The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

Hybrid Algorithm for early Detection of Water Pollution Impact on Environmental Indicators using Wavelet Techniques and RBF Neural Network Learning

Monireh Khayat[1]✉ | Rassoul Noorossana[2] | Paria Soleimani[1] | Sadigh Raissi[1]

1. Industrial Engineering Department, South Tehran Branch, Islamic Azad University, P.O. Box: 15847-43311, Tehran, Iran
2. Industrial Engineering Department, Iran University of Science and Technology, P.O.Box: 13114-16846, Tehran, Iran

Article Info

Article type:
Research Article

Article history:
Received: 2 February 2024
Revised: 27 May 2024
Accepted: 27 August 2024

Keywords:
Water Pollution, Water Quality, Environment, Anomaly, Neural Network, Wavelet, Time-Frequency Series

ABSTRACT

The present study examines the impact of water pollution on the environment with the aim of detecting early abnormalities or significant changes in the water pollution indicators. A hybrid algorithm based on wavelet techniques and radial basis function neural network learning using high-frequency surrogate relation is introduced. Important qualitative indicators such as phosphate, nitrate, and chemical oxygen demand (COD) in the water bodies have uncertainties with variations such as dependence, and effectiveness of physical and chemical factors. In the first step, the high-frequency time series of the main TP index is obtained through the surrogate model and compared with GARCH techniques. By using the wavelet transform, the noise components of the time series are removed and pre-processed. In the next step, it is created by using the neural network to identify the main characteristics of water quality. In the last step, the contamination threshold is calculated based on the estimated base pattern for analyzing statistical patterns. The results show that the proposed algorithm has high stability and accuracy because using the surrogate technique has extracted a more accurate model of the behavior of the required water variables. It can be used to manage surface runoff in watersheds to preserve the environment and improve water quality.

Cite this article: Khayat, M., Noorossana, R., Soleimani, P., & Raissi, S. (2024). Hybrid Algorithm for early Detection of Water Pollution Impact on Environmental Indicators using Wavelet Techniques and RBF Neural Network Learning. *Pollution*, 10 (4), 1074-1091.
https://doi.org/10.22059/poll.2024.372064.2246

DOI: https://doi.org/10.22059/poll.2024.372064.2246

INTRODUCTION

Water pollution is a critical environmental issue with profound impacts on ecosystems and human well-being. Directly affecting aquatic life, pollutants such as chemicals, heavy metals, and organic substances can compromise water quality and aquatic ecosystems. This pollution also extends its influence to food production, contributing to a decrease in agricultural yields due to the degradation of water sources. Human health is at risk through the consumption of contaminated water or food products. Furthermore, water pollution has detrimental effects on biodiversity, causing habitat destruction and alterations in ecosystems (Noori, 2017). Thus, addressing the commitment to preserving water resources and reducing water pollution is crucial for enhancing environmental sustainability and community well-being.

Along with the growth of the population in the world and development in various industries, water pollution has become one of the most fundamental problems in the world (Naderian *et al*. 2024). Not only the water shortage crisis threatens the countries, but all water sources are

*Corresponding Author Email: *st_m_khayat@azad.ac.ir*

exposed to the dangers of pollution. The environmental problems of water and the introduction of pollution into the human food chain cause thousands of people to suffer from social health diseases every year and cause billions of dollars in damage to the national economy and the development of countries around the world (Talebi, 2023; Azis & Alfarizi, 2023). The issue of uncertainty (Noori & Mirchi, 2021) and the probability of exceeding the standards of qualitative indicators from the set standard limits is a serious threat to the health of human life, fauna and flora, and water species.

Anomaly detection is the process of pattern recognition that is unexpected compared to normal behavior in observations (El-Shafeiy *et al*., 2023). Therefore, providing analyzer algorithms and predictor models to detect pollution or predict the quality status of drinking water sources is one of the challenges for researchers in the field of water quality system studies.

Water pollution is the major environmental impact affecting the study area. Among the various components of the environment, water is more related to other components of the environment due to its extensive cycle in nature. In this cycle, the water interacts with soil and vegetation and non-observance of environmental standards by humans can cause irreparable damage to the environment. Therefore, it is necessary to prevent damage to the environment by timely detection of water quality abnormalities.

This paper tries to introduce a structure based on data-driven to detect water quality anomalies by using neural network approaches. The proposed algorithm has used a new technique in data pre-processing due to the possibility of sensor noise or measurement noise entering the feature extraction section with better-quality data. Also, surrogate models have been used so that a more accurate model of the behavior of water requirement variables can be extracted. In the following sections, we will first introduce the study area, the Potomac River, then collect the river data, and perform pre-processing through the wavelet method. In the following, we focus on extracting the features and finally, classify through the neural network and present results.

MATERIALS AND METHODS

Subject: Potomac River

Potomac River with a length of about 405 miles (652 km), a catchment area of 14,700 square miles (38,000 Km^2), and an average discharge rate of 1.269 m^3 / s (US EPA, 2017). The expansion of the tributaries of this river in the vicinity of important adjacent has caused irreparable damages and losses to the environment. This issue has caused environmentalists to worry about the destructive environmental effects of EIA mining projects, including threats to drinking water.

In this research, the input data were randomly collected from two stations located in the upstream and downstream sub-branches. Station number 0163900 is located at Smith Greek in the northern branches of the river and station number 01646305 is located at Lower Man stream-MacLean (Figure (1)) (Byrand, 2010). The data collected from the first station was used to build the baseline of the proposed anomaly detection algorithm, and the data set from the second station was used to validate the proposed model.

Data acquisition: Surrogate measurement method

Surrogate models, in fact, are the discovery and expression of the relationship between the input and output variables of observations with a statistical approach. In many cases, qualitative or quantitative indicators in watersheds cannot be directly measured or monitored due to their dynamic nature or limitations in measuring equipment and sensors (Zhang *et al*., 2022). Surrogate measurements are appropriate for water quality variables that cannot be measured directly in situ or that are difficult to measure from high-resolution online observations (TP, TN, COD, etc.). High-frequency, in situ water quality monitoring data, can capture characteristic trends and

