A Comparative Study of Machine Learning Techniques for Water Potability Classification

Robert G. de Luna
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
rgdeluna@pup.edu.ph

Verna C. Magnaye
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
vcmagnaye@pup.edu.ph

Rose Anne L. Reaño
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
ralreano@pup.edu.ph

Karina L. Enriquez
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
klenriquez@pup.edu.ph

Joeben More R. Dalguntas
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
jmrdalguntas@iskolarngbayan.pup.edu.ph

Adrien Joshua M. Lizardo
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
ajmlizardo@iskolarngbayan.pup.edu.ph

Earl Stephen A. Molino
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
esamolino@iskolarngbayan.pup.edu.ph

Allen Andrew L. Pucyutan
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
aalpucyutan@iskolarngbayan.pup.edu.ph

Jayvee C. Solis
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
jvcsolis@iskolarngbayan.pup.edu.ph

David Ysmael D. Umali
Polytechnic University of the Philippines
Sto. Tomas City, Batangas, Philippines
smvsalcedo@iskolarngbayan.pup.edu.ph

***Abstract* — Water is an essential natural resource for life on Earth, and it is the foundation of all living things. However, water pollution is a growing environmental concern caused by human activities, such as improper waste disposal and the discharge of untreated sewage. The consequences of this problem on human health and aquatic life highlight the need for effective supervision and administration of water reserves. This research paper aims to utilize a machine learning approach to predict water quality and identify the most influential features affecting water potability. These features were obtained from three methods, namely Univariate Selection, Recursive Feature Elimination, and Feature Importance, to identify the most influential features. The study compares the performance of various classification algorithms, including K-Nearest Neighbor, Decision Tree, Random Forest, AdaBoost, XGBoost, Linear Discriminant Analysis, Gaussian Naïve Bayes, Logistic Regression, MLPClassifier, and ExtraTree Classifier, using evaluation criteria such as accuracy, precision, recall, F1 score, and computational efficiency. After conducting all these processes, ExtraTree Classifier achieved the highest accuracy of 89 % among the compared machine learning models. Overall, the results of this research may contribute to better public health outcomes and improved management of water resources.**

Keywords—water potability, machine learning, classification, random forest, xgboost, extratree classifier

I. INTRODUCTION

Making up 71% of Earth's surface area, water is considered as one of the most important natural resources on which the planet relies. It is the foundation of all life: humans, animals, and plants. Its use extends beyond drinking as it covers fields of commerce, agriculture, and global trade via seas and oceans. Water, the "universal solvent," can disperse more chemicals than any other substance in the world. As a result, it is rapidly tainted. It dissolves easily and blends with hazardous substances from ranches, villages, and industries, contaminating the water and having detrimental consequences.

Water is essential for all living things, and life cannot live without it. It is indicated that many human habits and activities have a considerable and frightening impact on the quality of water supplies. As urbanization accelerates, wastewater from industrial production will continue to increase, corrupting water [1]. These wastes will cause water pollution into aquatic ecosystems if not properly cleaned [2]. Making almost 80% of the sewage generated by human activity released into rivers and oceans, contaminating the water, and spreading more than 50 diseases [3]. Unfortunately, pollution of water sources has resulted in a decline in drinking water supplies.

Water quality has declined significantly in recent decades due to pollution and other concerns [4]. Also, drinking contaminated or unsterilized water can cause gastrointestinal sickness, nutritional absorption problems, and malnutrition. Other diseases caused by consuming contaminated water, listed by [5], comprised were cholera, enteric fever, Hepatitis E, Hepatitis A, Norovirus, Shigella, and Campylobacter, which spread to communities throughout Asia. These effects are especially noticeable in children [6]. According to [7, 8], diarrhea is generated by water that is contaminated, poor cleanliness, and bad hygiene practices. The polluted water contributed to the spread of this ailment [9]. It kills roughly 829,000 people each year. 5.3% of these deaths occur in children under the age of five, accounting for over 300,000 casualties in this range of age.

