Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1969 - 1994
Received 16 February 2021; Accepted 08 March 2021.

An Efficient Feature Selection with Weighted Extreme Learning Machine for Water Quality Prediction and Classification Model

J. Charles[1], G. Vinodhini[2], R. Nagarajan[3]

[1]Research scholar, Department of Computer and Information sciences, Annamalai University,Annamalai Nagar, India.

[2]Assistant Professor, Department of Information Technology, Annamalai University, Annamalai Nagar, India.

[3]Assistant Professor, Department of Computer and Information sciences, Annamalai University, Annamalai Nagar, India.

E-mail: jcharles.1404@gmail.com, vinodhini.g.t@gmail.com, rathinanagarajan@gmail.com

Abstract

Advanced urbanization and industrialization have resulted to a worsening of water quality significantly, and led to severe diseases. Water quality index (WQI) is a commonly employed measure to determine water quality, which is a costlier and lengthy process. The increasing effect of poor water quality leads to the requirement of latest machine learning (ML) models for automated water quality prediction. This paper aims to present a new feature selection with classification model for the proficient prediction of water quality in real time scenarios. The presented model involves three processes, such as preprocessing, feature selection, and classification. Primarily, the dataset is collected and preprocessing takes place to transform the actual dataset into a compatible format for classification process. The proposed method uses a quantum teaching and learning based optimization (TLBO) algorithm to select an optimal set of features, and thereby reduces the complexity level. Besides, the presented QTLBO-WELM model also uses a weighted extreme learning machine (WELM) model for classification process. In order to assess the predictive outcome of the presented model, a series of experiments were conducted on a collection of 35 groundwater samples from Dharmapuri district in Tamil Nadu. The experimental values showcased the betterment of the presented QTLBO-WELM model with the sensitivity of 96.29%, specificity of 94.10%, accuracy of 95.71%, precision of 98.11%, F-score of 97.17%, and kappa value of 88.82%.

Keywords: Water quality prediction, GIS, Machine learning, Sampling methods, imbalanced data

1. Introduction

Basically, water is one of the most significant sources needed for human survival; also it is polluted by massive pollutants. The progressive industrialization has resulted to the limitation of water quality. Inferior water quality is the major factor of escalation in harrowing infections. Based on the survey, in growing countries, numerous diseases are evolved from water where enormous disease and mortality have occurred [1]. In many decades, water pollution or contamination is one of the major problems. Green globalisation with no adequate cleanliness of water is not possible. Hence, the provided water quality has to be stored and observed frequently to ensure the scalable production of drinking water. Recently, pollution intensity results in enhanced pollutants, especially in sea waters at various rates which depend upon the Arctic aquatic ecology. Followed by, the issue of remote monitoring models used for predicting and classification of irregular phenomena on soil or water is a drastic study. Water is one of the natural factors used to enhance environmental growth as well as human activities.

In 2011, World Economic Forum (WEF) found the interconnected issues of water, energy, and food as threatening global risk and pointed that management of interrelated models with no consideration of global threat of severe issues. Thus, improper water management and natural resources has resulted to incline the civilizations. It is one of the major problems, especially in growing countries like Pakistan. Water Quality (WQ) is evaluated by cost-effective as well as time-consuming lab and statistical examination, that requires sample gathering, transport to labs, and reasonable count of time and estimation that is insignificant to communicable media and time is one of the essential objective polluted with disease-making waste. The terrible consequences of water pollution require robust and cheap alternative. Hence, the central premises of this study is to present and estimate the alternate model on supervised Machine Learning (ML) for effective detection of WQ in practical scenarios.

In case of evaluating WQ interms of ML, Shafi et al. [2] evaluated WQ under the application of ML models like Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (DNN), and k Nearest Neighbors (kNN), where DNN has gained better accuracy. Hence, evaluated WQ depends upon the 3 variables Turbidity, Temperature, and pH that are sampled on the basis of WHO. Under the application of 3 parameters and relate them with standard values are one of the major drawbacks while predicting WQ. Ahmad et al. [3]

applied feedforward neural networks (FFNN) and unification of various NN for the purpose of estimating WQI. With the help of backward removal and forward selection selective combination models, this has attained maximum R2 and MSE, correspondingly. The application of these parameters makes the solution immoderate by means of cheaper real-time model.

Sakizadeh [4] detected the WQI under the application of WQ variables and artificial neural network (ANN) with Bayesian regularization. As a result, better correlation coefficients have been accomplished. Abyaneh [5] forecasted the Chemical Oxygen Demand (COD) as well as Biochemical Oxygen Demand (BOD) with the help of 2 classical ML approaches such as ANN and multivariate linear regression. Around 4 variables are applied such as pH, temperature, total suspended solids (TSS), and total suspended (TS) for predicting COD and BOD. Ali and Qamar [6] employed unsupervised model of average linkage of hierarchical clustering for the purpose of classifying samples into water quality categories. Therefore, the main variable related to WQI has been predicted for learning and it does not apply any benchmark WQI for estimation of these predictions. Gazzaz et al. [7] employed ANN for detecting WQI with better variations. Here, massive number of variables has been utilized for predicting WQI, which is costlier when it is applied in IoT network. Rankovic et al. [8] found the Dissolved Oxygen (DO) with the help of feedforward neural network (FNN). Here, minimum parameters were applied for estimating DO, which further degrades the realistic WQI evaluation using Internet of Things (IoT) system.

Kut et al. [9] assessed the groundwater samples for matching the features of drinking water. Mitrovic´ et al. [10]forecasted WQ of Danube River (Serbia) by applying Monte Carlo optimized ANN. Here, 18 typical water quality parameters (WQPs) have been applied for inactive monitoring channels. Fijani et al. [11] deployed a system for practical observation of 2 WQP namely, chlorophyll-a (Chl-a) and DO. The development of 2-layer decomposition by applying CEEMDAN and VMD methodologies with LSSVM and ELM approaches. Wan et al. [12] managed to deploy 4-level pollution index on WQ of sea water and classification termed as WQ Classification Index (WQCI). Multi-Criteria Decision Making Models (MCDM) have been applied by Yousefi et al. [13] for estimating the quality of drinking water. The major objective of present study is to find a result for reducing possible errors emerged by using WQI model in classifying WQ classes.

Nnorom et al. [14] predicted the aqua physicochemical and toxic units of ground as well as surface water sources applied for domestic applications in Nigeria. In this approach, around 124 water samples have been acquired from natural springs, streams, boreholes, and hand-dug wells from both rural and urban regions. A brief work on WQ of Gomti River was performed by Kumar [15]. Also, the present works have shown that various scenarios were assumed in global changes like climate modification and population development. Roth et al. [16] examined the impacts of environmental change on water resources from transnational Blue Nile Basin (BNB) under the application of water details from predefined works. Awotwi et al. [17]implemented that cassava WQ units are more significant to estimate the quality of water. Hence, the former studies have carried out with various parameters and projected a classifier for frequent datasets which suits WQ observation significantly.Deep learning (DL) related models have accomplished challenging performance in massiveapplications [26-30].

This paper develops a novel feature selection with classification model for the proficient prediction of water quality in real-time scenarios. Initially, the dataset is gathered and preprocessing is carried out to convert the actual dataset into a compatible format for classification process. The proposed method involves quantum teaching and learning based optimization (QTLBO) algorithm for feature selection process. Additionally, the presented QTLBO-WELM model also uses a weighted extreme learning machine (WELM) model for classification process. To validate the predictive outcome of the presented model, a series of experiments were conducted on a collection of 35 groundwater samples from Dharmapuri district in Tamil Nadu.

2. The Proposed WTLBO-WELM Model

The workflow involved in the WTLBO-WELM method is depicted in Fig. 1. The figure states that the input data is preprocessed and then WTLBOalgorithm is employed as a feature selector. Afterward, the diminished feature subset is fed into the WELM model for classification purposes.

2.1. Data Collection

In case of data collection state, a blend of 35 samples is collected and examined in pre-monsoon and post-monsoon (November 2015) seasons for major ions from Palacode and Pennagaramtaluks. In this approach, a random sampling model has been employed for clustered water samples irrespective of land-based styles, geomorphology, and study area

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1969 - 1994
Received 16 February 2021; Accepted 08 March 2021.

topography. The instances are gathered in clear 1-L size polyethylene container sets. By referencing to WHO (2011) drinking water demands, the groundwater supremacy for drinking was estimated by evaluating WQI measures for soil water instances.

2.2. QTLBO-FS Model

Once the data is collected and preprocessed, the QTLBO-FS technique is employed to select an optimal set of features. Teaching-learning is defined as a significant process in which an individual manages for learning from alternate individuals in order to enhance themselves. The presented method named Teaching-Learning-Based Optimization (TLBO), which accelerates conventional TL concept of a classroom. The method triggers 2 basic modes of learning: (i) through a teacher (termed as teacher phase) and (ii) communicating with alternate learners (named as a learner phase).

