Computers, Materials & Continua
DOI: 10.32604/cmc.2022.029604
Article

Tech Science Press

Big Data Analytics with Artificial Intelligence Enabled Environmental Air Pollution Monitoring Framework

Manar Ahmed Hamza[1,*], Hadil Shaiba[2], Radwa Marzouk[3], Ahmad Alhindi[4], Mashael M. Asiri[5], Ishfaq Yaseen[1], Abdelwahed Motwakel[1] and Mohammed Rizwanullah[1]

[1]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, 16278, Saudi Arabia
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh, 11671, Saudi Arabia
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh, 11671, Saudi Arabia
[4]Department of Computer Sciences, College of Computing and Information System, Umm Al-Qura University, Saudi Arabia & Research and Innovation, Salla Holding Limited, Makkah, Saudi Arabia
[5]Department of Computer Science, College of Science and Arts, King Khalid University, Mahayil Asir, 62529, Saudi Arabia
*Corresponding Author: Manar Ahmed Hamza. Email: ma.hamza@psau.edu.sa
Received: 07 March 2022; Accepted: 06 April 2022

Abstract: Environmental sustainability is the rate of renewable resource harvesting, pollution control, and non-renewable resource exhaustion. Air pollution is a significant issue confronted by the environment particularly by highly populated countries like India. Due to increased population, the number of vehicles also continues to increase. Each vehicle has its individual emission rate; however, the issue arises when the emission rate crosses the standard value and the quality of the air gets degraded. Owing to the technological advances in machine learning (ML), it is possible to develop prediction approaches to monitor and control pollution using real time data. With the development of the Internet of Things (IoT) and Big Data Analytics (BDA), there is a huge paradigm shift in how environmental data are employed for sustainable cities and societies, especially by applying intelligent algorithms. In this view, this study develops an optimal AI based air quality prediction and classification (OAI-AQPC) model in big data environment. For handling big data from environmental monitoring, Hadoop MapReduce tool is employed. In addition, a predictive model is built using the hybridization of ARIMA and neural network (NN) called ARIMA-NN to predict the pollution level. For improving the performance of the ARIMA-NN algorithm, the parameter tuning process takes place using oppositional swallow swarm optimization (OSSO) algorithm. Finally, Adaptive neuro-fuzzy inference system (ANFIS) classifier is used to classify the air quality into pollutant and non-pollutant. A detailed experimental analysis is performed for highlighting the better prediction performance of the proposed ARIMA-NN method. The obtained outcomes pointed out the enhanced outcomes of the proposed OAI-AQPC technique over the recent state of art techniques.

Keywords: Sustainability; environmental air quality; predictive model; pollution monitoring; statistical models; artificial intelligence

1 Introduction

With the technological and economic improvement of cities, environmental pollution challenges are rising, like air, water, and noise pollution. Particularly, air pollution is a significant effect on human health over the disclosure of particulates and pollutants that has greater attention in air pollution and their effects amongst the academic research [1,2]. The major main cause related to air pollution comprises agriculture, burning of fossil fuels, natural disasters, residential heating, exhaust from industries and factories.

Recently, the Internet of Things (IoT) model has developed, enabling objects of daily lives with transceivers for digital communication, a suitable protocol stack, and microcontrollers empowers them to interact with each other, becomes an essential component of the Internet. Because of this accumulation, devices like surveillance cameras, home appliances, vehicles, and sensors could produce large numbers of data which could be consequently analyzed and utilized for developing novel applications. The current advancement in IoT techniques has enabled persons to improve effective systems consists of WiFi/Bluetooth for near regions, satellites for remote regions and mobile networks, multiple sensors connected wirelessly for monitoring distinct air pollutants in various fields. Such amount of data may require parallelization utilizing cloud computing and big data technique for facilitating nearby real-world monitoring [3] and analyzing evolving patterns.

The monitoring system allows gathering quality data that could be utilized for extracting deeper knowledge on pollution [4]. Air pollution predicting system enables to forecast the Air Quality Index (AQI), the values of pollutant, like carbon dioxide concentration (i.e., PM2.5, PM10, CO2, etc) or particle matter (PM), and identify higher pollution regions. Prediction air pollution system could assist government employs smart solutions and precautions for addressing air quality issues. Even though several models and solutions for forecasting air pollutions were introduced in the study, usually they are categorized into 2 classifications. Firstly, it tracks the transmission, generation, and dispersion processes of pollutants. The mathematical simulation generates the prediction result of this model. Next, it consists of deep learning (DL), statistical learning, and machine learning (ML) methods [5].

The current study shows that the conventional deterministic method struggles for capturing the nonlinear relations among the focuses of contaminant and their sources of dispersion and emission [6], particularly in a model application in region of complicated environment. In order to address the limitation of conventional method, the most significant method is to utilize statistical model on the basis of ML methods. Statistical methods do not consider the chemical and physical procedures and utilize past data for predicting air quality. Methods were trained on present measurement and utilized for estimating or forecasting concentration of air pollutants based on the prediction features (such as land use, meteorology, human activity, time, planetary boundary layer, elevation, pollutant covariates, and so on.). The simple statistical approach includes Time Series, Autoregressive Integrated Moving Average (ARIMA), and Regression model. Such analysis determines the relations among parameters according to the possibility and statistical average. Well stated regression could give moderate results. But, the reactions among influential factors and air pollutants are nonlinear, leads to complicated systems of air pollutant formation mechanism. Hence, more innovative statistic learning (or ML) approaches are desirable to account for accurate nonlinear modeling of air pollution. Such

as, Ensemble Learning, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) were employed to conquer nonlinear uncertainties and limitations for achieving an optimal predictive accuracy. Even though statistical methods don't explicitly simulate the environment process, they usually exhibit a high prediction efficiency compared to CTM on fin spatio-temporal scales in the existence of wide monitoring data [7].

In the application of big data services and technology in the procedure of ecological civilization construction and current environment protection, they could moderately utilize big data for solving few complex challenges in environment protection works 1) data collection and disclosure; 2) air quality prediction and earlier warnings; 3) utilizing big data acquisition technique for annualizing the cause of environmental pollution. The data analyses activity is performed by integrating different kinds of emission information of pollution sources and environment indicators. Using moderate prediction and scientific analysis of enterprise sewage intensity, its impacts on the surrounding environmental quality, and distribution of pollution sources, the environment treatment proposal is developed, and the impact of environment treatments is regularly monitored, and the treatment plans are continuously improving. The application and development of big data techniques could be stated to offer a novel method for humans to handle environmental challenges [8]. For environmentalists, it is essential to actively promoting the reasonable application of big data techniques in the region of environment protection, in order to obtain more independence in understanding the objective law of natural developments.

