Integration of advanced optimization algorithms into least-square support vector machine (LSSVM) for water quality index prediction

See Leng Chia, Min Yan Chia, Chai Hoon Koo and Yuk Feng Huang
Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tun Mustapha, Bandar Sungai Long, Kajang, Selangor Darul Ehsan, Malaysia
*Corresponding author. E-mail: kooch@utar.edu.my

ABSTRACT

Machine learning models hybridized with optimization algorithms have been applied to many real-life applications, including the prediction of water quality. However, the emergence of newly developed advanced algorithms can provide new scopes and possibilities for further enhancements. In this study, the least-square support vector machine (LSSVM) integrated with advanced optimization algorithms is presented, for the first time, in the prediction of the water quality index (WQI) at the Klang River of Malaysia. Thereafter, the LSSVM model using RBF kernel was optimized using the hybrid particle swarm optimization and genetic algorithm (HPSOGA), whale optimization based on self-adapting parameter adjustment and mix mutation strategy (SMWOA) as well as ameliorative moth-flame optimization (AMFO) separately. It was found that the SMWOA-LSSVM model had the better performance for WQI prediction by having the best achievement root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R²) and mean absolute percentage error (MAPE). Comprehensive comparison was done using the global performance indicator (GPI), whereby the SMWOA-LSSVM had the highest average score of 0.31. This could be attributed to the internal architecture of the SMWOA, which was catered to avoid local optima within short optimization period.

Key words: advanced optimization algorithm, hybrid optimized LSSVM model, kernel function, river water quality parameter, water quality index

HIGHLIGHTS

- Advanced optimization algorithms were applied, for the first time, in WQI prediction.
- LSSVM using RBF as kernel function was found to be the best model.
- All the hybrid LSSVM integrated with optimization algorithms had improved in accuracy, against the base models.
- SMWOA-LSSVM was found to be the most suitable hybrid model for WQI prediction at the Klang River.

INTRODUCTION

The water quality index (WQI) is a crucial indicator to assess the quality of different water bodies for better management (Mehta et al. 2018; Zhang 2019). The WQI is usually calculated using multiple water quality parameters that require on-site data collection. However, the process of obtaining water quality parameters through sampling at different points can be time consuming, not to mention the tedious work as well as the high financial cost incurred (Najah Ahmed et al. 2019). A more efficient measure to obtain WQI would be a step in the right direction to ensure effective repeated monitoring of the quality of water bodies, especially in regions where pollution is prone to happen frequently. This is especially vital to track down the pollution sources before water treatment plants and water supply of the respective regions are affected. Machine learning models fit this task extremely well due to their quick response towards the fluctuation of the water quality parameters.

Currently, the tree-based model, kernel-based model, fuzzy-based model and artificial neural network (ANN) model are the mainstream machine learning models that are used in water quality prediction (Rajaee et al. 2020). Asadollah et al. (2021) compared the performances of tree models and kernel-based support vector machine (SVM) in estimating WQI at the Lam Tsuen River in Hong Kong. Parsaie et al. (2018) assessed the ability of different ANN variants as well as the SVM...
for the prediction of different water quality parameters. However, it was found that the performance of such base models was not satisfactory and hence improvements needed to be carried out on the design of the machine learning models.

Kisi & Parmar (2016) utilized the least-square support vector machine (LSSVM) for the prediction of the chemical oxygen demand (COD) and compared its performance with the multivariate adaptive regression splines (MARS) as well as the M5 model tree. Similar work was also conducted at the Perak River Basin of Malaysia for WQI modelling (Leong et al. 2019). The authors of the mentioned works opined that the LSSVM model is well-suited for addressing problems with high non-linearity. Besides, the shortcomings of the LSSVM model were also detected and it was recommended that further improvement through hybridization in order to optimize its hyperparameters ought to be considered. This effort was done by numerous researchers through incorporating different optimization algorithms to automate the tuning of LSSVM model's hyperparameters (Bozorg-Haddad et al. 2017; Yaseen et al. 2018; Sun et al. 2019; Song et al. 2021).

