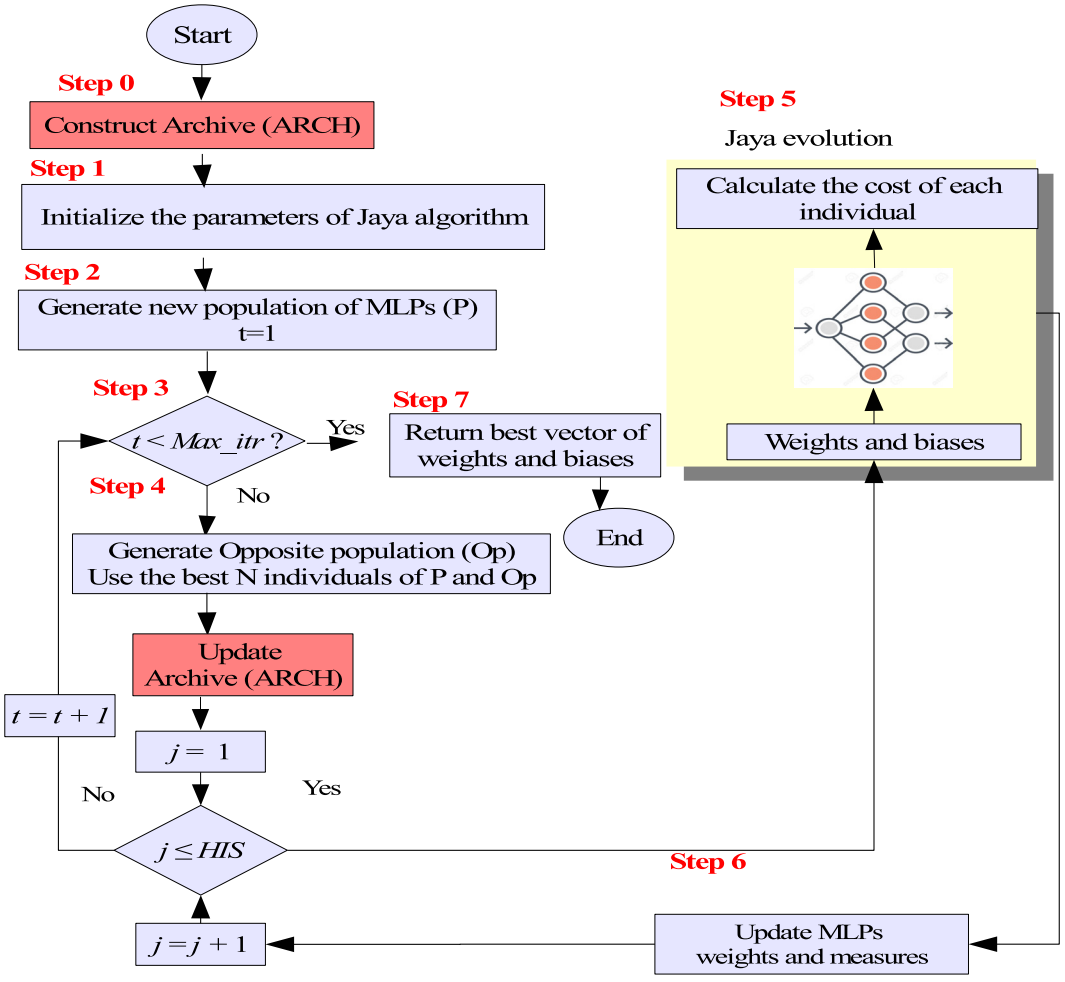

Figure 3: The steps of the proposed enhanced-Jaya-archive.

Algorithm 1. EOL Jaya for MLP pseudo-code

Initialize the parameters of Jaya (Max_{Itr}).
randomly initialize P solutions within the problem bounds (LB, UB)
Evaluate the initial P solutions
Set k = 0
while $k < Max_{Itr}$ **do**
  Select the best individual as the elite individual X_e from X_i
  Update the dynamic interval boundaries $[da_i, da_j]$ in X_i according to equation (14)
  Generate the opposite population
  Select the fittest P individuals from opposite populations and the population
  find the best X^{best} and worst X^{worst} solution in the current population
  for j = 1 to P **do**
    for i = 1 to $n + m$ **do**
      Calculate $X_{i,j}^{\text{new}}$ using equation (15)
      {Make sure that $X_{i,j}^{\text{new}}$ within the problem bounds}
      if ($X_{i,j}^{\text{new}} < LB_j$) **then**
        $X_{i,j}^{\text{new}} = LB_j$

```
            else
              X_{i,j}^{new} = UB_j
            end if {Replace the current solution in case of improvement}
          end for
          if f(X_i^{new}) < f(X_i) then
            X_i X_i^{new}
          end if
        end for
      k = K + 1
    end while{End of Generations}
    Return the best solution
```

Algorithm 2. EOL-Archive-Jaya for MLP pseudo-code

```
    Construct the archive ARCH matrix of size K × N, where K=P × A_r.
    Initialize the parameters of Jaya (Max_Itr).
    randomly initialize P solutions within the problem bounds (LB, UB)
    Evaluate the initial P solutions
    Set k = 0
    for run = 1 to #runs
      if run==1
        while k < Max_Itr do
          Select the best individual as the elite individual X_e from X_i
          Update the dynamic interval boundaries [da_i, da_j] in X_i according to equation (14)
          Generate the opposite population
          Select the fittest P individuals from opposite populations and the population
          find the best X^{best} and worst X^{worst} solution in the current population
          for j = 1 to P do
            for i = 1 to n + m do
              Calculate X_{i,j}^{new} using equation (15)
              {Make sure that X_{i,j}^{new} within the problem bounds}
              if (X_{i,j}^{new} < LB_j) then
                X_{i,j}^{new} = LB_j
              else
                X_{i,j}^{new} = UB_j
              end if {Replace the current solution in case of improvement}
            end for
            if f(X_i^{new}) < f(X_i) then
              X_i X_i^{new}
            end if
          end for
          k = k + 1
        end while {End of Generations}
      else
        copy the K solutions from ARCH.
        while k < Max_Itr do
```

Select the best individual as the elite individual X_e from X_i
Update the dynamic interval boundaries $[da_i, da_j]$ in X_i according to equation (14)
Generate the opposite population
Select the fittest P individuals from opposite populations and the population
find the best X^{best} and worst X^{worst} solution in the current population
for j = 1 to $P - K$
for i = 1 to $n + m$
Calculate $X_{i,j}^{\text{new}}$ using equation (15)
{Make sure that $X_{i,j}^{\text{new}}$ within the problem bounds}
if ($X_{i,j}^{\text{new}} < LB_j$)
$X_{i,j}^{\text{new}} = LB_j$
else
$X_{i,j}^{\text{new}} = UB_j$
end if{Replace the current solution in case of improvement}
end for
if $f(X_i^{\text{new}}) < f(X_i)$ **then**
$X_i X_i^{\text{new}}$
end if
end for
$k = k + 1$
end while {End of Generations}
end if
Update the Archive by feeding the *ARCH* by the best solutions from the current run.
end for {End of Runs}
Return the best solution

4.2 Air pollution data monitoring

The data used in this study are obtained the air quality monitoring stations of the Environment Public Authority in the State of Kuwait (K-EPA).

The State of Kuwait is situated in the northeastern region of the Arabian Pensinsula, bordered by Saudi Arabia to the South and West and Iraq to the north (Figure 4); Kuwait is one of the world's warmest locations since most of its land is superficially flat, and it is affected by a subtropical high-pressure system. It is marked by a huge temperature variation between summer and winter, with summer temperatures exceeding 50°C and winter temperatures falling below 7°C [86]. More than 4.5 million people live in Kuwait, and almost all of them are concentrated in Kuwait City, the center of the country's urban activity. Five governorates make up Kuwait: Ali Sabah Al Salem, Fahaheel, Jahra, Mansouria, and Rumaithya. The State of Kuwait has a coastline that looks out onto the northwestern portion of the Arabian Gulf, and Kuwait Bay, which is north of Kuwait City, serves as the nation's most significant shoreline due to its significant economic and historical significance [86] (Figure 5).

The K-EPA developed the air monitoring stations with the goal of monitoring Kuwait's air quality and preserving the present rates of various air quality factors within safe and acceptable ranges. Since its establishment in 1984, the network of stations designed specifically to track air quality has been supplying data continuously. The research uses an extensive dataset obtained from the air quality monitoring stations managed by the K-EPA. This dataset includes air quality data collected over a period of 3 years, namely, from January 1, 2018, to December 31, 2020. Data were obtained from 13 air monitoring stations strategically

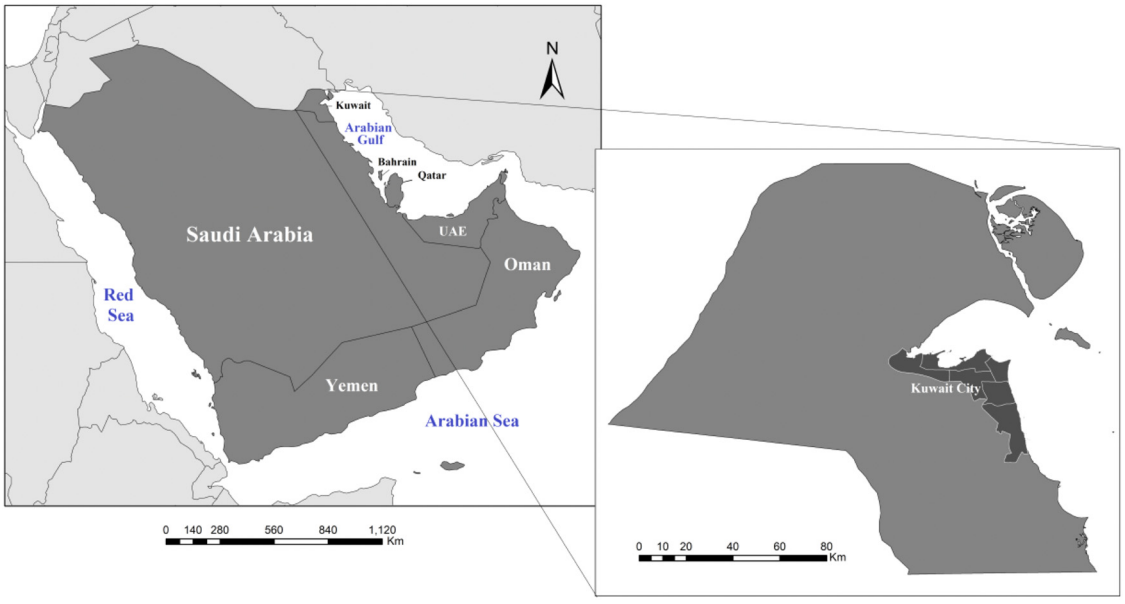

Figure 4: The Arabian Peninsula and the countries located in it, including Kuwait in the north-eastern part. Source: Kuwait Environment Public Authority (K-EPA) eMISK, 2020.

