JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Design of machine learning-based water quality prediction system with recursive feature elimination cross-validation

James Julian Ghossa[1,*], Annastya Bagas Dewantara[2], Fitri Wahyuni[3]
[1,3]Department of Mechanical Engineering, Universitas Pembangunan Nasional Veteran Jakarta
[2]Department of Electrical Engineering, Universitas Pembangunan Nasional Veteran Jakarta
[1,2,3]Jl. RS. Fatmawati Raya, Cilandak 12450, South Jakarta, Indonesia
[*]Corresponding email: zames@upnvj.ac.id

Received 1 June 2023, Revised 23 August 2023, Accepted 28 August 2023

Abstract — Lack of clean water has become a problem in the world, and it is estimated that by 2025, there will be 2.8 billion people who will experience a shortage of clean water. The high demand for clean water and the limited water sources with proper potency is one of the main reasons for the need for a device capable of measuring the potability level of water that is flexible to carry and does not require high costs in the manufacturing process. In this paper, the design of machine learning-based potability devices with recursive feature elimination with cross-validation (RFECV) is carried out as a guide in making the design of a water potability detection system, and the results obtained from RFECV with the Random Forest (RF) algorithm have a higher accuracy value. 15.71 % better than the RF model, 6.85 % better than the support vector machine (SVM) model, and 8.57 % better than the artificial neural network (ANN) model trained without RFECV. The water potability prediction system's design selection is based on feature elimination results in the RFECV process. It is based on a literature review on device selection. The proposed water potability detection system consists of ESP32 as the primary computing device, an electrochemical spectroscopy–based Al/PET sensor to detect sulfate values with a sensitivity of 0.874 Ω/ppm, PH4502C as a pH measuring instrument with an accuracy of up to 98.10 %, WD-35802-49 electrode. as a device for measuring hardness in water with a measurement range of 0.4—–40,000 ppm, a total dissolved solids sensor to determine the solids content in water with an accuracy of up to 97.80 %, as well as a carbon-based sensor for measuring chloramines with a reading capacity of 186 nA/ppm.

Keywords – artificial neural network, potability, random forest, RFECV, support vector machine

Copyright ©2023 JURNAL INFOTEL
All rights reserved.

I. Introduction

Clean water is a fundamental need for human daily life. Clean water resources are essential in various fields of health, economy, animal husbandry, and agriculture, so the quality of clean water resources is often used as a benchmark for a country. Based on data taken from the World Health Organization (WHO), there are 1.2 billion people in the world who still do not have access to sanitation services, and it is estimated that by 2025 there will be 2.8 billion people who will live without water in 48 countries [1].

The level of potability of clean water refers to the suitability level of water that humans can consume. The potability of drinking water has an important role in preventing disease and epidemics. An inspection process is necessary to ensure the quality of clean water. This is done to ensure that no hazardous materials are entering the body, such as arsenic material, which can harm reproductive health [2], harmful bacteria detected in water with non-standard pH [3], as well as digestive diseases and skin diseases from material chlorine and low content of low dissolved oxygen [4]. Therefore, a measurement of the level of potability of water needs to be done to ensure that the water consumed does not contain harmful contamination and is safe for the body.

Several studies have been conducted to measure the level of potability using odor sensors and microbiology instruments to determine the level of toxicity in cyanobacteria and potability level measurements using the CO2 sensor on the PIC32 [5], the use of Random Forest (RF) for measuring potability using the parameters pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes and Turbidity [6], water quality monitoring system on ARM Cortex-A53 based on IoT [7], water quality control monitoring system [8], The potentiality level measurement approach uses machine learning methods such as RF, support vector machines (SVM), and artificial neural networks (ANN) which have been carried out with successive accuracies of 0.70, 0.58, and 0.56 [9], [10].

This study aims to build a design of a device capable of predicting the potability of water using a machine learning approach on low-cost microcontrollers using the recursive feature elimination with cross-validation (RFECV) method as a feature elimination method to increase the cost efficiency of the design as well as a method to increase the effectiveness of the water potability detection system. Water quality and potability levels can be determined through predictions from the readings of several sensors to read several parameters in water, namely turbidity, solids, pH, conductivity, and total organic carbon that has gone through RFECV. Using the RFECV algorithm in selecting the design of the water potability level detection system, the results obtained in selecting water potability measurement instruments will be more accurate, effective, and efficient, and it is expected to be a solution to clean water problems in the future.

