XRF-TCG: WATER POTABILITY PREDICTION WITH XGBOOST, RANDOM FORESTS AND TEXT CONTENT GENERATION FOR AUTOMATED WATER QUALITY REPORTING

Abisheka PON [1*], Deisy C [2] , Sharmila P [3] and Eunice J [4]

[1] Department of Computer Science and Engineering

[1]Fatima Michael College of Engineering and Technology.

Madurai-625020. Tamilnadu. India

[2,3] Department of Information Technology,

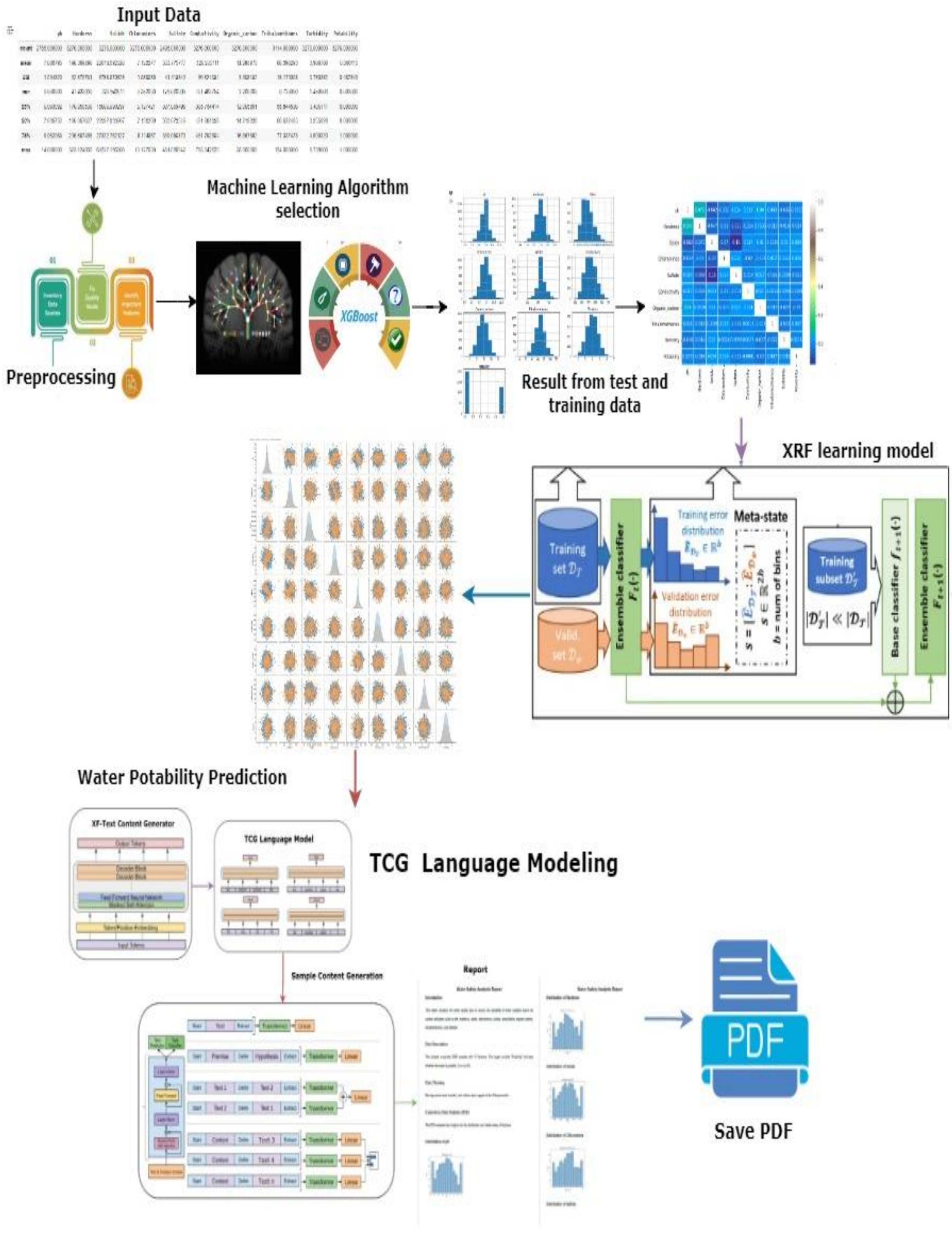
[4] Department of Civil Engineering,

[2,3,4]Thiagarajar College of Engineering.

Madurai-625015. Tamilnadu. India

[*]Corresponding author: Abisheka PON[1*]

Graphical Abstract


Input Data
Machine Learning Algorithm selection
Preprocessing
XGBoost
Result from test and training data
XRF learning model
Training set $\mathcal{D}_T$
Valid. set $\mathcal{D}_v$
Ensemble classifier $F_t(\cdot)$
Training error distribution
Validation error distribution
Meta-state
$s = [\hat{E}_{\mathcal{D}_T} : \hat{E}_{\mathcal{D}_v}]$
$s \in \mathbb{R}^{2b}$
b = num of bins
Training subset $\mathcal{D}'_T$
$|\mathcal{D}'_T| \ll |\mathcal{D}_T|$
Base classifier $f_{t+1}(\cdot)$
Ensemble classifier $F_{t+1}(\cdot)$
Water Potability Prediction
XF-Text Content Generator
TCG Language Model
TCG Language Modeling
Sample Content Generation
Report
Save PDF
PDF

ABSTRACT

Water potability is a critical public health issue, and timely reporting on water quality is essential for monitoring and ensuring safe drinking water. Existing water quality assessment methods often involve manual sampling and laboratory analysis, which can be time-consuming and resource-intensive. To overcome this, an automated real-time Ensemble model for water quality monitoring and report generation is proposed with machine learning and artificial intelligence. This methodology offers a robust solution for automated water quality monitoring and reporting, which is crucial for maintaining public health standards by producing detailed, shareable water potability reports. It also demonstrates the potential of combining transformer-based NLP and LLM text generation to automate technical report creation. The novelty proposed in this system integrates a comprehensive framework for generating water potability reports from numerical data by leveraging Large Language Models. Bayesian optimization is utilized to remove outliers in water potability data, comparing machine learning classifiers like Random Forest and XGBoost, and for hyper-parameter tuning to enhance performance. Further, Named Entity is used to generate coherent and contextually relevant textual content. Then the textual content is formatted into a readable document using Text Content Generator. This significantly reduces manual effort while enhancing report consistency and accuracy.

***Keywords**: Water Potability; Random Forest; XGBoost; Text Content Generator; TCG Language Model; Report Generation; Machine Learning.*

1. INTRODUCTION

Preserving drinking water's potability and safety is an important public health issue. Contaminated water can lead to severe health issues, making accurate and timely water quality monitoring essential (Khammar, H. et al. 2024). Manual sampling and laboratory analysis are common methods of assessing water quality, but they can be time-consuming and resource-intensive. With advancements in machine learning and artificial intelligence, there is a growing interest in developing automated systems for water quality monitoring and report generation. This study proposes an integrated framework that leverages Large Language Models (LLMs) and Transformer models for generating comprehensive water potability reports. The framework combines advanced data preprocessing techniques, multiple machine learning classifiers, and Bayesian optimization to ensure high accuracy and reliability in predicting water quality.

Recent research has demonstrated the possibilities of models for machine learning in enhancing water quality monitoring systems. For instance, [1] explored the use of Transformer models for Identifying irregularities in water quality data, achieving significant improvements in detection accuracy compared to traditional methods. Similarly, [8] integrated BERT and GPT-3 for comprehensive water quality reporting, highlighting the efficiency of combining language models with environmental data analysis.

