Interpretable Groundwater Quality Classification Using CatBoost and SHapley Additive exPlanations (SHAP)

A *Dissertation*

Submitted in partial fulfillment of the requirement for the award of degree of

MASTER OF TECHNOLOGY
in

COMPUTER SCIENCE AND ENGINEERING

Submitted to

RAJIV GANDHI PROUDYOGIKI VISHWAVIDHYALAYA,

BHOPAL (M.P.)

Submitted by

AVNI YADAV

ENROLLMENT NO.- 0546CS20MT04

Under the Guidance of

PROF. Jayshree Boaddh

Session: Jan -2026

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VAISHNAVI INSTITUTES OF TECHNOLOGY AND SCIENCE BHOPAL

VAISHNAVI INSTITUTES OF TECHNOLOGY AND SCIENCE BHOPAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Session: Jan - 2026

CERTIFICATE

This is to certify that the work embodies in this dissertation entitled " **Optimizing Ground Water Quality**" is the Bonafide research work carried out independently by **Avni Yadav (0546CS22MT04)** ,student of "**Master of Technology (Computer Science And Engineering)"** from Rajiv Gandhi Proudyogiki Vishwavidhyalaya Bhopal for the partial fullfilment of the requirements for the award of Degree OF Master Of Technology in Computer Science and Engineering ,and this dissertation has not formed previously the basis for the award of any degree ,diploma ,associate-ship, fellowship or any other similar ,to the best of out knowledge.

Guided by:

Prof. Jayshree Boaddh
HOD,CSE Department
VITS,Bhopal

Director/Principal

VAISHNAVI INSTITUTES OF TECHNOLOGY AND SCIENCE BHOPAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Session: Jan 2026

CERTIFICATE OF APPROVAL

This dissertation work entitled " **Optimizing Ground Water Quality "** being submitted by **Avni Yadav is** approved for the award of degree "**Master of Technology (Computer Science And Engineering)"** for which it has been submitted Rajeev Gandhi Prodyogiki Vishwavidhyalaya,Bhopal.

(Internal Examiner) **(External Examiner)**

Date: **Date:**

DECLARATION

I here by declare that the work, which is being presented in the dissertation entitle "Optimizing Ground water Quality"partial fulfillment of the requirementsfor the award of degree of **Masters of Technology** in **Computer Science and Engineering** submitted in the Department of **Computer Science And Engineering** in **Vaishnavi Institute of Technology And Science, Bhopal (M.P.).**Is an authentic record of my own work carried under the guidance of **Prof. Jayshree Boaddh.** I have not submitted the matter embodied in this report for award of any other degree.

I also declare that ..A check for Plagiarism is been carried out on the dissertation and is found within the acceptable limit and report of which is enclosed herewith..

AVNI YADAV
0546CS20MT04

DECLARATION ON PLAGIARISM

This report ensures that the work, which is being presented in the dissertation entitled **"Optimizing Ground Water Quality" is** partial fulfillment of the requirements for the award of degree of **Masters of Technology** in **Computer Science And Engineering is an authentic record of my own work carried out under the guidance of Prof. Jayshree boaddh, CSE Department. I have not** submitted the matter embodiesin this report for award of any other degree .I also declare that a check for plagiarism has been carried out and the dissertation is found within the acceptable limit ,report of which is enclosed herewith.

Prof. Jayshree Boaddh
HOD,CSE Department
VITS, Bhopal
Verified
Plagiarism

Director/Principal
Signature with seal

Avni yadav
0546CS20MT04

ACKNOWLEDGEMENT

At the outset I would like to express my deep sense of gratitude to my respected dissertation guide **Prof. Rahul Patidar** under whose able guidance my project saw the light of the day. I will be ever indebted to him for his morale boosting, support and encouragement.

I would like to thanks to Pro. Rahul Patidar , AP of computer Science & Engineering Department VITS Bhopal for his valuable guidance and motivation.

I extend my sincere thanks to **Prof. Jayshree Boaddh Head of the CSE Department VITS Bhopal**. It would have been impossible to complete the project work without her help. SHe made available all the facilities for the successful completion of the project.

I am also thankful to all department staff and fellow colleagues who have immensely helped and encouraged me to complete the project in time.

Finally, I am also very grateful to our Principal **Prof. Jayshree Boaddh**, and Group Director **Dr. Rajesh Kumar Yadav** for making all sort of technical and official work easy going for me on account of which I could complete this work.

AVNI YADAV

Enrollment No- 0546CS20MT04

प्रपत्र

मैं, अवनी यादव पुत्री श्री हेमराज यादवआयु 27 वर्ष, निवासी वार्ड क्रम 1बिशांखेड़ा ओबेदुल्लागंज, तहसील गोहरगंज, जिला रायसेन (म.प्र.) 464993की होकर शपथ पूर्वक निम्न कथन करती हूँ, कि:-

1. **यह कि मैं ने एम.टेक के विषय कंप्यूटर साइंस एंड इंजीनियरिंग सत्र 2020 में काउसलिंग के माध्यम से श्रेणी ओबीसी से वैष्णवी इंस्टीट्यूट ऑफ टेक्नोलॉजी एंड साइंस भोपाल, संस्था मे प्रवेश लिया था।**
2. **मैं 2020 से नियमित छात्रा के रूप में स्नातकोत्तर पाठ्यक्रम में अध्ययनरत थी ।**
3. **मैं घोषणा करती हूँ कि इस पाठ्यक्रम कि अवधि में किसी भी निजी अन्य क्षेत्र के संस्थान पूर्णकालीन रूप में कार्यरत नही थी।**

हस्ताक्षर शपथ ग्रहिता

गाइड एवं प्राचार्य द्वारा सत्यापित

हम सत्यापित करते हैं कि छात्रा का नाम अवनी यादव नामांकन क्रमांक 0546CS20MT04 द्वारा उपरोक्त भरी गई जानकारी प्रमाणित एवं सही है।

गाइड के हस्ताक्षर

संचालक / प्राचार्य

हस्ताक्षर / पदनाम सील सहित

संस्था वैष्णवी इंस्टीट्यूट ऑफ टेक्नोलॉजी एंड साइंस भोपाल

संस्था का कोड_____

दूरभाष क्रमांक_____

दिनांक:-

TABLE OF CONTENTS

List of Tables

List of Figures

Interpretable Groundwater Quality Classification Using CatBoost and SHapley Additive exPlanations (SHAP)

ABSTRACT

Groundwater quality evaluation is a prerequisite for the sustainable utilization of agricultural and drinking water in the areas where water sources are getting overexploited. The current study presents a framework that is based on machine learning for classifying groundwater suitability utilizing hydrochemical parameters that were obtained from post-monsoon groundwater samples taken in various districts of Telangana, India. The original multi-class irrigation suitability labels were reformulated into a binary classification problem to distinguish between suitable and unsuitable groundwater. A CatBoost classifier was used because of its capability to manage categorical variables as well as to synergize complex feature interactions. Various model performance measures including ROC-AUC, accuracy, F1-score, and confusion matrix analysis were utilized to evaluate the model. The model reached a test accuracy of 97.30% and an F1-score of 0.9375, meaning the predictive power and the balanced classification performance were indeed strong. Model interpretability was assessed using SHapley Additive exPlanations (SHAP) and permutation importance, which pinpointed the common hydrochemical parameters that have the greatest impact on groundwater suitability. The findings affirm that the proposed method is a reliable, interpretable, and pragmatic framework for the classification of groundwater quality and the decision support system in water resource management.

