Available Online at www.e-iph.co.uk
Indexed in Clarivate Analytics WoS, and ScienceOPEN

KICSS2023

Kedah International Conference on Social Science and Humanities
UiTM Kedah (Online), Malaysia, 21-22 June 2023:
2nd International Conference on Business, Finance, Management and Economics
(BIZFAME)

Predicting Kereh River's Water Quality: A comparative study of machine learning models

Norashikin Nasaruddin[1], Afida Ahmad[1*], Shahida Farhan Zakaria[1],
Ahmad Zia Ul-Saufie[2], Mohamed Syazwan Osman[3]
*Corresponding Author

[1*] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Kedah Branch, 08400 Merbok, Kedah, Malaysia
[2] School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam 40450, Selangor, Malaysia
[3] EMZI-UiTM Nanoparticles Colloids & Interface Industrial Research Laboratory (NANO-CORE), Chemical Engineering Studies, College of Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang, Permatang Pauh Campus, 13500 Pulau Pinang, Malaysia.

norashikin116@uitm.edu.my, *afidaahmad@uitm.edu.my, shahidafarhan@uitm.edu.my, ahmadzia101@uitm.edu.my, syazwan.osman@uitm.edu.my
Tel: +60175175881

Abstract

This study introduces a machine learning-based approach to forecast the water quality of the Kereh River and categorize it into 'polluted' or 'slightly polluted' classifications. This work employed three machine learning algorithms: decision tree, random forests (RF), and boosted regression tree, leveraging data spanning from 2010 to 2019. Through comparative analysis, the RF model emerged as the most efficient, boasting an accuracy of 97.30%, sensitivity of 100.00%, specificity of 94.74%, and precision of 95.00%. Notably, the RF model identified dissolved oxygen (DO) as the paramount variable influencing water quality predictions.

Keywords: Water quality; machine learning; decision tree; random forest

DOI: https://doi.org/10.21834/e-bpj.v8iSI15.5097

1.0 Introduction

Water is fundamental to the sustenance of all life forms. Despite Malaysia's rich water resources, the issue of water pollution poses a severe threat, potentially leading to water shortages (Gasim et al., 2013). Rapid economic expansion and urban development, particularly in areas like the Klang Valley and Langat Valley, are significant contributors to this deterioration in water quality (Rahman, 2021). Such activities are not only detrimental to ecological health but can also have far-reaching impacts on ecosystems and human well-being. The Malaysian government employs the water quality index to gauge pollution levels, using the national water quality standard (NWQS) to determine the suitability for water use. This index encompasses six parameters: dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH3-NL), suspended solid (SS), and pH. In a 2020 survey of 672 Malaysian rivers, 66% exhibited clean water quality, with 29% slightly polluted and 5% showing signs of pollution (Department of Environment Malaysia, 2020). Given the intricate nature of water quality index calculations, they are time-intensive and complex (Bui et al., 2020). However, the rise in sophisticated data and technological advancements make machine learning an increasingly viable tool for handling vast datasets and accurately modeling water quality predictions.

Centered on Penang's Kereh River as out case study model, this study sheds light on its current status as a highly polluted river, largely attributed to the extensive pig farming in Kampung Selamat, Tasek Gelugor, Penang, Malaysia (Dermawan, 2021). As of 2020, this area housed 77 licensed pig farms, rearing over 186,390 pigs. Pig waste discharge, a primary contamination source for Kereh River, has raised public concern regarding the river's water quality and environmental implications (Lee, 2020; Dermawan, 2021). To address

DOI: https://doi.org/10.21834/e-bpj.v8iSI15.5097

this, our study employs three machine learning models—decision tree (DT), random forest (RF), and boosted regression tree (BRT)—to forecast the water quality of the Kereh River. We aim to evaluate the efficacy of these machine learning techniques, given the diverse predictions and performance metrics in existing literature. This paper delineates the data collection and machine learning methodologies, with subsequent sections presenting results, discussions, and our conclusions."