located around Kuwait, supplemented by data from four intermittent mobile stations to provide thorough coverage and consistent data collection. These stations observe various air pollutants, such as PM with a diameter of less than 10 mu (PM-10), CO, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ground-level ozone (O_3). These pollutants are crucial for evaluating air quality and its effects on health and the environment. Data collection occurs at regular intervals, with each fixed station autonomously transmitting data to the database of the Environment Public Authority. This process guarantees continuous monitoring in real time. The dataset contains 24 h averages for PM-10, SO_2, and NO_2, as well as 8 h averages for O_3 and CO. The selected time periods were specifically aligned with worldwide standards and methods for measuring air quality. This ensures that daily exposure levels may be accurately assessed. In terms of the dataset's size, our research examined data from four stations over a 3-year duration, resulting in a comprehensive and reliable dataset for assessing the efficacy of various air quality imputation methods. The dataset contains a substantial number of hourly measurements for each pollutant, spanning all stations and the whole research period. The actual amount of data points may vary depending on the station and pollutant, owing to the natural fluctuation in environmental monitoring. Moreover, the preprocessing stages were carefully executed to guarantee the quality and usefulness of the data. The following steps were undertaken:

- Data Cleaning: The process of eliminating outliers and rectifying any evident flaws in the dataset to ensure that the data precisely represent the recorded air quality conditions.
- Missing Value Imputation: Utilizing sophisticated imputation methods to estimate missing values, which is a crucial process owing to periodic data gaps caused by equipment maintenance or failure. We based our approach on the idea that if fewer than 75% of the data for a particular pollutant were available on a certain day (i.e., less than 18 h of data), we considered the daily average as missing. This criterion corresponds to the criteria established by the US Environmental Protection Agency.
- Conversion to AQI: To make it easier to understand air quality data and its possible effects on health, the concentrations of pollutants were transformed into AQI values using the equation and breakpoints specified in the given material. This conversion adheres to established protocols and enables the uniform documentation of air quality.

In what follows, we elaborate on the dataset's specifics, which include the size of the dataset, frequency of data collection, and preprocessing steps undertaken.

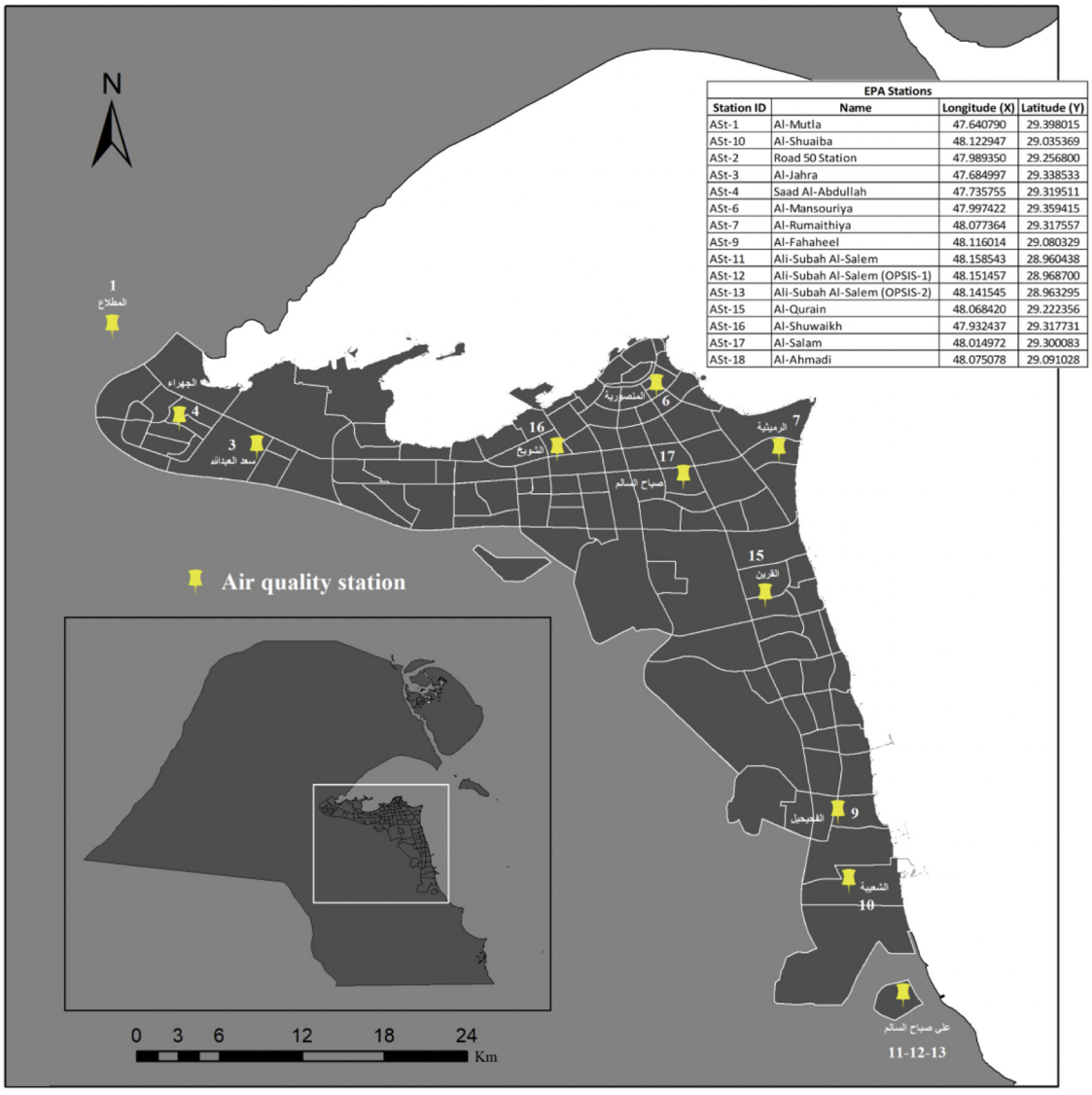

Figure 5: Locations of the air quality monitoring stations whose data was used in this study. Source: Kuwait Environment Public Authority (K-EPA) eMISK, 2020.

4.3 Datasets from Kuwait EPA

To compare and contrast the effectiveness of different imputation techniques for estimating missing values in the dataset, we used a real-time air quality monitoring dataset that was gathered from five locations in Kuwait from the Kuwait Environmental Public Authority (K-EPA). The dataset included information on the weather, time, and air quality.

- Air quality data: This set includes the following air pollutant variables: nitrogen dioxide (NO_2), carbon monoxide (CO), PM with a diameter of 10 μm or less (PM_{10}), sulfur dioxide (SO_2), and ozone (O_3).
- Meteorological data: This set consists of various meteorological parameters, such as temperature, humidity, wind direction, and WS.

Data on pollutants was gathered from Kuwait's Environmental Public Authority (K-EPA). Five environmental monitoring stations provided the data between January 1, 2018, and December 31, 2021. Each station's 24 h aggregate for PM_{10}, SO_2, and NO_2 as well as its 8 h aggregation for O_3 and *CO* were calculated using the

daily data. The micrograms per cubic meter ($\mu g/m^3$) units were used to measure all contaminants. The average is deemed missing if less than 75% of the data are provided (i.e., less than 6 h), according to US environmental protection agency [87]. We made use of the AQI generated by Al-Shayji et al. [88].

4.4 Converting data to AQI

It is challenging to compare the quality of the air in different regions and cities since there are so many different air quality indices in use around the world, each with its own concept and presentation [89,90]. The AQI, a metric used all over the world, was created to standardize the method of determining the level of air pollution using measured concentrations of particular ambient air pollutants. It educates the public and policymakers on the gravity of air pollution and the detrimental impact it can have on human health in an effort to protect the environment and human health. In addition, it is employed to assess pollution-reduction efforts and monitor changes in ambient air quality [91].

An enormous amount of data are provided on a time-specific basis by stations that monitor ambient air quality. These data are offered to interested parties as either AQI values or alternative indices with different timeframes, purposes, and other subindices based on epidemiological research [92–97]. These data concentrate on six major air pollutants: SO_2, NO_2, $PM_{2.5}$, carbon monoxide (CO), O_3, and PM_{10} [98,99]. Since the AQI values vary from 0 to 500, and their severity is linked to the amount of pollutants in the surrounding air, and a higher AQI number denotes potentially more serious health effects. For sensitive groups, the air quality is deemed unhealthy when the AQI is above 100 [100].

According to Johnson et al. [101], the AQI is described as a measurement of the air's quality in relation to any human need or to the needs of one or more biotic species [101,102]. The categories of the AQI are each assigned a number, and each slot is identified by a color. An index of the health risk associated with air quality is provided, ranging from a healthy level of zero to a very hazardous level of above 300.

The AQI is a standardized indicator and a tool for communication that gives a summary of the health hazards related to air pollution from gases and PM as well as ambient air quality [103]. Without being aware of the details of the underlying data, these indicators enable the stakeholders to monitor their local, national, and regional air quality. Offering information to the public and policymakers that enable stakeholders to take the required precautions to protect themselves from the negative impacts of air pollution is the main objective from the perspective of public health. A secondary objective is to increase public knowledge of the impacts of air pollution at current exposure levels to spur changes in both individual behavior and public policy [104,105].