II. Research Method

The potability instrument is a device used to determine the potability of water obtained through sensor readings to obtain information on parameters that determine water quality. Potability is a source of water that is safe to drink and does not contain harmful contaminants that are harmful to the body. In assessing the potability of water, several parameters are identified so that these parameters meet clean water quality standards. These standards are created by environmental and health organizations such as the WHO and the Environmental Protection Agency (EPA).

Table 1. Water Quality Standards by WHO and EPA

Parameter	Units	WHO	EPA
pH	pH	6.5-8.5	6.5-8.5
Total Dissolved Solids	mg/L	500 - 1000	500-1000
Turbidity	NTU	5	4
Conductivity	μS/cm	400	300
Total Organic Carbon	mg/L	-	0-2
Chloramines	mg/L	5	4
Sulfate	mg/L	250	500
Trihalomethanes	mg/L	0.1	0.1
Hardness	mg/L	300	120-170

RFECV has an important role in reducing the number of sensors used, as shown in Fig 1. with i representing the number of sensors and n_i represent the sensor at each index. In addition to impacting the accuracy of the resulting detection process, the RFECV method also has an important role in reducing the amount of power used, which increases the efficiency of device power usage, as represented in Fig. 1.

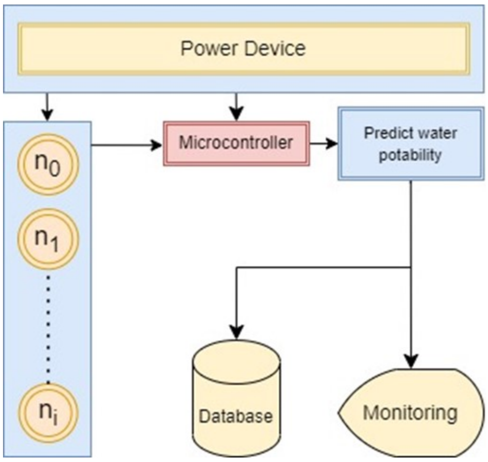

Fig. 1. System architectures.

A comparison of the performance of the proposed model was carried out against models that had been done previously using the same dataset based on evaluation metrics to determine the performance of the proposed model, as shown in Fig. 2. Through this comparison; it can be seen whether the use of RFEV has advantages over the models that have been done and what causes the models that have been made to have higher performance results than previous models.

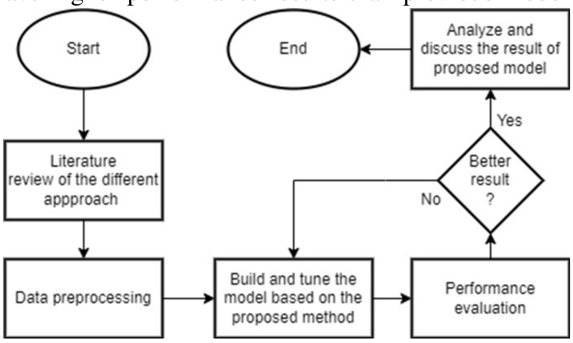
Start
End
Analyze and discuss the result of proposed model
Yes
Better result ?
No
Literature review of the different appproach
Data preprocessing
Build and tune the model based on the proposed method
Performance evaluation

Fig. 2. System architectures.

A. Dataset

The data used in this panel is data sourced from Kaggle [11], with a total of 3276 data samples. The data consists of 9 parameters shown in Table 1. The data is the result of synthesis data used to measure water potability. The value of the classification of water potability is divided into two binary parts, namely potable as a value of one (*1*) and not potable as a value of zero (*0*), referring to water quality standards in Table 1 with EPA quality standards.

The use of synthetic data in the machine learning training process has been carried out in various domains to prevent dataset imbalance and the presence of bias that tends to appear in real-world datasets [12], The use of synthetic datasets as primary data for the training process has reliability in predicting data actually with the impact of an insignificant decrease in accuracy and a low deviation size [13], [14].