In this paper, the researchers provide a machine learning approach that is based on previously collected data from a large number of water samples. This study aims to identify the features that have the greatest influence on water quality, predict water potability using machine learning, and compare multiple machine learning classification models to determine the best model for predicting water potability. These models were utilized as a tool to assist in interpreting the results and in decision-making. Finally, this research will contribute

some knowledge about the machine learning techniques needed to classify several datasets.

This study assumes that the water quality data used in this research will be successful and that the machine learning algorithms used are appropriate and effective for predicting water potability. Its focus is on identifying the features that have the greatest influence on water potability and the comparison of multiple classification models using machine learning techniques. The study utilizes previously collected data from a large number of water samples, and is limited to the geographical area where the samples were collected. Additionally, this study is limited to the use of machine learning algorithms and does not explore other techniques or methods for predicting water potability. This also does not address the practical implementation of the developed models in real-world scenarios. And is limited to self-generated dataset that may include bias, limited scope , quality issues, time and cost, and lack of diversity which should be taken into account when interpreting and analyzing the result.

By conducting the study, researchers can evaluate various machine learning methods to identify the most effective ways of assessing and monitoring water quality parameters that impact water potability. This comparison can help increase the accuracy and efficiency of water quality testing, contributing to better public health outcomes and improved management of water resources. In addition, the study can help develop more reliable and advanced methods for ensuring safe and clean drinking water. This is crucial for sustainable development and improving the quality of life for the country's population. The results of this study can serve as a reference for future researchers in similar contexts, helping to expand the knowledge base in this field.

II. Related Works

Using the correct approaches, classification in machine learning may yield an accurate prediction for a given input data[10]. To be usable and efficient, the raw dataset must be pre-processed [11]. Abuzir [12] eliminates the feature with poor correlation after pre-processing when their values are near to zero. These features will then be evaluated to several machine learning algorithms to determine which method has the best accuracy. With this way, [13, 14, 15, 16] identify the appropriate method. According to their findings, Random Forest, Decision Tree, and KNN outperform the other algorithms.

Water potability classification was not an entirely uncommon concept in machine learning. Establishing an effective classification model, however, remains challenging to this day. Poudel et al., [17] normalized X and unexpectedly standardized Y. They didn't balance the data and replace the missing value by its median. Using all features, they run the different algorithms like Artificial Neural Network, Logistic Regression, KNN, and Random Forest.

Nachaoui, M. et al. [18] normalized the dataset and replaced missing values with the dataset's mean. They did not compare how the substitution would affect the accuracy. They used all features and only use feature importance to classify the most important feature only after finishing the various machine learning models. Kaddoura [19] also use feature importance just for the objective of identifying the most significant feature, rather than for feature selection. The only difference is that she removed the missing values after data preparation because the total number of records collected is 2011.

Patel et al. [4] deals with missing data by replacing it with the mean value of the factor for which data is missing. It is then normalized and also balanced using Synthetic Minority Oversampling Technique (SMOTE). Their training data comprised 70% of the total dataset, while their testing data comprised only 30% of the total dataset. They also used all features to undergo Decision Tree, Random Forest, XGBoost, AdaBoost, and Support Vector Machine.

The study of Nataraj [20] drops the missing values and didn't normalize and didn't balance the data. She also only used 1000 samples on feature selection using feature importance and picked the top 3 of them. KNN is the only machine learning algorithm she used making the study doesn't have any comparison to other algorithms.

III. Methodology

The safety of drinking water is of utmost importance, as consuming contaminated water can lead to serious health issues. To ensure that water is safe for consumption, it is essential to classify it based on its potability. This classification process can be aided by machine learning techniques, which can create a model that takes into account various factors, such as pH levels, hardness and sulfate content, to determine water potability accurately.

However, there are several different machine learning algorithms that can be used for water potability classification, and it's not always clear which one is the most effective. That's why the goal of this study is to assess and compare multiple machine learning algorithms to determine which one is the most accurate and effective for classifying water potability. The research will use Fig. 1, a visual representation of the complete process, to choose the best machine learning algorithm for water potability categorization.