This study develops an optimal AI based air quality prediction and classification (OAI-AQPC) model in big data environment. To manage the big data from environmental monitoring, Hadoop MapReduce tool is used. Moreover, a hybridization of ARIMA and neural network (NN) called ARIMA-NN is designed to predict the pollution level, and the parameters in the AIMA-NN model are tuned by the use of oppositional swallow swarm optimization (OSSO) algorithm. At last, Adaptive neuro-fuzzy inference system (ANFIS) classifier is used to classify the air quality into pollutant and non-pollutant. The use of ARIMA-NN technique is utilized for predicting the air quality parameters and the ANFIS technique is employed to categorize the air quality into pollutants and non-pollutants depicts the novelty of the study. A wide range of simulation analyses is carried out and the results are inspected interms of different dimensions.

2 Literature Review

Kalajdjieski et al. [9] proposed a new methodology calculating 4 distinct frameworks which exploit camera images to evaluate the air pollution in this region. The projected method utilizes generative adversarial network and data augmentation methods for mitigating class imbalance problems. In Ghaemi et al. [10], a spatio-temporal method is proposed by a LaSVM based online approach. Meteorological data, geographical, Pollutant concentration parameters are repeatedly fed into the advanced online predicting method. The efficiency of the scheme is calculated by relating the predictive result of the AQI with conventional SVM method. Zhang et al. [11] proposed a predictive approach integrating Long Short-Term Memory networks (LSTM) and Graph Attention (GAT) model. In this case, the LSTM is utilized for defining the temporal correlation of historical data and GAT is applied to represent the spatial correlation amongst all the monitoring stations in the destination.

Honarvar et al. [12], developed a prediction method for PM predictions. The presented method contains many elements for integrating heterogeneous multiple sources of urban data and forecast the particulate matter according to the TL point of view, where regression and NN were leveraged as the heart of predictions. The result of the particulate matter predictions revealed that data sources can able

to predict accurate particulate matter. Zaree et al. [13] aim is to raise accuracy and speed in predicting location, effects of weather conditions on density of air pollution, and real levels of air pollution, a K-means clustering method with Mahout library is utilized as a big data mining tool on data sets of a city pulse project.

Shahbaz et al. [14] explored the new phenomenon of a BDA-EAP management scheme and presented a study of factors affecting adaption of this scheme. This study is depending on a TTF and unified concept of acceptance and UTAUT concept. A complete BDA-EAP management scheme is presented and the possible adaption speed of this scheme is calculated by transmitting structured forms to the employees of appropriate environment agency, yield 412 valid replies, utilizing structural equation modelling method. Al-Janabi et al. [15] designed a smart predictor for the concentration of air pollutants on the following 2 days relying on DL method utilizing RNN. The optimal structure for its process is later defined by a PSO method. The novel predictor dependent smart computation is based on unsupervised learning, viz., LSTM and optimization (viz., PSO) is known as SAQPM.

Shih et al. [16] proposed a PM2.5 instant predictive framework based on Spark big data architecture for handling large data from the LASS community. The Spark big data architecture presented in the research is separated into 3 models. The simulation result shows that the presented Spark big data ensemble predictive method in following 30-min predictions has an optimal efficiency (R2 up to 0.96), and the ensemble method has higher efficiency compared to individual ML method. Castelli et al. [17] employ a common ML technique, SVR, to predict particulate and pollutant levels and to forecast the AQI. Amongst the different tested alternates, RBF was the kind of kernel which allows SVR for obtaining the precise prediction. Utilizing the entire set of available parameters exposed an effective approach compared to FS utilizing PCA.

Chang et al. [18] proposed a semantic ETL architecture on cloud environment for predicting AQ. In the environment, they exploit ontology for concretizing the relation of PM 2.5 from different data sources and to combine such data with a similar idea but distinct naming to the unified database. They implemented the ETL architecture on the cloud environment that consists of storage nodes and computing nodes. In Zou et al. [19], an AQI predictive method (i.e., airQP-DNN) and its application are projected for addressing this problem. Primarily, this study contains 2 modules. The initial module is to forecast the upcoming AQI on the basis of DNN, weather predicting datasets, present meteorological datasets, and using historical air quality datasets. Next, it refers to an analysis of outdoor events route planning in Beijing, that could assist plan the routes for outdoor events according to the air using QP-DNN method, and allows users to enter the source and end point of the routes for the optimized path using a minimal collected AQI.

3 The Proposed Model

This study has devised a new OAI-AQPC technique to predict and classify air quality in the big data environment. The proposed model involves Hadoop Map Reduce tool to manage big data. As demonstrated in Fig. 1, the workflow of the OAI-AQPC technique encompasses different processes namely ARIMA-NN based prediction, OSSO based parameter optimization, and ANFIS based classification. The use of ARIMA-NN technique is utilized for predicting the air quality parameters and the ANFIS manner is employed to categorize the air quality into pollutants and non-pollutants. The detailed working of these processes is elaborated in the subsequent sections.

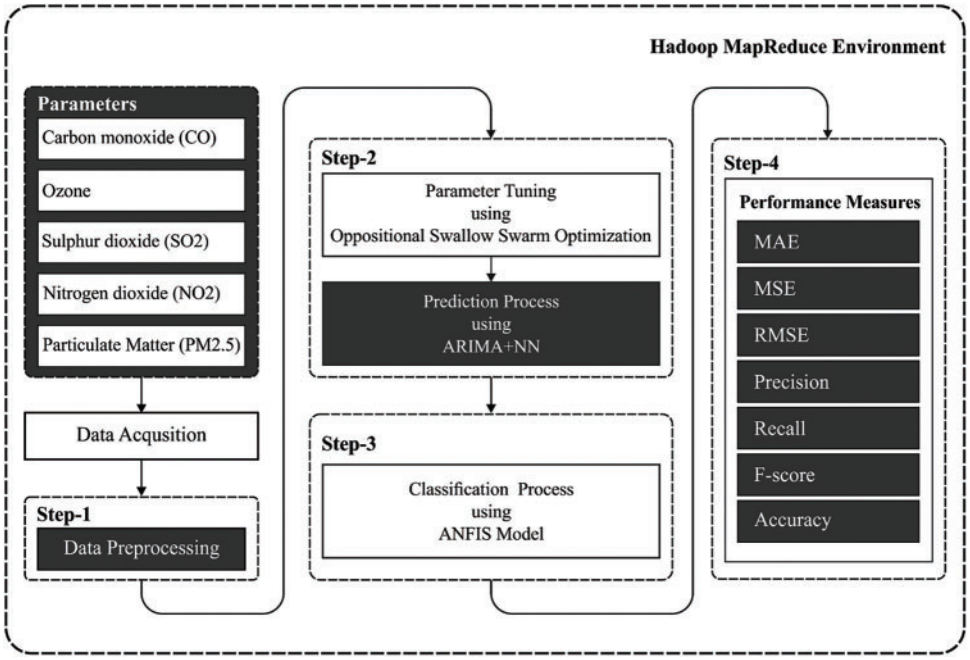

Figure 1: Overall process of OAI-AQPC model

3.1 Framework of Hadoop Map Reduce Model

The MR runtime environment is one of popular models utilized in the distributed processing system. As privative tool, its open-source counterpart, called Hadoop, was conventionally utilized in study field [20]. It was proposed for allowing distributed computation in an apparent manner for the programmer, as well as provide automated data partition and management, automatic job/resource scheduling, and fault tolerance. For using this system|, another process should be separated into 2 major phases: Reduce and Map. Firstly, it is dedicated to splitting the data to process, while the next one aggregates and collects the result. Moreover, the MR method is determined based on a fundamental data structure: the (key, value) pairs. The processing data, the intermediate and final result works based on (key, value) pair. For summarizing this process, Fig. 2 shows a usual MR program with its Map and Reduce phases. The MR system could be defined in the following.