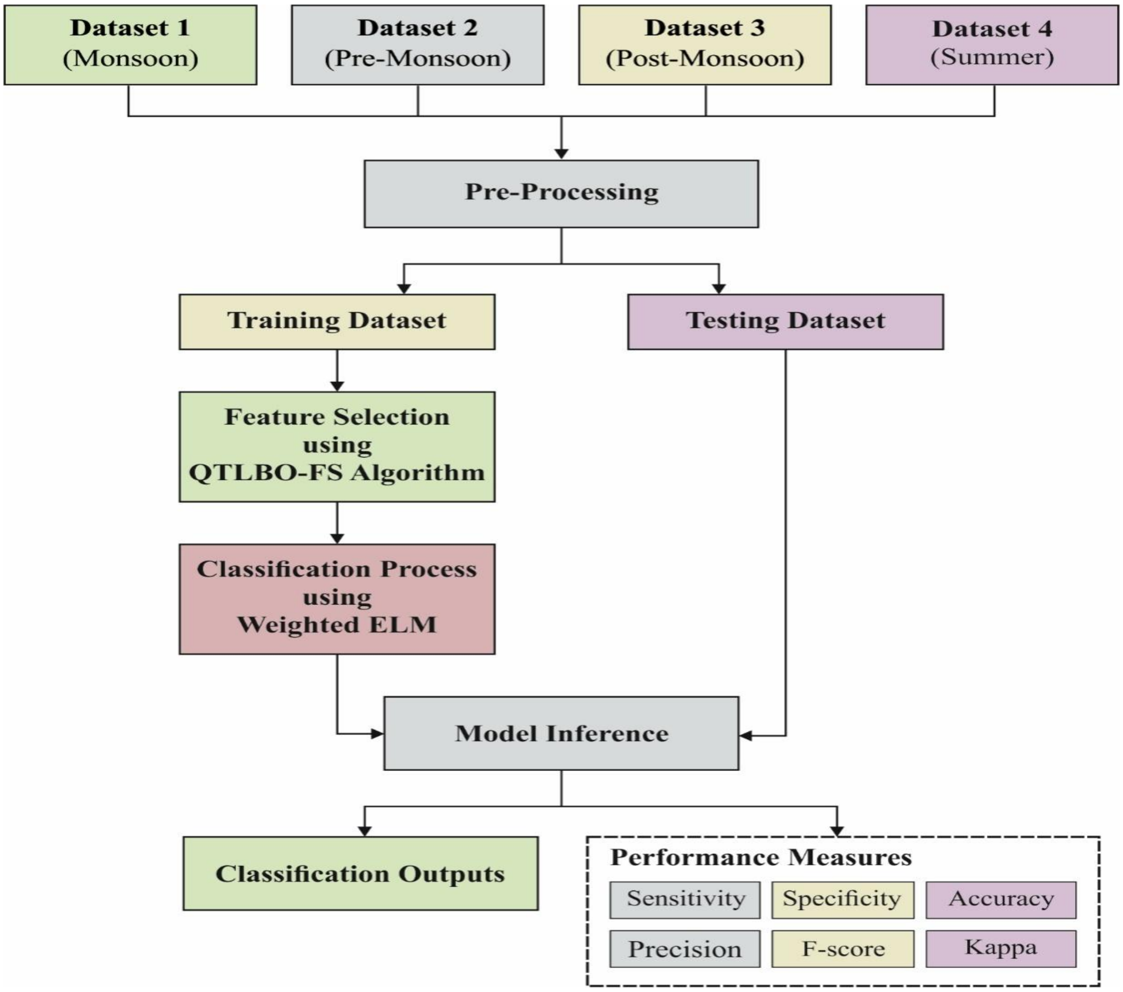

Fig. 1. Workflow of WTLBO-WELM Model

TLBO is defined as a population-related method, where set of students are assumed as population and various subjects provided for learners are analogous with diverse design parameters of optimization issues. Consequently, the learners are analogous to fitness value of optimization problems. An optimal solution from entire population is referred as a teacher.To improve the performance of the TLBO algorithm, quantum concept is introduced to it. Quantum processing depends upon the implication of data as quantum bits (Q-bit) and maximizes the benefits provided by superposition of states. The Q-bit is defined as a fundamental element of 2-state quantum computer [18]. The individual states, $|0\rangle$ and $|1\rangle$, the quantum state $|\Psi\rangle = a|0\rangle + b|1\rangle$ is defined as a linear superposition of unique states. The possible amplitudes, a and b are defined as complex weights of quantum particle within $|0\rangle$ and $|1\rangle$ correspondingly. In this approach, $|a|^2$ and $|b|^2$ gives the possibility of Q-bit to be in state $|0\rangle$ and $|1\rangle$correspondingly where:

$$|a|^2 + |b|^2 = 1 \tag{1}$$

When a demonstration is comprised of d states, afterward it can be referred to be the 2^d states simultaneously, with applied probability where the sum is 1. The d-length Q- bit individual is confused with a d-length binary vector under the application:

$$if\ |a_i|^2 < thresholdtheny_i \leftarrow 1\ elsey_i \leftarrow 0 \tag{2}$$

where y_i implies the parallel binary bit of observed Q-bit. A threshold sore is produced in random fashion. Collections of binary vectors are attained for gaining better fitness.

Variation metrics like selection, mutation, crossover, and quantum gates have been employed for upgrading the location of Q-bits in consecutive productions. Selection metrics such as Best, Random, Roulette, and Tournament have been applied for selecting these individuals that contribute in production of consecutive populations. The mutation operator is composed of changing the possibilities of number of Q-bits of individuals, as defined are the following. For instance, for a quantum individual P, mutated individual, P_{mut} at mutation point is implied as follows:

$$P = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & \cdots & a_d \\ b_1 & b_2 & b_3 & b_4 & b_5 & \cdots & b_d \end{pmatrix} P_{mut} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & \cdots & a_d \\ b_1 & b_2 & b_3 & b_4 & b_5 & \cdots & b_d \end{pmatrix} \tag{3}$$

The main aim of this work is a modification of QTLBOfor exploring the solution space of features to select the better subset of attributes. Hence, 2 phases are followed till reaching a termination condition. Generally, the termination condition is fixed as count of iterations. The optimal learners are referred as maximum solutions in specific iteration and computes with massive generations [19]. The sequential process of QTLBO-FSapproach is defined in the following:Initially, learners are trained using a teacher. For all iterations$k$, when ‘s’ number of features, $\{f = 1,2 \ldots. s\}$), ‘t’ count of samples (population, individuals, $\{i = 1,2 \ldots t\}$).Systematic explanation is defined in the following,

Algorithm 1:QTLBO-FS

Step 1: Initiatethe count of samples (binary population), count of features as $X_{f,i,k}$as well as a stopping criterion.

Step 2: Estimate the mean of every featurein case of learners as $M_{f,k}$.

Step 3: Identify the fitness of individuals.

$$Fitness(X_{f,i,k}) = Accuracy(X_{f,i,k})$$

Step 4: Upgrade learners using a Teacher (*Teacher Phase)*

a) Select the best learner (High fitness value) from the population as a Teacher.

b) Calculate the variance mean for these parameters interms of better individual as illustrated.

$$Diff_M ean_{f,i,k} = r_k(X_{f,ibest,k} - T_F M_{f,k})$$

where, $X_{f,ibest,k}$means the optimal individual with respect to f.X8 is a teaching factor using 2 and $r_k$implies a random value ranged from 0 to 1.

c) An optimal user is served as a teacher and trains the residual individuals. Upgrade a learner in population.

$$X'_{f,i,k} = 0\ if X_{f,i,k} + Diff_{Mean} f,k < 0.5$$

$$X'_{f,i,k} = 1\ if X_{f,i,k} + Diff_{Mean} f,k \geq 0.5$$

where, $X'_{f,i,k}$denotes a trained score of $X_{f,i,k}$.

d) When a result $X'_{f,i,k}$ is optimal than $X_{f,i,k}$, Repeat the advanced values; Else,

Replace a novel value.

Step5: Upgrade a learner using adjacent learners under the application of (*LearnerPhase*)

a) Select 2 samples U and V along with a condition $X'_{total\,-U,k} \neq X'_{total\,-V,k}$randomly.

where, $X'_{total\,-U,k}$, $X'_{total\,-V,k}$denotes the reformed parameters of $X_{total\,-U,k}$, $X_{total\,-V,k}$of U and V correspondingly.

b) When $X'_{total\,-U,k}$ is optimal than $X'_{total\,-V,k}$

$$X''_{f,U,k} = 0\ if X'_{f,U,k} + r_k(X'_{f,U,k} - X'_{f,V,k}) < \ 0.5$$

$$X''_{f,U,k} = 1\ if X'_{f,U,k} + r_k(X'_{f,U,k} - X'_{f,V,k}) \geq \ 0.5$$

Else,

$$X''_{f,U,k} = 0\ if X'_{f,U,k} + r_k(X'_{f,V,k} - X'_{f,U,k}) < 0.5$$

$$X''_{f,U,k} = 1 if X'_{f,U,k} + r_k(X'_{f,V,k} - X'_{f,U,k}) \geq 0.5$$

c) When $X''_{f,U,k}$ is optimal than $X'_{f,U,k}$

Then, follow the advanced value

Otherwise,

Replace the prior value.

