STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

Enhancing Surface Water Potability Assessment through a v-SVM Hybrid Statistical Learning Model

Ahmed Naziyah alkhateeb [1], Ahmed Mutlag Algboory [2,*], Zeina Ameer Hadied [3], Oday Esam Al-Saqal [4], Zakariya Yahya Algamal[5]

[1]*Department of Operation Research and Intelligent Techniques, University of Mosul, Iraq*
[2]*College of Physical Education and Sport Sciences, University of Samarra, Iraq*
[3]*University of Mosul, Mosul, Iraq*
[4]*Department of Sharia, University of Mosul, Mosul, Iraq*
[5]*Department of Statistics and Informatics, University of Mosul, Mosul, Iraq*

Abstract The overall aim of the study is to improve the classification of water as either drinkable (potable) or non-drinkable. Traditional laboratory monitoring is time consuming, expensive and inefficient in monitoring quickly or in large scale. Alternatives to machine learning may be faster, although their performance is identifiable by the hyperparameter tuning procedure, which entails the choice of model settings. This paper proposes a new hybrid algorithm, that combines v-Support Vector Regression (v-SVR) with the sparrow search algorithm (SSA) in order to achieve the task of hyperparameter tuning automatically. The hybrid model referred to as VSRM-SSA is tested on a dataset of the water quality in the form of 3,276 samples and 10 water quality items. The results depict that the VSRM-SSA model is much better than the others in terms of high accuracy in classification. In the case of the data to be trained on, the accuracy was as high as 97.1 percent with G-mean and MCC being equal to 0.966 and 0.961 respectively. The model has already demonstrated good generalization capability using the test data with the accuracy of 91.7% and G-mean of 0.912 and the MCC of 0.907. Obviously, these values are greater than those that have been obtained with random search, Bayesian optimization, cross-validation, or grid search. Also, VSRM-SSA is the fastest method to compute (109 seconds) of all the methods tested. All in all, the suggested VSRM-SSA model offers rapid, precise, and consistent water potability classification. Its sensitivity is high and its overall performance is equal making it promising in cases where the real-time water-quality and public-health is required.

Keywords Water potability, v-SVM, Machine learning, sparrow search algorithm

DOI: 10.19139/soic-2310-5070-3929

1. Introduction

Access to safe drinking water continues to be one of the most significant global challenges humanity faces, with nearly 2 billion people currently lacking a safely managed water service [1]. This global water crisis involves much more than water's availability, as there are intricate evaluation challenges for water quality especially while urbanization, industrialization and climate change put pressure on water resources [2, 3, 4, 21, 22]. Groundwater is especially vulnerable to pollution from metallic contaminants, pathogens and chemical contaminants [5, 6, 23, 24]. The economic effects are equally devastating, in that, degraded water quality can lead to reduced economic development potentially up to one third in the affected areas [7].

Conventional water quality monitoring methods, although they are very accurate, still face challenges that make them less effective in present-day monitoring situations. Laboratory-based analysis is usually time-consuming,

*Correspondence to: Ahmed Mutlag Algboory (Email: ahmed.mo@uosamarra.edu.iq). College of Physical Education and Sport Sciences, University of Samarra, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

labor-intensive, costly, and is often impractical for large or urgent analyses [8, 9, 25, 26]. The fluctuations of water quality indicators make monitoring more complicated, because the extent and timing of pollution impacts can differ greatly over geographic areas and times. The inherent limitations in laboratory conditions, particularly in developing countries, can also lead to inefficiencies in monitoring due to delays in equipment servicing and calibrating, staff training, and quality assurance [10, 11, 27, 28, 29, 30]. These restrictions create a strong demand for alternative methods for monitoring that can offer reliable, fast, and inexpensive water quality monitoring.

The introduction of machine learning methods has opened new avenues to these difficulties by offering the automated classification systems capable of handling enormous multi-dimensional data streams [12, 13, 14, 15]. The first trials of artificial intelligence in water quality evaluation demonstrated the power of neural networks and other techniques to simulate complex, non-linear relationships between quality indicators that traditional statistical methods were unable to successfully model [16, 17, 31, 32]. Recent comparison studies have demonstrated that machine learning methods, especially Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors, routinely outperformed the traditional approach to water potability classification [8, 18, 19, 20].

The latest machine learning techniques have shown exceptional effectiveness since they make use of the output of several algorithms combined and thus provide the performance that is better than any single methodological approach used [9]. Moreover, hybrid machine learning methods are already proving their worth as they are capable of taking up the complex algorithms and the data preprocessing techniques together for future applications. It has been demonstrated by researchers that such combinations as Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and SVM can improve the prediction accuracy by taking into account the noise and variability of the environmental data [40, 33, 34, 35].

SVM algorithms are among the topmost advanced machine learning methods offering and are characterized by a mix of features that include high accuracy, consistent results, and interpretability which together make them perfect for the water quality applications [41, 42]. The algorithm operates through determining the best hyperplanes which divide the different classes in the feature space of high dimensionality, using kernel functions to manage the non-linear relationships. As a result, SVMs offer a number of distinct benefits: greater level of generalization based on a limited amount of training data, ability to analyze high-dimensional datasets, ability to deal with noisy data, and a mathematical instruction which provides theoretical usefulness [43, 44, 45].

A big advantage of SVM is its robustness to noise and outliers which is a significant benefit for water quality applications with measure errors and rare events that introduce a lot of variability [46, 47]. Additionally, many algorithms require extensive pre-processing, while SVM algorithms inherently account for these things, making them suitable for real-world monitoring applications [48, 37, 38, 39]. Furthermore, the algorithm is mathematically based on optimization theory which provides excellent information about decision boundaries and support vectors, which can help create target management strategies [49, 50].

The v-support vector regression (v-SVR) model, developed by [1], is an extension of traditional SVR that introduces the parameter to control both the number of support vectors and the amount of training error. The performance of v-SVR depends strongly on several hyperparameters, and these settings influence how well the model finds the optimal solution. A common method for choosing these hyperparameters is grid search, which tests many possible combinations, usually together with cross-validation, to evaluate how accurately the model can make predictions [2].

Enhancing water potability classification is fully depending on how the machine learning method optimized. Relies on several hyperparameters, which are often chosen subjectively or determined by several approaches and can significantly affect v-SVR quality. It identifies hyperparameter tuning as crucial but challenging due to its impact on accuracy and time. The main contribution of our proposed algorithm is to hybrid the v-SVR with coati optimization algorithm, a meta-heuristic algorithm, to optimize these hyperparameters by efficiently exploring the hyperparameters space and avoiding local optima. This leads to more accurate and stable SVM algorithm performance in water potability classification. The objective of this work is not to propose a new optimization algorithm, but to conduct a focused application study of v-SVM tuned by SSA for water potability classification. Specifically, we: (i) apply SSA to tune v-SVM hyperparameters on a public water quality dataset, (ii) compare its classification performance and computational time against RS, BO, CV and GS, and (iii) investigate the robustness of the SSA-tuned v-SVM under different train–test splits.

2. Data description

The data utilized in this study was obtained from the publicly available open-data source at https://www.kaggle.com/ (Water QuaLity Prediction database 2025). This data set was used recently by Jose, Sulochana and Mol [3] and Elshewey, Youssef [4]. The dataset contains 3,276 water quality observations with 10 variables used to predict potability (safe for drinking or not). It includes measurements of pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity.