Fig. 3 shows a flowchart of the design determination process, which begins with the pre-processing stage

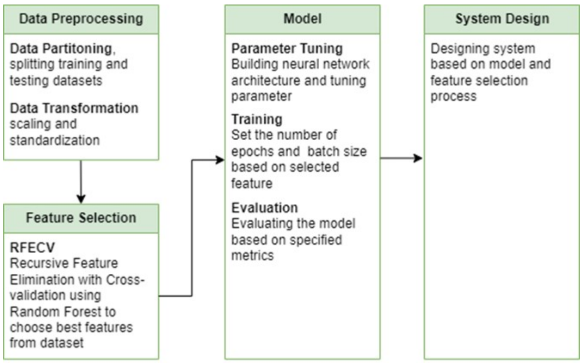
Data Preprocessing
Data Partitoning, splitting training and testing datasets
Data Transformation scaling and standardization
Feature Selection
RFECV Recursive Feature Elimination with Cross-validation using Random Forest to choose best features from dataset
Model
Parameter Tuning Building neural network architecture and tuning parameter
Training Set the number of epochs and batch size based on selected feature
Evaluation Evaluating the model based on specified metrics
System Design
Designing system based on model and feature selection process

Fig. 3. Data pipeline diagram.

by carrying out data imputation and standardization before entering the training stage. Data imputation is done to fill in the blank data to increase the accuracy and performance of the data [15]. The data imputation process is done by filling in the empty data with the average value of each data feature. After the data imputation process is done, the standardization process is done.

B. Data Standardization

Data standardization is carried out to avoid overscaling the input data features and ensure each feature is on the same scale. The data standardization process is carried out by calculating the mean-$\bar{X}$ and standard deviation s values of each feature using (1) and (2); after the mean and standard deviation values are obtained, the z-score value z can be calculated using (3).

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1}$$

$$S=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}} \tag{2}$$

$$z=\frac{X_i-\bar{X}}{s} \tag{3}$$

C. RFECV

RFECV is an algorithm used to determine the most optimal features from the dataset to get the highest accuracy [16].The use of RFECV is carried out by eliminating features that are not needed, thereby speeding up computation time and saving the power required to make predictions [17]. RFECV is done by conducting training from a dataset using a predetermined model. Random forest (RF) is the basic RFECV model because of RF's high accuracy on small datasets rather than neural networks [18]. The RF training process begins by calculating the $Gini$ impurity, $Gini_{t,j}$ valueof each decision tree on randomly selected features t from a subset of the dataset to determine which features will be used as root nodes and leaf nodes by calculating the probability of each categorical value $P_{t,j}$ For each feature input to the target feature j from total unique categorical values in the target class K, the value of $P_{t,j}$ It obtained by calculating the total number of each categorical value in the target class $n_(t,j)$ from given categorical feature values in class i and the total number of all categorical target values in the $i-index$ N_i is shown in (4) and (5).

$$P_{t,j}=\frac{n_{t,j}}{N_i} \tag{4}$$

$$Gini_{t,j}=1-\sum_{j=1}^{K}P_{t,j}^2 \tag{5}$$

The $Gini$ impurity of each definite value for each feature is calculated using (6) to obtain the weighted sum value of the $Gini$ to determine the effect of the feature on the target value, which can be used to determine which feature has an important role in determining the root node, child nodes, and leaf nodes of each decision tree from given total number of features C and the total number of all categorical target values in the classN_t.

$$Weighted\left(Gini_t\right)=\frac{1}{N_t}\sum_{i=1}^{C}N_i*Gini_{t,i} \tag{6}$$

Out–of–bag (OOB) values are data not used in the training process by decision trees. OOB error values are used to evaluate model performance and determine the important feature rank of the model. The OOB error is calculated by entering OOB data into a decision tree that is not used in the training data subset and comparing the results of the predictions with the actual data. The value of OOB is used to calculate the mean decrease in an impurity by calculating the average of the $Gini$ impurity in each feature in each decision tree. RFECV is done by setting the k-value and step value you want to use. After selecting the step and k-value, RFECV divides the data into k-folds with $k-1$ folds as test data. The average value of OOB error is calculated for all k-folds. The calculation results from the OOB error are used by the grid search to determine the best hyperparameters to determine how many features and which features need to be used to get the highest accuracy.