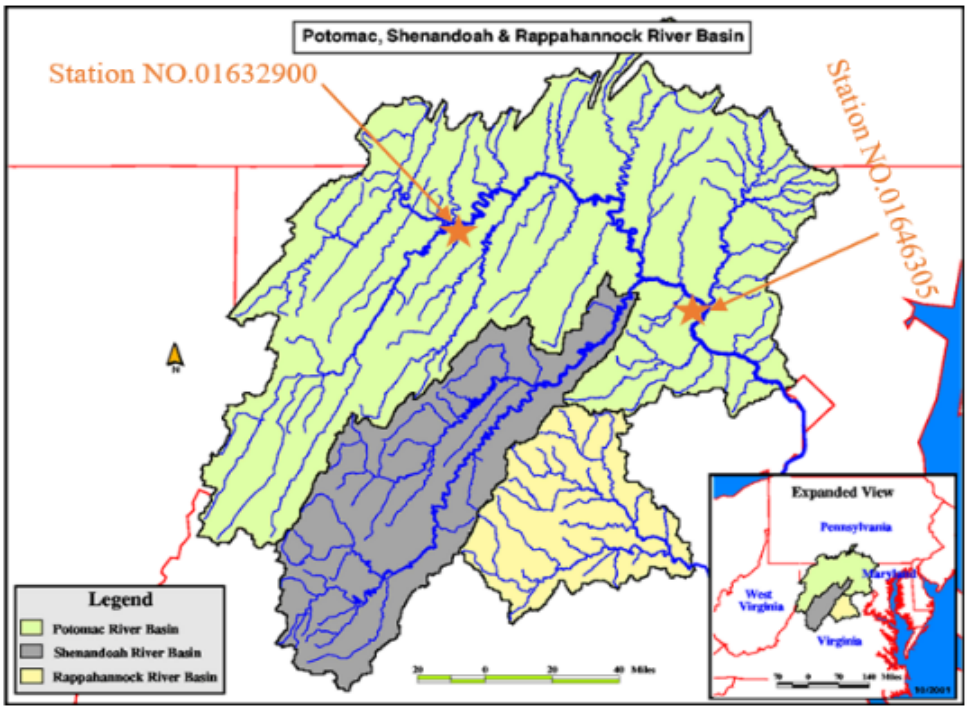

Fig. 1. Location of major river basins and the river input monitoring stations

Table 1. Parameters and formulas based on criteria such as MSE, and RMAE

	Index evaluation	Description of the formula	
1	MSE	$MSE = \sqrt[2]{\frac{\sum_i^m [y_i - \hat{y}_i]^2}{n-2}}$	
2	R^2	$R^2 = 1 - sse/ssy$	$\boldsymbol{sse} = \sum_i^m [\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i]^2, \boldsymbol{ssy} = \sum_i^m (\boldsymbol{y_i} - \bar{\boldsymbol{y}}_i)^2$
3	$RMAE$	$RMAE = \frac{1/n \cdot \sum_i^m \lvert y_i - \hat{y}_i \rvert 2}{\bar{y}_i} . 100$	

periods overlooked by traditional periodic sampling and is an emerging area of research (Kunz *et al*., 2017). As an example, suspending solids are a major contributor to unfiltered TN and TP in natural surface waters (Shi *et al*., 2018). Turbidity, the degree of relative clarity of a liquid, which is affected by suspended solids, has been used as an alternative measure of suspended solids concentration (Kuo *et al*., 2006). Specific conductivity is related to the concentration of TN and TP, which can affect the deposition and absorption of nitrogen and phosphorus. In addition, dissolved nitrogen and phosphorus are closely related to pH and water temperature. Therefore, TN and TP concentrations are related to water temperature (Christensen, 2001).

$$\widehat{y_i} = mx_i + b + e_i \qquad i = 1, 2, \ldots n \tag{1}$$

$$e_i = y_i - \widehat{y_i} \tag{2}$$

In equation 1, y_i is the *i*-th observation and *m* is the slope of the predictive equation, and

e_i is the error between the actual value of the *i*-th observation and the predicted value. The parameters and quality of the model are based on various statistical indicators such as the minimum squared error MSE, coefficient of determination R^2, squared mean squared error, and RMAE are estimated (Saghafi *et al*. 2009). Equation 2 represents a simple model of surrogates. In this equation, $\hat{y}$ is the dependent variable or response and the coefficients b_0 , b_1 are the constant value and slope of the line, respectively.

In addition, certain surrogate variables have significant relationships with the response variable. Relationships between surrogate measurements and water quality constituent concentrations are also site-specific and must be developed for each sensor node. Therefore, these relationships vary in different watersheds or locations.

Preprocessing: Wavelet Denoising

Wavelet technologies have been widely used for data-driven hydrological and environmental modeling (Erkyihun *et al*., 2016). Wavelet analysis can be used to transform an original time series into components (major trend (low-frequency parts) and noise (high-frequency parts)). In this study, the low-frequency part contains the sub-hourly variations in surrogate water quality time series (Shupe, 2017). The high-frequency part contains noise caused by measurement errors or the surrounding environment. The original product of water quality monitoring *S(i)* is a one-dimensional discrete time series expressed as follows:

$$S(i) = f(t) + \sigma \times e(t) \qquad t = 0,\ldots,n-1 \tag{3}$$

Where *f(t)* is the real-time series, *e(t)* is the time series of noise, *i* is the sampling time, and σ is the coefficient of noise (standard deviation). During wavelet denoising, the standard deviation can be replaced with 0. The corresponding discrete wavelet function can be expressed as follows:

$$\psi_{j,k}(t) = a_0^{\frac{-j}{2}} \psi\left(a_0^{-j} t - k b_0\right) \tag{4}$$

where a_0^j is the scaling factor, b_0 is the time factor, *t* is time, and *j* is an integer that represents the element number of the time series. The wavelet coefficients of discretization can be expressed as follows:

$$C_{j,k} = \int_{-\infty}^{\infty} f(t) \psi_{j,k}^{*}(t)\, dt \tag{5}$$

Additionally, the wavelet reconstruction formula can be expressed in the following form:

$$f(t) = C \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} C_{j,k} \psi_{j,k}(t) \tag{6}$$

where *C* is a constant. Moreover, $e(t) = C \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} C_{j,k} \psi_{j,k}(t)$.

Donoho's denoising technique

Donoho suggested a technique for denoising that attempts to remove responses by thresholding the wavelet coefficients at level *i* of the signal *X(i,k)* as shown by (Donoho & Johnstone, 1994). For each scale *i*, a threshold T_i is gained by equation (7):

$$T_i = \sigma_i \sqrt{2lnN} \tag{7}$$

Where N is the number of wavelet coefficients and σ_i is calculated by equation (8):

$$\sigma_i = Median\left\{\left|X_{i,1} - \tilde{X}_i\right|, \left|X_{i,2} - \tilde{X}_i\right|, \ldots\ldots, \left|X_{i,k} - \tilde{X}_i\right|\right\} / 0.6745 \tag{8}$$

Denoising is performed by hard thresholding of coefficients as given in equation (9):

$$X_{den}(i,k) = \begin{cases} X(i,k) & if \ \left|X_{i,k}\right| > T_i \\ 0 & if \ \left|X_{i,k}\right| > T_i \end{cases} \tag{9}$$

By applying the threshold T_i on the wavelet coefficients, the denoised coefficients X_{den} were gained. The denoised signals were reconstructed from these coefficients. The general preprocessing procedures are shown in Figure 3.

Feature extraction

The objective features for discriminating the responses to water quality such as the amount of estimated phosphate concentration, the flow rate of the river, and the amount of changes in water turbidity in the studied period are extracted in the time, frequency, and time-frequency domains. The observations extracted from the 60-day time interval in the form of online data from the

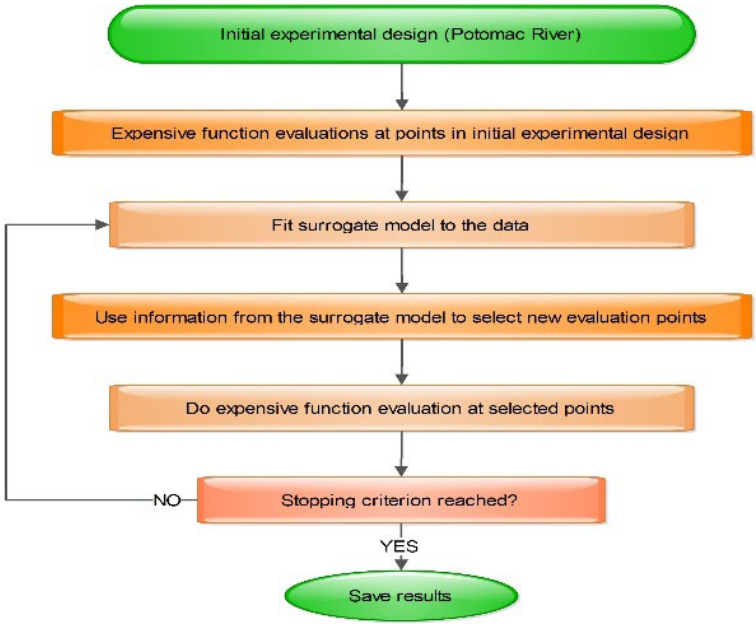

Fig. 2. Steps of a surrogate model algorithm

USGS site include the values of all three independent variables, TURB, WT, and Sc, which are shown in Figures 4, 5, and 6. As can be seen in Figure 4, the phosphate concentration has a peak in the 3549th observation, which corresponds to February 6, and the 3769th observation, which corresponds to February 9. Corresponding to these dates, on February 6, the river's discharge reached its highest value in the entire period, i.e., 43,000 cubic feet per second. Rainfall is one of the main factors in increasing the volume of the river flow, and subsequently, with the transport and washing of the salts in the soil of the lands adjacent to the river, indicators such as turbidity, which affect phosphate changes, also increase. However,

in the 2739th observation, which belonged to January 29 at 9:45 am, at the same time, other factors such as the river discharge was about 5770 cubic feet per second and the turbidity was also about 2.5 units. This paper can be caused by an anomaly event in a part of our monitored

