

Enhancing surface water quality prediction efficiency in northeastern thailand using machine learning

Surasit Uypatchawong, Nipaporn Chanamarn

Department of Computer, Faculty of Science and Technology, Sakon Nakhon Rajabhat University (SNRU), Sakon Nakhon, Thailand

Article Info

Article history:

Received Mar 17, 2024

Revised Jul 23, 2024

Accepted Jul 29, 2024

Keywords:

Bagging model
Machine learning
Prediction models
Suitable factors
Surface water quality

ABSTRACT

Water is the most vital resource for life and is necessary for most living creatures, including humans, to survive. Three rivers' surface water quality has been predicted by this study: the Chi river, the Mun river, and the Songkhram river. In the northeastern region of Thailand. The dataset is 881 samples and 13 factors. This study investigated various machine learning methods for predicting water quality, including neural networks (NN), support vector machines (SVM), decision trees (DT), Naive Bayes (NB), and K-nearest neighbors (KNN). Furthermore, this study was conducted to find suitable factors using correlation based feature selection, correlation coefficient, and information gain. And optimize the prediction model using the Bagging Approach. The result is found that the bagging model using the DT technique (BaggingDT) has better performance than all models with an accuracy value equal to 98.64%, precision value equal to 98.70%, recall value equal to 98.60%, F-measure value equal to 98.60% and RMSE value equal to 0.0961. The obtained factors and the most appropriate model can be used to develop a surface water quality standard predicting system.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nipaporn Chanamarn

Department of Computer, Faculty of Science and Technology, Sakon Nakhon Rajabhat University (SNRU)

That Choeng Chum, Muang, Sakon Nakhon, Thailand

Email: nipaporn@snru.ac.th

1. INTRODUCTION

Water is a necessary resource for all living things as well as the natural ecology [1]. Water resources undoubtedly play a vital role in our daily lives. Thus, one of the key concerns of environmental water management is the assessment of surface water quality [2]. The quality criteria of ambient water bodies, including rivers, lakes, and streams, are typically established to represent their level of quality. Additionally, water specifications for various purposes and usages are subject to their own set of rules [3]. The water quality of rivers is increasingly at risk from serious risks and degradation as a result of growing human activity related to industrialization and urbanization [4]. Consequently, for its varied applications and efficient management, the fundamentals regarding the quantity and quality of water should be known. Most ambient water bodies-rivers, lakes, and streams-must meet quality criteria. Determining the concentration of contaminants in the water is made easier by the quality of the water. Water quality is a field that has investigated a variety of examples. The northeastern region of Thailand has several significant rivers that are utilized in a variety of fields. Additionally, there are standards for water parameters for different uses and purposes.

This area of artificial intelligence also referred to as expert systems [5], [6], seeks to create methods that let computers learn from and generalize from pre-existing datasets. When there is a lot of data and it is

difficult for a human to analyze and form conclusions, these strategies are applied. It entails creating algorithms that let computers learn from data iteratively, adjusting their behavior to new information and patterns. Neural networks (NN): NN are a class of algorithms with pattern recognition capabilities, inspired by the structure and functioning of the human brain. They use a form of machine perception to understand sensory data by classifying or grouping unprocessed input [7], [8]. Support vector machines (SVM): SVM are supervised machine learning algorithms that are applied to regression and classification problems. It identifies a hyperplane in an N-dimensional space (where N represents the number of features) that distinctly classifies the data points [9], [10]. Decision trees (DT): DT are supervised regression and classification methods. Building a model that predicts a target variable using data-driven decision rules is the goal [11], [12]. Naive Bayes (NB): The “naive” assumption of independence between each pair of features in the Bayes theorem is the foundation of the Naive Bayes probabilistic classifier [13]. It is commonly used for text classification problems like spam filtering and sentiment analysis. K-nearest neighbors (KNN): KNN is a supervised learning algorithm used for regression and classification tasks. It works by finding the ‘K’ nearest data points in the training set to the new data point and makes predictions based on the majority class among its nearest neighbors (for classification) or the average value of its nearest neighbors (for regression) [14], [15]. These algorithms can be applied to various machine learning tasks depending on the characteristics of the data and the specific problem being addressed. They each have their strengths and weaknesses, and the choice of algorithm often depends on factors such as the size and complexity of the dataset, the interpretability of the model, and computational resources available.