D. Evaluation Metrics

Performance evaluation is a critical thing from the model testing stage. This aims to determine the performance of the model based on certain metrics; several metrics that are commonly used to measure model performance are shown in (7) up to (10) [19]. The five metrics are calculated by comparing the predicted label with the ground truth label. A true positive (TP) is when the predicted label has the same value as the positive ground truth label. A false positive (FP) is when the predicted label differs from the positive ground truth label. A true negative (TN) is when the predicted label matches the positive ground truth label. A false negative (FN) is when the predicted label differs from the negative ground truth label.

$$precission = \frac{TP}{TP + FP} \tag{7}$$

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$recall\,(TPR) = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times precission \times recall}{precission + recall} \tag{10}$$

AUC-ROC refers to the area under the receiver operating characteristic curve. It is a machine learning algorithm metric used to evaluate classification models. The ROC curve is a graph that plots the rate of true positives (TPR) against the false positive rate (FPR) to indicate the performance of a classification model at all classification criteria [20]. The AUC is obtained by measuring the two-dimensional Area under the entire ROC curve from (0,0) to (1,1). AUC is a metric that spans from 0 to 1 and offers an aggregate performance assessment over all possible classification thresholds using (9) and (11).

$$Sensitivity(FPR) = \frac{FP}{FP + TN} \tag{11}$$

III. Result

The data collected consists of nine numerical input data and 1 data label categorical potability, totaling 3,276 data lines data. Data preprocessing was done by identifying empty values and finding empty data in the pH, sulfate, and Trihalomethanes features, respectively, 491, 781, and 162 data. The data imputation process is carried out to fill in the blank data by filling in the blank data with the average feature value.

The high accuracy cannot be separated from the correlation between parameters, Fig. 4 shows the correlation between parameters with other parameters. Fig. 4 shows a strong correlation between the levels of pH and solids, pH and hardness, sulfate and hardness, and sulfate and solids, which are visualized through heatmap graphics. The existence of a correlation for each of these parameters will create a pattern that can be used by machine learning to determine the potability level of drinking water.

Islam et al. showed a correlation between solids and pH by sampling water bottles from 14 domestic brands in Dhaka City. They showed that the higher the value of the solid, the lower the pH value and the higher the conductivity value, and vice versa [21]. Price et al. also show the strong correlation between hardness and pH through zinc toxicity tests and different treatments of pH changes that affect the level of hardness of drinking water [22]. Research conducted by Kothan et al. indicates alkalinity, solids, and hardness correlate with each other through the concentration of sulfate ions [23]. A correlation between parameters described by the heatmap and research above illustrates the existence of related parameters in determining the level of potability.

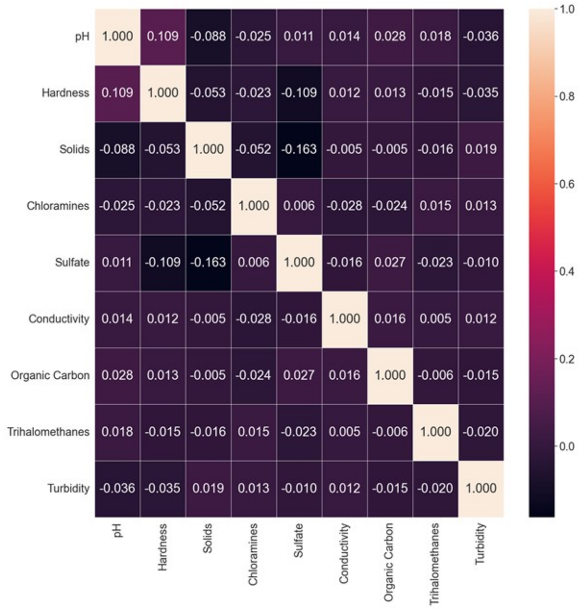

Fig. 4. Correlations between each feature.