2.0 Literature Review

There are several studies that used different machine learning models to model and predict water quality. Random forest (RF) is an alternative approach to predicting water quality. The method is an efficient tool for lowering modeling complexity, hence a suitable approach to developing appropriate management and mitigation strategies (Ali Khan et al., 2021). Virro et al. (2022) estimated Estonian water quality with RF and concluded that basic RF models could surpass or at minimum, match the precision of current process-based techniques. In another study, Behrouz et al. (2022) employed RF to predict the most influential parameters affecting stormwater quality from 314 stations throughout the United States. The performance of the model improved when numerous climatological and catchment attributes were utilized as inputs.

Another machine learning model is the decision tree (DT), which is advantageous considering its significant statistical reliability in data validation (Lu and Ma, 2020). Lu and Ma (2020) utilized decision tree-based machine learning combinations to estimate the short-term water quality of one of the most polluted rivers in the world, the Gales Creek in the Tualatin River. The models produced highly accurate predictions of which indicators contributed to the river water quality.

A regression tree (RT) is a machine-learning approach that is particularly valuable for extracting information from the data and producing an understandable predictive model (Huang et al., 2023). In this paper, the authors used a regression tree algorithm to predict the nitrite concentration that determines the water quality of a reactor. Motevalli et al. (2019) introduced an advanced RT technique and the Boosted Regression Tree, to model nitrate concentration in Tajan and Neka rivers. The model was more robust as it could be utilized even with different input types and missing values.

Several research have been employed to compare various machine learning models in projecting and evaluating water quality. For instance, Asadollah (2021) did a comparative study of machine learning models that compared the Extra Tree Regression (ETR) model with Support Vector Regression (SVR) and Decision Tree Regression (DTR) models to forecast the river water quality index of Lam Tsuen River. The experimental results documented that ETR was superior to SVR and DTR as the model produced more accurate and robust practical values. Ali Khan et al. (2021) proposed employing the Artificial Neural Network (ANN) and RF approaches to estimate the Indus River, Pakistan surface water salinity. The RF model documented superior performance compared to ANN in forecasting water quality parameters. In another report, Jeung et al. (2019) conjectured the first flush effects of stormwater runoffs in the Sang-mu District, Gwangju, South Korea, with RF and RT. The study found that the water quality level estimations were better with the RF algorithm than the RT. The Langat River's water quality index and classification were determined in a study that developed three machine learning models; ANN, DT, and Support Vector Machine (Shamsuddin et al., 2022). The comparative performance analysis demonstrated that the SVM technique was the most effective in forecasting Langat River water quality index. In summary, Decision Tree, Random Forests, and Boosted Regression Tree measures showed significant results in analyzing and predicting water quality. Limited studies were conducted on assessing and predicting the water quality of rivers in Malaysia which made this study substantial.

3.0 Methodology

3.1 Water Quality Data

The present study obtained secondary data from the Malaysian Department of Environment (DOE). The dataset, which was collected regularly from 2010 to 2019, contained various Kereh River water quality parameters. The Kereh River stations were situated in Penang and Kedah, from which a total of 132 observations were conducted in the current study. Malek et al. (2022) identified seven water quality measurements as among the most impactful predictors in estimating water quality. The present study employed DO, BOD, COD, SS, pH, NH3-NL, and temperature (TEMP; in °C) as the predictors to estimate Kereh River water quality as detailed in Table 1.

Table 1. The dataset description

No.	Variable	Role	Description	Unit
1.	River status	Target	River status classification	1 = Polluted, 0 = Slightly polluted
2.	DO	Input	Dissolved oxygen	mg/l
3.	BOD	Input	Biochemical oxygen demand	mg/l
4.	COD	Input	Chemical oxygen demand	mg/l
5.	SS	Input	Suspended solids	mg/l
6.	pH	Input	pH	unit
7.	NH3-NL	Input	Ammoniacal nitrogen	mg/l
8.	TEMP	Input	Temperature	°C

(Source: Malaysian Department of Environment (DOE))

The present study utilized the synthetic minority oversampling technique (SMOTE) to avoid imbalanced classification during predictive model developments. The dataset was initially filtered to only consider the minority class, which was the polluted river quality, with 37 data points. Subsequently, the k-nearest neighbors were determined (k = 5). An algorithm was then designated for a random nearest neighbor within the data, thus creating a new data point on the line between the two data points. The results before and after applying the SMOTE are illustrated in Fig. 1.

