Hindawi
Complexity
Volume 2021, Article ID 6643472, 16 pages
https://doi.org/10.1155/2021/6643472

Research Article

Improvement in Explicit Prediction of Water Quality Using Wavelet-Based LSSVR and M5pRT

Rashmi Bhardwaj [1] **and Aashima Bangia** [2]

[1]*University School of Basic and Applied Sciences (USBAS), Head, Non-Linear Dynamics Research Lab, GGS Indraprastha University, Delhi, India*
[2]*USBAS, GGS Indraprastha University, Delhi, India*

Correspondence should be addressed to Rashmi Bhardwaj; rashmib@ipu.ac.in

Received 3 November 2020; Revised 29 January 2021; Accepted 1 March 2021; Published 20 March 2021

Academic Editor: Zaher Mundher Yaseen

Imbalance in the pH of water reduces this precious resource as an extremely dangerous liquid for human health and plants' growth. Change in the pH levels of the drinkable water has majorly raised concern towards diverse health issues like heart problems, infant mortality rates, pigmentation of skin, and cholera outbreaks. Therefore, it is necessary to keep a check on essential water quality components that include acidic/basic nature of water. As per the US Environmental Protection Agency (USEPA), the drinkable water should have a pH level ranging from 6.5 to 8.5. Two sample situates have been identified wherever highly reported pollutants levels were found and have been analyzed through artificial intelligence (AI) techniques. It can be observed that wavelet denoised signals fed into the least squares support vector regression (LSSVR) and M5 prime regression tree (M5pRT) predicted more accurately on the basis of the performance errors that are as follows: (a) root mean squared error (RMSE); (b) mean squared error (MSE); (c) mean absolute error (MAE). On the basis of these errors, the coefficient of determination/goodness of fit (R^2) simulated for the prototypes is developed in this study. RMSE outcomes diminish on the whole on applying the training and forecasting data-division via WLSSVR and WM5pRT as compared with fitting the normalized data through LSSVR and M5pRT. These performance measures are essential to analyze the concentration levels of pH in the river streams at the identified sites of study. Thus, the observed pattern from this study may help for future estimation of the quality of water at their sources so that it prohibits the further increase in either acidic or basic salts which prove to be lethal for the environment. Thus, these predictors would be helpful towards formulation of strategies for protection of ecosystem and human health.

1. Introduction

Water sustains life. It is referred to be the most precious resource for all the living creatures on Earth. Each and every natural water source that is counted upon as fresh contains salts in varying concentrations. This results in increase in pH levels as the water flows through oceans, rivers, and waterways and finally gets consumed by either mammals or plants and trees in the ecological system. A stream/river evolving through the mountain watershed could contain as less as 50 parts per million (i.e., ppm) in total dissolved solutes. Ocean water averages about 35,000 ppm which is about 3.5% of the dissolved solutes. Gradually, this has risen the health and ecological concerns universally [1]. Mostly consumed for drinking by humans and animals alike plus plants' growth makes water an integral part of ecosystem.

Consuming water with unfair pH levels has developed devastating replications on human health. The WHO a standard health monitoring unit has laid clear instructions for maintaining the balanced pH levels of water used for human intake [2, 3]. This is because people consuming inappropriate quality of water are unaware to health effects caused by inappropriate levels of water. Thus, it had to be

clearly stated and identified that water pH should not be compromised as it is affecting numerous households on yearly basis. Drinking of such waters has led to cardiovascular, hypertension, epidermal, and other serious diseases.

In general, it has been observed that the health hazards, spread of waterborne diseases, and treatment costs have worsened due to imbalanced potential of hydrogen (pH) level. People consuming acidic/basic water have increased manifold and are unaware of its substantial negative impacts on their wellbeing in the longer run. It has over the time transformed into an added dimension in view of the health insecurities that possibly lead to increased financial liabilities. While, some prescribed limit of salts are essential for the human body to avoid iodine deficiency. Still, increased intake of salt on a regular basis is intolerable by the body as it is difficult to be absorbed by the body fluids present. The abrupt amount of iodine is undesirable and thus turns the body prone to hypertension, increased levels of blood pressure, heart stroke, and various other ailments. Therefore, universal authorities such as USEPA and WHO believe salt should be consumed by the body within limits considering health as a topmost priority.

Simulations through moving average combined with the wavelet model are carried out on rainfall data for forecasting noise [4]. The adaptive neuro-fuzzy system is applied to study the BOD of River Surma [5]. Prediction of dynamic indicators is applied on atmospheric pollutants [6]. AI techniques have been integrated with the wavelet decomposition for restructuring the data to predict the river water quality [7]. Large-scale info-data applications are explained with LSSVR [8]. Control and prediction of time series are done [9]. Simulated and forecasted surface flows via self-tuned ANN model are studied [10]. SVR with the kernel estimated short-term loading is studied [1]. The algorithm of GWO-ANFIS for prediction of hydropower generator is developed [11]. River-flow in Plata-basin attribution is studied [12]. Transboundary rivers from Romania were discussed regarding the water pollution and quality being affected [13]. Discharge of pollutants in a vegetated compound meandering river is studied [14]. DWT with ANN analyzed the short-term stream-flows [15]. Detailed development and analysis through neural networks are provided [16]. ANN technique is applied in various real-life applications such as biological and environmental phenomena [17]. Precipitations on monthly basis info using the neuro-fuzzy method are predicted [18]. Deep learning networks are designed to assess water quality of mariculture with accuracy [19]. Water quality of Karoun River via regression and ANFIS is forecasted [20]. Various neural network-based models such as ANN, BNN, and ANFIS modeling are discussed for groundwater level predictions [21]. Variations, i.e., seasonal along with spatial ones are studied for the quality of river Yamuna [2]. Hybrid of SSMD-whale optimization is devised for prediction of longitudinal dispersion coefficients [22]. The SVR algorithm for predicting river water quality is improved [23]. ANN-ANFIS carried out uncertainty analysis for assessment of gravel transport [24]. Optimal multigene programming simulated the dispersion coefficients [25]. Analysis is carried out through wavelet transform, genetics algo, and neural networks of monsoon floods [26]. WQI prediction is simulated through AI for studying groundwater systems [27]. Hourly records of ozone concentrations with the help of wavelet and ARIMA are forecasted [28]. Different machine learning-based hybrid models are carried out for estimating evapotranspiration in Iran [29]. Artificial intelligence methodologies on survey of long-term data from 2000–2020 for water quality are explored [3]. Decomposition mode ensemble modeling is analyzed for LSTM for streamflow forecasts [30].

This research article consists of case study and its dataset assessment in Section 2. Discussion of the mathematical model designing of LSSVR and M5pRT and further WD conjuncted to LSSVR and M5pRT procedure are given in Section 3. Performance measures for prediction are computed in Section 4 with algorithms for building models such as LSSVR, M5pRT, WLSSVR, and WM5pRT. Section 5 includes results observed from these hybrid models plus the errors are numerically simulated.

As per the literature survey carried out, none of the articles simulated the acidic or basic salts' presence in Yamuna waters through decomposition of wavelet (WD) with LSSVR and M5pRT for the sample sites considered in this study.

2. Case Study and Dataset Assessment for River Yamuna

The data consist of values from 2000 to 2019 of pH level at two major monitoring stations Nizamuddin Bridge in Old Delhi and Palla on the outskirts of Delhi as recorded by *Central Pollution Control Board* (*CPCB*). A total of 19 years' monthly values, i.e., from the year 2000 to 2019 have been trained and then simulated via intelligent learning regressive models LSSVR, M5pRT, WLSSVR, and WM5pRT in the study. For the conjuncted models WLSSVR and WM5pRT, first ten input data values are fed as the responses for the training and validation of the each of the datasets. It was detected that these proposed models give enriched efficiency and diminish error extreme sharply in contrast to existent classical models. Two stations have been numbered according to the flow of the river crossing stations in Table 1.

Starting at Yamunotri flowing till Allahabad, total extent of river Yamuna measures up to 1,376 kilometer with a total basin area as 3,66,223 km^2. The river tends to be practically dried-up in the region ranging from Hathnikund towards Delhi. The only source adding up waters is from the groundwater and small tributaries. From Hathnikund, the river reaches Delhi at Palla which has a spread of 224 km. Delhi itself dumps more than 58% of its unwanted garbage into these waters. Thus, the level of contamination is highest around *Delhi-NCR* geologically. The *Central Pollution Control Board (CPCB)* reported that adulterated expanse of Yamuna augmented from 500 to 600 km. Acidic or basic levels of water that have been highly reported into the river waters have been studied via pH turning greater than or lesser than 7 lying in the range 6.5–8 or outside, at the above listed stations. The location of Old Delhi Bridge and Palla can be spotted in the geological map given in Figure 1.

Table 1: Data sets for simulation of five hybrid models.

Cross validation	Data sets	Time period (monthwise recorded)	
		From	Up to
Data type 1	Old Delhi (Nizzamuddin Bridge)	Jan, 2000	Jan, 2019
Data type 2	Palla (outskirts of Delhi)	Jan, 2000	Jan, 2019

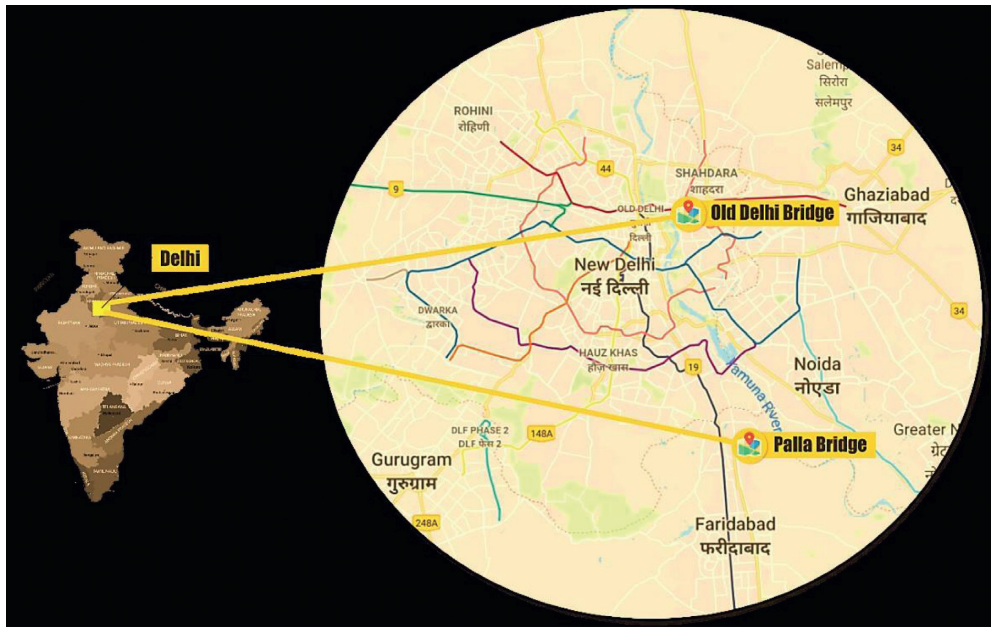


Figure 1: Map showing the two monitoring stations of river Yamuna.

