National College of Ireland

Comparative Analysis of Machine Learning and Deep Learning Models for Water Potability Prediction

MSc Research Project
Data Analytics

Pooja Sree Maniga
Student ID: 23243236

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Pooja Sree Maniga
Student ID:	23243236
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Anu Sahni
Submission Due Date:	12/12/24
Project Title:	Comparative Analysis of Machine Learning and Deep Learning Models for Water Potability Prediction
Word Count:	9700
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Pooja Sree Maniga
Date:	12th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Analysis of Machine Learning and Deep Learning Models for Water Potability Prediction

Poojasree Maniga
23243236
MSc in Data Analytics
National College of Ireland

Abstract

Water is one of the most important human needs, but with the rising incidences of waterborne diseases, there is a need for an efficient method of monitoring the water quality. Current approaches for evaluating the potability of water are time-consuming and may not be applicable in real-time. This research focuses on the following gaps and limitations in the prior research to examine the effectiveness of machine learning (ML) and deep learning (DL) models. The performance under different conditions was tested on two datasets: a dataset of 300,000 samples and low sampled and class imbalanced dataset of only 3,000 samples. Machine learning models like XGBoost and Random Forest along with various deep learning models like TabNet were used, while feature scaling, one-hot encoding, and handling missing values were done for the dataset. These findings showed that ML models, especially XGBoost, were more accurate with 97% and efficient than DL in predicting the outcomes because of their suitability in managing structured data. Although some DL models such as TabNet used provide good results, these models raise several problems in terms of their computational load and the required volume of data. This study compares traditional ML methods with state-of-the-art DL techniques in a systematic manner to provide best practices for water quality data.

1 Introduction

Water is among the most important resources in the globe as it is required by all individuals for survival and among the necessities of life is water. The problem of water-borne diseases has not been solved by better water treatment methods because most regions in the world cannot afford proper testing of water. Demographic increases and enhancement, especially in the cities, as well as inadequate water management and disposal of waste from agriculture and industries, have contributed to the increase in the stringency of water quality. Water quality prediction is valuable to protect people's health and maintain the environment when accurate information is provided at the right time (Wu and Wang; 2022; Dalal et al.; 2022). However, traditional methods for deriving WQI are accurate, require a lot of manual work, are time-consuming, and can only offer results at a later time. In this research, both ML and DL methods will be applied to formulate and solve these problems and also enhance water quality prediction (Saroja et al.; 2023; Ghosh et al.; 2023; Ainapure et al.; 2023).

In the case of water quality prediction, many researchers have demonstrated that the application of ML and DL algorithms is quite effective (Dawood et al.; 2021; Suleiman et al.; 2023; Abuzir and Abuzir; 2022; Patel et al.; 2022). They are very accurate like Random Forest and XGBoost but again highly dependent on features and thus, the model is limited to recently emerging features and newer contexts. On the other hand, the DL models such as LSTM and TabNet have the advantage of training and interpreting big and high dimensional data, and yet there is still a lot of room for improvement to outdo traditional models (Xu et al.; 2019). The prior literature also focuses on the application of larger datasets with water sources such as lakes, rivers, and groundwater to enhance the reliability and generalizability of the models Zhao (2021). To fill these gaps in this work, improved computational methods are employed and evaluated on large and varied datasets.

Research Question and Objectives: The primary research question this study seeks to answer is; How can one develop ML and DL models that can accurately and with good reliability forecast water quality across water sources? To answer this question, the goals of this research are as follows: compare the performance of different conventional ML, and DL techniques; define the best settings of architecture and hyperparameters for models; study the strengths and weaknesses of each approach in terms of interpretability and computational complexity. Moreover, this work seeks to address the problems of how to implement these models in practice, particularly in large-scale settings. The presence of variability in greater numbers of samples also increases the prospects of generalized estimates; the architectural evolution of DL such as TabNet can build better forecast models than conventional Machine Learning algorithms with due calibration; there are humongous trade-offs between desirability, readability, and computational complexity to choose from the most suitable models for specific uses.

In pursuing these goals, this work uses a dataset from a large set, at 300,000 samples, extracted from a 5-million-row population dataset. The unavailability of data is addressed, and feature scaling and normalization are also considered. The presented models are Decision Tree, Random Forest, XGBoost, LSTM, GRU, and TabNet are used. The enhancement of the parameters is used to achieve better performance, and the comparative assessment is used to show the pros and cons and the possibility of getting better solutions. Because this study only incorporates experiments with big data, it aims to fill the above gaps. The present work offers significant contributions to the utilization of computational ML and DL methods for water quality prediction through a thorough discussion of the performance differences for a range of data sets. It also provides fine-grained knowledge of how expanded DL architectures and diverse datasets support more precise and reliable predictions. Together, it provides solutions to the concerns of scale and explainability while serving as a bridge between novel concepts and real-world applications in water resource management and public health.

The specific organization of the report is one of the best ways to show that the ideas are interconnected to produce a chronological approach, and the report contains an introduction that involves the research problem, the relevance of the subject, and the objectives of the study. Besides explaining prior works, the literature review also points out a gap that forms the basis of the study. The nature and origin of the datasets, the preparatory processing of the data, the models used, and the assessment measures are discussed in the methodology section. Mainly, results and discussion sections related to the analysis of mechanisms of experiments, the comparison with the findings of other studies, and the indication of further research relevance. The final part of the study discusses the key

conclusions, the academic contributions of the study, and recommendations for future studies. This structure allows for a comprehensive understanding of this research project and its implications.

2 Related Work

Base papers: Various machine learning (ML) and deep learning (DL) based studies in relation to water quality prediction have been conducted and most of their interests include how sophisticated ML and DL models can further enhance their model accuracy, scalability, and generalizability on multiple attested datasets and enhanced methodologies. For example, Xu et al. (2019) and Zhao (2021) pointed out that the models are less robust when larger datasets were constructed from samples from different water sources, such as lakes, rivers, and groundwater. As noted by Zhang and Jin (2024), novel DL models like TabNet have not been explored for tabular data heavily in the field. Furthermore, as Alnaqeb et al. (2022) and Rustam et al. (2022) indicated, sampling techniques or weighted loss functions may represent promising areas for future study to address the problems posed by the class imbalance of the datasets.

2.1 Machine Learning Approaches

Decision Trees, Random Forests, and XGBoost have been identified as the most impactful models for water quality prediction. The researchers have proved that XGBoost is an accurate, effective classifier that surpasses others as the result of the gradient-boosting mechanism and the input's structure compatibility (Gill et al.; 2023). Likewise, for the LightGBM, the stability in handling the imbalanced data has been applauded, though containing a certain accuracy of 99.74 % in certain cases (Khan and See; 2024). Since Random Forest is an ensemble method, it provides high interpretability and high reliability for water quality prediction tasks (Whelan; 2022). Nonetheless, these models largely depend on features extracted from data and lack effective mechanisms to handle temporal dependencies, making them unsuitable for situations demanding sequential data analysis (Gao et al.; 2023).

Another successful type of model integration has been accomplished by bagging and boosting techniques. For example, CatBoost and LightGBM have shown great performance on water quality classification problems, proved by (Chakravarthy et al.; 2023). Further, the active learning methods have been combined with support vector machines of which the enhancement of binary classification has been noted (Zhao; 2021; Alnaqeb et al.; 2022). However, these methods have a high computational complexity and include much hyperparameter optimization, which becomes a problem in limited circumstances. This trade-off between accuracy and computational efficiency is an inherent characteristic of the algorithms, which deserve scalable solutions.