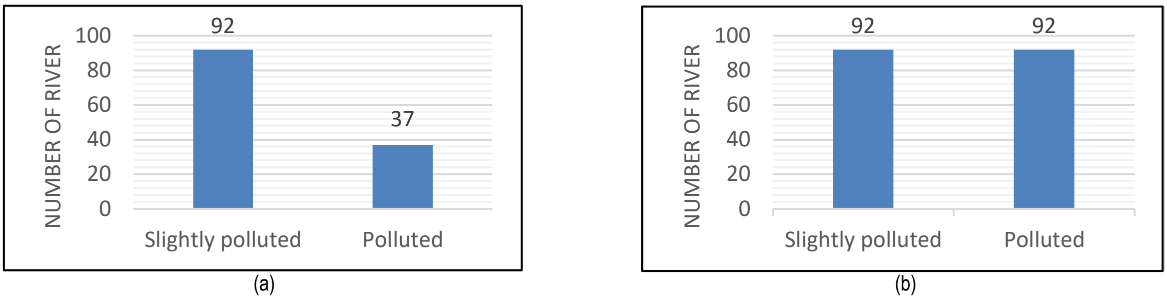


Fig. 1. (a) The original datasets; (b) SMOTE datasets.

Three tree-based machine learning models; DT, RF, and BRT, were utilized in the present study to estimate Kereh River water quality. Data exploration and pre-processing were conducted for comprehension and preparation of the data as inputs were performed before analysis. The step involved data exploration, detection of outliers, and missing values cleaning to avoid disturbances in the algorithm performance, which was applied in the next process. The data were then divided into training (80%) and validation (20%).

The present study applied the SMOTE approach to correct class imbalances during the data understanding stage. Subsequently, the data obtained from the machine learning models were analyzed with the R statistical software. Finally, the performances of the three machine learning models employed were compared to determine the technique that best predicted Kereh River water quality. Fig. 2 demonstrates the modeling process overview conducted in this study.