WHO (2006) encouraged countries to develop strategies to thoroughly examine their own local circumstances, taking into account the distinctive qualities of each location's objective, particularly AQI [106]. For the purposes of our study, we used the AQI generated by Al-Shayji et al. [88], which was created for the state of Kuwait and founded on United States Environmental Protection Agency (USEPA's) criteria [107]. A measurement of air quality used in our study is called the AQI. It gives details about how pure the air is. Then, we separated the values into ranges and gave each range a descriptor and a color code (yellow for moderate, red for unhealthy, green for good, maroon for hazardous, and purple for very unhealthy). Public health recommendations are provided for each AQI range. Each nation and pollutant has a different AQI. The concentration of each pollutant is converted into AQI using equation (16):

$$I_p = \frac{I_{\text{high}} - I_{\text{low}}}{C_{\text{high}} - C_{\text{low}}}(C_p - C_{\text{low}}) + I_{\text{low}}, \tag{16}$$

where I_p denotes the index (AQI) for pollutant p (i.e., NO_2, SO_2, … , etc.), C_p refers to the curtailed concentration of pollutant p, while C_{low} is the concentration breakpoint that is equal to or the less than C_p, i.e., $\leq C_p$, C_{high} is the concentration breakpoint that is equal to or greater than C_p, i.e., $\geq C_p$, I_{low} refers to the index breakpoint, that is, C_{low} (i.e., the AQI value matching C_{low}), while I_{high} is the index breakpoint that is C_{high} (i.e., the AQI value matching C_{high}) [107].

Table 2: Kuwait AQI values

Categories	AQI, sub-index I_{low}–I_{high}	O_3 (ppm), 8 h C_{low}–C_{high}	PM_{10} (μg/m^3), 24 h C_{low}–C_{high}	CO (ppm), 24 h C_{low}–C_{high}	SO_2 (ppm), 24 h C_{low}–C_{high}	NO_2 (ppm), 24 h C_{low}–C_{high}
Good	0–50	0.0–0.03	0.0–90	0.0–4.0	0.0–0.03	0.0–0.03
Moderate	51–100	0.031–0.06	90.1–350.0	4.1–8.0	0.031–0.06	0.04–0.05
Unhealthy (1)	101–150	0.061–0.092	350.1–431.1	8.1–11.7	0.061–0.182	0.06–0.30
Unhealthy (2)	151–200	0.093–0.124	431.4–512.5	11.8–15.4	0.183–0.304	0.31–0.55
Very unhealthy	201–300	0.125–0.374	512.6–675.0	15.5–30.4	0.305–0.604	0.56–1.04
Hazardous	301–500	0.375–0.504	675.1–1000	30.5–50.4	0.605–1.004	1.05–2.04

In this study, the AQI created by Al-Shayji et al. [88] for the State of Kuwait was used to evaluate the air quality. It was based on recommendations made by the USEPA [107]. The AQI is an index used to report on the current state of the air. It provides information on the air quality outside. To translate from pollutant concentration to AQI, Table 2 is used.

Here, we can demonstrate how to calculate the AQI using Table 2 and equation (16). The third column in Table 2 displays the 24 h PM_{10} range (low breakpoint (C_{low}) to high breakpoint (C_{high})). For "Good" air quality, this equates to an AQI score between 0 and 50. As a result, if the 24 h integrated PM_{10} concentration was 6.0 μg/m^3 (C_p), C_{high} would be 90.0 μg/m^3, C_{low} would be 0 μg/m^3, circumstances I_{high} would be 50, and I_{low} would be 0. The AQI range (I_{low}–I_{high}) corresponds to the PM_{10} range of 0 to 90 μg/m^3. Therefore, the AQI would be determined as follows for a daily average PM_{10} average concentration of 6.0 μg/m^3:

$$\text{AQI} = \frac{(50.0 - 0.0)}{(90.0 - 0.0)}(6.0 - 0.0) + 0.0 = 3.33.$$

4.5 Dealing with missing values

Monitoring the condition of the air quality in Kuwait is the responsibility of the environmental public authority (K-EPA). Most of the time, the air quality data from the five stations utilized in this study had missing data, which might lead to bias due to systematic errors between the unobserved and observed values [108]. To ensure that the data are of high quality, it is crucial to identify the best method for estimating the missing values. Results from incomplete data matrices may differ dramatically from those anticipated from a complete dataset [109]. Any data analysis' main goal should be to draw reliable conclusions about the population being studied. We conducted an in-depth comparison study of the various imputation techniques using the data obtained by K-EPA. Each dataset has missing data added to it under the assumptions of a general pattern for missing data and the three processes of missing data: MAR, MCAR, and MNAR. The MCAR assumption stated that each dataset's missing values were assigned at random. Information missing was more likely under the MAR hypothesis depending on the class property. The highest or lowest values of X_s were eliminated based on the MNAR hypothesis. The study's goal was to compare six different imputation techniques for MAR, MNAR, and MCAR in the case of missing data. We experimented with changing the proportions by 20, 5, 30, 10, and 40% to imitate the rates of missing data. We created multivariate missing data using a MAR, MNAR, or MCAR missing data mechanism by using the "*ampute*" function in the "*MICE*" package in R.

4.6 Multiple imputation using random forest method

Let us presume that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ is a $n \times p$-dimensional data matrix. The random forest technique is what we suggest using to impute missing observations. The missing value is handled by a built-in process in

the random forest algorithm. By balancing the frequency of values with the proximity of the bagging adjustment, this is accomplished. Consequently, after training on an initial set of mean data that are imputed, this creates a vast collection of de-correlated trees and averages them [110]. This method necessitates a comprehensive and beneficial response variable for forest training. Instead, we utilize a random forest model that has been trained on the dataset, where X is the matrix of the entire dataset, to directly predict the values of all the missing variables. $\mathbf{X}_s$ comprises all missing values at entries $\mathbf{i}_{\text{mis}}^{(s)} \subseteq \{1, \ldots, n\}$. Four sections of the dataset can be distinguished:

- $\mathbf{y}_{\text{obs}}^{(s)}$: the observed values of $\mathbf{X}_s$.
- $\mathbf{y}_{\text{mis}}^{(s)}$: the missing values of $\mathbf{X}_s$.
- $\mathbf{x}_{\text{obs}}^{(s)}$: the observations, $\mathbf{i}_{\text{obs}}^{(s)} = \{1, \ldots, n\} \backslash \mathbf{i}_{\text{mis}}^{(s)}$, that belong in the other variables $\mathbf{X}_s$.
- $\mathbf{x}_{\text{mis}}^{(s)}$: the observations, $\mathbf{i}_{\text{mis}}^{(s)}$, that belong in the other variables $\mathbf{X}_s$.

Note that $\mathbf{x}_{\text{obs}}^{(s)}$ and $\mathbf{x}_{\text{mis}}^{(s)}$ are not completely observed, as the index $\mathbf{i}_{\text{obs}}^{(s)}$ corresponds to the observed values of the variable $\mathbf{X}_s$.

The imputation process commences by making an initial estimate for the missing values in $\mathbf{X}$ using mean imputation or other suitable imputation methods, as suggested by Stekhoven et al. (2012). Following this, a decision is made on whether to order the predictors $\mathbf{X}s$, $s = 1, \ldots, p$, in an ascending or a descending order based on the quantity of missing data. The missing values for each variable $\mathbf{X}s$ are then imputed using a random forest (first fitting) with predictors $\mathbf{X}_{\text{obs}}^{(s)}$ and response $\mathbf{y}_{\text{obs}}^{(s)}$. The trained random forest is utilized to estimate the missing values $\mathbf{y}_{\text{mis}}^{(s)}$ from $\mathbf{x}_{\text{mis}}^{(s)}$. This imputation process is repeated until a specified stopping requirement is met. For a more detailed depiction of the missForest algorithm, refer to pseudo Algorithm 3.

Algorithm 3. Impute missing values with random forest, Stekhoven and Bühlmann [111]

Require: $\mathbf{X}$ is an $n \times p$ matrix. Set up the stopping criterion (γ)

set up initial guess for missing values;
$\mathbf{k}$ is the vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing the amount of missing values;
while not γ **do**
  $\mathbf{X}_{\text{old}}^{\text{imp}}$ stores the previously imputed matrix;
  for s in $\mathbf{k}$ **do**
    Fit a random forest: $\mathbf{y}_{\text{obs}}^{(s)} \sim \mathbf{x}_{\text{obs}}^{(s)}$;
    Predict $\mathbf{y}_{\text{mis}}^{(s)}$ using $\mathbf{x}_{\text{mis}}^{(s)}$;
    $\mathbf{X}_{\text{new}}^{\text{imp}}$ updates the imputed matrix using predicted $\mathbf{y}_{\text{mis}}^{(s)}$;
  end for
  update γ
end while
return the imputed matrix $\mathbf{X}^{\text{imp}}$

The last imputed data matrix differs from the prior one for the first time, with regard to both variable types, and the stopping requirement (γ) is satisfied at that point. According to this definition, the difference for the set of continuous variables $\mathbf{N}$ is given as follows:

$$\Delta_N = \frac{\sum_{j \in \mathbf{N}} (\mathbf{X}_{\text{new}}^{\text{imp}} - \mathbf{X}_{\text{old}}^{\text{imp}})^2}{\sum_{j \in \mathbf{N}} (\mathbf{X}_{\text{new}}^{\text{imp}})^2}, \tag{17}$$

and that for the set of categorical variables $\mathbf{F}$ is given as follows:

$$\Delta_F = \frac{\sum_{j \in \mathbf{F}} \sum_{i=1}^{n} \mathbf{I}_{\mathbf{X}_{\text{new}}^{\text{imp}} \neq \mathbf{X}_{\text{old}}^{\text{imp}}}}{\#NA}. \tag{18}$$

The imputation process begins with an $n \times p$ matrix $\mathbf{X}$, and a specified stopping criterion γ is set. The initial guess for missing values is determined, and a vector $\mathbf{k}$ is created, containing the sorted indices of columns in $\mathbf{X}$ based on increasing amounts of missing values. The previously imputed matrix is stored in $\mathbf{X}\text{old}^{\text{imp}}$. A random forest model is fitted using the observed values, where $\mathbf{y}\text{obs}^{(s)}$ represents the response variable, and $\mathbf{x}\text{obs}^{(s)}$ denotes the predictor variables. Utilizing this model, missing values $\mathbf{y}\text{mis}^{(s)}$ are predicted from the corresponding missing predictor values $\mathbf{x}\text{mis}^{(s)}$. The imputed matrix $\mathbf{X}\text{new}^{\text{imp}}$ is then updated with the predicted missing values. This process continues iteratively while adjusting the stopping criterion γ and updating the imputed matrix $\mathbf{X}^{\text{imp}}$ until the specified stopping requirement is met.

Here #NA is the number of missing values in the categorical variables $\mathbf{F}$.