Machine learning classifiers such as Random Forest, XGBoost, K-Nearest Neighbors (KNN), Recurrent Neural Networks (RNN), CatBoost, AdaBoost, Logistic Regression, and Decision Tree Classifier have been employed to predict various water quality parameters. [2] demonstrated the effectiveness of Transformer-based models in continuous tracking of water quality, illustrating the significance of prompt and precise forecasting in upholding water safety regulations. [3] highlighted the application of GPT-3 in generating environmental impact assessment reports, showcasing the capability of LLMs in producing high-quality, contextually relevant textual content. This study makes several significant contributions to the field of water quality monitoring and reporting:

- Integration of Advanced Models: By combining various machine learning classifiers and LLMs, we provide a comprehensive solution that leverages the strengths of different models for accurate water quality prediction and report generation.
- Enhanced Accuracy: The use of Bayesian Optimization for hyper-parameter tuning significantly enhances the accuracy of the predictive models.
- Automated Report Generation: The integration of GPT-3 for text generation and `fpdf` for PDF creation automates the report generation process, saving time and resources.
- Public Health Impact: This framework can contribute to the maintenance of public health standards and the safety of drinking water by providing a dependable and efficient system for water quality monitoring.

Access to potable water has numerous benefits! Some of the most significant advantages include, drinking adequate amounts of clean water offers numerous health advantages. It helps prevent waterborne diseases, boosts energy levels, improves skin health, supports kidney function, regulates body temperature, strengthens the immune system, aids in weight management, enhances mental clarity, supports athletic performance, and increases overall productivity. Essentially, staying hydrated is crucial for optimal physical and mental well-being.

The major contributions in predicting Water potability using Machine learning and generating report based on

- Integrates Machine Learning models for outlier prediction.

- Improves accuracy through Bayesian optimization and Ensemble.
- Automates report generation by NLP and LLM, saving time and resources.
- Reports significantly improve public health by ensuring safe drinking water.

The fundamental structure of this paper is comprised of the subsequent sections: Section 1 provides an overview of the contributions made by the paper. The research background conducted in the fields of machine learning algorithms and Text Content generation (TCG) with language processing and Report format using TCG-Language Model is listed in Section 2. The suggested methodology is described in Section 3. An experimental analysis of the suggested method is provided in Section 4. After that, Section 5 provides a thorough analysis of the results along with discussions based on the proposed model. The proposed TCG and Language Models for Water Potability Report Generation are explained in Section 6. Section 7 concludes the research and lays out the objectives for the future.

2. RELATED WORKS

The automation of data analysis and report generation tasks has significantly improved as a result of the rapid advances in machine learning and natural language processing (NLP). With an emphasis on water portability reporting specifically, this literature review discusses the latest advancements in transformer models, large language models (LLMs), and their applications in environmental monitoring. Since Transformers enable models to capture contextual relationships within text more effectively than earlier architectures, [20] have revolutionized natural language processing (NLP). One of the most well-known transformer models is BERT (Bidirectional Encoder Representations from Transformers) [19]. BERT has been applied extensively to tasks like question answering, sentiment analysis, and Named Entity Recognition (NER).Its bidirectional nature allows it to understand context from both directions, making it highly effective for extracting relevant information from text.

Recent studies have explored the application of BERT and other transformer models in environmental data analysis. For instance, [4] utilized a transformer-based model for water quality monitoring using remote sensing data. Their approach demonstrated the capability of transformers to handle large-scale environmental data and extract meaningful insights. Similarly, [17] applied BERT for water quality data analysis, highlighting its effectiveness in identifying key entities and patterns within large datasets. Large Language Models (LLMs), such as GPT-3 (Generative Pre-trained Transformer 3), have pushed the boundaries of text generation by producing human-like text based on a given prompt. GPT-3, developed by OpenAI [3] has shown remarkable capabilities in generating coherent and contextually relevant content across various domains.

The integration of LLMs in automated report generation has gained attention in recent years. For example, [18] enhanced environmental reports using GPT-3, demonstrating its ability to generate comprehensive and readable textual content from structured data. [20] Explored GPT-3 for generating environmental impact assessment reports, showing significant improvements in the quality and consistency of the generated documents. Environmental monitoring requires accurate and timely reporting to ensure public health and safety. Traditional methods often involve manual data processing and report generation, which can be time-consuming and prone to errors. The application of NLP techniques, particularly transformer models and LLMs, offers a promising solution to these challenges. [7] reviewed the applications of deep learning models in environmental monitoring, emphasizing the potential of transformer models to improve data analysis and reporting accuracy. [14] discussed the use of transformers for environmental data fusion and analysis, highlighting their ability to integrate and analyze diverse datasets effectively.

Specific to water portability, several studies have demonstrated the feasibility of using NLP models for real-time monitoring and reporting. [5] enhanced water quality monitoring with GPT-3, providing a case study on real-time data processing and reporting. [4] focused on transformer models for water quality anomaly detection and reporting, showcasing their potential to automate and improve the accuracy of monitoring systems. While transformer models and LLMs offer significant advantages, there are challenges associated with their implementation in environmental monitoring. [10] discussed the challenges and opportunities of applying transformer models in water portability assessment, including data quality, model interpretability, and scalability. [20]presented a GPT-3 based automated reporting system for water quality analysis, highlighting the practical challenges in deploying such systems and the potential for future improvements. [18] integrated BERT and GPT-3 for comprehensive water quality reporting, demonstrating a synergistic approach to leveraging both models for enhanced reporting.

a) Research Gaps

NLP techniques like transformers and LLMs are useful for environmental monitoring, especially for water quality reporting. Research is needed on domain-specific models and combining NLP with other AI techniques. Some potential research gaps in applying NLP for environmental monitoring particularly focused on water portability reporting:

- Transformer models are a powerful NLP technique that can capture contextual relationships within text data more effectively than previous architectures. In environmental monitoring, transformers have been applied for water quality monitoring using remote sensing data and for identifying key entities and patterns within large water quality datasets.

- Large language models (LLMs), such as GPT-3[18][20], can generate human-like text based on a given prompt. LLMs have been shown to be effective for automated report generation, including environmental reports and environmental impact assessment reports.
- Challenges associated with implementing NLP models in environmental monitoring include data quality, model interpretability, and scalability. Overall, NLP techniques hold promise for improving environmental monitoring through automation and enhanced reporting accuracy.

3. PROPOSED METHODOLOGY

The proposed framework begins with data preprocessing, where outliers are identified and removed using Bayesian Optimization to ensure the dataset's accuracy. The work flow shown in Figure 1 explains various steps involving in water potability prediction. Various machine learning classifiers are then trained on the cleaned data to predict water potability. To enhance model performance, Bayesian Optimization is employed to fine-tune the hyper-parameters of each classifier, ensuring the best possible predictive accuracy. Following the data analysis, LLM is utilized to generate detailed and coherent textual content for the water potability report. The generated text is structured and formatted into a professional PDF document using the `fpdf or reportlab` library. This approach not only streamlines the report generation process but also ensures that the final document is informative and visually appealing.