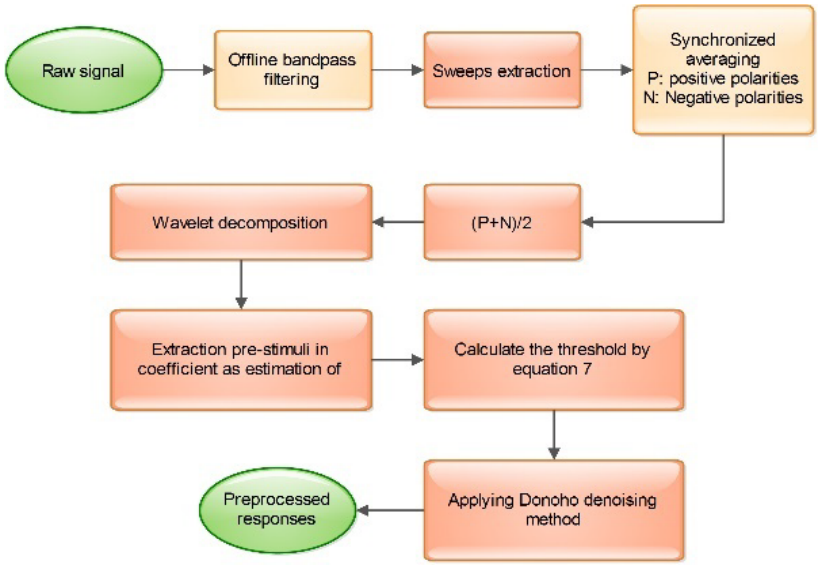

Fig. 3. The overall procedures of preprocessing

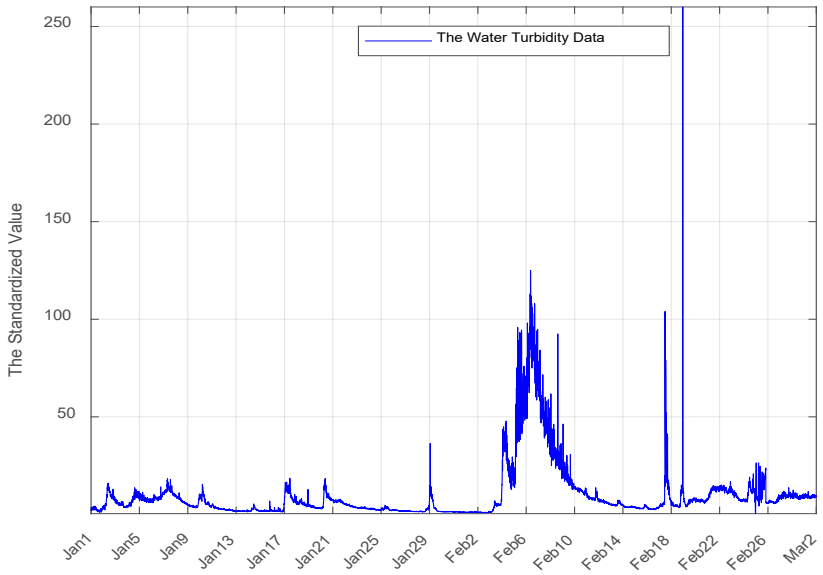

Fig. 4. The Water Turbidity (TURB) Data

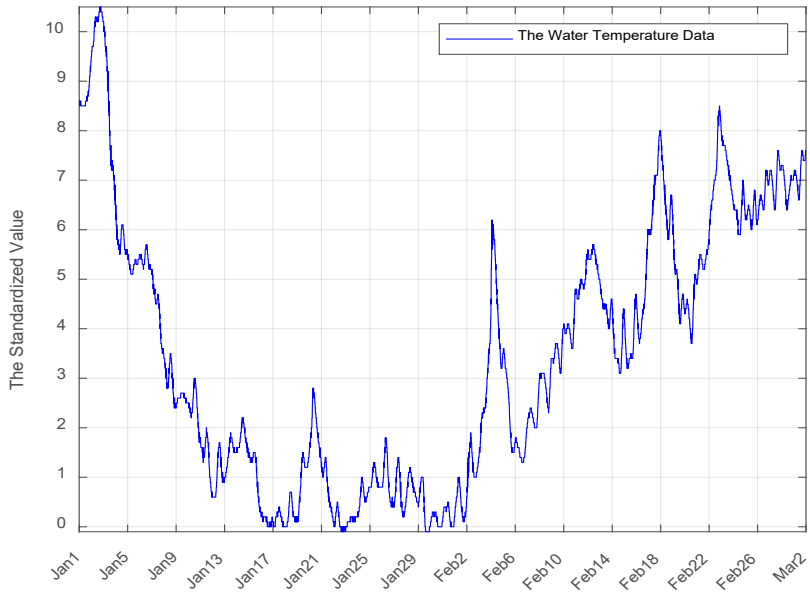

Fig. 5. The Water Temperature (WT) Data

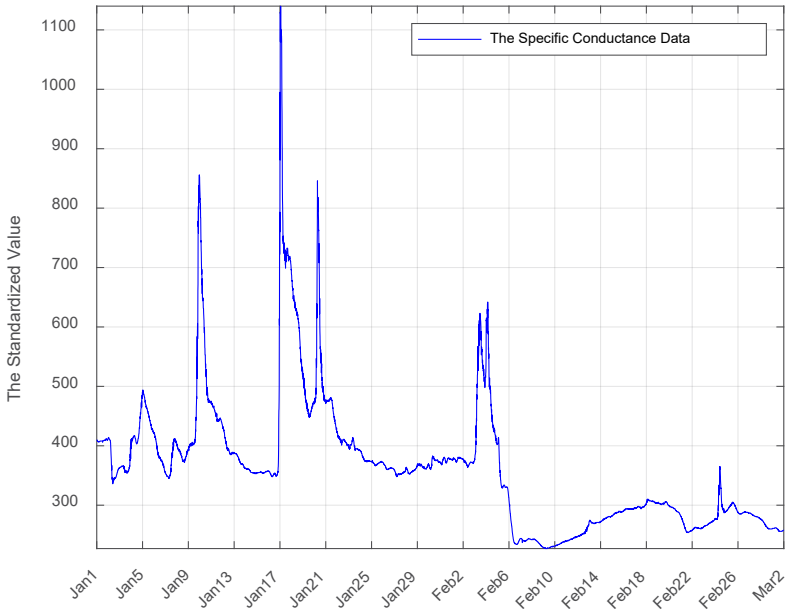

Fig. 6. The Specific Conductance (SC) Data

area for any reason. In the upcoming research, an attempt is made to distinguish between the nature of these two events so that environmental pollution can be prevented in the shortest possible time.