Keywords: *Groundwater quality, Machine learning, CatBoost, Irrigation suitability, Binary classification, SHAP.*

CHAPTER 1 INTRODUCTION

1.1 Background of Groundwater Resources

Groundwater is labeled as one of the most vital sources of freshwater. The support is not only limited to human life but also including agricultural and industrial development as well as ecological sustainability in many areas of the world. Groundwater is the water found under the Earth's crust in soil pores and rock fractures, which means that it is naturally protected and relatively stable compared to the surface water bodies like rivers, lakes, and reservoirs. Its subsurface storage means that it is less subjected to the direct impact of water loss through evaporation and it is also less affected by short-term climatic fluctuations, thus making it the region's most suitable water source, especially in the case of irregular rainfall patterns or drought situations.

On a global scale, groundwater accounts for almost half of the total drinking water needs of the human population and approximately forty percent of the entire irrigation water used in agriculture. In countries that are still developing like India, groundwater is the main source of supply for rural areas and agricultural fields. The significant socio-economic development that has been taking place in the last few decades marked by the increase in population, urbanization, and industrialization as well as the use of more agricultural inputs has increased the dependency on groundwater resources immensely. Groundwater extraction has become the primary way to meet the increasing demand as surface water becomes less available and less certain.

Groundwater resources are very important, and yet they are facing big problems both in terms of quantity and quality. The over-extraction of groundwater beyond the natural recharge rates has led to the decline of water tables in many parts of the world. This depletion not only puts the question of sustainability of the groundwater supplies at risk, but it also has a negative effect on the quality of water. One of the factors that contribute to this problem is the increase in the concentration of dissolved minerals as groundwater levels go down, and this, in turn, leads to salinity problems that affect both the quality of drinking water and the suitability of land for agricultural activities and the health of ecosystems. Groundwater pollution, unlike the surface water pollution, which is usually obvious and very quickly detectable, is a slow and hidden process. The underground characteristics of aquifers let the pollutants remain undetected for a long time, which of course,

makes it extremely hard and very expensive to carry out remedial actions after pollution has taken place. Hence, the early detection and the very reliable assessment of groundwater quality become very necessary.

The decline in the quality of groundwater is the result of a mix of both natural and human factors. The natural processes such as mineral dissolution, rock-water interactions, and geochemical evolution, among others, determine the baseline composition of groundwater. However, the human activities nowadays are significant contributors to the contamination of groundwater. Agriculture is a significant contributor to the problem with the chemical-heavy runoff that carries fertilizers and pesticides, and industrial discharges, urban sewage, landfill leachate and improper waste disposal are all responsible for drastic changes in groundwater chemistry. These pressures make it essential to develop advanced monitoring and predictive tools that will be able to support the sustainable groundwater management and informed decision-making.

1.2 Importance of Regional Groundwater Quality Assessment

One of the main factors that determine the quality of groundwater is its suitability for different uses such as irrigation and livestock consumption. In agricultural areas, groundwater is frequently applied to fields without any major treatment. The use of contaminated irrigation water can lead to the destruction of the soil, and loss of crops, and eventually the degradation of the land. In addition, the quantity and quality of groundwater determine the lifespan of the livestock. If the dissolved solids or the chemical constituents are too high, it may lead to health issues and lower productivity besides causing health disorders or lowering the productivity of the animals.

The spatial and temporal variability of the regional groundwater systems is substantial and is mainly influenced by the geological and soil characteristics, land use, and climatic conditions. This means that the different regions of groundwater quality are very complex and move gradually. In areas that are highly dependent on agriculture and groundwater supply, such variability impacts directly on crop planning, irrigation, and resource allocation. Therefore, systematic groundwater quality assessment and prediction at the regional level are necessary to achieve sustainable use.

Groundwater quality classification schemes are the main tool for categorizing water based on chemical characteristics and the potential use. Classification systems, which include salinity, sodium hazard, residual sodium carbonate, and total dissolved solids, allow for the identification

of water that is suitable, marginal, or unsuitable for irrigation and livestock use. Availability of accurate classification provides a basis for practical decision-making by farmers and water managers. Thus, they can decide on the type of crops, soil management practices, or groundwater use restrictions where necessary.

Given the increasing pressure on groundwater resources, there is a critical need for data-driven approaches that can analyze complex groundwater quality datasets and generate reliable predictions. Such approaches can assist in identifying vulnerable regions, understanding dominant influencing factors, and supporting proactive groundwater management strategies.

1.3 Limitations of Conventional Groundwater Quality Assessment

Conventional methods for groundwater quality assessment traditionally consist of taking samples for laboratory analysis of individual physicochemical parameters and subsequently comparing the results with the prescribed national or international standards. These methods, though very accurate, are still afflicted with several limitations when it comes to their application in large and complex groundwater systems because they only give measurements for specific parameters and locations at specific moments.

One of the most important drawbacks of the traditional assessment methods is their inability to represent the total impact of several parameters that interact with each other. Groundwater quality is determined by a set of factors that are quite complicated and that take non-linear forms among the constituents of the chemical stuff. Evaluating parameters one by one may not reveal the true quality of the water, especially when several factors together determine the case for irrigation or livestock use. Moreover, the availability of large datasets concerning groundwater quality is increasing, but traditional methods cannot efficiently deal with high-dimensional data because they are not designed that way.

Another limiting aspect is the lack of predictiveness. Where traditional methods give a picture of the groundwater quality at a particular time, these methods tell nothing about future conditions or trends. This characteristic makes it difficult for policymakers and water suppliers to implement preventive measures or foresee possible deterioration of water quality. Acompanying this, many classical statistical methods are based on the assumption of linear relationships among variables;

this is an unrealistic assumption in the case of groundwater systems that are subject to complex hydrogeochemical processes.

These limitations highlight the necessity for advanced analytical methods that can integrate multiple parameters, accommodate nonlinear relationships, and provide predictive insights to support sustainable groundwater management.

1.4 Data-Driven Groundwater Quality Classification

The adoption of data-driven approaches for groundwater quality evaluation has been made possible by recent advancements in data availability and computation. Data-driven classification models, as opposed to deterministic thresholds or subjective weighing schemes, utilize the observed patterns in historical data to classify different types of groundwater according to their quality.

Groundwater quality datasets usually consist of many chemical parameters, and often more than one of those parameters are interrelated and/or have nonlinear interactions. ML-based classification frameworks are the most appropriate ones for this type of data because they are able to infer complex relationships from the data without imposing any physical or chemical assumptions. Hence, data-driven models can effectively map the multidimensional quality measurements of groundwater into the quality classes that are relevant and can be applied in practice.

The use of binary classification frameworks which classify groundwater into large classes of suitable and unsuitable, is a very realistic way of handling things in the field. The classification of the water quality information is simplified thereby the essential distinctions that have to do with irrigation and livestock use are still retained. The da-ta-driven binary classification models can accurately assist groundwater management in the region by pointing out the areas where water quality is of little risk and the areas where caution or intervention is required.

1.5 Application of Machine Learning in Groundwater Quality Prediction

Environmental monitoring data becoming more and more accessible, alongside the enhancements of computational power, has brought the utilization of machine learning techniques in groundwater quality prediction to the forefront of the technology. The capability of machine learning algorithms

to unveil hidden patterns and relationships in massive datasets, so to be more fitted for the modeling of groundwater systems, which are typically defined by nonlinear behaviors and complex interactions among the variables.