Step 6: When the termination criteria is filled,

Then, generate the simulation outcome

Else, return for Step 2

For extended QTLBO-FSapproach, a population is represented with binary bits 1 or 0which depicts the existence or inexistence of a special attribute for a user. A length of binary string is similar to count of features with actual dataset. Then, in case of teacher phase mean value of a feature denotes the probability of certain features in a population. The variations mean refers the inclination of learners from teacher with respect to features application in a solution. Hence, the inclusion of variations refers a learner bit value which in turn decides the accessibility of particular features in consecutive population. In case of learner phase, a student with better knowledge (fitness value) would influenceneighboringstudents with specified conditions.

The major responsibility of QTLBO-FSmodel is to compute the Feature Selection (FS) for the purpose of disease analysis. It is evolved from student education process in a classroom. Initially, teacher grades the students with their knowledge, experiences, and academic performance, where the students are further trained on the basis of above-defined factors. The newly developed approach is relied on random searching model with2 phases namely, Teaching Phase and Learning Phase. Initially, the primary individuals are served as a teacher and train residual individuals by assuming the learners for enhancing the knowledge related on personal experience. Any individual with maximum accuracy or error rate would be termed as a teacher. Prediction methods have been applied for estimating the cost of a population in wrapper-related FS technologies. Classification accuracy is facilitated as a fitness value for maximization issues and classification error rates for minimization issues. A teacher helps the learners for developing significant solutions to optimal one using the difference mean. The different classification models $lil < e,$NB, SVM, $1 <$-Nearest Neighbors ($1 < -$ NN), Decision tree (DT), and DA for identifying classification errors and accuracies which is served as fitness values under the evaluation. The teaching aspect might be elected randomly among 1 and 2. Classification accuracy (CA) has been described as illustrated in Eq. (4).

$$CA = CorrectlyClassifiedInstances/TotalInstances \qquad (4)$$

Secondly, learner's skills might be enhanced by communicating between themselves. A learner in a solution space upgrades the solutions using learners based population. In general, a population is composed of essential individuals with optimal fitness measures and collection of features after various iterations of teacher as well as learner phases. A dataset with novel users might be applied for training classification methods in order to gain effective performance. Eventually, this model is sampled using various population sizes and optimal measures with massive number of individuals in a solution space for optimal convergence. Fig. 2 illustrates the flowchart of TLBO model.

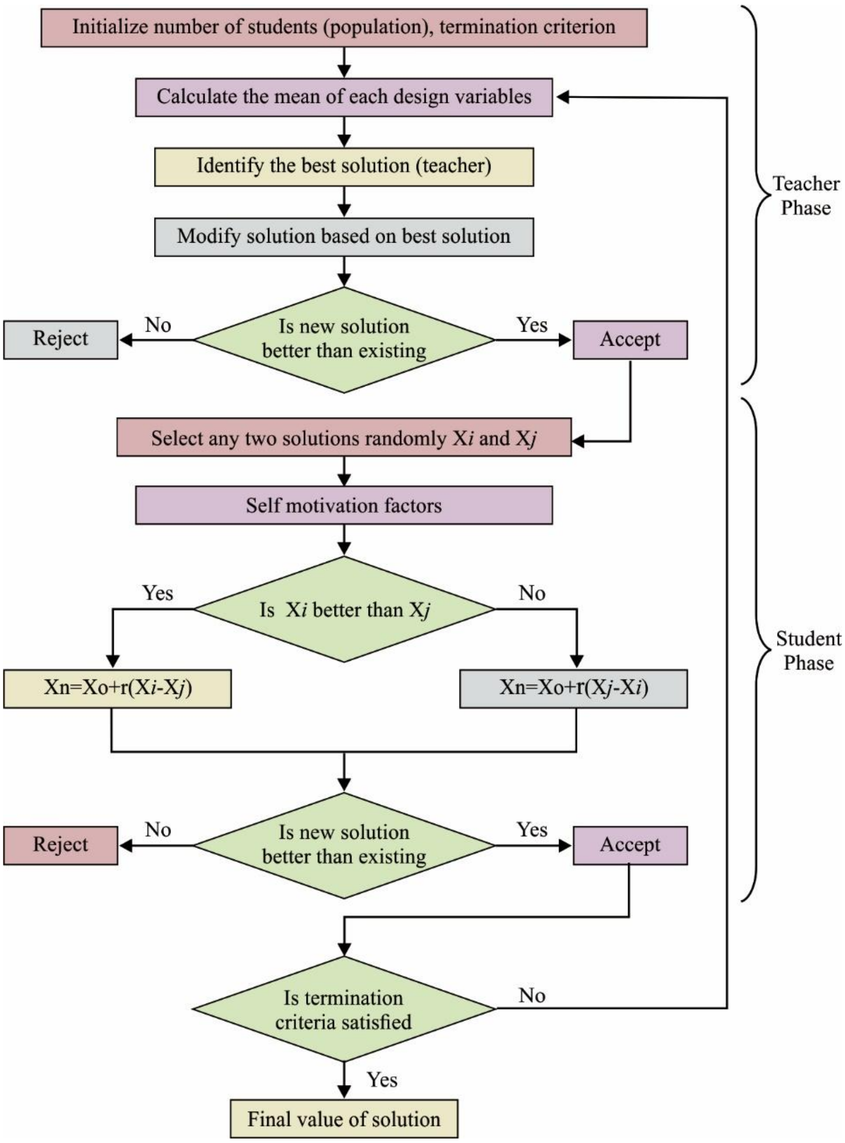
Initialize number of students (population), termination criterion
Calculate the mean of each design variables
Identify the best solution (teacher)
Modify solution based on best solution
Is new solution better than existing
No
Yes
Reject
Accept
Teacher Phase
Select any two solutions randomly Xi and Xj
Self motivation factors
Is Xi better than Xj
Yes
No
Xn=Xo+r(Xi-Xj)
Xn=Xo+r(Xj-Xi)
Student Phase
Is new solution better than existing
No
Yes
Reject
Accept
Is termination criteria satisfied
No
Yes
Final value of solution

Fig. 2. Flowchart of TLBO model

2.3. WELM Model for Classification

Once the feature reduced subset is obtained, the WELM model is employed to classify the water samples into ‘purity’ and ‘impurity’ classes. ELM has been developed for single-layer feed-forward networks (SLFNs), in which input weights of SLFN are produced in random fashion and final weights undergo training with batch learning method of least squares. It has ensured that SLFNs with arbitrarily hidden neurons as well as tunable final weights have global approximation and tremendous generalization function. Specifically, ELM surpasses the traditional learning models in training speed which has been employed extensively in face analysis, image computation as well as classification, electricity price categorization, energy commodity futures index detection, position fingerprinting model, protein sequence classification, and location categorization. The fundamental concept of unweighted ELM is defined in the following:

For input data $x \in R^n$, the simulation result of remarkable SLFN with L hidden nodes are reformed as,

$$H(x) = \sum_{\mathrm{i=1}}^{\mathrm{L}} \beta_i \, G(w_i, b_i, x), w_i \in R^n, \beta_i \in R^m,$$

where $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$ implies a weight vector linked ith hidden node to final nodes, and $G(w_i, b_i, x)$ defines the final result of ith hidden node.

Assume a collection of training pairs with N inputs $x_1, x_2, \dots, x_N$ (here $x_i \in R^n$), and N desired results $t_1, t_2, \dots, t_N$ (here $t_i \in R^m$), correspondingly. The numerical approach of an SLFN is

$$H\beta = T \tag{5}$$

with

$$H = \begin{bmatrix} G(w_1, b_1, x_1) & \cdots & G(w_L, b_L, x_1) \\ \vdots & \ddots & \vdots \\ G(w_1, b_1, x_N) & \cdots & G(w_L, b_L, x_N) \end{bmatrix} \tag{6}$$

In line with this, least squares support vector machine (LS-SVM), the final results are devised by identifying better solution in case of optimization issues:

$$Minimize \frac{1}{2} \|\beta\|^2 + \frac{C}{2} \|\varepsilon_i\|^2, \tag{7}$$

$$subjectto \varepsilon_i = H(x_i)\beta - t_i, i = 1,2, \dots, N, \tag{8}$$

where ε_i implies a training error vector of m output nodes by means of training instances and C refers a positive real regularization attribute.By resolving the optimization issues (7), least-square solution might be estimated as,

$$whenN < L{:}\, \beta = \widetilde{H}^{+}T = H^T \left(\frac{1}{C} I + HH^T\right)^{-1} T,$$

$$whenN \geq L : \ \beta = \widetilde{H}^{+}T = \left(\frac{1}{C} I + H^T H\right)^{-1} H^T T. \tag{9}$$

Along with that, traditional learning models where the working principle of un-weighted ELM depends upon a class distribution. It surpasses well using regular datasets, however irregular classification is highly complicated. Certainly, negative class intends to move isolating boundaries to the positive class for gaining maximum CA. Weighted ELM model [20] was deployed for resolving these issues. Fig. 3 shows the structure of WELM.

WELM is a cost-sensitive learning approach and weighted ELM allocates diverse weights for every sample for minimizing misclassification of positive instances and related cost errors. Followed by, the key responsibility of weighted ELM is to enhance the marginal distances:

$$Minimize\frac{1}{2}\|\beta\|^2+\frac{C}{2}W\|\varepsilon_i\|^2, \tag{10}$$

$$subjectto\varepsilon_i=H(x_i)\beta-t_i, i=1,2,\dots,N, \tag{11}$$

where W implies a misclassification cost matrix interms of class distribution.

