Classification of Water Quality Index Using Machine Learning Algorithm for Well Assessment: a Case Study in Dili, Timor-Leste

Zulmira Ximenes da Costa, Keisuke Ikeda, Takumi Nagawaki, Yuichi Nishida, Tamura Satoshi and Floris Cornelis Boogaard

September 6, 2024

Classification of Water Quality Index Using Machine Learning Algorithm for Well Assessment: A Case Study in Dili, Timor-Leste

1st da. Costa Zulmira Ximenes
Graduate School of Engineering,
Gifu University
1-1 Yanagido, Gifu 501-1193, Japan
Faculty of Engineering, science, and Technology
National University of Timor Lorosa'e
Hera, Timor Leste
ximenes.da.costa.zulmira.s1@s.gifu-u.ac.jp

2nd Keisuke Ikeda
Graduate School of Natural Science and Technology, Gifu University
1-1 Yanagido, Gifu 501-1193, Japan
ikeda@asr.info.gifu-u.ac.jp

3rd Takumi Nagawaki
Graduate School of Natural Science and Technology, Gifu University
1-1 Yanagido, Gifu 501-1193, Japan
nagawaki@asr.info.gifu-u.ac.jp

4th Yuichi Nishida
Graduate School of Natural Science and Technology, Gifu University
1-1 Yanagido, Gifu 501-1193, Japan
yuichi@asr.info.gifu-u.ac.jp

5th Tamura Satoshi
Faculty of Engineering, Gifu University
1-1 Yanagido, Gifu 501-1193, Japan
tamura@info.gifu-u.ac.jp

6th Floris Cornelis Boogaard
Research Centre for Built Environment NoorderRuimte,
Hanze University of Applied Sciences,
9747 AS Groningen, The Netherlands
Deltares
Daltonlaan 600, 3584 BK Utrecht
Postbus 85467, 3508 AL Utrecht,
The Netherlands
floris@noorderruimte.nl

Abstract— **This paper investigates to use of information technology, i.e. machine learning algorithms for water assessment in Timor-Leste. It is essential to assess groundwater quality to ensure the safety and availability of well water. The Water Quality Index (WQI) is the standard tool for assessing water quality, which can be calculated from physicochemical and microbiological parameters. However, in developing countries, it is sometimes difficult due to machine malfunctions and limited human resources. In such case, missing-value imputation and machine learning models are useful for classifying water samples into suitable or unsuitable with significant accuracy. Some imputation methods were tested, and four machine-learning algorithms were explored: logistic regression, support vector machine, random forest, and Gaussian naïve Bayes. We obtained a dataset with 368 observations from 26 groundwater sampling points in Dili, the capital city of Timor-Leste. According to experimental results, it is found that 64% of the water samples are suitable for human consumption. We also found k-NN imputation method and random forest method were the clear winners, achieving 96% accuracy with three-fold cross validation. The analysis revealed that some parameters significantly affected the classification results.**

Keywords—water quality index, missing value imputation, classification, machine learning.

I. Introduction

Water is one of the fundamental natural resources for human life, with around 71% of Earth's surface, while only 4% is freshwater and 0.5% is suitable for human consumption [1] . Water quality has emerged as a critical concern in some developing countries, impacting both public health and the environment while playing a pivotal role in fostering sustainable economic development and growth. The significance of water quality management extends beyond its immediate implications, influencing broader societal well-being and ecological sustainability. Nevertheless, based on a UN report, 2.1 billion people still lack access to drinking water and around 40% of the global population suffers from water scarcity [2]. Due to the increasing number of populations, urban development, and business activity, the community requires water supply demand and suitable water, such as quality and availability and easy to access.

Timor-Leste is a country located in the southeastern Asia. Timor-Leste is currently facing various challenges in terms of water quality and quantity supply. Many rivers mostly have water flow only in the wet season. To maintain the water quality and quantity, the government and all of Timorese need to provide a positive contribution to conserve water resources both surface and groundwater. Despite government-led efforts in water supply system management and water quality control, challenges persist, particularly in communities relying on their boreholes. The assessments of water quality in the water bodies, untreated wastewater, car washing, and community wastewater are directly discharged into the streams and groundwater wells and boreholes without any adequate treatment resulting in deterioration of water quality [3]. According to the previous study: Groundwater faces significant threats due to various natural and human-induced factors, including extensive agricultural activities, marine intrusion, population growth, and industrial development [4].