Time domain features

Time series refers to a sequence of observations of a phenomenon in regular time intervals (Kirchgässner *et al*., 2012). Analysis and modeling of time series are widely used in data-based algorithms that are designed to detect the anomaly of qualitative indicators (Chen *et al*., 2022; Lundgren, & Jung, 2022). Usually, it is not easy to find a suitable time series model to fit the collected data. One of the proposed strategies in the process of modeling the appropriate

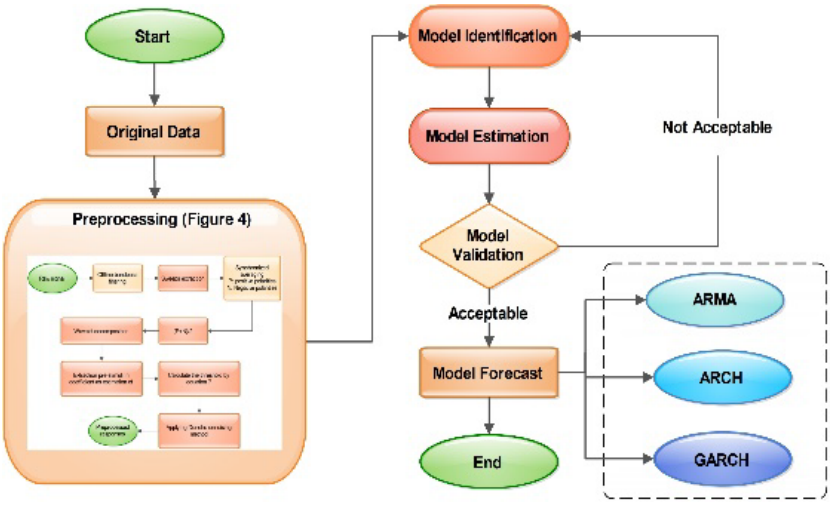

Fig. 7. Flowchart of time series methodology

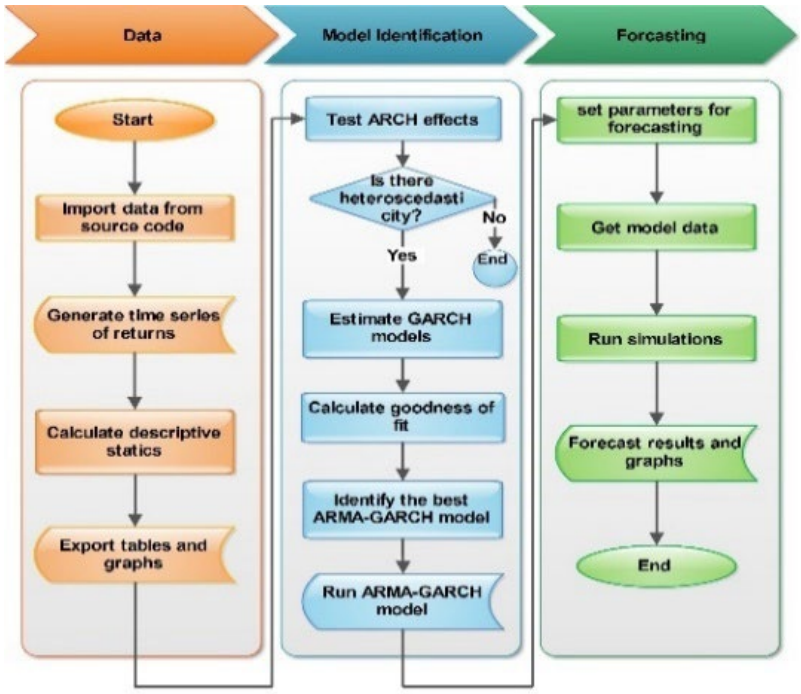

Fig. 8. Flowchart of GARCH modeling

time series is the method (Karasu & Altan, 2022; Jin *et al*., 2022). This method includes the following 3 steps.

1. **Identification:** In this step, keeping in mind the principle of parsimony, we select the appropriate model for the series of observations from among the available categories.

2. **Parameter Estimation:** In this step, we estimate the parameters of the model based on criteria such as the least squared error or the maximum precision.

3. Model Diagnostic Checking: Finally, in this section, the level of satisfaction resulting from the model fitting is measured. Figure 7, shows the type of time-series mythology. In the field of water resources management, instability and uncertainty of variables are very important. Therefore, autoregressive models with unstable conditional variance such as ARCH,

and developed models such as GARCH can be effective (Chen *et al*., 2022).

ARMA model: It is a linear model for time series that was proposed in (Makridakis & Hibon, 1997). Equation (10) represents an autoregressive-moving average process and includes two parts: past observations and past prediction errors.

$$ARMA(p,q): Y_t = c + \sum_{i=1}^{p} \varphi_i . Y_{t-i} + \sum_{i=1}^{q} \theta_i . \epsilon_{t\text{-}i} + \epsilon_t \quad (10)$$

Y_t is a variable that must be estimated based on previous observations, and the coefficients φ_i, θ_i, and c are model parameters, and p, q are the return order of observations to the previous steps. ϵ_t is a random sequence of error sentences with zero mean and constant variance.

In equation (11), ∇x_t is the regression operator, which, in the form of a differential equation, takes the observations one or more steps back, and as a result, a new series is formed on the right side of the equation, the average of which will be close to zero and the variance will be constant.

$$\nabla y_t = y_t - y_{t-1} \quad (11)$$

$$y_t = \log y_t - \log y_{t-1}$$

ARCH model: ARCH model or autoregressive model with unstable conditional variance was first introduced by Engel in 1982(Tan *et al*., 2022). It can be said that y_t in relation (12) is an ARCH type (p) process if the variance of the data is a weighted linear combination of the terms of the second power of the observations in the past steps and $a_i > 0$; $b_i > 0$.

$$y_t = a_0 + \sum_{i=1}^{p} a_i . y_{t-i} + \epsilon_t \quad (12)$$

In the real world, non-linear phenomena such as water quality variables and the changes and strong dependence of these indicators on other environmental and non-environmental factors generally raise issues such as volatility and momentary changes in the variance of observations (Song & Yao, 2022).

GARCH model: The difference between this model and ARCH is in the modeling of conditional and unstable variances (Sun *et al*., 2010). In other words, it can be said that the observations in a non-linear phenomenon follow the GARCH model when their variance applies to equation (13).

$$\sigma_t^2 = a + \sum_{i=1}^{p} b_i y_{t-1}^2 + \ldots + \sum_{j=1}^{q} c_j \sigma_{t-j}^2 \quad if \quad a > 0, b_i \geq 0, c_j \geq 0 \quad (13)$$

In the proposed algorithm, GARCH, ARMA, and ARCH models will be implemented on the output data from the surrogate equation of TP and TN. Finally, the time series with the best performance for each will be selected as the input of the main model. Figures 9 and 10 show the training and testing data based on the GARCH time series model.

Frequency domain features

The frequency domain is a common source of distinguishing features because it can provide information about the energy present at different frequencies in the signal. One common technique is to use the Fourier transform or another frequency domain transform to obtain a set