An approach that is widely used to improve the success of data mining initiatives is called cross-industry standard process for data mining (CRISP-DM) [16]. The technique outlines a flexible series of six stages that enable the construction and application of a DM model in an actual setting, helping to support business decisions see in Figure 1.

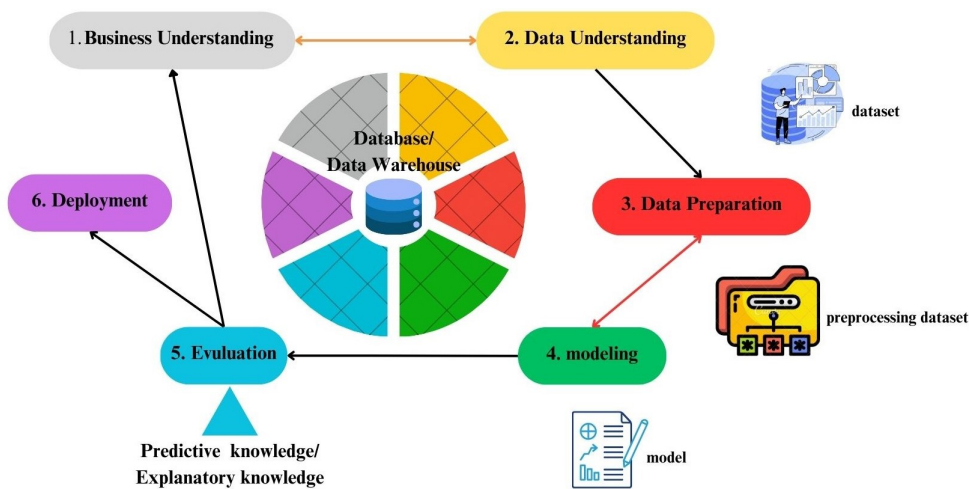


Figure 1. The CRISP-DM process model [16]

Surface water quality predictions have been made with success using machine learning. Nonetheless, there is a general dearth of research on machine learning for forecasting the water quality index (WQI). As a result, it is crucial to provide novel methods for assessing and, if feasible, forecasting the water quality. The WQI is the predominant approach for categorizing and conveying the current state of water quality [17]. Suwadi *et al.* [18] used ANN, SVM, NB, and RF for predicting WQI. The feature selection algorithms used successfully classified the most suitable attributes. The chosen features have the potential to address the classification challenge of predicting water quality that is important to this investigation and the dataset at hand. Shah *et al.* [19] proposed approach for analyzing surface water quality and modeling using gene expression, machine learning, and regression approaches. They presents the application of data-driven models, i.e., gene expression programming (GEP), artificial neural networks (ANN), multiple linear regression (MLR), and multiple nonlinear regression (MNLR), for estimating the total dissolve solids (TDS) and electrical conductivity

(EC) in the upper Indus river basin. All the models demonstrated a strong connection between the observed and simulated data. The GEP model was determined to be superior and surpassed all other techniques in performance. Chen *et al.* [20] employed 10 learning models (comprising 7 traditional models and 3 ensemble models) using extensive data from major rivers and lakes in China to investigate the potential key water parameters for enhancing future model predictions. The research that has been reviewed presents the findings of important factors and predicts using machine learning only. After that, there was no improvement in forecast efficiency. This research therefore has increased the efficiency of the experiment by finding important factors. Then came up with algorithm to experiment with these factors for better performance, which is state of the art.

This study was designed to achieve the following objectives:

1. To develop a predictive model using well-known machine learning techniques to predict surface water quality class and identify relevant factors.
2. To improve the efficiency of predictive models through various evaluation metrics and determine the best predictive model for this study.

This paper is organized as follows: section 2 method. Then, section 3 results and discussion. Finally, section 4 presents conclusions.

2. METHOD

The experiment consisted of six steps. According to the steps of CRISP-DM include: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, the proposed model shown as Figure 2.