3. v-Support vector algorithm

The use of SVM in the solution of different classification problems has been successful. However, the nonlinear regression problems have been addressed with the extension of the SVM to include the introduction of the $\varepsilon - insensitive$ loss function by [5].

Let n observations represent a trained dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,p}) \in \mathrm{R}^p$ represents a vector of the i^{th} feature, $y_i \in R$ for $i = 1, ..., n$ is the target variable, which is a quantitative variable, and $\varepsilon - insensitive$ loss function, the SVR can be obtained through solving the following optimization problem

$$\begin{array}{l} \min\limits_{\mathbf{w},b} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^n \left(\zeta_i + \tilde{\zeta}_i\right) \right\} \\ \text{S.T.} \quad \left\{ \begin{array}{l} y_i - (\mathbf{w} \bullet \varphi(\mathbf{x}_i) + b) \leq \varepsilon + \tilde{\zeta}_i \\ (\mathbf{w} \bullet \varphi(\mathbf{x}_i) + b) - y_i \leq \varepsilon + \zeta_i \\ \zeta_i, \tilde{\zeta}_i \geq 0, \end{array} \right. \end{array} \tag{1}$$

In which $C > 0$ is a penalized parameter that regulates the tradeoffs between the model complexity, and training error, ζ_i and $\tilde{\zeta}_i$are slack variables, $\varphi(\mathbf{x}_i)$ is a nonlinear mapping that is induced by a kernel function, $\mathbf{w}$ is a weight-vector and b is bias.

Then, Eq. (1) can be solved by the Lagrangian multipliers after reformulated it into its dual problem as

$$\begin{array}{l} \min\limits_{\tilde{\alpha},\alpha} \frac{1}{2} \sum_{i,j=1}^n (\tilde{\alpha}_i - \alpha_i)(\tilde{\alpha}_j - \alpha_j) K(\mathbf{x}_i, \mathbf{x}_j) + \varepsilon \sum_{i=1}^n (\tilde{\alpha}_i - \alpha_i) - \sum_{i=1}^n y_i(\tilde{\alpha}_i - \alpha_i) \\ \text{S.T.} \quad \left\{ \begin{array}{l} \sum_{i=1}^n (\alpha_i - \tilde{\alpha}_i) = 0 \\ 0 \leq \alpha_i, \tilde{\alpha}_i \leq C, \end{array} \right. \end{array} \tag{2}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$stands for kernel mapping, and α_i, $\tilde{\alpha}_i$are Lagrangian multipliers. The regression hyperplane for the underlying regression problem is then given by

$$y_i = f(\mathbf{x}_i) = \sum_{\mathbf{x}_i = \mathrm{SV}} (\tilde{\alpha}_i + \alpha_i) K(\mathbf{x}_i, \mathbf{x}_j) + b, \tag{3}$$

where SV is the support vectors set.

The original problem in v-SVR leads to convex quadratic programming with inequality constraints as [6, 7, 8, 9, 10, 11, 12]

$$\begin{array}{l} \min\limits_{\mathbf{w},b} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\left[v\varepsilon + \frac{1}{n}\sum_{i=1}^n \left(\zeta_i + \tilde{\zeta}_i\right)\right] \right\} \\ \text{S.T.} \quad \left\{ \begin{array}{l} y_i - (\mathbf{w} \bullet \varphi(\mathbf{x}_i) + b) \leq \varepsilon + \tilde{\zeta}_i \\ (\mathbf{w} \bullet \varphi(\mathbf{x}_i) + b) - y_i \leq \varepsilon + \zeta_i \\ \zeta_i, \tilde{\zeta}_i \geq 0, \ \varepsilon \geq 0, \end{array} \right. \end{array} \tag{4}$$

Equation (4) can be solved by the Lagrangian multipliers after reformulated it into its dual problem as follows:

$$\begin{array}{l} L(\mathbf{w}, b, \varepsilon, \zeta, \tilde{\zeta}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\left(\nu\varepsilon + \frac{1}{n}\sum_{i=1}^n (\zeta_i + \tilde{\zeta}_i)\right) - \sum_{i=1}^n \theta_i\zeta_i - \sum_{i=1}^n \tilde{\theta}_i\tilde{\zeta}_i - \gamma\varepsilon \\ + \sum_{i=1}^n \alpha_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b - y_i - \varepsilon - \zeta_i) + \sum_{i=1}^n \tilde{\alpha}_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b - y_i - \varepsilon - \tilde{\zeta}_i), \end{array} \tag{5}$$

where α_i, $\tilde{\alpha}_i$, θ_i, $\tilde{\theta}_i, \gamma \geq 0$ are Lagrange multipliers. The solution of Eq. (5) can be achieved by partially differentiating with respect to ζ_i, $\mathbf{w}$, b, ε,and $\tilde{\zeta}_i$ as

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} + \sum_{i=1}^{n} \alpha_i x_i - \sum_{i=1}^{n} \tilde{\alpha}_i x_i = 0 \\ \frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \tilde{\alpha}_i = 0 \\ \frac{\partial L}{\partial \varepsilon} = \frac{C}{n} \sum_{i=1}^{n} \nu - \gamma - \sum_{i=1}^{n} (\alpha_i + \tilde{\alpha}_i) = 0 \Rightarrow \\ \frac{\partial L}{\partial \zeta} = \sum_{i=1}^{n} \frac{C}{n} - \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \theta_i = 0 \\ \frac{\partial L}{\partial \tilde{\zeta}} = \sum_{i=1}^{n} \frac{C}{n} - \sum_{i=1}^{n} \tilde{\alpha}_i - \sum_{i=1}^{n} \tilde{\theta}_i = 0 \end{cases} \quad \begin{cases} \mathbf{w} = \sum_{i=1}^{n} (\tilde{\alpha}_i - \alpha_i) x_i \\ \sum_{i=1}^{n} (\tilde{\alpha}_i - \alpha_i) = 0 \\ \sum_{i=1}^{n} (\tilde{\alpha}_i - \alpha_i) = C\nu - \gamma \leq C\nu \\ \alpha_i = \frac{C}{n} - \theta_i \leq \frac{C}{n} \\ \tilde{\alpha}_i = \frac{C}{n} - \tilde{\theta}_i \leq \frac{C}{n} \end{cases} \tag{6}$$

Substituting Eq. (6) into Eq. (5), the Lagrange function can be rewritten as follows:

$$L = -\frac{1}{2} \sum_{i,j=1}^{n} (\tilde{\alpha}_i - \alpha_i)(\tilde{\alpha}_j - \alpha_j) K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} (\tilde{\alpha}_i - \alpha_i)\, y_i, \tag{7}$$

According to the Karush-Kuhn-Tucker (KKT) conditions the optimization problem in equation is. The solution of (7) is found by solving its dual [13, 14]. The last decision-making of the v-SVR model may be expressed as: after getting the optimal solution to the dual problem:

$$y_i = f(\mathbf{x}_i) = \sum_{i=1}^{n} (\tilde{\alpha}_i + \alpha_i)\, K(\mathbf{x}_i, \mathbf{x}_j) + b\,. \tag{8}$$

4. The proposed improving

A number of crucial settings have to be selected in SVR, which are called hyperparameters, in order to make the model functional. These are the parameter of penalty, the eps loss insensitive and the parameter of kernel. The choice of hyperparameters is very sensitive to the performance of SVR, and there is no precise mathematical process by which the best values can be chosen [15]. Because of this, choosing suitable hyperparameters is a major part of SVR research [15, 16, 17, 18, 19, 20, 40]. Numerous studies have attempted various methods of enhancing the performance of SVR by ensuring improved hyperparameters are chosen, and various nature-inspired optimization methods have been applied to the problem [18, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50]. But the majority of these approaches concentrate on the optimization of the hyperparameters and lack feature selection simultaneously [51, 52].