The prediction of groundwater quality through machine learning models has the chemical parameters of the water integrated all at once, so that the suitability of the water can be assessed in a comprehensive manner. Besides, machine learning methods have been proven to be more efficient than traditional statistical techniques in the aspects of noisy data handling, mixed data type accommodation, and nonlinear dependencies capturing. The advantages mentioned earlier have been the driving force for the growing popularity of machine learning methods in environmental modeling and water resource studies.

Ensemble-based algorithms, among the various machine learning techniques, are the ones that have been highly appreciated due to their capability to yield robust and predictive performance in the case of structured tabular data. To reduce the variance and enhance the generalization, the model combines several decision structures. Also, the current machine learning frameworks are increasingly focusing on interpretability, which allows the researchers to understand the individual features that have contributed to the predictions of the model. Interpretability is of utmost importance in the environmental applications, where transparency and trust in the model outcomes are necessary for the policy and management decisions.

1.6 Emergence of Data-Driven and Machine Learning Approaches

The technological progress in the field of environmental monitoring and data storage made it possible to have huge, multivariate groundwater quality datasets collected from various places and times. The presence of such data has led to the use of machine learning methods for groundwater quality evaluation and forecasting. Such models can recognize complicated patterns in the past data and also work with nonlinear relationships among variables without any strict statistical assumptions.

In studies on groundwater quality, machine learning techniques have outperformed traditional statistical models significantly, especially in the case of heterogeneous and high-dimensional data. These techniques can concurrently consider a variety of chemical parameters in the process of making predictions and can even do so reliably despite the presence of noise and missing values.

Thus, machine learning has become a strong and effective tool for, among others, groundwater quality classification and decision support.

On the other hand, machine learning models have their downsides, such as the necessity of thorough data preprocessing and the possibility of loss of interpretability. In environmental models, the openness of the model is of utmost importance because the stakeholders need clear and easy to understand explanations of why some groundwater samples are labeled as good or bad. This is the reason, the groundwater quality studies of the present day more and more often resort to the machine learning models that possess both high predictive performance and good interpretability, thus allowing accurate classification and at the same time giving meaningful insight into the underlying factors that drive the process.

1.7 Need for a Robust and Interpretable Classification Framework

A proficient groundwork quality forecasting system should have the ability to convert complex multivariate data into simple and usable classifications. Instead of depending only on continuous index values or individual parameter thresholds, classification-based approaches provide a more practicable way of categorizing groundwater into the two widely suitable and unsuitable classes. Such binary classification systems become an easier option for decision-making while still keeping the necessary information concerning irrigation and livestock usage.

A strong classification system must be able to consider both the quality parameter nature and the type being either categorical or numerical. A lot of the groundwater datasets are of the mixed type which consists of both the continuous chemical concentrations and the categorical descriptors. Models that can deal with such mixed data as a whole without going through a lot of manual encoding are highly beneficial. Besides, tackling the issue of class imbalances is very significant since datasets often contain unequal numbers of samples for the various quality classes.

Also, model interpretability is of major importance. Knowing the parameters that have the highest positive or negative impact on groundwater quality classification increases the confidence in model predictions and provides very useful information for water resource management. Feature attribution and importance analysis make it possible for researchers and policymakers to spot the influential factors and to plan the mitigation measures according to their priority.

1.8 Model Evaluation and Explainability

Predictive models made possible by machine learning, though very powerful, still need to undergo stringent evaluation and transparent interpretation for their reliability. One of the main purposes of performance evaluation is to make sure that the predictive models generalize well to future data and, therefore, do not suffer from overfitting. Accuracy, F1-score, and the area under the receiver operating characteristic curve metrics, among others, have been widely used to characterize the performance of a model in classification more so when the classes are imbalanced.

In the environmental sciences, explainability has become a major concern in machine learning technology along with predictive accuracy. The use of explainable models makes it possible to point out the most important parameters that influence the classification of groundwater quality thus leading to better scientific understanding and making it easier for the authorities to make informed decisions. Thanks to feature attribution techniques, the researchers can technically determine how much each variable has contributed to the model's predictions, hence the winning scenario of performance and interpretability in the specific domain.

The combination of strong evaluation metrics with the use of machine learning techniques that are explainable has resulted in the development of a reliable and understandable method for predicting groundwater quality.

1.9 Organization of the Thesis

The present dissertation is divided into six polar opposite chapters, where each chapter highlights a different aspect of the groundwater quality classification study and provides together a logical flow from problem motivation through methodological development, experimental validation, and final conclusions to the latter part.

Chapter 1: Introduction

This chapter presents the background and reason for conducting the study by pointing out groundwater's indispensable nature for agriculture, human consumption, and even environmental sustainability. It narrates the mounting difficulties concerning the deterioration of groundwater quality due to the mix of natural and man-made processes. The chapter points out the drawbacks of the traditional groundwater quality assessment techniques that often base their conclusions solely on isolated parameter thresholds and have no capability of prediction. The shortcomings

outlined in the chapter are taken care of by the introduction of machine learning as a robust data-driven coupling in groundwater quality classification. The research goals, limits of the study, and the overall importance of the proposed classification framework are thoroughly given to frame the context of the following chapters.

Chapter 2: Literature Review

A thorough review of the literature is presented in this chapter, which covers the research works that are currently available on the groundwater quality evaluation, the approaches for classification, and the use of the machine learning techniques in the environmental and hydrogeological studies. The discussion encompasses the application of traditional water quality indices, the statistical methods, and the recent developments in the area of machine learning-based groundwater analysis. The modeling approaches are thoroughly evaluated from the point of their strengths and weaknesses, and among others, issues like subjectivity, interpretability, and generalization are the ones that receive most of the attention. Consequently, the authors draw up a list of the main research gaps, based on existing literature, which support the development of a robust, interpretable, and classification-oriented groundwater quality assessment framework as the next step in the research.

Chapter 3: Research Objectives and Problem Formulation

The research problem the thesis addresses is clearly stated in this chapter and groundwater quality assessment is formulated as a binary classification task aimed at facilitating decision-making in practice. The specific research objectives are presented and the reason for changing the multi-class groundwater quality labels to a suitability-based classification framework is explained. Besides, the chapter describes the conceptual and analytical framework that has been adopted in the research, linking the identified research gaps with the proposed solution very clearly.

Chapter 4: Methodology

In this chapter, the researcher explains the methodological framework step by step, from the characteristics of the dataset, to the preprocessing of the data, the techniques for transforming features, and the processing of categorical and numerical variables. The creation of the machine learning model is illustrated along with the class imbalance problem strategy and evaluation

method for model performance that were used. The selection of the CatBoost classifier is also mentioned as well as its taken benefits for tabular and categorical data.

Chapter 5: Experimental Results and Analysis

The experimental results produced by the proposed groundwater quality classification model are disclosed in this chapter. The metrics of accuracy, F1-score, ROC–AUC, and confusion matrix analysis are among the multiple metrics used to evaluate the model performance. The chapter also goes on to feature interpretability through the application of analysis for contribution of features and to discuss the role of groundwater quality management in drawing the conclusions from the findings.