In recent years, a new class of optimization algorithms, known as the advanced optimization algorithms, have emerged due to the rapid development in soft computing technology. The hybrid of particle swarm optimization and genetic algorithm (HPSOGA) was proposed by Wang & Si (2010) with the intent to achieve mutual complementation between the particle swarm optimization (PSO) and the genetic algorithm (GA). The PSO has high convergence speed but often fails to adjust its velocity step size correctly. This can be solved by incorporating the GA, which can reflect the granularity of the search space via dynamic evolution or mutation. In exchange, the slow GA can be compensated by the efficient PSO. Besides, Tong (2020) introduced the whale optimization algorithm based on the self-adapting parameter adjustment and mix mutation strategy (SMWOA), which had evolved from the whale optimization algorithm (WOA) developed by Mirjalili & Lewis (2016). The SMWOA improves the original WOA by including self-adjustment parameters and mix mutation strategy. The former involves changing the progressive parameter in WOA into a self-adjustment parameter to ensure the global search ability remains constant, whereas the latter balances the trade-off between the ‘exploration’ and ‘exploitation’ phases to prevent premature convergence. Tian et al. (2019) developed the ameliorative moth-flame optimization (AMFO) based on the Kent chaotic map search strategy and dynamic inertia weight. The researchers are confident that such modifications will be able to help the original moth-flame optimization (MFO) to escape to converging into the local optima.

To this current day, there is no study or publication that involves the use of advanced optimization algorithms on machine learning models for the prediction of the WQI. Yet, many of the previous studies claimed that the integration of conventional optimization algorithms can be a promising meta-heuristic approach for boosting the performance of machine learning models (Chia et al. 2021). Therefore, this work proposes a superior hybrid LSSVM model for sustainable river water quality management. This study pioneered the integration of advanced optimization algorithm on the LSSVM model, whereby the resulting model shall be beneficial for water quality monitoring of the Klang River in Selangor, Malaysia. The specific objectives of this study are as follows:

1. Select the most suitable kernel function of the LSSVM model by comparing their prediction accuracies, to formulize the best base model for use.
2. Enhance the performance of the hybrid LSSVM models in WQI prediction through the integration of various advanced optimization algorithm (HPSOGA, SMWOA and AMFO).
3. Investigate how the combinations of input water quality parameters can affect the performance of the base and hybridized LSSVM model.

**METHODS**

**Study area and data**

The study was carried out at the Klang River in Selangor, Malaysia. The total length of the Klang River is 120 km and its basin covers an area of 1,280 km². The Klang River begins from the Ulu Gombak Forest Reserve and flows in the west direction until it discharges into the Straits of Malaka. Due to human activities, the Klang River has become one of the most polluted rivers in Selangor. The chosen water sampling station is the 1K08 station shown in Figure 1. The 1K08 station is located in the middle of Kuala Lumpur city where the surrounding land use is mainly commercial and the river water is deemed consistently polluted under the WQI standards of the Department of Environment, Malaysia (DOE). This has subsequently captured the interest of the authors to adopt the Klang River as the study site. The water quality data, which included dissolved oxygen (DO, %), biological oxygen demand (BOD, mg/L), chemical oxygen demand (COD, mg/L), suspended solids (SS, mg/L) ammoniacal nitrogen (NH₃-N, mg/L) and pH spans from year 1999 to year 2018 were provided by the DOE.
The machine learning models developed in this study were trained using supervised learning, where a target label had to be provided to the models. In this study, the WQI which acted as the target were calculated based on the standard equation published by the DOE, as shown in Equation (1) (Ahmad et al. 2016).

\[
WQI = 0.22SI_{DO} + 0.1SI_{BOD} + 0.16SI_{COD} + 0.16SI_{SS} + 0.15SI_{AN} + 0.12SI_{pH}
\]  

(1)

where \(SI_{DO}\) is the sub-index of DO, \(SI_{BOD}\) is the sub-index of BOD, \(SI_{COD}\) is the sub-index of COD, \(SI_{SS}\) is the sub-index of SS, \(SI_{AN}\) is the sub-index of NH\(_3\)-N and \(SI_{pH}\) is the sub-index of pH. The derivation of the sub-indices can be referred to in the guidelines published and can be accessed through the DOE portal.

**Data pre-processing**

The input data have to be normalized in order to avoid any effect of absolute scale (Feng et al. 2017). The normalization can be done by rescaling the inputs into the scale of 0–1 based on the minimum and maximum values. On top of that, the k-fold cross validation strategy was applied to reduce the risk of overfitting. The data were partitioned into five equal portions where one portion would be used as testing data for each fold.