Such methods that may be employed to calculate the value of the hyper parameters are randomized search (RS), Bayesian optimization (BO), cross-validation (CV) and grid search (GS). This was done prior to the best results on the selected criterion being returned on all possible combinations of the hyper parameters and also the combination of hyper parameters also returning the best results. However, they are computationally complex and they are yet to exhaust the entire hyperparameters combinations [53].

Consequently, more effective and better methods of optimizing hyper parameters of v- SVR ought to be acquired. During the last few years, metaheuristic optimization algorithms have been broadly applied to the issue of hyperparameter tuning [54].

In the past few years, researchers have come up with various new nature-based algorithms in order to expand and improve the range of exploration and use of the already available algorithms. A sparrow search algorithm (SSA) is one of the most popular algorithms of its new algorithms since it is extremely high-performin [55].

The SSA is a swarm based optimization algorithm that is based on the foraging and anti-predatory behaviorof sparrows. It models the smart collective action of the foraging sparrows to find food and escape predators [56]. SSA is between two major types of sparrows producers (leaders who find food) and scroungers (followers who follow producers to get food) with some sparrows serving as scouts or explorers. The algorithm repeats the optimization of solutions to the complex optimization problems by updating positions of individuals in the search space [57]. The SSA is mathematically computed updating the sparrows positions depending upon their positions and environmental factors within a number of equations describing their search and avoidance behavior:

4.1. First: Position update for producers:

$$y_{k.j}^{t+1} = \begin{cases} y_{k.j}^{t} \times \exp\left(\frac{-k}{\alpha \times iter}\right) & \mathrm{R}_2 < ST \\ y_{k.j}^{t} + qL & \mathrm{R}_2 \geq ST \end{cases} \tag{9}$$

where $y_{k.j}^{t+1}$ indicates the individual position of the k_{th} sparrow on the j_{th} dimension in the $(t+1)_{th}$ generation; $y_{k.j}^{t}$describes the position of the k_{th} sparrow on the j_{th} dimension in the t_{th} generation ;α is randomly generated from (0,1]; iter is the total number of generations; R_2 is the alarm value that belongs to [0,1]; ST is the safety threshold generated from [0.5,1]; denotes a random digit following the standard distribution and L expresses the matrix of 1*dim , where dim indicates the dimension of the dataset [57].

$$y_{k.j}^{t+1} = \begin{cases} q \times \exp\left(\frac{y_{worst}^{t} - x_{k.j}^{t}}{k^2}\right) & \mathrm{k} > \frac{n}{2} \\ y_{best}^{t+1} + \frac{1}{D} \sum_{D=1}^{D} \left(rand \times \left| y_{k.j}^{t} - y_{best}^{t+1} \right|\right) & \mathrm{k} \leq \frac{n}{2} \end{cases} \tag{10}$$

Where y_{worst}^{t}indicates the sparrow location having the worst fitness at the t_{th} generation and y_{best}^{t}denotes the location of the producer with the best fitness in the $(t+1)_{th}$ generation.

4.2. Second: Alarm or anti-predation behavior update

$$y_{k.j}^{t+1} = \begin{cases} y_{best}^{t} + \beta \left| y_{k.j}^{t} - y_{best}^{t} \right| & f_k > f_b \\ y_{k.j}^{t} + h\left(\frac{y_{k.j}^{t} - y_{worst}^{t}}{f_k - f_w + \varepsilon}\right) & f_k = f_b \end{cases} \tag{11}$$

where y_{best}^{t}represents the global optimal position in the t_{th} generation ; β represents the step size control factor ,which follows the normal distribution ;h represents the move direction of sparrow individuals ,which is generated from[-1,1],and it is also a factor of the step size control ; ε is constant used to avoid division be zero errors ; $f_k$indicates the fitness of the k_{th} sparrow; and the best and worst global are represented byf_b andf_w respectively.

To maximize the hyperparameters of v-SVR with the enhancement suggestion of SSA, the position vector X of each coati is determined as a dimension D vector which represents the position of the coati in the SSA. Consequently, the generated safety threshold would be X dependent on a specific configuration of the v-SVR and the dimensions of the X would be the hyperparameters of the v-SVR. Therefore, three points that each of the coatis in the swarm will be foraging will be identified. Our proposed improving is, therefore, as:

Step 1: $N_{sparrow}$ is initially defined as 30 and $T = 500$ is the maximum number of iterations.

Step 2: The location of every coati is randomly defined. The three positions are random, i.e. generated in a uniform distribution $C \sim U(0,7)$, $\sim U(0,4)$ and $v \sim U(0,1)$.

Step 3: The fitness function is determined as:

$$\text{fitness} = \max \text{CA}. \tag{12}$$

CA is the accuracy of classification.

Step 4: Coati positions are updated with the help of the Eq. (10) and Eq. (11), respectively. Step 5: Repeat steps 3 and 4 until a T is obtained.

5. Results and discussion

VSRM-SSA algorithm is tested against VSRM-RS, VSRM-BO, VSRM-CV and VSRM-GS so as to enhance the algorithm in classification of water potability. Accuracy of classification obtained in experiments is by classification accuracy (CA), G-mean and Mathew correlation coefficient (MCC). The CA, G-mean and MCC are defined respectively as: $\text{CA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$

$$\text{G} - \text{mean} = \sqrt{\frac{\text{TN}}{\text{FP} + \text{TN}} + \frac{\text{TP}}{\text{TP} + \text{FN}}} \tag{13}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP}+\mathrm{TN})-(\mathrm{FP}+\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{14}$$

where TP, TN, FP and FN are the values of true positive, true negative, false positive and false negative respectively of the confusion matrix. The larger the values of CA, G-mean and MCC, the stronger are the classification tasks.

Table 1 and Table 2 demonstrate CA, G-mean, and MCC of each method based on the division of the dataset into two parts (70 and 30 per cent) into training and testing data respectively. The outstanding ones are written in bold. The results on Table 1 are apparent with regard to CA that v-SVR using SSA, VSRM-SSA is very effective. There is the highest performance (more than 95) in the VSRM-SSA that denotes high performance of water potability classification based on the training data with balanced classification performance. The VSRM-SSA CA is 98.3 which is higher than the VSRM-RS CA (79.7), VSRM-BO CA (83.1) and VSRM-CV CA (83.2) and VSRM-GS CA (83.2).

Table 1. Classification results of the used methods for the training data.

	VSRM-RS	VSRM-BO	VSRM-CV	VSRM-GS	VSRM-SSA
CA	0.797	0.845	0.85	0.863	**0.985**
G-mean	0.782	0.828	0.833	0.846	**0.981**
MCC	0.777	0.825	0.83	0.843	**0.975**

Conversely, when comparing the VSRM-SSA to the other four methods on the Table 1 the VSRM-SSA had a higher value on the G-mean. The G-mean measure gives the harmonic response of sensitivity and specificity which presents the capability of the model to classify correctly both water that is potable and non-portable water, especially when there is an unequal distribution of data. The fact that the VSRM-RS was increased to 0.966 in VSRM-SSA also shows that VSRM-SSA maintains a good balance and is effective in various classes and not biased towards a particular class. This is essential to enhance reliability in classification of tasks of potability of water where the class imbalance is usual.