Similar to Eq. (9), the solutions of β can be attained:

$$whenN<L:\beta=H^T\left(\frac{1}{C}I+WHH^T\right)^{-1}WT,$$

$$whenN\geq L:\beta=\left(\frac{1}{C}I+H^TWH\right)^{-1}H^TWT. \tag{12}$$

In 2 cost matrices W_1 and W_2, have been presented.

$$WeightingschemeW_1:\ W_{i,i}=\frac{1}{\#(t_i)}, i=1,2,\dots,m, \tag{13}$$

where $\#(t_i)$ means the count of instances in class t_i. The irregular datasets accomplished a cardinal balance if a scheme W_1 has been applied.When the performance is compared, alternate weighting scheme

$$W_2:\begin{cases}W_{i,i}=\dfrac{1}{\#(t_i)}, if t_i>AVG\\ W_{i,i}=\dfrac{0.618}{\#(t_i)}, if t_i\leq AVG\end{cases}$$

where golden $ratio-O.618$—was applied and AVG shows the sample size. Basically, W_2 is defined as trade-off among unweighted ELM as well as weighted ELM W_1. Once the weighting scheme W_1 or W_2 is applied with unweighted ELM, weighted ELM limits a misclassification error correlated with minimum positive class. Followed by, weighting approach forces the boundary backward to a negative class where excess positive class samples are classified correctly have best weights as W_1 or W_2.

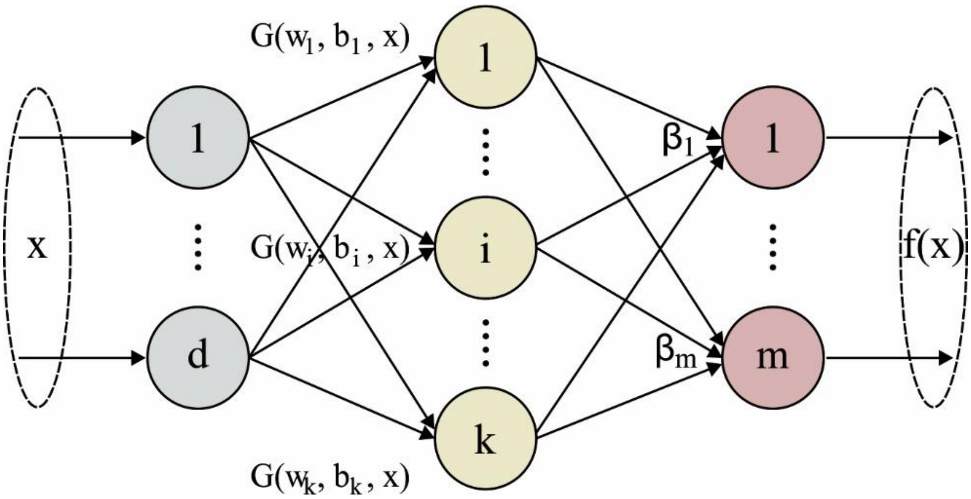

Fig. 3. Structure of WELM

3. Performance Validation

The presented method is accelerated with the help of Python tool and simulation outcomes are verified using 4 datasets. The working principle of the IF-ANFIS approach is sampled under the application of 4 datasets. The details relevant to a dataset are depicted in Table 1. Each dataset is composed of similar set of 15 parameters with 35 instances. The monsoon dataset contains 28 instances under ‘purity class’ and residual 7 samples are ‘impurity classes. Likewise, the count of ‘purity’ instances in post-monsoon, pre-monsoon, and summer dataset have total of 27, 25, and 27 instances. In line with this, the number of ‘impurity’ samples in the post-monsoon, pre-monsoon, and summer dataset is composed of a total of 8, 10, and 8 instances. The data relevant to the dataset is implied in Table 1 as well as related parameter information is provided in Table 2. Figs. 4-7depict the visualization outcome investigation of the variables present in the dataset.

Table 1 Dataset Description

No.	Dataset Name	Sources	No. of Attributes	No. of Instances	Purity/Impurity
1	Monsoon Dataset	Own	15	35	28/7
2	Post-Monsoon Dataset	Own	15	35	27/8
3	Pre-Monsoon Dataset	Own	15	35	25/10
4	Summer Dataset	Own	15	35	27/8

Table 2 Attributes in the Applied Dataset

Number	Attributes Description
1	Station ID
2	Longitude
3	Latitude
4	Ca-Si
5	Mg-Si
6	Na-Si
7	K-Si
8	Cl-Si
9	HCO3-Si
10	SO4-Si
11	PO4-Si
12	pH -Si
13	Ec-Si
14	TDS-Si
15	TH-SI
16	Target Class WQI [Purity/Impurity]