This paper focuses on the situation in the capital city Dili, where groundwater is the main water source for water supply. Groundwater quality has been evaluated and monitored for more than 10 years. Several research projects showed that shallow groundwater in Dili city is contaminated by dissolved solids and microbiological concentrations [5]. In the other work [6] the aquifer is a complex geological formation containing unconsolidated and moderately sorted silts and cobbles, bounded by a coastline and mountains. As the population rapidly increases, we have to address water quality issues comprehensively to safeguard the citizens' well-being.

Water quality is determined by a combination of physical, chemical, and microbiological parameters of water. The quality of groundwater varies in time and location, and current assessing methods rely heavily on manual sampling and laboratory analysis, which is time-consuming, resource-intensive, human skill and inaccurate manual computations. Moreover, traditional statistical techniques may not fully capture the intricate relationships between diverse water quality parameters, hindering the development of accurate predictive models. Applying an sophisticated analysis for water quality, we have encountered the issue of missing values in environmental data, which has a potential to significantly impact the accuracy and reliability of our research findings.

This work thus utilizes the Water Quality Index (WQI). By calculating the index, we can easily evaluate whether water is suitable or not. Though the WQI can be calculated deterministically as described later, in this work we introduce Machine Learning (ML) for water quality estimation. Our final goal is to leverage ML not only for estimating the quality in one sampling point, but also for modeling the whole water flow in Dili city. This paper is thus a first step to the goal. Several studies have explored the application of ML, specifically the Random Forest (RF) algorithm, for predicting and classifying water samples. These studies integrate various classifiers, including RF, to assess water purity, underscoring its effectiveness in monitoring water quality and protecting public health [5]. A research has shown that, among classifiers, RF accurately predicts the WQI, emphasizing its optimization through feature selection, such as dissolved oxygen and biochemical oxygen demand [6]. RF consistently outperformed the other models in terms of accuracy when predicting groundwater quality in India [8]. Nasir et al. further improved the classification accuracy of WQI with their model [7]. In this paper, we also explore the application of ML techniques to classify water data in Dili. By leveraging the power of ML algorithms, it is possible to analyze large datasets comprising multiple parameters and pollutants simultaneously, leading to more comprehensive and timely assessments of water quality. This study has the following main objectives: (i) to examine statistics of physicochemical and microbiology properties of well water, (ii) to calculate WQI in the urban area and visualize the results for further discussion, and (iii) to judge water samples using ML classifiers; water data are categorized into two classes: suitable or unsuitable for human consumption.

II. Data and Methods

This study aims to compute WQI and build ML models for classification using physicochemical and microbiology features. The methodology in this study including data acquisition, preprocessing and analysis, WQI computation, data splitting, classification model building, and evaluation process is illustrated in Figure 1.

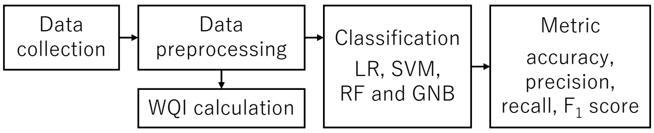

Data collection
Data preprocessing
WQI calculation
Classification LR, SVM, RF and GNB
Metric accuracy, precision, recall, F_1 score

Fig. 1. The framework of our proposed scheme.

A. Dataset

The whole dataset is split into training (80%) and test (20%) datasets. For the data analysis and model evaluation, 3-fold cross-validation is applied to tune hyperparameters in four ML models: Linear Regression (LR), Support Vector Machine (SVM), RF, and Gaussian Naïve Bayes (GNB).

Data collection: Water samples in Dili were obtained by the National Laboratory of Water Supply, Ministry of Public Works in Timor-Leste. The dataset contains 368 samples collected from 26 sampling points of groundwater from 2017 to 2018. From each sample water analysis results were obtained consisting of 16 physicochemical features and two microbiological parameters, as indicated in Table 1. Note that Dili has the wet season basically from November to May, and the dry season lasting from June to October.

Missing values: Due to some reasons, e.g. testing machine malfunction or lack of human resources, there are missing values in the dataset: pH (0.54%), TDS (1.09%), turbidity (0.54%), hardness (0.82%), NO3-N (2.45%), iron (2.99%), fluoride (0.27%), and total coliform (0.82%). Since the amount of missing data has a considerable impact, we need to investigate imputation schemes to fill in missing features. Among several imputation approaches, the preliminary experiment showed that the k-nearest neighbor method achieved the best performance to resolve the imputation for water quality.