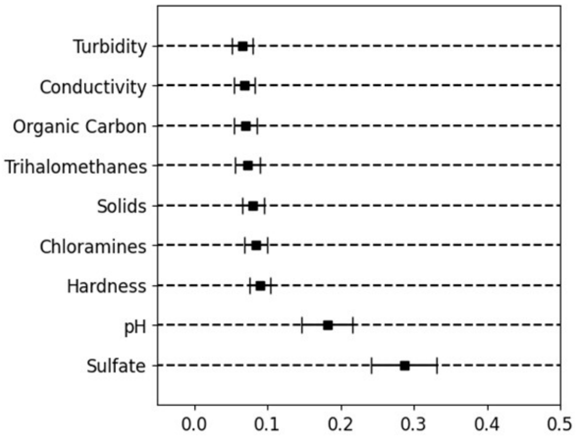

Fig. 5. RFECV mean decrease impurity for each feature.

The results obtained in the RFECV process by setting a value of k = 10 and a weight of step = 1 are shown in Fig. 6. Based on the data shown in Fig. 5 , the input sulfate feature has the highest influence in determining the water quality and potability level of a drink, and the turbidity feature has the lowest feature in knowing the level of potability.

Based on information from RFECV, the training process is carried out using five features with the RF algorithm. The results obtained from the training are shown in Fig. 7. Through a grid search on RFECV results, the highest accuracy was obtained when training used five features, sulfate, pH, hardness, solids, and chloramines, as shown in Fig. 5 using (7) up to (11). RFECV has a higher AUC result of 6.85 % higher than the SVM model and 8.57 % higher than the ANN

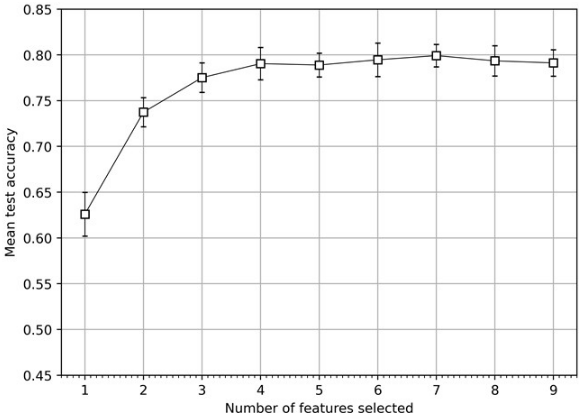

Fig. 6. RFECV accuracy result of testing data.

model. Without RFECV, it has 15.71 % higher accuracy than the RF model without RFECV.

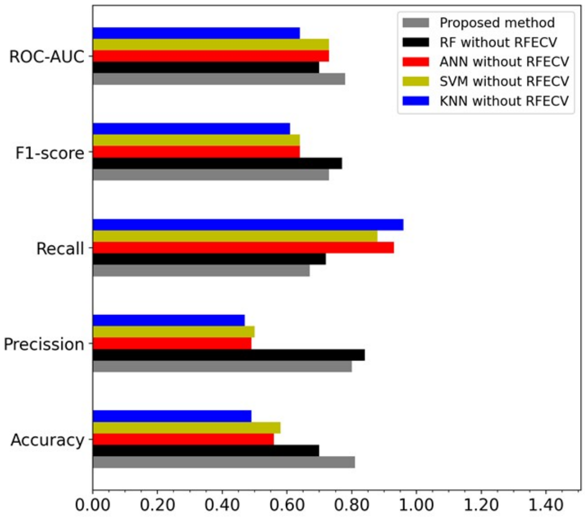

Fig. 7. Comparison of proposed method results toward other ML approaches.

IV. Discussion

This section discusses model effectiveness, System design, and challenge/opportunities.

A. Model Effectiveness

Models with high accuracy and AUC but lower precision, recall, and F1-score show that the model can predict negative classes but is not good at predicting positive classes. Various things can cause this. One of them is an imbalanced dataset on feature potability, model complexity that is too deep in predicting the model, which causes a lack of generalization of the model on the test dataset, and data quality that is noisy and contains outliers, which can affect the value of the evaluation metrics.