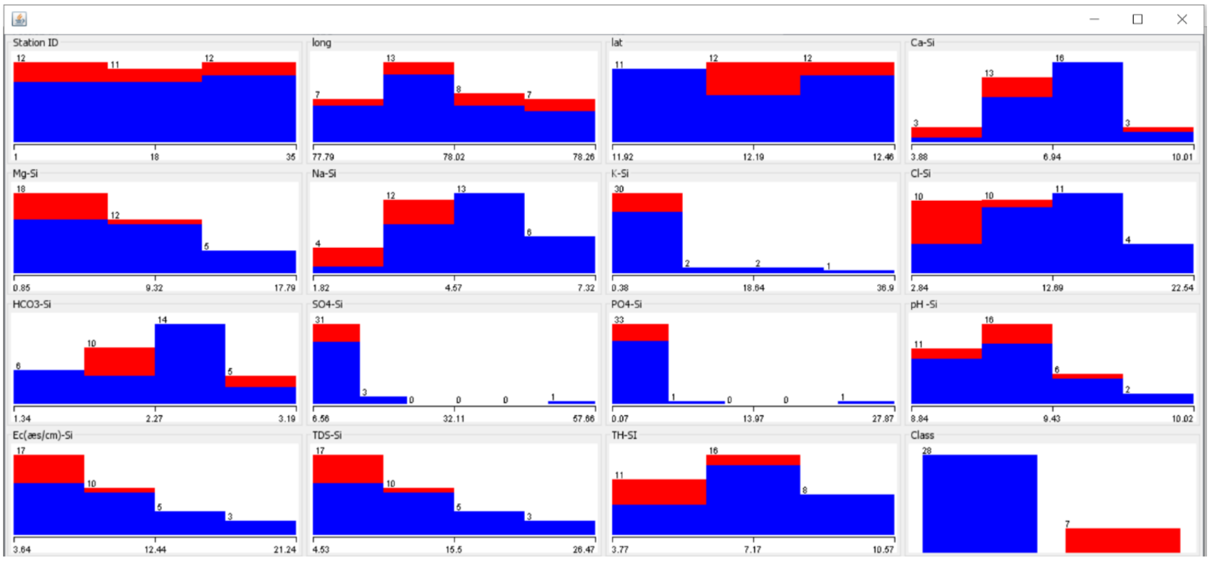

Fig. 4. Visualization of Attributes of Monsoon Dataset

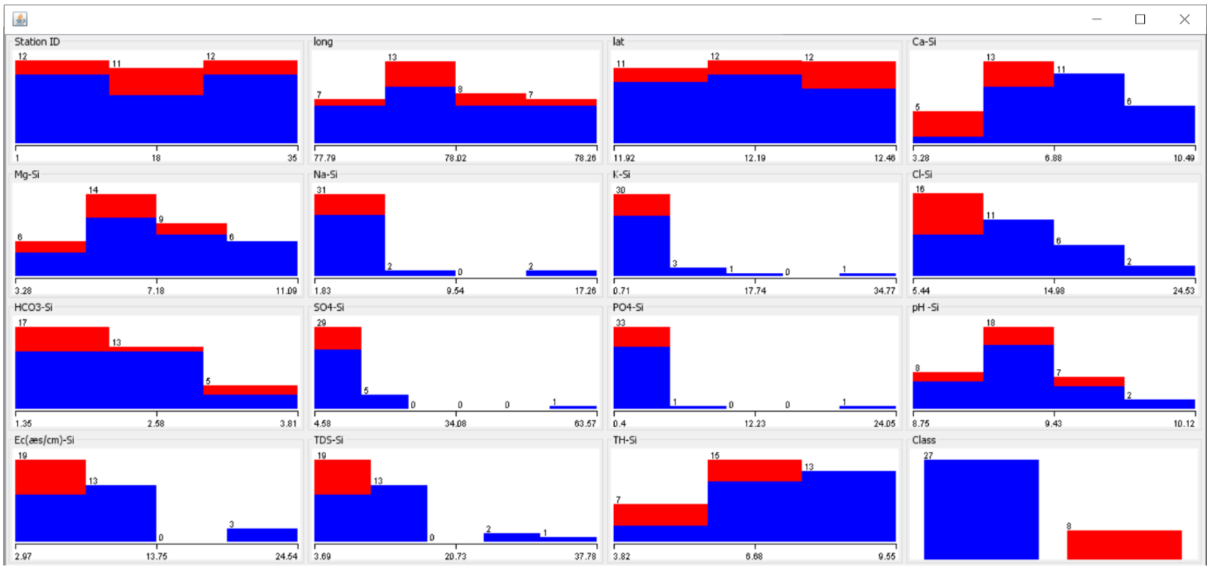

Fig. 5. Visualization of Attributes of Post-Monsoon Dataset

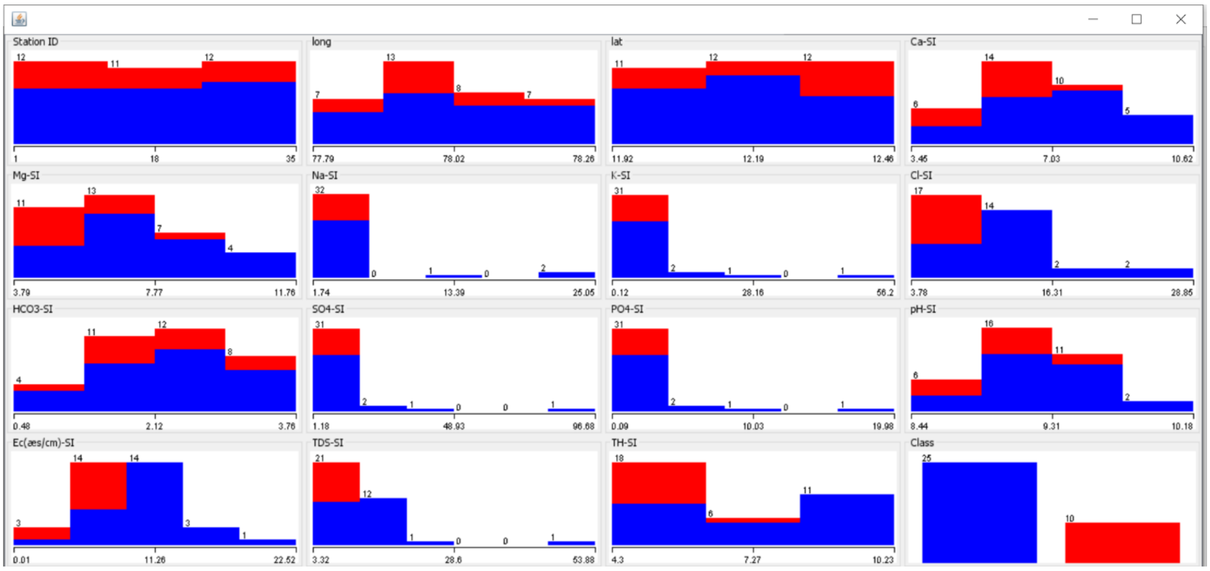

Fig. 6. Visualization of Attributes of Pre-Monsoon Dataset

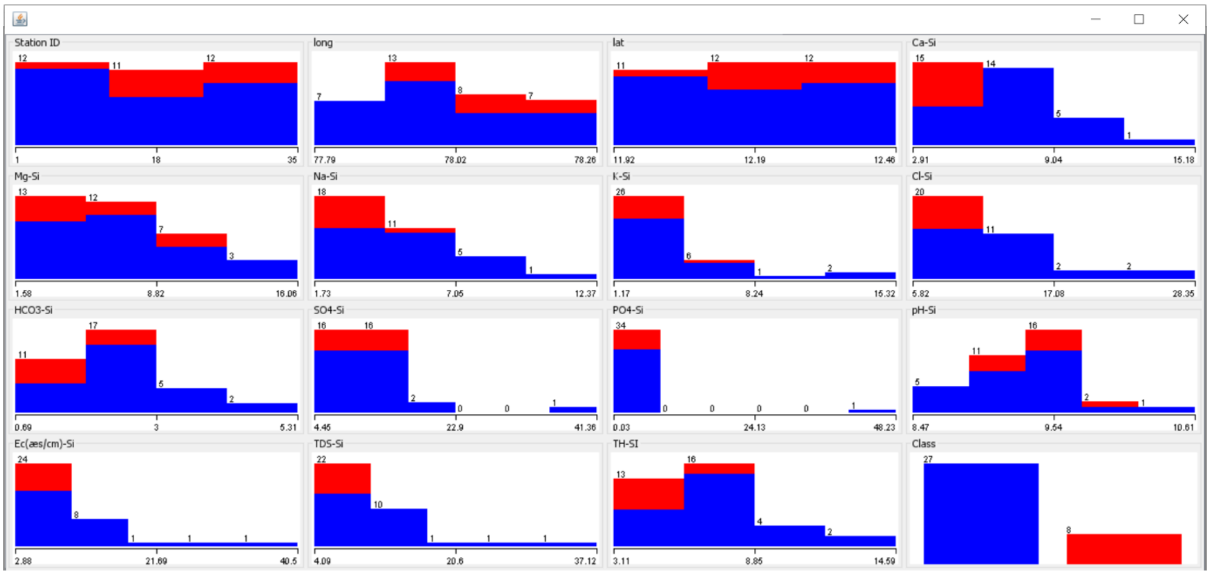

Fig. 7. Visualization of Attributes of Summer Dataset

Table 3 exhibits the summary of the FS process by the QTLBO-FS model on the applied dataset. The table values notified that the GA-FS technique has failed to reach effective results and resulted to a higher best cost of 0.1287. In addition, the PSO-FS algorithm has tried to perform well, but it also reached to a closer best cost of 0.1028. Though the GWO-FS algorithm has outperformed the previous FS models with the best cost of 0.0897, the presented QTLBO-FS model has reached to an optimal best cost of 0.0546.

Received 16 February 2021; Accepted 08 March 2021.

Table 3 Selected Features of Existing with Proposed QTLBO-FS Method

Methods	**Best Cost**	**Selected Features**
QTLBO-FS	0.0546	4, 6, 7, 8, 9, 12, 14, 15
GWO-FS	0.0897	2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14
PSO-FS	0.1028	1, 2, 4, 5, 6, 7, 9,11, 13, 14, 15
GA-FS	0.1287	1, 2, 6, 7, 8, 9, 11, 12, 14, 15

In this section, it examines the prediction results accomplished by the QTLBO-WELM model on the given 4 datasets. On the initial monsoon dataset, the QTLBO-WELM scheme has importantly divided 27 instances ‘purity’ and 7 instances as ‘impurity’. Likewise, on the post-monsoon dataset, the QTLBO-WELM framework has significantly categorized 27 instances ‘purity’ and 7 instances as ‘impurity’. Along with that, on the pre-monsoon dataset, QTLBO-WELM technique has certainly divided 24 samples ‘purity’ and 8 instances as ‘impurity’. Simultaneously, on the summer dataset, the QTLBO-WELM approach has efficiently categorized 27 instances ‘purity’ and 7 instances as ‘impurity’ as demonstrated in Table 4.

Table 4 Confusion Matrix of Proposed QTLBO-WELM Method on Applied Dataset

Classes	**Monsoon-Dataset**			**Post-Monsoon Dataset**			**Pre-Monsoon Dataset**			**Summer Dataset**		
	Purity	Impurity	Total	Purity	Impurity	Total	Purity	Impurity	Total	Purity	Impurity	Total
Purity	27	1	**28**	27	0	**27**	24	1	**25**	27	0	**27**
Impurity	0	7	**7**	1	7	**8**	2	8	**10**	1	7	**8**
Total	**27**	**8**	**35**	**28**	**7**	**35**	**26**	**9**	**35**	**28**	**7**	**35**

Table 5 and Fig. 8-9 investigate the prediction outcome analysis of the QTLBO-WELM model on the given dataset. The table scores have defined that the QTLBO-WELM methods has exhibited efficient classification function on all applied dataset. The QTLBO-WELM technique has gained higher sensitivity of 100%, specificity of 87.5%, accuracy of 97.14%, precision of 96.43%, F-score of 98.18%, and kappa value of 91.53% on the test monsoon dataset.

Table 5 Result Analysis of Proposed QTLBO-WELM Method on Applied Dataset

Dataset	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-score**	**Kappa**
Monsoon Dataset	100	87.50	97.14	96.43	98.18	91.53
Post-Monsoon Dataset	96.43	100	97.14	100	98.18	91.53
Pre-Monsoon Dataset	92.31	88.89	91.43	96.00	94.12	78.35
Summer Dataset	96.43	100	97.14	100	98.18	91.53
Average	**96.29**	**94.10**	**95.71**	**98.11**	**97.17**	**88.24**

In line with this, the QTLBO-WELM framework has accomplished maximum sensitivity of 96.43%, specificity of 100%, accuracy of 97.14%, precision of 100%, F-score of 98.18%, and kappa value of 91.53% on the test post-monsoon dataset. Subsequently, the QTLBO-WELM technique has attained best sensitivity of 92.31%, specificity of 88.89%, accuracy of 91.43%, precision of 96%, F-score of 94.12%, and kappa value of 78.35% on the test monsoon dataset. Followed by, the QTLBO-WELM approach has obtained superior sensitivity of 96.43%, specificity of 100%, accuracy of 97.14%, precision of 100%, F-score of 98.18%, and kappa value of 91.53% on the test monsoon dataset.