Normalization: Standardization is applied so that the transformed data would have a distribution of a mean of 0 and a standard deviation of 1. This step is conducted as an initial stage to prepare water quality data for ML. In the following formula, a standardized value Z_i is obtained as:

$$Z_i = \frac{X_i - \bar{X}}{S} \tag{1}$$

where X_i is an input value, $\bar{X}$ and S are mean and standard deviation of the original distribution, respectively.

Features selection: Feature selection is a prevalent technique aimed at mitigating the issue of irrelevant features. In this study, feature selection is conducted in two parts. The first part involves utilizing person correlation, and the second part is based on RF for selecting features. The RF algorithm consistently outperforms the other ML methods in selecting crucial features for data classification, demonstrating superior performance across all experimental groups [8]. Conversely, the enhanced feature simplification RF algorithm adeptly identifies features closely associated with wind turbine operating conditions, thereby enhancing models for monitoring wind turbine conditions [9]. The selected features are used for the binary classification; each water sample is categorized into suitable and unsuitable.

B. Water Quality Index Calculation

The WQI plays a crucial role in estimating the quality of water [10], which is a flexible, unbiased, and time-saving tool for evaluating drinking water suitability, aiding in prioritizing and maintaining water quality [11]. WQI can be calculated using the weighted arithmetic index method, consisting of the following three steps. First, we choose water quality parameters based on local regulations [12], shown in Table 2. Second, we apply standardization to the selected parameters, converting a measured value of each parameter to a common scale, typically from 0 to 1000. Simulnateously, each parameter value is checked using the standard threshold value. Finally, weights are assigned to the parameters based on their relative importance to overall water quality, under the restriction that the sum of all the weights should be one. Finally, the WQI can be calculated as:

$$WQI = \frac{\sum Q_i W_i}{\sum W_i} \tag{2}$$

where Q_i is a water quality rating of i-th parameter, and W_i is a corresponding weight factor. The rating Q_i is denoted as:

$$Q_i = 100\left[\frac{V_i}{S_i}\right] \tag{3}$$

where V_i is a monitored value in water quality data, and S_i a recommended standard for each parameter. The weight W_i is computed as follows:

$$W_i = \frac{w_i}{\sum_j w_j} \quad \text{, where } \; w_i = \frac{k}{S_i} \; \text{ and } \quad k = \frac{1}{\sum \left(\frac{1}{S_i}\right)} \qquad (4)$$

C. Machine Learning Model

We employ four ML-based classification algorithms to classify each groundwater sample data into the pre-defined categories. The classification algorithms used in this study are as follows:

Logistic Regression: The logistic function, or the sigmoid function, is a mathematical function used in the logistic regression. The LR maps any value to the range (0,1) and is the most common algorithm used for binary classification [13].

Support Vector Machine: The SVM is a supervised learning model used for classification tasks. It finds the optimal hyperplane that separates the classes in the feature space so as to maximize the distance between any two classes [14].

Random Forest: The RF is an ensemble learning method that combines multiple decision trees to improve predictive performance and control overfitting [15]. As mentioned, this model has been widely used in the related works.

Gaussian Naïve Bayes: The Bayes approach employs probabilistic statistics to classify data, and estimate outcomes. The GNB model uses prior and posterior probabilities to prevent from overfitting and bias [16].

D. Evaluation Metric

In this paper, the above four classifiers are evaluated using the test dataset and the following metrics. First, a confusion matrix is obtained. Secondly, accuracy, precision, recall, Area Under Curve (AUC), and F1 score are obtained, which are commonly used in the pertinent literature [17]:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (5)$$

$$Precision = \frac{TP}{TP + FP} \qquad (6)$$

$$Recall = \frac{TP}{TP + FN} \qquad (7)$$

$$F1\ Score = \frac{2 * Recall * Precision}{Recall + Precission} \qquad (8)$$

where TP means True Positive, TN indicates True Negative, FP is False Positive, and FN stands for False Negative.

III. RESULT AND DISCUSSION

A. Statistical Analysis

Given a new dataset, it is cricial to start from statistical analysis and visualization. Table 1 also indicates statistical properties of each parameter. Figure 2 depicts water quality status in each well in Dili. Red wells are unsuitable, while green ones are suitable.