In addition, MCC does not only take the accuracy into consideration, but the entire information of the confusion matrix such as true and false positives and negatives. MC is particularly strong in unbalanced classification. Based on Table 1, it is observed that MCC increases between 0.763 (VS-RM0) and 0.961 (VS-RM1), which confirms that VSRM-SSA has high and balanced power of water in terms of potability. High MCC indicates that the water potability classification of the model fits the reality classes strongly with little and no biased and random classification.

Moreover, VSRM-RS demonstrates the lowest performance as compared to the other five methods used in CA of 78.3, G-mean of 0.768 and MCC of 0.763. It shows a moderate overall accuracy but comparatively lower balance between sensitivity and specificity and lower correlation between predicted and actual classes. Instead, VSRM-BO has a higher prediction power and class balance with accuracy of 83.1% in water potability classification and better-balanced classification of 0.814 G-mean and 0.811 MCC than VSRM-RS. VSRM-CV, on the other hand, is just slightly better than VSRM-BO with CA of 83.6, G-mean of 0.819 and MCC of 0.816. This implies that tuning of the parameters in cross-validation leads to improved model performance and reliability. Associated with VSRM-GS, a better CA of 84.9, G-mean of 0.832, and MCC of 0.829 are better. The optimization of the hyperparameters through the grid search to fine-tuning of the hyperparameters enhances the accuracy and balance of the VSRM.

In relation to the testing data (Table 2), VSRM-SSA substantially outperforms other models with very high CA of 91.7%, G-mean of 0.912, and MCC of 0.907. It indicates excellent water potability classification accuracy, balanced sensitivity and specificity, and a strong correlation between prediction and ground truth, reflecting a robust and well-generalized model. High CA, G-mean, and MCC on testing data indicates that the SVM with SSA generalizes well beyond the training dataset, effectively capturing underlying patterns rather than overfitting to noise. This reliability means the SVM with SSA can make accurate water potability classification accuracy on unseen data, which is critical for practical applications.

To further highlight the performance of the v-SVR with SSA, the computational time in seconds is depicted in Figure 1 for the VSRM-SSA and the other four methods. Figure 1 reveals a clear downward trend in computational

Table 2. Classification results of the used methods for the testing data.

	VSRM-RS	VSRM-BO	VSRM-CV	VSRM-GS	VSRM-SSA
CA	0.742	0.79	0.795	0.808	0.932
G-mean	0.727	0.773	0.778	0.791	0.925
MCC	0.722	0.77	0.775	0.788	0.921

time from VSRM-RS to VSRM-SSA. This suggests that optimization techniques applied progressively reduce the time taken to tune the v-SVR algorithm. VSRM-SSA achieves the shortest computation time of 109 seconds, implying that this composite optimization algorithm effectively balances exploration and exploitation in parameter space or model structure, leading to faster training and evaluation without sacrificing accuracy. Conversely, VSRM-RS has the highest computational time at 175 seconds, indicating that random search for v-SVR hyperparameters, while straightforward, may involve many iterations or less efficient exploration of the parameter space, leading to longer runtimes.

On the other hand, VSRM-BO reduces time to 151 seconds, reflecting that Bayesian optimization efficiently focuses on promising parameter regions, reducing unnecessary computations compared to random search. Both VSRM-CV (136 seconds) and VSRM-GS (127 seconds) optimize v-SVR tuning by systematically validating parameter combinations with cross-validation or exhaustive grid search, balancing runtime efficiency and thoroughness.

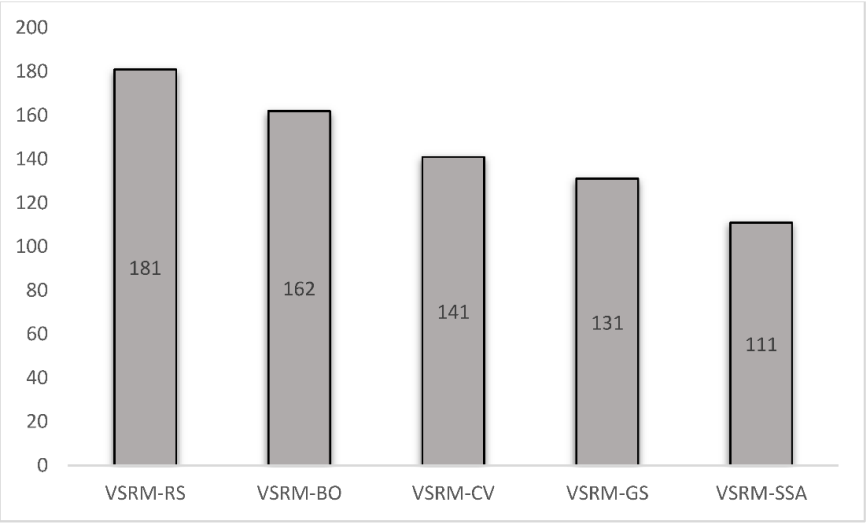


Figure 1. Computational time of the used methods for the training data.

To better check how well the VSRM-SSA method works, we also used a statistical test called the Wilcoxon signed-rank test. This test helps us see if the differences between VSRM-SSA and the other four methods are truly important and not just random. We applied the test using the G-mean values from the training results. The outcomes, shown in Table 3, indicate that VSRM-SSA performs significantly better than all other methods. All p-values were below 0.05, which means the improvements of VSRM-SSA are statistically meaningful.

In order to further emphasize the performance obtained with SVM using SSA, the different splitting data is considered as 90% training:10% testing, 80% training: testing 20%, 60% training: 40% testing, and 50% training: 50% testing. Figures 2-4 show the water potability classification performance of the VSRM-SSA, VSRM-RS, VSRM-BO, VSRM-CV, and VSRM-GS methods for the training and testing data in terms of CA, G-mean, and MCC. From these figures, it is clearly seen that VSRM-SSA consistently performs best across all splits. In terms

Table 3. p-values for the Wilcoxon signed-rank test of the VSRM-SSA results with four competitor methods.

Pairwise comparison	ρ-Value
VSRM-SSA vs VSRM-RS	0.0001
VSRM-SSA vs VSRM-BO	0.0005
VSRM-SSA vs VSRM-CV	0.0013
VSRM-SSA vs VSRM-GS	0.0029

of CA in the training data for example, the accuracy is significantly higher for VSRM-SSA at 90%:10% split up to at 70%:30%. This suggests that the VSRM-SSA improves the water potability classification accuracy more than others. In general, water potability classification accuracy is highest around the splits (70%:30%), especially for VSRM-SSA. Furthermore, the trend is less clear for other methods, but generally, accuracy decreases when training data is very small (50%:50%) or when very little test data is used (90%:10%). Additionally, related to the comparison among other methods, VSRM-GS, VSRM-CV, and VSRM-BO give moderately high water potability classification performance, with VSRM-GS usually slightly better than VSRM-CV and VSRM-BO. Conversely, VSRM-RS gives the lowest accuracy among the methods. In summary, increasing the fraction of training data generally improves classification accuracy, but having a reasonable test set size like 30% also helps in reliably evaluating performance.

The enhancements in performance made by the v-SVM based VSRM-SSA model, from the viewpoint of water quality and public health, carry significant implications. The main water safety determination by a classifier directly impacts the population's health. SVM models have already been acknowledged for their superb generalization ability, noise data immunity, and stable performance in large-dimensional settings [34, 35]. These traits make SVM a reliable ground for the classification of water potability, which is frequently defined by measurements with mistakes, variability, and overlapping patterns [55].