- First, the map function reads information and transforms records to a key value format. Transformation in this stage might employ any series of operations on every record beforehand transmitting the tuples through the network.
- Then, the output key is grouped and shuffled with a key value thus equivalent keys are gathered for creating a list of values. Later, Keys are divided and transmitted to the Reducer based on some key based system determined before.
- Lastly, the reducer performs this fusion on the list for ultimately generating an individual value for all pairs. As a new optimization, the reducers are also utilized as a combiner on the map output.

The enhancement decreases the overall number of data transmitted through the network by integrating every word created in the Map phase to an individual pair. Besides, taking into account MR as a processing model, this system could be viewed as a fusion procedure which permits merging

information schemes and partial models to a final fused result. Fusion of methods in MR has usually executed this ensemble approach which integrates various hypotheses via attachment/voting. As well, another proposal exists that exceeds ensemble learning, and provide result as an individual coalesced method. As logistic regression in Spark is made up of various subgradients that are eventually aggregated and locally computed.

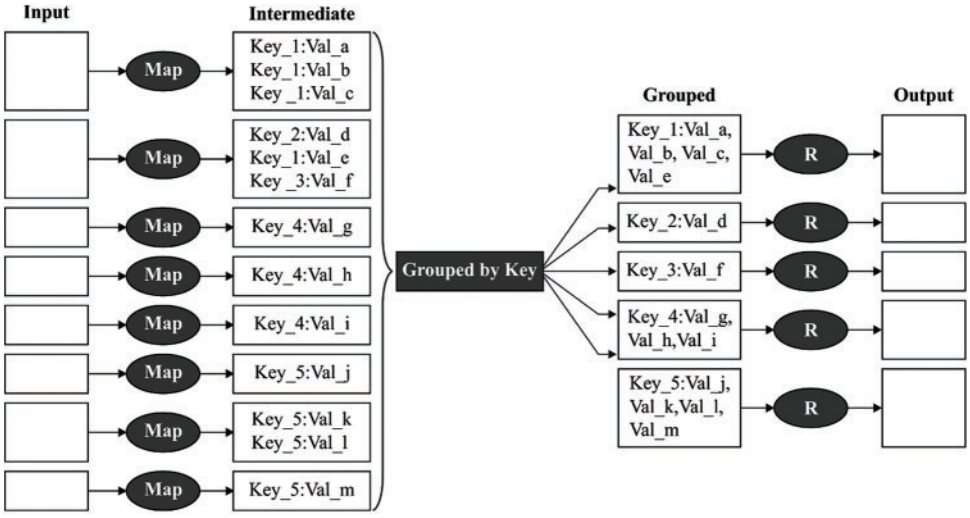

Figure 2: Framework of MR model

3.2 Prediction using ARIMA-NN Model

Primarily, the hybridization of ARIMA and NN takes place for the prediction of the pollution level in the environment. The ARIMA is initially proposed by Box and Jenkin in 1976. A common formula of successive variances at *dth* variance of X_t is given as follows:

$$\Delta^d X_t = (1 - B)^d X_t \tag{1}$$

where d implies the variance order and is generally 1 or 2, and B represents the backshift operator. The following variance at one-time lag is equivalent to,

$$\Delta^1 X_t = (1 - B)\, X_t = X_t - X_{t-1} \tag{2}$$

During this case, a common ARIMA (p, d, q) is written as follows [21]:

$$\Phi_p\,(B)\; W_t = \theta_q\,(B)\; e_t \tag{3}$$

where $\Phi_p(B)$ refers the auto-regressive operator of direction p, $\theta_q(B)$ is a moving average operator of order q, and $W_t = \Delta d X_t$.

The ARIMA model efficiency is estimated utilizing root mean square errors (RMSEs, Eq. (4)) and coefficient of resolve (R^2).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(X_t - X_0)^2}{n}} \tag{4}$$

where X_t signifies the estimated observation and X_0 refers the actual observation.

Amongst several NN frameworks, the feedforward BP network is mostly utilized. This network framework contains single hidden layer of neurons using a non-linear transfer function and an output layer of linear neurons using a linear transfer function. In the BP networks, $x_j = 1, .., N$ denotes the input parameters; $n_i (i = 1, \ldots, S)$ signifies the output of neurons in the hidden layer and y_t $(t = 1, \ldots L)$ indicates the output of NN. A NN should be trained for determining the values of weight that would generate the accurate output. In the training stage, a group of input data is utilized to train and present to the network several times. The efficiency of the network is tested afterward the training phase is ended. It is the direction where the efficiency function is reducing more quickly. It turned out that even though the function reduces more quickly with the negative of the gradient, it doesn't generate faster convergence [22]. Thus, the fundamental gradient descent training approach is ineffective because of its slower convergent speed and poor accuracy in prediction method. From an optimization perspective, training a NN could be deliberated as equal for minimizing a multi-variable global error function of the network weight. The SCG method was established for avoiding the time-consuming line search. The conventional BP method, usually applied in NN learning, calculates the gradient of global error function regarding the weights, (W^k), at all iterations and update the weight based on

$$W^{k+1} = W^{\mathrm{k}} - \alpha^k \nabla f\left(W^k\right) \tag{5}$$

The step size $\alpha^k > 0$ represents a user-selected learning rate parameter that affects the efficiency of learning process to a large amount. In each case, the BP approach might follow a zigzag path to the minimal, usual for a steepest gradient descent technique. In case D^k denotes the direction vector at iteration k, then the weight vectors are upgraded based on

$$W^{k+1} = W^k + \alpha^k D^k \tag{6}$$

The assumed values of W^k and D^k, certain values of α^k reduces the objective function more rapidly, which must be established. The evaluation of the optimal step size along SCG training method rises the learning speed and removes the dependency on crucial user-selected parameters. The major concept behindhand the process is the utilization of a factor ρ i.e., lowered/raised within all iterations at the time of implementation, look at the symbol of quantity δ exposes when the Hessian matrix isn't positive definite. The summary of SCG in NN is provided in following [23].