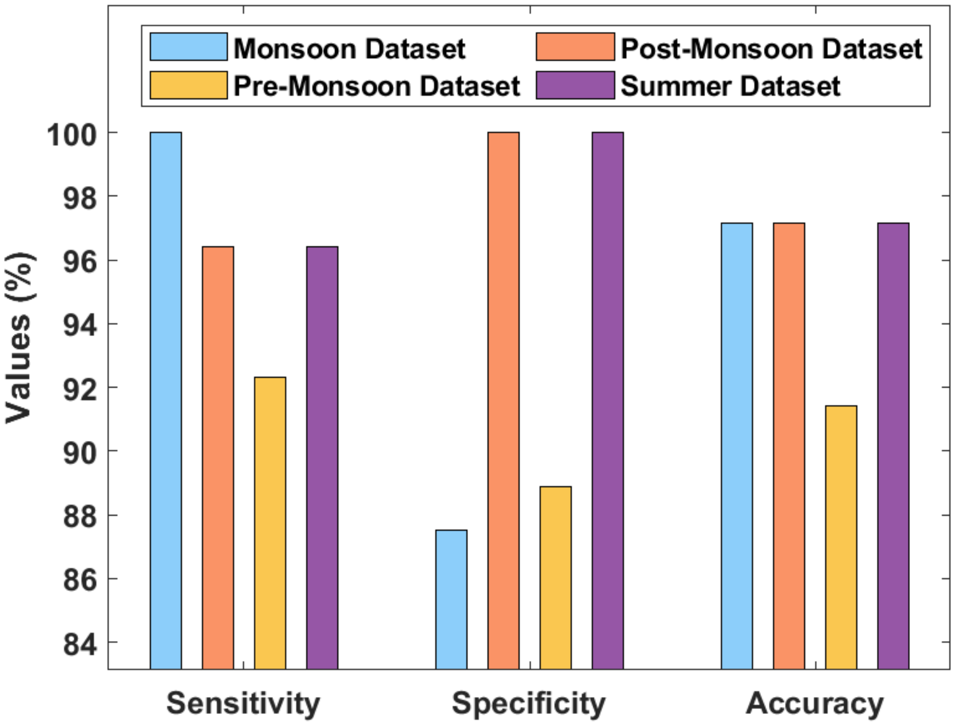

Fig. 8. Result analysis of QTLBO-WELM model

Received 16 February 2021; Accepted 08 March 2021.

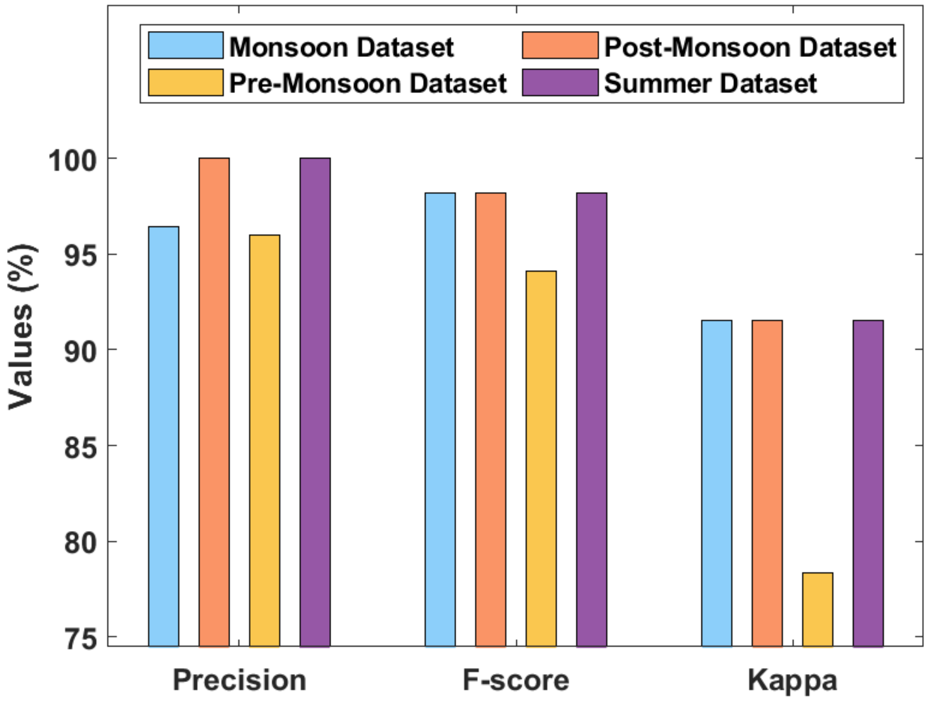

Fig. 8. Result analysis of QTLBO-WELM model

Fig. 10 investigates the average classifier results analysis of the QTLBO-WELM technique with the existing schemes. From the figure, it is implied that the QTLBO-WELM model has accomplished maximum average sensitivity of 96.29%, specificity of 94.1%, accuracy of 95.71%, precision of 98.11%, F-score of 97.17%, and kappa value of 88.24% on the test monsoon dataset.

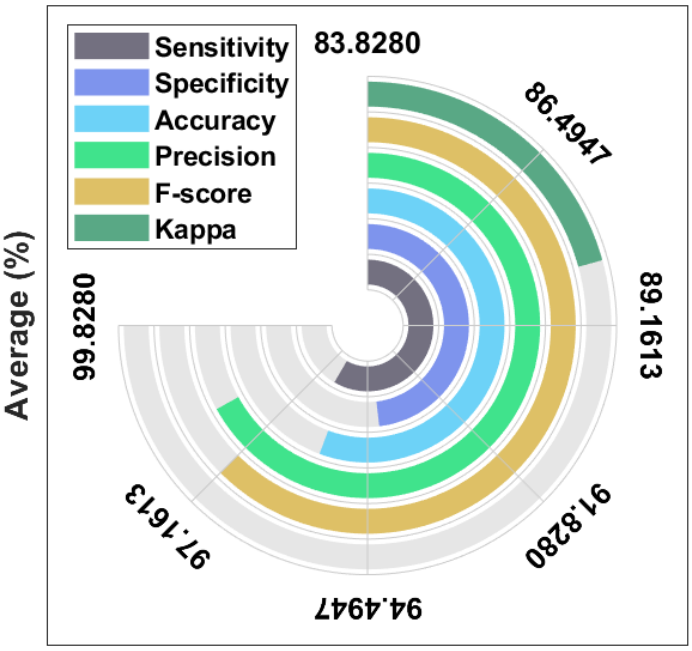

Fig. 10. Average analysis of QTLBO-WELM model with different measures

Received 16 February 2021; Accepted 08 March 2021.

Table 6 displays extensive results gained by the QTLBO-WELM with traditional approaches [21-25] by means of various measures.

Table 6 Result Analysis of Existing with Proposed QTLBO-WELM Method on Applied Dataset

Methods	Sensitivity	Specificity	Accuracy	Precision	F-score	Kappa
Proposed QTLBO-WELM	96.29	94.10	95.71	98.11	97.17	88.24
WELM	94.31	92.90	93.76	95.27	95.30	86.56
ANFIS	91.34	63.28	87.72	88.90	87.34	63.28
ANN	81.23	49.56	73.82	76.09	77.12	37.13
Decision Tree	80.72	63.21	87.69	87.70	79.97	45.78
C4.5	78.45	45.22	72.13	73.05	74.31	35.87
Random Forest	50.11	46.94	75.87	50.63	50.27	31.43
Naive Bayes	73.55	49.08	72.31	72.69	70.35	33.32
GBT Classifier	74.82	50.14	73.80	73.80	71.90	34.61
MLP	56.40	-	85.07	56.59	-	56.49
Bagging	-	-	67.41	-	-	-
K-Star	-	-	68.89	-	-	-
K-NN	-	-	72.00	-	-	-
SVM	-	-	87.10	-	-	-
Linear-SVM	-	-	76.03	-	-	-
Sigmoid-SVM	-	-	51.41	-	-	-

Fig. 11 examines the comparative accuracy analysis of the QTLBO-WELM scheme on the given dataset. From the figure, it is portrayed that the Sigmoid-SVM method acts as a poor performer by accomplishing low accuracy of 51.41%. Simultaneously, the Bagging model has resulted with considerable accuracy of 67.41%. Likewise, the K-Star scheme has attained moderate accuracy of 68.89%. Followed by, the K-NN technology has attained slightly better function with the accuracy of 72%. Concurrently, the C4.5 model has resulted reasonable outcomes with the accuracy of 72.13%. Meanwhile, the NB approach has demonstrated maximum results with the accuracy of 72.31%. In addition, the GBT classification technique has tried to perform better when compared with classical technologies with the accuracy of 73.80% whereas moderate accuracy of 73.82% was achieved by the ANN model. Also, the Linear-SVM method has showcased near optimal accuracy of 76.03%. Followed by, the MLP and SVM schemes have shown reasonable results with the accuracy of 85.07% and 87.1%. Therefore, the DT, ANFIS, and WELM methodologies have outperformed maximum outcomes with accuracy of 87.69%, 87.72%, and 93.76% respectively. Thus, the presented QTLBO-WELM technology has exhibited qualified results with the accuracy of 95.71%.

