IRAQI Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science (2026) 16 (2): 746-758

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN:2222-2995 E-ISSN:3079-3521
University of Kirkuk Journal For Administrative and Economic Science

Ahmed Lana Muhammed & Mhamad Aras Jalal. Assessing Quality of Water Using Twin Support Vector Regression for Total Dissolved Solids in Drinking Water. ***University of Kirkuk Journal For Administrative and Economic Science*** **(2026) 16 (2):746-758.**

Assessing Quality of Water Using Twin Support Vector Regression for Total Dissolved Solids in Drinking Water

Lana Muhammed Ahmed [1], Aras Jalal Mhamad [2]

[1,2] *1Statistics and Informatics, college of Administration and Economics, University of Sulaimani, Sulaimani, Iraq*

lana.ahmed@univsul.edu.iq [1], aras.mhamad@univsul.edu.iq [2]

Abstract: Drinking water quality assessment is essential for protecting public health and ensuring sustainable water resources. Accurate modeling of key indicators such as Total Dissolved Solids (TDS) remains challenging due to complex interactions among hydro chemical parameters. Reliable prediction of Total Dissolved Solids (TDS) is critical for quantitative assessment of drinking water quality. This study presents a statistically grounded framework using Twin Support Vector Regression (TSVR) for modeling TDS as a function of multiple hydro chemical indicators. Both linear and kernel-based TSVR models, including Polynomial and Radial Basis Function (RBF) kernels, are formulated and systematically optimized via hyperparameter tuning to minimize prediction error. Model performance is evaluated on real-world water quality datasets using standard statistical metrics, including mean squared error and coefficient of determination. Results indicate that Linear TSVR provides optimal generalization with minimal computational complexity, whereas RBF-TSVR effectively captures nonlinear dependencies among hydro chemical parameters. These findings confirm that TSVR constitutes a statistically efficient and computationally tractable alternative to classical Support Vector Regression, offering guidance for kernel selection, hyperparameter calibration, and robust predictive modeling in water quality assessment.

Keywords: Water quality, Total Dissolved Solids, Twin Support Vector Regression, Machine learning.

تقييم جودة المياه باستخدام انحدار متجهات الدعم التوأم لإجمالي المواد الصلبة الذائبة في مياه الشرب

الباحثة: لانه محمد احمد[١] ، أ.م.د. اراس جلال محمد [٢]

[١] *جامعة السليمانية، كلية الإدارة والاقتصاد*

lana.ahmed@univsul.edu.iq [1], aras.mhamad@univsul.edu.iq [2]

المستخلص: يعد تقييم جودة مياه الشرب أمراً ضرورياً لحماية الصحة العامة وضمان استدامة الموارد المائية. لا تزال النمذجة الدقيقة للمؤشرات الرئيسية، مثل إجمالي المواد الصلبة الذائبة(TDS) ، تشكل تحدياً بسبب التفاعلات المعقدة بين المعايير الهيدروكيميائية. ويعتبر التنبؤ الموثوق بإجمالي المواد الصلبة الذائبة أمراً حاسماً للتقييم الكمي لجودة مياه الشرب. تقدم هذه الدراسة إطاراً قائماً على الأسس الإحصائية باستخدام انحدار ناقل الدعم المزدوج (TSVR)لنمذجة (TDS) كدالة لمتغيرات هيدروكيميائية متعددة. تم صياغة نماذج (TSVR) الخطية والقائمة على الدوال اللبية(Kernel-based) ، بما في ذلك الدوال متعددة الحدود ودالة الأساس الشعاعي(RBF) ، وتحسينها بشكل منهجي من خلال ضبط المعلمات الفائقة لتقليل خطأ التنبؤ.

تم تقييم أداء النموذج على مجموعات بيانات حقيقية لجودة المياه باستخدام مقاييس إحصائية قياسية، بما في ذلك متوسط مربع الخطأ ومعامل التحديد. تشير النتائج إلى أن نموذج (TSVR) الخطي يوفر تعميماً مثالياً مع حد أدنى من التعقيد الحسابي، بينما نجح نموذج (RBF-TSVR) في التقاط الاعتمادات غير الخطية بين المعايير الهيدروكيميائية بفعالية. تؤكد هذه النتائج أن (TSVR) يشكل بديلاً فعالاً إحصائياً وسلساً حسابياً لنماذج انحدار ناقل الدعم الكلاسيكية، مما يوفر إرشادات لاختيار الدالة اللبية، ومعايرة المعلمات الفائقة، والنمذجة التنبؤية القوية في تقييم جودة المياه.

الكلمات المفتاحية: جودة المياه، إجمالي المواد الصلبة الذائبة، انحدار ناقل الدعم المزدوج(TSVR) ، التعلم الآلي.

Corresponding Author: E-mail: lana.ahmed@univsul.edu.iq

Introduction:

Surface water bodies, particularly rivers, they are at high risk. to pollution due to their continuous influence of natural processes and human activities. Climate conditions have main role in influencing the hydrochemical characteristics of rivers. Geological, and watershed practice (Singh et al., 2011). Therefore, it is essential that ongoing monitoring of our river water quality and prediction allows for better management of our lakes and rivers and usage of our available freshwater resources. The Support Vector Machine (SVM) is a supervised learning method used in classification and regression. It was developed by Vapnik in the early 1990s and utilizes kernel functions for modeling linear and nonlinear relationships. Its ability to generalize well makes SVM a very popular method.