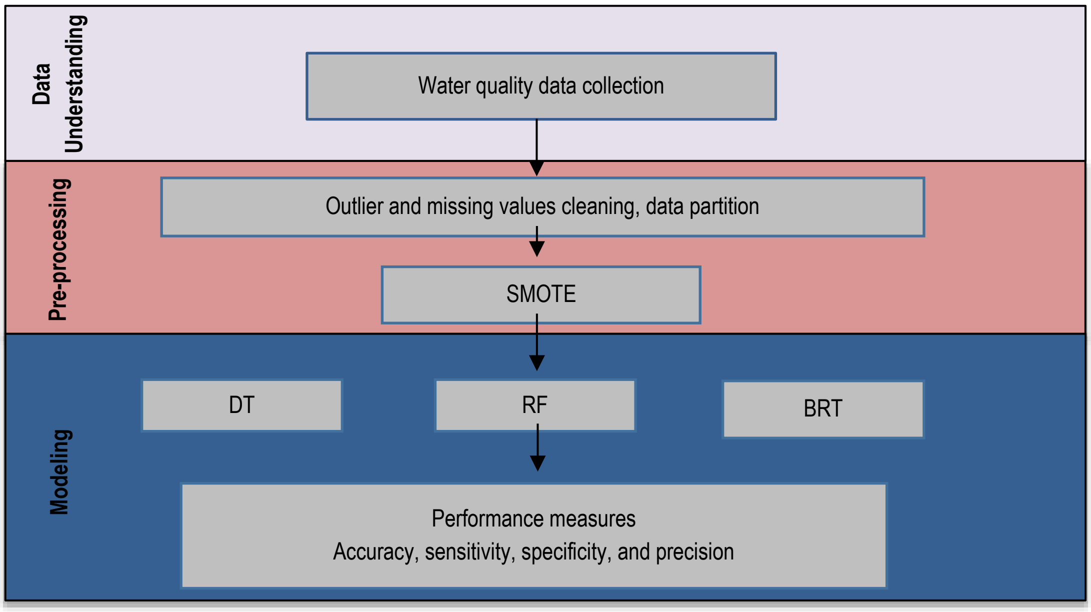

Data Understanding
Water quality data collection
Pre-processing
Outlier and missing values cleaning, data partition
SMOTE
Modeling
DT
RF
BRT
Performance measures
Accuracy, sensitivity, specificity, and precision

Fig. 2. The Kereh River water quality modeling process

3.2 The DT model

The DT method also referred to as supervised learning, was developed in the 1960s. Decision trees are one of the most effective data mining tools. The approach has also been widely utilized in several disciplines due to its ease of application, ambiguity-free properties, and robustness, even with missing data (Hastie et al., 2011). Moreover, DT is widely employed in classifications since it employs a tree structure model in the prediction (Liao and Sun, 2010).

In the DT method, the issues that require solving are represented as a tree, in which each leaf node represents a class label and an internal node denotes an attribute. The four most widely utilized algorithms for constructing decision trees are CART (classification and regression trees), C5.0, CHAID (chi-squared automatic interaction detection), and QUEST (quick, unbiased, efficient, statistical tree).

3.3 The RF method

Random forests are a group of learning techniques employed to solve classification, regression, and other difficulties by constructing numerous decision trees. For example, the output from RF indicates the class selected by most trees, while the mean or average predictions of individual trees represent regression values (Ho, 1998).

RF is quick and simple to utilize, produces high-quality predictions, and can process a large input variable number without overfitting (Hastie et al., 2011). In most cases, RF outperformed DT but was less accurate than gradient-boosted trees (Hastie et al., 2011). Nevertheless, the performance of the method might be impacted by data characteristics.

3.4 The BRT approach

Boosting is an advantageous machine-learning strategy for enhancing model precision. The technique is based on the principle of determining and averaging several rough rules of thumb is simpler than discovering a highly accurate prediction rule (Motevalli et al., 2019). BRT utilizes two algorithms; regression and regression (decision tree) trees and boosting, which construct and combine model sets. Hyperparameters are critical elements in learning algorithms that impact the performance and accuracy of a model. Learning rate and n_estimators are the two fundamental hyperparameters for gradient-boosting decision trees. Learning rates indicate how fast a model learns. The addition of each tree alters the overall model. The magnitude of the alteration is determined by the learning rate. A lower learning rate is attributable to slower learning. A model with a slower learning rate offers robustness and efficiency, and models that learn slowly are better in statistical learning.

3.5 Performance measures

Four performance measures were used to evaluate the three models utilized in this study as Malek et. al. (2022) suggested. The formula for each parameter, which were accuracy, sensitivity, specificity, and precision, is indicated in Equations (1). Based on the confusion matrix, the true-positives and negatives, and false-positives, and negatives are represented by true-positives (TP), true-negatives (TN), false-positives (FP), and false-negatives (FN), respectively.