2.2 Deep Learning Approaches

ANNs (Artificial Neural Networks) have been popular for the flexibility of their application and relatively simple implementation The authors Rustam et al. (2022) obtained 96 % accuracy on small datasets, thus showing that ANN can be used to predict water quality. However, the study shows that this accuracy of ANN deteriorates when applied to the larger and more diverse data set of 69%. The former leads to a decline in

the performance level of ANN since it is not capable of capturing complex, non-linear relationships for that additional optimization.

One of the most common reasons for using time-series forecasts is the existence of temporal dependencies, especially using Long Short-Term Memory (LSTM) networks. Xu et al. (2019) used LSTM in combination with SARIMA to improve the prediction of seasonal fluctuations in water quality and showed higher accuracy than statistical models. In the same way, Liu et al. (2024) proposed an improvement of LSTM using the Cuckoo Search algorithm, which yielded even higher performance indices, including lower error rates. However, these models require extensive computation and large datasets, which poses problems for implementation in areas of low-resource settings.

Hybrid Models: Autoencoders and CNN-BiLSTM: Some studies combine Autoencoder with LSTM, and CNN with BiLSTM to solve dimensionality and the multivariate correlation problem. Zhang and Jin (2024), for instance, used the AE-LSTM to reach a higher accuracy in the prediction of nitrogen and phosphorous levels. Likewise, Guo et al. (2023) proposed CNN-BiLSTM with attention mechanisms and outperformed other DL models in identifying spatial-temporal features in river water quality datasets. Although rich in potential, these models are computationally intensive and depend on the availability of feature and data preprocessing talent.

2.3 Emerging Techniques and Novel Approaches

Wavelet Transform-Based Models: Wu and Wang (2022) suggested the use of a hybrid ANN-Wavelet-LSTM process for the deconstruction of water quality signals into high and low-frequency bands. This approach was more accurate and less sensitive to overfit than the single models. But to implement it and especially preprocess the data it needs large amounts of domain knowledge.

Seasonal and Spatial Models: SARIMA-LSTM was used to consider seasonal fluctuations in water quality in Xu et al. (2019) and spatial-temporal features in AE-LSTM in Zhang and Jin (2024). Although these methods successfully captured the temporal behavior of the environment, they were often coupled with extensive data preprocessing and model parameter optimization to obtain the best results.

Attention Mechanisms and Advanced Feature Selection: Guo et al. (2023) have also stressed that due to attention mechanisms, critical features are emphasized in the time series datasets, and thus, interpretability, as well as accuracy, are improved. Likewise, Zhao (2021) paid attention to active learning and feature selection for enhancing SVM results because data knowledge is essential for model enhancement.

2.4 Challenges and Limitations in Current Literature

Some limitations still exist in water quality prediction. A limitation in developing datasets is a severe problem, especially when using region-specific datasets, which introduces significant limitations to the generalization of the results (Gill et al.; 2023; Zhang and Jin; 2024; Sreekumar et al.; 2024). Including various categories of water as it is done in this study solves this problem by increasing the range of applicability. One of the difficulties is the balance between the model's complexity and its ability to be easily understood. While there are highly interpretable models, for instance, Random Forest, deep learning models are non-interpretable, or as some like to refer to them, black boxes that make them less applicable in legal and practical environments. In addition,

Scalability and computational complexity are the major challenges that users must face Specific deep learning models such as CNN-BiLSTM and AE-LSTM consume much computational power and thus are not very suitable for real-time and low resource settings (Wu and Wang; 2022; Stepanova et al.; 2024).

2.5 Relevance and Contribution of Current Research

This research addresses these challenges by leveraging a large and diverse dataset of 300,000 rows and employing various machine learning and deep learning models. By optimizing performance through hyperparameter tuning and achieving full convergence, the study ensures robustness and efficiency. The inclusion of data from diverse water sources, such as lakes, rivers, and groundwater, enhances the ecological relevance and generalizability of the findings. Moreover, this research emphasizes a balance between accuracy and interpretability, making a significant contribution to the field. It effectively addresses the trade-offs identified in existing literature, paving the way for more scalable and practical solutions in water quality prediction.

3 Research methodology

This Research focus is on the Primary dataset as it is the inclusion of data from different water types, including lakes, rivers, and groundwater, increases the external validity and ecological potential of the results. The present research uses a broad and comparative strategy to investigate how well both ML and DL models predict water pot- ability. The selection of ML and DL models is justified by literature focusing on their ability to accurately predict structured/tabular data and to learn complex features (Gill et al., 2023; Whelan, 2022). Random Forest and XGBoost were selected since they are efficient in capturing non- linear relationships and in dealing with missing values while LSTM, GRU and TabNet were taken as powerful DL models to search for other under-lying features and interactions within the observation. This study aims to answer the research question interactions: What adjustments must be made to ML and DL models so that the forecast of water quality across diverse water sources is optimal and accurate? **Data Collection and Preprocessing** The main data set for this research includes 300,000 samples of physical and chemical attributes and continuous variables consisting of pH, iron, nitrate, chloride and other attributes. It also includes categorical features like watercolor and metadata such as the month having a significant impact on water quality as well as the time of the day. A second test dataset consisting of 3,000 samples was employed to analyze class distribution since the minority class which is the potable water is rare in this set. Each data processing step was performed to prepare the datasets for model training. In the case of the primary data set, rows containing missing values were dropped to exclude an impact of imputed data in the dataset to retain the close resemblance of the dataset to the original one. The procedure of imputing missing values for the secondary dataset was done using the Mean of the relevant column. Features like Color, Source, Month are categorical and were thus converted to numeric values using the label encoding technique. As for class distribution, the primary dataset was balanced so no additional sampling steps were required. But, the secondary dataset had problem of class imbalance and to rectify it, we randomly oversampled the minority class, which was potable water. Standardization of the features was also done to aid the models and bring the attributes to a common scale. In the case of ML, StandardScaler

was used, but in the case of the DL, we used MinMaxScaler to prepare the training data. **Model Selection and Techniques:** The Primary dataset study implemented a total of six models for comparative evaluation: For machine learning we use Decision Tree, Random Forest and XGBoost. For deep learning, we use LSTM, GRU and TabNet. First, we chose these models because they are suitable for structured/tabular data, have nonlinear relationships, and can handle missing values and feature interactions. For secondary dataset, we implemented models such as Random Forest, Support Vector Machines (SVM), and XGBoost. A DL model was constructed using TensorFlow's Keras API, employing neural network architecture with multiple layers. Machine Learning Models: For prediction tasks on structured data, with the ability to capture complex interactions and nonlinear relationship, Decision Tree, Random Forest and XGBoost were selected. In particular, XGBoost is fast at working with imbalanced datasets and resilient with noisy data (Whelan, 2022; Gill et al., 2023). Deep Learning Models: Although the water quality data does not explicitly have temporal dependencies, LSTM and GRU were included as win experiments to see if there were subtleties in the data that could be captured with latent sequential patterns that were not apparent. A deep learning model for tabular data, TabNet was included for its capability to tie their decision tree structures to deep learning algorithms to learn interactions between features in tabular data (Zhao, 2021). We investigated if these deep learning models could improve performance beyond that,even on non-sequential data.