This article studies at each location 229 data sets have been observed. This article studies 2 data types each having 229 data values.

3. Methodology

3.1. Least Square Support Vector Regression (LSSVR). Consider in general a function approximation problem that would be represented as follows:

$$f = g(x) = <\omega, x> + h = \sum_{j=1}^{K} w_j x_j + h, \quad f, h \in \mathbb{R}, x, \omega \in \mathbb{R}^K,$$
$$f(x) = \begin{bmatrix} \omega \\ h \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = \omega^T x + h, \quad x, \omega \in \mathbb{R}^{K+1}. \tag{1}$$

This problem can be solved efficiently by transforming into an optimization problem which is carried out through support vector regression (SVR) as follows:

$$\min_{\omega} \frac{1}{2}\|\omega\|^2, \tag{2}$$

where $\|\omega\|$ is themagnitude of normal to the surface to be estimated.

The measure of weights can be computed through:

$$g(x, \omega) = \sum_{j=1}^{K} \omega_j \cdot x^j, \quad x \in \mathbb{R}, \omega \in \mathbb{R}^K, \tag{3}$$

where K is the order of the polynomial.

Mainly, the constraint aims towards minimizing the performance measures that prevails in the predictors of the provided inputs and actual ones. Also, espouses ε-based loss function would penalize predictions farther than ε from anticipated output. Then, ε-value governs tube-width; a, that is, smaller the value, tolerance reduces towards simulation error and also affects number of support vectors that subsequently leads to sparsity of the solutions. If ε decreases, it leads to boundary to shift inwards. Thus, a greater number of target points around the boundary clearly indicate the increase in the number of support vectors. Similarly, the case of increasing ε fewer points around the boundary follows from the result.

Evolving on this technique, the least squares extension applied to the SVR modifies the minimization problem as follows:

$$\min \phi_2(\omega, h, \epsilon) = \frac{\gamma}{2}\omega^T \omega + \frac{\zeta}{2}\sum_{i=1}^{N} \epsilon_i^2, \tag{4}$$

and subjects to following constraints:

$$y_i\left[\omega^T \sigma(x_i) + h\right] = 1 - \epsilon_i, \quad i = 1, \ldots, K. \tag{5}$$

LSSVR representation involves explanation via the binary points as $y_i = \pm 1$

Now, with $y_i^2 = 1$,

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{K} (y_i \epsilon_i)^2 = \sum_{i=1}^{K} \epsilon_i^2 = \sum_{i=1}^{K} \left(y_i - \left(\omega^T \sigma(x_i) + h\right)\right)^2, \tag{6}$$

where $\epsilon_i = y_i - (\omega^T \sigma(x_i) + h)$

LSSVRs have been designed to tackle higher complicacies than SVR as put forward by Suykens. Objective function does not change much as compared with that of the existing SVR. Difference arises when ε-based loss function replaces the classical squared-loss function, and explaining every b_i coefficient becomes nonzero. Alongside, model proficiency increases on creating Lagrange multiplier that is obtained via resolving the Karush–Kuhn–Tucker (KKT) scheme. Solution of this system is carried out with the help of most standard approaches to solve sets-of linear equalities. SVR has three fine-tuning components defined whereas execution of LSSVR involves two such components. The prediction errors simulated are found to be least through LSSVR. It is said this prototype eradicates noises and moderates computational labor.

Remark 1. Thus, this LSSVR formulation modifies into the following:

$$\phi_2(\omega, h, \epsilon) = \gamma A_W + \zeta A_D,$$

with,

$$A_W = \frac{1}{2}\omega^T \omega \text{ and } A_D = \frac{1}{2}\sum_{i=1}^{K} \epsilon_i^2 = \frac{1}{2}\sum_{i=1}^{K} \left(y_i - \left(\omega^T \sigma(x_i) + h\right)\right)^2, \tag{7}$$

where γ and ζ are hyperparameters tuning amount of regularization w.r.t sum squared error (SSE).

Now, further LSSVR regressor solution can be obtained by the following Lagrangian function:

$$\left\{ L_m(\omega, h, \epsilon, \alpha) = \phi_2(\omega, \epsilon) - \sum_{i=1}^{K} \alpha_i \left\{ \left[\omega^T \sigma(x_i) + h\right] + \epsilon_i - y_i \right\} = \frac{1}{2}\omega^T \omega + \frac{\mu}{2}\sum_{i=1}^{K} \epsilon_i^2 - \sum_{i=1}^{K} \psi_i \left\{ \left[\omega^T \sigma(x_i) + h\right] + \epsilon_i - y_i \right\} \right\}. \tag{8}$$

Representing $\psi_i \in \mathbb{R}$ as Lagrange multipliers, minimality conditions can be counted in as follows:

$$\begin{aligned} &\left\{ \frac{\partial L_m}{\partial \omega} = 0 \quad \longrightarrow \omega = \sum_{i=1}^{N} \psi_i \sigma(x_i), \frac{\partial L_m}{\partial h} = 0 \longrightarrow \sum_{i=1}^{K} \psi_i = 0, \frac{\partial L_m}{\partial \epsilon_i} = 0 \longrightarrow \psi_i \right. \\ &\left. = \mu \epsilon_i, i = 1, \ldots, K, \frac{\partial L_m}{\partial \psi_i} = 0 \longrightarrow y_i = \omega^T \sigma(x_i) + h + \epsilon_i, i = 1, \ldots, K. \right. \end{aligned} \tag{9}$$

Now, it is imperative to eliminate ω and ϵ for the creation of the linear system as follows:

$$\begin{bmatrix} 0 & 1_K^T \\ 1_N & \Lambda + \mu^{-1} I_K \end{bmatrix} \begin{bmatrix} h \\ \psi \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix}. \tag{10}$$

Having $Y = [y_1, \ldots, y_N]^T, 1_N = [1, \ldots, 1]^T$ and $\psi = [\psi_1, \ldots, \psi_N]^T$;

I_N is the N dimensional identity matrix and $\Lambda_{ij} = \sigma(x_i)^T \sigma(x_j) = K_n(x_i, x_j)$ where $\Lambda \in \mathbb{R}^{N \times N}$ is the kernel matrix which can be linear kernel, polynomial kernel, multilayer kernel, or radial-basis function-based kernel. Thus, RBF kernel is defined as follows:

$$K_n\left(x_i, x_j\right) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{\phi^2} \right), \tag{11}$$

with $\phi \in R^+$ is the constant value.

The choice of kernel essentially determines the resultant regressor obtained from LSSVR as it normalizes data under study [14].

3.2. M5 Prime Regression Tree (M5pRT). Introduced by Quinlan in 1992, M5 model tree was established keeping in mind binary decision-tree that consisted of linear regression functions at terminal nodes referred to as leaf. The leaf stores relationship between independent and dependent variables. Such tree forming methodologies are based on the split and rule strategy that constructs a connection between independent and dependent variables. Tree models are also implemented on qualitative/quantitative corpora.

Theorem. The dividing criterion is basically standard deviation of the values of the subset formed on reaching the node to be taken as scale of error of that node plus computing estimated reduction error arising due to the process of testing carried out for each attribute at that particular node. Thus, standard deviation reduction (SDR) can be simulated from the following:

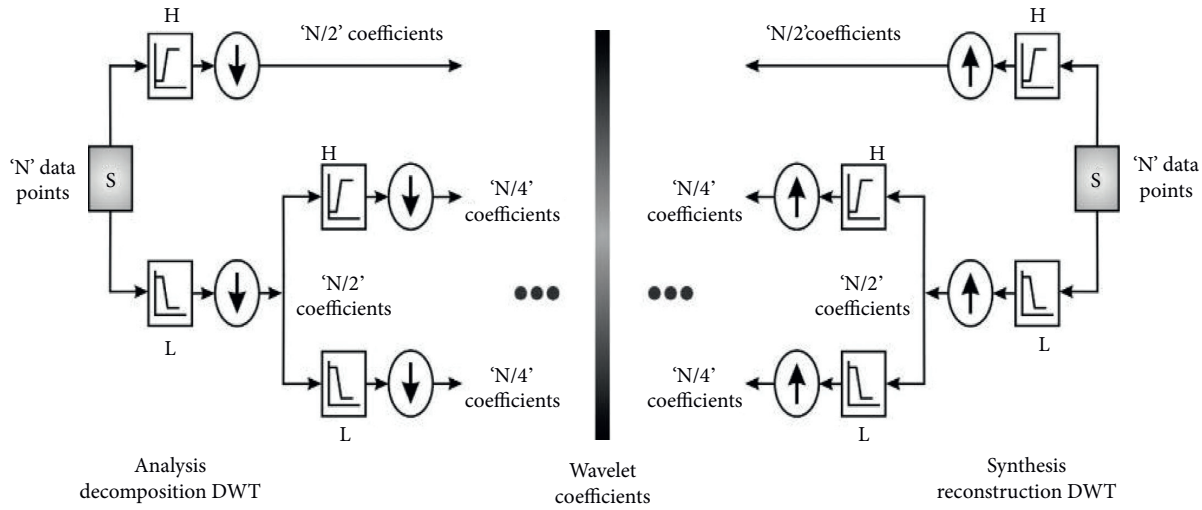


Figure 2: Wavelet study carried upon decomposition/synthesis and reconstruction through coefficients.

$$S\ DR = s\ d(T) - \sum_i \frac{|T_i|}{|T|} \times s\ d(T_i), \tag{12}$$

where *sd* is the standard deviation; T– is the set of instances that touch the node; and T_i is the sets obtained through node splitting w.r.t. the particular characteristic having value assigned to split.