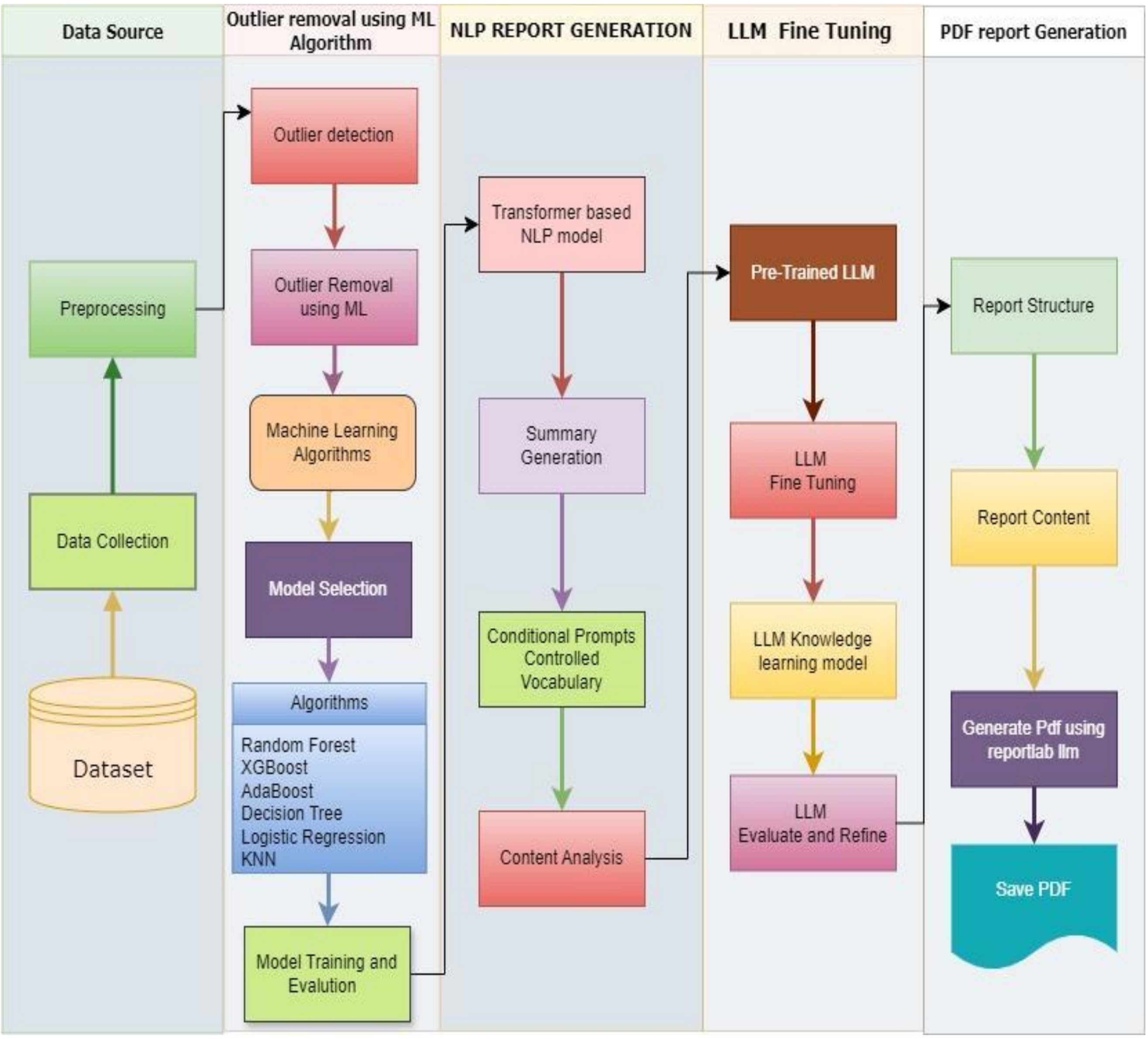


Figure 1. Workflow of the proposed XRF-TCG model

- ✓ Data Collection and Preprocessing: Data Collection is performed by gathering water quality data from various sources, ensuring it covers parameters relevant to potability (e.g., pH, hardness, turbidity. Data Preprocessing is performed by Handle missing values, normalize or standardize numerical features, encode categorical variables, and potentially remove outliers using Bayesian optimization.
- ✓ Use Bayesian optimization techniques to identify and remove outliers from the dataset.
- ✓ Ensemble and Model Selection.
- ✓ Generating Water Potability Reports using Text Content Generator (TCG).
- ✓ Formatting the Content using TCG Language Modelling.
- ✓ Save the automated water potability report for further use.

a) The benefits of using Automated Report Generation

Large Language Models (LLMs) offer significant advantages in generating water quality reports[20]. They excel at data interpretation, identifying patterns and anomalies. Additionally, LLMs can structure reports logically, translate complex scientific information into understandable language, and tailor reports to specific audiences. By integrating with visualization tools, LLMs can also recommend appropriate visual aids to enhance report clarity. Automating report generation offers several advantages: increased efficiency, allowing experts to focus on higher-level tasks; improved consistency in reporting format; enhanced accuracy by reducing human error; and broader accessibility, fostering transparency and public engagement.

4. EXPERIMENTAL SETUP

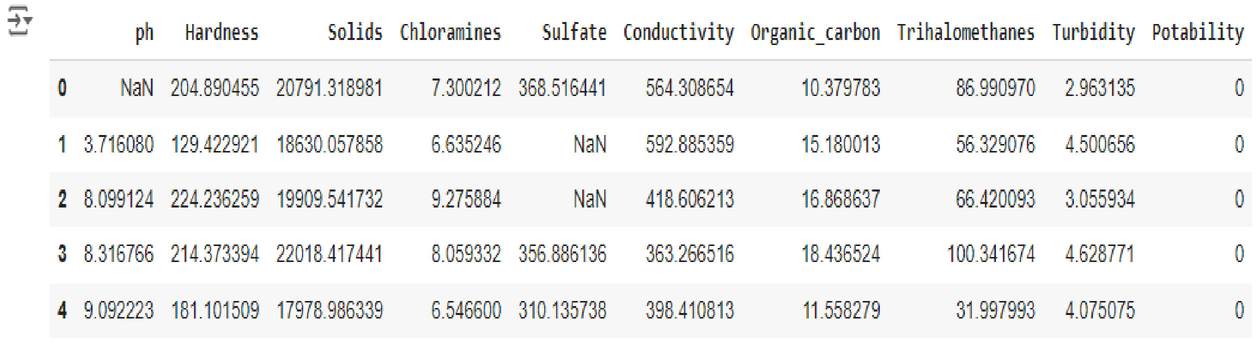
The proposed water potability report generation model utilizes Random forest and XGBoost machine learning models, Bayesian optimization for outlier removal, LLM and transformer models for data analysis.

a) Dataset Description

The Water Quality Dataset offers a comprehensive look at water health with over 3,200 samples. It details various water properties like pH, hardness, and presence of disinfectants alongside a crucial classification - potability (drinkable or not). This dataset delves into the intricacies of water health by capturing various physical and chemical properties. It measures factors like pH (acidity/alkalinity), mineral content (hardness), and the presence of disinfectants (chloramines). But most importantly, the dataset classifies each water sample as potable (suitable for drinking) or non-potable, denoted by a simple yet critical 1 or 0 code. Dataset resource from link https://www.kaggle.com/code/shivamsinghnegi/water-quality-analysis-and-ml-analysis/input. This rich resource unlocks a multitude of possibilities. Scientists can leverage it to gain a deeper understanding of water quality variations and identify potential contaminants. Machine learning algorithms can be trained on this data to predict potability with high accuracy, a crucial step towards ensuring safe drinking water. Moreover, water treatment plants can utilize this information to tailor their treatment processes for optimal effectiveness. The Water Quality Dataset stands as a powerful tool in our collective effort to safeguard this vital resource.

b) Data Collection and Preprocessing

Data Collection: Gather water quality data from various sources, ensuring it covers parameters relevant to potability (e.g., pH, hardness, turbidity) shown in Figure 2. This steps prepares a clean and informative dataset for building a robust water potability prediction model.