Data Analysis and Evaluation Metrics: The models' performance was evaluated using key metrics: precession, accuracy, precision, recall, and F1-score. For datasets that are imbalanced, these metrics were selected on account of their ability to provide a balanced evaluation. This defines Accuracy as it provides an overall measure of how many correct predictions are made, Precision and Recall, both concentrating on how well the model identifies potable vs non-potable water and how few of the latter are misclassified as the first. To compensate for any class imbalances, and a more complete evaluation of model performance, the F1 score was used, balancing precision and recall. We chose these metrics to verify that not only our models have high prediction accuracy, but also can deal with class imbalance, misclassification. Further, some hyper parameters were tuned using cross validation to prevent over fitting and ensure good generalization to data outside of the set of held out data. To further quantify how the Random Forest and XGBoost models related to water potability, feature im- portance analysis was performed to identify the water quality parameters that were most important in predicting water potability. Furthermore, this research also seeks to strike the balance between model interpretability and accuracy and as such make a noteworthy contribution to the literature. This approach efficiently manages the conflicts pointed out in prior work to enable the development of more flexible and efficient solutions for water quality prediction.

Statistical Techniques and Model Validation: Confusion matrices and classification reports were used to validate the models and judged how the performance of the models. Model optimization was important and grid search cv was used to find the best settings to hyper parameters for ML models. A custom training configuration for the deep learning models was used for LSTM and GRU with the 2 layers and size of 64 of the hidden, as well as early stopping for the TabNet to avoid overfitting. Interpretability and scalability provided singular focus for the analysis of the results with the models' ability to approach the class imbalance, noise and missing data challenges being essential in their applications. Specifically, this study seeks to demonstrate how well different models behaved upon being given larger, more varied datasets and if deep learning models

like TabNet could have a tangible advantage over their traditional ML counterparts in accurately predicting water quality.

Computational Resources and Equipment: Efficient model training on large datasets necessitated the use of multi core processors and GPUs in the computational environment used for training and evaluation of the model. The main programming language used in this thesis is Python and Scikit learn was used for the ML models and TensorFlow was used for the deep learning models. These tools built the models and helped test each model, but at the same time, they also made sure that computations were performed in the best and fastest manner possible.

limitations and strengths of the approach: However, the limitation of this methodology is that it's use of deep learning models such as LSTM and GRU as one of the main limitations of deep learning, which is normally applied to sequential data. In the case where there are no temporal dependencies in the dataset, these models might not give us significant gains over the ML models. Second, the time and computational complexity of hyperparameter tuning deep learning models may be too costly in resource constrained environments. This ap- proach, however, has the strong point that several models are used, which allows for a wide study of their performance in structured data. This research can solve by combiningML and DL model to get insight on the best way to predict water potability and make recommendation for future research focus on this field.

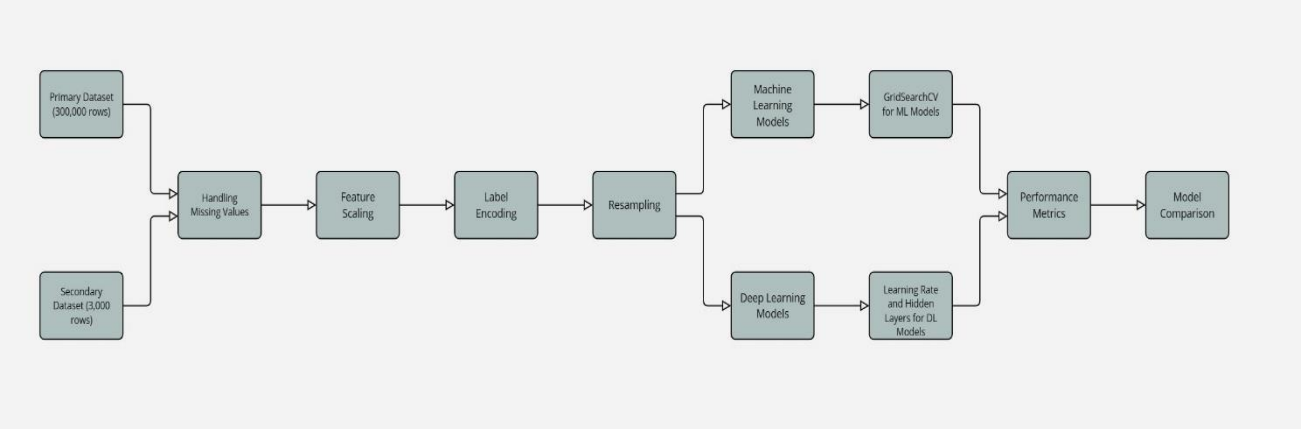


Fig 1: Workflow diagram

4 Design Specification

The design of the water potability prediction system is to incorporate the ML and DL models in the form of a scalable and efficient manner. The implementation remains at the structure level and the predictions are very strong with tabular data sets. It can be used for data preprocessing, model training, and evaluation while placing a significant focus on modularity and scalability. The system is built in Python with the use of Scikit-learn for implementing ML and TensorFlow for implementing DL. Large amount of computational power in the form of multi-core processors and GPUs is used to perform the computationally expensive operations such as training of Deep Learning models on big data. In the present design, several predictive models are applied, and all of them are customized to the specifics of the given dataset and its demands. Some ML models like Decision Trees, Random Forest, and XGBoost are incorporated because they are effective with structured data. These models do not require manual determination of features automatically and also work well with non-linear real-world relations. On the DL side, models such as LSTM, GRU, and TabNet are applied for potential pattern search in data. For instance, TabNet is specifically designed for tabular data and is a hybrid of tree-based methods and Deep Learning.

New Model/TabNet Description: Unlike existing deep learning models, e.g. LSTM and GRU, which are primarily designed for sequential data, TabNet has been engineered to recognise structured data, such as the water quality data in this study. Since it's a tree-based model with some deep learning coupled together, TabNet isa very good model for tabular problem handling when traditional neural networks might turn out to be a problem. It takes feature embeddings to learn and learn how to apply attention mechanisms to find important features and their interaction. Finally, TabNet can retain interpretability while achieving higher accuracy than other models, a useful quality for real world applications such as water quality prediction, where it is necessary for stakeholders to interpret the factors affecting predictions. The architecture of TabNet consists of a series of decision trees collaborative with deep neural networks to promote the ability of the model to learn feature interactions. Moreover, it is also built to handle missing values, as we often deal with real world data which have holes and it is important to cope with it. Using a sparse attention mechanism, the model is more efficient and interpretable than traditional neural networks.

Scalability and System Integration: Scalability was built into the system so more complex water quality parameters, or larger datasets, can easily be accommodated in the future. The system is built in a modular fashion, making it possible to accommodate new models is needed without reengineering much of the system. For instance, the system can be refined to use additional environmental data, such as weather patterns, or seasonal variations to make more precise predictions when further environmental in- formation becomes available. Modularity and scalability are emphasized to enable the system's deployment for real time water quality monitoring in municipal water supply systems as well as smaller rural settings. Additionally, the system's interpretability (via models such as Random Forest or TabNet) guarantees that decision makers can trust the predictions made by the model. The system architecture guarantees proper data management, preprocessing, and is- sues like missing values or categorical variables. It is applied to categorical variables where rows with missing values are deleted in order not to affect the results. Most of the existing models are evaluated using standard evaluation parameters which include accuracy, precision, recall, and F1-score which gives a clear insight of model performance. This design allows for accurate prediction of water quality while being scalable for larger data sets, and easily extendable in the future.

5 Implementation

5.1 About Datasets

In this study, two datasets were selected to predict water potability using different water quality attributes.