i) Initialization: At $\mathrm{k} = 0$, select a primary weight vector W°, and fix the primary direction vector to the negative gradient vector $D^o = G^o = -\nabla f\left(W^o\right)$. fix the scalar $0 < \sigma < 10^{-4}, 0 < \rho^0 \leq 10^{-6}$, $\overline{\rho}^o$, fix the boolean success = *true*.

ii) in case of success = true, evaluate next order data: $\sigma^k = \frac{\sigma}{|D^k|}$, $S_{\mathrm{k}} = \frac{\nabla f\left(W^k + \sigma^k D^k\right) - \nabla f\left(W^k\right)}{\sigma^k}$, $\delta^k =$transpose $\left(D^k\right) + S^k$

iii) Scale δ^k: $\delta^k = \delta^{\mathrm{k}} + (\rho^{\mathrm{k}} - \overline{\rho}^{\mathrm{k}})|D^{\mathrm{k}}|^2$, looking at the sign of δ^k for all iterations altering ρ^{k}. When $\delta^k \leq 0$, ρ^{k} & S_{k} is evaluated again.

iv) When $\delta^k \leq 0$ create the Hessian positive definite

v) $\overline{\rho} = 2(\rho^k - \delta^{\mathrm{k}}/|D^k|^2)$, $\delta^{\mathrm{k}} = -\delta^k + \rho^k |D^{\mathrm{k}}|^2$, $\rho^k = \overline{\rho}^k$

vi) Evaluate the step size: $\xi^k = trcmspose(D^{\mathrm{k}})G^k$, $\alpha^k = \xi^k/\delta^k$, the value of ρ^k directly measure the step size, bigger the ρ small the step size.

vii) Evaluate relation parameters c^k: $c^k = 2\delta^k \lfloor f(W^k) - f(W^k + \alpha^k D^k) \rfloor / (\xi^{\mathrm{k}})^2$

viii) Weight and direction upgrade: When $c^k \geq 0$, a success upgrade could be done: $W^{k+1} = W^k + \alpha^k D^k$, $G^{\mathrm{k}+1} = -\nabla f(W^{\mathrm{k}+1})$, $\overline{\rho}^{\mathrm{k}} = 0$, success = *true*. When k $modN = 0$ resume the process

using $D^{k+1} = G^{k+1}$ else $\beta^k = (|G^{k+1}|^2 - G^{k+1T}G^k)/\xi^k, D^{k+1} = G^{k+1} + \beta^k D^k$. If $c^k \geq 0.75$ reduces the scale variable to $\rho^k = \frac{1}{4}\rho^k$ else $\overline{\rho} = \rho$ success = *false*.

ix) When $c^k \geq 0.25$ rise the scale variable to $\rho^k = \rho^k + \delta^k(1 - c^k)/|D^k|^2$

x) Repetition: when the steepest descent direction $G^k \neq 0$, set $k = k + 1$ and return to step 2 otherwise end and return W^{k+1} as minimal.

3.3 Design of ARIMA-NN Model

The performance of air quality can be predicted using the OAI-AQPC technique. The estimation of the ARIMA model to complicate non-linear problems might not be sufficient. Alternatively, utilizing ANN for modeling linear problems has produced insufficient outcomes. A hybrid method containing nonlinear and linear modelling capabilities can be a better alternative to predict air quality data. With the integration of distinct methods, various features of fundamental patterns might be taken. Next, in hybrid method, the air quality time sequence can be made up of a linear nonlinear component and auto relation structure.

$$y_t = L_t + S_t \tag{7}$$

where L_c denotes the linear element and S_t signifies the nonlinear element. Those variables should be evaluated from the time sequence data. Firstly, the ARIMA method is utilized for capturing the linear element, later the residual from the linear method would have the nonlinear relation. The residual e_t at time t from the linear method is determined as follows

$$e_t = y_t - \hat{L}_t \tag{8}$$

where $\hat{L}_t$ denote the predictive values of ARIMA method at time t. The analytic check of the residual is significant for determining the sufficiency of ARIMA model. ARIMA models are not adequate when there is still linear relation structure remain in the residual. But, analytic check of the residual isn't capable of detecting nonlinear patterns in the time's sequence data. Therefore, though the residual passes the analytic check and the models are a sufficient one, still the method mightn't be sufficient because nonlinear relations haven't been modelled properly. Thus, the residual could be modelled with the help of ANNs for discovering nonlinear relationships. Using N input node, the ANN method for the residual would be given as

$$e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-N}) + \varepsilon_t \tag{9}$$

whereas f denotes a nonlinear function defined using a NN and ε_t represents the arbitrary error. Lastly, the integrated predictions would be

$$\hat{y}_t = \hat{L}_t + \hat{S}_t \tag{10}$$

where $\hat{S}_t$ denotes the prediction from Eq. (9).

3.4 Parameter Tuning using OSSO Algorithm

For enhancing the predictive efficiency of the AIRMA-NN model, the parameter tuning process gets executed by the use of OSSO algorithm. The SSO technique simulated as the combined effort of swallow and the interface amongst flock members has gained optimum outcomes. This technique was projected a metaheuristic approach dependent upon special properties of swallows are comprising fast flight, hunting skill, and intelligent social relation. At a glance, this technique has same as PSO but it can be unique features that could not be initiate in same techniques are containing the utilize of 3 kinds of particles: Explorer Particles (e_j), Aimless Particles (o_i), and Leader Particles (l_i), every of that is certain responsibility in the group. The e_j particle is responsible to search the problem space.

It can be carried out this exploring performance in the control of the number of parameters [24]. The particles utilize the subsequent formulas to explore and continue the path:

$$V_{HL_{i+1}} = V_{HL_j} + \alpha_{HL} rand()(e_{best} - e_j) + \beta_{HL} rand()(HL_j - e_j) \tag{11}$$

Eq. (11) illustrates the velocity vector variable in the path of global leaders.

$$\alpha_{HL} = \{if\ (e_j = 0 | |e_{best} = 0) - - > 1.5\} \tag{12}$$

Eqs. (12) and (13) compute the acceleration coefficients variable (α_{HL}) that directly impacts individual experience of all particles.