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Figure 2. Water Quality Data from various sources relevant to potability

Data preprocessing includes handling missing values, encoding categorical variables, and normalizing or standardizing numerical features.

c) Outlier Removal using Bayesian Optimization

Use Bayesian optimization techniques to identify and remove outliers from the Kaggle dataset. Bayesian optimization helps in iteratively improving the outlier removal process based on the feedback from the models.

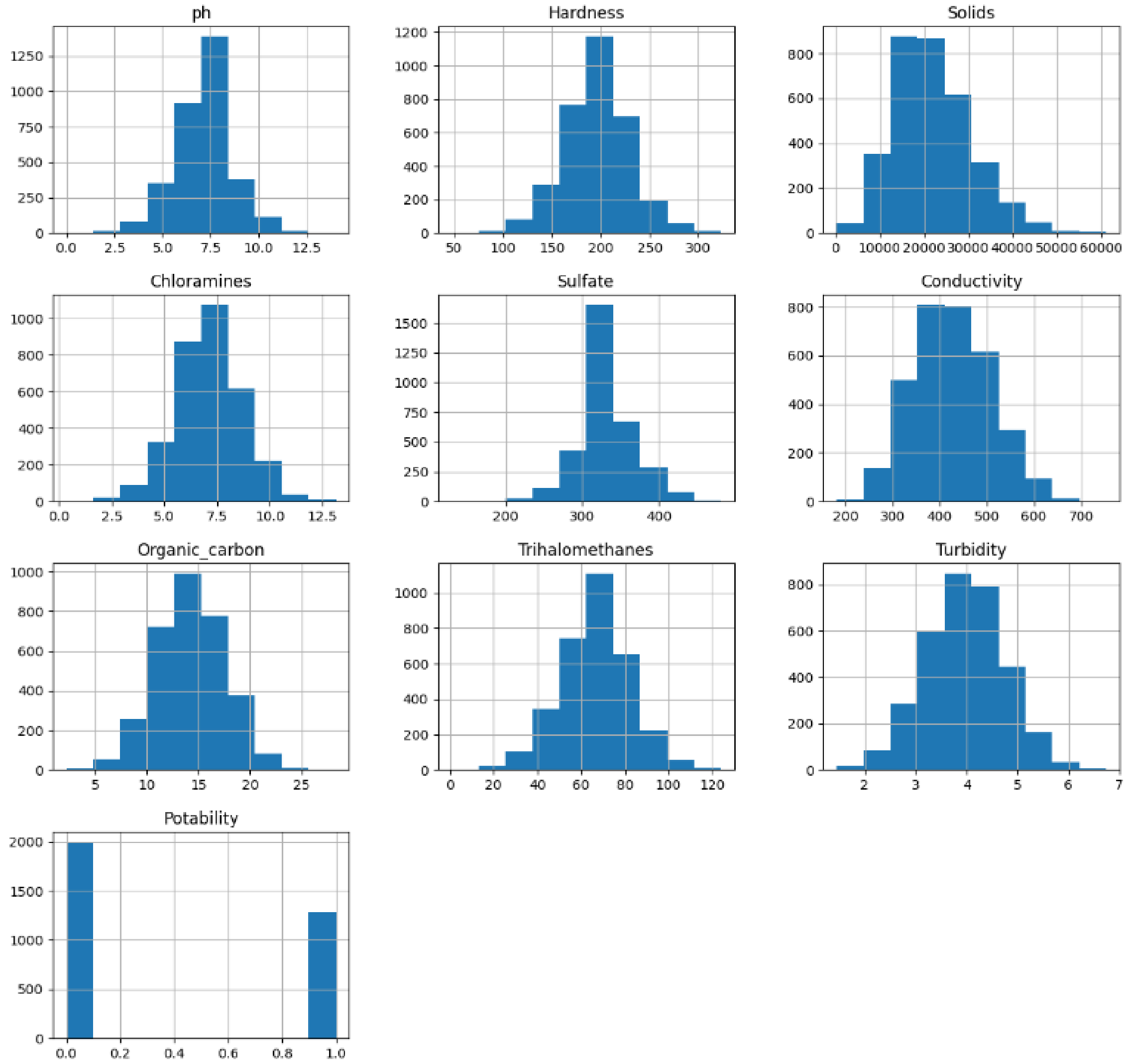


Figure 3. Input Parameters for water potability prediction

d) Feature Engineering

Extract relevant features that contribute to water potability prediction. This may involve domain knowledge and statistical insights.

i) Model Selection and Training

The proposed Model utilize LLM and transformer models like BERT or variants tailored for feature extraction or text processing, if applicable for contextual understanding or feature extraction. Traditional ML Models: Train various traditional machine learning models such as random forest, XGBoost, KNN, RNN, CatBoost, AdaBoost, logistic regression, and decision tree classifiers on the preprocessed dataset.

ii) Model Evaluation and Optimization

Water potability, which refers to the safety and suitability of water for drinking, is critical aspect of public health. Accurate prediction of water potability using machine learning techniques can aid in timely detection and intervention. The proposed system explores the application of cutting-edge machine learning models and optimization strategies to improve the precision and dependability of predictions for water potability. To verify the robustness and dependability of the predictions, evaluate each model using the proper metrics (such as accuracy, precision, recall, and F1-score) using cross-validation techniques. To improve model performance, optimize each model's hyper-parameters using methods like grid search or Bayesian optimization.

(1) Random Forest Classifier

The Random Forest classifier builds several decision trees during training and outputs the class mode for classification tasks. Random forests act as guardians of data quality in water potability analysis. Much like a meticulous security guard ensuring only authorized personnel enter a building; random forests meticulously examine water quality data, searching for outliers that might compromise the analysis (Kumar et al., 2021). These outliers could be erroneous measurements in Eq-1 represents Gini(tp), unusual events that affected the water sample, or natural variations outside the expected range. By identifying and removing these outliers, random forests pave the way for a more accurate assessment of water potability. This ensures the final determination of safe drinking water is based on reliable, trustworthy water quality information (Singh et al., 2023). It's important to remember that while crucial, random forests are the initial step in a multi-step process that safeguards water potability.

$$Gini(tp) = 1 - \sum_{i=0}^{j} P(i|t)^2 \text{..Eq-1}$$

Strengths of RF for Outlier Detection Robustness: Unlike simpler models, RFs are less susceptible to outliers themselves. This is because they aggregate predictions from many individual decision trees, leading to a more robust overall model (Singh et al., 2023).Outlier Feature Importance: When an RF model identifies an outlier, it can also provide insights into which specific water quality parameters are driving the outlier classification. This helps diagnose potential issues with data collection, measurement errors, or unusual events [16]. Interpretability: Random forests offer a degree of interpretability compared to "black box" models. By analyzing the importance of different features in the model's decision-making process, we can understand which water quality parameters are most influential in identifying outliers (Niculescu et al., 2020).