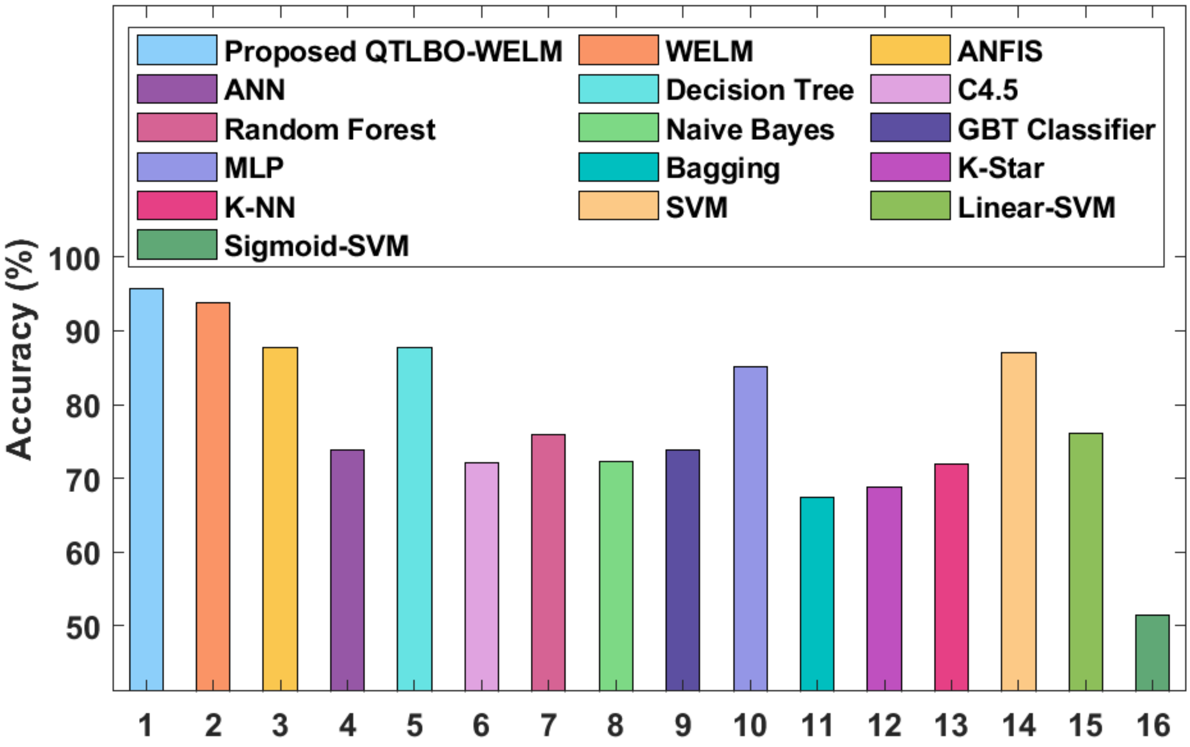

Fig. 11. Comparative analysis of QTLBO-WELM model interms of accuracy

Fig. 12 determined competing analysis of the QTLBO-WELM approach on the applied dataset with respect to sensitivity and specificity. The figure depicted that the RF model is served as inferior performer by gaining minimum sensitivity of 50.11% and specificity of 46.94%. Along with that, the MLP model has exhibited considerable results with sensitivity of 56.4%. Meantime, the NB method has accomplished acceptable sensitivity of 73.55% and specificity 49.08%. Similarly, the GBT classification method has accomplished reasonable sensitivity of 74.82% and specificity of 50.14%. Next, the C4.5 framework has attained moderate performance with the sensitivity of 78.45% and specificity of 45.22%. In line with this, the DT method has processed well even result with 80.72% and 63.21%.

Simultaneously, the ANN technique has exhibited useful results with the sensitivity of 81.23% and specificity of 49.56%. At the same time, the ANFIS and WELM approaches have demonstrated higher results with sensitivity of 91.34%, 94.31%, and specificity of 63.28%, 92.9%. Hence, the proposed QTLBO-WELM algorithms have displayed supreme results with the sensitivity of 96.29% and specificity of 94.10%.

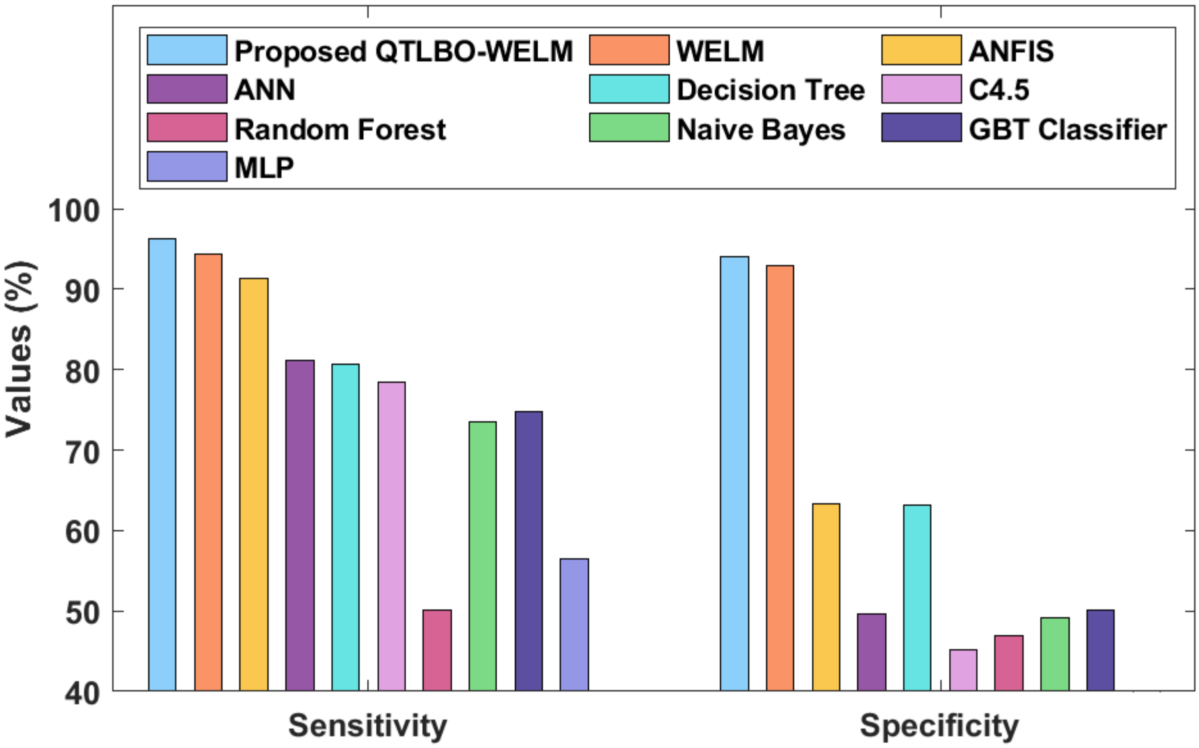

Fig. 12. Comparative analysis of QTLBO-WELM model interms of Sensitivity and specificity

Fig. 13 implies the comparative analysis of the QTLBO-WELM method on the applied dataset by means of precision, F-score, and kappa. The figure ensured that the RF model is facilitated as ineffective performance by gaining least precision of 50.63%, F-score of 50.27%, and kappa of 31.43%. As same as, the MLP model has provided considerable outcomes with precision of 56.59% and kappa of 56.49%. At the same time, the NB framework has reached acceptable precision of 72.69%, F-score of 70.35%, and kappa of 33.32%. In line with this, the C4.5 scheme has accomplished maximum precision of 73.05%, F-score of 74.31%, and kappa of 35.87%. Besides, the GBT classifier has gained maximum function with the precision of 73.8%, F-score of 71.9%, and kappa of 34.61%. Meanwhile, the ANN approach has depicted considerable outcomes with the precision of 76.09%, F-score of 77.12%, and kappa of 37.13%.

Received 16 February 2021; Accepted 08 March 2021.