After imputing the missing values, the performance is assessed using the normalized root MSE [112] for the continuous variables, defined by:

$$\text{NRMSE} = \sqrt{\frac{\text{mean}((\mathbf{X}^{\text{true}} - \mathbf{X}^{\text{imp}})^2)}{\text{var}(\mathbf{X}^{\text{true}})}}, \tag{19}$$

where the whole data matrix and the imputed data matrix are denoted by $\mathbf{X}^{\text{true}}$ and $\boldsymbol{X}^{\text{imp}}$, respectively. All predictors in this study are categorized as continuous observations. The empirical mean and variance computed over the missing values solely are denoted by the terms “mean” and “variance” in abbreviated form, respectively.

We employ the out-of-bag (OOB) estimate of a variable’s error when a RF_m is fitted to the portion that is seen on the variable. We average over the collection of variables of that type when we reach the halting requirement (γ), which gives us an approximation of the real imputation mistakes. By evaluating the absolute difference between the OOB imputation error estimate in each simulation run and the actual imputation error, we can determine how well this estimate performs. Almost any type of data can be imputed using the missForest approach. It can specifically handle multivariate data made up of concurrent continuous and categorical elements.

Both parameter adjustment and making assumptions about how the information would be distributed are not required by this method. Finally, missForest showed the smallest difference in prediction error when models employed imputed data, and it had the lowest imputation error for each incidence of missingness rates (10, 5, 20, 30, and 40%).

5 Experiments and results

In this section, we discuss the experimental results that have been carried out to evaluate the effectiveness of the proposed EOL-Jaya-MLP and EOL-Archive-Jaya algorithms in predicting the AQI. As mentioned earlier, the data used in this research come from Kuwait’s environmental public authority (K-EPA). It was collected from five monitoring stations across the country, namely, Ali-Sabah, Fahaheel, Jahra, Mansouria, and Rumaithiya for a three-year period spanning from January 1, 2018, to December 31, 2021. They include recordings of PM_{10}, SO_2, NO_2, O_3, and CO for each of the mentioned monitoring stations. The datasets have gone under a thorough inspection to handle the invalid and missing data as illustrated in Section 4.5.

In our data analysis, we adopted a prudent approach to splitting all datasets, allocating 10% for testing and 90% for training. To ensure a representative and balanced distribution of classes, we employed a technique called **stratified sampling**. This method involves computing the ratio of each class within the dataset and then achieving the train/test split percentage while considering these ratios. The primary advantages of using this strategy are twofold: first, it helps maintain the proportion of each class in the divided data, and second, it enhances the representation of minority classes.

As a result of employing this stratified sampling approach, both the training and testing subsets contain a balanced number of classes, ensuring a more robust and accurate model performance. By avoiding bias toward overrepresented classes and providing ample representation to the underrepresented ones, we achieve more reliable results in our analysis and predictions.

Note that in the realm of air pollution prediction using MLP models, the conventional emphasis on time complexity analysis may not be warranted for several compelling reasons. As air pollution data typically involve periodic updates rather than continuous streams which diminishing the significance of time complexity considerations. Also, the frequency of model updates and retraining is relatively low in air pollution prediction applications. The underlying factors influencing air quality, such as weather patterns and emissions from stationary sources, evolve gradually rather than abruptly. Furthermore, the primary focus in air pollution prediction often lies in the accuracy and interpretability of the model rather than its computational efficiency. Finally, the computational resources available for air pollution prediction tasks are often sufficient to accommodate moderately complex algorithms without significant performance bottlenecks.

5.1 Experimental settings

The proposed algorithm is evaluated against three swarm intelligence algorithms using identical datasets. All experiments are conducted using Python version 3.9 on a PC with the Windows operating system, equipped with an Intel R Xeon Silver 1.8 GHz CPU and 8 GB of RAM.

To ensure robustness and enhance reliability, the algorithms are executed independently for each dataset over 30 runs, maintaining a fixed number of 100 iterations for consistency. Throughout these runs, the population size of MLPs remains constant at 25 for all the comparative algorithms. This approach allows for thorough evaluation and consistent performance assessment across datasets, minimizing potential bias and variability in the results.

For consistency and fairness, we use the parameter settings recommended by the original researchers of each comparative method. Detailed information about these parameter settings can be found in Table 3.

By conducting these experiments under controlled conditions with standardized settings, we aim to provide a fair and comprehensive comparison of the proposed algorithm against the three swarm intelligence algorithms, leading to insightful and trustworthy results.

5.2 Descriptive results

Figures 6 and 7 present the simulation results obtained from running the developed ML algorithms (i.e., EOL-Jaya-MLP and EOL-Archive-Jaya) on the mentioned datasets. The results are then compared to the results

Table 3: Parameters settings of the comparative algorithms

Algorithm	Parameter	Settings
PSO [113]	Population size	30
	Acceleration constants (c_1 and c_2)	[2, 2]
	Inertia wesights (w)	[0.2, 0.9]
BA [114]	Population size	30
	Loudness (A_0)	0.5
	Pulse emission rate (r)	[0,1]
	Minimum frequency ($f_{\min}$)	0
	Maximum frequency ($f_{\max}$)	2
MFO [115]	b	1

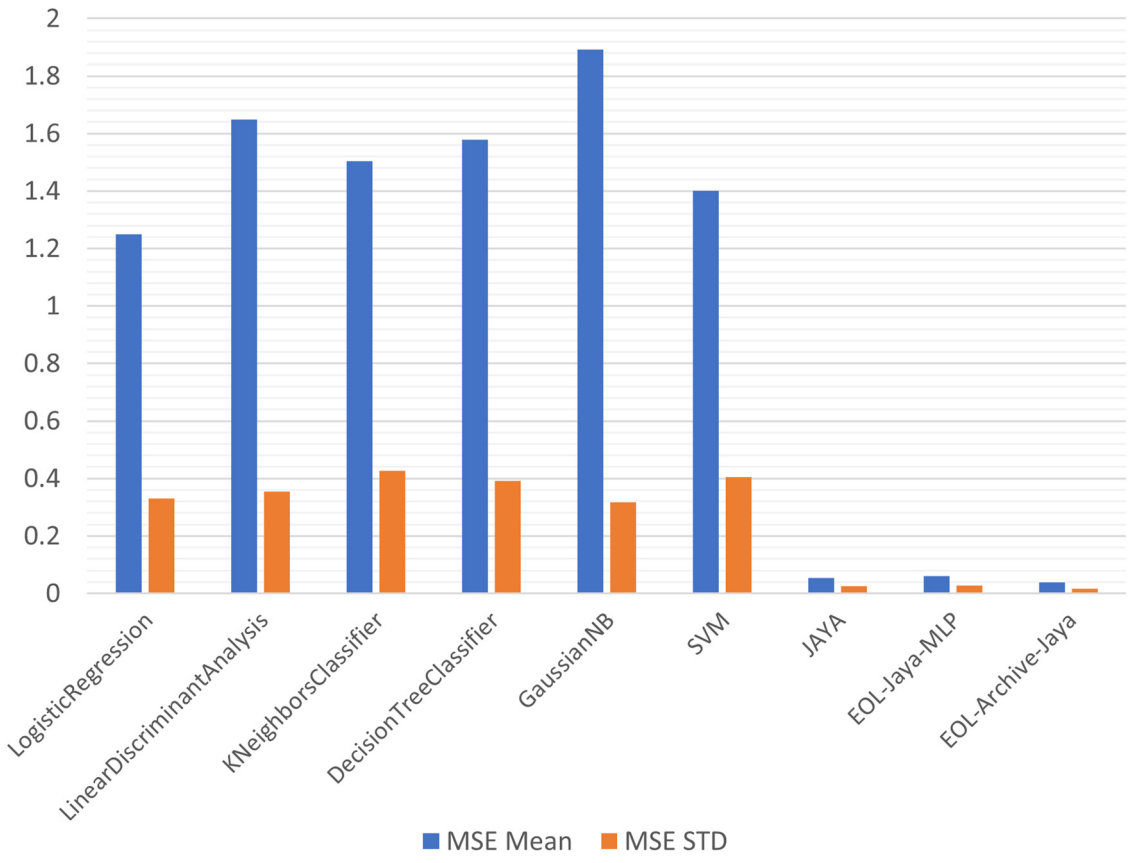

Figure 6: MSE mean and MSE STD of Jaya-based algorithms vs other classical ML algorithms for Rumaithiya station.

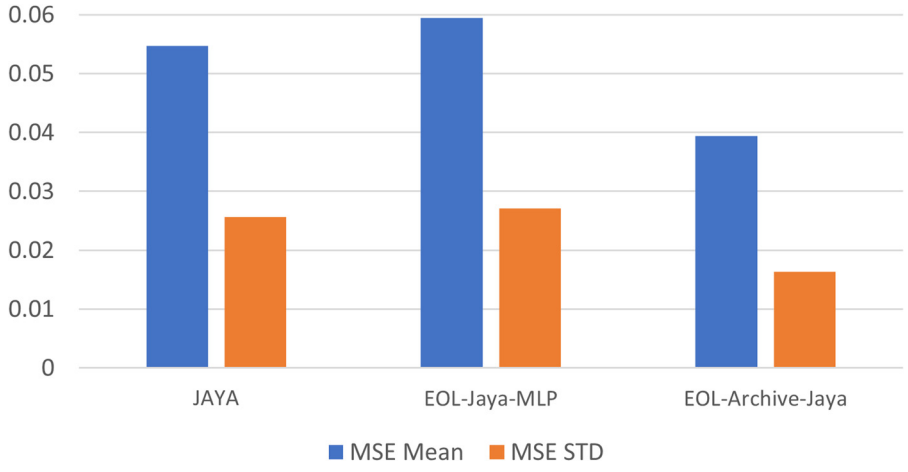

Figure 7: MSE mean and MSE STD of JAYA, EOL-Jaya-MLP and EOL-Archive-Jaya for Rumaithiya station.

obtained from some classical ML algorithms, namely, logistic regression, linear discriminant analysis, K-neighbors classifier, decision tree classifier, Gaussian NB, and SVM.

Figure 6 shows the MSE mean and MSE STD results of EOL-Archive-Jaya against other classical ML algorithms. It is obvious that Jaya-based algorithms (i.e., original Jaya, EOL-Jaya-MLP, and EOL-Archive-Jaya) outperform other ML techniques in general as they have the lowest MSE and lowest MSE STD.