We found several wells had problems in some parameters. For example, in the Cooperativo well, we observed high levels of alkalinity, calcium, hardness, and TDS. Elevated levels of manganese and iron were also found in four wells. NH3, NO2, and NO3 remained under the acceptable ranges at all sites. However, some water samples were heavily contaminated with coliform bacteria, especially E.coli and T.coli. We also found that groundwater turbidity was high due to inadequate well protection during the rainy season, and levels of TDS, hardness, iron, fluoride, calcium, manganese, sulfate, T.coli and E.coli often exceeded desirable limits based on WHO and NDWQS guidelines. The TDS range was from 73.3 mg/L to 2,321 mg/L, likely due to saline water from coastal sources, affected by agricultural activities and sewage disposal systems.

Correlation analysis was subsequently conducted to measure the relationship between two features. Figure 3 illustrates the result. A correlation between hardness and turbidity (0.60) means that increased mineral content leads to higher suspended solids. Sulfate is strongly correlated with both hardness (0.61) and turbidity (0.58), indicating that the presence of minerals affects these parameters. Note that T.coli and E.coli have higher correlations with WQI since water having any microbiology must be unsuitable.

B. Water Quality Index Calculation

The water supply must be determined according to whether water samples are suitable for human consumption, considering certain physical, chemical, and biological characteristics. In this study, in order to calculate WQI, we chose several parameters based on the national drinking water quality standard in Timor-Leste. The standard also refers to WHO guidelines [20], according to the ministry of health in Timor-Leste [19]. Microbiology parameters were fully selected because some research studies have shown that those parameters are crucial to estimate the contamination of the water [3], [5]. Consequently, we focused on 11 water quality parameters to compute WQI. Table 2 indicates the parameters and corresponding coefficients.

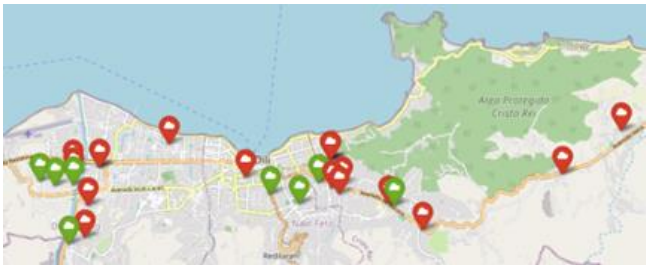
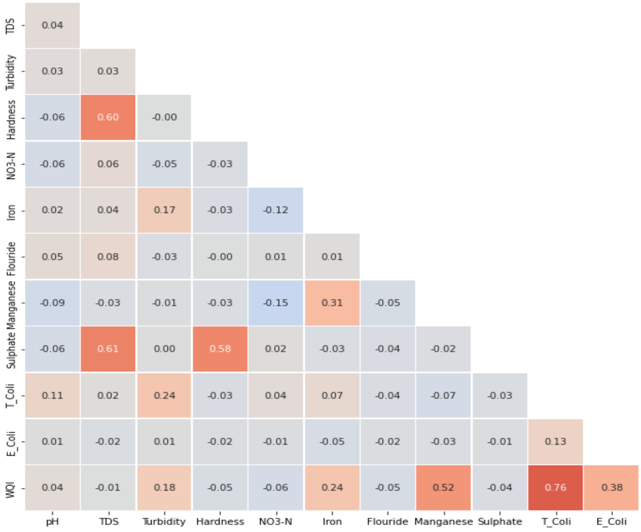


Fig.2. A map of water quality status in each well in Dili.

TABLE I. STATISTICAL SUMMARY OF WATER-QUALITY FEATURES.

Parameter	Mean	Std.dev.	Min	Max	WHO/Timor Leste Guidelines
pH	7.49	0.46	6.4	8.9	6.5-8.5
Temperature	28.29	1.22	22.3	32	NS (oC)
Conductivity	575.87	505.07	163	6901	NS (μS/cm)
TDS	292.4	222.58	73.3	2321	1000 (mg/L)
Salinity	0.29	0.29	0.1	3.8	NS (‰)
Turbidity	1.17	5.84	0.1	105	5 (NTU)
Hardness	207.65	421.81	22	5550	200 (mg/L)
Calcium	160.33	314.47	60	3700	NS (mg/L)
Alkalinity	157.13	277.46	65	4220	NS (mg/L)
NH3-N	0.28	0.23	0.1	1	1.5 (mg/L)
NO3-N	0.33	0.34	0.002	2.1	10 (mg/L)
NO2-N	0.01	0.01	0	0.065	1 (mg/L)
Iron	0.09	0.13	0	1.1	0.3 (mg/L)
Flouride	0.48	1.32	0	18	1.5 (mg/L)
Manganese	1.14	6.06	0	43	0.5 (mg/L)
Sulphate	39.36	64.64	2	1220	250 (mg/L)
T.Coli	20.29	47.62	0	150	$0^a/0.95^b$ CFU/100ml
E.Coli	3.38	18.55	0	150	$0^a/0.95^b$ CFU/100ml