**Least-square support vector machine**

The LSSVM model simplifies the quadratic approach in conventional SVM into linear equations as shown in Equation (2).

\[
f(x) = w^T \varphi(x) + b
\]  

(2)
where \( w^T \) is the transposed weight vector, \( \varphi(x) \) is the non-linear kernel function and \( b \) is the bias term. The cost function is used to determine \( w^T \) and \( b \), as shown in Equation (3).

\[
CF = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2
\]  

(3)

where \( \gamma \) is the regularization parameter, \( N \) is the number of data points and \( e_i \) is the residual vector for input data. The cost function is constrained by Equation (4).

\[
y_i = w^T \varphi(x_i) + b + e_i
\]  

(4)

where \( y_i \) is the output and \( x_i \) is the \( i \)th input. Equations (3) and (4) form a convex optimization problem which can be solved by the Lagrange Multiplier method illustrated in Equation (5).

\[
L(w, b, e, \alpha) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i(w^T \varphi(x_i) + b + e_i) - y_i
\]  

(5)

where \( \alpha \) is the Lagrange Multiplier and can be computed using Equation (6), which shows the partial derivative with consideration of \( w, b, e \) and \( \alpha \).

\[
y = \sum_{i=1}^{N} \alpha_i \varphi(x) \varphi(x_i) + b
\]  

(6)

Equations (7)–(9) show the linear, polynomial (degree of three) and RBF kernel functions, respectively.

\[
K(x, y) = x^T y + c, c \geq 0
\]  

(7)

\[
K(x, y) = (x^T y + c)^3, c \geq 0
\]  

(8)

\[
K(x, y) = \exp \left( -\frac{|x - x^T|^2}{\sigma^2} \right)
\]  

(9)

where \( c \) and \( \sigma \) are hyperparameters that have to be fine-tuned.

**Hybrid of particle swarm optimization and genetic algorithm**

The HPSOGA updates the swarm population by incorporating the selection, crossover and mutation of the GA. The initial swarm population is generated randomly within the search space. Subsequently, the swarm will be sorted according to root mean square error (RMSE), which acts as the fitness value for optimization. Only the best portion of the population will be selected for crossover as well as mutation, and the portion will grow as the iteration increases. The overall mechanism of the HPSOGA can be described mathematically as in Equations (10) and (11).

\[
v_{j, t+1} = w v_{j, t} + c_1 \text{rand}_1 (p_{best, t}(GA) - k_{j, t}) + c_2 \text{rand}_2 (g_{best}(GA) - k_{j, t})
\]  

(10)

\[
k_{j, t+1} = k_{j, t} + v_{j, t+1}, k_{\text{min}} < k_{j, t+1} < k_{\text{max}}
\]  

(11)

where \( j \) is the particle in the swarm population, \( t \) is the number of iteration, \( v_j \) is the velocity vector of the \( j \)th particle, \( w \) is the inertia weight, \( c_1 \) and \( c_2 \) are the learning rates, \( \text{rand}_1 \) and \( \text{rand}_2 \) are uniformly distributed random variables between 0 and 1, \( k_{j, t} \) is the position vector of the \( j \)th particle at iteration \( t \), \( p_{best}(GA) \) is the best solution after the GA operation and \( g_{best}(GA) \) is the global best solution of the swarm after the GA operation. For study purposes, the crossover rate is set at 0.7 and the mutation is set at 0.3.

**Whale optimization algorithm based on self-adapting parameter adjustment and mix mutation strategy**

The working principle of the SMWOWA can be expressed as shown in Equation (12) through Equation (20). Similarly, the RMSE was used as the fitness value to optimize the objective function.

\[
x_{ij} = lb_j + \text{rand} \cdot (ub_j - lb_j)
\]  

(12)
where $x_{ij}$ is the $i$th search agent (whale) in the $j$th dimension, rand is a random number between 0 and 1 whereas $lb_j$ and $ub_j$ are the lower and upper boundaries of the $j$th dimension. After applying the mix mutation strategy, the ‘exploration’ and ‘exploitation’ phases are dictated by the fitness value. During the ‘exploitation’ phase, the shrinking net bubble is modelled using Equation (13).

$$X'(t + 1) = X^*(t) + A$$  \hspace{1cm} (13)$$

where $X^*(t + 1)$ and $X^*(t)$ are the position of the whales at different iterations whereas $A$ is converted into a self-adapting parameter using Equations (14)–(16).

$$A_i = \text{normrnd}(\mu, d)$$  \hspace{1cm} (14)$$

$$
\mu = \frac{|x_{ij} - X_j^*|}{|ux_j - lx_j|}
\hspace{1cm} (15)$$

$$d = 1 - \frac{t}{t_{\text{max}}}
\hspace{1cm} (16)$$

where the parameter $A$ is generated randomly within a normal distribution that has a mean, $\mu$, and standard deviation, $d$. The value of $d$ slowly converges from 1 to 0 along the progress of the algorithm. When the SMWOA is not executing the ‘exploitation’ phase, the ‘exploration’ phase will take place as shown in Equations (17) and (18).