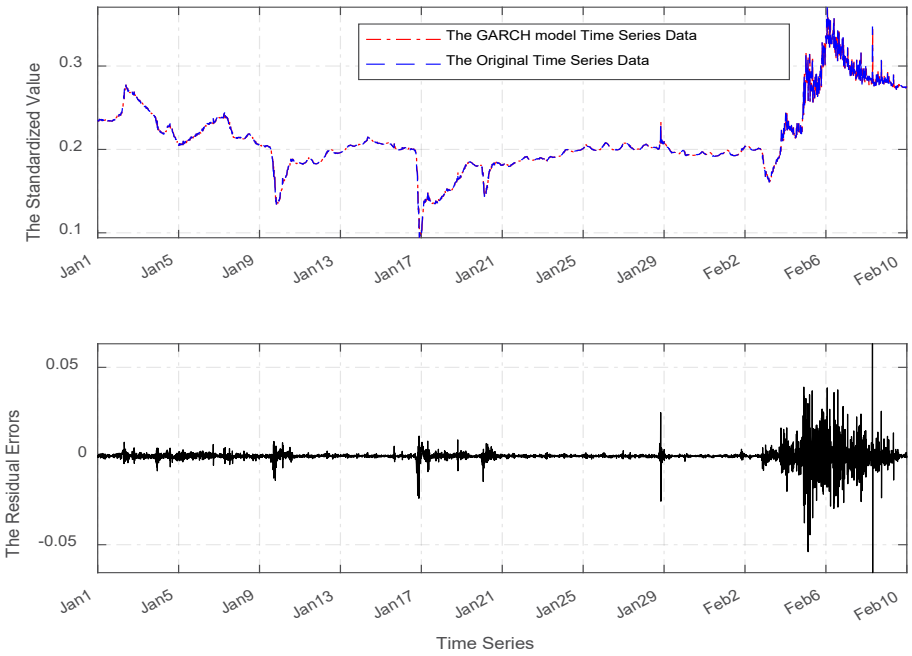

Fig. 9. Training of GARCH model time series

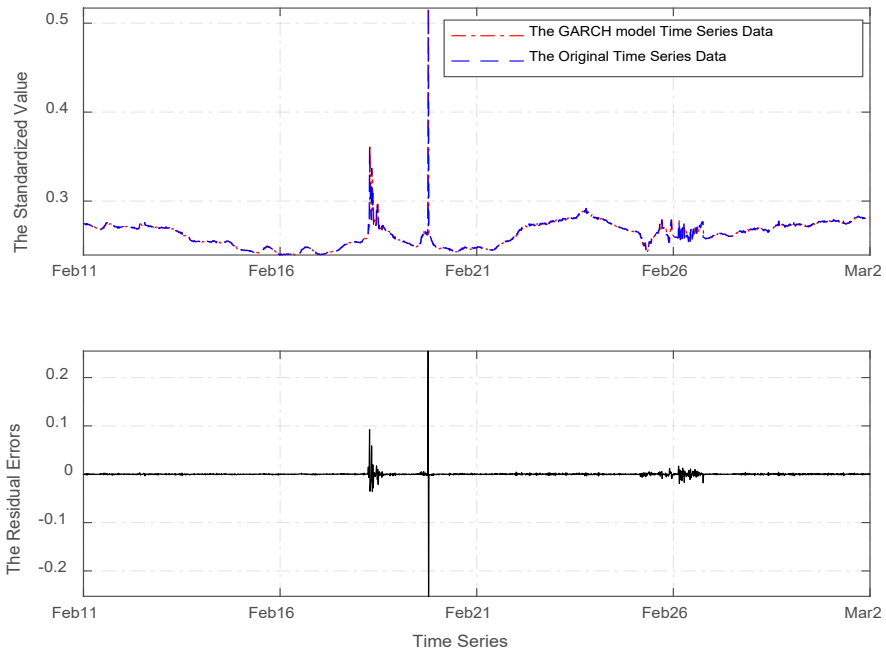

Fig. 10. Testing of GARCH model time series

of frequency components, and then apply feature selection or feature extraction algorithms to identify the most relevant features for a given classification task.

Time-Frequency domain features

The Fourier transform allows a distinctive transform of a signal from the time domain to the frequency domain. Several time-frequency analysis techniques have been developed and applied to anomaly diagnosis: e.g., short-time Fourier transform (STFT), wavelet transform

(WT), Hilbert-Huang transform (HHT), empirical mode decomposition (EMD), etc. The wavelet packet transform (WPT) is a time-frequency domain analysis approach that provides a great deal of freedom in analyzing non-stationary signals (Gokhale & Khanduja, 2010). WPT leads to a binary tree of orthonormal bases (the so-called nodes), which is shown in Figure 11.

Let us assume that A_{00} denotes the first parent node from WPT decomposition (i.e., the original signal). According to the representation of Figure 11, node A_{jk} is the kth node in the jth level of decomposition. The relation between the child ($A_{j+1\ 2k}$ and $A_{j+1\ 2k+1}$) and parent nodes (A_{jk}) is defined in Equation 14. This process continues until reaching the last level of decomposition.

$$A_{jk} = A_{j+1\ 2k} \oplus A_{j+1\ 2k+1} \tag{14}$$

The symbol $\oplus$ indicates the direct sum of two subspaces $A_{j+1\ 2k}$ and $A_{j+1\ 2k+1}$. The aim is to find the best subspaces that provide maximum dissimilarities between the three classes of responses. The detailed method proposed by this study is shown in Figure 12.

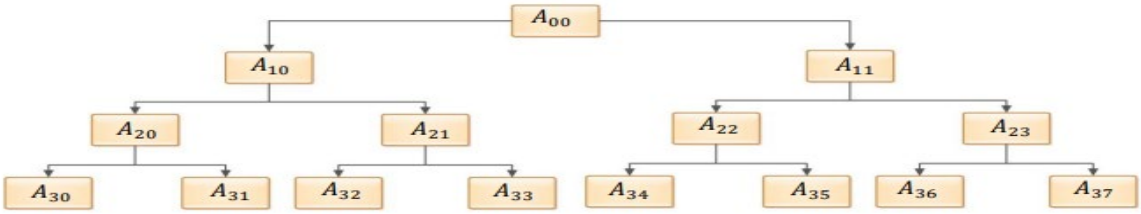

Fig. 11. Level 3 decomposition by using WPT

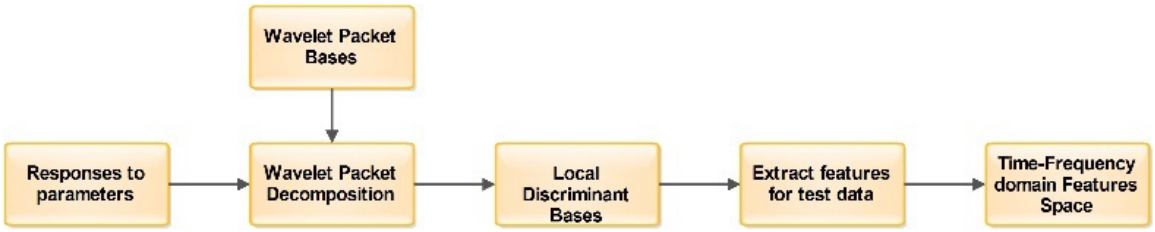

Fig. 12. The schematic representation of extracting time-frequency domain features

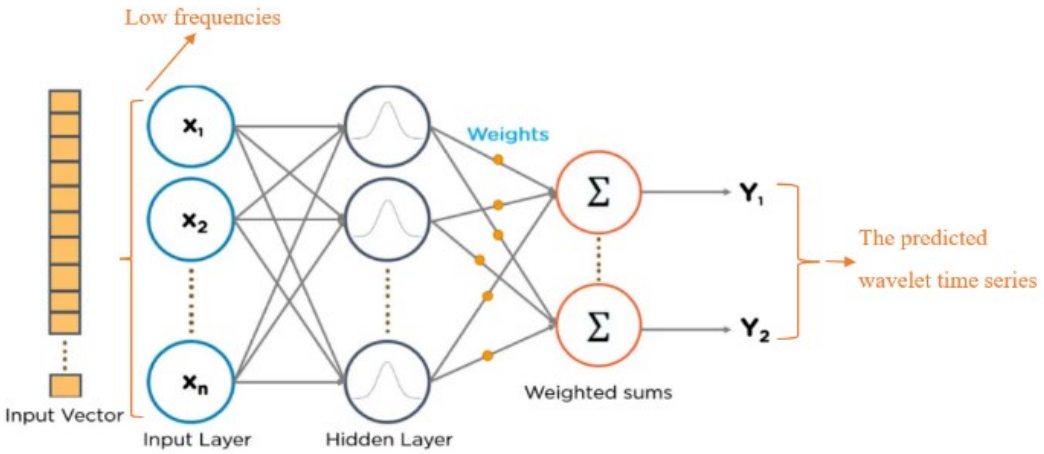

Fig. 13. RBF schematic

Classification

Classification neural networks used for feature categorization are very similar to anomaly-diagnosis networks, except that they only allow one output response for any input pattern, instead of allowing multiple faults to occur for a given set of operating conditions. In this paper, we used one of the algorithms that belong to the deep learning algorithms, RBFs, as they are a special type of feeder neural network that uses radial basis functions as activation functions.

