VOL. 20, NO. 20, OCTOBER 2025 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ENHANCING WATER POTABILITY PREDICTION: TUNED VS DEFAULT CAT BOOST MODELS

Arun Balaji S.[1,2], Subramoniam M.[1], Poornapushpakala S.[1] and Barani S.[1]
[1]Sathyabama Institute of Science and Technology, Chennai, India
[1]Faculty of Electrical and Electronics Engineering, Sathyabama Institute of Science and Technology, Chennai, India
[2]College of Engineering and Technology, University of Technology and Applied Sciences, Salalah, Oman
E-Mail: Arunbalaji.S@utas.edu.om

ABSTRACT

Water is a very important source for every living being on the earth. In such a scenario, access to clean and hygienic potable water remains a challenge in many parts of the world. Conventional water quality testing methods are expensive, and they need expertise. This challenge limits the scalability to test the water quality in environmental monitoring applications. To meet this challenge, a new attempt has been made by integrating the machine learning methods with chemical parameters to estimate the suitability of water for drinking purposes. The General and Tuned Cat Boost technique has been applied to the dataset, which contains the various physicochemical indicators. Various statistical parameters, such as accuracy, precision, recall, and F1-score, but also interpret their behaviour via confusion matrices, ROC curves, precision-recall plots, residual graphs, and feature importance scores, were evaluated for both models. On analysing these parameters, it was observed that the tuned model showed only marginal improvement over the general Cat Boost model. This paper concludes the observations of this study in a comprehensive framework that integrates performance, explainability, and model effectiveness for AI-empowered environmental intelligence systems.

Keywords: cat boost models, SMOTE, tuned cat boost model.

Manuscript Received 16 September 2025; Revised 5 January 2026; Published 20 January 2026

1. INTRODUCTION

Clean water is a fundamental requirement for all human beings in the world. Though there are rapid advancements in technology in every part of the world, access to safe drinking water is still lagging. Few studies showed that on average of 2 billion people are affected in a year with diarrhoea, dysentery, and cholera due to consuming contaminated water. The traditional methods involve analysing the physicochemical properties to estimate the contaminants in water and predicting its suitability for drinking. This method is time time-consuming process and also needs expertise. In this scenario, a machine learning model will support faster prediction of the Potability of drinking water using the sensor data. Though there are several machine learning algorithms in dominance in predicting data, the CatBoot algorithm developed by Yandex has shown outstanding performance with categorical and numerical datasets. It is also robust, thereby reducing the problem of overfitting. In this study, the performance of general and tuned Cat Boost methods was compared on a publicly available dataset predict the suitability of water for drinking purposes using the data of chemical concentrations in the water. The performance of both models was analysed visually and statistically. A similar kind of study has also been done by the researchers to predict the potability of water with other methods. The same has been discussed in the following sections.

2. LITERATURE SURVEY

Several methods have been developed by researchers to predict the quality of drinking water using machine learning algorithms. The recent advances in this topic of study have been discussed below.

Melchizedek I. Alipio(4) developed an IoT-based water quality monitoring system, which indicates the Potability of water in rural areas. In the study performed by the author, the data used are the physicochemical properties of water, which were collected using a portable sensor deployed at various sources. Ensemble learning methods were used to check the potability of water from the collected data. The advantage of this system is that it can communicate the water potability with minimum delay.

Jovita Biju *et al*. (12) performed a study on the potability of water using traditional and hybrid machine learning models. The values of parameters such as sulphate concentration, pH level of the water, Chloramines concentration, organic carbon content, water hardness, total dissolved solids in the water, electrical conductivity, Trihalomethanes concentration, and turbidity level are the features fed as input to the machine learning models. The statistical analysis done with both models showed that the hybrid model, which achieved an accuracy of 68 percent, performed better than the traditional machine learning model in predicting the water potability.

Husain Yusuf et al assessed the performance of various machine learning models in the classification of water quality using the physicochemical parameters. The study concluded that the Logistic regression model was

poor in predicting the water potability, as its accuracy achieved was only 47 percent. On the other hand, random forest achieved the highest accuracy of 83 percent in predicting the water potability. The same has also been confirmed in the study done by Rahan Manoj *et al*.(11). The Light model has attained 99.74% to the random forest model in the study done by Reem Alnaqeb *et al*. (9) Gradient Boosting Model performed better in the study done by Vishnu Sreekumar *et al*.(14) Similar study was done by Priya Thomas *et al*. (8). Logistic Regression, Random Forest, Decision Tree, KNN, and SVM on the dataset collected from the Central Pollution Control Board (CPCB) of India. The SVM model has produced the highest accuracy of 83 percent in their study. A study done with ensemble methods with all these classifiers and neural networks has achieved a much higher accuracy of 96.7 % in detecting water potability in the study done by Jigna K. Pandya (3)

Akilandeswari *et al*. (1) compared the performance of random forest and Boost for water potability using an open-source dataset. In this, the XG Boost achieved a higher accuracy of 87.22% than the random forest, which achieved an accuracy of 61.11%. A similar kind of study was done by Mehul Patel. (6) F1-score was the metric used by the author to evaluate the performance of various models. On examining various models, the XG boost outperforms better with an F1-score of 0.9798 as compared to other models.