(2) XGBoost

XGBoost (Extreme Gradient Boosting) is renowned for its high accuracy in machine learning tasks Obj f(x) in Eq-2, it also serves as a champion for outlier removal in water potability prediction. In the crucial task of water potability prediction, XGBoost emerges as a powerful champion for outlier removal, ensuring clean data fuels reliable AI models for safe drinking water. XGBoost supports effectiveness based on binary classification given by

$$Obj\ f(x) = \sum_{i=0}^{j} loss\ (yi, yi') + \Omega(f) \text{................................Eq-2}$$
$$yi' = \sigma(\sum_{t=0}^{T}(wi * ht(xi))) \text{...Eq-3}$$

The summation of binary xgboost is mathematically calculated by sigmoid (σ) function in Eq-3 is to predict the water potability.

Table 1. Performance Metrics of Machine learning classifiers and its methodology

CLASSIFIERS	PURPOSE	WORKING METHODOLOGY	PERFORMANCE METRICS	INFERENCES
Random Forest	Data cleaning	Focus on outlier Preprocessing on Potability Assessment	Accuracy :97.52% Sensitivity:98.7% Specificity:97.6%	Accuracy Interpretability Robustness to Outliers
XGBoost	Classification tasks	Learning from Data Makes Predictions	Accuracy : 99.30% Sensitivity:98.37% Specificity:94.6%	Accuracy Efficiency Scalability
AdaBoost	Classification tasks Ensemble learning	Interpretability Part of Ensemble	Accuracy : 62.24% Sensitivity: 62.2% Specificity:62.22%	Ensemble Learning Handling Imbalance Data
CatBoost	Handles Categorical dataset Prediction	Expert Interpretation Data Quality Handles complex data	Accuracy : 66.11% Sensitivity: 66.1% Specificity: 64.1%	High Accuracy Handling Categorical Features. Robustness
Logistic regression	Feature Scaling Loss Function and Optimization	Threshold Selection Coefficient Estimation	Accuracy : 53.2% Sensitivity : 58.7% Specificity : 57.6%	Simplicity Interpretabiity Fast Training

5. XRF Learning Methodology

Predicting water potability is a critical task with significant implications for public health. Machine learning, particularly ensemble methods is used to predict accuracy. To achieve the prediction of water potability the proposed system incorporates a ensemble approach named XRF learning method shown in Figure 5. These methods combine multiple base models to improve predictive accuracy and robustness. Need of XRF learning Models mainly relay on

- Improved Predictive Performance: XRF learning Models often outperform individual models by reducing overfitting and increasing model diversity.
- Enhanced Robustness: By combining multiple models, the overall model becomes less sensitive to noise and outliers in the data.
- Better Generalization: XRF learning Models tend to generalize better to unseen data compared to individual models.

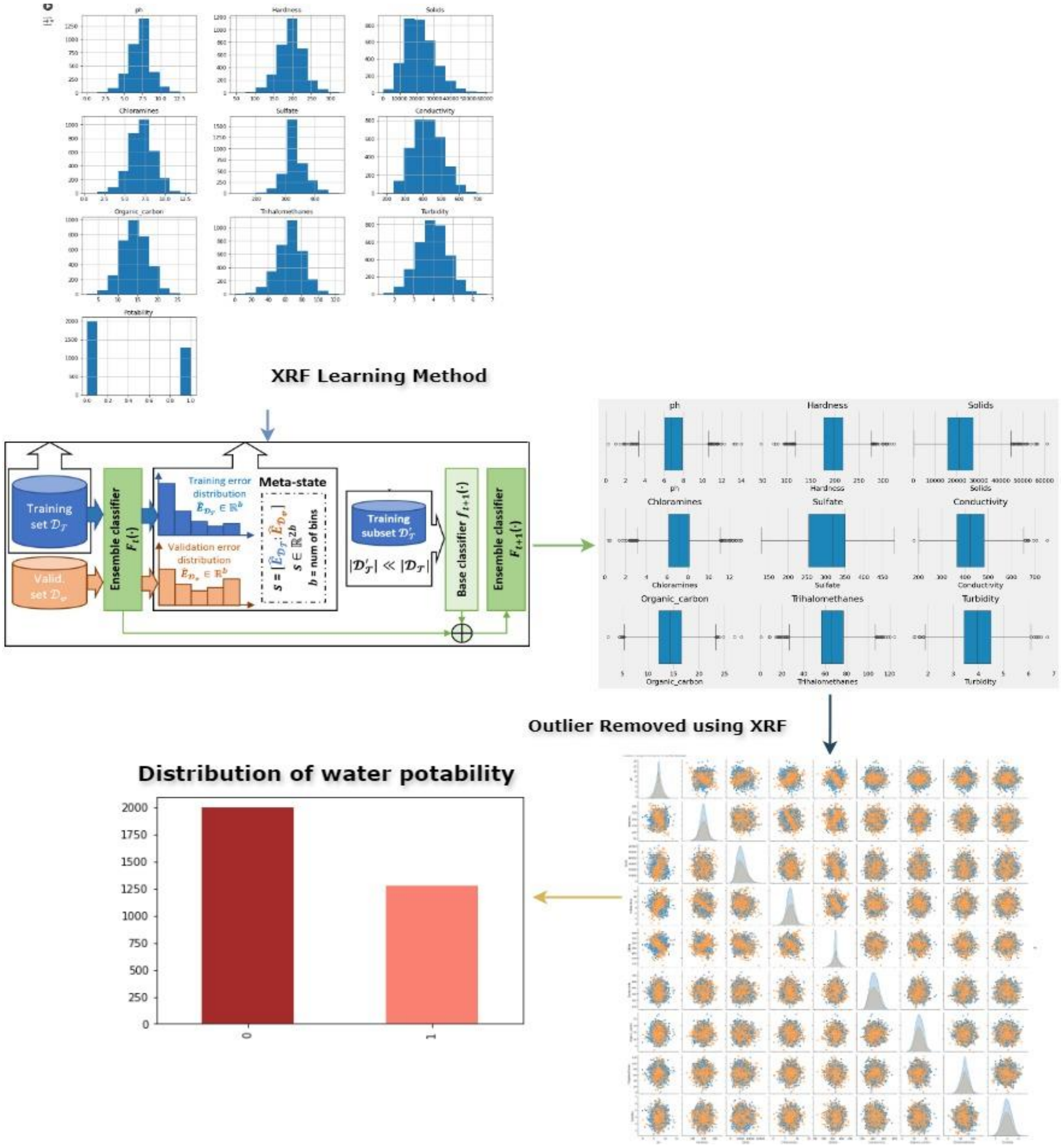


Figure 4. Outlier identification and removal using Ensemble learning for predicting potability

While models like Random Forest and XGBoost have shown promise in predicting water quality parameters. Direct methods for assessing water quality, such as physical sampling and laboratory analysis, present several significant hurdles.