To better analyze the results, we zoom in on the section depicting the different Jaya-based algorithms as shown in Figure 7. According to Figure 7, EOL-Archive-Jaya has the lowest MSE Mean and MSE STD among

others. This indicates that EOL-Archive-Jaya model, which is an EOL-Jaya-MLP enhanced by an external archive strategy, fits the data better than other models and has more precise predictions.

5.3 Performance measures

Since the different Jaya classes (i.e., original Jaya, EOL-Jaya-MLP, and EOL-Archive-Jaya) have proven to perform much better than other classical ML techniques in predicting the AQI as shown in Figure 6, in this section, we focus only on the three variant models of Jaya to further investigate their performance in forecasting the AQI. The prediction error statistics of the Jaya, EOL-Jaya-MLP, and EOL-Archive-Jaya are shown in Tables 4–7.

Table 4: Performance of EOL-Archive-Jaya compared to JAYA and EOL-Jaya-MLP based on MSE for all stations

MSE	JAYA	EOL-Jaya-MLP	EOL-Archive-Jaya
Ali-sabah	0.098549169	0.083169656	**0.074170456**
Fahaheel	0.119378516	0.105735534	**0.095300592**
Jahra	0.31658239	0.32439049	**0.259325914**
Mansouria	0.102748294	0.090753734	**0.032396835**
Rumaithiya	0.054699359	0.059473693	**0.039370602**

Values with lowest prediction error are in bold font.

Table 5: Performance of EOL-Archive-Jaya compared to JAYA and EOL-Jaya-MLP based on RMSE for all stations

RMSE	JAYA	EOL-Jaya-MLP	EOL-Archive-Jaya
Ali-sabah	0.308434005	0.282699631	**0.271488139**
Fahaheel	0.334131379	0.304285358	**0.308667509**
Jahra	0.560528787	0.566124564	**0.509240257**
Mansouria	0.313138389	0.292901687	**0.178576764**
Rumaithiya	0.22775206	0.237311398	**0.195375683**

Values with lowest prediction error are in bold font.

Table 6: Performance of EOL-Archive-Jaya compared to JAYA and EOL-Jaya-MLP based on RAE for all stations

MAE	JAYA	EOL-Jaya-MLP	EOL-Archive-Jaya
Ali-sabah	0.253542781	0.243039667	**0.218159042**
Fahaheel	0.288523129	**0.265927449**	0.26921946
Jahra	0.496854019	0.516403081	**0.495837055**
Mansouria	0.264683368	0.249060106	**0.138269146**
Rumaithiya	0.171692885	0.184846947	**0.161017751**

Values with lowest prediction error are in bold font.

Table 7: Performance of EOL-Archive-Jaya compared to JAYA and EOL-Jaya-MLP based on AARD for all stations

AARD	JAYA	EOL-Jaya-MLP	EOL-Archive-Jaya
Ali-sabah	50.37225028	47.72897369	**41.37735823**
Fahaheel	28.85231291	**26.59274492**	26.92194597
Jahra	55.13710813	60.87063964	**53.53279709**
Mansouria	43.0193467	40.06090119	**20.66533462**
Rumaithiya	54.54770488	49.55988076	**33.71705765**

Values with lowest prediction error are in bold font.

Specifically, Table 4 presents the MSE, which is calculated by equation (9) provided in Section 4. As shown in the table, MSE values of EOL-Archive-Jaya are always lower than those of other Jaya-based models for all stations which confirm that EOL-Archive-Jaya has an excellent learning ability and outperforms Jaya and EOL-Jaya-MLP.

Table 5 presents the results of RMSE of the three Jaya-based models. The RMSE is computed by equation (20).

$$\text{RMSE} = \sqrt{\left(\frac{1}{k}\right)\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}, \tag{20}$$

where y is the actual value, $\hat{y}$ is the predicted value, and k is the number of training samples.

As shown in Table 5, the results of RMSE confirm that EOL-Archive-Jaya always has a small deviation from the actual values for all stations, which indicate its excellent capabilities in anticipating the AQI.

In Table 6, we resent the mean absolute error (MAE) results of the three Jaya models. The MAE is defined by equation (21).

$$\text{MAE} = \sum_{i=1}^{k}|y_i - \hat{y}_i|, \tag{21}$$

where y is the actual value, $\hat{y}$ is the predicted value, and k is the number of training samples.

The results show that the MAE of EOL-Archive-Jaya is closer to "0" compared to other models for four of five stations, which again demonstrates the ability of EOL-Archive-Jaya to predict a more accurate value for AQI.

Table 7 presents the results of absolute average relative deviation (AARD) of the three Jaya models. The AARD is defined as follows:

$$\text{MAE} = \frac{1}{k}\sum_{i=1}^{k}\frac{|y_i - \hat{y}_i|}{y_i}, \tag{22}$$

where y is the actual value, $\hat{y}$ is the predicted value, and k is the number of training samples.

It is worth mentioning that the calculated AARD values indicate that the EOL-Archive-Jaya model is superior to Jaya and EOL-Jaya-MLP in forecasting AQI for four of five stations.

We now compare the performance of EOL-Archive-Jaya to other optimization algorithms based on MAE, MSE, RMSE, AARD, and Rumaithya station. The results show that EOL-Archive-Jaya performs better than other popular optimizers in predicting AQI for the mentioned station (Table 8).

In a nutshell, all the results presented in this section confirm the superiority of the EOL-Archive-Jaya model over other ML models and optimizers in predicting the AQI for the five stations under consideration.

5.4 Statistical validation

In this section, we validate the simulation results obtained by the Jaya, EOL-Jaya-MLP, and EOL-Archive-Jaya models using a statistical significance test, namely, one-way ANOVA. Please note that the significant test was performed under a P value <0.05.

Table 8: Comparison of the performance of EOL-Archive-Jaya compared to other optimization algorithms based on MAE, MSE, RMSE, AARD, and Rumaithya station

	BAT	MFO	PSO	JAYA	EOL-Jaya-MLP	EOL-Archive-Jaya
MAE	0.178287455	0.15761731	0.184585199	0.171692885	0.184846947	0.161017751
MSE	0.053456457	0.049604881	0.069248227	0.054699359	0.059473693	0.039370602
RMSE	0.222321186	0.213126235	0.256048884	0.22775206	0.237311398	0.195375683
AARD	51.23717721	57.15846342	70.42606898	54.54770488	49.55988076	33.71705765

Table 9: One-way-ANOVA Significant test results

Metric	Method	*N*	Mean	SD	SE	*P*
MSE	Jaya	30	0.099	0.039	0.007	
	EOL-Jaya-MLP	30	0.083	0.034	0.006	0.006
	EOL-Archive-Jaya	30	0.074	0.012	0.002	
RMSE	Jaya	30	0.308	0.059	0.011	
	EOL-Jaya-MLP	30	0.283	0.058	0.011	0.009
	EOL-Archive-Jaya	30	0.271	0.022	0.004	
MAE	Jaya	30	0.254	0.054	0.010	
	EOL-Jaya-MLP	30	0.243	0.048	0.009	0.001
	EOL-Archive-Jaya	30	0.218	0.020	0.004	
AARD	Jaya	30	50.372	15.836	2.891	
	EOL-Jaya-MLP	30	47.729	13.739	2.508	0.047
	EOL-Archive-Jaya	30	41.377	13.022	2.378	

The results in Table 9 confirms that there is a significant difference between the three Jaya methods for each used metric (i.e., MSE, RMSE, MAE, and AARD). The MSE for EOL-Archive-Jaya was significant among the other models (mean = 0.074, SD = 0.012, P value = 0.006). This indicates that EOL-Archive-Jaya performs better than the other models. Similarly, for other indicators (RMSE, MAE, and AARD), they all show that the EOL-Archive-Jaya model performs better than the two other models.

We also plot the results of the four indicators (i.e., MSE, RMSE, MAE, and AARD) for all stations in Figure 8. It is obvious from Figure 8 that the performance of EOL-Archive-Jaya is superior to that of the other two models, which, again, supports the results obtained in Sections 5.2 and 5.3.

A comparison between JAYA, E-JAYA, and E-JAYA-Archive with respect to MSE, RMSE, MAE, and AARD based on Table 9 is summarized as follows:

MSE and RMSE comparisons: E-JAYA and E-JAYA-Archive show no significant difference ($p > 0.05$), suggesting similar performance. However, both significantly outperform JAYA ($p < 0.001$), indicating lower mean and root MSEs.

MAE comparisons: Similar to MSE and RMSE, E-JAYA and E-JAYA-Archive are not significantly different, but both perform significantly better than JAYA, indicating lower mean absolute errors.

AARD comparisons: E-JAYA and E-JAYA-Archive show a borderline nonsignificant difference ($p = 0.057$), but both significantly differ from JAYA, with E-JAYA-Archive showing the most substantial difference ($p < 0.001$), indicating a lower average absolute relative difference.

The afore-mentioned comparisons indicate that the E-JAYA and E-JAYA-Archive methods are generally more effective than JAYA in predicting air quality, based on the metrics tested. The lack of significant differences between E-JAYA and E-JAYA-Archive in most comparisons suggests that both methods are comparable in performance, with the notable exception of AARD, where E-JAYA-Archive might offer a slight advantage.