TDS = Total Dissolved Solids, [a] WHO, [b] NDWQS [18], [19]



	pH	TDS	Turbidity	Hardness	NO3-N	Iron	Flouride	Manganese	Sulphate	T_Coli	E_Coli
TDS	0.04										
Turbidity	0.03	0.03									
Hardness	-0.06	0.60	-0.00								
NO3-N	-0.06	0.06	-0.05	-0.03							
Iron	0.02	0.04	0.17	-0.03	-0.12						
Flouride	0.05	0.08	-0.03	-0.00	0.01	0.01					
Manganese	-0.09	-0.03	-0.01	-0.03	-0.15	0.31	-0.05				
Sulphate	-0.06	0.61	0.00	0.58	0.02	-0.03	-0.04	-0.02			
T_Coli	0.11	0.02	0.24	-0.03	0.04	0.07	-0.04	-0.07	-0.03		
E_Coli	0.01	-0.02	0.01	-0.02	-0.01	-0.05	-0.02	-0.03	-0.01	0.13	
WQI	0.04	-0.01	0.18	-0.05	-0.06	0.24	-0.05	0.52	-0.04	0.76	0.38

Fig.3. A correlation analysis result.

After WQI is calculated, it is classified based on [21]. Table 3 shows five groups used in the classification. We can categorize water samples of which WQI are less than 50 are suitable for human consumption, while the others above 50 are unsuitable for drinking. We found the numbers of suitable and unsuitable samples were 237 and 131, respectively.

Let us go back to Figure 2. The GIS map shows that the contamination of drinking water quality in Dili has become a major challenge for both the community and the government in developing countries. The mapping of WQI status shows that 26 boreholes were classified to water suitability classes, however, the other 14 boreholes were polluted and unsuitable for drinking. This information is useful for decision-making by the government to obtain suitable water much more.

TABLE II. FEATURES AND PARAMETERS FOR WQI.

Parameters	Si	1/Si	Wi= k/Si
pH	8.5	0.118	0.0131
TDS	1000	0.001	0.0001
Turbidity	5	0.200	0.0223
Hardness	200	0.005	0.0006
NO3-N	50	0.020	0.0022
Iron	0.3	3.333	0.3723
Fluoride	1.5	0.667	0.0745
Manganese	0.4	2.500	0.2792
Sulphate	250	0.004	0.0004
T.Coli	0/0.95*	1.053	0.1176
E.Coli	0/0.95*	1.053	0.1176
11 features	Σ	8.954	1.0000
	k	0.112	

*NDWQS [19]: Ministry of Health-WHO

TABLE III. WATER QUALITY CLASSIFICATION BASED ON WQI [22].

WQI	Water Quality Range	# samples	
0-25	Excellent water quality	178	237
26-50	Good water quality	59	
51-75	Poor water quality	18	131
76-100	Very poor water quality	7	
100+	Unsuitable for drinking purposes	106	

C. Machine Learning Model Comparison

a) Accuracy comparison with different imputation strategies

First of all, we applied the ML models to analyze the dataset comprising 336 water quality records, ensuring that no values were missing. As mentioned, we used split training and test datasets, while no k-fold validation was carried out. Table 4 summaries classification performance. It is found that RF and SVM obtained approximately 90% accuracy, followed by LR and GNB.

In practical situation, unfortunately, missing values are sometimes observed due to some reasons. We subsequently evaluated the models by incorporating samples having missing values. Four imputation schemes were then tested, to clarify which strategy is the best for our data. The other experimental setup was the same as the previous experiment. Table 4 also includes the results. There is not a big difference among the imputation schemes, however, the experimental results demonstrated that the k-NN imputation achieved better accuracy than the others, particularly for RF; the accuracy of RF-based classification model with k-NN missing-value imputation, was roughly 96%. Figure 4 also depicts confusion matrices.