Water Quality Dataset 1: This dataset has 300,000 samples which are quite large enough to build accurate prediction models. Various attributes like pH, iron, nitrate, chloride, lead, and zinc in water, a categorical data type that includes watercolor, and metadata which includes month, day, time of day, and target variable which is water potability. Although the data is large and contains features originating from different sources, there are some gaps within some of the features like “Source” and” temperature-related” attributes that need preprocessing for handling missing values. Such a data set provides perfect ground for analyzing temporal and chemical factors, which affect water potability.

Water Quality Dataset 2: This subset has 3,200 samples with features like pH, hardness, solids, chloramines, sulfate, and potability. In contrast to the first dataset, there is an imbalance in the dataset. The smaller number of instances and the class imbalance are a problem for model development, especially for identifying patterns for the minority potable class, however, these are also the great potential for using resampling, data augmentation, or weighted modeling techniques to combat these problems effectively.

5.2 Data Preprocessing for Water Quality Dataset 1:

Null Value Handling: For the Water Quality Dataset 1, features were inspected specifically for missing values. The analysis revealed that most features included null values in the dataset. Instead of imputing missing values, rows containing the missing values were deleted. The rationale for this decision was not to contaminate the model with imputed values, which would probably distort the data distribution. The removal of rows containing null values made the dataset as close to the actual structure as possible, eliminating the influences of possibly erroneous or imputed data by the model.

Class Distribution: The dataset is balanced with the number of samples of both classes, potable and non-potable. Based on this balance, no further resampling was performed since the class distribution did not create an imbalance that could challenge model performance.

Categorical Data Encoding: To handle categorical data as “Color”, “Source” and “Month”, label encoding was used to handle them. This converted all string values to numeric form which makes data more applicable for machine learning algorithms. This encoding was done that the categorical features did not affect the learning of the model and made proper predictions.

5.3 Data Preprocessing for Water Quality Dataset 2:

Null Value Handling: For the Water Quality Dataset 2, only a few missing values were detected. These missing values were filled using the mean of their respective columns. Given the small number of missing values, this method was considered appropriate, as it maintained the overall data structure without introducing significant bias.

Class Imbalance: The main complication in this dataset was class imbalance: Class 1, or the potable class, is the minority. To overcome this problem, resampling strategies were used in this study. The samples of the minority class were randomly augmented to equal the samples in the majority class (non-potable). The majority and the minority classes were then resampled so that the model would be trained on a balanced dataset, which is important for the minority class samples.

Categorical Data Encoding: Like in Dataset 1, all the categorical variables were then transformed into numerical format using label encoding. This was done to ensure that the machine learning models did not have issues handling categorical data in that they mistake for string values.

Finally, for preprocessing, the rows with missing values were removed from Water Quality Dataset 1 since removing was not deemed appropriate; and, with a balanced class distribution, no oversampling or under sampling was performed. For Water Quality Dataset 2, the missing data was imputed using the mean and oversampling was done to handle the class imbalance problem. Categorical features in both datasets were addressed using label encoding to be compatible with the machine learning algorithms. Finally, the last stage of implementation was training and validating the models with the datasets processed and then testing models on data which is unseen. Six models were developed and trained: the same three ML models (Decision Tree, Random Forest, and XGBoost) and three DL models (LSTM, GRU, and TabNet). To get the highest possible performance out of these models, they were fine-tuned using hyperparameter optimisation. The resulting models were able to reach full pattern extraction to meaningful patterns and accurate potable water predictions. Performance metrics were then computed based on how well our predictions sat at the test set by making predictions for each model. The outputs from this stage were the trained models themselves, together with their respective evaluation report, confusion matrix and graphical representation of performance comparison.

Outputs Produced: The implementation produced several outputs that were essential for answering the research question and achieving the study's objectives: Trained Models: Six prediction models were built and trained, using traditional ML and Deep Neural Network, where XGBoost consistently performed the best for the tabular/structured data. • Performance Metrics: In each model, the measures, including accuracy, precision, recollection and F1-measure, were derived to offer an overall view of model efficiency as well. • Transformed Data: The feature scale normalization was used to normalize the features, label encoding was used for the target variable and missing values were handled. These transformations made the data suitable for the chosen algorithms and prepare to train them. • Visualization Outputs: Qualitative results included a confusion matrix of each model, feature importance for models like Random Forest and XGBoost models, and loss curve for the LSTM and GRU model. These visualizations were useful in making sense and comparing the above results.

Tools and Technologies Used: The implementation used a high-performance sincere environment to handle large datasets and the actual training of deep learning models. The tools and technologies used include: • Programming Language: Python was the main language selected as the most flexible and having the richest ecosystem of libraries for ML and DL. • ML Frameworks: In terms of executing most traditional ML models such as Decision Trees, Random Forest, and XGBoost, Scikit-learn was used, and its parameters were tuned. • DL Frameworks: TensorFlow and Keras were used for building and training the presented DL models including LSTM, GRU, and TabNet. • Visualization Libraries: Matplotlib and Seaborn libraries were used to develop the plots required in the data analysis and to quantify the performance. • Computational Resources: DL models' training was performed using multi-core processors and GPUs since these operating modes were critical for managing the fundamental dataset appropriately and training the models.

6 Discussion/ Evaluation of Models

Using *Water Quality Dataset 1* for predicting water potability, several machine learning (ML) and deep learning (DL) models were chosen to compare their effectiveness for the given task.

6.1 Evaluation Metrics

In this study, various evaluation measures for the machine learning models for water potability prediction, accuracy, precision, recall, and F1-score are employed.

- Accuracy is the measure of the total correct estimate giving a broad view of how it performs. However, in the case of class imbalance, accuracy may not be a good measure to assess the performance of a given model.

- Precision calibrates the true positive rate, that is the right classification of potable water, to the total positive, which is important to avoid misclassification of non-potable water to potable water because it poses some danger.

- Recall, in contrast, calculates the model ability to get all the samples of potable water right meaning as many safe water samples as possible will be correctly tagged.

  F1-score, the harmonic mean of precision and recall, is used to overcome the biased result between the two measures as well as is advantageous when there is a class imbalance in the data set. interpreting string values as continuous variables. These metrics combined guarantee that the developed model offers accurate and safe predictions regarding water quality.

6.2 Choice of Machine Learning Classifiers

- An XGBoost model was implemented because of its effectiveness and speed in training and making predictions on structured/tabular data. It is an optimized gradient boosting algorithm that can easily handle data imbalances which could be a major concern whenever the classes to be used, such as potable and non-potable water, are not balanced. It's useful in capturing non-linear relations and interactions between features such as pH, Iron chloride, etc. which may affect the potability of water. XGBoost can also handle missing values and noisy data and therefore a good candidate for this problem

- The Decision Tree classifier was chosen for water potability prediction, the model should determine the relations between several features, and decision trees segment the data according to these features. Such an ability to capture complex interactions is useful when deciding whether a water sample is potable or not. While the decision trees can be easily overfitting, especially on a small sample space, it is highly understandable and visualize how graphical features such as high levels of Iron or Turbidity affect the classification.

- Random Forest Classifier is particularly useful for problems such as water potability since there are several features influencing it. Also, Random Forest does not suffer from noisy or missing data and the feature importance analysis shows which variable is most important for the model to decide whether the water is potable or not

6.3 Choice of Deep Learning Models

LSTMs are built for sequential data and for capturing long range dependencies. Even though the water quality data does not possess temporal dependencies, LSTMs were contemplated as part of an exploration of the deep learning approach. Rustam et al. (2022) showed that LSTM and GRU could perform well with large size datasets for water quality prediction classification problems. Drawing from those insights, LSTM and GRU models are used in my study with the aim to explore the possibility of capturing the complex latent sequential patterns that might exist in the data. Although the data is not explicitly sequential, this dataset contains features like "Month," "Day," and "Time of Day" could potentially exhibit subtle trends or periodic variations due to seasonal factors such as heavy rains/ drought that these models are designed to detect.