Jefferson Johan (2) compared the performance of default and tuned classification results of XGBoost, KNN, Random Forest, SVC, and ADABoost models with the open-source water potability dataset. The study concluded that tuned SVC performed better among all these models, with a precision of 77.32, than the general model, which produced a precision of 69.66.

Rohan Manoj (11) concludes that Random Forest (accuracy 79.60%) in Machine Learning methods performs better than deep learning methods (accuracy 70.94%) in predicting the water potability. Statistical methods were used by Varun Mishra to analyse the factor that makes water potable. The results concluded that the feature attributes that decide the water potability should meet a specific standard. The tuned SVM model produced the highest classification rate among various other machine learning algorithms in the study done by Roy Hendro Siburian (10) on water samples, whereas the ARIMA model showed better results in another study done by Selvaraj.

Most of the studies were performed by developing a generalised learning model using the features of the collected water samples. The developed model was then evaluated using the statistical parameters. The study on developing a robust model in predicting the water potability is still lagging. Hence, an attempt is made to fulfil the gaps in such research. This study is done to evaluate the robustness of the Cat Boost model in identifying the water potability. The methodology adopted for the study is given in the following section.

3. METHODOLOGY

This section describes the experimental setup followed for this study. The topics such as data acquisition, preprocessing of data for class imbalance, implementation of general and tuned Cat Boost models, and the process of analysis were discussed. The various steps adopted in this study have been depicted as a block diagram, which is given in Figure-1.

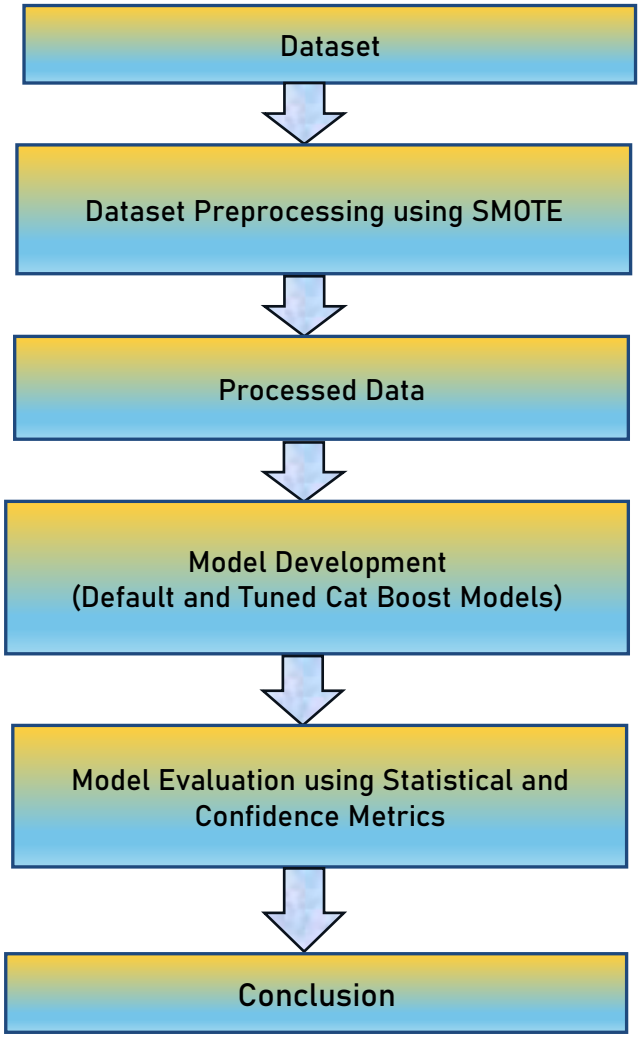

Figure-1. Block diagram of the proposed study.

3.1 Dataset Description

The dataset used for this study was taken from a publicly available dataset, which has around 3,276 observations with water samples characterized by physicochemical properties (15). The various features available in the dataset are pH, hardness, total dissolved solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. Potability is the target variable fixed to conclude the results from the observations of physicochemical properties of water. This target variable is given as binary to indicate whether the sample is safe to drink (1) or not suitable to drink (0).

3.2 Data Preprocessing

Missing values or improper values in a dataset may result in misinterpretation by the classifier. To avoid this, the missing value has to be eliminated or has to be filled with appropriate values. This practice is very common before deploying a dataset into a Machine learning model. This dataset also contains several missing values, particularly in pH, Sulphate, and Trihalomethanes. These missing values are filled with median values. The reason for choosing median values over normalization or standardization is that they are less sensitive to outliers. This also conserves the central tendency, original scale, and relationships of features.