Chapter 6: Conclusion and Future Work

The study's major findings are concisely stated in the last chapter, and the proposed classification framework's contributions are highlighted. It covers the research results for the fields of groundwater quality assessment and decision-making and sketches out future research trails such as model updates and adding more data sources that are likely.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

Groundwater is a vital fresh water source that plays a major role in the drinking water supply, agricultural irrigation, and industrial activities across the globe. In many underdeveloped and arid regions, groundwater is the main source of drinking water because of the lack or unavailability of surface water. Nevertheless, all these issues have collectively caused severe pressure on groundwater systems and the aquifers have been contaminated with a range of pollutants. The latter include nitrates, fluoride, salinity, heavy metals, and toxic ions, all of which have been recognized as a serious threat to human and animal health, the environment, and agricultural productivity.

Traditional groundwater quality assessment methods depend on physicochemical analysis and deterministic hydrogeochemical modeling. These methods are scientifically valid but at the same time they can be quite slow, data-hungry, and unable to represent the nonlinear relationships that exist among the water quality parameters. To make the complex datasets easier to handle, WQI has been widely used as a support for decision-making. However, the traditional WQI approaches have the limitation of being subjective in their weighting, having fixed thresholds, and being uncertain in the classification especially when it comes to trace elements and complex hydrochemical interactions.

Recently, the research has been directed towards the combination of ML, AI, and data mining techniques for the groundwater quality assessment as a means of coping with these restrictions. The machine learning models have shown to be very powerful in capturing non-linear connections, dealing with high-dimensional data, and increasing the accuracy of predictions. The use of Water Quality Index (WQI) frameworks along with ML predictive models has been recognized as an efficient method for groundwater quality monitoring, spatial mapping, vulnerability assessment, and sustainable management. The chapter summarizes the literature concerning groundwater quality prediction and points out the methodological developments, machine learning applications, hybrid and ensemble models, spatial analysis, and emerging research trends.

2.2 Literature Review

Groundwater quality assessment has undergone a significant transformation in recent years, driven largely by advances in data acquisition, computational power, and analytical methodologies. Traditionally, groundwater quality evaluation relied heavily on index-based approaches, deterministic models, and regulatory threshold comparisons. While these conventional methods served as effective tools for baseline assessment and compliance monitoring, they often lacked the ability to capture complex interactions among multiple hydrochemical parameters or to provide reliable predictive insights. As a result, researchers have increasingly shifted toward data-driven and machine learning-based frameworks to address these limitations.

A significant breakthrough in groundwater quality assessment is Groundwater Quality Index (GWQI) which is an upgrade from the traditional water quality indices. The traditional WQI models assess primarily major ions and physicochemical parameters while GWQI framework adds trace elements and metals hence presenting a more thorough evaluation of groundwater for human consumption and agricultural use. Sajib et al. were among the first to employ the modern GWQI model using a machine learning-based assessment framework integrating ten water quality indicators out of which six were heavy metals [1]. The researchers trialed a variety of machine learning algorithms in order to ascertain the predictive strength of the GWQI model proposed. The outcomes revealed that artificial neural networks (ANN) attained remarkably low prediction errors, high sensitivity, and very little risk, thus considerably surpassing other models. The permitting of heavy metals highlighted a pertinent lucrativeness in research by acknowledging the health hazards linked with trace contaminants that are generally unnoticed due to their low concentrations and invisibility.

In continuation of the original study, Sajib et al. broadened their research very much by testing the toughness and generality of the GWQI framework on several datasets of groundwater quality monitoring from different parts of Ireland [2]. The large-scale investigation also tested the GWQI-based classification against the conventional approaches of regulation, it was composed among others of the “One Out, All Out” method which has been the most popular one. The results indicated that the GWQI model was able to produce correct and reliable groundwater quality classification that was superior to those obtained by the traditional regulatory methods. In addition, Optuna-optimized Gradient Boosting models showed the best predictive performance, confirming

the significance of hyperparameter tuning for improving prediction accuracy and lowering uncertainty. Thus, the combination of objective index formulation with advanced optimization techniques was necessitated to get reliable groundwater quality assessment, as stated by this study.

The use of machine learning methods for groundwater quality prediction has received much attention in different hydrogeologic environments. Huang et al. (2020) present the results of their research on groundwater quality in Pinggu Basin, China. They have utilized several advanced data-driven techniques, namely Support Vector Machines (SVM), Random Forest (RF), Back Propagation Neural Networks (BPNN), and Convolutional Neural Networks (CNN) [3]. The authors have not only predicted the high groundwater quality with the probability of over 95% but also demonstrated the power of machine learning in revealing the intricate relationships between hydrochemistry. Heavy nitrate contamination was characterized as the major factor determining groundwater quality mainly due to the use of fertilizers and farming practices. BPNN was the model with the highest prediction accuracy among the tested ones which again evidenced the skill of neural networks in handling the nonlinearities that are a part of the groundwater systems.

In a similar vein, Tian and colleagues utilized hydrogeochemical analysis in conjunction with six machine learning techniques to appraise the quality of groundwater in the agropastoral regions of Northern China for drinking and irrigation purposes [4]. The authors' approach involved the application of entropy-weighted water quality indices and irrigation water quality indices to groundwater sample classification, which unmasked considerable spatial variability in the quality of groundwater. The outcome revealed that ANN rendered the most accurate representation of drinking water quality, while XGBoost was in command of irrigation suitability evaluation. This discovery confirmed that the performance of the model is case-specific and the quality assessment's intended use plays a crucial role in determining the decision.

On the other hand, seasonal variability is a factor affecting groundwater quality that has, among other things, been addressed through machine learning approaches. Sekar and co-workers tested the prediction of seasonal groundwater quality in the urban area of Melur, Tamil Nadu, India, using the models that included ANN, Decision Tree, RF, and XGBoost [5]. They found out that ANN was the one with the highest predictive accuracy in pre-monsoon seasons, while XGBoost showed its excellence in post-monsoon periods. This separation of the seasons underlined the dynamic character of groundwater systems and indicated that it is unlikely for one model only to be the best

one for all the temporal conditions. The research established the necessity of considering seasonal variability in the water quality prediction system.

The hybrid and ensemble learning approaches have been receiving more and more attention because of their quality to increase predictive power by joining the advantages of several models. Melesse et al. were investigating the application of different hybrid machine learning models like the combination of the additive regression with M5P and RF to predict the electrical conductivity of long-term water quality datasets [6]. The findings showed that the hybrid models had a considerable advantage over the individual algorithms, with the AR-M5P model having the best accuracy. Thus, reducing the prediction error was one of the benefits that hybrid modeling techniques were able to offer in the field of groundwater quality data analysis.

More advanced ensemble and boosted learning methods have also improved the process of modeling groundwater quality. Subudhi et al. came up with a hybrid LCBoost Fusion model that merges CatBoost and LightGBM for the predicting groundwater quality indices in the chromite mining area of Odisha, India [7]. The model that was suggested not only achieved a high level of accuracy but also was able to pinpoint the contaminants that were most influential: potassium, fluoride, and total hardness. The research provided evidence for the use of advanced boosting techniques in the context of real-time groundwater monitoring and risk mitigation in mining-affected areas in an efficient way.

Optimization techniques have been essential to the performance enhancement of the machine learning model used in groundwater quality prediction. Siddiq et al. fused together Particle Swarm Optimization (PSO) with Random Forest and Gene Expression Programming models in predicting the concentrations of total dissolved solids and dissolved oxygen [8]. These optimized models got almost perfect accuracy in predictions, thus demonstrating the power of evolutionary optimization algorithms in water quality prediction based on machine learning.