The Support Vector Regression (SVR) is an extension of SVM when solving regression problems and has been successfully implemented in the field of water quality modeling. Such as, (Singh et al.,2011) used (SVC) Support Vector Classification and SVR on data collected over 15 years for several locations regarding surface water quality and found that SVR had a high level of accuracy in prediction for (BOD) biochemical oxygen demand, as measured by correlation coefficients of 0.907 to 0.952 and a low RMSE, and that non-linear methods performed better than standard linear methods. In spite of its high predictive performance, classical SVR solving a single large quadratic programming problem (QPP), which can be computationally demanding for large datasets. This limitation led (Jayadeva, et al., 2018) create the Twin Support Vector Machine (TSVM), which find two nonparallel hyperplanes instead of one, this allows significant reductions in computational complexity. This original concept was expanded to regression problems in the form of Twin Support Vector Regression (TSVR). TSVR method find two nonparallel regression functions that define the upper and lower bounds of the ε-insensitive zone. By solving two smaller optimization problems, Thus, the TSVR method produces fast training due to solving two smaller optimization problems as opposed to one large quadratic programming problem like the classical SVR algorithm while maintaining high-quality predictions (Peng, 2010), recent study concludes once again that employing a variety of optimization techniques can increase the performance of the SVR (Support Vector Regression) model. Methods of optimization that can be utilized are parameter selection and model generalization (Mhamad et al., 2025) and Genetic Algorithm based optimization for improving predictive accuracy in cross sectional datasets (Taher et al., 2025). Due to the excellent efficiency and learnability features of TSVR, many recent improvements have been developed as a result of increasing interest from researchers. Improvements of stability, speed, and capability of via noise are of particular focus at this time. (Huang et al.,2018) provided a comprehensive review of TSVR, including a description of its basis, major improvements, opportunities, current applications,

and plans for future research and development. (Khemchandani et al.,2015) developed a new model of regression called the Twin Support Vector Regression, this is achieved by deriving the model for regression directly from the classification framework and by solving two smaller Quadratic Programming Problems (QPPs) so the model trains at a faster rate when compared with traditional SVR methods but provides the same accuracy level as conventional support vector regressions. Additionally. (Zhang et al.,2020) have developed (GL-TLSSVR) Gauss–Laplace Twin Least Squares Support Vector Regression to solve mixed-noise distributions that frequently occur with real-world data sets for wind speed forecasting. The results show an improved capability for the GL-TLSSVR as compared with other methods.

The purpose of this study was to develop and evaluate double support vector regression (TSVR) models to predict total dissolved solids (TDS) in drinking water. We implemented both linear and non-linear (kernel based) TSVR models as two types of TSVR models within our integrated modelling environment (IME). Additionally, the study involved systematically optimizing hyperparameters and comparing the performances of the TSVR models and the classical SVR models, with the goal being to identify the best performing method of predicting water quality. By doing so, our objective was to demonstrate that the use of TSVR-based models is appropriate for simulating complex hadrochemical processes in addition to practical applications in water quality management.

1st: Methodology

1- Twin Support Vector Regression (TSVR)

TSVR is an efficient regression method. The way that it works similar to TSVM, both methods find two separate planes or (lines) instead of only one. But they have a different goal. Because TSVM is worked for classification, while TSVR is used for regression (peng,2010). These two boundaries function forms the lower boundary of the regression, and the other function forms the upper boundary are ε-insensitive, that allow a small margin (ε) where errors do not matter. The average of the two boundary functions that give final regression output. The first section of the optimization tries to make each boundary near to the training data, while keeping at least a distance ε from the data. If the boundary gets too close to the data, the slack variables measure these errors, and the second part of the objective punishes this error (peng,2010).

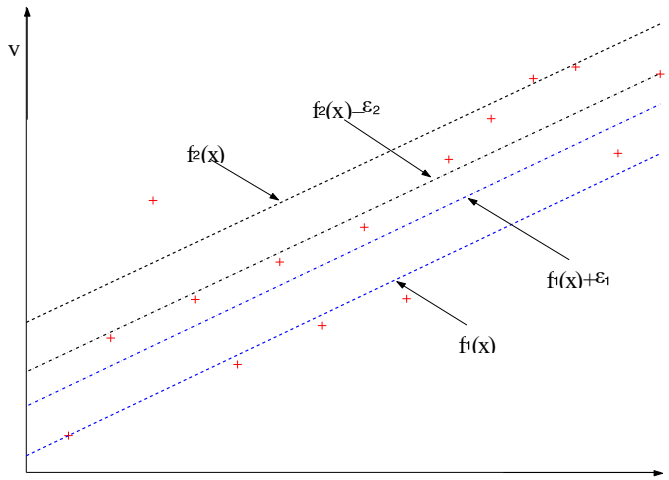

Figure (1): The geometric interpretation for TSVR (peng 2010).

TSVR solves the following pair of quadratic programming problems (peng 2010): Given training data (xi,yi), i=1,2,…,li = 1, where xi∈R^n and yi∈R , TSVR solves two smaller optimization problems:

Lower boundary function:

Min $_{w_1,b_1,\ \xi}$ $\frac{1}{2}\|(Y-e\varepsilon_1-(Aw_1+eb_1)\|^2+C_1\|\xi\|$ ...*(TSVR1)*

s.t. $(Y-(Aw_1+eb_1)\geq e\varepsilon_1-\xi,\quad \xi\geq 0$ ... *(1)*

boundary function:

Min $_{w_2,b_2,\ \eta}$ $\frac{1}{2}\|(Y+e\varepsilon_2-(Aw_2+eb_2)\|^2+C_2\|\eta\|$ ... (TSVR*2)*

s.t. $(Aw_2+eb_2)-Y\geq e\varepsilon_2-\eta\ ,\quad \eta\geq 0$ ...*(2)*

Where:

- A is the input matrix
- Y is the output vector
- C_1 and C_2 are regularization parameters
- ϵ_1 and ϵ_2 define the ε-insensitive tube

The final regression model is:

$$f(x)=\frac{1}{2}\ (f_1(x)+f_2(x))=\frac{1}{2}(w_1+w_2)^T x+\frac{1}{2}(b_1+b_2) \quad ...(3)$$

2- Kernel Twin Support Vector Regression

A kernel function performs a type of mathematical methods that allows an TSVR to take data that is naturally one dimensional and treat it as if it were mapped into a higher space, in simple terms, kernels map data from a low-dimensional space into a higher-dimensional one, enabling more complex separation boundaries with the aim of extend our results to nonlinear regressors, we assess the following kernel generated functions instead of linear functions (Peng,2010).