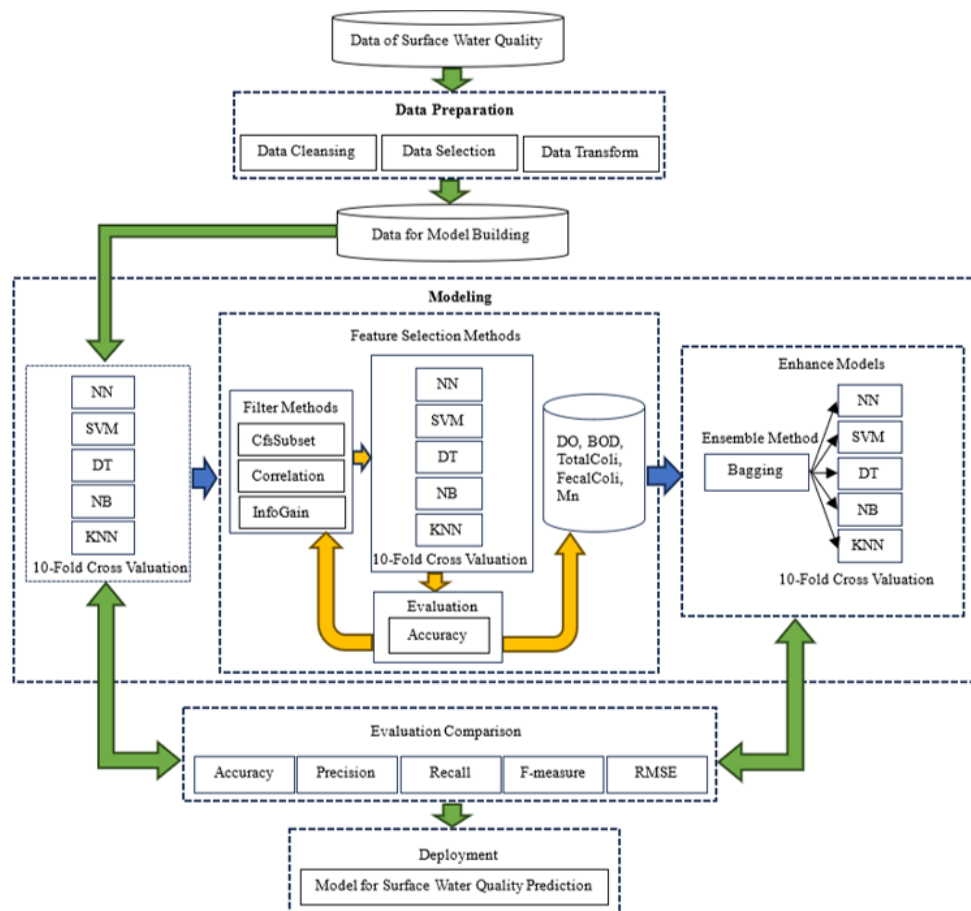


Figure 2. Working diagram of the proposed model

2.1. Business understanding

The surface water quality data set was obtained from the Pollution Control Department, Ministry of Natural Resources and Environment, Thailand. It is surface water quality data for three rivers: the Chi river, the Mun river, and the Songkhram river. In the northeastern region of Thailand. Data collected from 2011 - 2017. There are 5 classes of surface water quality standards. In this research, we received only 3 classes of information that the Pollution Control Department has collected: types 2, 3, and 4.

2.2. Data understanding

Selecting the water quality dataset is a prerequisite for model construction and involves several key factors: gathering essential parameters that impact water quality, identifying the number of data samples, and defining class labels for each data sample in the dataset. The datasets used in this work consist of 13 indicator parameters. These parameters are temperature (Temp), pH value (pH), dissolved oxygen (DO), biological oxygen demand (BOD), Total Coliform Bacteria (TCB), Fecal Coliform Bacteria (FCB), Nitrates (NO₃), ammonia (NH₃), Cadmium (Cd), Manganese (Mn), Nickel (Ni), Zinc (Zn) and Copper (Cu). For each data sample in the dataset, the WQI is initially calculated and a class label is applied, which ranges from “excellent” to “heavily polluted” (5 - 1), shown as Figure 3.