Remark 2. Splitting process dismisses as and when outcomes for every instance which touch the node vary only slightly else if some instance remain.

The M5 algorithm over a period of time got extended into M5'. It was designed to substitute conventional regression structured by existing trees as M5' based on their splitting through *if-then* rules. Thus, this prototype is designed for dividing response realm into multiple subdomains plus linear regressive model to be fitted at every subdomain. M5' constructs regressors' tree on recursively splitting rule on the standard deviation calculated for class values that would influence nodes and error measures at each particular one. Attributes which maximize predictable inaccuracy that decrease get opted for splitting at the node. As branching process, data in child nodes (subtree or smaller nodes) have fewer SD than parent nodes (greater nodes). Reasonably eliminating all possible tree-forms, one that would have the maximum estimated error reduction is finalized.

3.3. Multiresolution-Based Discrete Wavelet Denoising

Theorem 1. *Wavelet is an apt balance of sine-cosine waves comprehending characteristics that would vary around zero and also lies within an interval domain. Wavelet-function is developed into father wavelet (ϕ) and mother wavelet (ψ) holding properties as follows:*

$$\int_{-\infty}^{\infty} \phi(x)dx = 1 \text{ and } \int_{-\infty}^{\infty} \psi(x)dx = 0. \tag{13}$$

Remark 3. By integration of amplified dyadic along with integral transformations, mother-father wavelets are transformed into the wavelet family as follows:

$$\phi_{j,k}(x) = 2^{j/2}\phi\left(2^j x - k\right) \text{ and } \psi_{j,k}(x) = 2^{j/2}\psi\left(2^j x - k\right) \tag{14}$$

3.3.1. Wavelet Decomposition Algorithm. It can be demarcated as follows: $c_{j-1,k} = \sum_l a_{j-1,l-2k} c_{j,l}$ and $d_{j-1,k} = \sum_l b_{j-1,l-2k} c_{j,l}$ that embrace high-low pass filters accordingly.

3.3.2. Wavelet-Reconstruction Algorithm. Representation of $c_{j,l} = \sum_k p_{j,l-2k} c_{j-1,k} + \sum_k q_{j,l-2k} d_{j-1,k}$ with filters is as follows.

Both Wavelet decomposition and reconstruction processes can be together observed in Figure 2 as it clearly shows the analysis through decomposition and synthesis through reconstruction in DWTs.

3.4. Wavelet Least Squares Support Vector Regression (WLSSVR).

The following flowchart in Figure 3 clearly puts forward every step involved in the simulation of the responses obtained after training through WD into

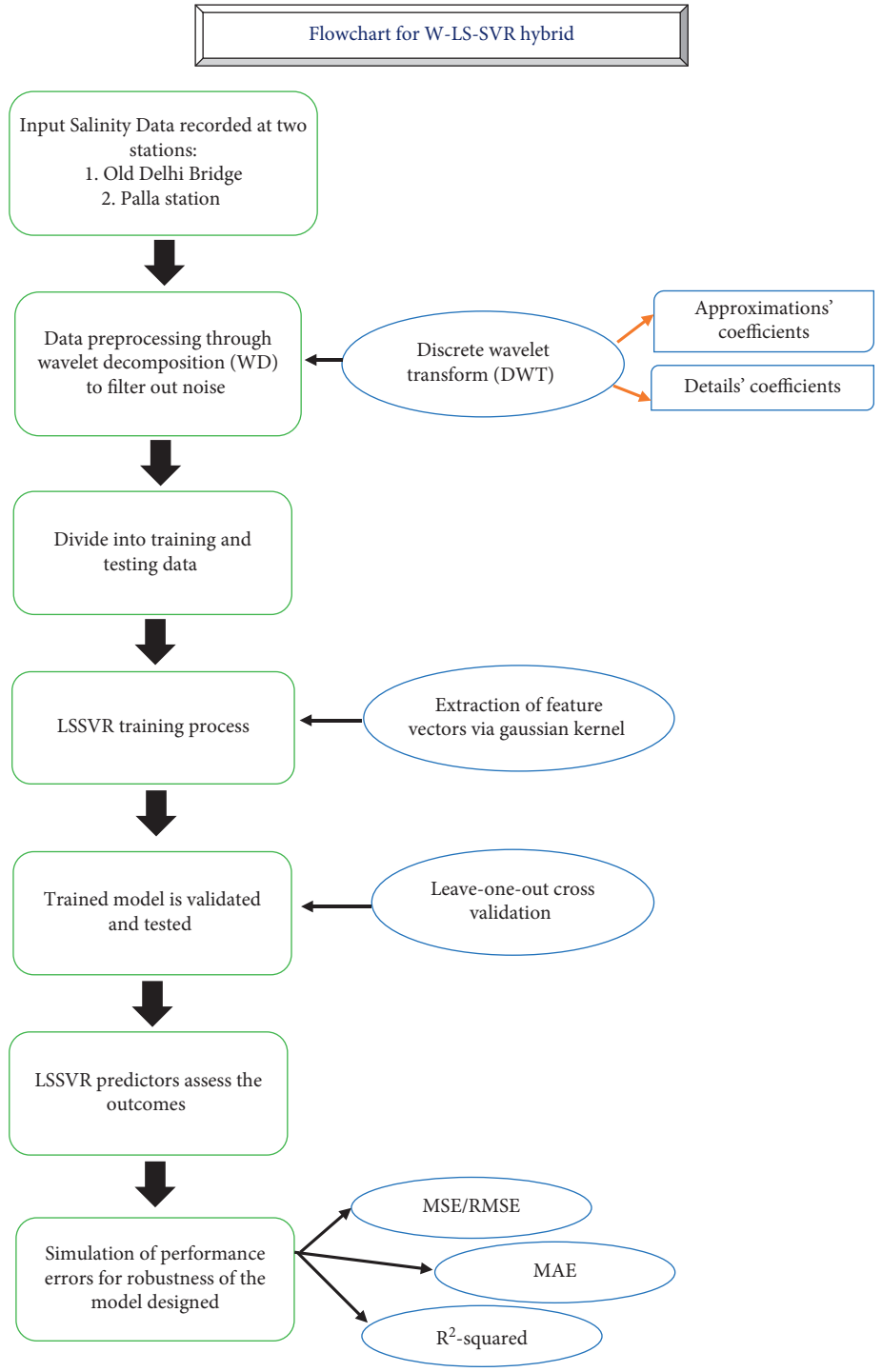


Figure 3: Flowchart of wavelet LSSVR prototype.

approximations and details and then fed into the least squares support vector regressors' setup where Gaussian kernel is an integral part of solving the so formed optimization formulation. WD filters out noise which can be also understood as outliers so as to compute results with better accuracy.

3.5. Wavelet M5 Prime Regression Tree (WM5pRT). Following flowchart in Figure 4 clearly puts forward every step involved in the simulation of the responses obtained after training through WD into approximations and details and then fed into the M5 prime regressors' tree setup where

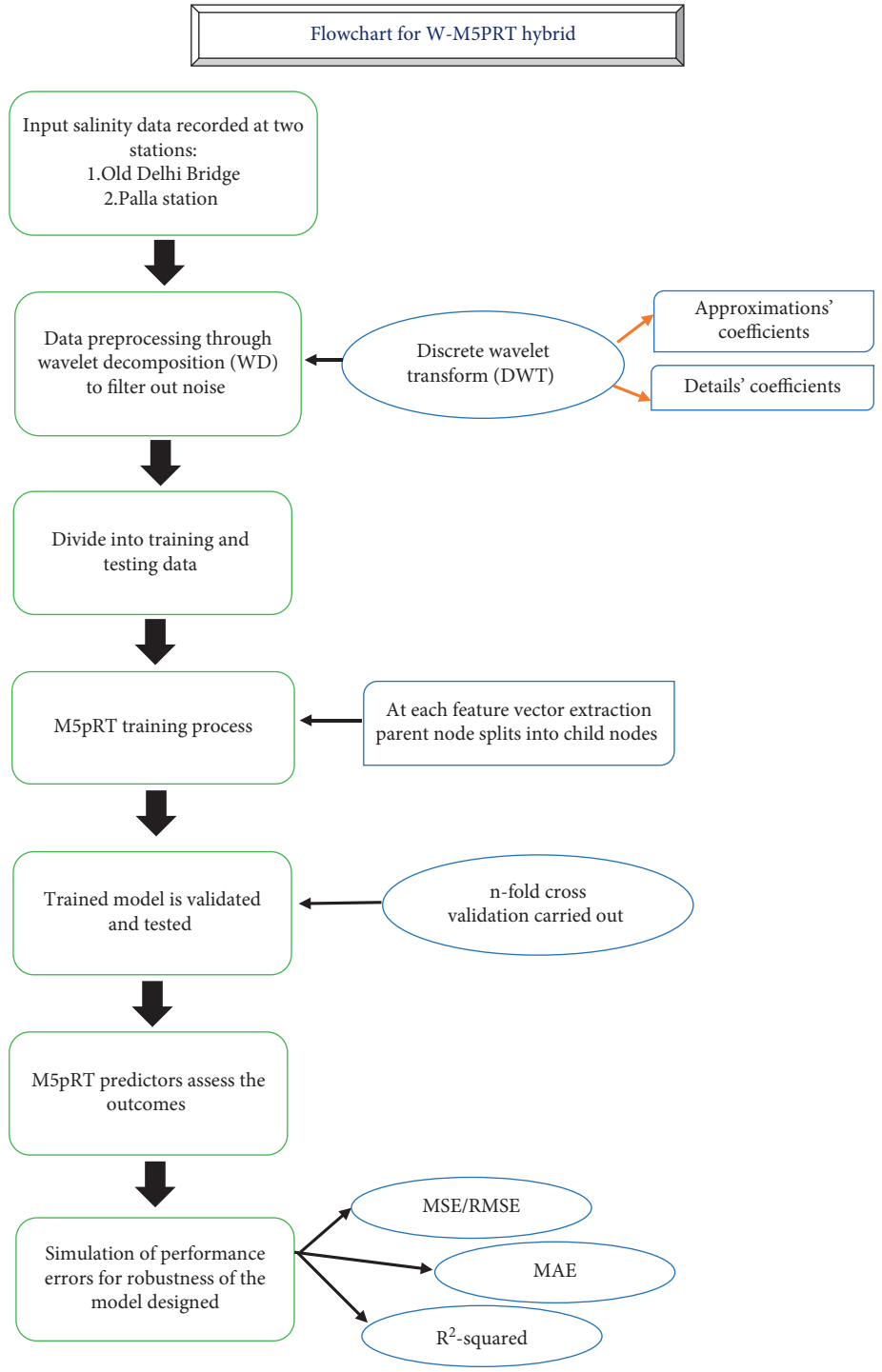


Figure 4: Flowchart of wavelet M5pRT prototype.

at every feature, f extracted the parent node splits into the child nodes and thus a tree formation takes place.