3.3 Handling Class Imbalance with SMOTE

The dataset contains a total of 3276 observations obtained from the collected water samples. Among these, 1998 samples belong to the non-potable target class and the rest 1278 belong to the potable target class. This imbalance can cause biased prediction by the model and can also result in poor accuracy. There are several data imbalance methods available, and SMOTE (Synthetic Minority Over-sampling Technique) is one among them. SMOTE has proven promising results when dealing with binary classification. This algorithm creates synthetic data points for the minority class. This is done by interpolating between existing examples of the minority class in feature space. By adding new synthetic points to the existing dataset, SMOTE aids in class distribution. Thus, the final dataset used for training contains 1998 samples from each target class. The graph given in Figure-2 shows the class distribution before and after applying SMOTE on the raw data set.

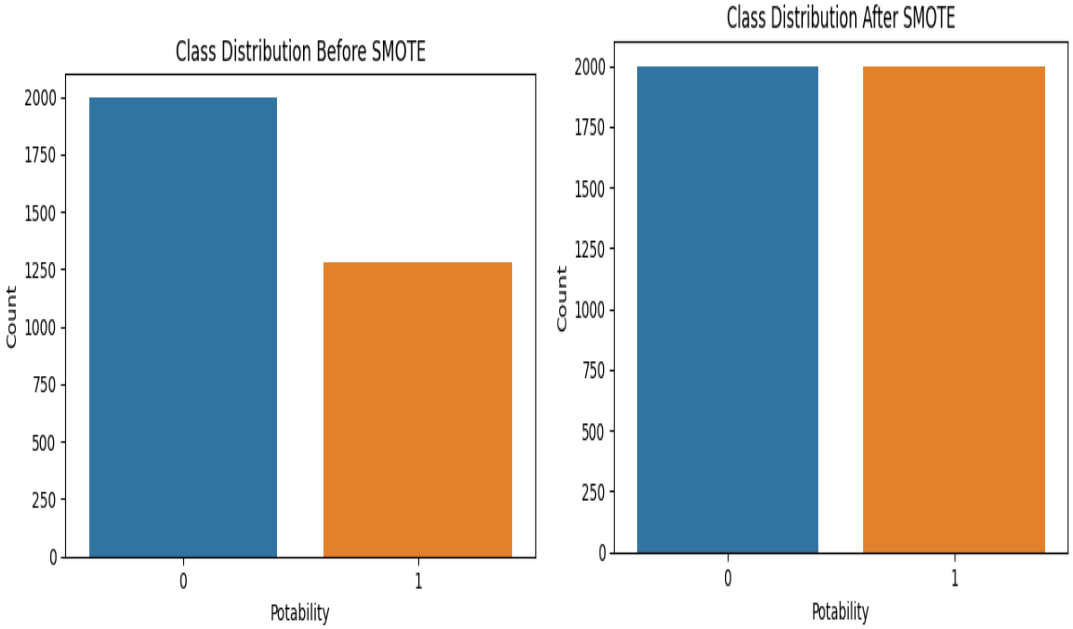

Figure-2. Comparison chart class distribution with and without SMOTE.

3.4 Model Training and Development

The dataset whose classes have been balanced using the SMOTE method is used to develop the model. The Cat Boost model was planned in this phase of the study because this model does not need any kind of categorical encoding. Even though the missing values in this dataset have been filled with median values, the added advantage of the Cat Boost model is its capability to handle the missing values internally. A description of architecture of Cat Boost is given in section 3.4.1.

3.4.1 Cat boost architecture

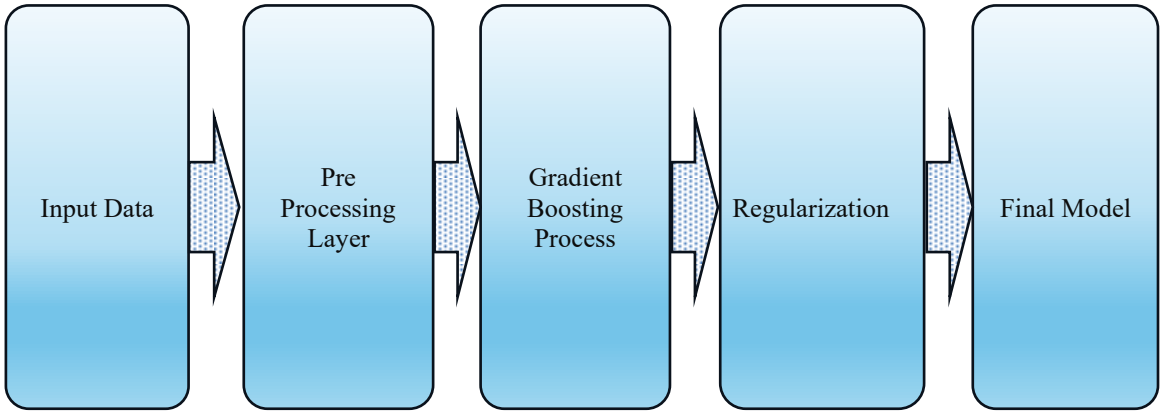

Figure-3. Cat Boost Model Architecture.