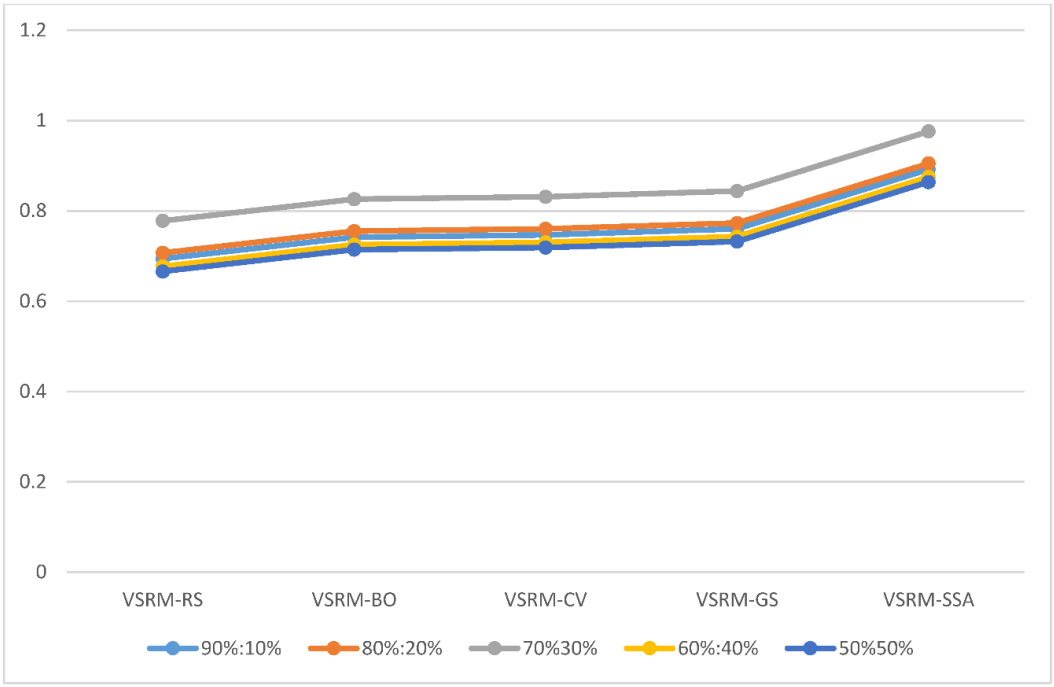
This research version of the v-SVM model supports the idea that hyperparameter tuning via SSA is the best way of potential SVM-based systems. The model's performance has been tested with a 91.7% accuracy, and it has yielded very high G-mean (0.912) and MCC (0.907), which are the signs of good sensitivity and specificity expressed. Practically speaking, this means that the classifier is able to identify and correctly classify more than 90% of the contaminated water samples while at the same time being very precise in marking the clean ones. A fully balanced detection ability such as this one is very critical since to assign unsafe water to safe category (false negatives) may expose people to serious health risks through waterborne diseases, whereas the opposite (false positives) might lead to unnecessary measures being taken and ultimately, the costs of operation going up.

The v-SVM framework, detrimental through SSA, lessens two types of mistakes by creating a clear-cut decision boundary that divides the two categories better than classical tuning methods. The remarkable performance in terms of MCC and G-mean shows that the model is reliable even if there is class imbalance, which is a common problem in environmental datasets. Such a situation reflects that the model is both accurate and trustworthy in distinguishing correctly both less and more number of instances classes.

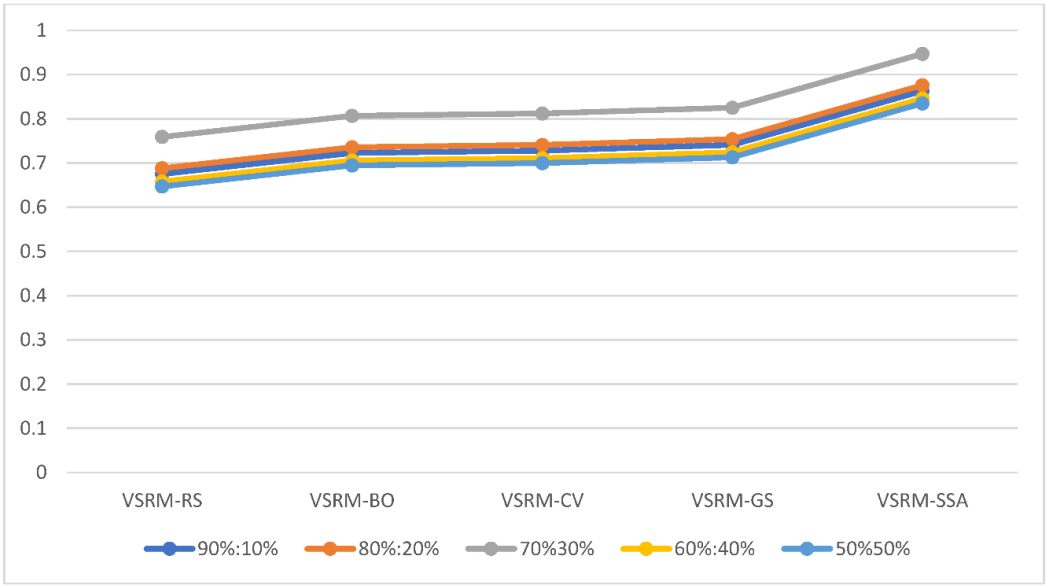
In real-life situations, a v-SVM-based trustworthy classifier can significantly support the water quality real-time monitoring systems. Thanks to its computation time of only 109 seconds, the VSRM-SSA model is the most suitable option for automated monitoring and alarm systems where quick decisions are critical. The pairing of speed and accuracy gives the environmental agencies an opportunity to spot pollution very fast and take the necessary actions, thus, reducing the risk of large-scale exposure.

Also, the interpretability of decision functions of support vector machine (SVM) allows water authorities to fully understand which water-quality parameters (like pH, turbidity, or sulfate) have positive or negative impacts on the predictions of potability. This may lead to a smoother transition to the implementation of evidence-based policies and water treatment practices. The outputs of the model could be used by municipal authorities and public health agencies to rank areas for further investigation, distribute testing resources optimally, and observe contamination patterns more efficiently.

To summarize, the remarkable and steady performance of the v-SVM-SSA model has demonstrated that using advanced SVM-based optimization in a water potability classification system can greatly meet the growing demand