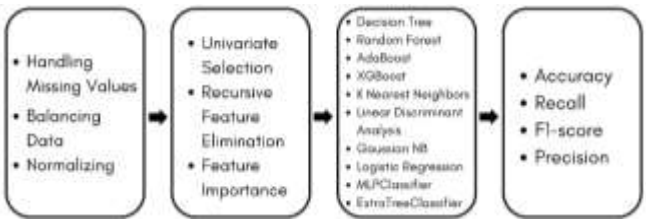

Fig. 1. Machine Learning Process for Classification of Water Potability

A. Data Collection and Preprocessing

The first step of the research involves collecting relevant data on water quality parameters. The dataset that was used in this study came from the Department of Public Health Engineering in Rajshahi Branch, Bangladesh, and it includes the 10 variables which are pH, hardness, total solids dissolve, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability [21]. The data was gathered from 3276 various waters.

After obtaining the dataset, the data is being preprocessed by determining the missing values, balancing it to avoid biasing, and normalizing. Dropping the missing values is the approach used in this study. Using SMOTE, the data is being balanced. In order to know if balancing the data will help improve the accuracy, it will also be compared to the accuracy of the unbalanced model. All of these models will be normalized to make the dataset uniform.

B. *Feature Selection*

Feature selection is a crucial step in the classification process, as it involves identifying the most significant features in a dataset that contribute the most to the task. This reduces the complexity of the data and enhances the models' performance. To achieve this, various feature selection techniques such as Univariate Selection (UVS), Recursive Feature Elimination (RFE), and Feature Importance were employed. By applying these techniques, the researcher can identify the features that are most essential in determining water potability.

Once the different feature selection techniques have been applied, the most commonly recurring feature among the results will be used for model creation. This ensures that the model is based on the most important features, enhancing its accuracy and reliability in classifying water potability.

C. *Model Training*

There are various machine learning algorithms for classifying water potability, depending on the size of the dataset, the complexity of the problem, and the performance metrics of interest.

In this step, we will apply different machine learning algorithms to classify water potability based on the selected features. The algorithms we will explore include decision trees, random forest, KNN, AdaBoost, XGBoost, Linear Discriminant Analysis, Gaussian NB, Logistic Regression, and MLPClassifier. We will use the Scikit-learn and XGBoost libraries in Python to implement these algorithms.

Ray and Sarker proposed all these algorithms: K-Nearest Neighbor, Decision Tree, Random Forest, AdaBoost, XGBoost, Linear Discriminant Analysis, Gaussian Naïve Bayes, Logistic Regression, MLPClassifier, and Extra Tree Classifier.

D. *Model Evaluation*

Following the selection of the model, the training dataset is used to train the model, if one exists. For the testing set, metrics like accuracy, precision, sensitivity, and F1-score are used to assess the model's performance. Hyperparameters are modified using approaches such as GridSearch to improve the model. Finally, we compare the performance of each algorithm to determine the most accurate and efficient method of identifying water potability.

IV. Results and Discussion

In this study, missing data were handled by dropping it. This method will be implemented throughout the entire duration of the research. Additionally, it is deemed justifiable to eliminate missing values considering the subject matter under consideration, which is the potability of water. There is also a lack of certainty if the potability of water will remain the same if the missing values are replaced with the mean or mode values. According to [24], at the multivariate level for groundwater, greater turbidity values suggest an acceptable water pH. This statement of [24] reveal that certain characteristics of water exhibit a correlation with the efficacy of excluding missing values. As such, it is deemed appropriate to utilize this approach for dropping missing values in this study.

Following the elimination of the missing values, the number of rows in the dataset is reduced to 2011. The subsequent measure to be undertaken is data balancing to prevent any potential bias in the dataset. An evaluation of the efficiency of this technique in improving the accuracy of the model will be conducted by comparing it with an unbalanced model. This comparative analysis will determine the most effective method for classifying water potability.