The architecture of Cat Boost or Categorical Boosting is shown in Figure-3. This is a kind of gradient boosting algorithm, which was developed by Yandex (16). This Cat Boost uses a unique boosting technique called ordered boosting to minimise the target leakage and prediction shift. This is done by building a series of models on training data using different permutations. This will ensure that the model's approximation for any of the samples will not rely on its target value. All these processes were done in the preprocessing layer. In the next stage, ordered boosting, which is a type of gradient boosting, is done. The ordered boosting is given by (Equation). This ordered boosting avoids the usage of future information and largely reduces overfitting, especially when the size of the datasets is small. Similar to one-hot encoding, Cat Boost uses a technique called mean encoding or ordered target statistics to convert each categorical value. It reduces the regularized loss function, which is in the form (Equation). Similarly, for binary classification, it uses log loss, which is given by (Equation). It uses the statistics given in the equation to convert each categorical value from the previously seen rows. As only the current target is used to calculate the mean sample, this method avoids overfitting and also ensures robust encoding even for intermittent categories. Through the above-discussed steps, the model resembles an ensemble of decision trees in which the error that occurred in a tree will be reduced in the subsequent tree. Two kinds of Cat Boost algorithm were used in this study, viz general Cat Boost model and other one is Tuned Cat Boost model.

The general Cat Boost Model uses the default hyperparameter settings. The different hyperparameter settings are iterations, learning rate, depth, and regularization. The iteration value decides the number of decision trees to be built sequentially. A larger iteration value may lead to overfitting if proper regularization is not done. Learning rate is a scaling factor that minimizes the influence of each new tree before adding it to the model. This value can be used in the range between 0.001-0.1, but it will be good to use low values with higher iterations to build robust models. Depth is the factor that decides the maximum depth of each tree in the ensemble model. The range of depth is usually between 4-10. If high depth values are used, the model can build complex relationships between the parameters. It can also lead to overfitting. On the other hand, lesser values will build shallow trees, which may lead to underfitting. L2 regularization is the term that is added to the objective function to avoid overfitting. It helps to make the model more robust to the noise present in the data, thereby aiding in smooth prediction by the model. The general range for this parameter lies between 1-10. Higher values are generally used for noisy data, of overfitting is expected in the model. The default values of iterations, learning rate, depth, and L2regulaization in the general Cat Boost model are 1000, 0.03, 6, and 3, respectively.

3.4.2 Tuned cat boost model

This study is done by adjusting the hyperparameters of the Cat Boost algorithm discussed in the previous section and thereby evaluating the model performance. The hyperparameters are the one that makes an impact on the performance of a machine learning model. It balances between underfitting and overfitting by controlling the learning dynamics. Hence, this analysis is important. In this study, a four-range from the low to high in each hyperparameter was chosen.

Grid search method was used to identify the optimal combination of hyperparameters from 4x4x4x4=256 combinations of these parameters. This method creates a Cartesian product for the given parameter combination. Then, for every combination, the model is trained with the training data. The holdout validation technique was then applied, and the scoring metrics, such as accuracy, F1-score, ROC, and AUC, were computed for each combination. The F1 score decides the best combination of hyperparameters. With this method, the best combination of hyperparameters identified from the accuracy and F1 score was 'Iterations': 300, 'Learning rate': 0.1, 'Depth': 8, 'L2_leaf_reg': 1.

4. RESULTS AND DISCUSSIONS

To evaluate the performance of the developed model hold-out evaluation model was used. This method was chosen because of its simplicity and its unbiased

evaluation. The entire data set was split into two, with 80 percent of the data used for training and 20 percent of the data used for testing. The performance metrics, such as Accuracy, Precision, Recall, and F1-Score, which are derived from the confusion matrix given in Figure-4, and Confidence-based metrics such as Receiver Operating Characteristic - Area Under Curve (ROC AUC) and Precision-Recall Area Under Curve (PR AUC). The ROC AUC is best when the data is balanced, and the PR AUC will be better to analyse when the dataset is imbalanced. In order to have a comprehensive evaluation, both these methods are used for analysis. A summary of these evaluation metrics is given below.