Moreover, Figure 9 displays the distribution of the average absolute relative difference (AARD) for three methods: E-JAYA, E-JAYA-Archive, and JAYA. The form of each violin corresponds to the distribution of data points at various AARD levels, with broader sections indicating a higher density of data. The median value is shown by the centre dot in each violin, while the interquartile range (IQR) is represented by the thick line, which shows the range of the middle 50% of the values. In addition, individual data points are graphed, offering a deeper understanding of the distribution beyond the aggregate numbers. The mean AARD values indicate that E-JAYA-Archive has a more focused distribution of lower AARD values and the lowest mean AARD (approximately 35.24), suggesting superior performance in this metric compared to E-JAYA (mean approximately 40.99) and JAYA (mean approximately 46.39). JAYA exhibits the widest range of values and the highest mean AARD, implying less consistency and inferior performance. The sample sizes for each approach are as follows: E-JAYA and JAYA each have 150 data points, while E-JAYA-Archive has 120 data points. The statistical significance of the observed differences between the techniques was assessed by analysing descriptive statistics over numerous stations and contaminants, providing a thorough comparison. The EOL-Jaya-MLP

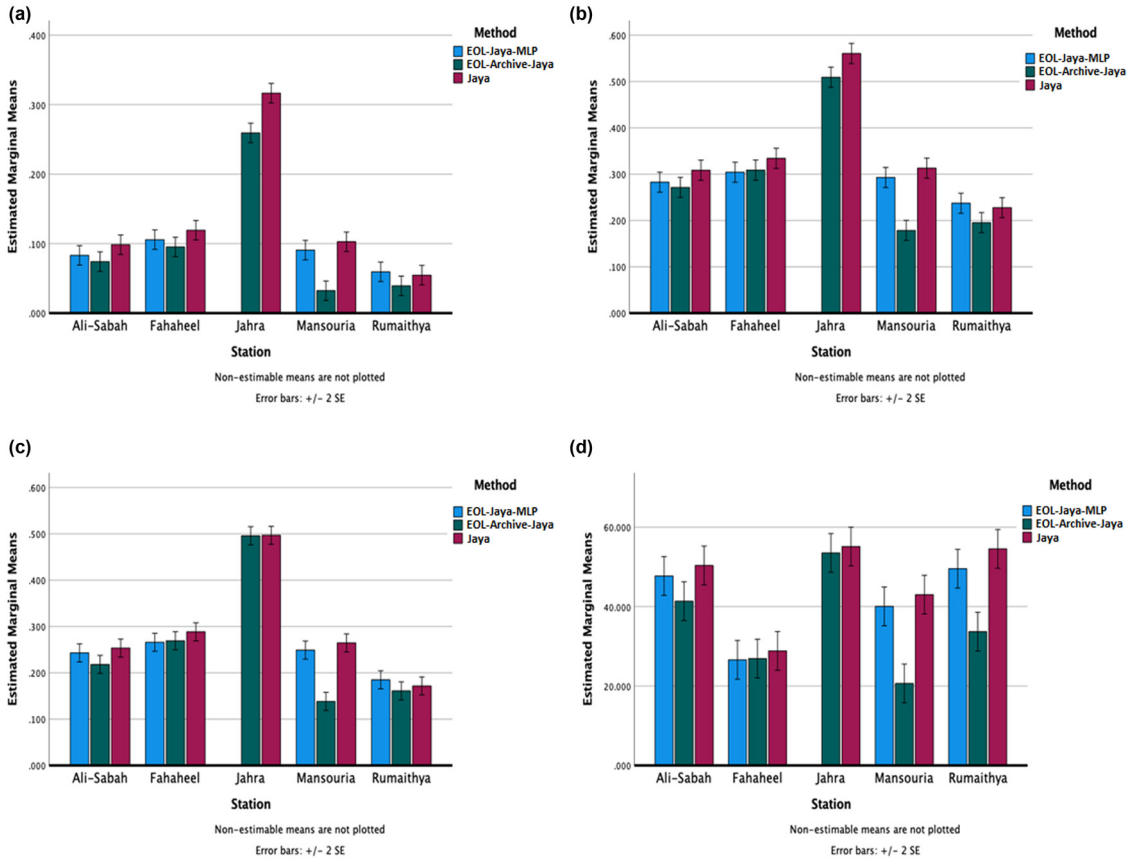

Figure 8: MSE, RMSE, MAE, and AARD statistical results for all stations. (a) Estimated marginal means of MSE, (b) estimated marginal means of RMSE, (c) estimated marginal means of MAE, and (d) estimated marginal means of AARD.

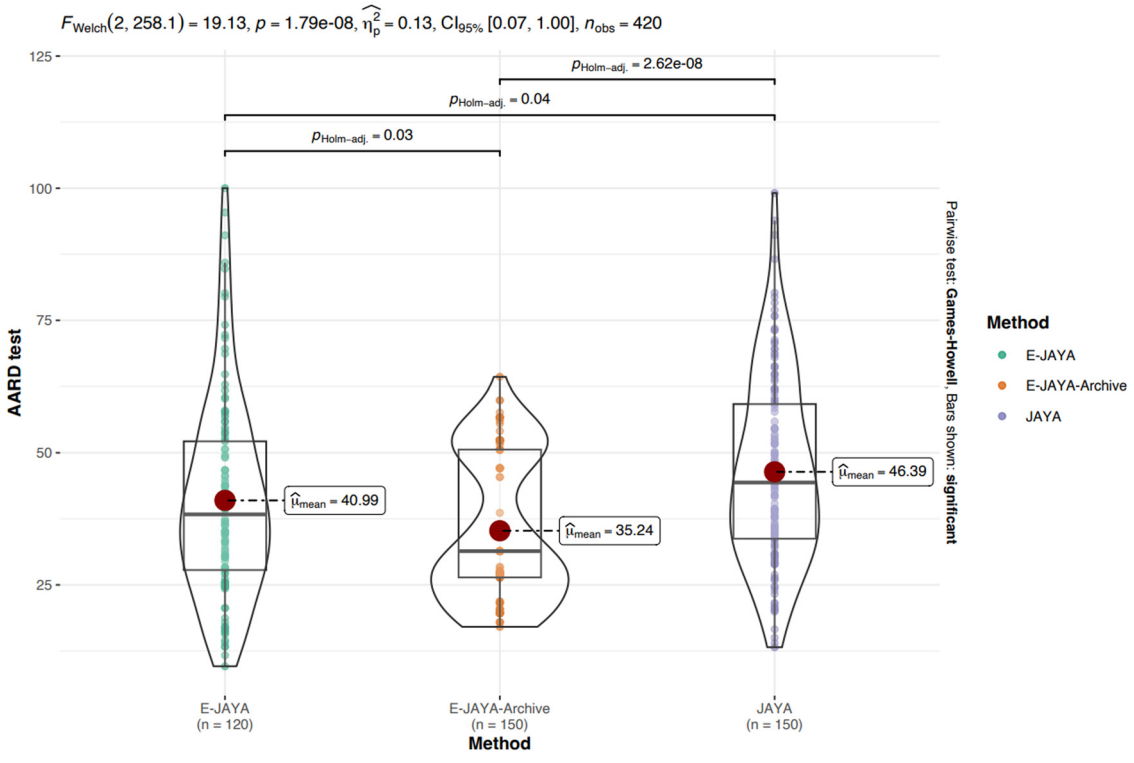

Figure 9: Distribution of the average absolute relative difference (AARD) for three methods: E-JAYA, E-JAYA-Archive, and JAYA.

technique demonstrates stability and dependability in forecasting air quality, as shown by consistently decreased error metrics (MSE, RMSE, MAE, and AARD) across multiple stations. The results emphasise the significance of choosing suitable optimisation algorithms to enhance air quality prediction models, a critical aspect in the development of successful methods for environmental monitoring and control.

6 Conclusion and future work

In this study, we explored the power of the Jaya algorithm, which mimics the survival of the fittest strategy, for optimization problems. We chose this algorithm due to its unique ability to strike a balance between exploring diverse search space niches and exploiting each niche effectively.

To enhance the performance of the Jaya algorithm and maintain local optima while preserving population diversity, we introduced two improvements: elite opposition-based learning (EOL-Jaya-MLP) and an archive of best solutions (EOL-Archive-Jaya). These new algorithms were specifically designed for the selection and training of MLPs.

Our focus was on predicting urban air quality, a critical aspect of mitigating air pollution and improving the well-being of city residents. We utilized a comprehensive 3-year dataset collected from five air quality monitoring stations and created multivariate missing data using various mechanisms (MAR, MNAR, and MCAR).

To ensure fair experimentation, all datasets were normalized and split into 10% for testing and 90% for training, with a stratified approach maintaining balanced class representation in the split. The varying number of features and class labels in each dataset required adjusting the number of inputs, hidden, and output nodes for each MLP.

We compared the results of our proposed techniques (EOL-Jaya-MLP and EOL-Archive-Jaya) against the original Jaya algorithm and three other swarm optimization algorithms (PSO, BA, and MFO). In addition, we compared our techniques against six ML techniques. Notably, EOL-Archive-Jaya demonstrated exceptional accuracy, surpassing other comparative methods. Furthermore, the proposed algorithm exhibited superior convergence, outperforming comparable methods in terms of speed and avoiding local optima thanks to its diverse techniques.

Moving forward, we plan to apply the proposed algorithm to real-world applications, expanding its potential and usefulness. Furthermore, we envision hybridizing the algorithm with other local search algorithms to improve its exploitation capabilities further. With its promising results and versatile applications, the EOL-Archive-Jaya algorithm shows great promise in tackling complex optimization and ML challenges.

Acknowledgment: We would like to thank Kuwait Environment Public Authority for providing the air pollution data.

Funding information: This publication was made possible by the support of the AUK Open Access Publishing Fund.

Author contributions: Abu Doush, Sultan, and Alsaber develop the conceptualization, Methodology, and Software. Sultan and Abdullah did the validation and experiements. Alkandari and Alsaber did formal analysis and visualization. All authors reviewed the manuscript.

Conflict of interest: The authors declare that they have no conflict of interest.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Kumar P, Druckman A, Gallagher J, Gatersleben B, Allison S, Eisenman TS, et al. The nexus between air pollution, green infrastructure and human health. Environ Int. 2019;133:105181.

[2] Li X, Zheng W, Yin L, Yin Z, Song L, Tian X. Influence of social-economic activities on air pollutants in Beijing, China. Open Geosci. 2017;9(1):314–21.

[3] Zheng W, Li X, Yin L, Wang Y. Spatiotemporal heterogeneity of urban air pollution in China based on spatial analysis. Rendiconti Lincei. 2016;27(2):351–6.

[4] Chen X, Yin L, Fan Y, Song L, Ji T, Liu Y, et al. Temporal evolution characteristics of PM2. 5 concentration based on continuous wavelet transform. Sci Total Environ. 2020;699:134244.

[5] Wu D, Xu Y, Zhang S. Will joint regional air pollution control be more cost-effective? An empirical study of China's Beijing-Tianjin-Hebei region. J Environ Manag. 2015;149:27–36.

[6] Rahman MM, McConnell R, Schlaerth H, Ko J, Silva S, Lurmann FW, et al. The effects of coexposure to extremes of heat and particulate air pollution on mortality in California: implications for climate change. Am J Resp Critical Care Med. 2022;206(9):1117–27.