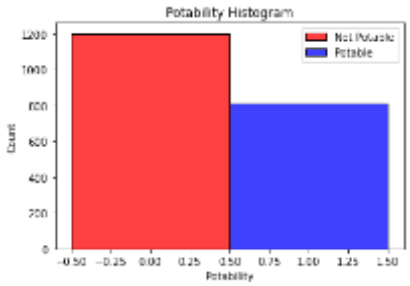

Fig. 2. Unbalanced Potability Ratio

According to the figure presented, there is a larger proportion of non-potable water than potable water. This model will be used on the unbalanced dataset. Before proceeding, the data was balanced with SMOTE to ensure that potable and non-potable water were represented in equal volumes as shown in Fig. 3.

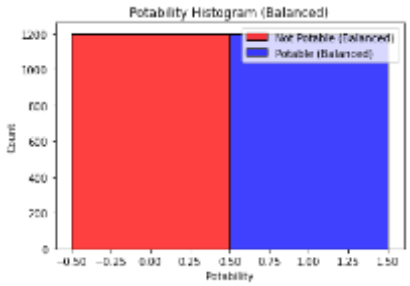

Fig. 3. Balanced Potability Ratio

Following the completion of the data preprocessing stage, the researcher proceeded to undertake feature selection utilizing three distinct techniques: Feature Importance, Univariate Selection (UVS), and Recursive Feature Elimination (RFE) in order to determine the most vital features. The researcher will also compare the accuracy of the model that has five (5) features and four (4) features.

The result in Fig. 4 in identifying the four (4) vital features for unbalanced dataset shows UVS method identifies Solids, Turbidity, Chloramines, and Conductivity as the most important features, while RFE selects Sulfate, Chloramines, Hardness and ph. Lastly, Feature Importance results are Hardness, Chloramines, pH, and Sulfate. As can be seen, the four features that occur most frequently across all three methods are pH, Chloramines, Hardness, and Sulfate.

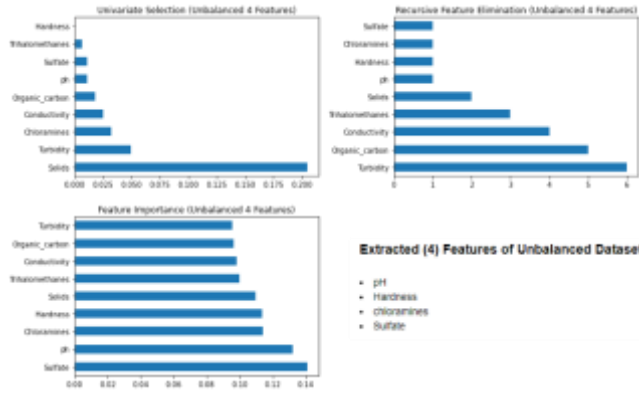

Fig. 4. Four Feature Selection (Unbalanced)

The results of finding the five (5) most important features in an unbalanced dataset in Fig. 5 reveal that the UVS technique added Organic Carbon to the previous four features, while RFE and Feature Importance both added Solids. pH, Chloramines, Hardness, Sulfate, and Solids are the five traits that appear most frequently across all three approaches.

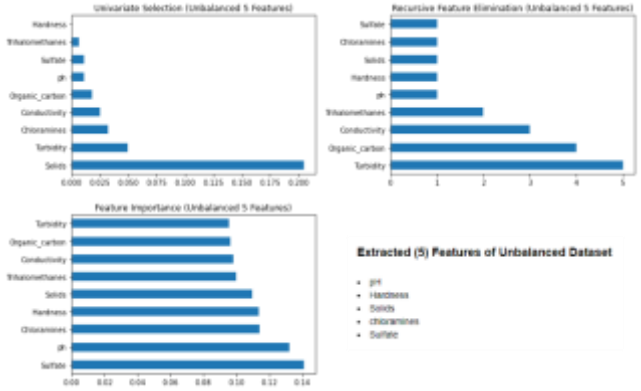

Fig. 5. Five Feature Selection (Unbalanced)

The four (4) prominent features selected for a balanced dataset as shown in Fig. 6 using the UVS method are pH, Turbidity, Chloramines, and Solids. Sulfate, Chloramines, Hardness, and pH are the chosen features by RFE, while the Feature Importance method identifies Hardness, Chloramines, pH, and Sulfate as the best features.