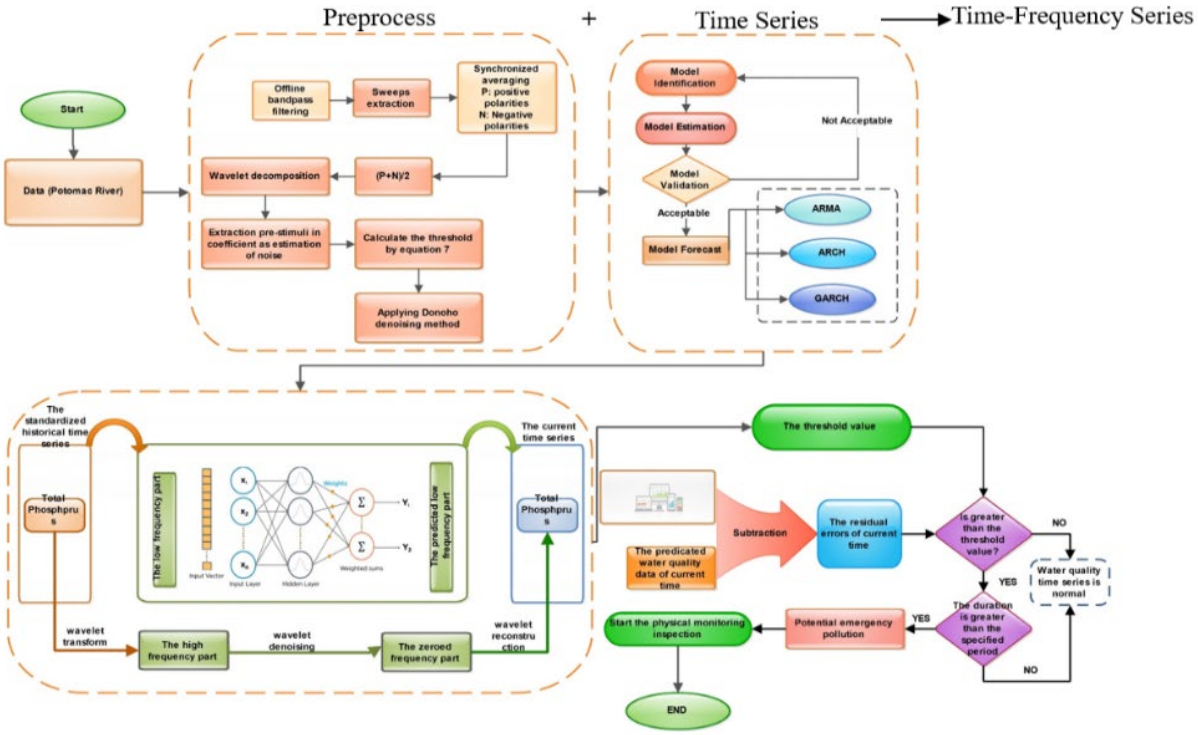

Fig. 14. Flow chart of the proposed combined approach of water quality anomaly detection

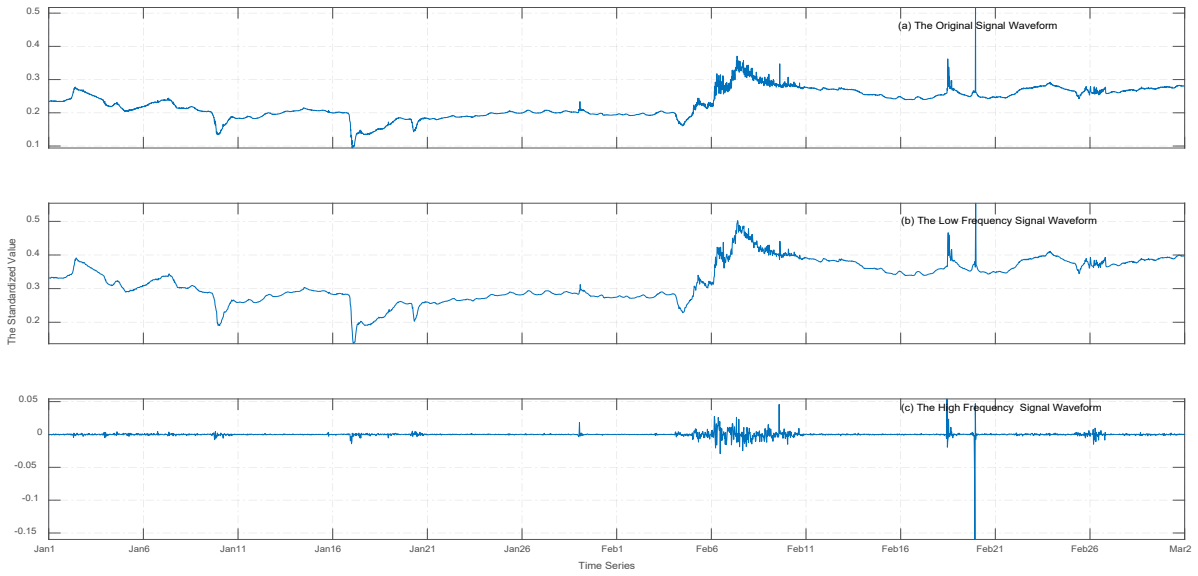

Fig. 15. Wavelet transform results for the original time series of surrogate TP

Radial basis function (RBF)

RBF networks are a common type of use in artificial neural networks for function approximation problems. RBF networks are distinguished from other neural networks due to their global approximation and fast learning speed. The main advantage of the RBF network is that it has only one hidden layer and uses the radial basis function as the activation function. Input vectors that are more similar to the prototype return a result closer to 1. There are different possible choices of similarity functions, but the most popular is based on the Gaussian. Equation (15) is the equation for a Gaussian with a one-dimensional input.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{15}$$

Where x is the input, μ is the mean, and σ is the standard deviation. Given an observation x, and using K neurons, the output is gained by Equation 16:

$$y = \phi(x) = \sum_{i=1}^{K} \omega_i \exp\left(-\gamma x - \mu_i^2\right) \tag{16}$$

Where ω_i is the weight of the ith neuron and μ_i is the center vector of the same neuron. Given the Gaussian decay, changing the parameters of one neuron has only a small effect on input values that are far away from the center of that neuron.

VALIDATION AND EVALUATION

Step 1: description of sampling and data collection: In the upcoming study, the continuous sampling in 15-minute time intervals, the results of measuring the variables are available at the measuring station number 01646500 (sensors and equipment) installed in this section, 18 parameters are calibrated with the help of sondes and calibrated probes, have registered. Considering that the focus of this study is on monitoring and anomaly detection from the perspective of phosphate in surface water, only key factors such as turbidity, specific conductance, and temperature are taken from the site.

Step 2: extraction of phosphate concentration data based on surrogate relationship: In real environments and when faced with the analysis of qualitative issues, the surrogate estimation method can significantly contribute to the development and flexibility of the estimator model. The concentration of the phosphate index as a highly variable and highly correlated variable that cannot be directly measured has a significant relationship with the turbidity factor, temperature, and specific electrical conductivity.

$$TP = 0.00103TUR + 0.0057WT - \log_{10} SC + 0.776 \tag{17}$$

The interpretation of equation 17 shows that the coefficients predicting phosphate changes in equation 17 with a confidence factor of 96%, R^2, were correct and significant variables. The observations extracted from the time interval of 60 days in the form of online data from the (Gokhale & Khanduja, 2010; Jastram, 2014) include the values of all three independent variables TURB, WT, Sc. Figures 4, 5, and 6 show the estimated phosphate concentration, the trend of the river discharge volume, and the amount of water turbidity changes in the studied period.