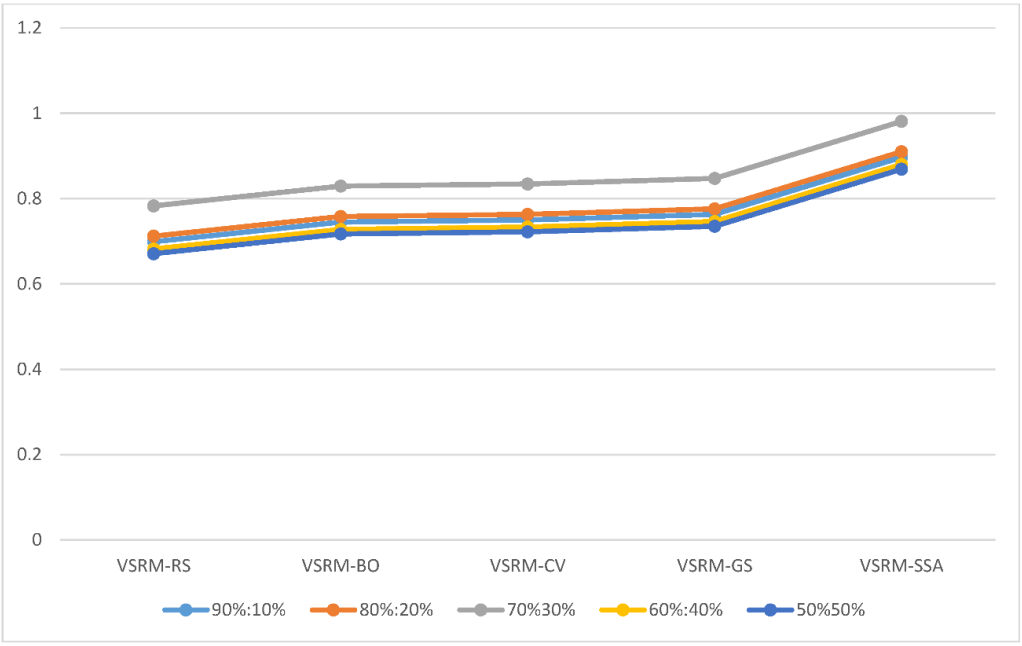
Training data



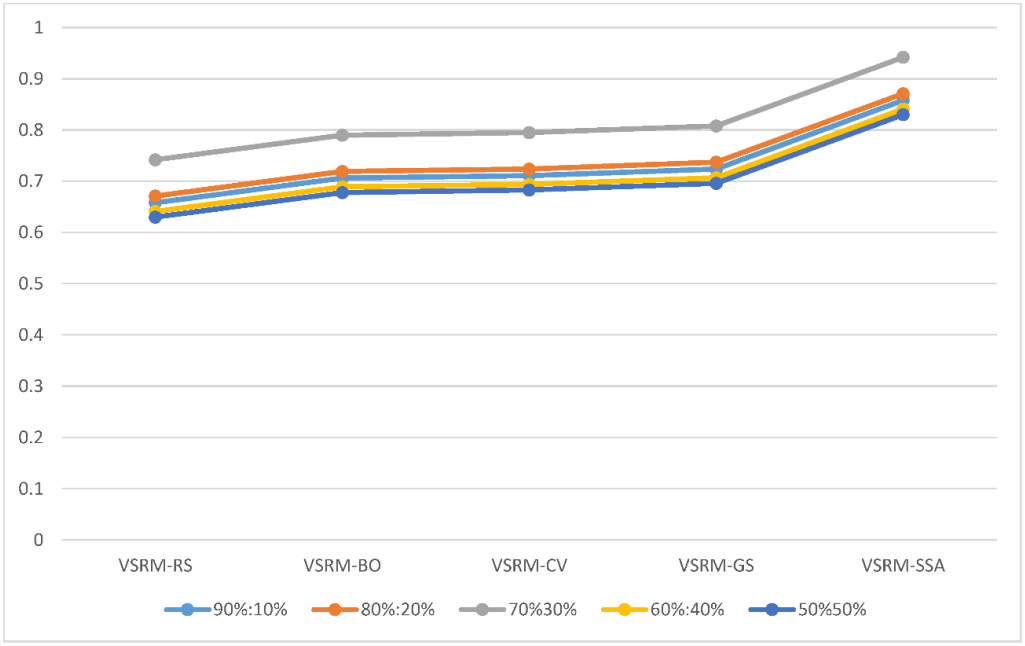
Testing data

Figure 2. Water potability classification performance under different splitting percentage in terms of CA.

for fast, trustworthy, and automated systems. The model's high accuracy and robustness under various conditions



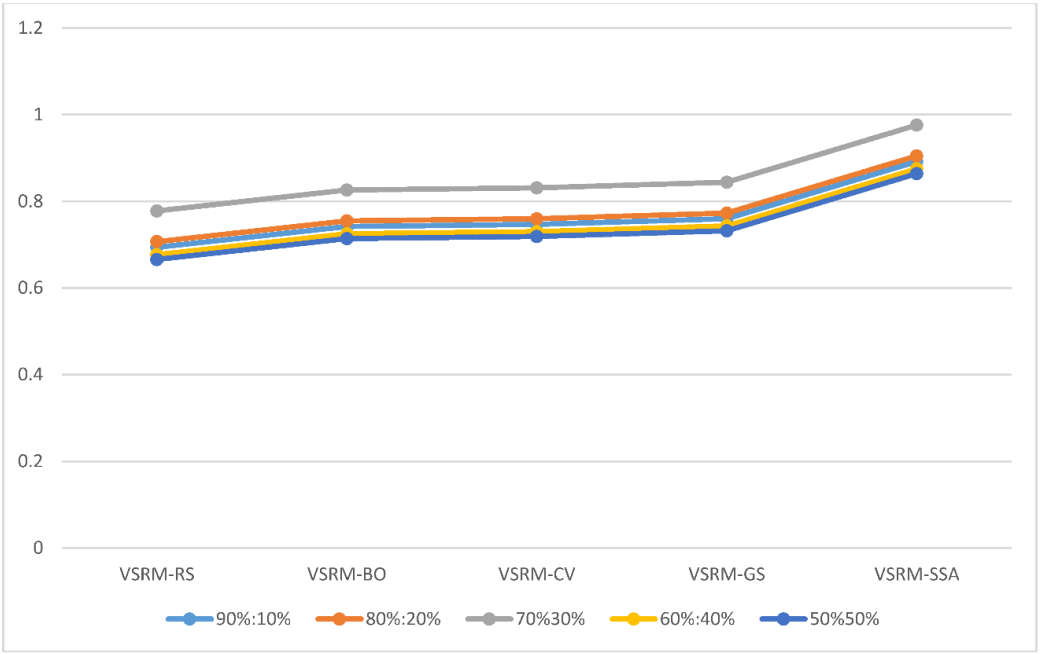
Training data



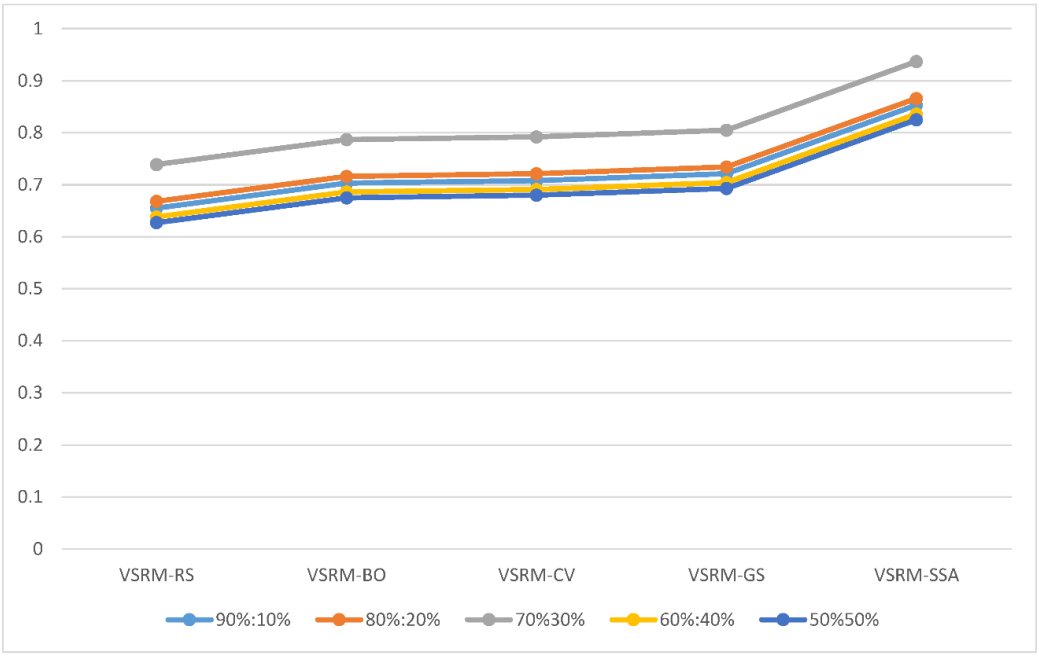
Testing data

Figure 3. Water potability classification performance under different splitting percentage in terms of G-mean.

offer an excellent and credible resource for water quality management and, consequently, the protection of public health.