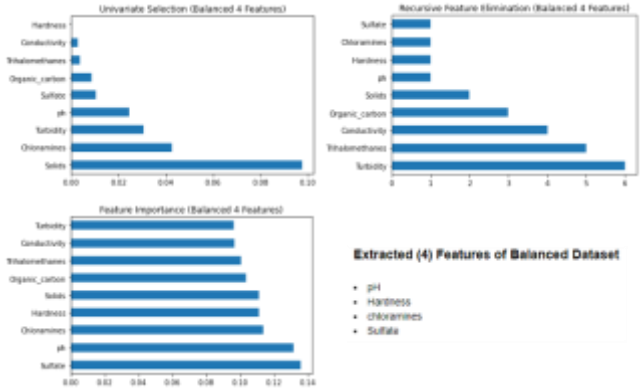

Fig. 6. Four Feature Selection (Balanced)

Overall, the most frequently identified features across all methods are pH, Hardness, Chloramines, and Sulfate.

The five (5) features that are selected as shown in Fig. 7 in the balanced dataset shows that UVS picked the previous four features and Sulfate. Both RFE and Feature importance added Solids, making the five most frequently occurring features are pH, Hardness, Chloramines, Sulfate, and Solids.

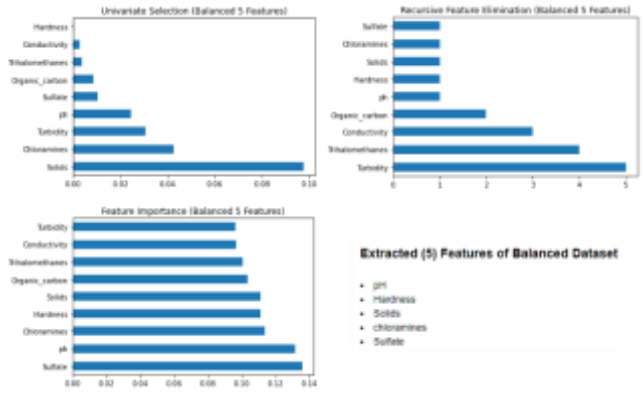

Fig. 7. Five Feature Selection (Balanced)

These features are then utilized in different machine learning algorithms to predict water potability. The use of different algorithms is critical as they help in analyzing and evaluating the best results of the predictions based on the selected features. The table below will show the accuracy of different machine learning models used for classification.

Parameters of the machine learning models are optimized and evaluated their performance in comparison to each other. This is an essential step in determining the most effective model for classifying water potability. By optimizing the models' parameters, we can fine-tune their performance and potentially improve their accuracy. Once the models have been optimized, we can evaluate and compare their performance to determine which one is the best fit for our classification task.

TABLE 1. COMPARISON OF ACCURACY OF OPTIMIZED-ALL-FEATURE MODEL

Optimized (All Features)						
	Unbalanced			Balanced		
Model	Holdout	Cross-Validation	Standard Deviation	Holdout	Cross-Validation	Standard Deviation
KNN	0.70	0.59	0.02	0.80	0.77	0.03
DT	0.65	0.63	0.03	0.79	0.80	0.03
RF	0.72	0.68	0.02	0.85	0.87	0.02
ADA	0.63	0.61	0.02	0.66	0.68	0.03
XGB	0.67	0.65	0.03	0.83	0.84	0.02
LDA	0.64	0.60	0.00	0.49	0.49	0.03
GNB	0.65	0.62	0.02	0.54	0.57	0.03
LR	0.63	0.60	0.00	0.48	0.50	0.02
MLPC	0.64	0.59	0.01	0.48	0.50	0.01
ETC	0.71	0.69	0.02	0.88	0.89	0.02

After optimization, it can be seen in Table 1 that the models who achieved the better accuracy were ETC, RF, XGB, and DT. These models have a better accuracy on balanced dataset compared to its unbalanced dataset. ETC before was 69% became 89%, RF from 68% became 87%, XGB from 65% to 84%, and 63% became 80% on DT.