$$X_i(t + 1) = X_{\text{other}}(t) + A \cdot \text{Dis}$$  \hspace{1cm} (17)$$

$$\text{Dis} = |C \cdot X_{\text{other}}(t) - X_i(t)|$$  \hspace{1cm} (18)$$

where $\text{Dis}$ is the distance between $X_i(t)$ and $X_{\text{other}}(t)$ and $C$ is a random coefficient that ranges between 0 and 2. In the spiral updating position, the position of the whales are updated using Equations (19) and (20).

$$X_i(t + 1) = \text{Dis} \cdot e^{bl} \cos(2\pi L) + X^*(t)$$  \hspace{1cm} (19)$$

$$\text{Dis} = |X^*(t) - X_i(t)|$$  \hspace{1cm} (20)$$

where $\text{Dis}$ is the distance between $X_i(t)$ and $X^*(t)$, $b$ is the spiral constant and $L$ is a random number between $-1$ and 1. The encircling process (with shrinkage) is repeated until termination of the algorithm.

**Ameliorative moth-flame optimization**

The AMFO is an improvisation based on the original moth-flame optimization (MFO) by using the Kent chaotic strategy and dynamic inertia weight. The moth population and flames is first initialised as shown in Equations (21) and (22).

$$M = \begin{bmatrix}
    m_{1,1} & \cdots & m_{1,j} \\
    \vdots & \ddots & \vdots \\
    m_{i,1} & \cdots & m_{i,j} 
\end{bmatrix}
\hspace{1cm} (21)$$

$$F = \begin{bmatrix}
    f_{1,1} & \cdots & f_{1,j} \\
    \vdots & \ddots & \vdots \\
    f_{i,1} & \cdots & f_{i,j} 
\end{bmatrix}
\hspace{1cm} (22)$$

where $m$ and $f$ are moths and flames, respectively. The first value of the chaotic sequence, $Z_0$ is obtained using Equation (23).

$$Z_0 = \frac{M_i - M_{\text{min}}}{M_{\text{max}} - M_{\text{min}}}$$  \hspace{1cm} (23)$$
where $M_i$ is the best moth, whereas $M_{\text{max}}$ and $M_{\text{min}}$ are the moth with the largest and smallest values, respectively. The chaotic sequence is then generated using Equation (24).

\[
Z_{i+1} = \begin{cases} 
\frac{Z_i}{a}, & 0 < Z_i < a \\
(1 - Z_i)(1 - a), & a \leq Z_i < 1
\end{cases}
\]  

where $a = 0.4$. Finally, the individual location $U_i$ is produced by Equation (25). If $U_i$ has a better fitness value than $M_i$, then it will replace $M_i$ and the $F_i$ will be updated using Equations (26) and (27), which incorporate the dynamic inertia weight, $w_{ij}$.

\[
U_i = M_i + \frac{M_{\text{max}} - M_{\text{min}}}{2} \cdot (2Z_i - 1)
\]

\[
M_{i+1} = w_{ij}[f_j - M_{i}]e^{bt} \cdot \cos(2\pi t) + (1 - w_{ij})F_j
\]

\[
w_{ij} = \frac{\exp(f_j) - \mu}{2.4 + \exp\left(-\frac{f_j}{\mu}\right)}
\]

where $\mu$ is the weightage parameter, $b$ is the spiral flight shape and $t$ is the current iteration. The adaptive number of flames, $N_t$ is updated using Equation (28).

\[
N_t = \text{round}\left(N - t \frac{N - 1}{T}\right)
\]

where $N$ is the initial number of flames, $t$ is the current iteration and $T$ is the maximum iteration. Similar to the SMWOA, the AMFO would iteratively operate on until it satisfies the pre-set termination criteria.

The various resultant hybrid LSSVM models developed via the appropriate integration of the HPSOGA, SMWOA or AMFO, respectively were labelled as HPSOGA-LSSVM, SMWOA-LSSVM and AMFO-LSSVM. For all the advanced optimization algorithms, 10 search agents are assigned with a maximum iteration of 100.

**Performance evaluation**

In order to assess the performance of the base and hybrid models, several performance evaluation metrics were used. The mean absolute error (MAE), root mean square error (RMSE) and the mean absolute percentage error (MAPE) were used to compare the difference between the predicted and actual values of WQI, with RMSE functions to capture possible large errors. On the other hand, the coefficient of determination ($R^2$) measured how well the model is fitted to the data. The equation for calculating RMSE, MAE, MAPE and $R^2$ are shown in Equations (29)–(32), respectively.