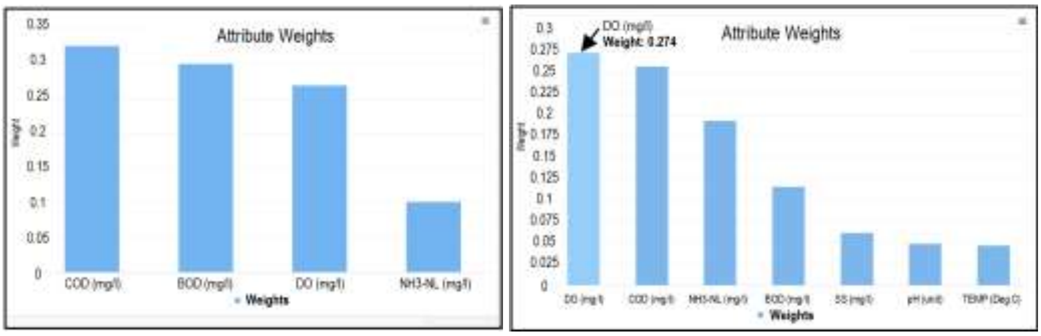
$$\begin{aligned} Accuracy &= \frac{TP + TN}{\left(TP + FP + TN + FN\right)} \\ Sensitivity &= \frac{TP}{\left(TN + FP\right)} \\ Specificity &= \frac{TN}{\left(TN + FP\right)} \\ Precision &= \frac{TP}{\left(TP + FP\right)} \end{aligned} \qquad (1)$$

4.0 Findings

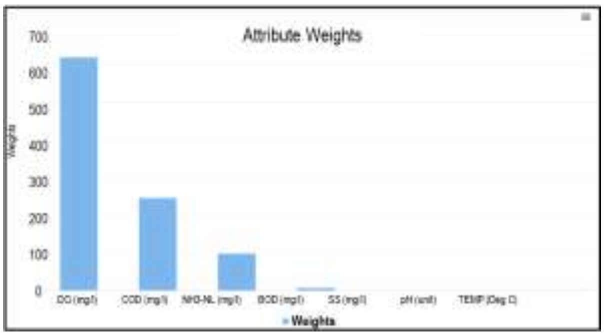
This section includes the results of Kereh River quality predictions determined with three machine learning models. A total of 132 water quality observation data were obtained from the Penang/Kedah Kereh River basin from January 2010 to November 2019. In this study, the river status class of the Kereh River was ascertained. The descriptive statistics of all independent variables are listed in Table 2.

Table 2. The descriptive statistics of the independent variables (IVs) employed

The IV	Minimum	Maximum	Mean	Standard deviation	Skewness	Kurtosis
DO (mg/l)	0.15	8.42	3.08	1.64	0.35	0.14
BOD (mg/l)	2.00	61.00	13.18	9.15	2.23	7.52
COD (mg/l)	5.00	173.00	38.99	24.86	2.49	9.15
SS (mg/l)	6.00	217.00	55.93	37.08	1.80	4.78
pH (unit)	5.63	7.63	6.82	0.38	−0.56	0.39
NH_3-NL (mg/l)	0.06	44.50	8.59	8.23	1.75	3.48
TEMP (°C)	25.51	32.85	29.43	1.57	0.14	−0.27



(a) (b)



(c)

Fig. 3. (a) The variables' importance of the seven parameters employed based on the DT; (b) The variables' importance of the seven parameters employed based on the RF; (c) The variables' importance of the seven parameters employed based on the BRT models.

Fig. 3 demonstrates the importance of the seven indicators utilized in the machine learning models employed in the present study. The variable that primarily impacted the river status quality in the DT model was COD (0.3269), followed by BOD (0.3011), DO (0.2701), and NH3-NL (0.1019). Based on the RF technique, the most influential variable was DO (0.2739), COD (0.2577), NH3-NL (0.1943), BOD (0.1164), SS (0.0620), pH (0.0489), and TEMP (0.0468), while DO (618.87) impacted the results the most in the BRT approach, followed by COD (245.65), NH3-NL (98.04), BOD (7.66), pH (0.05), SS (0.00013), and TEMP (0.00000).

4.1 Performance comparison

The performance of each model employed in the current study was measured in terms of accuracy, sensitivity, specificity, and precision (see Table 3). The RF model yielded the best accuracy (97.37%) and sensitivity (100.00%). Nevertheless, ideal specificity and precision were recorded by DT and BRT at 100.00%, while RF documented a 94.74% specificity and 95.00% precision.