TABLE 2. COMPARISON OF ACCURACY OF OPTIMIZED-FOUR-FEATURE MODEL

Optimized (4 Features)						
	Unbalanced			Balanced		
Model	Holdout	Cross-Validation	Standard Deviation	Holdout	Cross-Validation	Standard Deviation
KNN	0.69	0.60	0.02	0.79	0.79	0.02
DT	0.64	0.63	0.04	0.76	0.80	0.03
RF	0.71	0.69	0.03	0.84	0.85	0.03
ADA	0.63	0.62	0.02	0.64	0.65	0.02
XGB	0.67	0.66	0.03	0.81	0.81	0.03
LDA	0.63	0.60	0.00	0.52	0.50	0.03
GNB	0.65	0.63	0.02	0.55	0.58	0.02
LR	0.63	0.60	0.00	0.52	0.49	0.03
MLPC	0.69	0.62	0.03	0.54	0.56	0.03
ETC	0.73	0.69	0.02	0.85	0.87	0.02

As per the findings presented in Table 2, the most precise model is the balanced ETC model with an accuracy of 87%, followed by the RF model with an accuracy of 85%. The XGB and DT models trail behind with accuracies of 81% and 80%, respectively. The analysis also indicates that the all-feature model continues to perform better than the 4-feature model. However, it's worth noting that the 4-feature model performs reasonably well, with only a 2% difference in accuracy from the all-feature model.

TABLE 3. COMPARISON OF ACCURACY OF OPTIMIZED-FIVE-FEATURE MODEL

Optimized (5 Features)						
	Unbalanced			Balanced		
Model	Holdout	Cross-Validation	Standard Deviation	Holdout	Cross-Validation	Standard Deviation
KNN	0.73	0.59	0.02	0.81	0.77	0.03
DT	0.66	0.62	0.04	0.77	0.80	0.02
RF	0.72	0.70	0.02	0.85	0.86	0.02
ADA	0.62	0.59	0.03	0.68	0.65	0.03
XGB	0.67	0.67	0.03	0.82	0.83	0.03
LDA	0.64	0.60	0.00	0.48	0.50	0.02
GNB	0.65	0.63	0.02	0.54	0.58	0.03
LR	0.63	0.60	0.00	0.46	0.51	0.03
MLPC	0.68	0.61	0.01	0.48	0.51	0.02
ETC	0.72	0.70	0.03	0.86	0.89	0.01

Table 3 shows the performance of the optimized five-feature model with a balanced dataset is better than on an unbalanced dataset, and it achieves accuracy levels that are comparable to those of the optimized model that utilizes all features. Among the models, ETC produced the highest accuracy of 89%, followed by RF with an accuracy of 86%. XGB achieved an accuracy level of 83%, while DT achieved an accuracy of 80%.

Based on the information presented in the tables above, it can be determined that balancing the dataset can improve the accuracy of a water potability classification model. Moreover, a model that employs only 5 features can perform comparably to the models that utilize all available features. The utilization of a model that employs only five features can potentially reduce the likelihood of overfitting the data. Thus, to make a valid comparison with other related works, the most suitable model would be the balanced-five-feature model.

TABLE 4. COMPARISON OF PERFORMANCE METRICS ON DIFFERENT MODELS

MODELS	ACCURACY	PRECISION	RECALL	F1-SCORE
KNN	77%	73%	85%	78%
DT	80%	76%	89%	82%
RF	86%	84%	88%	86%
ADA	65%	65%	65%	65%
XGB	83%	81%	88%	84%
LDA	50%	50%	49%	49%
GNB	58%	60%	45%	51%
LR	51%	51%	47%	49%
MLPC	51%	54%	47%	45%
ETC	89%	91%	86%	88%

The findings in Table 4 indicate that Extra Trees Classifier outperformed other models with an accuracy rating of 89%, Precision ratings of 91%, and F1-Score of 88%. On the other hand, Decision Tree had the highest Recall score of 89%. Meanwhile, Linear Discriminant Analysis got the lowest scores.

The findings indicate that balancing the dataset will greatly help the model the achieve high accuracy. In addition, the accuracy of the model with five selected features, namely pH, Chloramines, Hardness, Sulfate and Solids, is superior to that of the model with all features and four features. Consequently, these five features will serve as the default features for various machine learning algorithms. Furthermore, the researchers are continually optimizing the machine learning algorithm through GridSearch and RandomSearch techniques. After performing the optimization process, there was a significant increase in accuracy for different machine learning algorithms, demonstrating good accuracy ranging from 51% to 89%.