	Temp(w)	pH	DO(mg/l)	BOD(mg/l)	Total Coli(MPN/100ml)	Fecal Coli(MPN/100ml)	NO3-N(mg/l)	NH3-N(mg/l)	Cd(ug/L)	Mn(mg/L)	Ni(ug/L)	Zn(mg/L)	Cu(ug/L)	Level
0	26.53	7.31	5.2	1.4	230.0	230.0	0.133	0.1	0.49	0.037	0.1	0.059	0.49	2.0
1	27.62	6.66	5.0	1.4	2400.0	430.0	0.188	0.1	0.49	0.069	0.1	0.004	0.49	2.0
2	27.62	6.37	5.6	1.3	150.0	150.0	0.112	0.2	0.49	0.140	0.1	0.120	0.49	2.0
3	28.8	6.33	4.2	0.9	930.0	430.0	0.091	0.1	0.49	0.049	0.1	0.007	0.49	2.0
4	28.67	6.18	4.7	1.1	430.0	430.0	0.118	0.1	0.49	0.140	0.1	0.120	0.49	2.0

Figure 3. The dataset of surface water quality data for experiment

2.3. Data Preparation

A total of 881 records datasets were subjected to data cleaning, data transform, and data selection, as shown in Table 1.

2.4. Modeling

This experiment uses NN, SVM, DT, NB, K-NN with specified parameters. NN technique uses momentum=0.2, learning rate=0.3 and hidden nodes are 3, 4, 5, 6, 7, 8, 9, 10, and 11. SVM technique uses Kernel = PolyKernel and PukKernel. And K-NN uses K value = 3 in comparative efficacy trials. To divide data in training and testing, use the 10-fold cross-valuation method. Once the best model has been obtained, it is then used for initial testing to find important features of water quality prediction. Then use seature selection methods like filter methods, namely correlation based feature selection, correlation coefficient, and information gain methods. Feature selection was carried out using the Waikato environment for knowledge analysis (Weka) [21]. Correlation-based feature selection is used in this study with the CfsSubsetEval, an attribute evaluator in WEKA that assesses the value of a subset of attributes by taking into account the individual predictive capacity of each feature as well as the degree of redundancy between them. Then experiment with all 5 factors obtained from feature selection methods and experiment with 5 techniques, then compare the Accuracy values to see which model has the highest performance and by how many factors. Then optimize the prediction model. By using the bagging approach ensemble method technique by testing all 5 models. Bagging, also known as bootstrap aggregating, is an ensemble technique that entails the independent training of numerous models on random subsets of the data, followed by the aggregation of their predictions through voting or averaging [22], [23]. This experiment shown as Figure 2.

Table 1. Summary of results from data preparation

No.	Variables	Description	Data
1	Temp	Temperature	Min=22.23 Max=38.00 Mean=29.32 S.D.=2.49 Min=0.20
2	pH	pH Value	Max=31.50 Mean=7.24 S.D.=0.98 Min=0.00
3	DO	Dissolved oxygen (mg/l)	Max=14.05 Mean=5.74 S.D.=1.61 Min=0.00
4	BOD	Biological oxygen demand (mg/l)	Max=7.90 Mean=1.63 S.D.=1.02 Min=2.00
5	TCB	Total Coliform Bacteria (MPN/100ml)	Max=160,900 Mean=4,432.62 S.D.=15,987.47 Min=0.00
6	FCB	Fecal Coliform Bacteria (MPN/100ml)	Max=160,900 Mean=1,600.17 S.D.=11,502.25 Min=0.00
7	NO3	Nitrate (mg/l)	Max=11.43 Mean=0.79 S.D.=0.82 Min=0.00
8	NH3	ammonia (mg/l)	Max=7.60 Mean=0.37 S.D.=0.40 Min=0.01
9	Cd	Cadmium (ug/L)	Max=6.00 Mean=0.49 S.D.=0.32 Min=0.00
10	Mn	Manganese (mg/L)	Max=4.75 Mean=0.14 S.D.=0.16 Min=0.01
11	Ni	Nickel (ug/L)	Max=20.00 Mean=1.34 S.D.=3.14 Min=0.00
12	Zn	Zinc (mg/L)	Max=1.67 Mean=0.12 S.D.=0.14 Min=0.00
13	Cu	Copper (ug/L)	Max=6.00 Mean=0.48 S.D.=0.36
14	Level (class)	Level of surface water quality	Level 2=392 records Level 3=224 records Level 4=265 records

2.5. Evaluation

The performance of the proposed approach has been assessed and evaluated using various indicators and criteria. The evaluation metrics implemented are accuracy, precision, recall, F-measure, and RMSE [9], [24]. Accuracy: accuracy is computed as (1):

$$Accuracy = \frac{TP + TN}{P + N} \quad (1)$$

P represents the total number of positive cases, N represents the total number of negative cases, TP represents the total number of correctly identified positive cases, and TN represents the total number of correctly identified negative cases. Precision and recall are essential performance parameters in machine learning for the identification and classification of patterns [25].