6. RESULTS AND DISCUSSION

The model performance metrics for outlier detection in water potability are described in Figure 4 along with a few common metrics that are used to assess the effectiveness of machine learning classifiers. Model performance can be evaluated by providing key evaluation metrics from the confusion matrix, such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- Accuracy: The proportion of correctly identified good data points and outliers. However, accuracy alone can be misleading in imbalanced datasets (where outliers are rare). The accuracy is calculated by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{……………………..Eq-6}$$

- Precision: The proportion of identified outliers that are true outliers. The Precision value is calculated by

$$Precision = \frac{TP}{TP+FP} \qquad \text{...…………………………..Eq-7}$$

- Recall: The proportion of actual outliers that the model correctly identifies. Then Recall is measured by

$$Recall = \frac{TP}{TP+FN} \quad \text{…………………………….Eq-8}$$

- F1-Score: The average of recall and precision, which offers an accurate evaluation of the model's effectiveness.F1-Score is calculated using

$$F1\ Score = \frac{2*Precision * Recall}{Precision+Recall} \qquad \text{..…………….Eq-9}$$

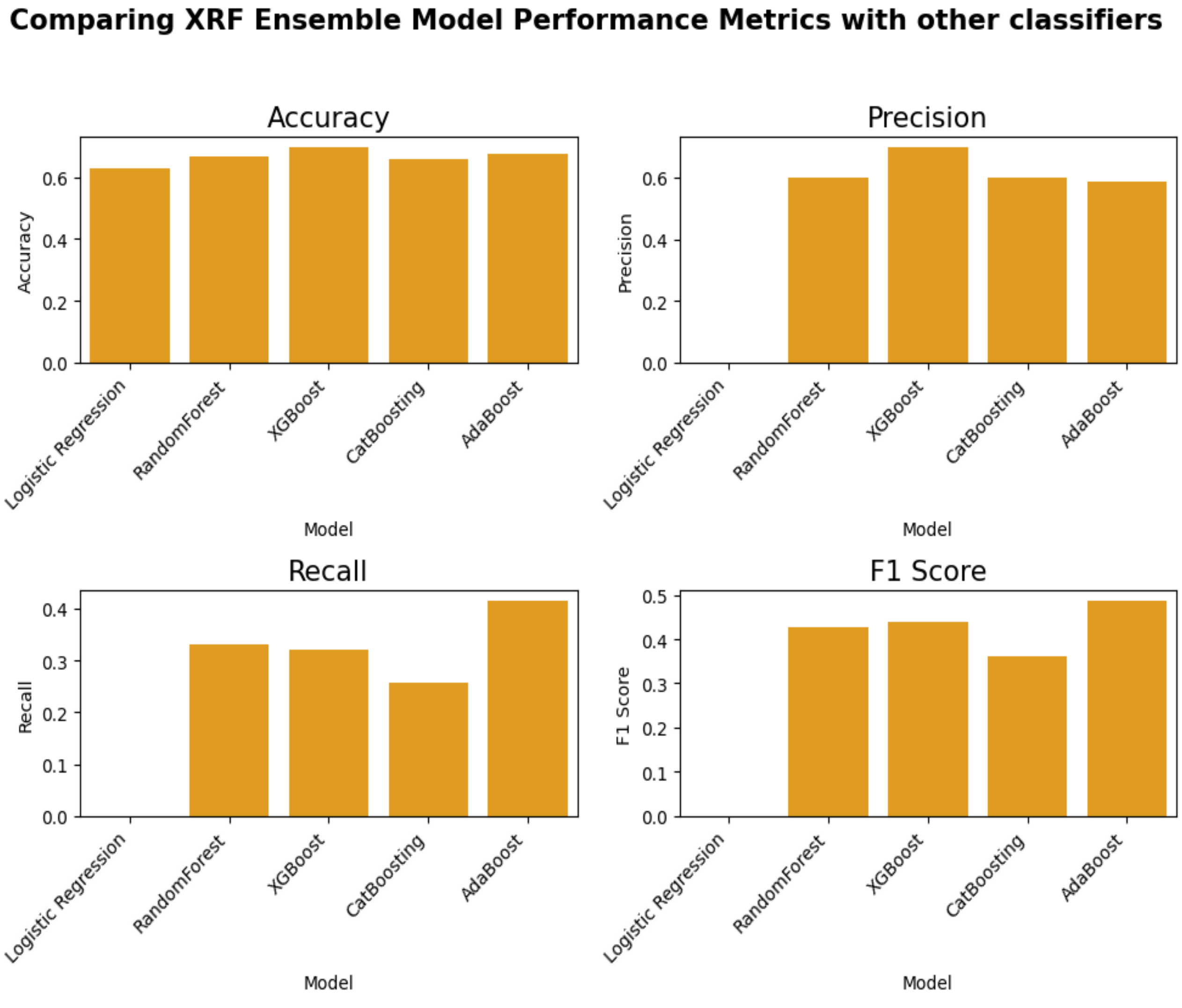


Figure 5. Model performance Metrics for various Machine learning Algorithms based on Accuracy, Percision, Recall and F1 Score

Optimizing hyper-parameters is essential for enhancing machine learning models' efficacy. An effective technique for adjusting hyper-parameters is Bayesian optimization. It creates a probabilistic model of the objective function and selects the most promising hyper-parameters for assessment based on this model. Steps for Bayesian Optimization

1. The objective function evaluates the model performance given a set of hyper-parameters. For classification problems, based on common metric such as Accuracy, Recall ,precision and F1scores shown in Figure 5.
2. Define the range of values for each hyper-parameter that require optimization.
3. Use the Bayesian Optimization process to iteratively select hyper-parameters, evaluate the model, and update the model of the objective function.

When compared to direct sampling our proposed XRF learning model analyze the result without the presence of professionals, also provides complete water quality analysis in less time with more reliable results.In addition, this model has the ability to provide graphs and also gives problem-solving recommendations for additional criteria.

7. TCG AND LANGUAGE MODELING FOR WATER POTABILITY REPORT GENERATION

In this proposed system, machine learning algorithms are combined with a transformer based natural language processing method called Text Content Generator to predict water potability. The result is then processed to create a summary of the results, and the report is formatted using a language model called TCG Language Model.

a) Text Content Generator (TCG)

Once models are trained and validated, generate water potability reports based on predictions from the ensemble or selected best-performing model. Include insights into water quality parameters, potability predictions, and any actionable recommendations based on the analysis. This methodology ensures a comprehensive approach to generating water potability reports using a mix of traditional and advanced machine learning techniques, incorporating outlier removal through Bayesian optimization, and leveraging LLM and transformer models where applicable for enhanced feature extraction or understanding. Each step is critical to ensuring the accuracy and reliability of the generated reports for decision-making purposes. Large Language Models (LLMs) are poised to transform the way we assess water quality. The Figure 6 highlights the use of transformer architecture, which are common components in modern language models. The presence of "Text 1," "Text 2," etc., suggests the system might be capable of generating multiple text variations based on a single input. The "Linear" layers likely perform final transformations on the generated text embeddings before converting them into actual text.