The result outperformed other related studies. Patel et al. [4] utilized five machine learning algorithms including DT, RF, XGBoost, Ada Boost, and SVC to classify water potability and achieved an accuracy of 81%. Poudel, D. et al. [15] used only four algorithms namely ANN, KNN, Logistic Regression, and RF, and obtained an accuracy of 70.42% using RF. In contrast, Nataraj, R. [20] had the lowest accuracy (66%) among the five related works as they used only KNN without applying hyperparameter tuning, manually coded five iterations of different parameters to optimize their previous model. Kaddoura, S. [19] and Nachaoui, M. et al. [18] obtained different results. [19] used SVC and obtained an accuracy of 73.1% with a low precision but high recall, meaning that they usually predict the water as potable even if it is not. On the other hand, [18]

developed their own approach and achieved an accuracy of 67.37% with high precision but low recall, implying that they frequently classify water as unfit to drink even when it is.

V. Conclusions

This study aims to assess and monitor water quality parameters that impact water potability using machine learning methods. The results demonstrate that machine learning classification models can accurately interpret and predict water quality, thus contributing to better public health outcomes and improved management of water resources.

Based on the results presented in the previous tables, it can be inferred that balancing the dataset can significantly enhance the accuracy of a water potability classification model. In addition, a model that employs only five features can perform similarly to models that utilize all available features, which minimizes the risk of overfitting. Among all the models that were developed in this study, the ExtraTree Classifier was found to be the most effective, followed by the Random Forest.

Moreover, the study's results were compared with those of previous research, revealing that the accuracy achieved in this study surpassed those of previous studies. This is mainly due to the fact that the study employed balanced and unbalanced datasets, utilized broad machine learning models with different features, and evaluated their performance. Overall, this study successfully developed a machine learning model that can accurately classify water potability, with the assistance of the ExtraTree Classifier.

VI. Acknowledgment

The authors would like to acknowledge the support and resources provided by Polytechnic University of the Philippines (PUP). The facilities, funding, and access to relevant literature and research materials have greatly contributed to the success of this project.

References

[1] Wu, H., Gai, Z., Guo, Y., Li, Y., Hao, Y., and Lu, Z. N. (2020). Does Environmental Pollution Inhibit Urbanization in China? A New Perspective through Residents' Medical and Health Costs. Environ. Res. 182 (Mar.), 109128–109128.9. doi:10.1016/j.envres.2020.109128

[2] Lin, L., Yang, H.; Xu, X. (2022, June). Effects of water pollution on human health and disease heterogeneity: A Review. Frontiers. Retrieved from https://doi.org/10.3389/fenvs.2022.880246

[3] Chowdhary, P., Bharagava, R. N., Mishra, S., and Khan, N. (2020). Role of Industries in Water Scarcity and its Adverse Effects on Environment and Human Health. Environ. Concerns Sustain. Dev., 235–256. doi:10.1007/978-981-13-5889-0_12

[4] Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S., Ratna, R. (2022, September). A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. Computational Intelligence and Neuroscience. Retrieved from https://doi.org/10.1155/2022/9283293

[5] Luby, S. P., Davis, J., Brown, R. R., Gorelick, S. M., & Wong, T. H. F. (2019, August 2). Broad approaches to cholera control in Asia: Water, sanitation and handwashing. Vaccine. Retrieved January 20, 2023,https://www.sciencedirect.com/science/article/pii/S0264410X19310059

[6] Stübner, C., Miniscalco, C., Nielsen, C., Gillberg, C., Jakobsson, K., & Ebel, M. (n.d.). Developmental language disorders in preschool children after high exposure to perfluoroalkyl infstances from contaminated drinking water in Ronneby, Sweden. LWW. Retrieved January 20, 2023, https://journals.lww.com/environepidem/Fulltext/2023/02000/Developmental_language_disorders_in_preschool.6.aspx