$f_1(x)=K(x^T, A^T)w_1+b_1, f_2(x)=K(x^T, A^T)w_2+b_2$ … (4)

A detailed review of widely used kernel functions in SVR is showed, with emphasis on identifying suitable kernels for effective nonlinear modeling (Yu et al.,2024).

- Linear Kernel Function: $k(xi.\ xj)=xi^T xj$
- Polynomial Kernel Function: $k(xi.\ xj)=(1+xi^T.\ xj)^d$
- Radial Basis Function (RBF): $k(xi.\ xj)=\exp(-\gamma\|x-xi\|^2)$

The kernel choice determines the model's ability to learn complex nonlinear pattern.

3- Training the Model

During training, the optimization problem is solved the best weights w and bias b. Only the data points outside the ε-tube (support vectors) significantly influence the resulting regression function (Peng,2010).

4- Making prediction

The process of using the trained regression model to estimate the output value y^ for new, unseen input data x (Zhang et al.,2011).

$$y^{\wedge}=f(x)=w^T x+b \quad \ldots (5)$$

5- Evaluate precision of Models

To test the performance and accuracy of proposed model, some measurements and statistical tests are used. Such as, mean square error, root of mean square error, (Aziz et al., 2023)

A. Mean Squared Error (MSE)

MSE measures the average squared difference between the predicted and actual values. A lower MSE indicates better model performance (Aziz et al., 2023). It is calculated as:

$$MSE=\frac{1}{n}-\sum_{i=1}^{n}(yi-\hat{y}_i)^2 \quad \ldots (6)$$

Where yi is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. MSE will fit better of the model to the data when it has lower value.it indicates that the prediction value is near to actual value, it squares the error because of this it can sensitive to outlier (Aziz et al., 2023).

B. Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and represents the average magnitude of prediction errors in the same units as the target variable (A Jiménez al et 2025).

$$\text{RMSE}= \sqrt{\frac{1}{n} \sum_{i=1}^{n}(yi - \hat{y}_i)^2} \quad \dots (7)$$

C. Coefficient of Determination (R^2)

R^2 evaluates how well the model explains the variance of the dependent variable. Values closer to 1 indicate better predictive performance (A Jiménez al et 2025).

$$R^2= 1- \frac{\sum(yi-\hat{y})^2}{\sum(yi-\bar{y})^2} \quad \dots (8)$$

where $\bar{y}$ is the mean of the observed values.

2nd: Data analysis and Results

1- Data Description

To analyze the predictive ability of the suggested method, the dataset water quality monitoring station in Sulaymaniyah, Iraq, was collected from January to November 2025 from a. The dataset Contains of 575 water samples, each Including chemical measurements of water quality. It containing of (8) features, where (7) input variables (Cl, T.Alk, Ca^{2+}, Mg^{2+}, T.H, EC, pH) used for prediction and final is a target variable T.D.S (Total Dissolved Solids), This study utilized Three kinds of Twin Support Vector Regression (TSVR) , Linear TSVR, Polynomial TSVR and Radial Basis Function (RBF) TSVR. The models were applied in R programming language using standard libraries of machine learning. A train-test split (80%-20%) was Implemented To evaluate how comprehensive the model will be. For both Polynomial and RBF kernels, hyperparameter tuning was performed for best value of C and σ. All features were scaled, and outliers were Limited using the 1.5 × IQR rule to reduce their impact on model training, as shown in Table 1.

Table (1): Dataset Description

Attribute	Description	Data Type	Possible Values / Range
Cl	Chloride concentration (Chloride ion content in water)	Numeric	0–500 mg/L
T.Alk	Total alkalinity (Water buffering capacity	Numeric	10–300 mg/L
Ca^{2+}	Calcium ion concentration	Numeric	5–200 mg/L
Mg^{2+}	Magnesium ion concentration	Numeric	2–100 mg/L
T.H	Total hardness (sum of Ca^{2+} and Mg^{2+}) Measure of water hardness	Numeric	50–400 mg/L
EC	Electrical conductivity (Ability of water to conduct electricity)	Numeric	100–2000 µS/cm
pH	Acidity or alkalinity (Measure of acidity or alkalinity)	Numeric	6–9
T.D.S	Total Dissolved Solids(Total dissolved inorganic substances)	Numeric	50–1500 mg/L

2- Linear TSVR Model (Baseline)

A. Hyperparameter Sensitivity Analysis

Hyperparameter tuning was conducted for all TSVR variants in order to conduct a fair and optimal comparison of models. A grid search method was used to tune hyperparameters. The regularization parameter C was chosen from a set of candidates {0.01,0.1,1,10,100} for each modelling variant.

The regularization parameter C controls the trade-off between modelling complexity and prediction error; low values of C increase the strength of the regularization, and increase the likelihood of model underfitting, while high values reduce the amount of regularization which allows more model flexibility and increased likelihood of overfitting. Model performance was compared using metrics MSE, RMSE and R^2 on the hold-out test set, with the best hyperparameter values selected based on minimizing the RMSE and maximizing the R^2.