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall is determined by utilizing the accurate column of the real data to calculate the model and F-measure, which may be represented as the harmonic mean of recall and precision. The process of calculating recall and F-measure is demonstrated in (3) and (4).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Root mean squared error (RMSE) the metric calculates the square root of the mean squared error in the forecasts. The performance improves as the value decreases. A number of 0 indicates perfect prediction [15]. The RMSE is computed as (5).

$$RSME = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (5)$$

2.6. Deployment

The obtained factors and the most appropriate model can be used to develop a surface water quality standard forecasting system. This system can be used by students, researchers, or the general public to monitor and predict water quality effectively. The deployment of such a system involves integrating the model into an accessible platform, such as a web-based application or mobile app, which allows users to input relevant data and receive real-time predictions. Additionally, the system can be enhanced with user-friendly features such as data visualization tools and alerts for water quality thresholds, ensuring comprehensive and practical usability.

3. RESULTS AND DISCUSSION

In this section, the researcher has presented the experimental results into 3 parts: prediction results, which is the experimental results using the traditional model, the second part is the feature selection method, which is the part to find suitable factors, and the third part is optimize the prediction model using the bagging approach.

3.1. Prediction results

The present study employs NN, SVM, DT, NB, and K-NN models with specified parameters. The 10-fold cross-validation method is employed to partition data into training and testing sets. The experimental results are shown in Table 2.

Table 2. Performance of the used machine learning models to predict surface water quality

Models	Accuracy	Precision	Recall	F-measure	RMSE
NN	88.76	88.60	88.80	88.50	0.2443
SVM	82.75	82.40	82.70	82.30	0.3431
DT	92.20	92.20	92.20	92.20	0.102
NB	70.26	75.60	70.30	70.30	0.4023
K-NN	75.37	75.20	75.40	74.70	0.3376

Table 2 demonstrates that the DT model has better performance than all models with an accuracy value equal to 92.20%, precision value equal to 92.20%, recall value equal to 92.20%, F-measure value equal to 92.20%, and RMSE value equal to 0.1023, followed by NN, SVM, K-NN, and NB respectively. It is noted that the performance of the DT model is very superior as compared to the other models. The DT classifier exhibited exceptional performance, with an accuracy rate of 92.20%. One possible reason is that the DT algorithm successfully identified mutually exhaustive and exclusive rules for each class, enabling the construction of an effective decision tree. However, the NB has shown the poorest performance.

3.2. Feature selection method

Then, an experiment was conducted to find suitable factors using correlation based feature selection, correlation coefficient, and information gain. Using the DT technique in the beginning, it was found that the correlation based feature selection method had 3 features: DO, BOD, and TCB. The correlation coefficient method has 8 features: DO, BOD, TCB, FCB, Mn, Temp, NO₃, and pH. And the information gain method has 5 features: DO, BOD, TCB, FCB, and Mn. Then, using the factors that were tested with the NN, SVM, DT, NB, and K-NN techniques, the efficiency is shown in Table 3.

Table 3. Performance of the used machine learning models to predict surface water quality using factors obtained from feature selection methods

Models	Accuracy		
	3 Features	5 Features	8 Features
NN	92.05	92.96	91.71
SVMs	87.51	87.51	85.47
DT	95.80	96.63	94.07
NB	76.73	70.60	72.53
K-NN	92.96	93.07	82.41

From Table 3, it was found that every model had increased efficiency from the original using all 13 factors, and it was found that the DT model gave better performance in terms of accuracy than the other models. and the highest accuracy value is equal to 96.63% at 5 factors: DO, BOD, TCB, FCB, and Mn using the information gain method. For the purpose of predicting surface water quality in the future, information on these 5 factors should be thoroughly gathered and using the information gain method.