$$\alpha_{HL} = \begin{cases} \text{if } (e_i < e_{best}) \&\& (< HL_i) \rightarrow \dfrac{rand() \cdot e_i}{e_i \cdot \mathrm{e_{best}}} e_i, e_{best} \neq 0 \\ \text{if } (e_i < e_{best}) \&\& (e_i > HL_i) \rightarrow \dfrac{2rand() \cdot e_i}{\dfrac{1}{(2 \cdot \mathrm{e_i})}} e_i \neq 0 \\ \text{if } (e_i > e_{best}) \rightarrow \dfrac{e_{best}}{\dfrac{1}{(2 \cdot \mathrm{rand}())}} \end{cases} \tag{13}$$

$$\beta_{HL} = \{if\ (e_j = 0 | |e_{best} = 0) - - > 1.5\} \tag{14}$$

$$\beta_{HL} = \begin{cases} if\,(e_i < e_{best}) \&\& (e_i < HL_i) \rightarrow \frac{rand() \cdot e_i}{e_i \cdot HL_i} e_i, HL_i \neq 0 \\ if\,(e_i < e_{best}) \&\& (e_i > HL_i) \rightarrow \frac{2rand() \cdot e_i}{\frac{1}{(2 \cdot \mathrm{e_i})}} e_i \neq 0 \\ \text{if } (e_i > e_{best}) \rightarrow \frac{HL_i}{\frac{1}{(2 \cdot \mathrm{rand}())}} \end{cases} \tag{15}$$

Eqs. (14) and (15) computes the acceleration coefficient variables (β_{HL}) that directly affects the combined experiences of all particles. Actually, these 2 acceleration coefficients are quantified allowing for the place of all particles relative to optimum individual experience and global leaders. The o_i particles utilize the subsequent Eq. (16) for arbitrary movements:

$$o_{i+1} = o_i + \left[rand\,(\{-1, 1\}) * \frac{rand(\min_s, \max_s)}{1 + rand\,()} \right] \tag{16}$$

In SSO technique, there are 2 kinds of leaders: the local as well as global leaders. The particles are separated as to groups. The particles in all groups are frequently same. Afterward, an optimum particle in all groups is elected and is named as local leader. Then, an optimum particle amongst the local leaders is selected and is named as global leader. The particle alteration its way and converge based on place of these particles.

To enhance the convergence rate of the SSO algorithm, the OSSO algorithm is designed by the incorporation of OBL concept. It performs by exploring all directions in the search space namely actual and opposite solutions. The opposite number x is defined as a real number in the range of $x \in [lb, ub]$. The opposite number of x is represented by $\tilde{x}$ [25]:

$$\tilde{x} = lb + ub - x \tag{17}$$

To generalize it, every searching agent and the opposite solution is represented as follows.

$$x = [x_1, x_2, x_3, \ldots x_D] \tag{18}$$

$$\tilde{x} = [\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \ldots, \tilde{x}_D] \tag{19}$$

The value of all the components in $\tilde{x}$ can be represented using Eq. (20):

$$\tilde{x}_j = lb_j + ub_j - x_j where\ j = 1, 2, 3, \ldots, D \tag{20}$$

If the fitness value $f\left(\tilde{x}\right)$ of the opposite solution exceeds the $f\left(x\right)$ of the original solution x, then $x = \tilde{x}$; else $x = x$.

3.5 ANFIS Based Classification

Once the air quality is predicted by the ARIMA-NN technique, the next stage involves the classification of air quality parameters into pollutants and non-pollutants using the ANFIS classifier. The ANFIS model integrates the processing ability of the ANN and high-level logical ability of the FIS technique. Alternatively, the ANFOS model is derived by adding the FIS into the adaptive network model. Due to the proficient learning and reasonability ability of the ANFIS technique, it is employed for the classification of air pollution monitoring in this study [26]. It comprises different characteristics which help to accomplish improved performance. The ANN model can determine the patterns by adapting to the platform with the learning abilities. In addition, the FLS technique integrates the expert's opinion to make decisions. The capability of handling past data and expert opinion makes it adaptable to unusual situations. It is called an intelligent model due to the variables and fuzzy rules involved that are determined by the ANN model in an intelligent way.

4 Performance Validation

This section inspects the predictive and classification results analysis of the OAI-AQPC technique. The OAI-AQPC technique is simulated using Python 3.6.5 tool. The OAI-AQPC technique is validated using a dataset comprising different air quality parameters such as Carbon monoxide (CO), Ozone, Sulphur dioxide (SO2), Nitrogen dioxide (NO2), and Particulate Matter (PM2.5). The dataset holds a total of 22321 instances. The AQI values are grouped into 6 different classes as shown in Fig. 3. In addition, the good and moderate values come under 'Non-Pollutant' class (15738 instances) and the remaining values come under 'Pollutant' class (6583 instances). Besides, the statistical values of the dataset are given in Tab. 1.

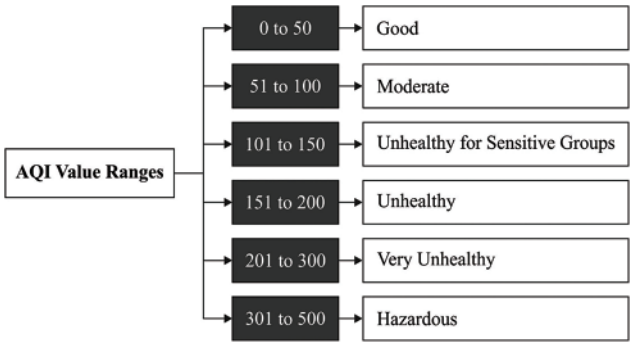

Figure 3: Six different classes

Table 1: Statistics of applied dataset

Variables	Min.	Max.	Mean
CO	0.06	2.17	0.27
Ozone	0.00	0.08	0.03
SO2	−0.57	158.37	1.07
NO2	0.28	55.58	12.00
PM2.5	−2.78	240.00	7.71

A brief forecasting results analysis of the ARIMA-NN technique is examined in Tab. 2 and Fig. 4 under varying training and validation sets. From the results, it is ensured that the ARIMA-NN technique has forecasted the variables effectively with the minimum MSE, MSE, and RMSE values. For instance, the ARIMA-NN technique predicted the 'CO' variable with the lower RMSE of 0.170 and 0.490 on the applied training and validation sets respectively. In addition, the ARIMA-NN approach predicted the 'Ozone' variable with the minimum RMSE of 0.105 and 0.247 on the applied training and validation sets correspondingly. Along with that, the ARIMA-NN manner predicted the 'SO2' variable with the lesser RMSE of 0.363 and 0.732 on the applied training and validation sets correspondingly. Moreover, the ARIMA-NN method predicted the 'NO2' variable with the least RMSE of 0.134 and 0.300 on the applied training and validation sets respectively. At last, the ARIMA-NN methodology forecast the 'PM2.5' variable with the minimal RMSE of 0.170 and 0.490 on the applied training and validation sets correspondingly.