Table (2): Accuracy of Linear TSVR for different values of the regularization parameter C.

C	MSE	RMSE	R^2
0.01	0.079647	0.282219	0.918249
0.1	0.041594	0.203947	0.957307
1	0.04034	0.200848	0.958595
10	0.040325	0.200811	0.95861
100	0.040325	0.20081	0.95861

The Linear TSVR model accuracy for various values of the regularization parameter C. The Impacts represent a clear Enhancement in prediction as C increases from 0.01 to 1, after which the performance stabilizes. This Suggests that values of C ≥1 provide the necessary model flexibility, while smaller values may result in less consistency.3.2.2 Training and Testing Performance. Utilized the optimal hyperparameter (C=100) because it provides the minimum RMSE with stable results, the TSVR linear model was tested on both training and testing datasets. Table 3 illustrate the final effect. The nearest agreement between training and testing metrics Implies perfect generalization, without any signs of exaggeration. In fact, the RMSE test below shows that the linear model effectively captures the main patterns in the data.

Table (3): Linear TSVR performance on training and testing datasets

Dataset	MSE	RMSE	R^2
Train	0.040325	0.20081	0.95861
Test	0.007509	0.08665	0.99268

3- Polynomial TSVR

A. Hyperparameter Sensitivity Analysis

The polynomial TSVR models' behavior is influenced by both the regularization parameter C and the polynomial degree (Zhang et al.,2011). In this study, the polynomial degree was fixed at 2 to avoid extreme model complexity, while a grid search was performed over C= {0.01,0.1,1,10,100}. Table 4. demonstrate the test performance metrics (MSE, RMSE, and R^2) for various values of C. The smaller values of C provide better generalization. Specifically, C=0.01 Obtains the lowest RMSE and highest R^2, whereas higher values of C improve training performance but degrade test performance, suggesting overfitting.

Table (4): Test performance of Polynomial TSVR (degree = 2) for different values of C

C	MSE	RMSE	R2
0.01	0.094414	0.307269	0.907967
0.1	0.106077	0.325695	0.896599
1	0.724292	0.851054	0.293977
10	0.149678	0.386882	0.854098
100	0.119342	0.34546	0.883668

The lowest RMSE and highest R^2 are reached at C=0.01, indicating superior generalization. Higher C values lead to overfitting.

B. Training and Testing Performance

The Evaluation of Polynomial TSVR was done with the ideal hyperparameter values at C = 0.01, Degree = 2 using the resulting outputs from both the training and testing data sets, as shown below in Table 5. Training data performance was excellent (R2 = 0.9815, RMSE = 0.1342) while that of test data was not as good (R2 = 0.8837, RMSE = 0.3455), indicating that the model has limited ability to generalize to the test dataset due to its increased flexibility from the use of polynomial representations of features.

Table (5): Performance of Polynomial TSVR Results for both Training and Testing

Dataset	MSE	RMSE	R^2
Train	0.018	0.1342	0.9815
Test	0.1193	0.3455	0.8837

C. Kernel Matrix Analysis

A polynomial kernel matrix was created from the standardized training input data. The polynomial kernel demonstrates a larger magnitude and greater variability than the linear kernel because of its ability to amplify high-order feature interactions when compared with the linear kernel. Hence, the polynomial kernel may provide improved accuracy. Unfortunately, the flexibility of the polynomial kernel also creates additional susceptibility to noise.