TABLE IV. CLASSIFICATION ACCURACY IN EACH ML MODEL WITH DIFFRENT IMPUTATION METHODS.

Model	Accuracy				
	w/o imp.	w/ imputation (N=368)			
	(N=336)	Removing	Mean	Median	k-NN
LR	0.86	0.87	0.87	0.87	0.87
SVM	0.90	0.88	0.88	0.88	0.89
RF	0.91	0.92	0.92	0.92	0.96
GNB	0.85	0.89	0.89	0.89	0.85

LR ACC 86%		Predicted	
		T	F
Actual	T	163	20
	F	31	154

SVM ACC 89%		Predicted	
		T	F
Actual	T	169	14
	F	26	159

RF ACC 97%		Predicted	
		T	F
Actual	T	181	9
	F	4	174

GNB ACC 85%		Predicted	
		T	F
Actual	T	161	23
	F	34	150

Fig.4. Confusion matrices in each ML model.

b) Cross Validation

The k-fold cross validation was conducted to check the performance of the ML model for developing high-accuracy classification. Three-fold cross validation was carried out in this paper. Classification was repeated three times with different subsets (A: data subset 1, B: data subset 2 and C: data subset 3). Then the average of the classification accuracy was obtained. Table 5 shows the results.

TABLE V. CROSS-VALIDATION RESULTS OF RF MODEL.

RF	Training	Testing	Accuracy	Precision	Recall	F1 score
Fold 1	A + C	B	0.96	0.95	0.95	0.95
Fold 2	B + C	A	0.96	0.95	0.96	0.95
Fold 3	A + B	C	0.95	0.95	0.95	0.95
		Average	**0.957**	**0.952**	**0.951**	**0.950**

We further checked the model and results. First, we tried to optimize model hyperparameters. In the cross-validation, the RF model with default parameters achieved an accuracy of 96.5%. It is found that the model achieved 97.9% after the tuning. As well known, hyperparameter tuning improves model reliability and performance, avoiding overfitting. We also investigated the learning curve for training and validation datasets, according to training data size (see Figure 5). The small gap between training and validation scores indicates good generalization and minimal overfitting. Overall the learning curve confirms the RF classifiers' robustness and strong performance on this dataset.

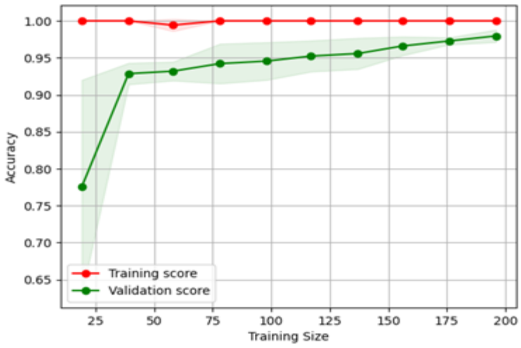


Fig.5. Learning curve of our RF model.

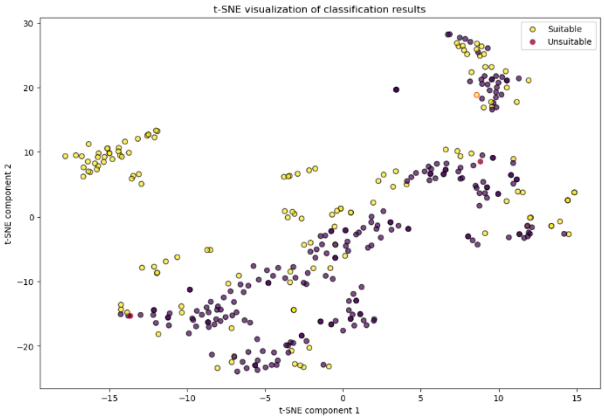


Fig.6. Visualisation of classification results.

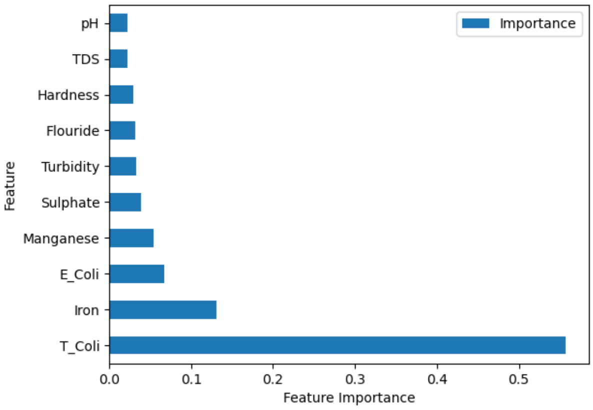


Fig.7. Feature importance in our RF classifier.

c) Visualization and Feature Importance

We conducted visualization to classification results. We obtained Figure 6 by applying t-SNE to visualize the map of the suitability of water quality and its misclassification. We can see some clusters of unsuitable water samples, while the other samples and suitable ones make one large cluster.