This RFECV model proves that feature reduction from the dataset on detection has a higher accuracy than the model trained using the entire dataset. Reducing features in the dataset also reduces computation time, costs, and the power load to be used, so in addition to increasing the tool's effectiveness, this RFECV technique is proven to increase the cost efficiency of the design.

B. System Designs

The model from the random forest will learn patterns from the data used in the potency measuring device. ESP32 is used in the deployment process for its architecture, which is built using a 32-bit RISCT Tensilica Xtensa LX106 MCU, which has been equipped with FPU and DSP and has a clock speed of 240 MHz and 520 kb SRAM [24]. ESP32's ability to perform machine learning-based computing has been proven through traffic decrease detection and object detection based on convolutional neural networks b25. Testing the inference time of ESP32 in computing Deep Learning on 50 input layers and 200 hidden layers has a speed of 1,599 μs [26]; machine learning algorithm tests on ESP32 have been carried out in various fields, such as autonomous vehicles, hand gesture recognition, and speech recognition with various models, such as RF, SVM, and ANN [27], This makes ESP32 a major computational component in the design of water potability level detection systems.

To be able to adapt the machine learning model into ESP32, the model is converted into a form that can be interpreted by embedded systems, taking into account limited resources such as power consumption and memory allocation from SRAM, ROM, or ROM [28], by converting the model to in special formats such as converting into the C++ model from the converted tflite model or into the tinyML model [29]. Based on the RFECV results, it is known that the five features that need to be needed to determine the potability level of water are sulfate, pH, hardness, solids, and chloramines. Sensor selection is made based on the parameters shown in Table 2, considering measurements' accuracy, range, and sensitivity.

C. Challenge and Opportunities

The approach to designing a water potability detection system using the machine learning method is known to bring benefits in cost efficiency and the effectiveness of the tool's predictive ability. One of the challenges of a system design approach using machine learning is the use of large amounts of data and good data quality. System design based on data can be used to design more flexible and more resistant systems under certain conditions.

The challenge in the design process using a machine learning approach is that it often requires domain-specific knowledge, which is often misinterpreted by machine learning models and makes the model inaccurate. Besides making designs based only on data, it can cause bias and misleading predictions, this is dangerous, especially in the design of a health system that can impact the health of its users.

Table 2. Sensor Devices for Each arameter

Parameter	Sensor	References
Sulfate	The electrochemical spectroscopy-based device made with a laser-inscribed technique with - material polyethylene terephthalate film coated with a thin layer of aluminum film (Al/PET) can read sulfate in water at 0.1 ppm – 1000 ppm with a sensitivity of 0.874 ω/ppm	[24]
pH	PH4502C, an electrochemical electrode made of Ag/AgCl and KCl solution, with an accuracy of 98.10%.	[25]
Hardness	WD-35802-49, an ion-selective electrode that could measure Ca2+ values, was converted into an electrical signal with a range of measurement from 0.4 ppm – 40.000 ppm.	[26]
Solids	TDS sensor, stainless steel that is used to measure the conductivity of dissolved solids that proof had an accuracy of 97.80%.	[27]
Chloramines	A carbon-based free chlorine sensor with 186 nA/ppm.	[28]

V. Conclusion

Based on this research, it is known that the use of the RFEV method in the design selection process has a significant impact on system effectiveness. The use of RFECV can reduce nine parameters to five parameters. The use of RFECV with the RF algorithm has a better accuracy value of 15.71 % than the RF model, 6.85 % better than the SVM model, and 8.57 % better than the AN model trained without RFECV, with values of accuracy, precision, recall, $F1$-score, and AUC respectively 0.81, 80, 0.67, 0.73, and 0.78.

The water potability prediction system's design selection is based on feature elimination results in the RFECV process. It is based on a literature review on device selection. The proposed water potability detection system consists of ESP32 as the main computing device, electrochemical spectroscopy-based Al/PET sensor to detect sulfate values with a sensitivity of 0.874 Ω/ppm, PH4502C as a pH measuring device with an accuracy of up to 98.10 %, WD–35802–49 electrode. as a device for measuring the hardness level in water with a measurement range of 0.4 - 40,000 ppm, a total dissolved solids sensor to determine the solids content in water with an accuracy of up to 97.80%, as well as a carbon-based sensor for measuring chloramines with a reading capability of 186 nA/ppm.