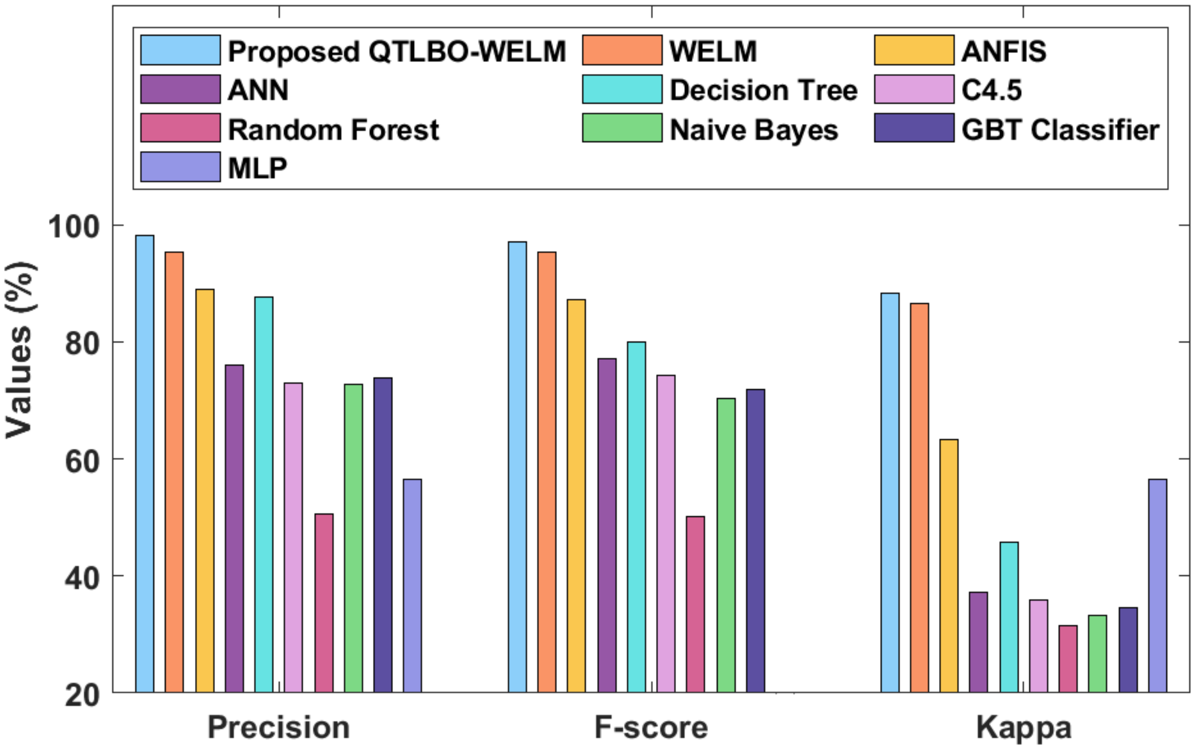

Fig. 13. Comparative analysis of QTLBO-WELM model interms of precision, F-score, and kappa

However, the DT approach has attained better results with precision of 87.7%, F-score of 79.97%, and kappa of 37.13%. Afterward, the ANFIS frameworks have outperformed reasonable outcomes with precision of 88.9%, F-score of 87.34%, and kappa of 63.28%. But, the WELM technology has provided moderate results with precision of 95.27%, F-score of 95.3%, and kappa of 86.56%. Thus, presented QTLBO-WELM technique has exhibited supreme results with the precision of 98.11%, F-score of 97.17%, and kappa of 88.24%.

4. Conclusion

This paper has presented a feature subset selection with classification model, called QTLBO-KELM model for water quality prediction. The presented model involves three processes, such as preprocessing, feature selection, and classification. The input data is preprocessed and then WTLBO algorithm is employed as a feature selector. Afterward, the reduced feature subset is fed into the WELM model for classification purposes. In order to assess the predictive outcome of the presented model, a series of experiments were conducted on a collection of 35 groundwater samples from Dharmapuri district in Tamil Nadu. The experimental values showcased the betterment of the presented QTLBO-WELM model with the sensitivity of 96.29%, specificity of 94.10%, accuracy of 95.71%, precision of 98.11%, F-score of 97.17%, and kappa value of 88.82%. In future, the predictive performance can be further increased by the use of deep learning (DL) approaches.

References

[1] Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality index (WQI) for the Loktak Lake in India. Appl. Water Sci. 2017, 7, 2907–2918.

[2] Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.

[3] Ahmad, Z.; Rahim, N.; Bahadori, A.; Zhang, J. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int. J. River Basin Manag. 2017, 15, 79–87

[4] Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ. 2016, 2, 8.

[5] Abyaneh, H.Z. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J. Environ. Health Sci. Eng. 2014, 12, 40.

[6] Ali, M.; Qamar, A.M. Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan. In Proceedings of the Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, Pakistan, 10–12 September 2013; pp. 108–113.

[7] Gazzaz, N.M.; Yusoff, M.K.; Aris, A.Z.; Juahir, H.; Ramli, M.F. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar. Pollut. Bull. 2012, 64, 2409–2420.

[8] Rankovi´c, V.; Radulovi´c, J.; Radojevi´c, I.; Ostoji´c, A.; Comi´c, L. Neural network modeling of dissolved oxygen ˇ in the Gruža reservoir, Serbia. Ecol. Model. 2010, 221, 1239–1244.

[9] K.M.K. Kut, A. Sarswat, J. Bundschuh, D. Mohan, Water as key to the sustainable development goals of South Sudan–a water quality assessment of eastern Equatoria state, Groundwater Sustain. Dev. 8 (2019) 255–270.

[10] T. Mitrovic,´ D. Antanasijevic,´ S. Lazovic,´ A. Peric-Gruji ´ c,´ M. Ristic,´ Virtual water quality monitoring at inactive monitoring sites using Monte Carlo optimized

artificial neural networks: a case study of Danube river (Serbia), Sci. Total Environ. 654 (2019) 1000–1009.

[11] E. Fijani, R. Barzegar, R. Deo, E. Tziritis, S. Konstantinos, Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters, Sci. Total Environ. 648 (2019) 839–853.

[12] R. Wan, F. Meng, E. Su, W. Fu, Q. Wang, Development of a classification scheme for evaluating water quality in marine environment receiving treated municipal effluent by an integrated biomarker approach in Meretrixmeretrix, Ecol. Indic. 93 (2018) 697–703.

[13] H. Yousefi, S. Zahedi, M.H. Niksokhan, Modifying the analysis made by water quality index using multi-criteria decision making methods, J. Afr. Earth Sci. 138 (2018) 309–318.

[14] I.C. Nnorom, U. Ewuzie, S.O. Eze, Multivariate statistical approach and water quality assessment of natural springs and other drinking water sources in Southeastern Nigeria, Heliyon 5 (1) (2019) e01123.

[15] P. Kumar, Simulation of Gomti River (Lucknow City, India) future water quality under different mitigation strategies, Heliyon 4 (12) (2018) e01074.

[16] V. Roth, T. Lemann, G. Zeleke, A.T. Subhatu, T.K. Nigussie, H. Hurni, Effects of climate change on water resources in the upper Blue Nile Basin of Ethiopia, Heliyon 4 (9) (2018) e00771.

[17] A. Awotwi, M.A. Bediako, E. Harris, E.K. Forkuo, Water quality changes associated with cassava production: case study of white Volta basin, Heliyon 2 (8) (2016) e00149.

[18] Agrawal, R.K., Kaur, B. and Sharma, S., 2020. Quantum based whale optimization algorithm for wrapper feature selection. *Applied Soft Computing*, *89*, p.106092.

[19] Allam, M. and Nandhini, M., 2018. Optimal feature selection using binary teaching learning based optimization algorithm. *Journal of King Saud University-Computer and Information Sciences*.

[20] Lu, C., Ke, H., Zhang, G., Mei, Y. and Xu, H., 2019. An improved weighted extreme learning machine for imbalanced data classification. *Memetic Computing*, *11*(1), pp.27-34.

[21] Lerios, J.L. and Villarica, M.V., 2019. Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir. *International Journal of Mechanical Engineering and Robotics Research*, *8*(6).

[22] Muhammad, S.Y., Makhtar, M., Rozaimee, A., Aziz, A.A. and Jamal, A.A., 2015. Classification model for water quality using machine learning techniques. *International Journal of software engineering and its applications*, *9*(6), pp.45-52.

[23] Babbar, R. and Babbar, S., 2017. Predicting river water quality index using data mining techniques. *Environmental Earth Sciences*, *76*(14), p.504.

[24] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R. and García-Nieto, J., 2019. Efficient water quality prediction using supervised Machine Learning. *Water*, *11*(11), p.2210.

[25] Liao, Y., Xu, J. and Wang, W., 2011. A method of water quality assessment based on biomonitoring and multiclass support vector machine. *Procedia Environmental Sciences*, *10*, pp.451-457.

[26] Metawa, N., Pustokhina, I. V., Pustokhin, D. A., Shankar, K., & Elhoseny, M. (2021). Computational Intelligence-Based Financial Crisis Prediction Model Using Feature Subset Selection with Optimal Deep Belief Network. Big Data.

[27] Le, DN., Parvathy, V.S., Gupta, D. et al. IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification. Int. J. Mach. Learn. & Cyber. (2021). https://doi.org/10.1007/s13042-020-01248-7

[28] Rajagopal, A., Ramachandran, A., Shankar, K., Khari, M., Jha, S., & Joshi, G. P. Optimal routing strategy based on extreme learning machine with beetle antennae search algorithm for Low Earth Orbit satellite communication networks. International Journal of Satellite Communications and Networking.

[29] K. Shankar, E. Perumal, M. Elhoseny and P. T. Nguyen, "An iot-cloud based intelligent computer-aided diagnosis of diabetic retinopathy stage classification using deep learning approach," Computers, Materials & Continua, vol. 66, no.2, pp. 1665–1680, 2021.

[30] Denis A. Pustokhin, Irina V. Pustokhina, Phuoc Nguyen Dinh, Son Van Phan, GiaNhu Nguyen, Gyanendra Prasad Joshi & Shankar K. (2020) An effective deep residual network based class attention layer with bidirectional LSTM for diagnosis and classification of COVID-19, Journal of Applied Statistics, DOI: 10.1080/02664763.2020.1849057