An RF model provides feature importance information in addition to the classification results. Figure 7 illustrates the result. The predictor importance for the model shows that T.coli is the most important feature, followed by iron, E.Coli, Manganese, sulphate, and turbidity.

D. Discussion

The studies conducted in different regions, including India, Malaysia, Bangladesh, Algeria, Pakistan, and Timor-Leste, highlight the effectiveness of different machine learning techniques for water quality prediction. Radhakrishnan and Pilla [23] in India achieved the highest accuracy of 98.5% using a decision tree algorithm on lake and river data. Malek et al. [24] in Malaysia used gradient boosting on river water and achieved an accuracy of 94.9%, while Gupta and Mishra [25] also used gradient boosting on river water and achieved perfect accuracy. In Bangladesh, Khan et al. [26] demonstrated the highest accuracies with gradient boosting (100%) and principal component regression (95%) for lake water. In Algeria, Derdour et al. [27] found SVM is highly effective for groundwater prediction with an accuracy of 95.4%. In Pakistan, Mohd et al. [23] applied RF to canal water with an accuracy of 91%. Our RF model in Timor-Leste for groundwater achieved an accuracy of 96%. These results suggest that gradient boosting and random forest are consistently high-performing models across water types and regions, on the other hand, the choice of the best model may depend on the specific characteristics of water, in addition to water, soil and climate environments. We need to investigate ML models carefully in order to find the best solution for new water data.

IV. CONCLUSION

This research involves monitoring water quality data from 26 wells in Dili, Timor-Leste over two years, focusing on 11 parameters. Groundwater samples exhibited considerable variation in turbidity, hardness, TDS, E.coli and T.coli concentrations, with values exceeding the desirable and permissible limits set by WHO and NDQWQ. The suitability in the map of water quality status indicates 64% of water samples are safe for human consumption. Regarding the imputation method of missing value, k-NN is the best method for the imputation of missing value for our dataset. In order to accomplish water suitability classification, we tested four machine learning classifiers. All the models performed well, in particular the RF has significantly better prediction performance with 96% accuracy than that of LR, GNB, and SVM. The model was validated using the cross-validation method. We also investigated feature importance and visualization results obtained from the RF model.

In future work, we will apply the proposed model to predict the water quality to different datasets in Timor-Leste. We have a plan to incorporate the water data and the other data e.g. rainfall data, humidity data to further improve the performance. Apart from the engineering standpoints, to provide safe water to all the citizens in Timor-Leste, we continue long-term monitoring of environmental changes like sea-level rise, urbanisation, and any climate influences, as well as having discussions using the results.

ACKNOWLEDGMENT

The authors extend their sincere gratitude to the Japan International Cooperation Agency (JICA), the Graduate Studies of Human Resources Development in Science, Technology and Innovation for providing generous scholarship funding enabling the first author to conduct this research. We also thank the Ministry of Public Works of Timor-Leste, particularly the National Direction of Water Supply, for their invaluable contribution of essential data for this study.

REFERENCES

[1] "The value of water – for our survival, peace and prosperity," https://www.un.org/development/desa/en/news/sustainable/the-value-of-water.html. accessed on 17 July 2024

[2] UNICEF, "water-and-sanitation/drinking-water/," https://data.unicef.org/topic/water-and-sanitation/drinking-water/. accessed on 17 July 2024

[3] Z. X. da Costa *et al.*, "Wastewater management strategy for resilient cities - case study: challenges and opportunities for planning a sustainable Timor-Leste," *Land (Basel)*, vol. 13, no. 6, p. 799, 2024, doi: 10.3390/land13060799.

[4] S. Aouiti *et al.*, "Selected case studies on the environment of the mediterranean and surrounding regions groundwater quality assessment for different uses using various water quality indices in semi-arid region of central Tunisia", doi: 10.1007/s11356-020-11149-5/Published.