- Even though these models are not ideal for non-sequential, by considering above reasons, experimented with these state-of-the-art models. In case some temporal features were not easily recognizable in water quality data (for instance, daily fluctuations or monthly trends), LSTM would pick them. However, since there are no clear time-based dependencies in the dataset, LSTM may not bring much improvement over and above other machine learning methods, though it was used to test the possibility of using deep learning for this problem. - Like LSTMs, GRUs are also considered a recurrent neural network (RNN) architecture used to train data with sequential information. They are less complex than LSTMs in terms of computational structure, which means that their

training is quicker. For this task, GRUs were considered to evaluate the deep learning methods to find out if there exist temporal patterns in the data. Like LSTM, if the dataset had temporal or sequential characteristics, it was possible that GRU could identify those characteristics. Nonetheless, since there is no clear temporal structure in the dataset, GRU may not perform better than typical machine learning algorithms, although it was

experimented with to compare deep learning methods. • TabNet is a deep learning model that is aimed at tabular data and is highly relevant to predicting water potability given chemical and physical characteristics. Despite these traditional artificial neural networks

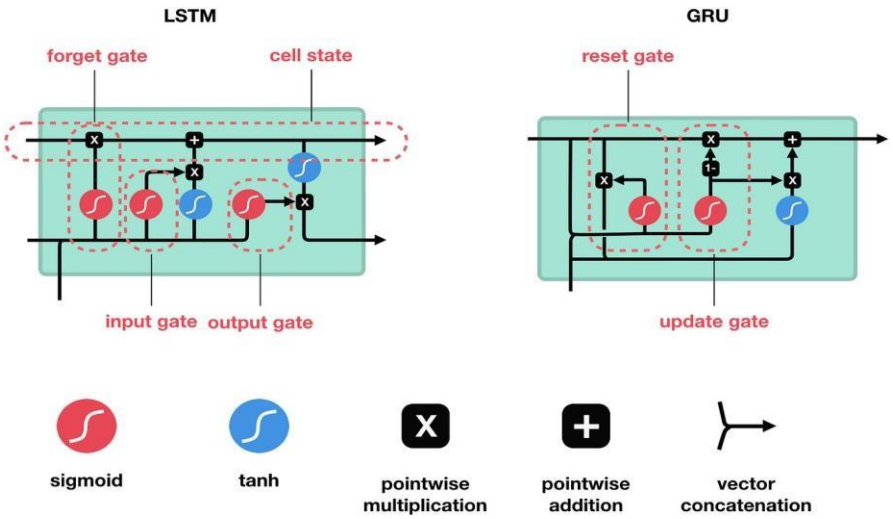


Figure 2: Architecture of LSTM and GRU

are not designed to handle structured data; this is where TabNet comes in, as it takes decision trees and adds deep learning to pick up more intricate relationships. Indeed, it can work with the missing values, and features of different types, and can discover complex interactions between features. TabNet was adopted to investigate whether it can enhance performance against standard machine learning algorithms. Its architecture is optimized for use with tabular data, and it could potentially provide for more detailed learning from the data, especially if there exist interactions between features that cannot be easily described using classical techniques

6.4 Reasoning for Model Selection

XGBoost, Random Forest, and Decision Trees were chosen because these models were previously known to work well with structured tabular data and are known to capture complex relations between the features which are ideal for predicting water potability. These models are insensitive to noise and imbalance of classes and do not demand a lot of computational capacity. Besides, they are more easily interpreted to give insights into the correlation between different water quality parameters and portability.

For this purpose, LSTM, GRU, and TabNet neural network models were considered primarily to identify the potential benefits of neural networks for this problem. These models are more appropriate for learning from large datasets and are presumed to work even better if there are interactions in the data. However, since there is no inherent sequential or temporal structure in the dataset, the deep learning models were then trained to check whether they could discover any latent interactions that other machine learning models are unable to model.

To achieve a balance between the levels of detail and complexity of the data and the accuracy of the water potability prediction the use of both machine learning and deep learning models were employed, and various models were tested.

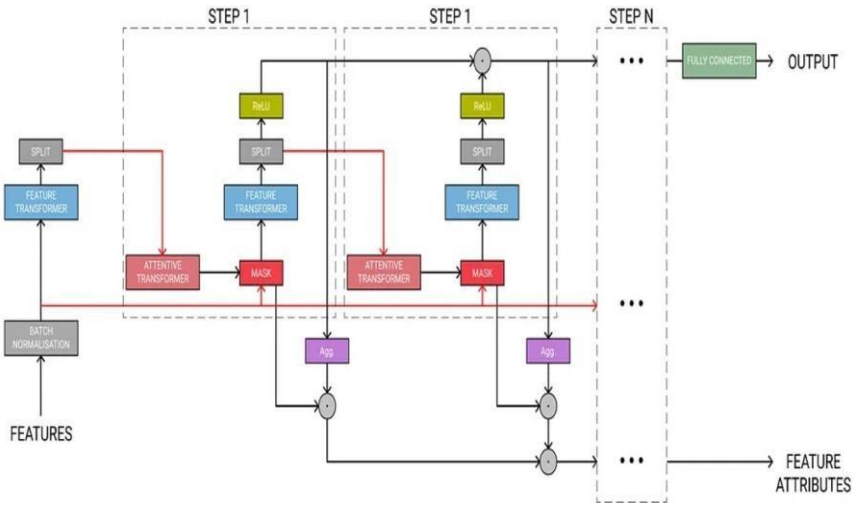


Figure 3: Architecture of TabNet

7 Evaluation and Discussion of Results

The evaluation of several ML (machine learning) and DL (deep learning) algorithms for the water potability prediction task, provides metrics, including accuracy, precision, recall, and F1-score. The discussion in the following section reveals the comparison of each of the model's performances and highlights their accuracy in predicting whether the water is potable or otherwise.

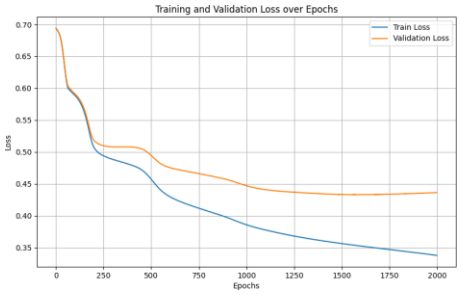
Model	Accuracy	Precision	Recall	F1-Score
Machine Learning Models				
XGBoost	**0.9708**	**0.97**	**0.97**	**0.97**
Decision Tree	0.9380	0.94	0.94	0.94
Random Forest	0.9018	0.91	0.90	0.90
Logistic Regression	0.6548	0.65	0.65	0.65
Deep Learning Models				
LSTM	0.8314	0.83	0.83	0.83
GRU	0.8186	0.82	0.82	0.82
TabNet	0.8416	0.84	0.84	0.84

Table 1: Model Performance Comparison for Water Potability Prediction (for all the metrics higher the values better the model). All the values reported will be in range 0 and 1