References

[1] T. Pl. M. Editors, “Clean water should be recognized as a human right,” *PLoS Med*, vol. 6, no. 6, p. e1000102, Jun. 2009, doi: 10.1371/JOURNAL.PMED.1000102.

[2] M. L. Kile, E. G. Rodrigues, M. Mazumdar, C. B. Dobson, N. Diao, M. Golam, Q. Quamruzzaman, M. Rahman, and D. C. Christiani, “A prospective cohort study of the association between drinking water arsenic exposure and self-reported maternal health symptoms during pregnancy in Bangladesh,” *Environ Health*, vol. 13, no. 1, pp. 1—–13, Apr. 2014, doi: 10.1186/1476-069X-13-29/TABLES/6.

[3] F. Oluwafemi and M. E. Oluwole,“Microbiological examination of sachet water due to a cholera outbreak in Ibadan, Nigeria,”*Open J Med Microbiol*, vol. 2012, no. 03, pp. 115—–120, Sep. 2012, doi: 10.4236/OJMM.2012.23017.

[4] J. N. Halder and M. N. Islam, “Water pollution and its impact on the human health” *Journal of Environment and Human*, vol. 2, no. 1, 2015, doi: 10.15764/EH.2015.01005.

[5] J. W. Gardner, H. W. Shin, E. L. Hines, and C. S. Dow, “An electronic nose system for monitoring the quality of potable water,” *Sens Actuators B Chem*, vol. 69, no. 3, pp. 336—–341, Oct. 2000, doi: 10.1016/S0925-4005(00)00482-2.

[6] J. Patel, C. Amipara, T. A. Ahanger, K. Ladhva, R. K. Gupta, H. O. Alsaab, Y. S. Althobaiti, and R. Ratna, “A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI,” *Computational Intelligence and Neuroscience*, vol. 2022, 9283293, doi: 10.1155/2022/9283293.

[7] R. Kondle, S . Dastagiri, and Mv. Lakshmaiah,“Implementation of IoT in embedded systems for real time water quality monitoring,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 4S2, pp. 1—–4, Mar. 2020, doi: 10.35940/IJITEE.D1021.0394S220.

[8] A. Menacho, P. Plaza, E. S. Cristóbal, R. Gil, F. García, C. Pérez, and M. Castro, “Arduino-based water analysis pocket lab,” in *2021 World Engineering Education Forum/Global Engineering Deans Council (WEEF/GEDC)*, 15-18 November 2021, Madrid, Spain, pp. 205—–210, 2021, doi: 10.1109/WEEF/GEDC53299.2021.9657377.

[9] L. Pacheco and S. Kaddoura, “Evaluation of machine learning algorithm on drinking water quality for better sustainability,” *Sustainability*, vol. 14, no. 18, p. 11478, 2022, doi: 10.3390/SU141811478.

[10] D. Poudel, D. Shrestha, S. Bhattarai, and A. Ghimire, “Comparison of machine learning algorithms in statistically imputed water potability dataset,” *Journal of Innovations in Engineering Education*, vol. 5, no. 1, pp. 38—–46, 2022, doi: 10.3126/JIEE.V5I1.42265.

[11] Kaggle, “Water Quality.” https://www.kaggle.com/datasets/adityakadiwal/water-potability (accessed March. 14, 2023).

[12] B. N. Jacobsen, “Machine learning and the politics of synthetic data,” *Big Data & Society*, vol. 10, no. 1, 2023, doi: 10.1177/20539517221145372.

[13] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, “Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing,” *JMIR Med Inform*, vol. 8, no. 7, p. e18910, Jul. 2020, doi: 10.2196/18910.

[14] S. Baressi Šegota, N. Anđelić, M. Šercer, and H. Meštrić, “Dynamics modeling of industrial robotic manipulators: A machine learning approach based on synthetic data, *Mathematics*, vol. 10, no. 7, p. 1174, 2022, doi: 10.3390/MATH10071174.

[15] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of Performance of Data Imputation Methods for Numeric Dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913—–933, 2019, doi: 10.1080/08839514.2019.1637138.