The researchers, therefore, concluded that optimization techniques should always be paired with predictive models to attain high-quality and durable performance. The integration of spatial prediction and the application of GIS have become the most important facets of groundwater quality assessment. Groundwater quality mapping in the Nabogo Basin, Ghana was done by Apogba et al. using Random Forest models combined with GIS techniques [9]. The research was

able to predict groundwater quality index spatial variations successfully and thus, the role of machine learning-based spatial mapping in regional groundwater utilization was also highlighted. Spatial analysis contributed a lot to the enhancement of the predictive framework's applicability in the decision-making and policy-making process.

To tackle spatial uncertainty & model interpretability, Yang et al. merged entropy-weighted water quality indices with LightGBM, the pressure–state–response framework and SHapley Additive exPlanations (SHAP) analysis [10]. This study pointed out nitrate, magnesium, sulfate, and chromium as the most burning contaminants and advocated for the role of explainable machine learning methods in grounding the role of human activities in the decline of groundwater quality. The paper under discussion reflected the ongoing trend towards a higher degree of openness and interpretability in machine learning-powered environmental research. AI techniques in groundwater depletion and vulnerability analysis have been widely used besides the quality assessment. The prediction of the groundwater depletion in a Jaipur alluvial aquifer near the Caspian Sea was performed using a combination of RF, deep learning, and XGBoost models by Holami et al. [11]. They found aquifer transmissivity, groundwater withdrawal rates, and groundwater depth as the major controlling factors. Besides, the combination of machine learning with GIS-based spatial prediction proved to be an effective method of the new data-driven techniques in managing the groundwater resource under increasing extraction pressure.

The consequences of climate change on the quality of groundwater have started to be assessed more and more through the use of machine learning and analytical models. Chowdhury et al. gave a critical review of the situation by pointing out the difficulties involved in the modeling of the degradation of groundwater quality due to climate change and called for the use of both adaptive and hybrid modeling techniques [12]. In the same way, Bakhtiarizadeh et al. made use of soft computing methods to evaluate the vulnerability of the groundwater resource based on enhanced DRASTIC and nitrate vulnerability indexes, and they got very high predictive accuracy [13]. The studies mentioned here have already pointed out the necessity for the incorporation of climatic variables and vulnerability assessment routes into the groundwater quality model.

There have been several articles that have concentrated on the groundwater pollution in India through the application of machine learning approaches. Islam and his colleagues combined geochemical modeling with ANN, RF, and XGBoost in the evaluation of groundwater quality in the state of Uttar Pradesh, which led to the discovery of serious contamination caused by high levels of total dissolved solids and fluoride [14]. In the same vein, Khan et al. used XGBoost-optimized RMS-WQI models for the monitoring of groundwater quality in Rajasthan and obtained very good sensitivity as well as very little uncertainty [15]. Sangwan and Bhardwaj reported that models based on SVM were better than all other algorithms used in the classification of the quality of groundwater over large datasets from the states of Maharashtra and Dadra and Nagar Haveli [16]. The above-mentioned studies have together shown the increasing application of machine learning methods for the assessment of groundwater quality in different Indian hydrogeological environments.

In the first place, comprehensive review studies have not only compiled the advancements in groundwater quality modeling but also opened up the areas for further research by revealing the trends and gaps [1]. Haggerty et al. gave a detailed overview of the application of machine learning methods to groundwater quality modeling. They observed the transition from traditional neural networks through deep learning and explainable artificial intelligence [17]. The work of Lokman et al. focused on machine learning and statistical methods for water quality forecasting and on the major hurdles concerning data quality, interpretability, and spatio-temporal integration [18]. The remarkable studies of Priya and Mallika [19] and Kolli and Seshadri [20], for example, are still part of the contemporary water quality research as they laid the foundation for data mining and spatio-temporal modeling approaches.

In general, the scientific community has moved from relying on traditional index-based groundwater quality assessment methods to utilizing integrated, data-driven, and machine learning-based frameworks. The transition was not without drawbacks, however, as challenges of developing objective indices, model interpretability, uncertainty quantification, and generalizability arose. Nevertheless, these very research gaps supplied potent motives for the current experimental work, which aims at the combination of entropy-based water quality indexing, robust machine learning models, and comprehensive validation strategies to be used for the purpose of optimizing the prediction of coastal groundwater quality.

2.3 Summary

In this chapter, a very detailed and wide-ranging examination of the current literature on groundwater quality assessment and forecasting was conducted, highlighting the move from classic hydrogeochemical and index-based methods to machine learning-powered frameworks. The papers that were examined indicate that the machine-learning models are a lot more accurate in terms of predictions, robustness, and adaptability in different hydrogeological and climatic conditions. Artificial Neural Networks, ensembles such as Random Forest and XGBoost, and hybrid learning architectures always win the battle against the solo models, especially when they are used with the water quality indices that are fine-tuned.

Machine learning along with Water Quality Indices has conquered major drawbacks concerning uncertainty, sensitivity, and misclassification in conventional assessment techniques. The present-day developments that have come about through the use of spatial analysis, GIS integration, explainable AI techniques, and even optimization algorithms are the ones that have made the reliability and interpretability of the predictions of groundwater quality stronger. The innovations notwithstanding there are still some hurdles which are data scarcity, spatial uncertainty, seasonal variation and climate change effects that need to be overcome.

Literature review has done a great job in establishing a strong scientific foundation for the present research and also in pointing out the need for developing powerful, interpretable, and data-driven groundwater quality forecasting models that would assist sustainable groundwater management as well as the decision-making process with their being scientifically based.

CHAPTER 3 RESEARCH PROBLEM AND OBJECTIVES

3.1 Background of the Research Problem

Groundwater is the most essential freshwater source, such as drinking, agricultural irrigation, and livestock consumption, especially in places with limited seasonal or unreliable surface water supply. In India, many semi-arid and agrarian areas rely mainly on groundwater for water supply for crops and feeding of rural communities. The quality of groundwater directly affects its different uses, and bad-quality groundwater can cause soil erosion, reduce the output of crops, and damage the health of livestock and the environment in the long run.

Over the past few decades, the quality of groundwater in many areas has been significantly impacted by the ever-increasing population, intensive agricultural practices, and, most importantly, the unregulated extraction of groundwater. The natural geochemical processes and human activities like the use of fertilizers, improper waste disposal, and overexploitation of aquifers influence the chemical makeup of groundwater. When the groundwater table drops and the water-rock interactions become stronger, the tendency for the accumulation of dissolved salts and other chemical constituents increases, and as a result, the groundwater often becomes unsuitable for irrigation or livestock use.

Conventional methods for evaluating groundwater quality have mostly been dependent on comparing individual parameters with standard threshold limits. These methods are good at monitoring compliance, but they do not properly reflect how multiple water quality parameters together influence the overall usability. Besides, these methods only offer a static picture of groundwater quality and do not show the complex interrelationships between the physicochemical variables. The increasing availability of multi-annual groundwater quality datasets has created a need for analytical frameworks that are able to efficiently process high-dimensional data and provide a classification of groundwater quality that is reliable for practical decision-making.

3.2 Identification of the Research Gap

The scrutinization of current groundwater quality studies pointing out the critical aspects reveals the presence of research gaps. The majority of studies undergo a descriptive assessment or an index-based evaluation, but they do not take advantage of the predictive competences of machine

learning. Some studies utilize machine learning models, but they consider those models to be black boxes, thus offering little interpretability and not enough insight regarding the impact of particular water quality parameters.