Table 2: Forecasting results of optimal ARIMA-NN model on applied dataset

Variables	MAE	MSE	RMSE
CO			
Training set	0.118	0.029	0.170
Validation set	0.297	0.240	0.490
Ozone			
Training set	0.065	0.011	0.105
Validation set	0.184	0.061	0.247
SO2			
Training set	0.226	0.132	0.363
Validation set	0.446	0.536	0.732
NO2			
Training set	0.103	0.018	0.134
Validation set	0.225	0.090	0.300
PM 2.5			
Training set	0.194	0.073	0.270
Validation set	0.375	0.332	0.576

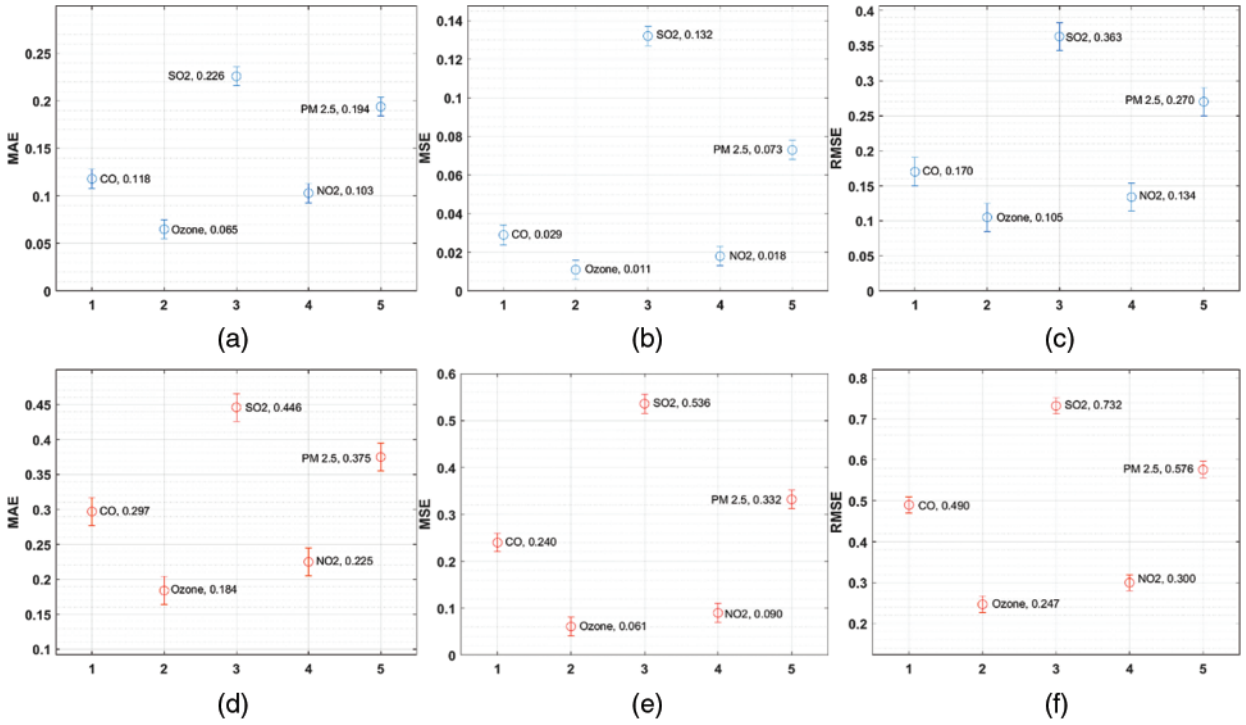

Figure 4: (a–c) Training set, (d–f) Validation set: Result analysis of ARIMA-NN model with different measures

Fig. 5 demonstrates the set of five confusion matrices produced by the OAI-AQPC technique under five distinct runs. Fig. 5a depicts the confusion matrix of the OAI-AQPC technique under the execution of run-1. The figure exhibited that the OAI-AQPC technique has classified the 13337 instances into Not Polluted class and 5983 instances into Polluted class. Simultaneously, Fig. 5b showcases the confusion matrix of the OAI-AQPC approach under the execution of run-2. The figure demonstrated that the OAI-AQPC manner has classified the 13537 instances into Not Polluted class and 5975 instances into Polluted class. Concurrently, Fig. 5c illustrates the confusion matrix of the OAI-AQPC manner under the execution of run-3. The figure outperformed that the OAI-AQPC technique has classified the 13458 instances into Not Polluted class and 6081 instances into Polluted class. In the meantime, Fig. 5d portrays the confusion matrix of the OAI-AQPC methodology under the execution of run-4. The figure demonstrated that the OAI-AQPC method has classified the 13698 instances into Not Polluted class and 6073 instances into Polluted class. Lastly, Fig. 5e illustrates the confusion matrix of the OAI-AQPC method under the execution of run-5. The figure showcased that the OAI-AQPC manner has classified the 13755 instances into Not Polluted class and 6103 instances into Polluted class.

Tab. 3 and Fig. 6 offer the classification results analysis of the OAI-AQPC technique interms of different measures. From the table, it is demonstrated that the OAI-AQPC technique has gained effective classification outcomes on the applied dataset. For instance, under fold-1, the OAI-AQPC technique has obtained a precision of 0.982, recall of 0.963, accuracy of 0.962, and F-score of 0.972.