[7] Zanocco C, Flora J, Boudet H. Disparities in self-reported extreme weather impacts by race, ethnicity, and income in the United States. PLOS Climate. 2022;1(6):e0000026.

[8] Meister K, Johansson C, Forsberg B. Estimated short-term effects of coarse particles on daily mortality in Stockholm, Sweden. Environ Health Perspectives. 2012;120(3):431–6.

[9] Nakharutai N, Traisathit P, Thongsak N, Supasri T, Srikummoon P, Thumronglaohapun S, et al. Impact of residential concentration of PM2. 5 analyzed as time-varying covariate on the survival rate of lung cancer patients: A 15-Year hospital-based study in upper Northern Thailand. Int J Environ Res Public Health. 2022;19(8):4521.

[10] Zhang NN, Ma F, Qin CB, Li YF. Spatiotemporal trends in PM2. 5 levels from 2013 to 2017 and regional demarcations for joint prevention and control of atmospheric pollution in China. Chemosphere. 2018;210:1176–84.

[11] GBD 2016 Risk Factors Collaborators and others. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1345.

[12] Chen R, Kan H, Chen B, Huang W, Bai Z, Song G, et al. Association of particulate air pollution with daily mortality: the China Air Pollution and Health Effects Study. Am J Epidemiol. 2012;175(11):1173–81.

[13] Ramadan A. Detailed analysis of power generation and water desalination sector emissions-part 1: criteria pollutants and BTEX. Int J Environ Sci Technol. 2022;19(2):763–74.

[14] Thomas RJ, Turkelboom F. An integrated livelihoods-based approach to combat desertification in marginal drylands. In: The future of drylands. Tunis, Tunisia: Springer; 2008. p. 631–46.

[15] Al-Dousari A, Ramadan A, Al-Qattan A, Al-Ateeqi S, Dashti H, Ahmed M, et al. Cost and effect of native vegetation change on aeolian sand, dust, microclimate and sustainable energy in Kuwait. J. Taibah Univ. Sci. 2020;14(1):628–39.

[16] Nanney R, Fryrear D, Zobeck T. Wind erosion prediction and control. Water Sci Technol. 1993;28(3–5):519–27.

[17] Al-Dousari A, Al-Nassar W, Al-Hemoud A, Alsaleh A, Ramadan A, Al-Dousari N, et al. Solar and wind energy: challenges and solutions in desert regions. Energy. 2019;176:184–94.

[18] Al-Dousari A, Doronzo D, Ahmed M. Types, indications and impact evaluation of sand and dust storms trajectories in the Arabian Gulf. Sustainability. 2017;9(9):1526.

[19] Blott SJ, Al-Dousari AM, Pye K, Saye SE. Three-dimensional characterization of sand grain shape and surface texture using a nitrogen gas adsorption technique. J Sediment Res. 2004;74(1):156–9.

[20] Al-Dousari A, Al-Enezi A, Al-Awadhi J. Textural variations within different representative types of dune sediments in Kuwait. Arab J Geosci. 2008;1(1):17–31.

[21] Organización Mundial de la Salud and Weltgesundheitsorganisation and World Health Organization and European Centre for Environment. WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization; 2021.

[22] Peng M, Zhang H, Evans RD, Zhong X, Yang K. Actual air pollution, environmental transparency, and the perception of air pollution in China. J Environ Develop. 2019;28(1):78–105.

[23] Alsaber A, Alsahli R, Al-Sultan A, Abu Doush I, Sultan K, Alkandary D, et al. Evaluation of various machine learning prediction methods for particulate matter PM 10 in Kuwait. Int J Inform Technol. 2023;15(8):4505–19.

[24] Doush IA, Sultan K, Alsaber A, Alkandari D, Abdullah A. Improving neural network using Jaya algorithm with opposite learning for air quality prediction. In: Intelligent Systems Conference. Springer; 2023. p. 597–606.

[25] Fahad S, Sonmez O, Saud S, Wang D, Wu C, Adnan M, et al. Plant growth regulators for climate-smart agriculture. Boca Raton: CRC Press; 2021.

[26] Fahad S, Sonmez O, Saud S, Wang D, Wu C, Adnan M, et al. Sustainable soil and land management and climate change. Boca Raton: CRC Press; 2021.

[27] Fahad S, Sonmez O, Saud S, Wang D, Wu C, Adnan M, et al. Climate change and plants: biodiversity, growth and interactions. Boca Raton: CRC Press; 2021.

[28] Rybarczyk Y, Zalakeviciute R. Assessing the COVID-19 impact on air quality: A machine learning approach. Geophys Res Letters. 2021;48(4):e2020GL091202.
[29] Al-Jamimi HA, Al-Azani S, Saleh TA. Supervised machine learning techniques in the desulfurization of oil products for environmental protection: A review. Process Safety Environ Protection. 2018;120:57–71.
[30] Karthikeyani S, Rathi S. A survey on air quality prediction using traditional statistics method. Int J Sci Res Comput Sci Eng Inf Technol. 2020;6:942–6.
[31] Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159–75.
[32] Wang CY, Zhang WY, Wang JJ, Zhao WF. The prediction of SO2 pollutant concentration using a RBF neural network. In: Applied mechanics and materials. vol. 55. Baech, Switzerland: Trans Tech Publ; 2011. p. 1392–6.
[33] Cai M, Yin Y, Xie M. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. Transport Res Part D Transport Environ. 2009;14(1):32–41.
[34] Castelli M, Clemente FM, Popovič A, Silva S, Vanneschi L. A machine learning approach to predict air quality in California. Complexity. 2020;2020:8049504.
[35] Ganesh SS, Arulmozhivarman P, Tatavarti R. Forecasting air quality index using an ensemble of artificial neural networks and regression models. J Intelligent Syst. 2019;28(5):893–903.
[36] Rao R. Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. Int J Industr Eng Comput. 2016;7(1):19–34.
[37] Singh SP, Prakash T, Singh V, Babu MG. Analytic hierarchy process based automatic generation control of multi-area interconnected power system using Jaya algorithm. Eng Appl Artif Intell. 2017;60:35–44.
[38] Yu K, Qu B, Yue C, Ge S, Chen X, Liang J. A performance-guided JAYA algorithm for parameters identification of photovoltaic cell and module. Appl Energy. 2019;237:241–57.
[39] Degertekin S, Lamberti L, Ugur I. Sizing, layout and topology design optimization of truss structures using the Jaya algorithm. Appl Soft Comput. 2018;70:903–28.
[40] Wang L, Huang C. A novel elite opposition-based Jaya algorithm for parameter estimation of photovoltaic cell models. Optik. 2018;155:351–6.
[41] Wang SH, Phillips P, Dong ZC, Zhang YD. Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. Neurocomputing. 2018;272:668–76.
[42] Rao R, More K, Taler J, Ocłoń P. Dimensional optimization of a micro-channel heat sink using Jaya algorithm. Appl Thermal Eng. 2016;103:572–82.
[43] Warid W, Hizam H, Mariun N, Abdul-Wahab NI. Optimal power flow using the Jaya algorithm. Energies. 2016;9(9):678.
[44] Gao K, Yang F, Zhou M, Pan Q, Suganthan PN. Flexible job-shop rescheduling for new job insertion by using discrete Jaya algorithm. IEEE Trans Cybernet. 2018;49(5):1944–55.
[45] Wu C, He Y. Solving the set-union knapsack problem by a novel hybrid Jaya algorithm. Soft Comput. 2020;24(3):1883–902.
[46] Awadallah MA, Al-Betar MA, Hammouri AI, Alomari OA. Binary JAYA algorithm with adaptive mutation for feature selection. Arab J Sci Eng. 2020;45:1–16.
[47] Rao RV, Saroj A. Economic optimization of shell-and-tube heat exchanger using Jaya algorithm with maintenance consideration. Appl Thermal Eng. 2017;116:473–87.
[48] Farah A, Belazi A. A novel chaotic Jaya algorithm for unconstrained numerical optimization. Nonlinear Dyn. 2018;93(3):1451–80.
[49] Rao R, More K. Design optimization and analysis of selected thermal devices using self-adaptive Jaya algorithm. Energy Conversion Management. 2017;140:24–35.
[50] Rao RV, Saroj A. A self-adaptive multi-population based Jaya algorithm for engineering optimization. Swarm Evolut Comput. 2017;37:1–26.
[51] Lacroix B, Molina D, Herrera F. Region-based memetic algorithm with archive for multimodal optimisation. Inform Sci. 2016;367:719–46.
[52] Zhang YH, Gong YJ, Chen WN, Zhan ZH, Zhang J. A generic archive technique for enhancing the niching performance of evolutionary computation. In: 2014 IEEE Symposium on Swarm Intelligence. IEEE; 2014. p. 1–8.
[53] Kundu S, Biswas S, Das S, Suganthan PN. Crowding-based local differential evolution with speciation-based memory archive for dynamic multimodal optimization. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation; 2013. p. 33–40.
[54] Wang ZJ, Zhan ZH, Lin Y, Yu WJ, Yuan HQ, Gu TL, et al. Dual-strategy differential evolution with affinity propagation clustering for multimodal optimization problems. IEEE Trans Evolut Comput. 2017;22(6):894–908.
[55] Sheng W, Wang X, Wang Z, Li Q, Chen Y. Adaptive memetic differential evolution with niching competition and supporting archive strategies for multimodal optimization. Inform Sci. 2021;573:316–31.
[56] Turky AM, Abdullah S. A multi-population harmony search algorithm with external archive for dynamic optimization problems. Inform Sci. 2014;272:84–95.
[57] Kalra S, Rahnamayan S, Deb K. Enhancing clearing-based niching method using delaunay triangulation. In:2017 IEEE Congress on Evolutionary Computation (CEC). IEEE; 2017. p. 2328–37.
[58] Hosseini Nejad Takhti A, Saffari A, Martín D, Khishe M, Mohammadi M, et al. Classification of marine mammals using the trained multilayer perceptron neural network with the whale algorithm developed with the fuzzy system. Comput Intell Neurosci. 2022;2022:1–21.