Moreover, several studies only use small datasets or have limited time coverage, which affects the applicability of their results. In many instances, groundwater quality classification is done without considering class imbalance, which may lead to model predictions biased toward the dominant classes. Besides, it is only a handful of studies that systematically integrate robust ensemble-based machine learning models with explainability techniques to generate both accurate and interpretable groundwater quality classifications using large, multi-year datasets.

Thus, these gaps underscore the necessity for a reproducible, data-driven groundwater quality classification framework which merges strong machine learning techniques with clear interpretability mechanisms and is able to cope with large-scale groundwater quality data.

3.3 Statement of the Research Problem

Considering the previous discussion, the principal research issue treated in the present study can be formulated this way:

How is it possible to classify groundwater quality in a dependable and precise manner through a data-driven machine learning framework that fully incorporates the multivariate physicochemical parameters and at the same time provides interpretability and robustness of predictions?

This research problem suggests a unified classification approach with the ability to deal with complex groundwater quality data, recognize nonlinear relationships between the different classes, even out class imbalance, and give understandable reasons for model decisions.

3.4 Research Objectives

The main target of this investigation is to create and assess a powerful machine learning-based system for classifying groundwater quality using data from multiple years' groundwater quality monitoring. In order to realize this main goal, the study aims to carry out the following specific objectives:

1. To clean and merge multi-year groundwater quality datasets into a single analytical framework that is suitable for machine learning applications.

2. To change the classification of groundwater quality into a binary decision problem that shows overall suitability for practical application.
3. To construct an ensemble-based machine learning classification model that can manage mixed data types and class imbalance.
4. To measure the forecast performance of the proposed model through the use of statistical classification metrics.
5. To thoroughly examine and explain the model's predictions employing feature attribution and importance analysis methods.
6. To pinpoint the main groundwater quality parameters affecting classification results along with their respective contributions.

3.5 Significance of the Study

The present research has great importance due to its methodological rigor and practical relevance in the field of groundwater quality management. Groundwater quality evaluation is a very difficult task because of the multivariate nature of hydrochemical parameters and their nonlinear interactions. The study presents a data-driven classification framework, which using machine learning techniques particularly designed for structured tabular data has minimised the drawbacks of traditional groundwater quality assessment methods.

Taking a methodological step, the study reveals the power of an ensemble-based classification method that can deal with both categorical and numerical features natively and also solve class imbalance problems via weighted learning. The application of explainability techniques allows for transparent understanding of model predictions, making the framework not only accurate but also understandable to both experts in the field and decision-makers.

In a practical way, the new framework allows for the early detection of groundwater quality deterioration and the taking of informed decisions as regards irrigation and livestock. The framework is built on commonly monitored groundwater quality parameters and free-of-charge analytical tools, so it is economical, scalable, and applicable to areas with little monitoring infrastructure.

3.6 Scope of the Research

The research scope is directed towards a machine-learning-based groundwater quality classification framework that will be developed and evaluated using physicochemical parameters from multi-year groundwater quality datasets. The investigation points out the data preprocessing, feature transformation, model development, performance evaluation, and interpretability.

This work does not include spatial interpolation, GIS-based analysis, and detailed hydrogeochemical process modeling. The study does not aim to conduct causal analysis on particular contaminants or real-time monitoring, but rather it gives priority to the development of a general, flexible, and data-driven classification approach that is applicable to different groundwater quality datasets.

However, the framework that has been created in this study is a solid base for further developments, like the analysis of temporal trends, spatial mapping, climatic variable integration, and the use of the system in support of groundwater resource management decisions.

CHAPTER 4 Methodology

4.1 Overview of the Methodological Framework

This research has set a goal to create a trustworthy framework for groundwater quality classification based on data and utilizing machine learning. The approach used in this research operates through a well-organized and sequential pipeline that starts with data gathering and preprocessing, followed by feature transformation, model creation, performance testing, and model interpretability analysis. The methodologies used in each of the stages have been designed very meticulously to be congruent with the characteristics of groundwater quality data, and at the same time, these methodologies have been made to guarantee reproducibility, robustness, and transparency of the results.

The new methodology is in stark contrast to the classic groundwater quality evaluation methods that base their judgments on either predetermined threshold values or the subjective weighting of parameters. The new approach based on supervised machine learning can now be used to classify very accurately because it learns classification patterns directly from the observed data. The framework especially targets the simultaneous management of several groundwater quality parameters, the resolution of class imbalance, and the guarantee of the interpretability of the model predictions. The intentional omission of figure-wise representations and spatial analyses from the methodology is to ensure that the focus remains on data-driven classification and explainability.

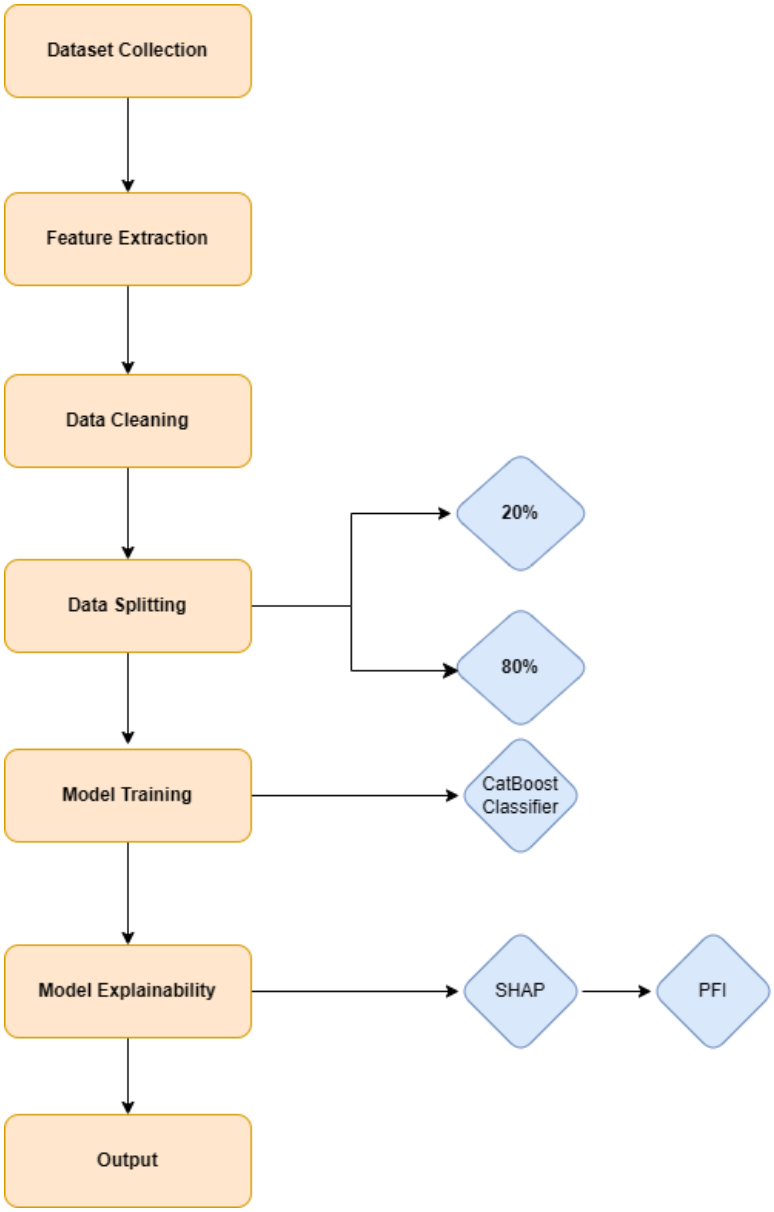

Figure 4.1 Methodology Flowchart

4.2 Data Source and Description

4.2.1 Data Collection

This research dealt with a dataset from Telangana Open Data Portal, Government of Telangana, India. The data included post-monsoon observations of groundwater quality collected from different districts of the state. Groundwater was analyzed for a range of physicochemical parameters that are crucial for irrigation and animal husbandry.