[5] M. Ximenes *et al.*, "Initial observations of water quality indicators in the unconfined shallow aquifer in Dili city, Timor-Leste: suggestions for its management," *Environmental Earth Sciences*, vol. 77, no. 19, 2018, doi: 10.1007/s12665-018-7902-8.

[6] D. Pinto *et al.*, "Groundwater environment in Dili, Timor-Leste," in *Groundwater Environment in Asian Cities*, Elsevier, 2016, pp. 263–286. doi: 10.1016/B978-0-12-803166-7.00012-X.

[7] N. Nasir *et al.*, "Water quality classification using machine learning algorithms," *Journal of Water Process Engineering*, vol. 48, p. 102920, 2022, doi: 10.1016/j.jwpe.2022.102920.

[8] R.-C. Chen *et al.*, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, p. 52, 2020, doi: 10.1186/s40537-020-00327-4.

[9] K. Mei *et al.*, "Modeling of feature selection based on random forest algorithm and pearson correlation coefficient," *Journal of Physics: Conference Series*, vol. 2219, no. 1, 2022, doi: 10.1088/1742-6596/2219/1/012046.

[10] S. Tyagi *et al.*, "Water quality assessment in terms of water quality index," *American Journal of Water Resources*, vol. 1, no. 3, pp. 34–38, 2020, doi: 10.12691/ajwr-1-3-3.

[11] S. Mukate *et al.*, "Development of new integrated water quality index (IWQI) model to evaluate the drinking suitability of water," *Ecological Indicators*, vol. 101, pp. 348–354, 2019, doi: 10.1016/j.ecolind.2019.01.034.

[12] Guidelines, "National drinking water quality government of republic democratic Timor-Leste," 2016.

[13] D. W. Hosmer *et al.*, "Applied Logistic Regression," Wiley, 2013. doi: 10.1002/9781118548387.

[14] H. Li, Z. Lü *et al.*, "Support vector machine for structural reliability analysis," *Applied Mathematics and Mechanics*, vol. 27, no. 10, pp. 1295–1303, 2006, doi: 10.1007/s10483-006-1001-z.

[15] L. Breiman, "Random Forests," 2001.

[16] H. Zhang, "The optimality of naive bayes," https://www.aaai.org.

[17] S. Moccia *et al.*, "Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics," *Elsevier Ireland Ltd*. 2018, doi: 10.1016/j.cmpb.2018.02.001.

[18] WHO., "Guidelines for the Safe Use of Wastewater, Excreta and Greywater, Volume 1 : Policy and Regulatory Aspects," https://www.who.int/publications/i/item/9241546824.

[19] Guidelines, "National drinking water quality government of republic democratic Timor-Leste," 2016.

[20] "Guidelines for drinking-water quality 4th edition incorporating the first addendum."

[21] S. Tyagi *et al.*, "Water quality assessment in terms of water quality index," *American Journal of Water Resources*, vol. 1, no. 3, pp. 34–38, 2020, doi: 10.12691/ajwr-1-3-3.

[22] R. Das Kangabam *et al.*, "Development of a water quality index (WQI) for the Loktak Lake in India," *Applied Water Science*, vol. 7, no. 6, pp. 2907–2918, 2017, doi: 10.1007/s13201-017-0579-4.

[23] M. Y. Shams *et al.*, "Water quality prediction using machine learning models based on grid search method," *Multimed Tools Applications*, vol. 83, no. 12, pp. 35307–35334, 2023, doi: 10.1007/s11042-023-16737-4.

[24] N. H. A. Malek *et al.*, "Prediction of water quality classification of the Kelantan river Basin, Malaysia, using machine learning techniques," *Water (Basel)*, vol. 14, no. 7, p. 1067, 2022, doi: 10.3390/w14071067.

[25] D. Gupta *et al.* "Development of entropy-river water quality index for predicting water quality classification through machine learning approach," *Stochastic Environmental Research and Risk Assessment*, vol. 37, no. 11, pp. 4249–4271, 2023, doi: 10.1007/s00477-023-02506-0.

[26] Md. S. Islam Khan *et al.*, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 4773–4781, 2022, doi: 10.1016/j.jksuci.2021.06.003.

[27] A. Derdour *et al.*, "Designing efficient and sustainable predictions of water quality indexes at the regional scale using machine learning algorithms," *Water (Basel)*, vol. 14, no. 18, p. 2801, 2022, doi: 10.3390/w14182801.