4. Performance Measures

For estimation of performance, each of the hybrid models' forecasting errors is computed for a comparison to understand which of the hybrids best suits the info under study. So, with regard to this, model responses recorded are used to simulate statistical measures referred to as the computational errors represented through root mean squared error (RMSE), mean absolute error (MAE), and also coefficient of determination (R^2) [14].

$$\text{MSE} = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2}{n}, \tag{15}$$

where y_i denotes the actual quantity; $\widehat{y}_i$ denotes the predicted assessment; and n is the no. of days in prediction.

Table 2: Comparison of prediction errors for data recorded two stations.

Learning methods	Stations	Performance measures			Goodness of fit statistic
		RMSE	MSE	MAE	R^2
LSSVR	Palla (outer Delhi)	7.7988	60.8711	7.7934	0.7260
	Old Delhi bridge	7.5796	57.4510	7.557	0.8420
M5pRT	Palla (outer Delhi)	77.740	6.1284e03	63.5206	0.4814
	Old Delhi bridge	42.464	1.8566e03	35.9128	0.4225
WLSSVR	Palla (outer Delhi)	7.7980	60.8208	7.7919	0.8640
	Old Delhi bridge	7.5714	57.4779	7.5765	0.8759
WM5pRT	Palla (outer Delhi)	101.1626	1.001e04	68.7353	0.6800
	Old Delhi bridge	49.5543	2.5661e03	42.0013	0.6079

The root mean squared error is square root for MSE.

$$\text{RMSE} = \sqrt{\text{MSE}}, \qquad R^2 = 1 - \frac{\text{SSE}}{\text{SST}}. \tag{16}$$

Having sum squared errors of regression fitting, $SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$, and sum squared errors of the actual data fitting, $SST = \sum_{i=1}^{n} (y_i - \overline{y}_i)^2$,

$$MAE = \frac{1}{N}\left(\sum_{i=1}^{N} |y_i - \widehat{y}_i|\right), \tag{17}$$

where y_i denotes the actual quantity, $\widehat{y}_i$ denotes the predicted assessment, and n is the no. of days in prediction in all the performance errors above.

5. Results and Discussion

Intelligent learning algorithms, namely, WLSSVR and WM5pRT are computed from the monthly basis data provided via Central Pollution Control Board, CPCB, for the pH recorded and noted at two discussed sample sites. This study helps in understanding through comparison of three neuronal models which one could improve the performance of the model-based structures and with time and would be cost effective. Table 2 represents the errors: MSE, RMSE, and MAE along with the fitness measure, i.e., R^2 for the explained for four models for the two stations: Old Bridge and Palla. At the station Palla, following graphs analyze the data and forecast with the help of responses. Figure 5 shows the decomposition of the wavelet-form signals according to Db8 into approximations (A_3) and details (D_1, D_2, and D_3) to filter out the noise which takes care of all kinds of nonlinearities for the extensive analysis. Wavelet filtered neuronal fuzzified inferences' data are divided into training and testing data for better predictions. Figure 6 graphs the linear fit on daily basis at Palla. It can be observed that pH values range from 6.8 to 8.8 and mostly data lie above the pH level of 7. Figures 7(a) and 7(b) clearly demonstrate the regression fitting through LSSVR and linear fit of LSSVR trained values separately. Here, LSSVR trains in the pH level range of 7.4 to 8.4 which is in the drinkable range as prescribed. Figures 8(a) and 8(b) clearly validate regression fitting through M5pRT and linear fit of M5pRT trained values separately. Values are concentrated from pH levels 7 to 8. Figures 9(a) and 9(b)

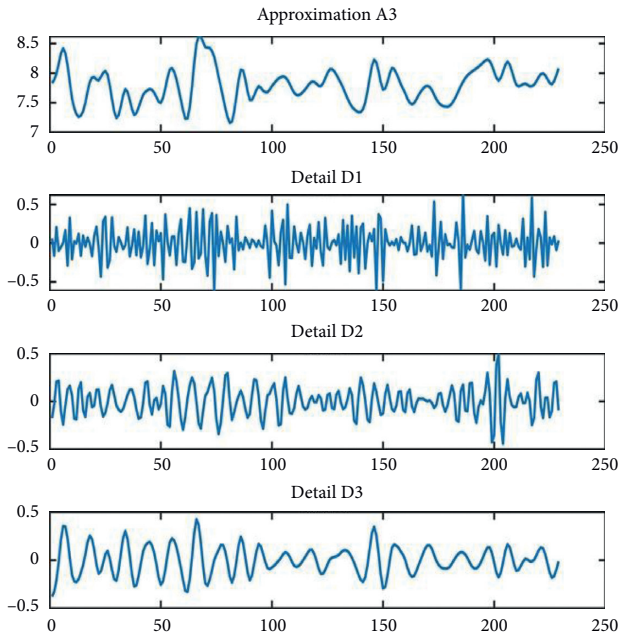


Figure 5: Wavelet decomposition of pH values at Palla.

demonstrate the regression fitting through WLSSVR and linear fit of WLSSVR trained values separately. WLSSVR captures most of the values concentrated around pH range 7.5 to 8.5 even though outliers can be seen around 6.5–7 and 9–9.5. Figures 10(a) and 10(b) clearly demonstrate regression fitting through WM5pRT and linear fit of WM5pRT trained values separately. Trains data and concentration lie from pH range 7 to 8.5 as for every feature extracted, and the parent node divides into child nodes and thus tree gets created.

Now, at the station Bridge with the help of daily data of values of pH levels, it can be observed whether the pH changes at this point due to various external factors and the effect of pH on the adjoining areas. Figure 11 shows the decomposition of the wavelet-form signals according to Db8 into approximations (A_3) and details (D_1, D_2, and D_3) to filter out the noise which takes care of all kinds of nonlinearities for the extensive analysis. Wavelet filtered neuronal fuzzified inferences' data are divided into training and testing data for better predictions. Figure 12 graphs the linear fit on daily basis at Old Delhi Bridge. It can be observed that linear fit line has pH values 7.4 to a little over 7.8.

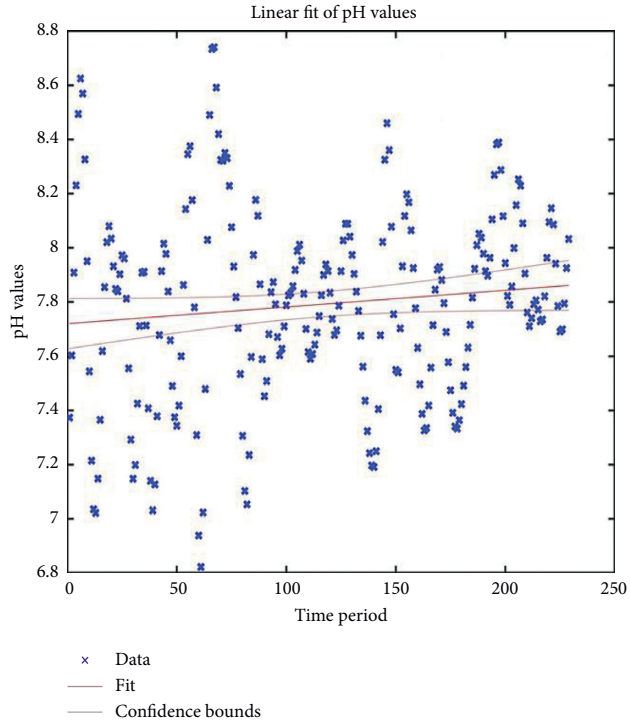
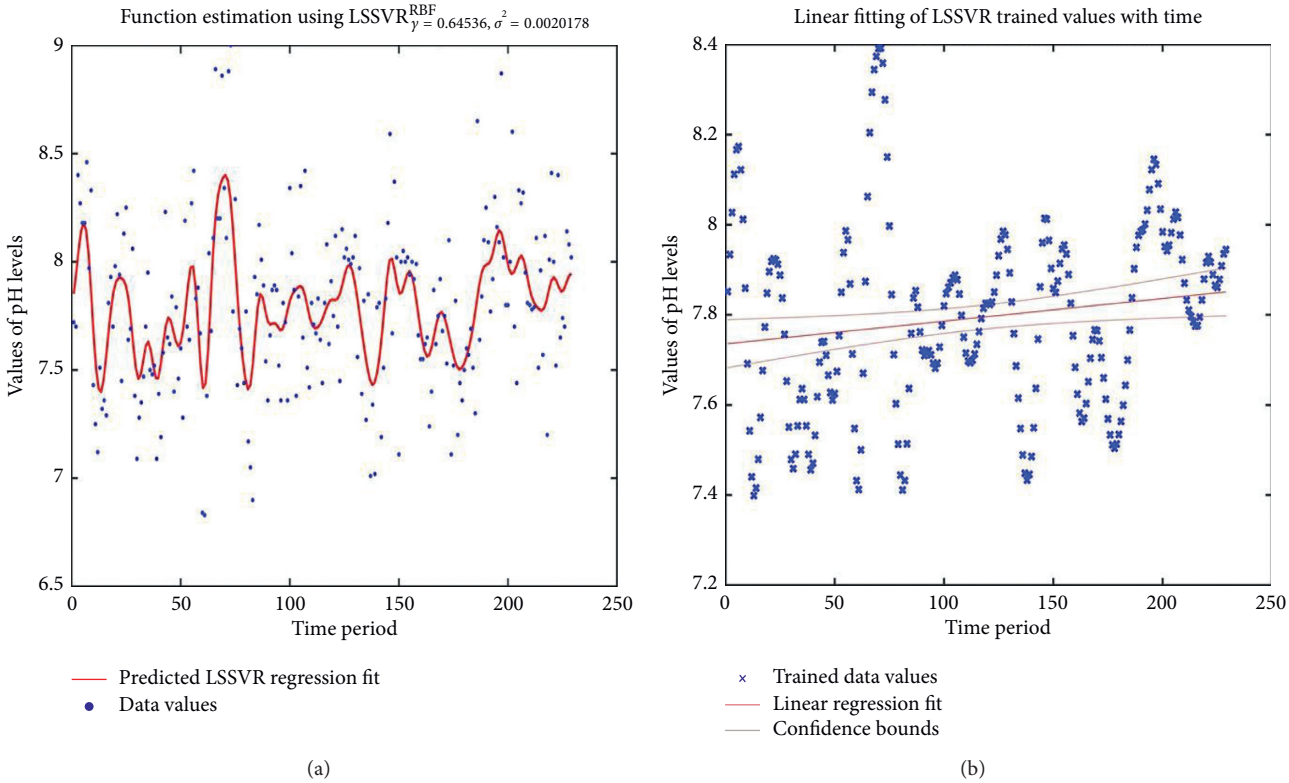