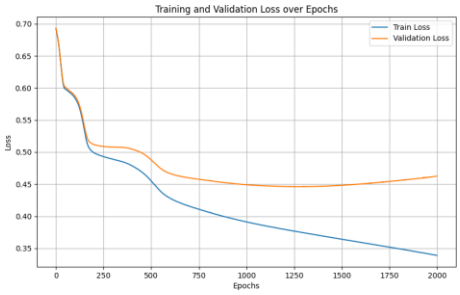
7.1 Machine Learning Models

7.1.1 XGBoost Classifier:

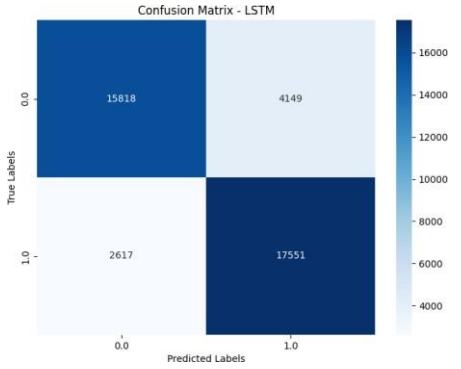
XGBoost model has performed efficiently, by achieving an accuracy of **97.08%**; for the other three metrics, the result was **0.97** of precision, recall, and F1-score. These statistics show that the model has a high level of reliability in repetitively recognizing both potable and non-potable water samples. This makes XGBoost very strong because it can



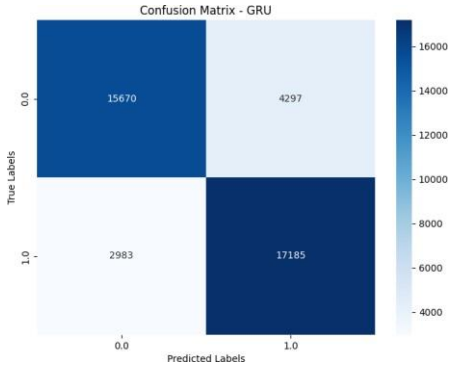
(a) Training and validation loss of LSTM



(b) Training and validation loss of GRU



(c) Confusion matrix of LSTM model



(d) Confusion matrix of GRU model

Figure 4: Training and validation loss of LSTM and GRU models with Confusion matrices

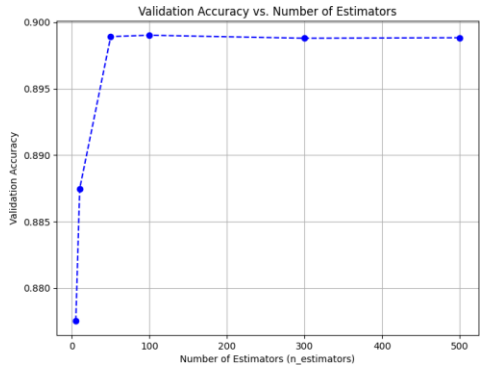
also handle non-linear relations, missing features, and instances of class imbalance. Its performance demonstrates why it remains one of the go-to models for structured, tabular data and why it is the best model for the water potability prediction task.

7.1.2 Decision Tree Classifier:

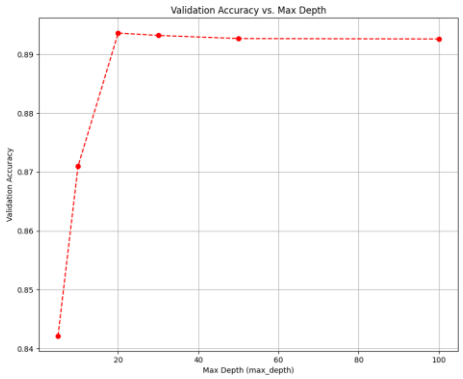
The Decision Tree model had the highest reliability with weighted-average precision, recall, and F1 measure of **0.94** and accuracy of **93.80%**. Its main strength is, therefore, the ability to provide clear and easy-to-interpret results in terms of the factors affecting water potability. Although less accurate than the XGBoost model, the Decision Trees are easier to interpret to discover feature importance and to make a fast decision. However, the model is not very robust since it over splits data and can be highly sensitive to small datasets which hampers its' generalization ability.

7.1.3 Random Forest Classifier:

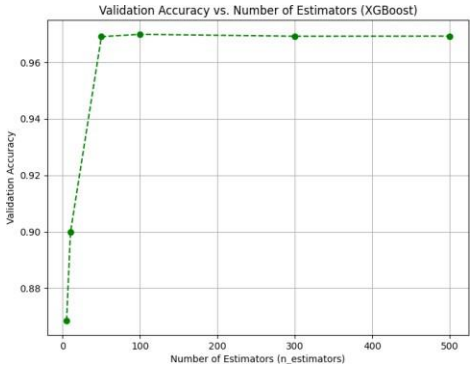
Random Forest acquired an accuracy of **90.18**, and the precision, recall, and F1 score equaled **0.91, 0.90, 0.90**. The Random Forest is overfitting resistance and has a better performance in capturing various feature interactions. Nonetheless, it is a little slower than XGBoost, which might be attributed to the inadequate optimization of hyperparameters. However, it still stands as a stable model for medium-sized datasets and can be enhanced using fine-tuning and sampling approaches.



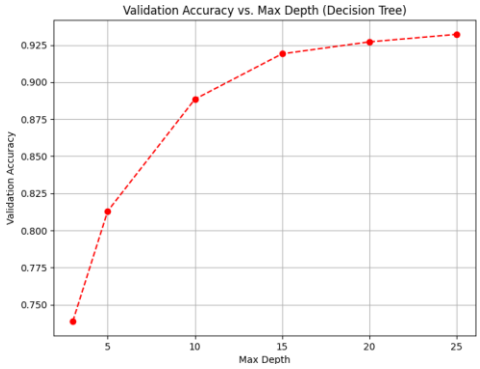
(a) Validation accuracy vs Number of estimators on validation data for random forest model



(b) Validation accuracy vs maximum depth on validation data for random forest model



(c) Validation accuracy vs Number of estimators on validation data for XGBoost



(d) Validation accuracy vs maximum depth on validation data for random forest model

Figure 5: Plots of hyper parameter tuning on various model

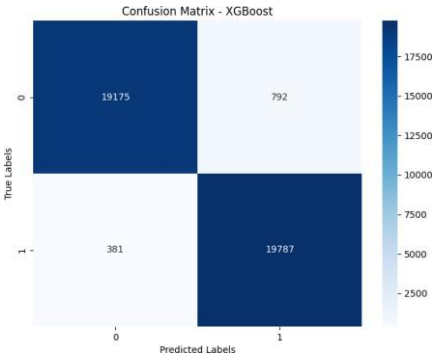
7.1.4 Logistic Regression:

Logistic Regression was the least effective, achieving accuracy, precision, recall, and F1-score of **0.65** each. The model making of Logistic Regression is a straight line and found unfit to model the non-linearity present in the data set, a weakness for this problem. It is also used mainly to determine a reference measure for comparing the execution of a process within an organization.

7.2 Deep Learning Models

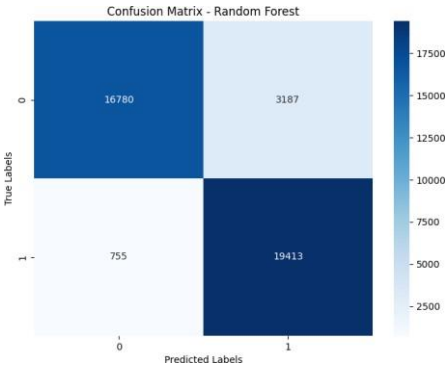
7.2.1 Long Short-Term Memory (LSTM):

The LSTM model provided a weighted average of precision, recall, and the F1 score of **0.83** and the accuracy of **83.14 %**. Though it performs slightly better than GRU, the accuracy is far below the standard machine learning algorithms such as XGBoost and Decision Trees. LSTM has this sequential property that does not go well with the current data set because it is not temporally divided. This moderate performance indicates that LSTM can be quite inefficient for non-sequential tabular data.