Table 3. The performance measures of each machine learning model

Model	Accuracy	Sensitivity	Specificity	Precision
DT	94.74%	89.47%	100.00%	100.00%
BRT	92.11%	84.21%	100.00%	100.00%
RF	97.37%	100.00%	94.74%	95.00%

4.2 The ROC Comparison

The current study employed ROC charts to compare the performance of the machine learning methods utilized to predict the Kereh River water quality level. Based on the ROC charts (see Fig. 4), RF was the ideal model compared to the DT and BRT as its curve was the closest to the top left corner of the plot.

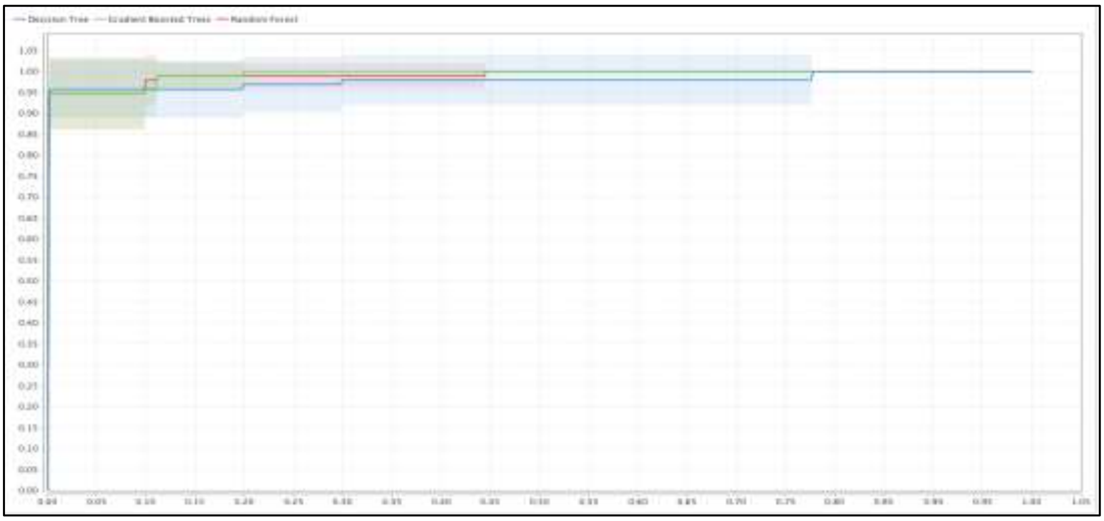


Fig. 4. The ROC charts of the machine learning models employed in the present study

5.0 Discussion

Considering that traditional classifiers commonly favor the majority class with numerous observations, the water quality data obtained in this study was imbalanced, hence requiring attention. The SMOTE method reportedly has positively affected imbalanced data issues (Blagus & Lusa, 2013). Consequently, this study utilized the SMOTE technique to correct the imbalances detected and proven to improve the model performances. Nevertheless, improvements to the SMOTE algorithm have been investigated. For instance, Uyun and Sulistyowati (2020) recorded increased accuracy from 83.3% to 98.8% when they integrated SMOTE and bootstrapping to overcome imbalance class issues. Future studies might consider employing improved SMOTE approaches, to enhance the effectiveness and applicability of SMOTE and continue to advance the field of imbalanced classification (Nemade et al., 2023).

Xu et al., (2020) proved that there is no unique model that fits all sites of water quality, and the reason can be attributed to the different conditions and settings of each location. In this study, the results demonstrated that RF was the best model for predicting Kereh River water. This result is similar to research done by Shaziayani et al. (2022) and Malek et al. (2022) where Decision-tree-based ensemble models, including RF and gradient boosting outperformed single decision-tree methods. The results take into consideration multiple decision trees and employ the output averages to produce better models. The results contradicted the work by Sirikarin and Khonthapagdee (2023) where it was found that extreme gradient boosting with the SMOTE technique achieved the highest score in classifying Thailand's water quality in comparison to RF.