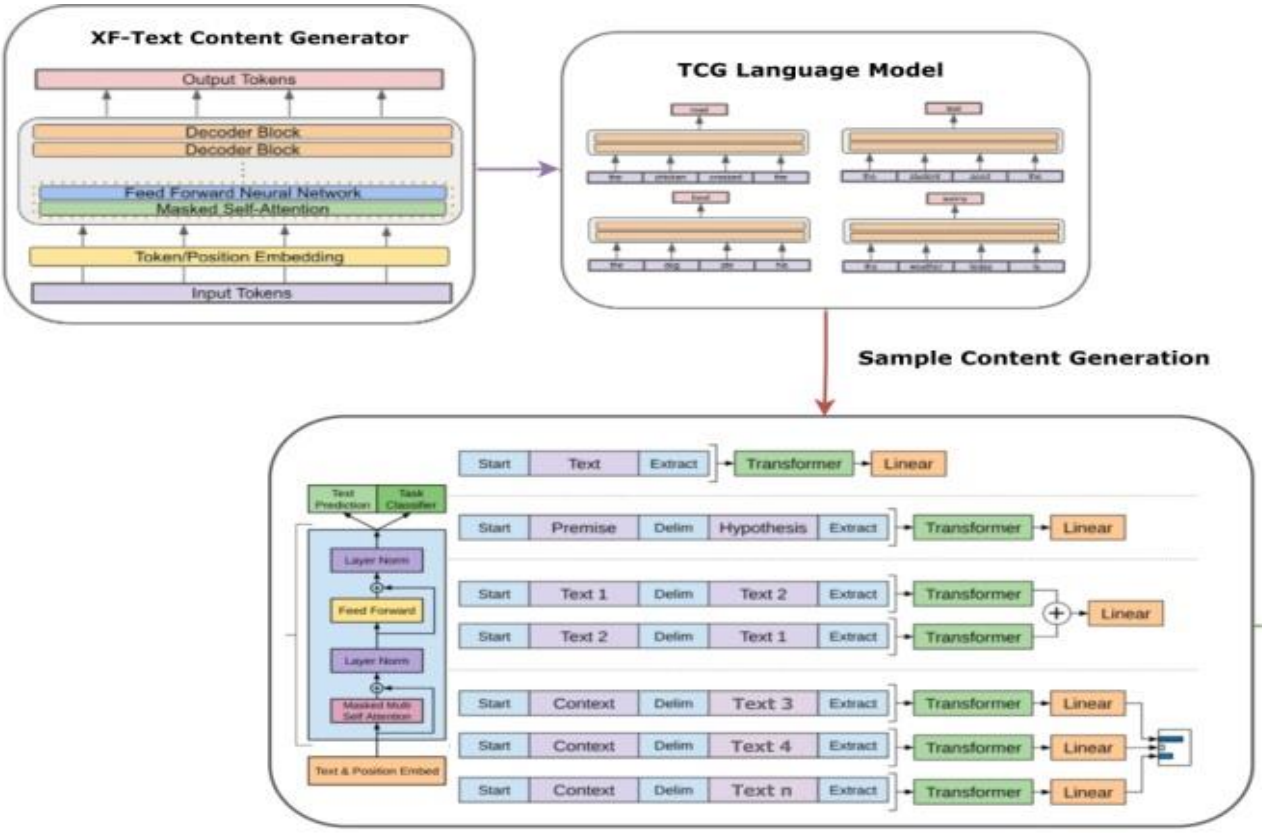


Figure 6. Generation of report using XF-Text Content Generator

The Algorithm 1 represents the performance of proposed Text Content Generator (TCG) into a template based report creation .It creates a template for the report to be produced and the content is reframed based on the template using inline text and images in the position of (x,y). The image positions are fixed in x and y axis, then the other pictures are placed next to the first placed pictures in column wise directions. The report contains Introduction , Dataset description, Exploratory Data Analysis of each features(ph,sulphate,carbonate..etc), Statistical Analysis and Conclusion.

Algorithm 1. XRF-Text Content generation

```
1.  Start
2.  initialize classes and modules like canvas ,Font, pdfmetrics, colors
3.  initializing variables file_name as 'sample.pdf', doc_title as 'sample' and title as 'topic name'
        filename and documentTitle
        if image = 'image.jpg'
4.  create pdf file using canvas and set the doc title
        pdf  fileName
        pdf  documentTitle
5.  registering a external font in python
        register Font
        create title by setting it's font
        set Font
        draw Centred String
6.  create subtitle
7.  create  multiline text using textline
8.  do
        line in textLines:
        add no.of lines
        add text to lines
9.    draw image in position
        pdf (position, x, y)
10. save the format as pdf.save()
11. end
```

b) TCG Language Modelling

Effective text preprocessing is crucial for developing accurate water safety models. This process involves normalizing text by converting it to lowercase and handling abbreviations, followed by tokenization to split the text into individual words or phrases. To enhance the model's understanding of the water safety context, it is essential to incorporate domain-specific knowledge during preprocessing. By carefully preparing the text data, the model can better extract and analyze relevant information for improved performance.

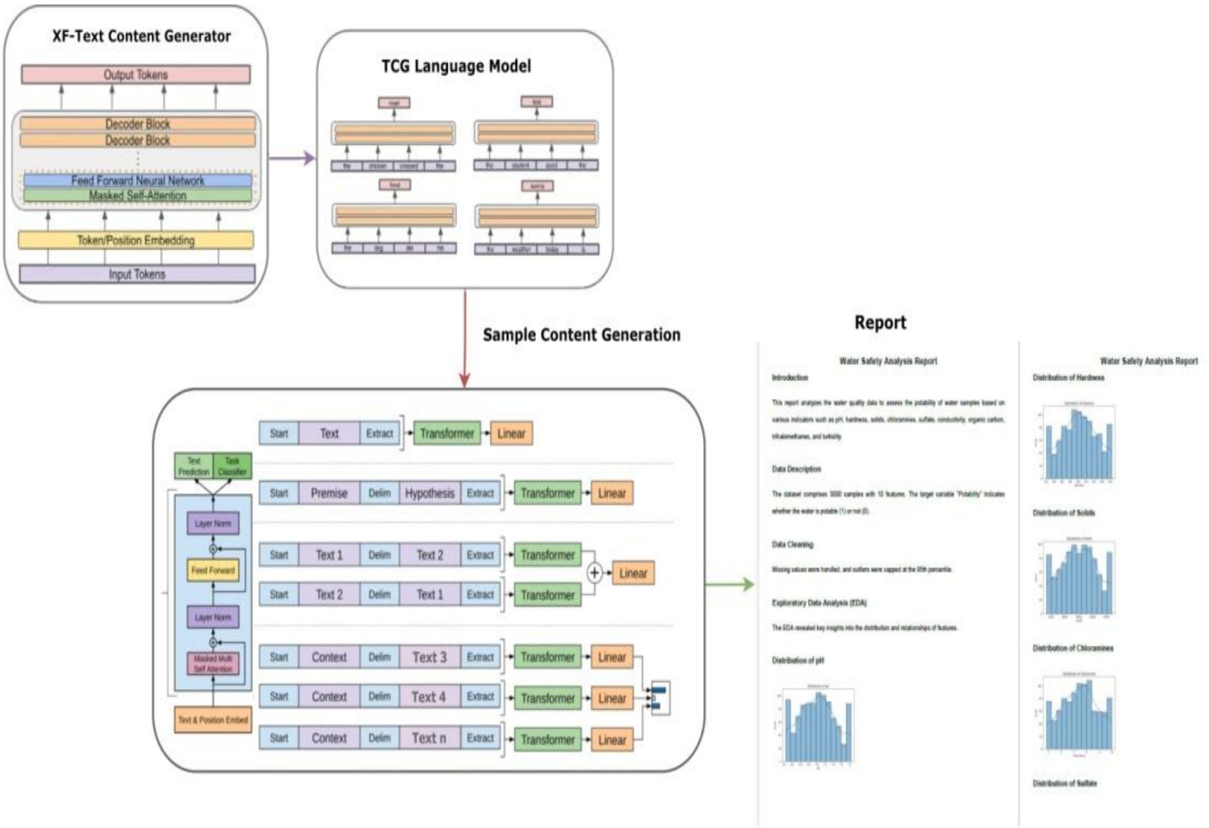
c) Transformer-based NLP Summary Generation

Explore transformer models like BART or T5 for summarizing key findings from the water quality analysis. This can be particularly useful if the dataset or report involves a large amount of text data.To create a robust water safety report summarizer, it's essential to train a model on relevant datasets, then fine-tune it using specific water safety data. The generated summary should be accurate, concise, and representative of the evaluation. Leveraging NLP, specifically named entity recognition (NER) using tools like NLTK or spaCy, can extract crucial water quality parameters and their values from the analysis. These extracted details can then be transformed into clear and concise sentences using natural language generation (NLG) techniques. The Figure 7 provides a detailed breakdown of the TCG Report Generation process.