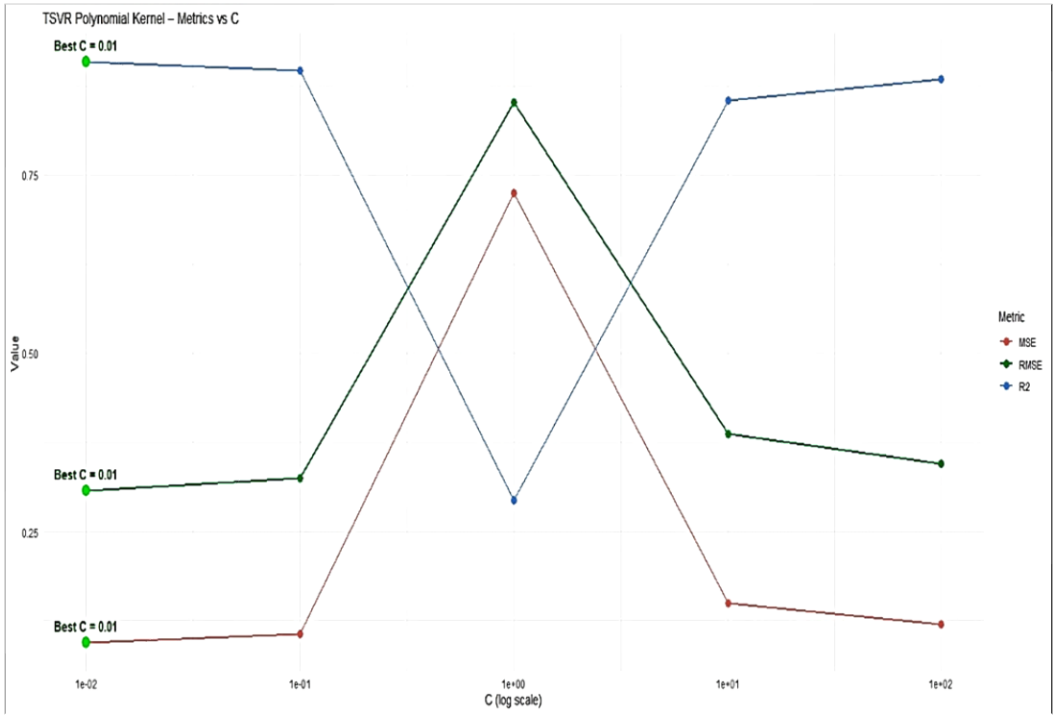

Figure (2): Polynomial Kernel Matrices for TSVR

A graphical representation illustrates how MSE, RMSE, and R-squared for Polynomial TSVR model change with respect to the regularization parameter C (in log scale). The optimal values were obtained with C=0.01, resulting in lowest MSE and RMSE; highest R-squared value indicating good prediction accuracy using this model. Increasing C resulted in increased MSE & RMSE; decreasing R-squared value this shows poorer predictive abilities due to overfitting and/or loss of generalization ability of the model. This trend shows large polynomial kernels will produce optimal results with minimal regularization (C) values. Large C-value results in the polynomial kernel overfitting the training dataset with less predictive ability on new, unseen data

4- Radial Basis Function (RBF) TSVR

A. RBF TSVR Hyperparameter Analysis

The creation of a new parameter for the RBF kernel, σ, modifies its width and determines how 'local' a model is. This value was first determined without considering the effects of different values of the regularization parameter C (that was fixed as 0.1). Where small values of σ (0.1–0.5) had poor predictive performance (high error, and negative or a very low R^2), this occurred because the models were using overly localized kernels, resulting in underfitting. As σ increased, the performance of the model improved, and the best results were achieved for the σ = 2 case within this limited scenario. The effect of varying σ on the performance of RBF TSVR with C = 0.1 can be seen in Table 6.

Table (6): Performance effect of σ on RBF TSVR (C = 0.1)

sigma	MSE	RMSE	R^2
0.1	1.079868	1.039167	-0.05263
0.5	0.932201	0.965505	0.091313
1	0.696687	0.834678	0.320886
2	0.143411	0.378697	0.860206

As result in Table 6, adding σ significantly increases the predictive performance, when C is controlled. Small result of σ values in poor accuracy and underfitting, while larger result of σ values improve model generalization. The best performance in this work is produced at σ = 2, indicating the importance of selecting an appropriate kernel width

B. Joint Hyperparameter Optimization (C, σ)

The purpose of conducting a grid search on both C and σ, to find the optimum settings for both parameters, was achieved through compiling test results from each parameter combination into Table 7 showing MSE, RMSE and R^2 performance metrics for every combination tested, and generating a corresponding R^2 heatmap as shown in Figure 3.

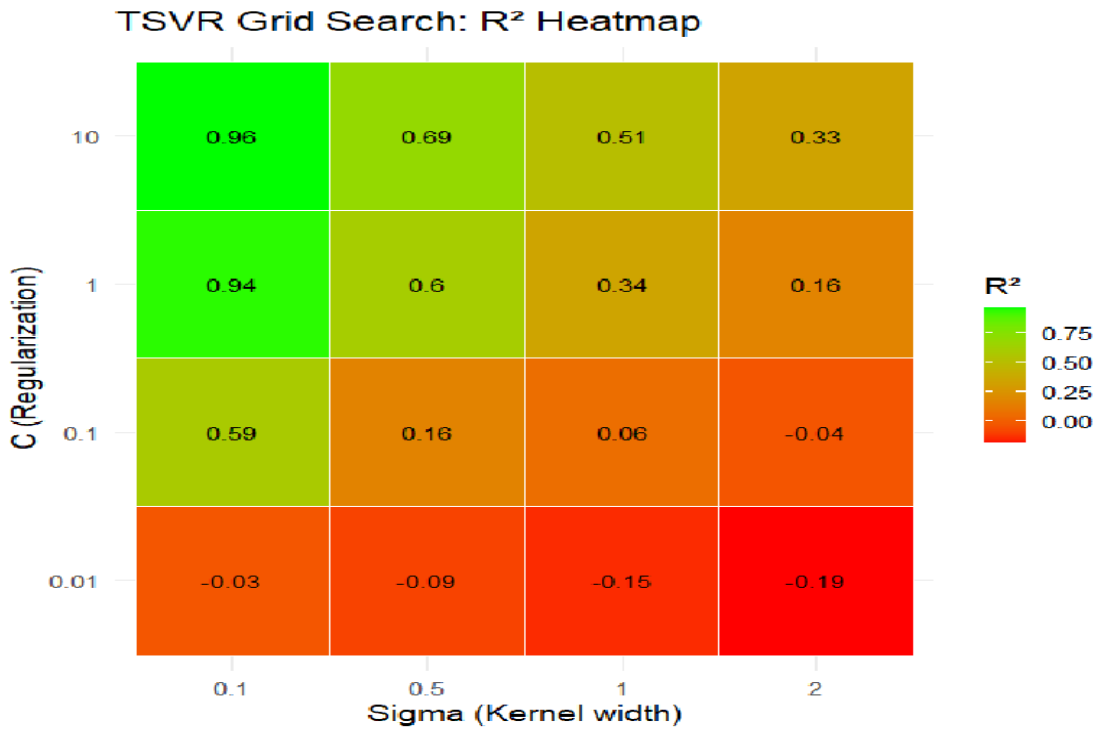

Figure (3): Performance R^2 Heatmap of RBF TSVR for Joint Optimization of C and σ.