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{\text{actual},i} - y_{\text{predicted},i})^2}
\]

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_{\text{actual},i} - y_{\text{predicted},i}|
\]

\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_{\text{actual},i} - y_{\text{predicted},i}|}{y_{\text{actual},i}}
\]

\[
R^2 = 1 - \left(\frac{\sum_{i=1}^{N} (y_{\text{actual},i} - \bar{y}_{\text{predicted}})^2}{\sum_{i=1}^{N} (y_{\text{actual},i} - \bar{y}_{\text{actual}})^2}\right)^2
\]

where $y_{\text{actual}}$ and $y_{\text{predicted}}$ each represent the actual and predicted values of WQI. However, individual performance evaluation metrics cannot assess the model from all aspects and it is difficult to make comparison among the models. Therefore,
the global performance indicator (GPI) was used as the comprehensive metric to compare the models (Despotovic et al. 2015), as shown in Equation (33).

\[
GPI = \sum_{i=1}^{4} \alpha_j (\bar{y}_j - y_{ij})
\]

(33)

where \( \alpha_j \) is the coefficient of the \( j \)th indicator (equals to 1 for indicators except of \( R^2 \), which is \(-1\), \( \bar{y}_j \) is the scaled median value of the \( j \)th indicator and \( y_{ij} \) is the scaled \( j \)th indicator of the \( i \)th model. The overall work flow of this study is presented in Figure 2.

![Figure 2 | Methodology of the study.](http://iwaponline.com/ws/article-pdf/22/2/1951/1009250/ws022021951.pdf)
RESULTS AND DISCUSSION

Preliminary screening and effect of kernel functions

During the preliminary screening, a total of 63 different combinations of input parameters were tested for their suitability in the development of the LSSVM models. The best combinations of the different number of input parameters were selected to be presented in this paper. Initially, all the six water quality parameters were fed into the LSSVM models for WQI prediction (namely C6). Then, the number of input parameters were reduced one at a time to create the C5, C4, C3, C2 and C1 arbitrary combinations accordingly. The pH input data was first discounted, followed next by the NH3-N, SS, BOD and lastly COD until when DO is the sole input parameter left, as shown in Table 1. The performance of LSSVM models using different kernel functions and input combinations is summarized in Table 2.

The three oxygen-based water quality parameters, namely DO, COD and BOD, appear to be the most important inputs for accurate modelling and prediction of WQI. This was due to the fact that the oxygen dynamics in the river water reflected the suitability of the water for the aqua-ecosystem. The oxygen-based parameters were followed by the SS which represented the inorganic composition, and possibly the turbidity of the river water. Meanwhile, the NH3-N (organic representation) and pH were deemed the least important parameters, and had minimal influence on the accuracy of models.

From Table 2, it can be clearly seen that the LSSVM model using the RBF kernel function, which had the best performance in predicting the WQI. This can be observed from the lower error metrics and also the higher values of $R^2$ as well as the GPI. From the results, one can observe that the linear kernel function failed to predict the WQI satisfactorily. This could be due to the highly non-linear nature of the interactions among the water quality parameters. When the kernel functions were switched to polynomial (degree of 3) and RBF, the accuracy of the prediction improved drastically, with the effect of input combinations becoming more prominent. The increase in the number of input parameters from C1 to C6 evidently enhanced the performance of the LSSVM models, as shown by the increasing value of the GPI. Thus, the LSSVM model with the RBF kernel function was selected as the base model for hybridization purposes at the next stage, as it had achieved the higher average GPI value (0.08) compared to those of LSSVM models with linear (−2.32) and polynomial (−1.22) kernel functions. This could be due to the Gaussian function, which can map the input better into the feature space.

Optimized LSSVM models

In this next stage, the LSSVM model with RBF kernel function (known as the base model) was hybridized using the various advanced optimization algorithms. To be exact, the respective advanced optimization algorithms (HPSOGA, SMWOA and AMFO) were each separately integrated outside of the LSSVM model, so that they could receive feedback from the LSSVM models in terms of the fitness value. Based on the fitness value received, the advanced optimization algorithms would react by searching for the global optima for the hyperparameters of the LSSVM model. Since the RBF was the chosen kernel function, the hyperparameters tuned were $\gamma$ and $\sigma$ (refer to Equations (3) and (9)), in which the boundaries (search spaces) were set to be in between 0.001 and 100. Table 3 summarizes the results of the performance of the three resultant optimized LSSVM models, namely the HPSOGA-LSSVM, SMWOA-LSSVM and AMFO-LSSVM.