Figure 6: Linear regression fit of pH values at Palla.



(a) (b)

Figure 7: (a) LSSVR regression fit of pH data at Palla. (b) Linear regression fit of LSSVR trained pH values at Palla.

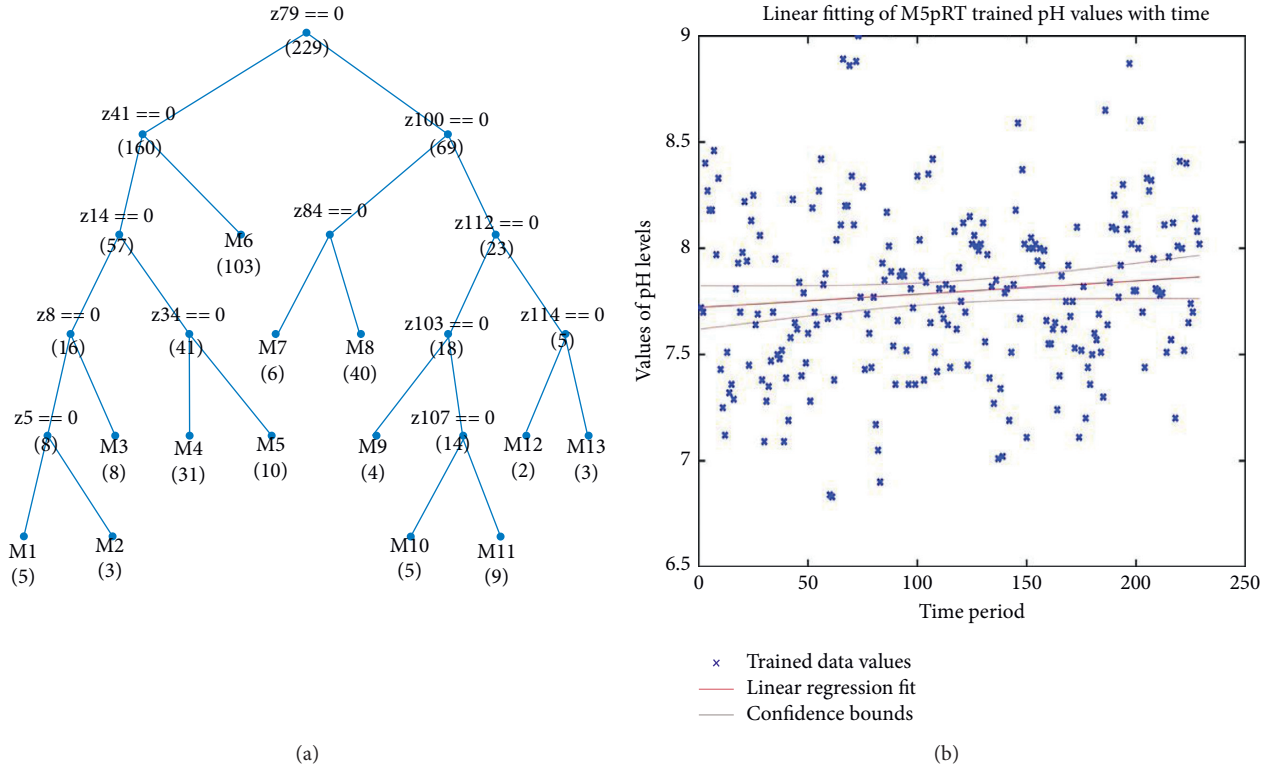


Figure 8: (a) Regression tree simulated via M5pRT at Palla. (b) Linear regression fit of M5pRT trained pH data at Palla.

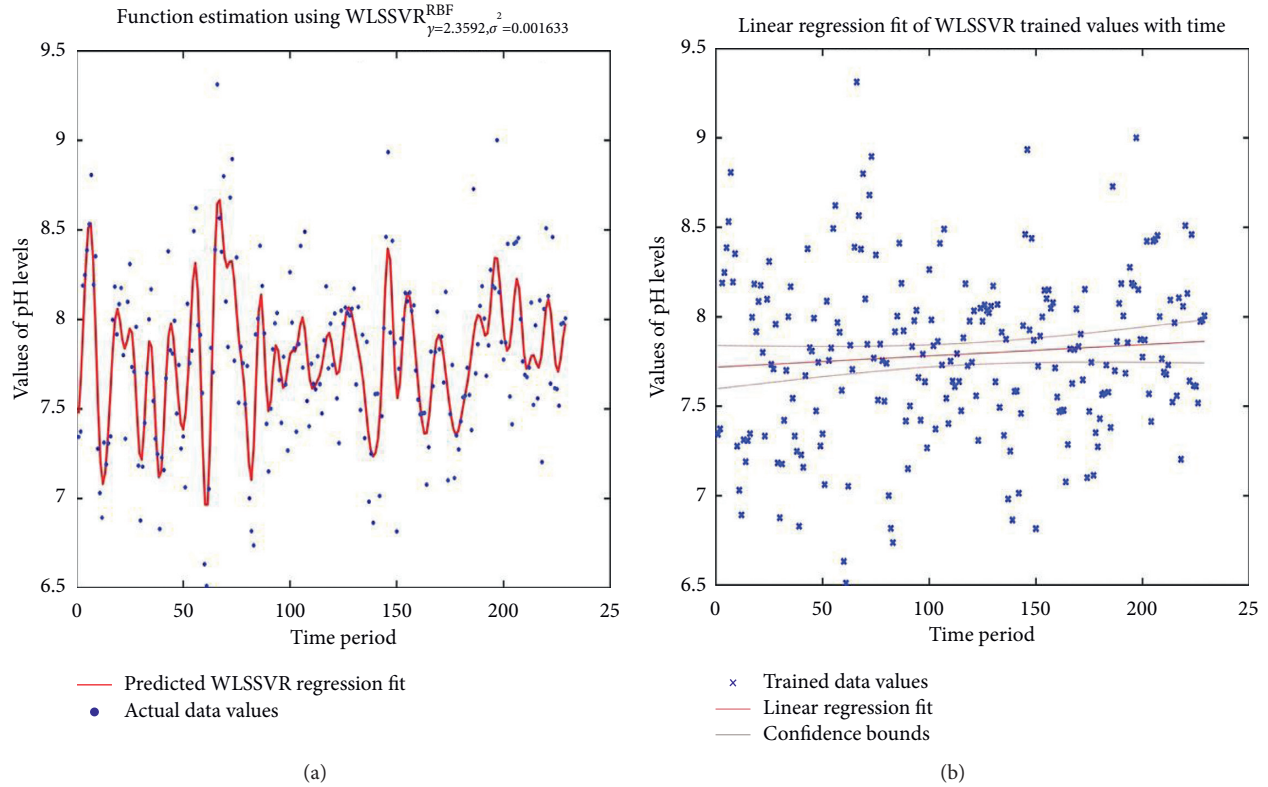


Figure 9: (a) WLSSVR regression fit of pH values at Palla. (b) Linear regression fit of WLSSVR trained pH values at Palla.