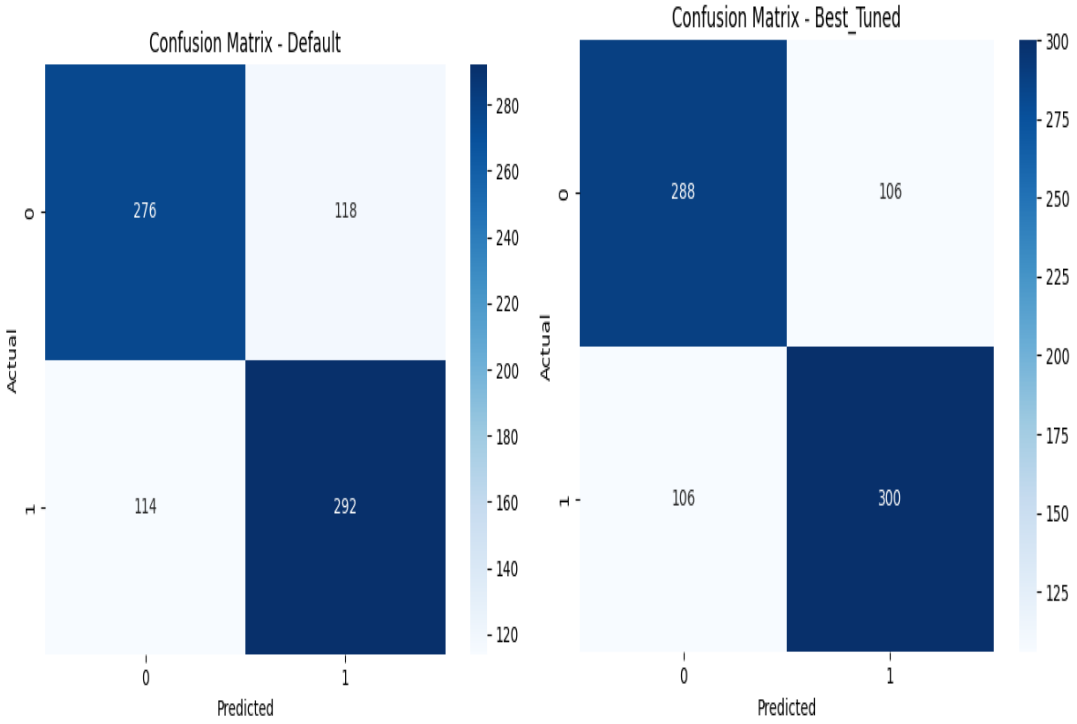

Figure-4. Confusion Matrix Obtained for General and Best Tuned Cat Boost Models.

Accuracy

It is the ratio between the number of correctly predicted observations to the total number of observations in the dataset. The formula to evaluate the accuracy using the confusion matrix
Is given by

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where,
TP - True Positive
TN - True Negative
FP - False Positive
FN - False Negative

Precision

It is defined as the ratio of correctly predicted positive class to the total number of positives in the dataset. Precision is also called positive predictive value. The formula to evaluate precision from the confusion matrix is

Precision = TP / (TP + FP)

Recall

It is defined as the ratio of correctly predicted positive class to the actual positive class in the dataset. Recall is also called as sensitivity or true positive rate. The formula to calculate recall from the confusion matrix is given by

Recall = TP / (TP + FN)

F1-Score

F1-Score balances between the precision and recall by taking the harmonic mean between these two metrics. It is much effective for averaging ratios as this harmonic mean is not much influenced by large outliers and is much sensitive to small values. The formula to evaluate the F1-score from the confusion matrix is given by

F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

The comparison plots obtained for these evaluation metrics with the general Cat Boost and the best-tuned hyperparameters Cat Boost model are given in Figure-5.