[59] Qiao W, Khishe M, Ravakhah S. Underwater targets classification using local wavelet acoustic pattern and multilayer perceptron neural network optimized by modified Whale optimization algorithm. Ocean Eng. 2021;219:108415.
[60] Kosarirad H, Ghasempour Nejati M, Saffari A, Khishe M, Mohammadi M, et al. Feature selection and training multilayer perceptron neural networks using grasshopper optimization algorithm for design optimal classifier of big data sonar. J Sensors. 2022;2022:1–14.
[61] Khishe M, Safari A. Classification of sonar targets using an MLP neural network trained by dragonfly algorithm. Wireless Personal Commun. 2019;108(4):2241–60.
[62] Kaveh M, Khishe M, Mosavi M. Design and implementation of a neighborhood search biogeography-based optimization trainer for classifying sonar dataset using multilayer perceptron neural network. Analog Integrated Circuits Signal Proces. 2019;100:405–28.
[63] Wu J, Khishe M, Mohammadi M, Karim SHT, Shams M. Acoustic detection and recognition of dolphins using swarm intelligence neural networks. Appl Ocean Res. 2021;115:102837.
[64] Tian Y, Khishe M, Karimi R, Hashemzadeh E, Pakdel Azar O. Underwater image detection and recognition using radial basis function neural networks and chimp optimization algorithm. Circuits Syst Signal Proces. 2023;42(7):3963–82.
[65] Khishe M, Ebrahimi E, Goldani A. Designing a sonar system with the ability of classifying active and passive acoustic targets based on the evolutionary neural network. Iran: The Faculty and Research Institute of Basic Sciences of Imam Hossein (AS) University; 2020.
[66] Khishe M, Mosavi M. Classification of underwater acoustical dataset using neural network trained by chimp optimization algorithm. Appl Acoustics. 2020;157:107005.
[67] Saffari A, Khishe M. Classification of marine mammals using trained multilayer perceptron neural network with whale algorithm developed with fuzzy system. United Kingdom: Hindawi; 2020.
[68] Lu C, Gao L, Yi J. Grey wolf optimizer with cellular topological structure. Expert Syst Appl. 2018;107:89–114.
[69] Seng D, Zhang Q, Zhang X, Chen G, Chen X. Spatiotemporal prediction of air quality based on LSTM neural network. Alexandr Eng J. 2021;60(2):2021–32.
[70] Cordova CH, Portocarrero MNL, Salas R, Torres R, Rodrigues PC, López-Gonzales JL. Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru. Sci Rep. 2021;11(1):1–19.
[71] Aljanabi M, Shkoukani M, Hijjawi M. Ground-level ozone prediction using machine learning techniques: A case study in Amman. Jordan Int J Automat Comput. 2020;17(5):667–77.
[72] Al-Rashed A, Al-Mutairi N, Al Attar M. Air pollution analysis in Kuwait using a statistical technique (CUSUM). Int J Geosci. 2019;10(3):254–94.
[73] Zhao X, Zhang R, Wu JL, Chang PC. A deep recurrent neural network for air quality classification. J Inf Hiding Multim Signal Process. 2018;9(2):346–54.
[74] Ebrahimi-Khusfi Z, Taghizadeh-Mehrjardi R, Nafarzadegan AR. Accuracy, uncertainty, and interpretability assessments of ANFIS models to predict dust concentration in semi-arid regions. Environ Sci Pollution Res. 2021;28(6):6796–810.
[75] Maleki H, Sorooshian A, Goudarzi G, Baboli Z, Tahmasebi Birgani Y, Rahmati M. Air pollution prediction by using an artificial neural network model. Clean Technol Environ Policy. 2019;21(6):1341–52.
[76] Yang H, Liu Z, Li G. A new hybrid optimization prediction model for PM2. 5 concentration considering other air pollutants and meteorological conditions. Chemosphere. 2022;307:135798.
[77] Espinosa R, Palma J, Jiménez F, Kamińska J, Sciavicco G, Lucena-Sánchez E. A time series forecasting based multi-criteria methodology for air quality prediction. Appl Soft Comput. 2021;113:107850.
[78] Freeman BS, Taylor G, Gharabaghi B, Thé J. Forecasting air quality time series using deep learning. J Air Waste Manag Assoc. 2018;68(8):866–86.
[79] Sun K, Huang SH, Wong DSH, Jang SS. Design and application of a variable selection method for multilayer perceptron neural network with LASSO. IEEE Trans Neural Networks Learn Syst. 2016;28(6):1386–96.
[80] Tsantekidis A, Passalis N, Tefas A, Kanniainen J, Gabbouj M, Iosifidis A. Forecasting stock prices from the limit order book using convolutional neural networks. In: 2017 IEEE 19th conference on business informatics (CBI). vol. 1. IEEE; 2017. p. 7–12.
[81] Abu-Doush I, Ahmed B, Awadallah MA, Al-Betar MA, Rababaah AR, Enhancing multilayer perceptron neural network using archive-based harris hawks optimizer to predict gold prices. J King Saud Univ Comput Inf Sci. 2023;35(5):101557.
[82] Rao RV, Rai DP. Optimisation of welding processes using quasi-oppositional-based Jaya algorithm. J Experiment Theoretic Artif Intel. 2017;29(5):1099–117.
[83] Sihwail R, Omar K, Ariffin KAZ, Tubishat M. Improved harris hawks optimization using elite opposition-based learning and novel search mechanism for feature selection. IEEE Access. 2020;8:121127–45.
[84] Abed-alguni BH, Paul D. Island-based cuckoo search with elite opposition-based learning and multiple mutation methods for solving optimization problems. Soft Comput. 2022;26(7):3293–312.
[85] Yildiz BS, Pholdee N, Bureerat S, Yildiz AR, Sait SM. Enhanced grasshopper optimization algorithm using elite opposition-based learning for solving real-world engineering problems. Eng Comput. 2022;38(5):4207–19.
[86] Albanai JA. The effect of the coronavirus pandemic (Covid-19) on the tropospheric NO2 concentration changes in the Arabian Peninsula and the State of Kuwait. Arab J GIS;14(1):36–68.
[87] Agency UEP. National ambient air quality standards for ozone; Final rule. Fed Regist. 2015;80(206):65–292.

[88] Al-Shayji K, Lababidi H, Al-Rushoud D, Al-Adwani H. Development of a fuzzy air quality performance indicator. Kuwait J Sci Eng. 2008;35:101–26.
[89] Van den Elshout S, Léger K, Nussio F. Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative. Environ Int. 2008;34(5):720–6.
[90] Kanchan AKG, Goyal P, Benitez-Garcia SE, Kanda I, Okazaki Y, Wakamatsu S, et al. A review on air quality indexing system. Asian J Atmos Environ. 2015;9(2):101–13.
[91] Plaia A, Ruggieri M. Air quality indices: a review. Rev Environ Sci Bio/Technol. 2011;10(2):165–79.
[92] Cheng WL, Chen YS, Zhang J, Lyons T, Pai JL, Chang SH. Comparison of the revised air quality index with the PSI and AQI indices. Sci Total Environ. 2007;382(2–3):191–8.
[93] Murena F. Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. Atmospheric Environ. 2004;38(36):6195–202.
[94] Al-Fadhli AA. Ambient air quality assessment of twelve inhabited areas in the state of Kuwait between years 2011-2014. Int J Chem Eng Appl. 2017;8(5):340–3.
[95] Kumar A, Goyal P. Forecasting of daily air quality index in Delhi. Sci Total Environ. 2011;409(24):5517–23.
[96] Van den Elshout S, Léger K, Heich H. CAQI common air quality index–update with PM_{10} and sensitivity analysis. Sci Total Environ. 2014;488:461–8.
[97] Chen R, Wang X, Meng X, Hua J, Zhou Z, Chen B, et al. Communicating air pollution-related health risks to the public: An application of the Air Quality Health Index in Shanghai, China. Environ Int. 2013;51:168–73.
[98] Mintz D. Guidelines for the reporting of daily air quality—air quality index (AQI). Washington: United States Environmental Protection Agency; 2006.
[99] Mintz D. Technical assistance document for the reporting of daily air quality-the air quality index (AQI). Tech Research Triangle Park, US Environmental Protection Agency. 2009.
[100] World Health Organization and others WHO air quality guidelines global update 2005: Report on a Working Group Meeting, Bonn, Germany, 18–20 October 2005. World Health Organization. Regional Office for Europe; 2005.
[101] Johnson M, Isakov V, Touma J, Mukerjee S, Özkaynak H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmospheric Environ. 2010;44(30):3660–8.
[102] Johnson DL, Ambrose SH, Bassett TJ, Bowen ML, Crummey DE, Isaacson JS, et al. Meanings of environmental terms. J Environ Quality. 1997;26(3):581–9.
[103] Kowalska M, Ośródka L, Klejnowski K, Zejda JE, Krajny E, Wojtylak M. Air quality index and its significance in environmental health risk communication. Archives Environ Protect. 2009;35(1):13–21.
[104] Doan M, East C. A proposed air quality index for urban areas. Water Air Soil Pollution. 1977;8(4):441–51.
[105] Stieb DM, Doiron MS, Blagden P, Burnett RT. Estimating the public health burden attributable to air pollution: an illustration using the development of an alternative air quality index. J Toxicol Environ Health Part A. 2005;68(13–14):1275–88.
[106] Pruss-Ustun A, Corvalán CF and World Health Organization and others. Preventing disease through healthy environments: toward an estimate of the environmental burden of disease. World Health Organization; 2006.
[107] Fitz-Simons T. Guideline for reporting of daily air quality: Air Quality Index (AQI). Research Triangle Park, NC (United States): Environmental Protection Agency, Office of Air Quality Planning and Standards; 1999.
[108] Alsaber A, Pan J, Al-Herz A, Alkandary DS, Al-Hurban A, Setiya P, et al. Influence of ambient air pollution on rheumatoid arthritis disease activity score Index. Int J Environ Res Public Health. 2020;17(2):416.
[109] Forbes D, Hawthorne G, Elliott P, McHugh T, Biddle D, Creamer M, et al. A concise measure of anger in combat-related posttraumatic stress disorder. J Traumatic Stress Official Publication Int Soc Traumatic Stress Studies. 2004;17(3):249–56.
[110] Breiman L. Random forests. Machine Learn. 2001;45(1):5–32.
[111] Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.
[112] Oba S, Sato Ma, Takemasa I, Monden M, Matsubara K, Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003;19(16):2088–96.
[113] Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95-International Conference on Neural Networks. vol. 4. IEEE; 1995. p. 1942–8.
[114] Yang XS. A new metaheuristic bat-inspired algorithm. In: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010). Springer; 2010. p. 65–74.
[115] Mirjalili S. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-based Syst. 2015;89:228–49.