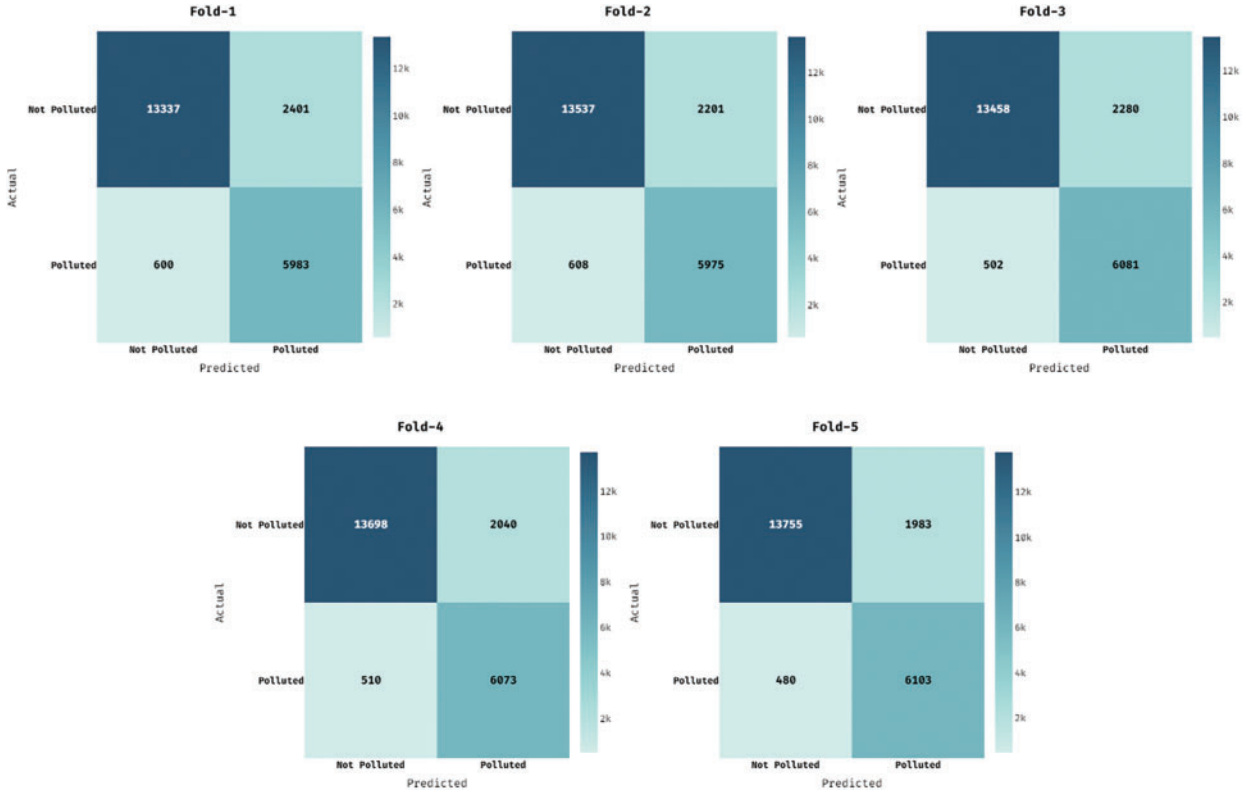

Figure 5: Confusion matrices of OAI-AQPC method

Table 3: Results analysis of proposed OAI-AQPC model in terms of various measures

No. of folds	Precision	Recall	Accuracy	F-score
Fold-1	0.982	0.963	0.962	0.972
Fold-2	0.983	0.969	0.966	0.976
Fold-3	0.983	0.964	0.963	0.974
Fold-4	0.982	0.963	0.961	0.972
Fold-5	0.983	0.961	0.961	0.972
Average	0.982	0.964	0.962	0.973

Also, under fold-2, the OAI-AQPC approach has gained a precision of 0.983, recall of 0.969, accuracy of 0.966, and F-score of 0.976. Additionally, under fold-3, the OAI-AQPC attained has achieved a precision of 0.983, recall of 0.964, accuracy of 0.963, and F-score of 0.974. Moreover, under fold-4, the OAI-AQPC methodology has gained a precision of 0.982, recall of 0.963, accuracy of 0.961, and F-score of 0.972. Furthermore, under fold-5, the OAI-AQPC methodology has gained a precision of 0.983, recall of 0.961, accuracy of 0.961, and F-score of 0.972.

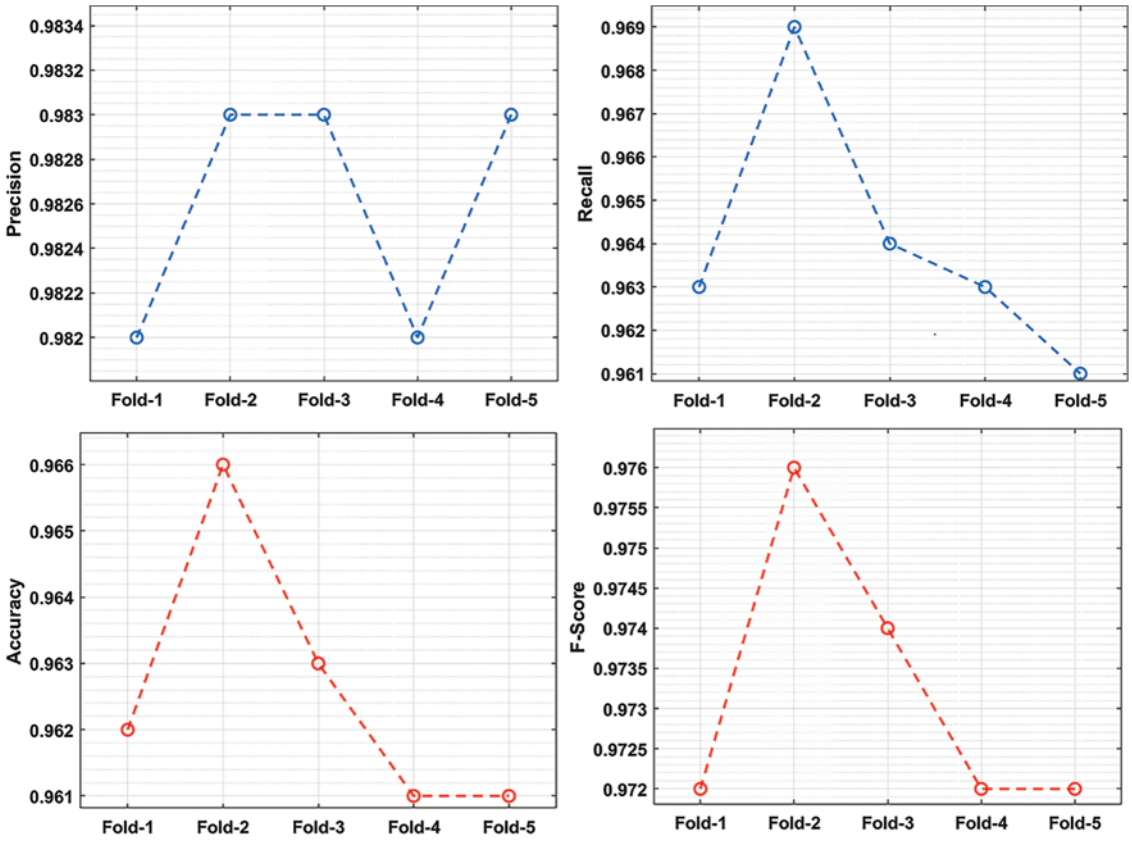

Figure 6: Classification results analysis of the OQI-AQPC technique

In order to guarantee the supremacy of the OAI-AQPC technique, a detailed comparison study is made in Tab. 4 [17]. The obtained results exemplified that the DT approach has showcased poor outcomes with the least accuracy of 0.874 whereas the SVM model has gained somewhat enhanced performance with an accuracy of 0.917. In addition, the ANN and PCA SVR-RBF techniques have showcased moderately closer outcomes with the accuracy of 0.945 and 0.952. Besides, the SVR-RBF technique has resulted in a near optimal outcome by offering high accuracy of 0.960. However, the proposed OAI-AQPC technique has depicted better performance over the other algorithms with maximum precision of 0.982, recall of 0.964, accuracy of 0.962, and F-score of 0.973.

Table 4: Results analysis of existing with proposed OAI-AQPC model in terms of various measures

Methods	Precision	Recall	Accuracy	F-score
PCA SVR-RBF	0.607	0.608	0.952	0.607
SVR-RBF	0.618	0.620	0.960	0.619
Decision tree	0.908	0.694	0.874	0.586
SVM	0.924	0.725	0.917	0.745
ANN	0.945	0.763	0.945	0.788
OAI-AQPC	0.982	0.964	0.962	0.973

From the above mentioned tables and figures, it can be stated that the OAI-AQPC technique has been found to be an appropriate air quality prediction and classification in the real time environment.