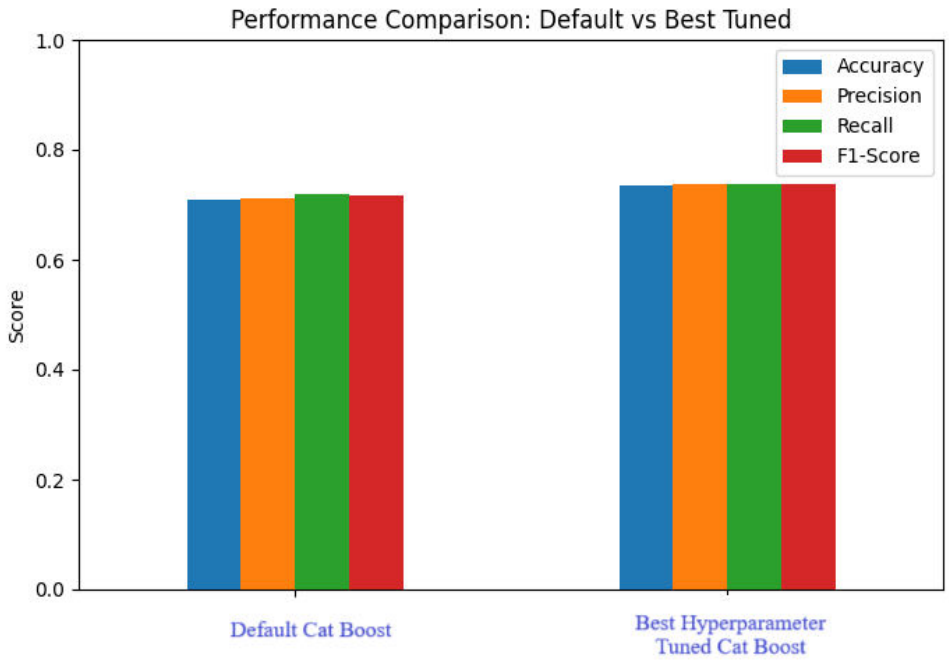

Figure-5. Comparison Chart of Evaluation Metrics Obtained with Default and Tuned Cat Boost Models.

The above figure shows the classification performance between the general Cat Boost and Tuned Cat Boost Models. The evaluation of these two models is done using metrics such as Accuracy, Precision, Recall, and F1-Score. From the chart, it can be observed that the tuned Cat Boost model (Accuracy:0.735, Precision:0.738, Recall: 0.738, and F1-Score:0.738) performs better than General or Default Cat Boost models (Accuracy:0.71, Precision:0.712, Recall: 0.719, and F1-Score:0.715). Even though the tuned Cat Boost model has performed well, there is only a slight difference in the evaluation metric values from the default Cat Boost model. This shows that the general Cat Boost model is robust, and careful tuning has to be done for further improvements. The advantages, such as Improvement in Accuracy, Increase in Recall, and improved F1-score, are obtained with hyperparameter tuning.

4.1 Evaluation with Confidence-Based Metrics

In binary classification problems, especially on working with imbalanced datasets, evaluating models with general evaluation metrics such as Accuracy, Precision, Recall, and F1-Score will not provide the complete information. In such scenarios, confidence-based evaluation metrics such as Receiver Operating Characteristic - Area under Curve (ROC AUC) and Precision-Recall Area under Curve (PR AUC) will provide more precise information about model performance across varying decision thresholds.

4.1.1 Receiver operating characteristic - area under curve (ROC AUC)

ROC is the graph plotted between the True Positive Rate and the False Positive Rate for various classification thresholds. The area under the ROC will help the model to quantify between the positive and negative classes. The relation between the Area under the Curve and the classifier performance is given below.

- AUC = 1.0: Perfect classifier
- AUC = 0.5: No discriminative ability (random guessing)
- AUC < 0.5: Less than random (Not a good classifier model)

The ROC-AUC graphs obtained for the general and tuned Cat Boost model is given in Figure-6.

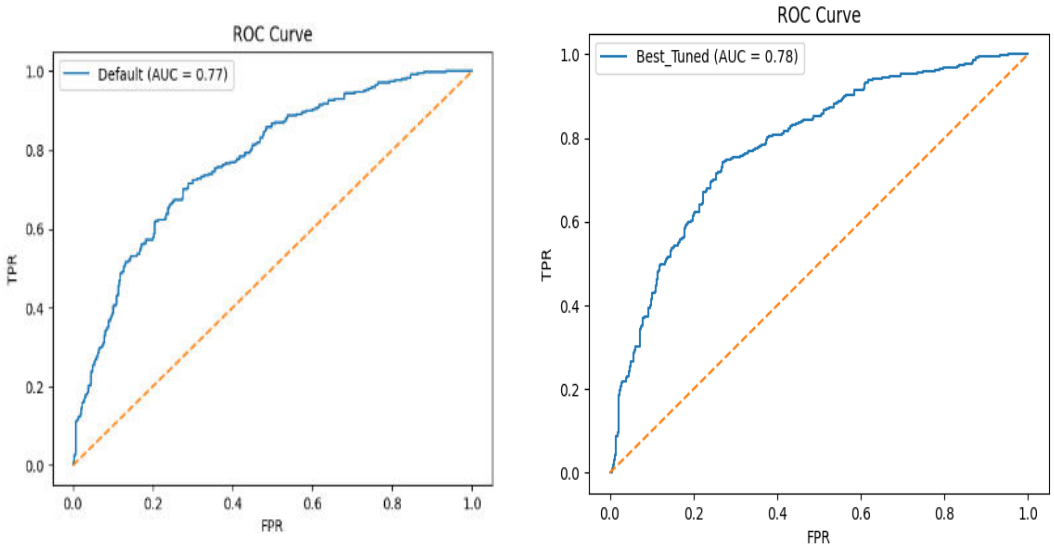

Figure-6. ROC-AUC plot for Default and Tuned Cat Boost Models.