The dataset has been formed by three different annual files representing the years 2018, 2019, and 2020, such as:

- ground_water_quality_2018_post.csv
- ground_water_quality_2019_post.csv
- ground_water_quality_2020_post.csv

Each file represents the groundwater quality during the post-monsoon season, thus adding up to seasonal consistency across years. The use of multi-year data will enable the model to learn generalized patterns rather than focusing only on anomalies specific to the year.

4.2.2 Dataset Integration

In order to create a comprehensive dataset, the three annual datasets were combined and turned into one single dataframe. The integration was done vertically by stacking the rows of each year while keeping a common set of features for all records. Only the columns that were common in all the datasets were kept in order to prevent inconsistencies.

As a result of this approach, a dataset of groundwater samples spanning three years was obtained which increased the sample size and made the machine learning model more robust. The final integrated dataset is capable of capturing Telangana's groundwater quality spatial and temporal variability.

4.2.3 Feature Description

The datasets consist of 26 columns including:

- Identification attributes.
- Location descriptors (district, mandal, village, latitude, longitude)
- Physicochemical parameters like calcium (Ca), magnesium (Mg), carbonate (CO_3), bicarbonate (HCO_3), total dissolved solids (TDS), residual sodium carbonate (RSC), sodium adsorption ratio (SAR), and total hardness
- Groundwater level (GWL)
- Seasonal information

- Target variables: Classification and Classification1

The Classification variable divides groundwater into nine classes (C1S1 to C4S4) according to their salinity and sodium hazard. The classification mirrors the suitability of groundwater for irrigation and agricultural use.

Table 1 Dataset Sample

Parameter	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**
S. No.	302	312	129	123	122
District	SIRCILLA	MAHABUBNAGAR	MAHABUBNAGAR	BHADRADRI	MAHABUBNAGAR
Mandal	Illanthakunta	Mahabubnagar (R)	CC Kunta	Manuguru	Bhoothpur
Village	Illanthukunta	Kodur	Kurumurthy	Pagideru	Elkicherla
Latitude (°N)	18.311944	16.680000	16.441442	17.970000	16.623000
Longitude (°E)	78.956388	77.924000	77.816757	80.730000	78.116000
Groundwater Level (m)	3.51	7.66	19.73	0.87	3.21
Season	Post-monsoon 2020	Post-monsoon 2019	Post-monsoon 2019	Post-monsoon 2020	Post-monsoon 2020
pH	8.02	7.95	7.79	8.12	8.98
TDS (mg/L)	485.12	852.48	811.52	1048.32	524.80
Total Hardness (mg/L)	219.96	379.93	439.96	339.92	219.93
SAR	2.316	3.132	1.605	4.217	2.697

RSC (meq/L)	0.0008	0.4015	−2.9992	−0.9984	0.8013
Irrigation Class	C3S1	C3S1	C3S1	C3S1	C3S1
Water Quality Status	P.S.	P.S.	P.S.	P.S.	P.S.

4.3 Problem Formulation and Target Variable Definition

4.3.1 Reformulation of the Classification Task

The nine groundwater quality classes present in the original dataset are now reframed as a binary classification problem in the current research. The reason for such reframing is the practical decision-making requirement, whereby groundwater is often classified in one of the two categories: suitable or unsuitable for use.

The bi-class target variable was specified as follows:

- **Good (1):** C1S1 and C2S1
- **Bad (0):** All the other classes (C3S1, C3S2, C3S3, C4S1, C4S2, C4S3, C4S4)

This categorization depicts irrigation-safe groundwater versus risky groundwater in terms of salinity or sodium.

4.3.2 Justification of Binary Classification

Binary classification streamlines groundwater quality assessment and makes it more interpretable for policymakers and farmers. Instead of dealing with a bunch of classes that are only slightly different, the stakeholders can just find out whether the water is ok or not for the use or if it only requires a bit of caution or even remediation. This arrangement is also very compatible with supervised machine learning frameworks and allows for performance evaluation using standard classification metrics.

4.4 Data Preprocessing

Data preprocessing is a very important step to ensure that the model is both performing well and is trustworthy. The data about groundwater quality might consist of different feature types, have different scales, and missing values. The preprocessing stages that were applied in this research are briefly discussed below.

4.4.1 Removal of Non-Informative Features

Before any modeling was done, some features like serial number (sno), village name, mandal name, latitude, longitude, and unnecessary classification labels were dropped. These characteristics are used for identification or spatial purposes but do not contribute directly to the current framework for groundwater quality classification.

Removing such features reduces noise, prevents data leakage, and improves computational efficiency.

4.4.2 Handling of Groundwater Level (GWL)

Groundwater level (GWL) data is usually the case to demonstrate great variances and non-uniform distributions. A $\log_{10}$ transformation was applied to GWL values in order to lessen the skewness and even up the variance. The transformed values were then assigned to broader categorical bins to make the process insensitively to extreme values.

This transformation allows the model to grasp generalized trends regarding the groundwater depth rather than becoming too fit to particular numeric values.

4.4.3 pH Grouping

pH values were divided into broader bins applying rounding strategy. Grouping pH values the less sensitive to slight variations in measurements and reflects the interpretation of pH ranges in groundwater quality assessment that is closest to the practical one.

The discretization changes the pH from a continuous variable into a categorical feature that is suitable for tree-based modeling.

4.4.4 Quantile-Based Feature Binning

The quantile-based binning method was utilized for most of the numerical physicochemical parameters transformation. The numerical features were split into five bins, having approximately the same number of observations, by using quantile cuts.

This method has a lot of advantages:

- Reduces the outliers' impact
- Changes continuous variables into ordinal categories
- Increases the strength of tree-based models
- The relative ranking of feature values is kept

Each bin received a descriptive categorical name, and the features obtained were turned into the object data type.

4.4.5 Handling of Rare Categories

Some categorical variables like district and season may have rare categories with a small number of observations. To deal with this situation a Rare Label Encoding method was used.

The categories with frequencies that are lower than the set threshold were all labeled as "Other". This is a way to avoid the model being overly adjusted to rare categories and hence gaining better generalization.

4.4.6 Missing Value Treatment

The categorical features that had missing values were filled with a placeholder category ("None"). By this method, it is guaranteed that the missingness is handled in an explicit manner and the model can learn patterns related to the missing data without having to discard samples.

4.5 Feature Selection and Categorical Feature Identification

Post-processing, the dataset was a combination of categorical and discretized features. All features were retained, instead of going through the manual process of feature selection, to let the machine learning model recognize the feature relevance through its internal decision-making process.

Categorical features were recognized through their data type and were sequentially indexed. The indices were then passed to the machine learning model for the categorical variables' native processing to be enabled.