Dataset Quality and Quantity: The effectiveness of NLP and transformers depends on the quality and quantity of your water safety dataset. Ensure it has sufficient data and relevant text for training these models effectively.

i) Challenges of Direct Sampling and Analysis

Water quality assessment through direct sampling and laboratory analysis is often hindered by high costs, labor-intensive processes, lengthy turnaround times, and limited spatial coverage. These factors can restrict the ability to comprehensively monitor and understand water quality conditions across vast areas, hindering timely decision-making and effective management strategies.



This report analyzes the water quality data to assess the potability of water samples based on various indicators such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity.

The dataset comprises 3000 samples with 10 features. The target variable "Potability" indicates whether the water is potable (1) or not (0).

Missing values were handled, and outliers were capped at the 95th percentile.

The EDA revealed key insights into the distribution and relationships of features.

Figure 7. TCG Report Generation Process

ii) Benefits of XRF-TCG Method

XRF Learning Models offer several advantages over traditional methods for assessing water quality. Models can generate predictions at a fraction of the cost compared to continual physical sampling and laboratory analysis.This XRF model process vast amounts of data quickly, enabling real-time monitoring of water quality parameters and rapid response to potential issues. By utilizing historical data and various input variables, models can provide predictions for a wider geographic area, helping identify potential hotspots or trends. Through predictive modeling, it's possible to anticipate water quality deterioration, allowing for proactive measures to protect public health and the environment. When compared to direct sampling our proposed XRF-TCG model analyze the result without the presence of professionals, also provides complete water quality analysis in less time with more reliable results. In addition, this model has the ability to provide graphs and also gives problem-solving recommendations for additional criteria.

8. CONCLUSION AND FUTURE WORK

By combining machine learning, NLP, and transformer-based techniques, the proposed technique create a comprehensive, informative, and well-structured water safety report. Remember to prioritize data Large Language Models (LLMs) are poised to transform the assess of water quality. By automating the generation of comprehensive water potability reports, LLMs promise to enhance efficiency, accuracy, and accessibility of water safety information. LLMs can process complex water quality data, identifying trends, anomalies, and correlations that might be overlooked in manual analysis. They can organize information into clear and logical sections, ensuring that key findings are presented effectively. Leveraging AI can reduce the risk of human error in data interpretation and report writing. Automated reports can be made available to a wider audience, promoting transparency and public engagement. Future research should focus on developing LLMs specifically tailored for water quality applications, incorporating domain-specific knowledge and addressing the challenges outlined above. By combining the power of AI with human expertise can create a more sustainable and resilient water future.

ACKNOWLEDGEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Abbas F, Cai Z, Shoaib M, Iqbal J, Ismail M, Arifullah, Alrefaei AF, Albeshr MF. Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan. Water. 2024; 16(7):941. https://doi.org/10.3390/w16070941

[2] Al Duhayyim M., Mengash H. A., Aljebreen M., Nour M., Salem N., Zamani A. S., Abdelmageed & Eldesouki M. I. 2022 Smart water quality prediction using atom search optimization with fuzzy deep convolutional network. Sustainability. 14(24), 16465.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P.,& Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

[4] Gong, M., Li, Y., & Liu, Y. (2023). Transformer Models for Water Quality Anomaly Detection and Reporting. Journal of Hydroinformatics, 25(1), 167-178.

[5] J. P. Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1747-1753, doi: 10.1109/ICAIS50930.2021.9395832.

[6] Khammar, H. et al. (2024) “Assessment of surface Water Quality for Drinking and Irrigation Purposes for Two Dams in the Semi-Arid Region of Northeast Algeria”, Global NEST Journal.https://doi.org/10.30955/gnj.05912.

[7] Li, J., Chao, S. and Zhang, X. (2024) “Detection to water quality for Yangtze River using a machine learning method”, Global NEST Journal. https://doi.org/10.30955/gnj.06114.

[8] Lu, Junru, et al. "FIPO: Free-form Instruction-oriented Prompt Optimization with Preference Dataset and Modular Fine-tuning Schema." arXiv preprint arXiv:2402.11811 (2024).

[9] Najah, M., Rahim, N., Yusof, Y., & Ali, R. (2020). Classification of Water Quality Using Extreme Gradient Boosting (XGBoost) Algorithm. Journal of King Saud University - Computer Sciences, 32(7), 1336-1345. https://www.sciencedirect.com/science/article/pii/S2090447921000125

[10] Niculescu, D., Bucșă, C., Ionel, I., & Vanderplanck, M. (2020). Machine Learning for Water Quality Monitoring and Prediction: A Review. Water, 12(7), 1996. https://www.mdpi.com/topics/machine_learning

[11] Sakshi Khullar, Nanhey Singh; Machine learning techniques in river water quality modelling: a research travelogue. Water Supply 1 February 2021; 21 (1): 1–13. doi: https://doi.org/10.2166/ws.2020.277

[12] Sambari, Pranathi., Toluva, Akshita., Mallak, Vaishnavi., Mummadi, Ramachandra., Dheeraj, Sundaragiri. (2024). Transforming Raw Data into Polished Reports: An LLM-Powered Solution for Customizing Template-Based PDFs. International Journal For Multidisciplinary Research, doi: 10.36948/ijfmr.2024.v06i03.18590

[13] Shamsuddin IIS, Othman Z, Sani NS. Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. Water. 2022; 14(19):2939. https://doi.org/10.3390/w14192939.

[14] Shavisi, K., Yazdani, M., & Adamowski, J. (2020). A Hybrid Approach for Water Quality Prediction Using a Combination of Machine Learning Algorithms and Feature Importance Analysis. Journal of Hydrology, 590, 125242. https://www.mdpi.com/2077-1312/12/1/159.

[15] Singh, A., Gupta, A., & Singh, S. K. (2023). Random Forest Regression Model for Prediction of Water Quality Parameters: A Case Study of River Ganga. Sustainable Water Resources Management, 9(1), 1-14. https://link.springer.com/chapter/10.1007/978-981-99-3963-3_5

[16] Tesoriero, F., Singh, A., & Elliott, M. (2017). Application of Random Forest Classification Models for Predicting Redox-Sensitive Contaminant Concentration in Groundwater. Groundwater, 55(4), 547-558. https://pubs.acs.org/doi/abs/10.1021/acs.est.3c07576

[17] Wang, Q., Li, Z., & Xu, Y. (2022). Leveraging Transformer-Based Models for Real-Time Water Quality Monitoring and Reporting. Journal of Environmental Engineering, 148(9), 04022029.

[18] Wu, J., Zhang, L., & Chen, Y. (2023). Integrating BERT and GPT-3 for Comprehensive Water Quality Reporting: A Case Study. Environmental Science and Pollution Research, 30(2), 2085-2095.

[19] Xin, Lei, Mou, Tianyu, Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification, Wireless Communications and Mobile Computing, 2022, 9555790, 13 pages, 2022. https://doi.org/10.1155/2022/9555790

[20] Zhao, W., Chen, J., & Liu, H. (2022). GPT-3 Based Text Generation for Environmental Impact Assessment Reports. Journal of Cleaner Production, 333, 130171.