Step 3: Time series modeling: Measurement and collection of observations related to the date of January 1, 2022, to March 2, 2022, and for each day in 15-minute intervals, 96 observations for each variable such as turbidity, electrical conductivity, temperature, etc. have been made.

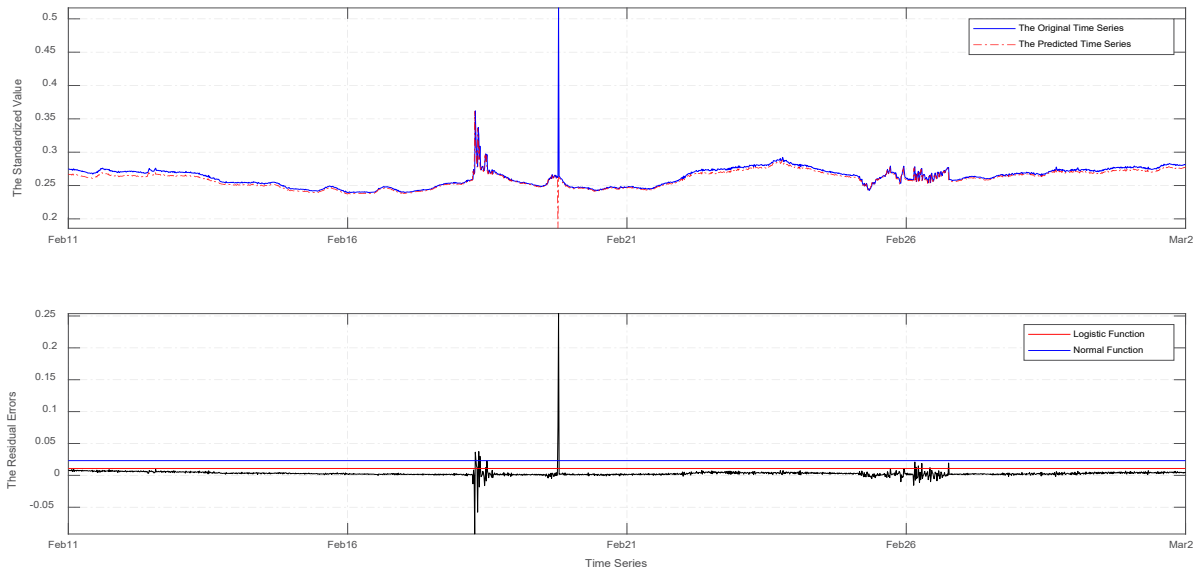

Fig. 16. Standardized TP time series and the prediction of residual errors during the test period (20 days)

The data is divided into two parts, training and testing. For the educational data part, it was used from January 1 to February 22, and for the test part, from February 21 to March 11. For each day of this interval, 96 observations of each index have been measured. After calculating the phosphate concentration index based on the surrogate relationship stated in equation 17, based on the total number of observations collected, three models of the time series mentioned in the modeling topic for phosphate concentration data ARCH, GARCH, and ARMA were applied. (The training data set is 3892 phosphates that were sampled at a distance of 15 minutes from station number 01646500 of the river.)

Step 4: Construction of anomaly detection reference (Construction Base Line Pattern): This section gives the model the ability to detect anomalies in online data with sufficient training from the input data. Diagram (14) shows the relationship of the whole structure of the proposed model in the model. The data collected for training must first be cleaned and de-noised. By decomposing the original time series into two high-frequency and low-frequency parts, the wavelet function prepares the low-frequency components to enter the RBF neural network. In this section, the threshold level of anomaly detection has been extracted and calculated by comparing the trained time series and the real time series, and the decision criterion has been set. If we consider class membership as positive and non-membership as negative; Facing these cases together causes the occurrence of four situations.

As we know, the use of wavelet transform in time-frequency series analysis allows us to extract more detailed information from the dynamics of signals and to better recognize important time and frequency points. Figure 15 shows the results of wavelet transformation for the initial time series of TP surrogate. From the difference of the two-time series of the original data and the time-frequency series trained from the RBF neural network, the residual time-frequency series of the error shown in Figure 16 is obtained.

The relationship (18) represents the logistic probability distribution function and the relationships (19) and (20) are the lower and upper limits of this distribution function. μ is the position parameter of the shape and s is the scale, the width, or shrinkage of the shape. In this study, μ=0.001 and s=0.019. Therefore, the lower limit of the pollution threshold is $Thresh_{inf} = -0.098$ and the upper limit of the threshold is $Theresh_{sup} = 0.100$.

$$f\left(x,\mu,s\right)=\frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)} \tag{18}$$

$$Thresh_{inf}=\mu+\ln\left(\frac{p_{inf}}{1-p_{inf}}\right)s=\mu-5.29s \tag{19}$$

$$Thresh_{sup}=\mu+\ln\left(\frac{p_{sup}}{1-p_{sup}}\right)s=\mu+5.29s \tag{20}$$

The sample should be a member of the real class and be recognized as a member of the same class, which is the correct positive state of TP. The instance is a member of the real class, but it is not recognized as a member of the same class, which is the false negative state of FN. The instance is not a member of the real class and is not recognized as a member of the same class, which is the correct negative state of TN. The sample is not a member of the real class, but it is recognized as a member of the same class, which is a false positive state of FP. After the implementation of the classification algorithm, according to the mentioned explanations and definitions, the performance of a classifier can be checked with the help of confusion matrix as shown below.

The accuracy parameter is the most widely used, the most basic, and the simplest measure of the quality of a classifier. This parameter shows the amount of correct recognition of the

Table 2. The results of the classification based on the real information available

		Predicted Label	
		Positive	**Negative**
Actual Label	Positive	TP	FN
	Negative	FP	TN

Table 3. The results of the evaluation parameters with 2 different scenarios

		RBF-Sen. I		Wavelet-RBF-Sen. I		RBF-Sen. II		Wavelet-RBF-Sen. II	
		Predicted		Predicted		Predicted		Predicted	
		P	N	P	N	P	N	P	N
Actual	P	966	77	994	16	975	70	1001	3
	N	92	785	20	890	60	815	8	908

	Scenario I		Scenario II	
	RBF	**Wavelet-RBF**	**RBF**	**Wavelet-RBF**
Accuracy	0.912	0.981	0.932	0.994
Misclassification Rate	0.088	0.019	0.068	0.006
Sensitivity (TPR)	0.926	0.984	0.933	0.997
Specificity	0.895	0.978	0.931	0.991

classifier in a total of two categories. Its calculation formula according to the matrix of Table 2 will be as follows.

$$Accuracy = \left(TP + TN\right) / \left(TP + FN + FP + TN\right) \tag{21}$$

The next criterion is the Misclassification Rate. This measure shows how much the classification algorithm was wrong in predicting the real positive and negative outputs. By adding the false positive and negative values together and dividing this sum by the total number of values in the data set, this value can be calculated. Its calculation formula is as follows.

$$Mis\,Classification\,Rate = \left(FP + FN\right) / \left(TP + FN + FP + TN\right) \tag{22}$$

The third criterion is sensitivity, which is also known as True Positive Rate (TPR). Sensitivity means the proportion of positive cases that the classifier has correctly identified as positive samples. This parameter is calculated as follows:

$$Sensitivity\left(TPR\right) = TP / \left(TP + FN\right) \tag{23}$$

In fact, when the researcher uses this parameter as an evaluation parameter for his classifier, his goal is to achieve the highest accuracy in detecting samples of the positive class. The last criterion studied in this research is specificity, which can be described as the ability of the algorithm to predict the true negative of each existing category. According to the definition, Spec is the ratio of true negative cases of TN to the total number of negative diagnoses of the classification. Its calculation formula is in the form of a relationship.

$$Specificity = TN / \left(FP + TN\right) \tag{24}$$

In this research to check the effectiveness of the proposed algorithm's anomaly detection, an unusual event between 10 am and 12 am on February 11 to March 2, and February 11 to March 2, was considered and implemented. The first scenario (I) was carried out by doubling the concentration of elements compared to the original time series in this period. The second scenario (II) was also created in the same period by tripling the concentration compared to the original time series. The neural network method and combined Wavelet-neural network method were used in these scenarios to check water quality. The results of the evaluation parameters can be seen in the table below.