3.3. Optimize the prediction model using the bagging approach

From the accuracy in section 3.2, the researcher conducted an experiment to increase the efficiency of the model using the bagging ensemble using 5 factors. The empirical findings are displayed in Table 4. The bagging approach, which combines multiple models to improve stability and accuracy, demonstrated significant enhancements in prediction performance.

Table 4. Performance of the used bagging ensemble models

Ensemble models	Accuracy	Precision	Recall	F-measure	RMSE
BaggingNN	93.64	93.60	93.60	93.60	0.1904
BaggingSVMs	88.20	88.40	88.20	88.10	0.3172
BaggingDT	98.64	98.70	98.60	98.60	0.0961
BaggingNB	71.28	75.50	71.30	70.60	0.3835
BaggingK-NN	92.28	92.30	92.30	92.20	0.2025

Table 4 demonstrates that the bagging model using the DT technique (BaggingDT) has better performance than all models with an accuracy value equal to 98.64%, precision value equal to 98.70%, recall value equal to 98.60%, F-measure value equal to 98.60% and RMSE value equal to 0.0961. The models are BaggingNN, BaggingK-NN, BaggingSVM, and BaggingNB, respectively.

This section presents the experimental results used to predict the surface water quality. DT model is a model that is suitable for predicting surface water quality, it can be observed from Table 2. To increase the efficiency of prediction from this work, we experimented with finding important features using the information

gain method with 5 features that had the highest efficiency, shown in Table 3. In addition, there has been an experiment using the bagging approach method to enhance efficiency. This results in increased efficiency from all models. The empirical findings are displayed in Table 4. From all the experiments in this research, it is known that the DT model is still a suitable model for predicting surface water quality. The prediction efficiency can be improved by finding important features and using the bagging approach, which compares the efficiency of all experiments, as shown in Figures 4 and 5. The results of this experiment are in line with the research objectives.