[16] E.-H. Huang, H.-W. Hu, W.-L. Jheng, K.-Y. Chen, C.-H. Liu, H.-Y. Chi, T.-W. Chang, C.-Y. Wu, C.-H. Un, H.-M. Lin, C.-W. Chen, and J.-F. Wang, “Feature selection for intradialytic blood pressure value prediction using GRU-based method under RFECV algorithm,” in *2021 9th International Conference on Orange Technology (ICOT)*, 16-17 December 2021, Tainan, Taiwan, doi: 10.1109/ICOT54518.2021.9680645.

[17] R. A. Mowri, M. Siddula, and K. Roy, “A comparative performance analysis of explainable machine learning models with and without RFECV feature selection technique towards

ransomware classification," *ArXiv*, Dec. 2022, doi: 10.1109/ACCESS.2017.DOI.

[18] T. Han, D. Jiang, Q. Zhao, L. Wang, and K. Yin, "Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery," *Transactions of the Institute of Measurement and Control*, vol. 40, no. 8, pp. 2681—–2693, 2017, doi: 10.1177/0142331217708242.

[19] B. J. Erickson and F. Kitamura, "Performance metrics for machine learning models," *Radiol Artif Intell*, vol. 3, no. 3, 2021, doi: 10.1148/RYAI.2021200126.

[20] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit*, vol. 30, no. 7, pp. 1145—–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

[21] R. Islam, S. M. Faysal, M. R. Amin, F. M. Juliana, M. J. Islam, M. J. Alam, M. N. Hossain, and M. Asaduzzaman, "Assessment of pH and total dissolved substances (TDS) in the commercially available bottled drinking water," *IOSR Journal of Nursing and Health Science*, vol. 6, no. 5, pp. 35—–40, doi: 10.9790/1959-0605093540.

[22] G. A. V. Price, J. L. Stauber, A. Holland, D. J. Koppel, E. J. V. Genderen, A. C. Ryan, and D. F. Jolley, "The influence of hardness at varying pH on zinc toxicity and lability to a freshwater microalga, Chlorella sp.," *Environ Sci Process Impacts*, vol. 24, no. 5, pp. 783—–793, May 2022, doi: 10.1039/D2EM00063F.

[23] V. Kothari, S. Vij, S. K. Sharma, and N. Gupta, "Correlation of various water quality parameters and water quality index of districts of Uttarakhand," *Environmental and Sustainability Indicators*, vol. 9, p. 100093, Feb. 2021, doi: 10.1016/J.INDIC.2020.100093.

[24] J. Ivković and J. Lužija Ivković, "Analysis of the performance of the new generation of 32-bit microcontrollers for IoT and big data application," *ICIST 2017 - 7th International Conference on Information Society and Techology*, Kopaonik, Serbia, 2017.

[25] A. N. Kokoulin, A. I. Tur, A. A. Yuzhakov, and A. I. Knyazev, "Hierarchical convolutional neural network architecture in distributed facial recognition system," in *Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2019*, pp. 258—–262, Feb. 2019, doi: 10.1109/EICONRUS.2019.8656727.

[26] M. Z. H. Zim, "TinyML: Analysis of Xtensa LX6 microprocessor for Neural Network Applications by ESP32 SoC,"*ArXiv*, 2021, doi: 10.13140/rg.2.2.28602.11204.

[27] P. P. Ray, "A review on TinyML: State-of-the-art and prospects," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1595—–1623, Apr. 2022, doi: 10.1016/J.JKSUCI.2021.11.019.

[28] R. David, J. Duke, A. Jain, V. J. Reddi, N. Jeffries, J. Li, N. Kreeger, I. Nappier, M. Natraj, S. Regev, R. Rhodes, T. Wang, and P. Warden, "TensorFlow lite micro: Embedded machine learning on TinyML systems," *ArXiv*, 2021.

[29] I. Katsidimas, T. Kotzakolios, S. Nikoletseas, S. H. Panagiotou, and C. Tsakonas, "Smart objects: Impact localization powered by TinyML," *SenSys 2022*, November 6–9, 2022, Boston, MA, USA, doi: 10.1145/3560905.3568298.