CONCLUSION

This article is about a hybrid algorithm for early detection of water pollution impact on environmental indicators using wavelet techniques and RBF neural network learning, which was carried out on the Potomac River. Data have been extracted based on target characteristics to detect responses to water quality, such as the estimated phosphate concentration, river discharge, and water turbidity changes in the studied period in time, frequency, and time-frequency, all of which are real data. Domains of the observations extracted from the 60-day period are taken in the form of online data from the USGS website including the values of all three independent variables TURB, WT, and Sc. A hybrid algorithm based on wavelet techniques and RBF neural network learning using high-frequency surrogate relation is introduced. Important qualitative indicators such as phosphate, nitrate, and COD in the real environment have uncertainties with variations such as dependence, effectiveness of physical and chemical factors, and environmental noise. In the first step, the high-frequency time series of the main TP index was obtained through

the surrogate model and compared with GARCH techniques. By using wavelet transform, time series noise components were removed and pre-processing was done. In the next step, it was created using the neural network to identify the main characteristics of water quality. In the last step, the contamination threshold was estimated based on the basic model and calculated for the analysis of statistical models. The results show that the proposed algorithm has high stability and accuracy, and since the data are real and compared with different methods in the previous section, the proposed algorithm can be used to manage surface runoff in watersheds to protect the environment from pollution and improve water quality.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Byrand, K. (2010). Nature and History in the Potomac Country: From Hunter–Gatherers to the Age of Jefferson.Journal of Historical Geography, 2(36), 233-234.

Chen, H., Li, Q., & Zhu, F. (2022). A new class of integer-valued GARCH models for time series of bounded counts with extra-binomial variation. AStA Advances in Statistical Analysis, 106(2), 243-270.

Chen, N., Tu, H., Duan, X., Hu, L., & Guo, C. (2023). Semisupervised anomaly detection of multivariate time series based on a variational autoencoder. Applied Intelligence, 53(5), 6074-6098.

Christensen, V. G. (2001). Characterization of surface-water quality based on real-time monitoring and regression analysis, Quivira National Wildlife Refuge, south-central Kansas, December 1998 through June 2001 (No. 1-4248). US Department of the Interior, US Geological Survey.

Donoho, D. L., & Johnstone, I. M. (1994, November). Threshold selection for wavelet shrinkage of noisy data. In Proceedings of 16th annual international conference of the IEEE engineering in medicine and biology society (Vol. 1, pp. A24-A25). IEEE.

El-Shafeiy, E., Alsabaan, M., Ibrahem, M. I., & Elwahsh, H. (2023). Real-Time Anomaly Detection for Water Quality Sensor Monitoring Based on Multivariate Deep Learning Technique. Sensors, 23(20), 8613.

Erkyihun, S. T., Rajagopalan, B., Zagona, E., Lall, U., & Nowak, K. (2016). Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators. Water Resources Research, 52(5), 4061-4077.

Gokhale, M. Y., & Khanduja, D. K. (2010). Time domain signal analysis using wavelet packet decomposition approach. Int'l J. of Communications, Network and System Sciences, 3(03), 321.

Jastram, J. D. (2014). Streamflow, water quality, and aquatic macroinvertebrates of selected streams in Fairfax County, Virginia, 2007-12 (No. 2014-5073). US Geological Survey.

Jin, X. B., Gong, W. T., Kong, J. L., Bai, Y. T., & Su, T. L. (2022). PFVAE: a planar flow-based variational

auto-encoder prediction model for time series data. Mathematics, 10(4), 610.
Karasu, S., & Altan, A. (2022). Crude oil time series prediction model based on LSTM network with chaotic Henry gas solubility optimization. Energy, 242, 122964.
Kirchgässner, G., Wolters, J., & Hassler, U. (2012). Introduction to modern time series analysis. Springer Science & Business Media.
Kunz, J. V., Hensley, R., Brase, L., Borchardt, D., & Rode, M. (2017). High frequency measurements of reach scale nitrogen uptake in a fourth order river with contrasting hydromorphology and variable water chemistry (Weiße E lster, G ermany). Water Resources Research, 53(1), 328-343.
Kuo, J. T., Wang, Y. Y., & Lung, W. S. (2006). A hybrid neural–genetic algorithm for reservoir water quality management. Water research, 40(7), 1367-1376.
Lundgren, A., & Jung, D. (2022). Data-driven fault diagnosis analysis and open-set classification of time-series data. Control Engineering Practice, 121, 105006.
Makridakis, S., & Hibon, M. (1997). ARMA models and the Box–Jenkins methodology. Journal of forecasting, 16(3),147-163.
Naderian, D., Noori, R., Heggy, E., Bateni, S. M., Bhattarai, R., Nohegar, A., & Sharma, S. (2024). A water quality database for global lakes. Resources, Conservation and Recycling, 202, 107401.
Noori, R., Ghiasi, B., Sheikhian, H., & Adamowski, J. F. (2017). Estimation of the dispersion coefficient in natural rivers using a granular computing model. Journal of Hydraulic Engineering, 143(5), 04017001.
Noori, R., Mirchi, A., Hooshyaripor, F., Bhattarai, R., Haghighi, A. T., & Kløve, B. (2021). Reliability of functional forms for calculation of longitudinal dispersion coefficient in rivers. Science of The Total Environment, 791, 148394.
Saghafi, B., Hassaniz, A., Noori, R., & Bustos, M. G. (2009). Artificial neural networks and regression analysis for predicting faulting in jointed concrete pavements considering base condition. International Journal of Pavement Research and Technology, 2(1), 20-25.
Shi, B., Wang, P., Jiang, J., & Liu, R. (2018). Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. Science of the Total Environment, 610, 1390-1399.
Shupe, S. M. (2017). High resolution stream water quality assessment in the Vancouver, British Columbia region: a citizen science study. Science of the Total Environment, 603, 745-759.
Song, C., & Yao, L. (2022). Application of artificial intelligence based on synchrosqueezed wavelet transform and improved deep extreme learning machine in water quality prediction. Environmental Science and Pollution Research, 29(25), 38066-38082.
Sun, Y., Babovic, V., & Chan, E. S. (2010). Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory. Journal of Hydrology, 395(1-2), 109-116.
Talebi, M. (2023). Water Crisis in Iran and Its Security Consequences. Journal of Hydraulic Structures, 8(4), 17-28.
Tan, W. Y., Lai, S. H., Teo, F. Y., & El-Shafie, A. (2022). State-of-the-Art Development of Two-Waves Artificial Intelligence Modeling Techniques for River Streamflow Forecasting. Archives of Computational Methods in Engineering, 29(7), 5185-5211.
US EPA, (2017). National hydrography dataset high-resolution flowline data. The national map. https://www.data.gov/, Accessed date: 20 May 2017.
USGS Surface-Water Data for the Nation, https://waterdata.usgs.gov/nwis/sw.
Zhang, Y. F., & Thorburn, P. J. (2022). A deep surrogate model with spatio-temporal awareness for water quality sensor measurement. Expert Systems with Applications, 200, 116914.