Figure 3 demonstrate the combined effect of the RBF kernel width σ and the regularization parameter C on TSVR performance in terms of R^2. (C = 0.01) is a very small result in negative R^2 across all σ values, indicating underfitting because of excessive regularization. Increases C, predictive performance enhances significantly, specifically for smaller σ values. At C = 10 and σ = 0.1, is the best result reached, suggesting that weaker regularization joined with a localized kernel enables the model to better record nonlinear patterns in the data. by contrast, at fixed C rising σ and decrease R^2, indicating lower model flexibility. Overall, the results emphasize the importance of jointly tuning C and σ to reach optimal TSVR result.

Table (7): Performance RBF TSVR for different combinations of C and σ.

C	sigma	MSE	RMSE	R^2
0.01	0.1	1.061509	1.030296	-0.03473
0.01	0.5	1.122231	1.059354	-0.09392
0.01	1	1.178038	1.085375	-0.14832
0.01	2	1.22003	1.10455	-0.18926
0.1	0.1	0.424677	0.651673	0.586035
0.1	0.5	0.861469	0.928154	0.160261
0.1	1	0.965712	0.982706	0.058647
0.1	2	1.066312	1.032624	-0.03942
1	0.1	0.062859	0.250718	0.938726
1	0.5	0.407838	0.638622	0.60245
1	1	0.67202	0.819768	0.344931
1	2	0.857965	0.926264	0.163676
10	0.1	0.042567	0.206318	0.958506
10	0.5	0.322163	0.567594	0.685964
10	1	0.507017	0.712051	0.505772
10	2	0.682669	0.826238	0.334551

The influence of regularization variable C on model results is very evident as the value of C increases and the reduction in both MSE and RMSE takes place while also seeing an increase in R^2, therefore indicating that there was an increase in how accurately the model was able to predict what would be observed. The data suggests that this is even more pronounced for smaller values of σ in which the model was able to recognize the structure within the provided data set much better. Conversely, when C = 0.01 (very low), results in large errors and negative R^2, which indicates that the model is underfitting the data due to too much regularization. Moderate levels of C give a more balanced degree of complexity and generalization.

The width of the Kernel parameter σ also has an important role in the performance of the model whereby smaller values of σ result in lower levels of error metrics and higher R^2 indicating a better fit (non-linear relationship example). Conversely, when σ is too large (i.e. higher than 0.8), the performance of the model is degraded due to excessive smoothing and the associated loss of model flexibility.

The maximum performance results for the model occurred using a value of C = 10 and σ = 0.1, which gave rise to the lowest RMSE 0.2063 and the highest R^2 0.9585. Overall, these results emphasize the need to consider both regularization parameter C and Kernel Width parameter σ

C. Training and Testing Performance

Utilizing the optimal hyperparameters (C = 10, σ = 0.1), the RBF TSVR model was analyzing on training and testing datasets. Table 8 illustrate the results. Training and testing performance are well balanced, Pointing to good generalization.

Table (8): Performance of Training and testing RBF TSVR

Dataset	MSE	RMSE	R^2
Train	0.072787	0.26979	0.925291
Test	0.062859	0.250718	0.938726

The RBF kernel is the same for accuracy in both datasets. Therefore, the RBF kernel proved to be a good representation and method to validate non-linear trends and has also been demonstrated in its use of established hyperparameter settings (C=10 and σ=0.1), hence how well the RBF kernel can properly model non-linear correlations and accurately predict in various datasets. Evidence of excellent generalization and indicating overfitting is found in the small discrepancy between the training and the testing errors. Additionally, the R^2, MSE and RMSE values produced as a result of using this model indicates very low levels of error and have produced an accurate representation of correlations existing in both data set.

D. Kernel Matrix Characteristics

Using the RBF Kernel, non-linear correlation can be seen between predictors. As shown in Tables 9 and 10, a small 6x6 section of the kernel matrix (i.e., KERNEL MATRICES) is presented for each dataset, Train and Test respectively. The fact that the RBF Kernel has diagonal dominance indicates it has been correctly constructed, and because of the off-diagonal elements, we can deduce that there are Non-linear Relationships between predictors in each instance (i.e., observations). An additional advantage of using the RBF Kernel is its larger dynamic range than either the linear or polynomial kernels. Therefore, it is able to learn both local and global patterns

Table (9): RBF training kernel matrix (first 6×6 block)

	V1	V2	V3	V4	V5	V6
1	1.00E+00	1.78E-20	9.57E-21	4.10E-22	3.15E-18	2.35E-20
2	1.78E-20	1.00E+00	2.57E-18	4.42E-12	8.08E-20	6.46E-11
3	9.57E-21	2.57E-18	1.00E+00	2.46E-02	4.15E-01	2.46E-02
4	4.10E-22	4.42E-12	2.46E-02	1.00E+00	1.66E-03	8.55E-01
5	3.15E-18	8.08E-20	4.15E-01	1.66E-03	1.00E+00	2.02E-03
6	2.35E-20	6.46E-11	2.46E-02	8.55E-01	2.02E-03	1.00E+00

The RBF training kernel matrix indicates that the diagonal entries = 1 which means that each observation & itself are identical to one another. The values of the off-diagonal entries differ from each other; thus, some will be near zero & others will be intermediate (0.415, 0.855). This implies that some of the observations are exhibiting non-linear relationships with respect to each other. The range of values indicates that the RBF kernel can model both local similarities (between closely located observations) and global patterns across all observations in the dataset. As such, the use of this kernel has allowed the model to learn complex non-linear relationships resulting in good prediction accuracy.