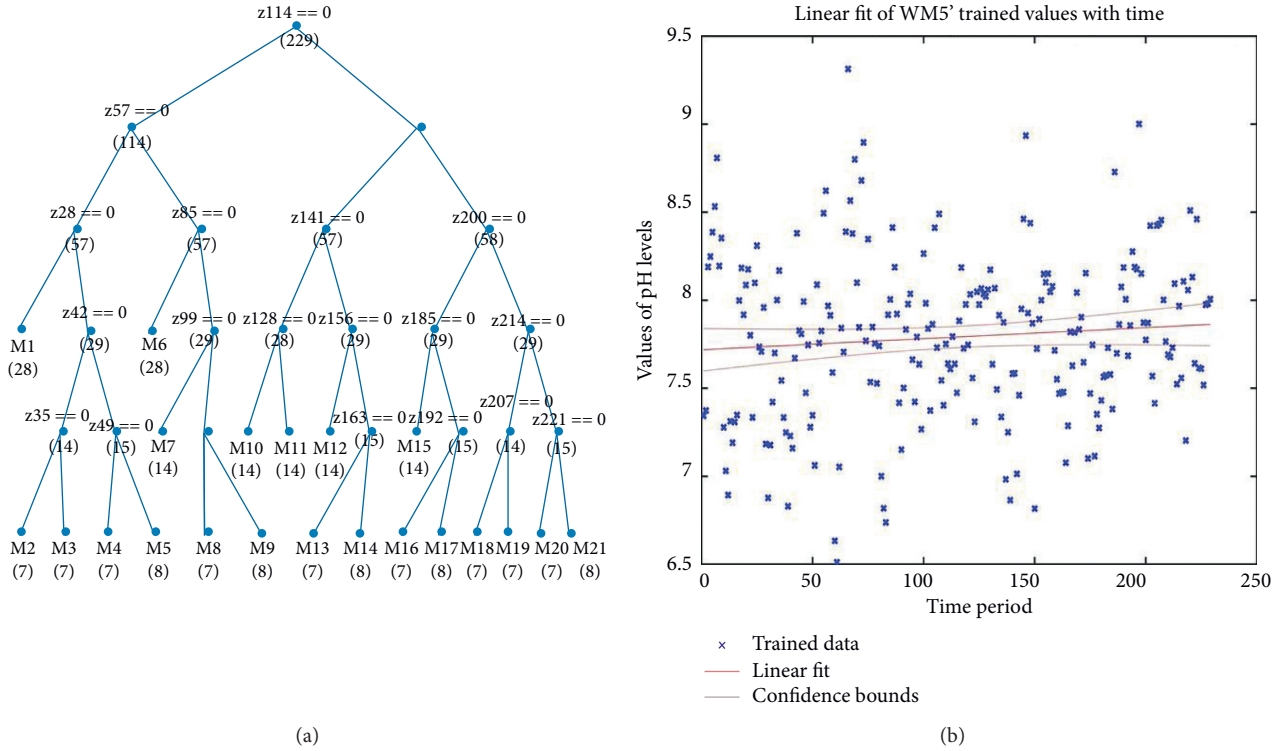


Figure 10: (a) WM5pRT regression tree of pH values at Palla. (b) Linear regression fit of WM5pRT trained pH values at Palla.

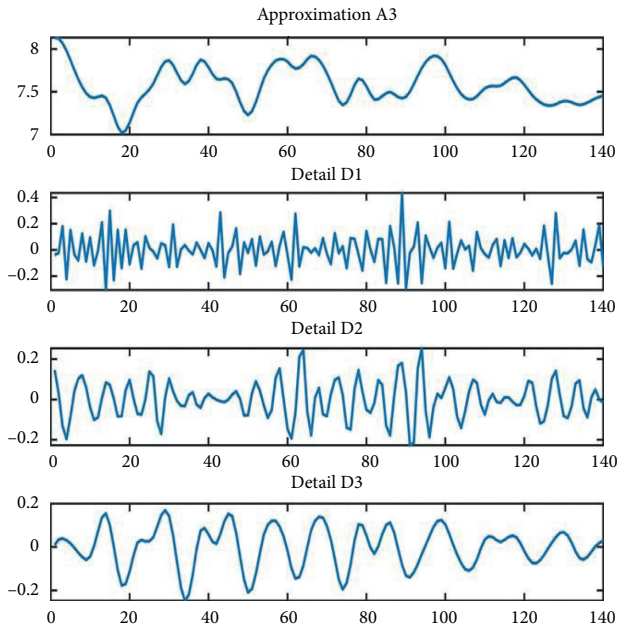


Figure 11: Wavelet decomposition for pH values at Bridge.

Figures 13(a) and 13(b) clearly demonstrate the regression fitting through LSSVR and linear fit of LSSVR trained values separately. Here, LSSVR trains in the pH level range of 7.2 to above 8 upto 8.4. Figures 14(a) and 14(b) clearly validate regression fitting through M5pRT and linear fit of M5pRT trained values separately. Values concentrated for pH levels 7.6 to 7.8. Figures 15(a) and 15(b) demonstrate the regression fitting through WLSSVR and linear fit of WLSSVR trained values separately. WLSSVR captures most of the values concentrated around pH levels 7 to 8 even though outliers can be seen around 6.5–7 and 8–8.2 and decreases with time. Figures 16(a) and 16(b) clearly demonstrate regression fitting through WM5pRT and linear fit of WM5pRT trained values separately. Trains data and

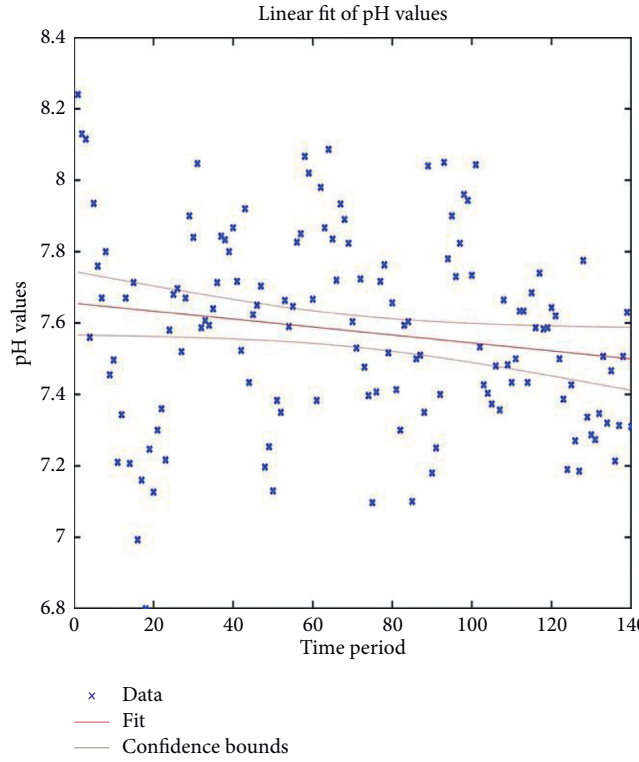


Figure 12: Linear regression fit of pH values at Bridge.

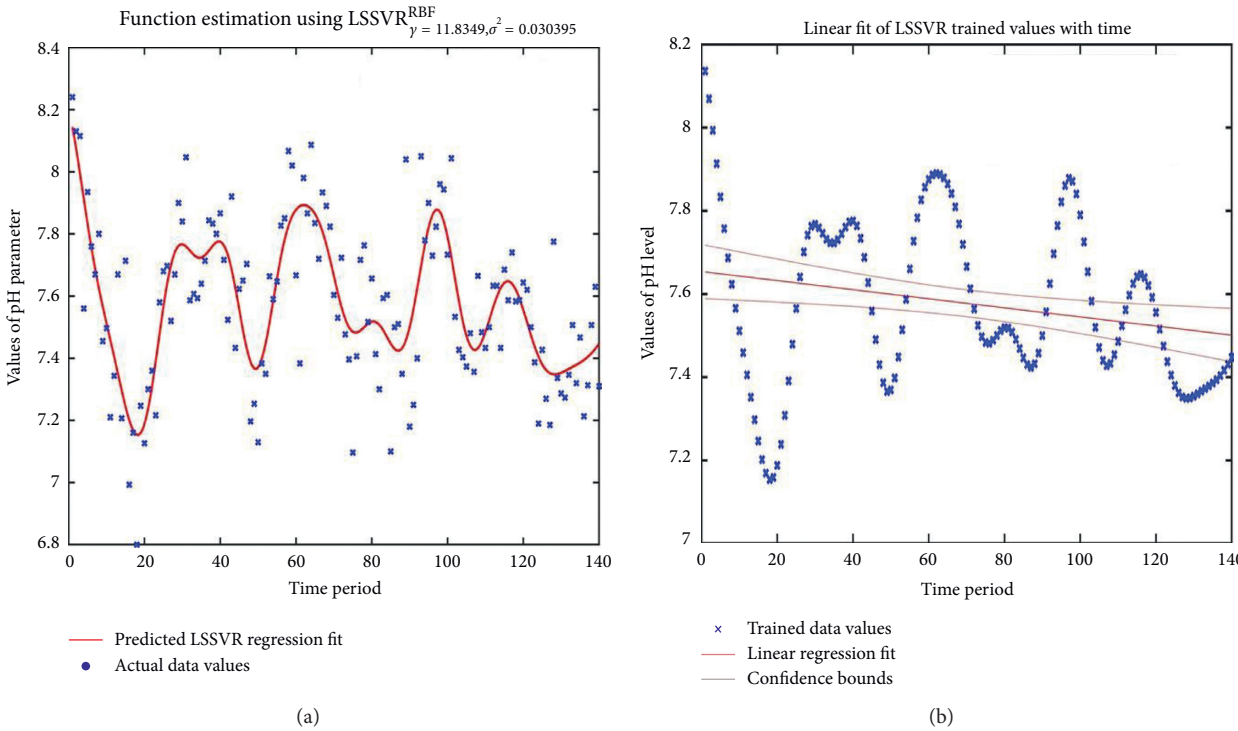


Figure 13: (a) LSSVR regression fit of pH data at Bridge. (b) Linear regression fit of LSSVR trained pH values at Bridge.

concentration can be observed from pH 7.2 to 8 as for every feature extracted, the parent node divides into child nodes, and thus tree gets created. Table 2 compares prediction errors RMSE, MSE, and MAE and also records goodness of fit (R^2) statistic for data recorded at two stations through different learning methods.