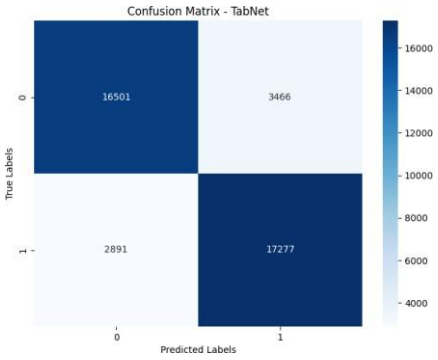


(a) Confusion matrix of XGBoost model

(b) Confusion matrix of Decision tree model



(c) Confusion matrix of Random Forest model



(d) Confusion matrix of TabNet model

Figure 5: Plots of confusion matrices on various model

7.2.2 Gated Recurrent Unit (GRU):

The GRU model was slightly worse than LSTM and its accuracy was **81.86** with such specificity, sensitivity, and F1 measure as **0.82**. Like LSTM, GRU addresses temporal data which makes it unsuitable to model non-temporal relations in the dataset. Comparatively, GRU is less time-consuming than LSTM; however, its efficiency shows that using recurrent architectures in non-recurring problems is not without difficulty.

7.2.3 TabNet:

TabNet achieved higher performance among the deep learning models with an accuracy value of **84.16 %**, precision, recall value, and F1-score of **0.84**. Due to its specific design for processing tabular data, it was able to encode many cross-feature interactions effectively. However, it is still lower than the results of a standard ML algorithm, XGBoost, which means that classical algorithms are still better for this dataset.

7.3 Machine Learning Models Outperform Deep Learning Models

The results prove that the efficiency of ML (machine learning) models is significantly higher than the efficiency of DL (deep learning) models for water potability prediction. The baseline models of XGBoost and Decision Trees provided a significantly better accuracy of the models and even the parameters of precision, recall, and F1-score balanced the results.

The main reason for such differences is the disparity of the given dataset. The water quality data is numeric, two-dimensional, and has no temporality or spatiality component. Such datasets are best trained with Classical ML models since these are built to work with non-linear data and can work with noisy or even incomplete data. Of these, the XGBoost gradient boosting approach offers an effective tool to improve performance making it the leader.

In contrast, deep learning models such as LSTM and GRU are designed for a sequential or time series dataset which the water quality does not possess. Therefore, these models failed to define some worthwhile patterns and had a considerably lower predictive accuracy. Slightly outperformed by TabNet, which is designed for tabular data, it was clear that XGBoost was once again surpassing DL models due to their need for more complex data with complex patterns.

This comparison shows how crucial it is to choose models according to the datasets, based on their characteristics. DL models are very useful for some specific data types (image, text, or time-series data), but there can be no big improvements for structured tabular data, where XGBoost already shows excellent performance.

7.4 Results on water quality dataset 2

The second dataset reveals that even though various resampling methods have been applied to address the problem of class imbalance; the final performance of the model is not quite satisfactory because of the scarcity of samples which range from 3000 samples. Logistic Regression again proved rather weak as it only yielded 50% accuracy due to the high level of difficulty of the overloaded and imbalanced data set. This was even worse than Random sampling, though at a slightly higher figure of 66.6% Random Forest was still unable to provide accurate predictions. The same can be said about XGBoost which achieved an accuracy of 63.9% and despite having an opportunity to outperform Random Forest in most cases, it failed to do that and to tackle the problems resulting from the minority class.

The other deep learning models showed similar performance with a test accuracy of 66.50%. Yet, the models were unable to fully utilize the resampled dataset: the performance of the proposed approach slightly outperformed other machine learning methods. These results suggest that resampling techniques can help to reduce im- balance to some extent but they are not able to overcome problems with the dataset and perhaps feature engineering or using advanced methods specifically for imbalance will be necessary. Further hyperparameter optimization of all the models, including XGBoost, Random Forest, Logistic Regression, and Decision Trees, revealed that, in general, the outcomes differ only slightly for the same reason – the smaller number of samples in the sample set. Among the tuning parameters of XGBoost, the numeracy of estimators, learning rate, and maximum depth were chosen, while for Random Forest, the estimators,

tree depth, and splitting criteria were selected randomly. For the first choice of features, Logistic Regression was optimized for its regularization parameter, and Decision Trees for the maximum depth of decision trees. Still, due to a relatively small training sample size, there was not enough evidence for the models to generalize, proving the idea that even the best hyperparameter selection cannot solve the problems arising from small-sample data.

8 Hyperparameter tuning

Hyperparameter tuning is crucial when designing machine learning and deep learning models because it involves selecting the right combination of parameters. In the Decision Tree classifier, the parameter was set as 'max_depth' to manage the intricacy of the tree. The grid search applied a range from 3 to 25 in order not to overfit and, at the same time, not to underfit. For 'XGBoost' the most important hyperparameter was the 'n_estimators', which gives the number of boosting stages to be conducted and was varied from 5 to 500. This tuning made it possible to achieve enough boosting rounds to capture intricate patterns without overfitting. Other works carried out included tuning of Logistic Regression for the regularization parameter C ranging from 0.01 to 100 to have control over complexity and accuracy.

The LSTM, GRU, and TabNet models were all trained with different parameters, and the best results are presented here. Both LSTM and GRU models had two layers, a hidden size of 64, and a learning rate of 0.001. These settings were chosen to provide a sufficient capacity for learning patterns while avoiding the model's complexity. The TabNet model was trained for 50 epochs with early stopping and stopped at the 43rd epoch to avoid overfitting the model. For Random Forest, the parameters that were tuned are 'n_estimators ', 'max_depth', and 'criterion' These parameters are on the number of trees, depth of a single tree, and the split criterion respectively. In the tuning of each model, an effort was made to optimize both the computation time and the overfitting of the model.

9 Conclusion and Future work

9.1 Conclusion

This research presents an empirical analysis that compares ML (machine learning) and DL (deep learning) in predicting water potability. Based on the comparison of the presented models, XGBoost can achieve the highest accuracy and stability in processing structured data in comparison with the ML models, namely, Random Forest and Decision Trees, and DL models, namely, LSTM and GRU. ML approaches are known to excel in managing non-linear association, formatted table format data, and missing values making such tasks such as water potability prediction feasible. On the other hand, DL models failed to perform well because the target dataset does not possess temporal characteristics which are the core competency of most DL models. The results of the second experiment using a second, less balanced dataset reiterated these observations as even when applying extensive resampling and hyperparameter tuning, the models simply could not perform well. These findings have highlighted how computational approaches can revolutionize

public health and water resource management by providing easily scalable and highly accurate methods for tracking water quality. This study advances Patel et al, (2023) research by utilizing a larger and more diverse dataset that includes more attributes like Water Source, Water Temperature, and Air Temperature, making it more realistic water quality dataset compared to prior studies. A key contribution is also like using TabNet, an advanced deep learning model specifically optimized for tabular datasets. This study demonstrates TabNet's ability to handle structured data more effectively than traditional deep learning models, highlighting its potential for tasks involving non-sequential tabular data. By comparing traditional models like XGBoost, Random Forest, and Decision Trees with deep learning models on water datasets, this study concludes that ML models outperformed DL models while also showing scope in research on more DL models to enhance the performance.