If comparison is made between Tables 2 and 3, the evidence that the advance optimization algorithms could improve the base LSSVM model is explicit. Taking the C6 case as an example, the RMSE, MAE, $R^2$, MAPE and GPI of base LSSVM model were 6.18, 4.25, 0.87, 8.66% and 0.24, respectively. On the other hand, the same metrics for the HPSOGA-LSSVM, SMWOA-LSSVM and AMFO-LSSVM models were 3.96, 2.86, 0.90, 5.69% and 0.57, 3.93, 2.86, 0.90, 5.62% and 0.57 as

<table>
<thead>
<tr>
<th>Combination</th>
<th>Water quality parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>DO</td>
</tr>
<tr>
<td>C2</td>
<td>DO, COD</td>
</tr>
<tr>
<td>C3</td>
<td>DO, BOD, COD</td>
</tr>
<tr>
<td>C4</td>
<td>DO, BOD, COD, SS</td>
</tr>
<tr>
<td>C5</td>
<td>DO, BOD, COD, SS, NH3-N</td>
</tr>
<tr>
<td>C6</td>
<td>DO, BOD, COD, SS, NH3-N, pH</td>
</tr>
</tbody>
</table>
well as 4.94, 3.46, 0.90, 6.96% and 0.44, respectively. Similar observations were obtained for the models trained with different input combinations. The dispersion and distribution of the search agents initialized by the advanced optimization algorithms resolved the constraints caused by the random selection of initial point by the base LSSVM model, thus resulting in better WQI predictions.

Although the HPSOGA-LSSVM, SMWOA-LSSVM and AMFO-LSSVM models outperformed their base LSSVM model counterpart, a comparison is still needed to examine the efficiency and effectiveness of each of the individual advanced optimization algorithms. The HPSOGA-LSSVM model achieved an average GPI value of 0.24, whereas the average GPI value for SMWOA-LSSVM and AMFO-LSSVM models are 0.31 and 0.26, respectively. From the macro aspect (in terms of GPI), the SMWOA-LSSVM model could predict the WQI better than the HPSOGA-LSSVM models. This could be due to the nature of the SMWOA, which incorporates ‘progressive algorithms’, whereby the potential search space is gradually shrinking as the iteration increases. This noble characteristic is not to be found in the HPSOGA, which focuses on the survival of fitter swarms, or that of the AMFO, which focuses on the optimization’s computational efficiency instead of the final results. In order to avoid premature convergence in the shrinking search space, the SMWOA further adopts the rebalancing between the ‘exploration’ and ‘exploitation’ phases so that the global optimum could be located without lengthening the optimization time. In other words, the SMWOA actually shows its upper hand over the HPSOGA (retaining fitter swarm analogous to shrinking potential search space) and the AMFO (rebalancing the phases to avoid premature convergence analogous to optimizing the computational efficiency). This in turn, has resulted in the better performance of the SMWOA-LSSVM model as compared to the other optimized models.