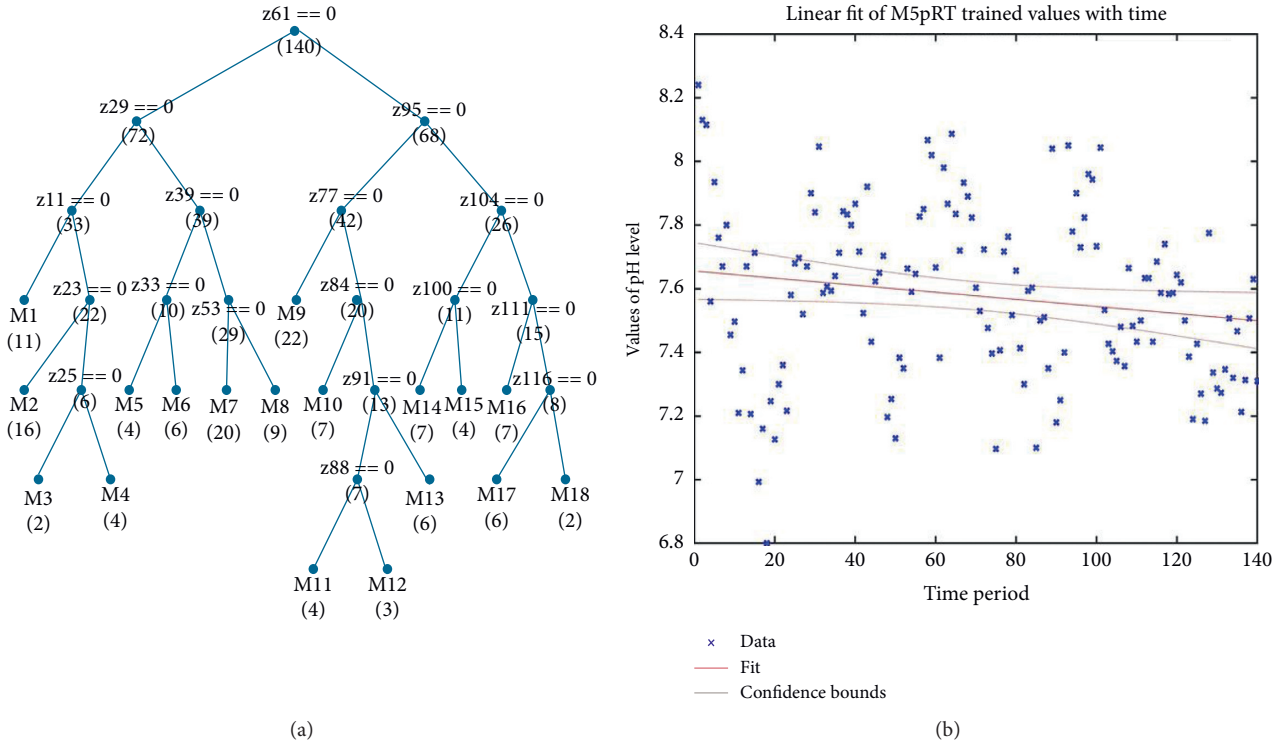


Figure 14: (a) Regression tree simulated via M5pRT at Bridge. (b) Linear regression fit of M5pRT trained data at Bridge.

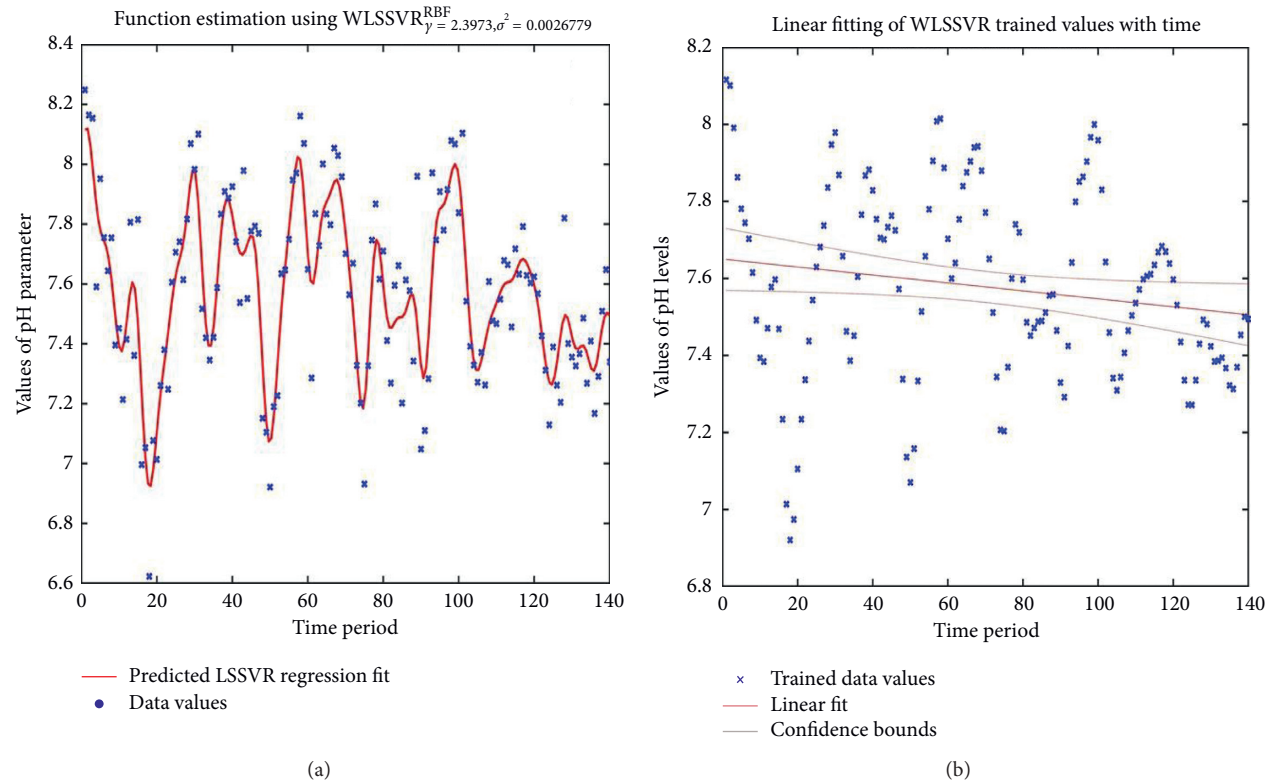
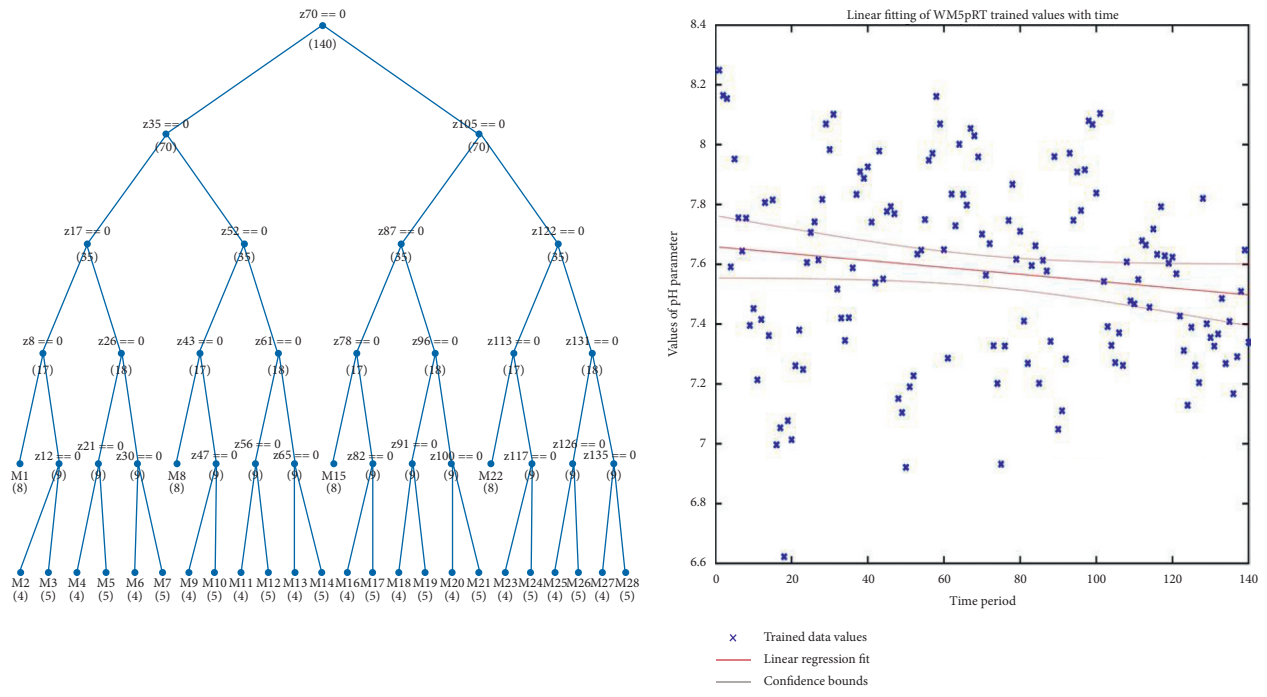


Figure 15: (a) WLSSVR regression fit of pH values at Bridge. (b) Linear regression fit of WLSSVR trained pH values at Bridge.

For LSSVR, RMSEs of Palla and Old Bridge are 7.7988 and 7.5796, respectively, which depict lesser error in simulation for Old Delhi Bridge. Similarly, MSE values are 60.8711 and 57.4510; MAEs are computed to be 7.7934 and 7.557 which shows lesser error in prediction for Old Delhi Bridge. Now, R-squared values are more accurate for Old



(a) (b)

Figure 16: (a) WM5p regression tree of pH values at Bridge. (b) Linear regression fit of WM5pRT trained pH values at Bridge.

Table 3: Various studies using different intelligent models.

Year	Author	Description	Parameters simulated	Results limited to
2014	Akrami SA. et al.	Rainfall data analyzing using moving average (MA) model and wavelet multiresolution intelligent model for noise evaluation to improve the forecasting accuracy	Wavelet transform (WT), moving average (MA)	RMSE and R^2 computed for MA at various levels of WT.
2018	Mahmoodabadi M. and Rezaei Arshad R.	Evaluated water quality parameters of the Karoun River using a regression approach and adaptive neuro-fuzzy inference system	Mann–Kendall regression, ANFIS	RMSE, MAE, and R^2 computed for the ANFIS model.
2019	Salazar L. et al.	Hourly ozone concentrations predicted using wavelets and ARIMA models	Haar discrete wavelet transform (HDWT), ARIMA	MSE and MSPE computed for ARIMA, HDWT, and combine model. The combined model performed better.
2019	Dehghani M. et al.	Predicted hydropower generation using the grey wolf optimization adaptive neuro-fuzzy inference system	ANFIS and GWO-ANFIS	RMSE, MAE, R^2, relative error, and confidence index computed for ANFIS and GWO-ANFIS. GWO-ANFIS was observed to be better.
2020	Seifi A. and Riahi H.	Estimated daily reference evapotranspiration using hybrid gamma test-least square support vector machine, gamma test-ANN, and gamma test-ANFIS models in an arid area of Iran	LSSVM, ANN, and ANFIS (all with gamma parameter)	RMSE, MAE, and R^2 computed for LSSVR, ANN, and ANFIS. LSSVR performed well overall.
2020	Present study Bhardwaj R., Bangia A.	Improved explicit prediction of river water quality using wavelet- based LSSVR and M5pRT	LSSVR, M5pRT, WLSSVR, WM5pRT	MSE, RMSE, MAE, and R^2 computed for LSSVR, M5pRT, WLSSVR, and WM5pRT. Wavelet conjuncted LSSVR and M5pRT observed to be better prediction models in our study.