Table (10): RBF test kernel matrix (first 6×6 block)

	V1	V2	V3	V4	V5	V6
1	2.84E-19	8.79E-19	0.584224	0.00648	0.848706	0.007246
2	3.59E-09	2.93E-08	3.26E-06	2.3E-05	1.56E-05	0.000116
3	1.53E-22	0.136937	4.5E-11	5.93E-09	4.20E-16	5.60E-08
4	5.65E-08	6.09E-08	1.13E-06	1.38E-05	3.18E-06	7.93E-05
5	4.69E-20	1.10E-19	0.74637	0.006554	0.760494	0.006162
6	4.13E-19	5.73E-18	0.280434	0.013392	0.318659	0.010912

The RBF Test contains pairs of test samples. The diagonal entries are close to 1 because each sample has absolute similarity to itself; however, the off-diagonal entries represent very little similarity for some sample pairs and greater amounts of similarity for other sample pairs (e.g. 0.280, 0.584, and 0.848). Therefore, since the RBF kernel captures the nonlinear relationships between pairs of samples, the longer the distance between two samples along the coordinate system of the RBF kernel, the less similar they are to one another. The RBF test kernel holds on to high and low similarities for those samples, thus confirming once again that the RBF kernel is able to represent nonlinear patterns in the test data through strong and weak similarities across different

datasets. The strong and weak similarities are consistent with the overall prediction accuracy of the RBF kernel when tested against other datasets, making it an excellent predictor of test data based on previous predictor knowledge for other datasets.

E. Comparison of TSVR Variants

Test results in Table 11 show a comparison of three methods: linear, polynomial, and radial basis function (RBF) support vector regression (SVR). Among these three methods, SVR with a linear kernel produced the greatest accuracy when predicting test data using test data set, suggesting that most of the relationships responsible for total dissolved solids (TDS) behaviors are fundamentally linear. While RBF SVR has a strong ability to model nonlinearity as well as provide reasonable generalization from training to testing datasets, polynomial SVR suffers from overfitting of training data and has lower accuracy when predicting test datasets.

Table (11): Comparison of Linear, Polynomial, and RBF TSVR models

Model	Dataset	C	Degree / σ	MSE	RMSE	R^2
Linear TSVR	Train	100	Linear	0.040325	0.20081	0.95861
Linear TSVR	Test	100	Linear	0.007509	0.086654	0.99268
Polynomial TSVR	Train	0.01	Degree=2	0.018014	0.134217	0.98151
Polynomial TSVR	Test	0.01	Degree=2	0.119342	0.34546	0.883668
RBF TSVR	Train	10	σ=0.1	0.072787	0.26979	0.925291
RBF TSVR	Test	10	σ=0.1	0.062859	0.250718	0.938726

On the test set, Linear TSVR achieved the following scores: lowest mean squared error (0.0075), lowest root mean squared error (0.0867), and highest coefficient of determination (0.9927). This suggests that the majority of the relationships involved in total dissolved solids (TDS) are linear. RBF TSVR performed reasonably well, although with slightly larger errors (MSE: 0.0629, RMSE: 0.2507) and a good R^2 (0.9387). This indicates that RBF TSVR was capable of modelling potential nonlinear patterns but also generalizing well between the training and test sets. In contrast, Polynomial TSVR has shown that it has good performance on the training set (0.9815), but was unable to perform well on the test set (0.8837); thus, indicating that the polynomial model has a tendency to over-fit the data at these parameter settings. Overall, this comparison shows that linear models can adequately explain TDS for this data set, while RBF SVR provides an alternative methodology for developing nonlinear models when they exist. However, Polynomial SVR is less reliable because at these parameter settings it exhibits greater tendencies toward over-fitting and lower predictive ability when tested on out-of-sample data.

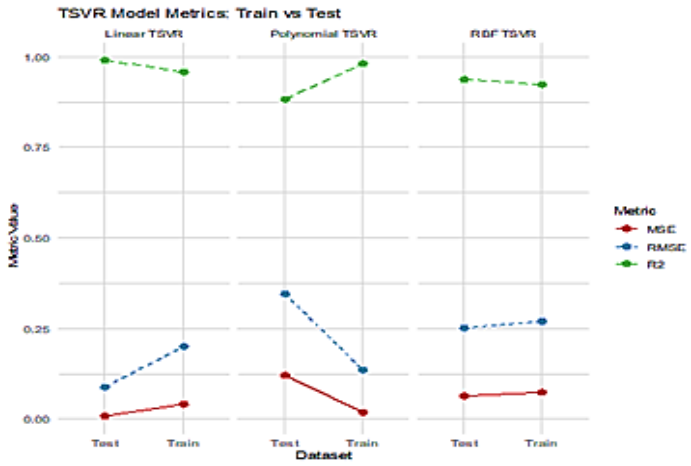

Figure (4): Comparison of TSVR model performance (Linear, Polynomial, and RBF) on training and testing datasets.

The figure shows MSE, RMSE, and R^2 values for each TSVR variant. Solid green lines represent R^2, dotted blue lines demonstrate RMSE, and solid red lines represent MSE. Linear TSVR has a strong generalization ability based on testing on both training and test sets as evidenced by the high R2 values and low errors. Polynomial TSVR has less gap between the training and test sets than the RBF model; this points to a possibility of the polynomial model being overly fit to the training data, while the RBF model appears to be well balanced with moderate errors on both the training data and the test data and has a strong nonlinear capability.