Training data



Testing data

Figure 4. Water potability classification performance under different splitting percentage in terms of MCC.

6. Conclusion

A hybrid ML model called VSRM-SSA was presented in this paper, which integrates v-Support Vector Regression with SSA to classify water potability better. The results indicate that the adoption of SSA for hyperparameter tuning greatly boosts the performance of v-SVR. In every test, VSRM-SSA was the one producing the highest classification results of the compared methods. The training dataset accuracy was 97.1%, G-mean was 0.966, and MCC was 0.961. The testing dataset also showed a strong generalization with 91.7% accuracy, 0.912 G-mean, and 0.907 MCC. The mentioned advances signify that SSA not only moves through the parameter space well but also avoids local optima better than random search, Bayesian optimization, grid search, or cross-validation. Moreover, the model was the one demonstrating the shortest computation time (109 seconds), indicating that great accuracy can be obtained at the same time with no increase in computational cost. The above-mentioned characteristic of VSRM-SSA makes it appropriate for the real-time or near real-time monitoring systems. To sum up, the results indicate that the developed VSRM-SSA framework delivers a prompt, precise, and stable solution for colorimetric classification of drinking-water safety. Its excellent performance on various metrics indicates that it could be an effective instrument for the monitoring of environmental changes and protection of human health. Future studies may consider implementing this hybrid technique on more extensive datasets, various environmental conditions, or coupling it with automated sensor systems.

REFERENCES

1. B. Schölkopf, et al., *New support vector algorithms*, Neural Computation, vol. 12, no. 5, pp. 1207–1245, 2000.
2. H. Kaneko, and K. Funatsu, *Fast optimization of hyperparameters for support vector regression models with highly predictive ability*, Chemometrics and Intelligent Laboratory Systems, vol. 142, pp. 64–69, 2015.
3. D. J. Jose, C. H. Sulochana, and I. J. Mol, *Sustainable water quality prediction using adaptive spider monkey optimization with Bi-LSTM*, Earth Science Informatics, vol. 18, no. 3, pp. 1–20, 2025.
4. A. M. Elshewey, et al., *Water potability classification based on hybrid stacked model and feature selection*, Environmental Science and Pollution Research, vol. 32, no. 13, pp. 7933–7949, 2025.
5. V. N. Vapnik, *An overview of statistical learning theory*, IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988–999, 1999.
6. P. Y. Hao, *Pair-ν-SVR: A novel and efficient pairing nu-support vector regression algorithm*, IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 11, pp. 2503–2515, 2017.
7. D. Kong, et al., *Tool wear monitoring based on kernel principal component analysis and ν-support vector regression*, The International Journal of Advanced Manufacturing Technology, vol. 89, no. 1–4, pp. 175–190, 2016.
8. N. Li, et al., *Force-based tool condition monitoring for turning process using ν-support vector regression*, The International Journal of Advanced Manufacturing Technology, vol. 91, no. 1–4, pp. 351–361, 2016.
9. Y. Liu, and G. Pender, *A flood inundation modelling using ν-support vector machine regression model*, Engineering Applications of Artificial Intelligence, vol. 46, pp. 223–231, 2015.
10. X. Teng, H. Dong, and X. Zhou, *Adaptive feature selection using v-shaped binary particle swarm optimization*, PLoS One, vol. 12, no. 3, e0173907, 2017.
11. J. Wang, et al., *Improved ν-support vector regression model based on variable selection and brain storm optimization for stock price forecasting*, Applied Soft Computing, vol. 49, pp. 164–178, 2016.
12. Y. Zhang, and Y. Xie, *Forecasting of short-term freeway volume with ν-support vector machines*, Transportation Research Record, vol. 2024, no. 1, pp. 92–99, 2007.
13. O. M. Ismael, O. S. Qasim, and Z. Y. Algamal, *Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in ν-support vector regression based on opposition-based learning*, Journal of Chemometrics, vol. 34, no. 11, 2020.
14. O. M. Ismael, O. S. Qasim, and Z. Y. Algamal, *Improving parameters of ν-support vector regression with feature selection in parallel by using quasi-oppositional and Harris hawks optimization algorithm*, Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, vol. 14, no. 2, pp. 113–118, 2024.
15. P. Tsirikoglou, et al., *A hyperparameters selection technique for support vector regression models*, Applied Soft Computing, vol. 61, pp. 139–148, 2017.
16. J. S. Chou, and A. D. Pham, *Nature-inspired metaheuristic optimization in least squares support vector regression for obtaining bridge scour information*, Information Sciences, vol. 399, pp. 64–80, 2017.
17. R. Laref, et al., *On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications*, Chemometrics and Intelligent Laboratory Systems, vol. 184, pp. 22–27, 2019.
18. S. Li, H. Fang, and X. Liu, *Parameter optimization of support vector regression based on sine cosine algorithm*, Expert Systems with Applications, vol. 91, pp. 63–77, 2018.
19. V. Cherkassky, and Y. Ma, *Practical selection of SVM parameters and noise estimation for SVM regression*, Neural Networks, vol. 17, no. 1, pp. 113–126, 2004.
20. K. Ito, and R. Nakano, *Optimizing support vector regression hyperparameters based on cross-validation*, Proceedings of the International Joint Conference on Neural Networks, IEEE, 2003.