5 Conclusion

This paper has introduced an effective OAI-AQPC technique to predict and classify air quality on big data platforms. The OAI-AQPC technique encompasses different processes namely ARIMA-NN based prediction, OSSO based parameter optimization, and ANFIS based classification. The use of ARIMA-NN technique is utilized for predicting the air quality parameters and the ANFIS approach is employed to categorize the air quality into pollutants and non-pollutants. Besides, the application of OSSO algorithm to optimally adjust the parameters of the ARIMA-NN technique also results in improved air quality predictive outcomes. A wide range of simulation analyses is carried out and the results are inspected interms of different dimensions. The simulation values showcased the better performance of the OAI-AQPC technique over the other state of art predictive models interms of different evaluation parameters. In future, the performance of the OAI-AQPC technique is extended to the design of deep learning architectures for prediction and classification purposes.

Funding Statement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under grant number (RGP2/45/43). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R135), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4270206DSR02).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Castelli, F. M. Clemente, A. Popovič, S. Silva and L. Vanneschi, "A machine learning approach to predict air quality in California," *Complexity*, vol. 2020, no. 332, pp. 1–23, 2020.

[2] L. Pimpin, L. Retat, D. Fecht, L. D. Preux, F. Sassi *et al.,* "Estimating the costs of air pollution to the National Health Service and social care: An assessment and forecast up to 2035," *PLOS Medicine*, vol. 15, no. 7, pp. e1002602, 2018.

[3] E. Zdravevski, P. Lameski, C. Apanowicz and D. Ślęzak, "From big data to business analytics: The case study of churn prediction," *Applied Soft Computing*, vol. 90, no. 4, pp. 106164, 2020.

[4] J. Kalajdjieski, M. Korunoski, B. R. Stojkoska and K. Trivodaliev, "Smart city air pollution monitoring and prediction: A case study of Skopje," in *Proc. of the Int. Conf. on ICT Innovations*, Skopje, North Macedonia, pp. 15–27, 2020.

[5] J. Fan, Q. Li, J. Hou, X. Feng, H. Karimian *et al.,* "A spatiotemporal prediction framework for air pollution based on deep RNN," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-4/W2, pp. 15–22, 2017.

[6] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approaches for outdoor air quality modelling: A systematic review," *Applied Sciences*, vol. 8, no. 12, pp. 2570, 2018.

[7] M. Ritter, M. D. Müller, M. Y. Tsai and E. Parlow, "Air pollution modeling over very complex terrain: An evaluation of WRF-Chem over Switzerland for two 1-year periods," *Atmospheric Research*, vol. 132-133, no. D04314, pp. 209–222, 2013.

[8] Y. Huang, Q. Zhao, Q. Zhou and W. Jiang, "Air quality forecast monitoring and its impact on brain health based on big data and the Internet of Things," *IEEE Access*, vol. 6, pp. 78678–78688, 2018.

[9] J. Kalajdjieski, E. Zdravevski, R. Corizzo, P. Lameski, S. Kalajdziski *et al.,* "Air pollution prediction with multi-modal data and deep neural networks," *Remote Sensing*, vol. 12, no. 24, pp. 4142, 2020.
[10] Z. Ghaemi, A. Alimohammadi and M. Farnaghi, "LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran," *Environmental Monitoring and Assessment*, vol. 190, no. 5, pp. 300, 2018.
[11] K. Zhang, X. Zhang, H. Song, H. Pan and B. Wang, "Air quality prediction model based on spatiotemporal data analysis and metalearning," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–11, 2021.
[12] A. R. Honarvar and A. Sami, "Towards sustainable smart city by particulate matter prediction using urban big data, excluding expensive air pollution infrastructures," *Big Data Research*, vol. 17, no. 1, pp. 56–65, 2019.
[13] T. Zaree and A. R. Honarvar, "Improvement of air pollution prediction in a smart city and its correlation with weather conditions using metrological big data," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 3, pp. 1302–1313, 2018.
[14] M. Shahbaz, C. Gao, L. Zhai, F. Shahzad and I. Khan, "Environmental air pollution management system: Predicting user adoption behavior of big data analytics," *Technology in Society*, vol. 64, pp. 101473, 2021.
[15] S. A. Janabi, M. Mohammad and A. Al-Sultan, "A new method for prediction of air pollution based on intelligent computation," *Soft Computing*, vol. 24, no. 1, pp. 661–680, 2020.
[16] D. H. Shih, T. H. To, L. S. P. Nguyen, T. W. Wu and W. T. You, "Design of a spark big data framework for PM2.5 air pollution forecasting," *International Journal of Environmental Research and Public Health*, vol. 18, no. 13, pp. 7087, 2021.
[17] M. Castelli, F. M. Clemente, A. Popovič, S. Silva and L. Vanneschi, "A machine learning approach to predict air quality in California," *Complexity*, vol. 2020, no. 332, pp. 1–23, 2020.
[18] Y. S. Chang, K. M. Lin, Y. T. Tsai, Y. R. Zeng and C. X. Hung, "Big data platform for air quality analysis and prediction," in *2018 27th Wireless and Optical Communication Conf. (WOCC)*, Hualien, Taiwan, pp. 1–3, 2018.
[19] Z. Zou, T. Cai and K. Cao, "An urban big data-based air quality index prediction: A case study of routes planning for outdoor activities in Beijing," *Environment and Planning B: Urban Analytics and City Science*, vol. 47, no. 6, pp. 948–963, 2020.
[20] S. R. Gallego, A. Fernández, S. García, M. Chen and F. Herrera, "Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce," *Information Fusion*, vol. 42, no. 4, pp. 51–61, 2018.
[21] M. Alsharif, M. Younes and J. Kim, "Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea," *Symmetry*, vol. 11, no. 2, pp. 240, 2019.
[22] L. Yu, S. Wang and K. K. Lai, "Foreign-exchange-rate forecasting with artificial neural networks," in: *International Series in Operations Research & Management Science Book Series*, vol. 107, Boston, MA: Springer US, 2007.
[23] D.Ö. Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 4, pp. 586–594, 2010.
[24] M. Neshat and G. Sepidname, "A new hybrid optimization method inspired from swarm intelligence: Fuzzy adaptive swallow swarm optimization algorithm (FASSO)," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 339–350, 2015.
[25] X. Ma, F. Liu, Y. Qi, M. Gong, M. Yin *et al.,* "MOEA/D with opposition-based learning for multiobjective optimization problem," *Neurocomputing*, vol. 146, no. 6, pp. 48–64, 2014.
[26] H. Alawad, M. An and S. Kaewunruen, "Utilizing an adaptive neuro-fuzzy inference system (ANFIS) for overcrowding level risk assessment in railway stations," *Applied Sciences*, vol. 10, no. 15, pp. 5156, 2020.