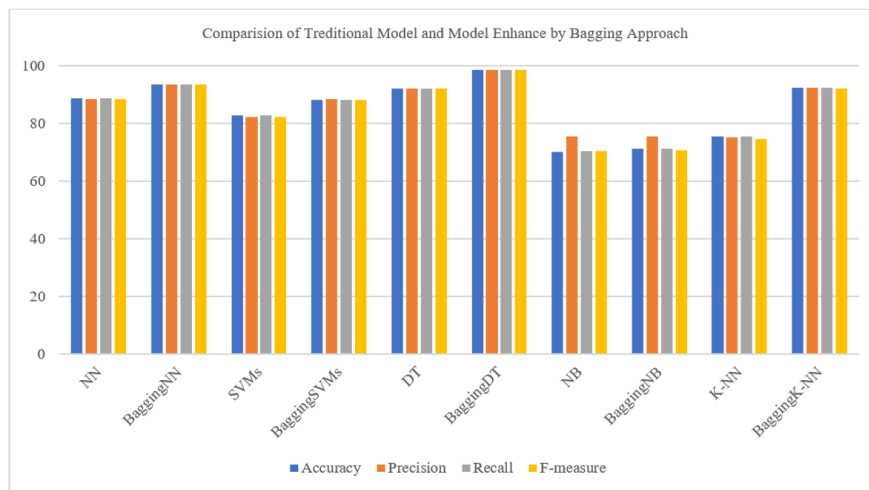


Figure 4. Comparison of traditional model and model enhance by bagging approach

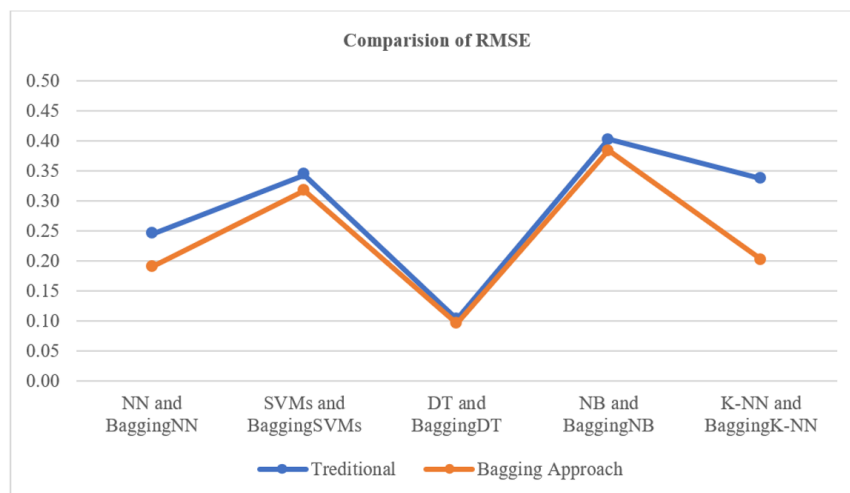


Figure 5. Comparison of RMSE

4. CONCLUSION





An investigation was conducted to assess the suitability of several machine learning models for predicting surface water quality. This is important in applying the model to the benefits of related organizations. In this proposed methodology enhances surface water quality prediction efficiency in northeastern Thailand using machine learning, namely, NN, SVM, DT, NB, and KNN models. Moreover, feature selection methods such as correlation based feature selection, correlation coefficient, and information gain used to find suitable factors. And optimize the prediction model using the bagging approach. The findings indicate that the DT

model outperforms all other models, value in accuracy of 92.20%. However, the information gain method is the most appropriate way to find suitable factors. The result is found that the DT model has better performance than all models. The accuracy value increased to 96.63%. And Bagging model using the DT technique (baggingDT) has better performance than all models. The accuracy value increased to 98.64%. In future research, the created models will be utilized to predict the water quality in Thailand for various water types, based on an analysis of the effectiveness of the suggested model for predicting surface water quality.