21. Al-Fakih, A. M., Algamal, Z. Y., & Qasim, M. K. *An improved opposition-based crow search algorithm for biodegradable material classification* SAR and QSAR in Environmental Research, vol.33, no.5, p. 403-415, 2022.
22. Alharthi, A. M., Kadir, D. H., Al-Fakih, A. M., Algamal, Z. Y., Al-Thanoon, N. A., & Qasim, M. K. *Quantitative structure-property relationship modelling for predicting retention indices of essential oils based on an improved horse herd optimization algorithm*, SAR and QSAR in Environmental Research, vol . 34 , no .10 , p. 831-846, 2023.
23. Lukman, A. F., Dawoud, I., Kibria, B. G., Algamal, Z. Y., & Aladeitan, B. *A new ridge-type estimator for the gamma regression model* Scientifica, vol. 1, p.5545356, 2021.
24. Naziyah, A. A., & Algamal, Z. Y. *Jackknifed Liu-type estimator in Poisson regression model* Journal of the Iranian Statistical Society, vol.19, no.1, p. 21-37, 2020.
25. Algamal, Z., & Ali, H. M. *An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression* , Electronic Journal of Applied Statistical Analysis, vol . 10 , no . 1 , 242-256 , 2017.
26. Qasim, M. K., Algamal, Z. Y., & Ali, H. M. *A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine*, SAR and QSAR in Environmental Research, vol. 29, no.(7), p.517-527, 2018
27. Awwad, F. A., Odeniyi, K. A., Dawoud, I., Algamal, Z. Y., Abonazel, M. R., Kibria, B. G., & Eldin, E. T. *New two-parameter estimators for the logistic regression model with multicollinearity*, WSEAS Transactions on Mathematics, vol. 21, p.403-414, 2022
28. Al-Taweel, Y., & Algamal, Z. Y. *Some almost unbiased ridge regression estimators for the zero-inflated negative binomial regression model* Periodicals of Engineering and Natural Sciences, vol.8, no.1, p. 248-255, 2020.
29. Algamal, Z. Y., & Lee, M. H. *Applying penalized binary logistic regression with correlation based elastic net for variables selection*, Journal of Modern Applied Statistical Methods, vol. 14, no.(1), p.15, 2015
30. Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. *High-dimensional QSAR modelling using penalized linear regression model with L 1/2-norm*, SAR and QSAR in Environmental Research, vol. 27, no.(9), p.703-719, 2016
31. Ewees, A. A., Algamal, Z. Y., Abualigah, L., Al-Qaness, M. A., Yousri, D., Ghoniem, R. M., Abd Elaziz, M. *A cox proportional-hazards model based on an improved aquila optimizer with whale optimization algorithm operators*, Mathematics, vol. 10, no.(8), p.1273, 2022.
32. Algamal, Z. Y., Qasim, M. K., Lee, M. H., & Ali, H. T. M. *QSAR model for predicting neuraminidase inhibitors of influenza A viruses (H1N1) based on adaptive grasshopper optimization algorithm*, SAR and QSAR in Environmental Research, vol.31, no.11, p.803-814, 2020.
33. Shamany, R., Alobaidi, N. N., & Algamal, Z. Y. *A new two-parameter estimator for the inverse Gaussian regression model with application in chemometrics*, Electronic Journal of Applied Statistical Analysis, vol. 12, no.(2), p.453-464, 2019
34. Kahya, M. A., Altamir, S. A., & Algamal, Z. Y. *Improving firefly algorithm-based logistic regression for feature selection*, Journal of Interdisciplinary Mathematics, vol. 22, 1577-1581, 2019.
35. Alharthi, A. M., Kadir, D. H., Al-Fakih, A. M., Algamal, Z. Y., Al-Thanoon, N. A., & Qasim, M. K. *Improving golden jackel optimization algorithm: An application of chemical data classification*, Chemometrics and Intelligent Laboratory Systems, vol.250, 105149, 2024.
36. Shehab, Z. N., Algamal, Z. Y., & Farhhan, A. F. *Water Quality Assessment and Spatial-Temporal Variation Analysis in Tigris River, Mosul, Iraq During 2023-2024* Egyptian Journal of Aquatic Biology and Fisheries, vol. 29, no.6, p.3607-3626, 2025.
37. Ibrahim, O. S., & Algamal, Z. Y. *Hybrid V-Support Vector Regression as Statistical Methodology for Forecasting Daily Natural Gas Prices* Industrial Engineering & Management Systems, vol. 24, no.4, p. 694-705, 2025.
38. Shehab, Z. N., Algamal, Z. Y., & Faisal, R. M. *Forecasting PM2.5 Daily Concentration in Baghdad, Iraq Based on Improving Random Forest Algorithm* Iraqi National Journal of Earth Science, vol. 26, no.1, p. 180 - 192, 2026.
39. Basheer, G. T., Waleed, S., & Algamal, Z. Y. *Improving Parameters estimation of a truncated Poisson regression model based on meta-heuristic optimization algorithms* Frontiers in Applied Mathematics and Statistics, vol. 12, 1744058, 2026.
40. M. M. Al-hashimi, and A. N. Alkhateeb, *Spatial analysis of brain and other CNS cancers incidence in Iraq during 2000–2015*, Malaysian Journal of Public Health Medicine, vol. 20, no. 3, pp. 27–34, 2020.
41. C. H. Wu, G. H. Tzeng, and R. H. Lin, *A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression*, Expert Systems with Applications, vol. 36, no. 3, pp. 4725–4735, 2009.
42. A. Kazem, et al., *Support vector regression with chaos-based firefly algorithm for stock market price forecasting*, Applied Soft Computing, vol. 13, no. 2, pp. 947–958, 2013.
43. M. N. Amar, and N. Zeraibi, *Application of hybrid support vector regression artificial bee colony for prediction of MMP in CO2-EOR process*, Petroleum, 2018.
44. C. F. Huang, *A hybrid stock selection model using genetic algorithms and support vector regression*, Applied Soft Computing, vol. 12, no. 2, pp. 807–818, 2012.
45. C. T. Cheng, et al., *Optimizing hydropower reservoir operation using hybrid genetic algorithm and chaos*, Water Resources Management, vol. 22, no. 7, pp. 895–909, 2007.
46. W. C. Hong, et al., *SVR with hybrid chaotic genetic algorithms for tourism demand forecasting*, Applied Soft Computing, vol. 11, no. 2, pp. 1881–1890, 2011.
47. J. Cheng, J. Qian, and Y. N. Guo, *Adaptive chaotic cultural algorithm for hyperparameters selection of support vector regression*, International Conference on Intelligent Computing, Springer, 2009.
48. B. Üstün, et al., *Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization*, Analytica Chimica Acta, vol. 544, no. 1–2, pp. 292–305, 2005.
49. J. Zhang, et al., *Optimization enhanced genetic algorithm-support vector regression for the prediction of compound retention indices in gas chromatography*, Neurocomputing, vol. 240, pp. 183–190, 2017.
50. G. Cao, and L. Wu, *Support vector regression with fruit fly optimization algorithm for seasonal electricity consumption forecasting*, Energy, vol. 115, pp. 734–745, 2016.
51. A. M. Algboory, A. N. Alkhateeb, and Z. Y. Algamal, *A modified jackknife Liu-type estimator for the gamma regression models data*, Statistics, Optimization & Information Computing, vol. 14, no. 5, pp. 2142–2154, 2025.

52. A. N. Alkhateeb, and Q. N. N. Al-Qazaz, *Variable selection in Weibull accelerated survival model based on chaotic sand cat swarm algorithm*, Statistics, Optimization & Information Computing, vol. 13, no. 5, pp. 2105–2118, 2025.
53. E. Abdulsaed, M. Alabbas, and R. Khudeyer, *Hyperparameter optimization for convolutional neural networks using the salp swarm algorithm*, Informatica, vol. 47, no. 9, 2023.
54. Z. Y. Algamal, et al., *Improving grasshopper optimization algorithm for hyperparameters estimation and feature selection in support vector regression*, Chemometrics and Intelligent Laboratory Systems, vol. 208, 104196, 2021.
55. M. Dehghani, et al., *Coati optimization algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems*, Knowledge-Based Systems, vol. 259, 110011, 2023.
56. J. Xue, and B. Shen, *A novel swarm intelligence optimization approach: Sparrow search algorithm*, Systems Science & Control Engineering, vol. 8, no. 1, pp. 22–34, 2020.
57. L. Sun, et al., *BSSFS: Binary sparrow search algorithm for feature selection*, International Journal of Machine Learning and Cybernetics, vol. 14, no. 8, pp. 2633–2657, 2023.
58. L. Lian, *An improved sparrow search algorithm using chaotic opposition-based learning and hybrid updating rules*, Concurrency and Computation: Practice and Experience, vol. 36, no. 14, 2024.