F. Final Model Selection Discussion

Evaluating the predictive performance of TSVR models uses three typical regression metrics (MSE, RMSE, & R^2). RMSE was used as the primary metric because it shows the error of prediction in original TDS scale, while R^2 explains the ratio of the difference between the actual TDS values and predicted TDS values. Through the comparison of linear, polynomial, and RBF TSVR models, various kernel functions showed different behaviors. The linear TSVR consistently demonstrated the lowest RMSE and highest R^2 in the test dataset. This results from linear prediction being the best way to model how input variables relate to TDS. In addition, linear kernels are easier to compute than other types and make applicable in more practical situations on larger datasets. Though the polynomial based TSVR had a strong fit on the training data, it did not perform as well on the testing data with increasing regularization parameter values. This indicates a sensitivity to hyperparameter choices as well as a tendency toward overfitting from higher complexity models. The RBF based TSVR had strong nonlinear modelling capabilities when properly tuned. The hyperparameter optimize results show that using hyperparameters $C = 10$ and $\sigma = 0.1$ results in the optimal configuration yielding competitive accuracy and consistent generalization on both the training and testing datasets. For the above reasons, the linear based TSVR was selected as the final model; it performed best in terms of accuracy, robustness, as well as being the most computationally efficient. The RBF based TSVR would still be a good option to model localized nonlinear relationships, while the polynomial kernel is not recommended for this dataset due to overfitting issues.

Conclusion:

This study comprehensively assessed the predictive capability of Twin Support Vector Regression (TSVR) models for Total Dissolved Solids (TDS) in drinking water using a dataset of 575 samples characterized by eight key hydrochemical features, including chloride, alkalinity, calcium, magnesium, total hardness, electrical conductivity, and pH. Three TSVR variants (Linear, Polynomial (degree = 2), and Radial Basis Function (RBF)) were implemented, with systematic hyperparameter optimization performed via grid search, and model performance evaluated using standard regression metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The analysis demonstrates that Linear TSVR provides the most consistent and reliable predictive performance, achieving high R^2 and low RMSE on both training and testing datasets, indicating strong generalization and computational efficiency for practical applications. Polynomial TSVR, while demonstrating excellent fit to the training data, exhibited overfitting and reduced predictive reliability on the testing set, highlighting its sensitivity to hyperparameter selection and the increased risk associated with higher model complexity. The RBF TSVR successfully captured nonlinear relationships among hydrochemical variables, achieving robust performance across datasets, although at the cost of higher computational demand. Collectively, these findings establish Linear TSVR as a statistically robust and computationally efficient model for TDS prediction, with RBF TSVR.

References:

1. Altamimi, M., 2022. Big data in e-government: Classification and prediction using machine learning algorithms. *Iraqi Journal of Intelligent Computing and Informatics (IJICI)*, 1(2), pp.41–55.
2. Aziz, A.A., Mahmood, H.O.F., Rahim, S.A., Maaroof, R.S. and Taher, H.A., 2023. Using optimizing parameters support vector regression model to predict potassium ratio in carp fish. *Journal of Survey in Fisheries Sciences*, 10(3S), pp.4931–4937.
3. Huang, H., Wei, X. and Zhou, Y., 2018. Twin support vector machines: A survey. *Neurocomputing*, 300, pp.34–43. https://doi.org/10.1016/j.neucom.2018.01.093
4. Jayadeva, Khemchandani, R. and Chandra, S., 2018. *Twin support vector machines: Models, extensions and applications*. Springer Publishing Company.
5. Jiménez-Macías, A., Muñoz-Merino, P.J., Moreno-Marcos, P.M. and Kloos, C.D., 2025. Evaluation of traditional machine learning algorithms for featuring educational exercises. *Applied Intelligence*, *55*(6), pp.1-25.
6. Khemchandani, R., Goyal, K. and Chandra, S., 2015. Twin support vector machine based regression. In: *Proceedings of the Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, IEEE, pp.1–6. https://doi.org/10.1109/ICAPR.2015.7050651
7. Mhamad, A.J., Taher, H.A. and Taha, A.A., 2025. A Hybrid BAT-SVR Methodology for Forecasting Iraq's Interest Rates of Commercial Bank (IRCB) Time Series Data: Hybrid BAT-SVR Methodology. *Academic Journal of International University of Erbil*, *2*(03), pp.347-355.
8. Patle, A. and Chouhan, D.S., 2013. SVM kernel functions for classification. In: *Proceedings of the International Conference on Advances in Technology and Engineering (ICATE)*, IEEE, pp.1–9. https://doi.org/10.1109/ICAdTE.2013.6524743
9. Peng, X., 2010. TSVR: An efficient twin support vector machine for regression. *Neural Networks*, 23(3), pp.365–372.
10. Singh, K.P., Basant, N. and Gupta, S., 2011. Support vector machines in water quality management. *Analytica Chimica Acta*, 703(2), pp.152–162.
11. Taher, H.A., Mhamad, A.J. and Taha, A.A., 2025. Optimization of Support Vector Regression for Improved Prediction Accuracy in Cross-Sectional Data Using Genetic Algorithms. *Raparin Journal of Humanities (RJH)*, *12*(4), pp.842-864.
12. Tomar, D. and Agarwal, S., 2015. Twin support vector machine: A review from 2007 to 2014. *Egyptian Informatics Journal*, 16(1), pp.55–69. https://doi.org/10.1016/j.eij.2014.12.003
13. Vapnik, V.N., 1997. The support vector method. In: *International Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, pp.261–271.
14. Yu, S., Miao, J., & Qin, F. (2024). *Twin support vector regression for characterizing uncertainty in surface reconstruction*. **Scientific Reports, 14**, Article 19612. https://doi.org/10.1038/s41598-024-70109-y
15. Zhang, L. and Lin, C., 2011. Twin support vector machines for regression. *Neurocomputing*, 74(13–15), pp.2030–2037.
16. Zhang, S., Liu, C., Wang, W. and Chang, B., 2020. Twin least square support vector regression model based on Gauss–Laplace mixed noise feature with its application in wind speed prediction. *Entropy*, 22(10), p.1102. https://doi.org/10.3390/e22101102