### Table 2 | Performance of the LSSVM using different kernel functions for WQI prediction

<table>
<thead>
<tr>
<th>LSSVM (Linear)</th>
<th>Combination</th>
<th>RMSE</th>
<th>MAE</th>
<th>R^2</th>
<th>MAPE (%)</th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>19.12</td>
<td>13.58</td>
<td>0.66</td>
<td>29.42</td>
<td>–1.93</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>15.33</td>
<td>10.72</td>
<td>0.53</td>
<td>20.62</td>
<td>–1.40</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>22.74</td>
<td>16.29</td>
<td>0.48</td>
<td>31.71</td>
<td>–2.55</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>22.25</td>
<td>17.31</td>
<td>0.39</td>
<td>32.54</td>
<td>–2.73</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>20.90</td>
<td>16.77</td>
<td>0.39</td>
<td>31.17</td>
<td>–2.58</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>21.27</td>
<td>17.47</td>
<td>0.33</td>
<td>31.80</td>
<td>–2.73</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LSSVM (Polynomial)</th>
<th>Combination</th>
<th>RMSE</th>
<th>MAE</th>
<th>R^2</th>
<th>MAPE (%)</th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>25.71</td>
<td>17.72</td>
<td>0.00</td>
<td>34.61</td>
<td>–3.41</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>17.24</td>
<td>12.73</td>
<td>0.28</td>
<td>23.22</td>
<td>–1.98</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>14.58</td>
<td>9.68</td>
<td>0.40</td>
<td>17.41</td>
<td>–1.32</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>8.78</td>
<td>6.24</td>
<td>0.70</td>
<td>11.32</td>
<td>–0.29</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>7.40</td>
<td>5.54</td>
<td>0.77</td>
<td>10.63</td>
<td>–0.08</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>7.35</td>
<td>5.49</td>
<td>0.64</td>
<td>10.79</td>
<td>–0.22</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LSSVM (RBF)</th>
<th>Combination</th>
<th>RMSE</th>
<th>MAE</th>
<th>R^2</th>
<th>MAPE (%)</th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>8.28</td>
<td>6.57</td>
<td>0.52</td>
<td>13.08</td>
<td>–0.55</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>6.73</td>
<td>5.11</td>
<td>0.76</td>
<td>10.17</td>
<td>–0.02</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>6.01</td>
<td>4.55</td>
<td>0.80</td>
<td>8.78</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>5.55</td>
<td>3.83</td>
<td>0.89</td>
<td>7.58</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>5.88</td>
<td>3.92</td>
<td>0.91</td>
<td>8.24</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>6.18</td>
<td>4.25</td>
<td>0.87</td>
<td>8.66</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>
From the micro scale point of view, as the number of input water quality parameters increase, the performance of the optimized LSSVM models would also be improved. This property was actually inherited from the base LSSVM model using the RBF kernel function. However, the improvements to the optimized model due to the increase in number of input parameters would reach an optimum after which it became less significant when further parameter inputs are made, say for example from C4 to C6. That is to say, after the utilization of the four water quality parameter suites (DO, BOD, COD and SS), the five-parameter option (plus NH3-N) and six-parameter option (plus NH3-N and pH) were not reflected in significantly improved performance of the optimization models. This finding can be taken positively as this indicates that the optimized LSSVM models could predict equally well, regardless of the number of input parameters, as long as at least four water quality parameters are fed into the optimized LSSVM models, particularly the HPSOGA-LSSVM and SMWOA-LSSVM models. That is to say the optimal number of inputs would be four. On top of that, the MAPE of the HPSOGA-LSSVM and SMWOA-LSSVM models using C4, C5 and C6 are at most 6%. Although DO appears to be the most important water quality parameter for WQI prediction using the optimized LSSVM models; however, WQI prediction depending solely on DO was not reliable due to the high error incurred. In other words, complementary water quality parameters such as the BOD, COD and SS are mandatory for good prediction, irrespective of the model used. At this stage, using the optimized LSSVM models, at least three water quality parameters are confidently needed in order to achieve WQI prediction with MAPE less than 10%.

The comparison between all the LSSVM models (base and optimized) in this study is illustrated in Figure 3. The performance of the advanced optimization algorithm improved the base LSSVM model, regardless of the input combinations fed. The SMWOA-LSSVM model stood out among the hybridized model and is suggested to be used as a predictive model for WQI at the Klang River.

### Table 3 | Performance of hybrid optimized LSSVM models for WQI predicting

<table>
<thead>
<tr>
<th>Combination</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>MAPE (%)</th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>7.51</td>
<td>5.86</td>
<td>0.63</td>
<td>11.76</td>
<td>−0.30</td>
</tr>
<tr>
<td>C2</td>
<td>7.81</td>
<td>6.17</td>
<td>0.74</td>
<td>12.15</td>
<td>−0.23</td>
</tr>
<tr>
<td>C3</td>
<td>5.23</td>
<td>3.87</td>
<td>0.84</td>
<td>7.53</td>
<td>0.31</td>
</tr>
<tr>
<td>C4</td>
<td>4.39</td>
<td>3.08</td>
<td>0.90</td>
<td>6.03</td>
<td>0.52</td>
</tr>
<tr>
<td>C5</td>
<td>4.12</td>
<td>2.96</td>
<td>0.91</td>
<td>5.84</td>
<td>0.56</td>
</tr>
<tr>
<td>C6</td>
<td>3.96</td>
<td>2.86</td>
<td>0.90</td>
<td>5.69</td>
<td>0.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Combination</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>MAPE (%)</th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>6.97</td>
<td>5.46</td>
<td>0.66</td>
<td>10.46</td>
<td>−0.17</td>
</tr>
<tr>
<td>C2</td>
<td>6.67</td>
<td>5.06</td>
<td>0.78</td>
<td>10.22</td>
<td>0.01</td>
</tr>
<tr>
<td>C3</td>
<td>5.23</td>
<td>3.87</td>
<td>0.84</td>
<td>7.53</td>
<td>0.31</td>
</tr>
<tr>
<td>C4</td>
<td>4.33</td>
<td>3.07</td>
<td>0.90</td>
<td>6.01</td>
<td>0.53</td>
</tr>
<tr>
<td>C5</td>
<td>4.10</td>
<td>2.95</td>
<td>0.92</td>
<td>5.83</td>
<td>0.57</td>
</tr>
<tr>
<td>C6</td>
<td>3.95</td>
<td>2.90</td>
<td>0.90</td>
<td>5.62</td>
<td>0.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Combination</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>MAPE (%)</th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>7.09</td>
<td>5.67</td>
<td>0.65</td>
<td>10.80</td>
<td>−0.21</td>
</tr>
<tr>
<td>C2</td>
<td>6.67</td>
<td>5.06</td>
<td>0.78</td>
<td>10.22</td>
<td>0.01</td>
</tr>
<tr>
<td>C3</td>
<td>4.96</td>
<td>3.71</td>
<td>0.85</td>
<td>7.26</td>
<td>0.36</td>
</tr>
<tr>
<td>C4</td>
<td>5.18</td>
<td>3.93</td>
<td>0.91</td>
<td>7.66</td>
<td>0.38</td>
</tr>
<tr>
<td>C5</td>
<td>4.14</td>
<td>3.01</td>
<td>0.91</td>
<td>5.93</td>
<td>0.55</td>
</tr>
<tr>
<td>C6</td>
<td>4.94</td>
<td>3.46</td>
<td>0.90</td>
<td>6.96</td>
<td>0.44</td>
</tr>
</tbody>
</table>
A check on similar studies that had been carried out at the Klang River using other machine learning models is made here. According to the results reported by Hameed et al. (2016), although the RBF neural network performed better than the SMWOA-LSSVM model of this work (both using C6 as input combination; however, the accuracy of the RBF neural network