In the above ROC-AUC plot, for both models, the curve bends pointedly towards the upper-left corner, which indicates that the model has strong discriminative ability. The diagonal dashed line in the graph represents random guessing (AUC = 0.5), and the ROC curve lies above this line. This indicates that the model is better. The AUC of 0.77 for default Cat Boost and the AUC of 0.78 for tuned Cat Boost indicate strong performance of both models in distinguishing potable and non-potable water samples.

The marginal increase in AUC value of tuned Cat Boost from the general Cat Boost indicates the robustness of the general Cat Boost model. This robustness is the key factor in deploying a model for risk-sensitive applications like this water quality assessment.

4.1.2 Precision-recall area under curve (PR AUC)

The PR-AUC is the graph plotted between Precision and Recall for various classification thresholds. This graph will help to conclude the model's ability to identify the positive instances while preserving low false positives. This analysis is very valuable while handling imbalanced datasets. The PR-AUC graphs obtained for the general and tuned Cat Boost model are given in Figure-7.

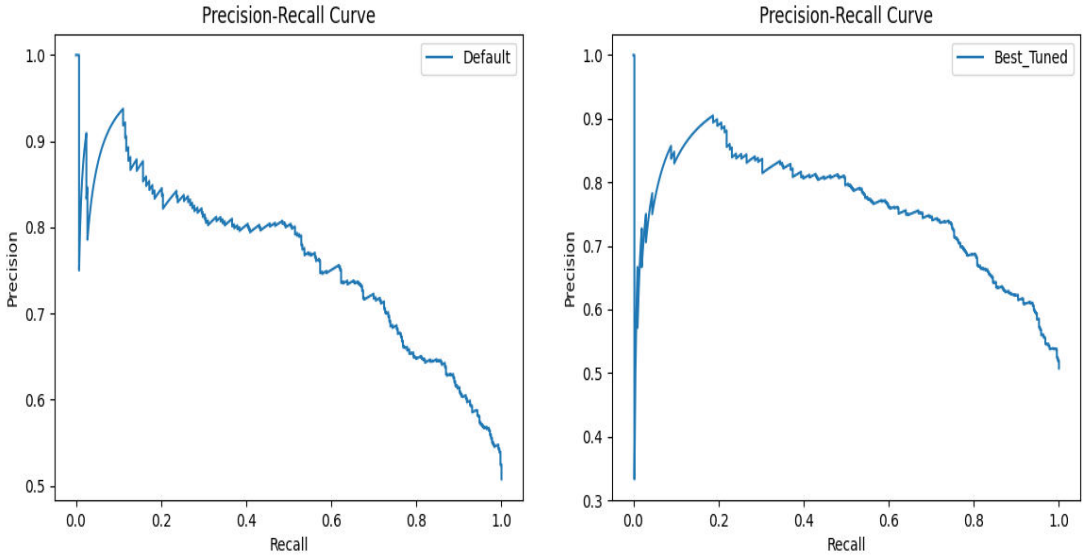

Figure-7. PR-AUC plot for Default and Tuned Cat Boost Models.

In the above PR-AUC plot, the PR curve for the tuned model establishes higher overall precision over a broad range of recall values, mostly in the low to mid recall region (recall < 0.5). Precision starts near 1.0 for very low recall, peaks around recall 0.2-0.3, and then gradually decreases as recall increases. This indicates that the tuned model can preserve higher precision while retrieving more relevant instances than the default model.

This indicates that the tuned model has improved performance and better discrimination between classes compared to the general Cat Boost configuration. This graph also shows the effectiveness of model tuning to achieve a better balance between precision and recall.

5. CONCLUSIONS

The objective of the research is to develop and evaluate the performance of a machine learning model that accurately predicts the potability of water using the physicochemical features. As a part of the research Cat Boost model was attempted for the study. Initially SMOTE technique was applied to the dataset to balance the class imbalance. The balanced data set was used to develop general and tuned Cat Boost models. For the tuned Cat Boost model, 256 combinations of parameters were applied using the grid search method, and the best combination was chosen using the F1-score. The statistical metrics and confidence metrics were used to evaluate the performance of both models. From the observations of the metrics, it was concluded that the tuned model shows improved performance over the general model. The general model also proves its robustness with the values obtained for the evaluation metrics. This property showed that the Cat Boost algorithm can be deployed with confidence in complex decision-making problems. The methodology and evaluation framework followed here can serve as a pattern for future research in the applications of environmental monitoring and intelligent decision systems.

REFERENCES

[1] A. Akilandeswari, V. Bura, P. Sadagopan, and T. Tirumalaikumari. 2023. Comparative Analysis of XGBoost and Random Forest for Predicting Water Potability. in Proc. 2023 2nd Int. Conf. Automation, Computing and Renewable Systems (ICACRS), doi: 10.1109/ICACRS58579.2023.10405315.