9.2 Future Work

Future studies can extend this work on the development of complex models that combine the interpretability of the ML approaches with the flexibility and adaptability of the DL structures. For instance, models that incorporate decision tree algorithms in collaboration with the neural networks, like tree-based embeddings or attention mechanisms could be able to detect the intricate feature interaction while at the same time remaining interpretable. In addition, the generalization and reliability of the predictions can be enhanced if more diverse and more real-time data from water sources such as rivers, lakes, and groundwater are included in the study.

Another potential direction is the application of unsupervised learning techniques for feature extraction and potentially helpful in minimizing the amount of preprocessing required. AutoML (Automated machine learning) also has great potential to ease processes of hyperparameters tuning and model selection, which is now consuming much computational time and effort to find the best configurations. Possible approaches to address the issues of smaller and imbalanced data in the second dataset of this study would be to consider new resampling techniques or synthetic data simulations. Moreover, integrating real-time data pipelines into these models could enhance the possibility of producing dynamic predictions, as well as enhance their practical applicability for large scale in public health and water management systems. Such developments could extend toward enhancing the precision, effectiveness, and interactiveness of computational models in approaching essential issues.

References

1. Abuzir, S. Y. and Abuzir, Y. S. (2022). Machine learning for water quality classification,
2. *Water Quality Research Journal* **57**(3): 152–164.

3. Ainapure, B., Baheti, N., Buch, J., Appasani, B., Jha, A. V. and Srinivasulu, A. (2023). Drinking water potability prediction using machine learning approaches: a case study of indian rivers, *Water Practice & Technology* **18**(12): 3004–3020.

4. Alnaqeb, R., Alrashdi, F., Alketbi, K. and Ismail, H. (2022). Machine learning-based wa- ter potability prediction, *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, pp. 1–6.

5. Alnaqeb, R., Alrashdi, F., Alketbi, K. and Ismail, H. (2023). Machine learning-based water potability prediction.

6. Baliyan, N., Jain, S. and Sharma, N. (2024). Comparative study of the state-of-the-art machine learning algorithms in predicting potability of drinking water, *International Conference on Deep Learning and Visual Artificial Intelligence*, Springer, pp. 353–366.

7. Chakravarthy, S. R., Bharanidharan, N. and Venkatesan, K. (2023). Prediction of water quality using softmax-elm optimized with adaptive crow-search algorithm, *IEEE Access***11**: 96.54.

8. Dalal, S., Onyema, E. M., Romero, C. A. T., Ndufeiya-Kumasi, L. C., Maryann, D. C., Nnedimkpa, A. J. and Bhatia, T. K. (2022). Machine learning-based forecasting of pot- ability of drinking water through adaptive boosting model, *Open Chemistry* **20**(1): 816–828.

9. Dawood, T., Elwakil, E., Novoa, H. M. and Delgado, J. F. G. (2021). Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks, *Journal of Cleaner Production* **291**: 125266.

10. Gao, Z., Chen, J. and Wang, Q. (2023). A novel multivariate time series prediction of crucial water quality parameters with lstm networks, *Journal of Contaminant Hydrology* **259**.

11. Ghosh, H., Tusher, M. A., Rahat, I. S., Khasim, S. and Mohanty, S. N. (2023). Water quality assessment through predictive machine learning, *International Conference on Intelligent Computing and Networking*, Springer, pp. 77–88.

12. Gill, K. et al. (2023). Water quality prediction using xgboost, random forest, and gaus- siannb, *IEEE GCITC* .

13. Guo, Z. et al. (2023). Cnn-bilstm for water quality prediction with attention mechanism,
14. *IEEE SWC* .

15. Khan, Y. and See, C. S. (2024). Predicting and analyzing water quality using machine learning.

16. Liu, L., Zhao, Z. and Wang, X. (2024). A water quality prediction method based on lstmneural network optimized by cuckoo search algorithm.

17. Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S. and Ratna, R. (2022). A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable ai, *Computational Intelligence and Neuroscience* **2022**(1): 9283293.

18. Rustam, F., Ishaq, A. and Kokab, S. T. (2022). Artificial neural network model for water quality prediction, *MDPI Water* **14**(21): 3359.

19. Saroja, Haseena and Dharshini, S. (2023). Deep learning approach for prediction and classification of potable water, *Analytical Sciences* **39**(7): 1179–1189.

20. Sreekumar, V., Ihsan, F., Reghuram, S. and Sarath, S. (2024). A detailed analysis of machine learning models to predict water potability, *2024 15th*

International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, pp. 1–6.

21. Stepanova, E., Orlov, V., Kukartsev, V., Pinchuk, I. and Suprun, E. (2024). Machine learning approaches for water potability prediction: Addressing class imbalance with smote, *BIO Web of Conferences*, Vol. 116, EDP Sciences, p. 03024.
22. Wu, Z. and Wang, Z. (2022). Hybrid artificial neural network-wavelet-transform approach for water quality prediction, *Journal of Environmental Management*, 313, p. 115000.
23. Zhao, Y. (2021). Active learning and feature selection for enhancing SVM results in water quality prediction, *Environmental Modelling & Software*, 137, p. 104935.
24. Zhang, X. and Jin, Y. (2024). Autoencoder-LSTM-based water quality prediction for seasonal and spatial variations, *Journal of Water Research*, 198, pp. 102304.
25. Xu, Y., Li, H. and Wang, J. (2019). Seasonal water quality prediction using SARIMA-LSTM, *Environmental Science & Technology*, 53(23), pp. 13523–13532.
26. Guo, L., Zhang, Q. and Huang, F. (2023). Enhancing river water quality prediction using CNN-BiLSTM with attention mechanisms, *IEEE Transactions on Neural Networks and Learning Systems*, 34(5), pp. 2235–2244.
27. Liu, J., Wang, R. and Zhang, W. (2024). A hybrid AE-LSTM model for predicting water quality in river basins, *Environmental Pollution*, 311, p. 120623.
28. Chakravarthy, S. R., Bharanidharan, N. and Venkatesan, K. (2023). Adaptive crow-search optimized water quality prediction model using Softmax-ELM, *IEEE Access*, 11, pp. 65034–65047.
29. Stepanova, E., Orlov, V., Kukartsev, V., Pinchuk, I. and Suprun, E. (2024). Addressing class imbalance with SMOTE for water potability prediction, *BIO Web of Conferences*, 116, p. 03024.
30. Suleiman, A. A., Yousafzai, A. K. and Zubair, M. (2023). Comparative analysis of machine learning and deep learning models for groundwater potability classification, *Engineering Proceedings* **56**(1): 249.
31. Zhang, H. and Jin, K. (2024). Research on water quality prediction method based onae-lstm.
32. Zhao, R. (2021). Water potability prediction with svm and ann, *ICBAR* .
33. Zhu, M., Wang, J. and Yang, X. (2022). A review of the application of machine learning in water quality evaluation, *Eco-Environment & Health* **1**: 107–116.
34. Whelan, S. (2022). Predicting river water quality parameters using supervised ml techniques: Uk.
35. Wu, J. and Wang, Z. (2022). Hybrid ann-wt-lstm model for water quality prediction, *Water* **14**: 610.
36. Xu, R. et al. (2019). Sarima-lstm for mangrove ecosystem water quality prediction, *IEEEHPCC* .