The machine learning models employed in this study could aid in predicting the next river quality status. Furthermore, the model algorithms included the most critical features in water quality estimation. Based on the RF and BRT models, DO was the most influential variable on Kereh River water quality, while COD was the second-most significant factor. The DT model indicated that COD was the most critical variable, followed by BOD. Similarly, studies done by Suwadi et al. (2022), on the Langat Basin in Selangor also showed that DO is an important feature for predicting WQI followed by BOD. Moreover, Sirikarin and Khonthapagdee (2023) also found that BOD is the most significant factor in determining water quality in Thailand. In conclusion, important features to be included to predict water quality are DO, BOD, and COD quality to assist local authorities in performing necessary actions to reduce pollution.

6.0 Conclusion and Recommendation

In this study, three tree-based machine learning methods were utilized to predict the Kereh River water quality, which was classified as polluted or slightly polluted. The main contribution of this study is to assess the quality of the river at an early stage using machine learning methods. If polluted river water quality is detected, immediate preventive measures can be taken. This is important because the main economic activities of rivers in Malaysia are fishing and agriculture since these activities are completely dependent on the quality of river water. The current study proposed employing seven input variables in the DT, BRT, and RF models. The RF approach predicted the Kereh River water quality most efficiently. The technique recorded the highest prediction accuracy based on the accuracy and sensitivity measures. Moreover, the findings were supported by the ROC chart obtained. According to the model, the most crucial variable in predicting the river water quality was DO, followed by COD, NH3-NL, BOD, SS, pH, and TEMP. The second-best method was the DT method, followed by the BRT model.

For this research, the limitation of the study lies in the fact that the results obtained from this study are specific to the Kereh River only, and thus, consideration should be taken when extending the results to other regions where wide variations of water quality parameters are present. As a recommendation, the study scope should be broader with consideration of other water quality parameters, for example, conductivity, salinity, turbidity, Phosphorus, and Escherichia coli. In addition, to minimize the standard error, future research projects should focus on imbalanced learning and rare event detection, since environment data are highly imbalanced.

Acknowledgment

The authors would like to express gratitude to Geran Dalaman Negeri (GDN): 600-TNCPI 5/3/DDN (02) (004/2020), Universiti Teknologi MARA (UiTM), Malaysia for providing financial support throughout this study.

Paper Contribution to Related Field of Study

This paper contributes to the field of water quality management using machine learning methods.

References

Ali Khan, M., Izhar Shah, M., Faisal Javed, M., Ijaz Khan, M., Rasheed, S., El-Shorbagy, M. A., Roshdy El-Zahar, E., & Malik, M. Y. (2022). Application of random forest for modeling of surface water salinity. *Ain Shams Engineering Journal*, 13(4). https://doi.org/10.1016/j.asej.2021.11.004

Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering*, 9(1), https://doi.org/10.1016/j.jece.2020.104599

Behrouz, M. S., Yazdi, M. N., & Sample, D. J. (2022). Using Random Forest, a machine learning approach to predict nitrogen, phosphorus, and sediment event mean concentrations in urban runoff. *Journal of Environmental Management*, 317, 115412. https://doi.org/https://doi.org/10.1016/j.jenvman.2022.115412

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14. https://doi.org/10.1186/1471-2105-14-106

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, 721. https://doi.org/10.1016/j.scitotenv.2020.137612

Department of Environment Malaysia. (2020). Environmental Quality Report 2020. Official Portal of the Department of Environment, 36–105. Retrieved from https://www.doe.gov.my/portalv1/wp-content/uploads/formidable/5/Kualiti-Air-Sungai.pdf

Dermawan, A. (2021, February 4), Main cause of Sg Kreh pollution? Pig farming activities in Kg Selamat, say NGOs, https://www.nst.com.my/news/nation/2021/02/663027/main-cause-sg-kreh-pollution-pig-farming-activities-kg-selamat-say-ngos. (Accessed: 22 October 2022)