**Figure 3** | Comparison of base and optimized LSSVM models.

A check on similar studies that had been carried out at the Klang River using other machine learning models is made here. According to the results reported by Hameed et al. (2016), although the RBF neural network performed better than the SMWOA-LSSVM model of this work (both using C6 as input combination; however, the accuracy of the RBF neural network...
degraded as the number of input water quality parameters decreased. For inputs with limited parameters, the SMWOA-LSSVM model of this work performed better than the RBF neural network as well as the back-propagation neural network. Besides, the SMWOA-LSSVM model developed in this study also performed similarly to the hybridized random forest, conditional random forest, random forest generator and extreme gradient boosting models reported by Tiyasha et al. (2021), although the models in the latter study used significantly more input water quality parameters. This means that the LSSVM models hybridized using advanced optimization algorithms are more resilient towards low numbers of input water quality parameters, which is more useful for river quality monitoring and management.

**CONCLUSIONS**

In this study, the three different novel hybrid LSSVM models based on the individual separate integration of three advanced optimization algorithms, namely the HPSOGA, SMWOA and AMFO with the initially selected base model (LSSVM with the RBF kernel function) were developed for WQI prediction at station 1K08 of the Klang River in Kuala Lumpur. It was found that the optimized LSSVM models had better performance compared to the standalone base model, as portrayed by the lower RMSE, MAE and MAPE as well as higher $R^2$ and GPI. Next, among the optimized LSSVM models, the SMWOA-LSSVM had the best performance, followed by the HPSOGA-LSSVM and then AMFO-LSSVM. The RMSE, MAE, $R^2$, MAPE and $R^2$ of WQI prediction of the SMWOA-LSSVM are 3.93, 2.90, 0.90, 5.62% and 0.57, which exhibits great significant improvement based on its base counterpart (6.18, 4.25, 0.87, 8.66% and 0.24) when C6 was used as the input combination. Similar observations were seen for the other input combinations. The SMWOA encompasses the progressive reduction of search space (faster convergence) as well as the trade-off between the ‘exploration’ and ‘exploitation’ phases (prevent premature convergence). This can help to tune the hyperparameters of the LSSVM so that minimal loss can be achieved with optimum performance of WQI prediction. On the contrary, the noble characteristics found in the SMWOA that are not present in both the HPSOGA and AMFO led to the relatively weaker performance of the latter two models. In general, the SMWOA-LSSVM produced reliable predictions of the WQI at station 1K08 of the Klang River, as long as the minimal input of the DO, BOD and COD are included in the input combination during the training of the optimized LSSVM model. Having said that, the validity of this study is still being constrained at this one sampling station along the Klang River. More comprehensive work, which involves greater spatial variability, should be undertaken in future and would undoubtedly justify the integration of advanced optimization algorithms onto the LSSVM model.

**ACKNOWLEDGEMENTS**

This research was funded by Universiti Tunku Abdul Rahman Research Fund (UTAR), Malaysia through Universiti Tunku Abdul Rahman Research Fund under project number IPSR/RMC/UTARRF/2020-C2/K03. The authors would also like to express their gratitude to the Department of Environment, Malaysia for their provision of water quality data.

**DATA AVAILABILITY STATEMENT**

Data cannot be made publicly available; readers should contact the corresponding author for details.

**REFERENCES**


First received 4 July 2021; accepted in revised form 27 August 2021. Available online 13 September 2021.