[2] J. Johan, J. Orlanda, W. S. Pangestu, W. T. Priyanto, M. Meiliana, and S. Achmad. 2023. Application of Data Mining Techniques and Hyperparameter Tuning for Accurate Water Potability Classification. in Proc. 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), Jakarta, Indonesia, doi: 10.1109/ICAICTA59291.2023.10390049.

[3] J. K. Pandya, S. S. Khandelwal, R. K. Tipu, and K. S. Pandya. 2025. Advancing Water Quality Management: An Integrated Approach Using Ensemble Machine Learning and Real-Time Interactive Visualization. IEEE Access, 13: 92406-92428, 2025, doi: 10.1109/ACCESS.2025.3573589.

[4] M. Alipio. 2020. Data-driven IoT-based Water Quality Monitoring and Potability Classification System in Rural Areas. in Proc. 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, doi: 10.1109/ICTC49870.2020.9289505

[5] M. I. K. Haq, F. D. Ramadhan, F. Az-Zahra, L. Kurniawati, and A. Helen. 2021. Classification of Water Potability Using Machine Learning Algorithms. In Proc. 2021 International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA), pp. 1-5, doi: 10.1109/ICAIBDA53487.2021.9689727.

[6] M. S. Patel, A. Gupta, P. Kumar, and R. Sharma, 2024. A Detailed Analysis of Machine Learning Models to Predict Water Potability, in Proc. 2024 IEEE International Conference on Electronics, Communications and Information (ICEI), pp. 631-639, doi: 10.1109/ICEI63732.2024.10917239.

[7] P. Selvaraj and H. Shalma. 2024. ARIMA Modeling for Reliable Potable Water Identification and Quality Prediction. In Proc. 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), doi: 10.1109/ICOSEC61587.2024.10722505.

[8] P. Thomas. 2025. Ensuring Clean Water through Machine Learning Driven Water Quality Prediction. in Proc. 2025 IEEE International Conference on ICICCS, doi: 10.1109/ICICCS65191.2025.10985149.

[9] R. Alnaqeb, F. Alrashdi, K. Alketbi, and H. Ismail. 2022. Machine Learning-based Water Potability Prediction. In Proc. 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, UAE, doi: 10.1109/AICCSA56895.2022.10017579.

[10] R. H. Siburian, M. K. M. Nasution, and J. T. Tarigan. 2025. Optimization of KNN, SVM, and SVM Kernel in Water Potability Prediction with Hyperparameter Approach. in Proc. 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI), doi: 10.1109/ICoCSETI63724.2025.11019055.

[11] R. Manoj, S. Abhishek, B. Prathap Nair, A. T., and N. Prabhu Ramlal. 2023. A Contemporary Method of Assessing Water Quality based on the Fusion of

Predictive Analytics and Deep Structured Learning. in Proc. 2023 8th International Conference on Communication and Electronics Systems (ICCES), pp. 961-967, doi: 10.1109/ICCES57224.2023.10192729.

[12] R. Manoj, S. Abhishek, B. Prathap Nair, A. T. , and N. Prabhu Ramlal. 2023. A Contemporary Method of Assessing Water Quality based on the Fusion of Predictive Analytics and Deep Structured Learning. in Proc. 2023 8th International Conference on Communication and Electronics Systems (ICCES), pp. 961-967, doi: 10.1109/ICCES57224.2023.10192729.

[13] S. K. Biju, C. Badgujar, A. Poulose, and H. F. Thomas. 2024. Hybrid Horizons: Advancing Water Potability Prediction through Hybrid Machine Learning. in Proc. 2024 IEEE Int. Conf. on Ubiquitous and Future Networks (ICUFN), doi: 10.1109/ICUFN61752.2024.10625242.

[14] V. Mishra, S. Upadhyay, P. Jhaba, S. Prajapati, and K. K. Gola. 2023. An Examination of Guaranteeing the Safety of Drinking Water using Machine Learning. in Proc. 2023 IEEE International Conference on Self-Sustainable Artificial Intelligence Systems (ICSSAS), doi: 10.1109/ICSSAS57918.2023.10331671.

[15] V. Sreekumar, F. Ihsan, S. Reghuram, and S. Sarath. 2024. A Detailed Analysis of Machine Learning Models to Predict Water Potability. in Proc. 2024 15th International Conference on Computing, Communication and Networking Technologies (ICCCNT), doi: 10.1109/ICCCNT61001.2024.10725826.

[16] https://www.kaggle.com/code/nimapourmoradi/water-potability/input

[17] Prokhorenkova L, Gusev G, Vorobev A. 2018. Cat Boost: unbiased boosting with categorical features. Advances in neural information processing systems. 31.