REFERENCES

- [1] N. L. Kushwaha *et al.*, "Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection," *Ecological Informatics*, vol. 75, p. 102122, 2023, doi: 10.1016/j.ecoinf.2023.102122.
- [2] N. H. Than, C. D. Ly, and P. V. Tat, "The performance of classification and forecasting Dong Nai River water quality for sustainable water resources management using neural network techniques," *Journal of Hydrology*, vol. 596, p. 126099, 2021, doi: 10.1016/j.jhydrol.2021.126099.
- [3] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water quality prediction using artificial intelligence algorithms," *Applied Bionics and Biomechanics*, vol. 2020, pp. 1–12, 2020, doi: 10.1155/2020/6659314.
- [4] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang, "Water quality prediction based on machine learning and comprehensive weighting methods," *Entropy*, vol. 25, no. 8, 2023, doi: 10.3390/e25081186.
- [5] T. Freiesleben and T. Grote, "Beyond generalization: a theory of robustness in machine learning," *Synthese*, vol. 202, pp. 1–28, 2023.
- [6] D. M. Gaba, "Artificial intelligence and expert systems," in *Control and Automation in Anaesthesia*, Springer, 2022, pp. 22–36.
- [7] P. Mohapatra *et al.*, "Artificial neural network based prediction and optimization of centelloside content in *Centella asiatica*: a comparison between multilayer perceptron (MLP) and radial basis function (RBF) algorithms for soil and climatic parameter," *South African Journal of Botany*, vol. 160, pp. 571–585, 2023, doi: 10.1016/j.sajb.2023.07.019.
- [8] P. Sharma and G. Kaur, "Review on data mining techniques for prediction of chronic kidney disease," *International Journal of Engineering Trends and Technology*, vol. 63, no. 1, pp. 58–60, 2018, doi: 10.14445/22315381/ijett-v63p209.
- [9] M. Aljanabi, R. Hayder, S. Talib, A. H. Ali, M. A. Mohammed, and T. Sutikno, "Distributed denial of service attack defense system-based auto machine learning algorithm," *Bulletin of Electrical Engineering and Informatics (BEEE)*, vol. 12, no. 1, pp. 544–551, 2023, doi: 10.11591/eei.v12i1.4537.
- [10] W. Sun, J. Yu, Y. Kang, S. Kadry, and Y. Nam, "Virtual reality-based visual interaction: a framework for classification of ethnic clothing totem patterns," *IEEE Access*, vol. 9, pp. 81512–81526, 2021, doi: 10.1109/ACCESS.2021.3086333.
- [11] M. R. Mahmood, M. B. Abdulrazzaq, S. R. M. Zeebaree, A. K. Ibrahim, R. R. Zebari, and H. I. Dino, "Classification techniques' performance evaluation for facial expression recognition," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 21, no. 2, pp. 1176–1184, 2020, doi: 10.11591/ijeecs.v21.i2.pp1176-1184.
- [12] M. M. El Sherbiny, E. Abdelhalim, H. El-Din Mostafa, and M. M. El-Seddik, "Classification of chronic kidney disease based on machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 32, no. 2, pp. 945–955, 2023, doi: 10.11591/ijeecs.v32.i2.pp945-955.
- [13] A. Rawal and B. Lal, "Predictive model for admission uncertainty in high education using Naïve Bayes classifier," *Journal of Indian Business Research*, vol. 15, no. 2, pp. 262–277, 2023, doi: 10.1108/JIBR-08-2022-0209.
- [14] X. H. Tay *et al.*, "An entropy-based directed random walk for cancer classification using gene expression data based on bi-random walk on two separated networks," *Genes*, vol. 14, no. 3, 2023, doi: 10.3390/genes14030574.
- [15] F. Khan, C. Ncube, L. K. Ramasamy, S. Kadry, and Y. Nam, "A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning," *IEEE Access*, vol. 8, pp. 119710–119719, 2020, doi: 10.1109/ACCESS.2020.3003785.
- [16] P. Chapman *et al.*, "Step-by-step data mining guide," *SPSS inc*, vol. 78, no. 13, pp. 1–78, 2000, [Online]. Available: <http://www.crisp-dm.org/CRISPWP-0800.pdf>.
- [17] R. Babbar and S. Babbar, "Predicting river water quality index using data mining techniques," *Environmental Earth Sciences*, vol. 76, no. 14, pp. 1–15, 2017, doi: 10.1007/s12665-017-6845-9.
- [18] N. A. Suwadi *et al.*, "An optimized approach for predicting water quality features based on machine learning," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, 2022, doi: 10.1155/2022/3397972.
- [19] M. I. Shah, M. F. Javed, and T. Abunama, "Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques," *Environmental Science and Pollution Research*, vol. 28, no. 11, pp. 13202–13220, 2021, doi: 10.1007/s11356-020-11490-9.
- [20] K. Chen *et al.*, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," *Water Research*, vol. 171, p. 115454, 2020.
- [21] Q. Qiao, A. Yunusa-Kaltungo, and R. E. Edwards, "Feature selection strategy for machine learning methods in building energy consumption prediction," *Energy Reports*, vol. 8, pp. 13621–13654, 2022, doi: 10.1016/j.egy.2022.10.125.
- [22] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, 1996.
- [23] M. O. Arowolo, M. O. Adebisi, A. A. Adebisi, and O. J. Okesola, "Predicting RNA-Seq data using genetic algorithm and ensemble classification algorithms," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 21, no. 2, pp. 1073–1081, 2020, doi: 10.11591/ijeecs.v21.i2.pp1073-1081.
- [24] D. Manikandan and J. Dhillippan, "Machine learning approach for intrusion detection system using dimensionality reduction," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 34, no. 1, pp. 430–440, 2024, doi: 10.11591/ijeecs.v34.i1.pp430-440.
- [25] N. Z. Tawfeeq, O. G. Ghazal, and W. S. Abed, "Using data mining techniques to extract students' attitudes toward e-learning," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, pp. 1037–1048, 2022, doi: 10.11591/ijeecs.v28.i2.pp1037-1048.

BIOGRAPHIES OF AUTHORS

Surasit Uypatchawong     is currently working as a lecturer in the Department of Computer, Faculty of Science and Technology at Sakon Nakhon Rajabhat University (SNRU). He Holds a Master degree in Computer Science at Khon Kaen University. His research areas are image processing, biometrics, machine learning, data mining business intelligence, deep learning, natural language processing, and pattern recognition. He can be contacted at email: surasit@snru.ac.th.



Nipaporn Chanamarn     is currently working as a lecturer in the Department of Computer, Faculty of Science and Technology at Sakon Nakhon Rajabhat University (SNRU). She received a Ph.D. degree in Computer Science at Naresuan University. Her research activities are in the areas of machine learning, data mining, education data mining, business intelligence, neural networks, decision trees, decision support systems, data science, deep learning, web programming, virtual reality, and augmented reality technology. He can be contacted at email: nipaporn@snru.ac.th.