[7] United Nations Educational, Scientific and Cultural Organization. (2021). United Nations World Water Development Report 2021: Valuing Water. Retrieved from: https://unesdoc.unesco.org/ark:/48223/pf0000375724_eng

[8] Ren, T., Yuyan, J., Huan, L., Yingxin, P., Dongmei, T., Qiumei, D., & Zhong, Y. (2022, November 8). Investigation on an outbreak of bacillary dysentery due to infection of Shigella Sonnei in a town of Guangxi Province. Dialogues in Health. Retrieved January 20, 2023, fromhttps://www.sciencedirect.com/science/article/pii/S2772653322000727

[9] Dawood, T., Elwakil, E., Novoa, H. M., & Delgado, J. F. G. (2020, November). Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks. Journal of Cleaner Production. Retrieved from 10.1016/j.jclepro.2020.125266

[10] DataCamp. (2022, February 17). Classification in Machine Learning: What Is It and How Does It Work? [Blog post]. Retrieved from https://www.datacamp.com/blog/classification-machine-learning

[11] Dalal, S., Onyema, E. M., Romero, C. A. T., Ndufeiya-Kumasi, L. C., Maryann, D. C., Nnedimkpa, A. J., Bhatia, T. K. (2022, January). Machine learning-based forecasting of potability of drinking water through adaptive boosting model. De Gruyter. Retrieved from https://doi.org/10.1515/chem-2022-0187

[12] Abuzir, S. Y., & Abuzir, Y. S. (2022, August). Machine Learning for Water Quality Classification. Water Quality Research Journal. Retrieved from https://doi.org/10.2166/wqrj.2022.004

[13] Kuthe, A., Bhake, C., Bhoyar, V., Yenurkar, A., Khandekar, V., & Gawale, K. (2022). Water quality analysis using machine learning. International Journal for Research in Applied Science and Engineering Technology, 10(12), 581–585. Retrieved from https://doi.org/10.22214/ijraset.2022.47944

[14] Abraham, A., Livingston, D., Guerra, I., & Yang, J. (2022, September). Exploring the Application of Machine Learning Algorithms to Water Quality Analysis. IEEE Explore. Retrieved from https://doi.org/10.1109/BCD54882.2022.9900636

[15] Kurra, S. S., Naidu, S. G., Chowdala, S., Yellanki, S. C., Sunanda, E. (2022, May). Water quality prediction using machine learning - IRJMETS. International Research Journal of Modernization in Engineering Technology and Science. Retrieved from https://www.irjmets.com/uploadedfiles/paper/issue_5_may_2022/22391/final/fin_irjmets1651989957.pdf

[16] Anbuchezhian, M., Venkataraman, R., Kumuthavalli, V., & Phil, M. (2018, November). Water Quality Analysis and Prediction using Machine Learning Algorithms. Journal of Emerging Technologies and Innovative Research. Retrieved from https://www.jetir.org/papers/JETIR1811966.pdf

[17] Poudel, D., Shrestha, D., Bhattara, S., Ghimire, A. (2022, February). Comparison of Machine Learning Algorithms in Statistically Imputed Water Potability Dataset. ResearchGate. Retrieved from https://doi.org/10.13140/RG.2.2.25767.21925

[18] Nachaoui, M., Lyaqini, S., & Chaouch, M. (2023). Indicating if water is safe for human consumption using an enhanced machine learning approach. Statistics, Optimization & Information Computing, 11(1), 70-81. Retrieved from https://doi.org/10.19139/soic-2310-5070-1703

[19] Kaddoura, S. (2022). Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. Sustainability, 14(18), 11478. Retrieved from https://doi.org/10.3390/su141811478

[20] Nataraj, R. (2021, December). Application of Machine Learning to Predict Water Potability. View of application of machine learning to predict water potability. Retrieved from http://ijmaa.in/index.php/ijmaa/article/view/29

[21] Pal, O. K. (2022, June). The Quality of Drinkable Water using Machine Learning Techniques. Ijaers.com. Retrieved from https://doi.org/10.22161/ijaers.96.2

[22] S. Ray, "A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.