Gasim, M. B., Al-Badaii, F., & Shuhaimi-Othman, M. (2013). Water Quality Assessment of the Semenyih River, Selangor, Malaysia. *Journal of Chemistry*, 2013, 871056. https://doi.org/10.1155/2013/871056

Hastie, T., Tibshirani, R., & Friedman, J. (2011). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (9780387848570): Trevor Hastie, Robert Tibshirani, Jerome Friedman: Books. In The elements of statistical learning: data mining, inference, and prediction.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8). https://doi.org/10.1109/34.709601

Huang Y., Su R., Bu Y., & Ma B. (2023). A predictive model for determining the nitrite concentration in the effluent of an anammox reactor using ensemble regression tree algorithm, *Chemosphere*, 339, 139553, https://doi.org/10.1016/j.chemosphere.2023.139553

Jeung, M., Baek, S., Beom, J., Cho, K. H., Her, Y., & Yoon, K. (2019). Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. *Journal of Hydrology*, 575. https://doi.org/10.1016/j.jhydrol.2019.05.079

Lee Goi, C. (2020). The river water quality before and during the Movement Control Order (MCO) in Malaysia. Case Studies in Chemical and Environmental Engineering, 2. https://doi.org/10.1016/j.cscee.2020.100027

Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249, 126169. https://doi.org/https://doi.org/10.1016/j.chemosphere.2020.126169

Malek, N. H. A., Yaacob, W. F. W., Nasir, S. A. M., & Shaadan, N. (2022). Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, *Using Machine Learning Techniques*. Water (Switzerland), 14(7). https://doi.org/10.3390/w14071067, Ministry of Environment and Water. (2020).

Motevalli, A., Naghibi, S. A., Hashemi, H., Berndtsson, R., Pradhan, B., & Gholami, V. (2019). Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. *Journal of Cleaner Production*, 228, 1248-1263.

Nemade, B. ., Bharadi, V. ., Alegavi, S. S., & Marakarkandy, B. (2023). A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons. *International Journal of Intelligent Systems and Applications in Engineering*, 11(9s), 790–803. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/3268

Rahman, H. A. (2021). Water Issues in Malaysia. *International Journal of Academic Research in Business and Social Sciences*, 11(8), 860-875.

Shamsuddin, I.I., Othman, Z., & Sani, N.S. (2022). Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. Water 14(19):2939. https://doi.org/10.3390/w14192939.

Shaziayani, W. N., Ul-Saufie, A. Z., Mutalib, S., Mohamad Noor, N., & Zainordin, N. S. (2022). Classification Prediction of PM10 Concentration Using a Tree-Based Machine Learning Approach. *Atmosphere*, 13(4). https://doi.org/10.3390/atmos13040538

Sirikarin, K., & Khonthapagdee, S. (2023). Machine Learning Techniques for Water Quality Classification of Thailand's Rivers. 20th International Joint Conference on Computer Science and Software Engineering (JCSSE), Phitsanulok, Thailand, 2023, pp. 470-475, doi: 10.1109/JCSSE58229.2023.10202008.

Suwadi, N. A., Derbali, M., Sani, N. S., Lam, M. C., Arshad, H., Khan, I., & Kim, K. (2022). An Optimized Approach for Predicting Water Quality Features Based on Machine Learning. *Wireless Communications and Mobile Computing*, 2022, 3397972. https://doi.org/10.1155/2022/3397972

Uyun, S., & Sulistyowati, E. (2020). Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes. *International Journal of Electrical and Computer Engineering*, 10(4). https://doi.org/10.11591/ijece.v10i4.pp4331-4339

Virro, H., Kmoch, A., Vainu, M., & Uuemaa, E. (2022). Random forest-based modeling of stream nutrients at national level in a data-scarce region. *Science of The Total Environment*, 840, 156613. https://doi.org/https://doi.org/10.1016/j.scitotenv.2022.156613

Xu, T., Coco, G., & Neale, M. (2020). A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Research*, 115788. doi:10.1016/j.watres.2020.115788