Bridge as it would have detected lesser pollutants than at Palla while for WLSSVR, RMSEs of Palla and Bridge are 7.2990 and 7.5714, respectively; similarly, for MSEs; MAE values are 7.7919 and 7.5765. R^2 value at Old Bridge is 0.8759 more close to 1, i.e., WLSSVR predicts better responses for Old Bridge. Considering the M5pRT and WM5pRT model, all performance statistics have outcomes in favour of Palla station on the basis of performance as the data are training and validating through WM5pRT in comparison to M5PRT. It can be detected that MAE outcomes demonstrate lesser reduction compared with RMSE. Overall, MAE has lesser variation for the delta-error outputs through WLSSVR and WM5pRT. Thus, the proposed model, WM5pRT is good for estimation and simulation of the pH at Palla station whereas WLSSVR works better for Old Delhi Bridge station. Error tolerance set for all the learning prototypes is equivalent to $10^{(-4)}$ and 500 epochs of training. The forward and backward phase training depends on the termination condition which is when R^2 (goodness of fit) improvement converges below threshold. Table 3 tabulates findings of various authors that were referred in designing the algorithms for this study.

6. Conclusion

In this proposed study, forecasting pH level of Yamuna River provides improved accurateness on appointing decompositions of wavelets into approximations and details. Evaluation clarifies that the novel algorithm provides precise predictors for estimation. Nonlinearity of data is incompetent for training without the help of LSSVR and M5pRT. Thus, the decomposition of wavelet-form signals into details (D_1, D_2, and D_3) along with approximations' coefficients (A_1, A_2, and A_3) simulation has dynamic role in computation of the concentration of acidic/basic salt levels for river waters. It is observed that the anticipated prototype applying WLSSVR and WM5pRT is better than applying tools such as LSSVR and M5pRT, respectively, as the exactness grows. It can be determined that these additional wavelet layer in models filter-out disturbances while moderating computational-labour. Also, it is believed that the motivation to balance the pH level in the river basins is because there is high consumption of water for various chores and environmental natural processes. It can be better understood if the hydrological background having complexities is studied. These hybrid models with appropriate modifications can be used for training and predicting that would help estimating other water quality parameters such as BOD, DO, and COD at various other monitoring stations allocated by the designated authorities. It can be observed that it is best if water dependency on river basin can be reduced by shifting to rainwater harvestings and also to manage fresh surface-water, ground-water resources, and expansion of the concept of rainwater harvesting as a sustainable solution for future generations. For water quality changes that may happen due to unforeseen naturally occurring events or some man-made hazards, these models will have to be redesigned. Some modifications in the artificial networks will accommodate those factors in the layers of the hybrid models in the best possible manner to achieve optimal quality of the river water. This will help in consumption of clean and healthy drinking water and other purposes to humans and the vast ecosystem dependent on water.

Data Availability

Data were sourced from the Central Pollution Control Board (CPCB).

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank G.G.S. Indraprastha University for providing financial support and research facilities for this work.

References

[1] K. Chansaengkrachang, A. Luadsong, and N. Aschariyaphotha, "A study of the time lags of the Indian ocean dipole and rainfall over Thailand by using the cross wavelet analysis," *Arabian Journal for Science and Engineering*, vol. 40, no. 1, pp. 215–225, 2015.

[2] P. Mandal, R. Upadhyay, and A. Hasan, "Seasonal and spatial variation of Yamuna River water quality in Delhi, India," *Environmental Modeling & Assessment*, vol. 170, no. 1-4, pp. 661–670, 2010.

[3] Tiyasha, T. M. Tung, and Z. M. Yaseen, "A survey on river water quality modelling using artificial intelligence models: 2000-2020," *Journal of Hydrology*, vol. 585, Article ID 124670, 2020.

[4] S. A. Akrami, A. El-Shafie, M. Naseri, and C. A. G. Santos, "Rainfall data analyzing using moving average (MA) model and wavelet multi-resolution intelligent model for noise evaluation to improve the forecasting accuracy," *Neural Computing and Applications*, vol. 25, no. 7-8, pp. 1853–1861, 2014.

[5] A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," *Journal of King Saud University—Engineering Sciences*, vol. 29, no. 3, pp. 237–243, 2017.

[6] R. Bhardwaj and A. Bangia, "Dynamic indicator for the prediction of atmospheric pollutants," *Asian Journal of Water, Environment and Pollution*, vol. 16, no. 4, pp. 39–50, 2019.

[7] A. Bangia, R. Bhardwaj, and K. V. Jayakumar, "River water quality estimation through artificial intelligence conjuncted with wavelet decomposition," in *Advances in Intelligent Systems and Computing (AISC) Volume 979. Numerical Optimization in Engineering and Sciences*, J. Kacprzyk, D. Dutta, and B. Mahanty, Eds., pp. 159–166, Springer, Berlin, Germany, 2020.

[8] K. D. Brabanter, *Least Square Support Vector Regression with Applications to Large-Scale Data: A Statistical Approach*, Katholieke Universiteit Leuven, Belgium, Europe, 2011.

[9] G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*, Holden Day, San Fran-Cisco, CA, USA, 1976.
[10] O. Bozorg-Haddad, M. Zarezadeh-Mehrizi, M. Abdi-Dehkordi, H. A. Loáiciga, and M. A. Mariño, "A self-tuning ANN model for simulation and forecasting of surface flows," *Water Resources Management*, vol. 30, no. 9, pp. 2907–2929, 2016.
[11] M. Dehghani, H. Riahi-Madvar, F. Hooshyaripor et al., "Prediction of hydropower generation using grey wolf optimization adaptive neuro-fuzzy inference system," *Energies*, vol. 12, no. 2, p. 289, 2019.
[12] M. E. Doyle and V. R. Barros, "Attribution of the river flow growth in the Plata basin," *International Journal of Climatology*, vol. 31, no. 15, pp. 2234–2248, 2011.
[13] A.-M. Dunca, "Water pollution and water quality assessment of major transboundary rivers from banat (Romania)," *Journal of Chemistry*, vol. 2018, Article ID 9073763, , 2018.
[14] M. Farzadkhoo, A. Keshavarzi, H. Hamidifar, and M. Javan, "Sudden pollutant discharge in vegetated compound meandering rivers," *Catena*, vol. 182, Article ID 104155, 2019.
[15] P. K. d. M. M. Freire, C. A. G. Santos, and G. B. L. d. Silva, "Analysis of the use of discrete wavelet transforms coupled with ANN for short-term streamflow forecasting," *Applied Soft Computing*, vol. 80, pp. 494–505, 2019.
[16] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*, PWS Publishing Company, Boston, MA, USA, 1996.
[17] G. Hanrahan, *Artificial Neural Network in Biological and Environmental Analysis*, Taylor and Francis Group, New York, NY, USA, 2011.
[18] C. Jeong, J.-Y. Shin, T. Kim, and J.-H. Heo, "Monthly precipitation forecasting with a neuro-fuzzy model," *Water Resources Management*, vol. 26, no. 15, pp. 4467–4483, 2012.
[19] J. Liu, C. Yu, Z. Hu et al., "Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network," *IEEE Access*, vol. 8, pp. 24784–24798, 2020.
[20] R. Mahmoodabadi and R. Rezaei Arshad, "Long-term evaluation of water quality parameters of the Karoun River using a regression approach and the adaptive neuro-fuzzy inference system," *Marine Pollution Bulletin*, vol. 126, pp. 372–380, 2018.
[21] S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction," *Environmental Earth Sciences*, vol. 71, no. 7, pp. 3147–3160, 2014.
[22] R. Memarzadeh, H. Ghayoumi Zadeh, M. Dehghani, H. Riahi-Madvar, A. Seifi, and S. M. Mortazavi, "A novel equation for longitudinal dispersion coefficient prediction based on the hybrid of SSMD and whale optimization algorithm," *Science of The Total Environment*, vol. 716, p. 137007, 2020.
[23] C. Min, "An improved recurrent support vector regression algorithm for water quality prediction," *Journal of Computational Information*, vol. 12, pp. 4455–4462, 2011.
[24] H. Riahi-Madvar and A. Seifi, "Uncertainty analysis in bed load transport prediction of gravel bed rivers by ANN and ANFIS," *Arabian Journal of Geosciences*, vol. 11, no. 21, p. 688, 2018.
[25] H. Riahi-Madvar, M. Dehghani, A. Seifi, and V. P. Singh, "Pareto optimal multigene genetic programming for prediction of longitudinal dispersion coefficient," *Water Resources Management*, vol. 33, no. 3, pp. 905–921, 2019.
[26] R. R. Sahay and A. Srivastava, "Predicting monsoon floods in rivers embedding wavelet transform, genetic algorithm and neural network," *Water Resources Management*, vol. 28, no. 2, pp. 301–317, 2014.
[27] M. Sakizadeh, "Artificial intelligence for the prediction of water quality index in groundwater systems," *Modeling Earth Systems and Environment*, vol. 2, no. 1, p. 8, 2016.
[28] L. Salazar, O. Nicolis, F. Ruggeri, J. Kiseľák, and M. Stehlík, "Predicting hourly ozone concentrations using wavelets and ARIMA models," *Neural Computing and Applications*, vol. 31, no. 8, pp. 4331–4340, 2019.
[29] H. Riahi and H. Riahi-Madvar, "Estimating daily reference evapotranspiration using hybrid gamma test-least square support vector machine, gamma test-ANN, and gamma test-ANFIS models in an arid area of Iran," *Journal of Water and Climate Change*, vol. 11, no. 1, pp. 217–240, 2020.
[30] G. Zuo, J. Luo, N. Wang, Y. Lian, and X. He, "Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting," *Journal of Hydrology*, vol. 585, Article ID 124776, 2020.