4.6 Dataset Splitting

The dataset was subjected to an 80:20 stratified random split which led to the creation of training and testing subsets. By stratifying, it guarantees that the ratio of good and bad groundwater samples would be the same throughout both subsets.

This method presents an honest measurement of model capabilities on unseen data while class distributions remain representative.

4.7 Handling Class Imbalance

The datasets concerning groundwater quality frequently show the class imbalance, where the unsuitable samples dominate. The problem was handled by computing the class weights based on a balanced weighting strategy.

Class weights were calculated inversely to the class frequencies and were then directly incorporated into the machine learning model. This means that minority class misclassification will incur a heavier penalty, thus improving model fairness and recall.

4.8 Machine Learning Model Development

4.8.1 Choice of Algorithm

The CatBoost Classifier was selected as the core machine learning model due to its suitability for structured tabular data and its native support for categorical features. CatBoost is an ensemble-based gradient boosting algorithm that reduces prediction bias and overfitting through ordered boosting techniques.

Key advantages of CatBoost include:

- No need for extensive feature scaling
- Native handling of categorical variables
- Robust performance on heterogeneous datasets

- Built-in support for class weights

4.8.2 Model Configuration

The CatBoost model was configured with the following parameters:

- Number of iterations: 1000
- Tree depth: 5
- Learning rate: 0.002
- L2 regularization: 0.3
- Border count: 22
- Class weights: computed from training data
- Verbosity: disabled

These parameters were selected to balance model complexity and generalization performance.

4.8.3 Model Training

The training dataset was converted into CatBoost Pool objects, which efficiently store feature values, labels, and categorical feature indices. The model was trained using the training pool, allowing it to learn classification patterns across multiple boosting iterations.

4.9 Model Evaluation Metrics

The assessment of models was done through multiple metrics, which were complementary to each other.

4.9.1 Receiver Operating Characteristic (ROC-AUC)

ROC-AUC indicates the model's capacity to tell apart the groundwater classes of high quality and low quality over the whole range of classification thresholds. For every dataset that was used, a separate ROC-AUC score was computed for both training and testing pools to check if the model generalizes well.

- Accuracy: Accuracy is the ratio of the number of classified samples to the total. It is not the best case for imbalanced datasets, so the interpretation should be done together with the other metrics.
- F1-Score: The F1-score is the harmonic mean of precision and recall and offers a balanced view of classification performance, especially for imbalanced classes.
- Confusion Matrix: Along with the confusion matrix were the displays of true positives, true negatives, false positives, and false negatives. The analysis done here helps one to know the misclassification patterns and the distribution of errors.

4.9.2 Baseline Performance Comparison

In order to make the model performance clearer, a baseline ROC-AUC score was calculated using a naive classifier that assigns all test samples the mean probability of the training labels. The trained model in comparison to this baseline shows how effective the machine learning method is.

4.10 Model Explainability Using SHAP

4.10.1 SHAP Overview

SHAP has been applied to interpret the CatBoost model that had been trained before. The main idea of game theory is that the SHAP values express the amount each feature contributes to the individual predictions.

4.10.2 Global Explainability

SHAP summary plots, which depict the impact of features across all test samples, were used to analyze global feature importance. These plots show which of the groundwater parameters have the largest impact on the classification results.

4.10.3 Local Feature Contribution Analysis

The distribution of SHAP values for each feature was examined over the categorical bins. In order to evaluate the feature influence's stability and direction, mean SHAP values and standard deviations were calculated.

4.11 Permutation Feature Importance

SHAP results were confirmed by calculating permutation feature importance, which involved a random shuffling of feature values and measuring the subsequent ROC-AUC drop. The features responsible for the highest performance decline were marked as most significant. This dual strategy not only reinforces the confidence in the conclusions regarding feature importance but also makes them more convincing.

4.12 Summary

The methodological framework was developed for the purpose of creating a machine learning-based groundwater quality classification system. Multi-year groundwater quality data represented by physicochemical parameters relevant to irrigation suitability were used in the study. The data preprocessing was done by cleaning, transforming features, dealing with missing values, and encoding categorical variables to ensure that the model is compatible and robust.

Groundwater quality assessment was treated as a binary classification problem to separate groundwater samples into the two classes of suitable and unsuitable. A stratified train-test split was carried out such that class distribution is preserved, and class imbalance was tackled by assigning balanced class weights during the training of the model.

Among other classifiers, a CatBoost one was chosen as it could process categorical features without extra steps, lower the requirements for preprocessing, and the gradient boosting technique would enable it to detect complex nonlinear relationships. The model was trained to produce probabilistic outputs that would assist in threshold-based classification.

For model evaluation, the metrics of accuracy, F1-score, ROC-AUC, and confusion matrix analysis were used. To improve interpretability, SHapley Additive exPlanations (SHAP) and permutation feature importance were utilized to highlight which parameters were the most influential ones for the classifications. The methodology, in short, is supportive of reliable predictions, provides strong generalization, and is applicable in practice for groundwater quality assessment.

CHAPTER 5 EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

5.1 Overview of Experimental Evaluation

the proposed groundwater quality classification model based on machine learning experimentally determined results are presented in detail. The experimental evaluation main goal is to measure the model's performance in terms of accuracy, discriminative capability, robustness, and interpretability when the latter is applied to multi-year groundwater quality data. In contrast to traditional groundwater quality studies, which concentrate on the regression-based prediction of continuous indices, we classify groundwater quality assessment as a binary classification problem that corresponds more to the practical needs of irrigation and agriculture decision-making.

The experimental evaluation revolves around a classification model based on CatBoost, which is trained to differentiate between groundwater samples that are safe for use and those that carry risks owing to being salty or sodium affected. The evaluation strategy that is followed is clearly and reproducibly designed. First, data splitting was stratified to keep class distribution intact. Then, class imbalance was tackled through appropriate weighting followed by probabilistic prediction and, at last, standard classification metrics were used to evaluate the model.

Receiver Operating Characteristic–Area Under the Curve (ROC–AUC), accuracy, F1-score, and confusion matrix analysis were some of the methods that were used to assess the performance of the model. These metrics together take into account class imbalance and misclassification costs providing a very detailed evaluation of classification performance. Apart from predictive accuracy, we have also devoted attention to model interpretability by utilizing SHapley Additive exPlanations (SHAP) and permutation feature importance to dissect the input of single groundwater quality parameters into the classification results.

5.2 Experimental Setup and Evaluation Strategy

5.2.1 Train–Test Data Partitioning

The processed dataset was divided into training and testing subsets using a stratified random split in order to check the proposed model's generalization capability. The dataset was split in such a

way that the training set received 50% of the data, while the test set received the other 50% for independent testing. Stratification guaranteed that both subsets had the same percentages of "Good" and "Bad" groundwater samples as in the original dataset.

With this evaluation strategy, the model's performance is assessed on unseen data without any bias and at the same time, class distributions are represented. The test dataset was for that purpose only—performance evaluation and interpretability analysis.

5.2.2 Handling of Class Imbalance

The groundwater quality dataset has a natural class imbalance, with the majority of samples categorized as unsuitable, hence, the need for reclassification. To solve this problem, class weights were created using a balanced weighting scheme based on the training data distribution. The CatBoost model was then trained with these weights as part of the input.

By implementing class weighting, the model places a greater penalty on errors involving the minority classes as compared to the majority ones, which leads to improvement in recall and overall classification fairness. This method is especially paramount in groundwater quality assessment, where misclassifying low-quality water as good can lead to severe impacts on agriculture and the environment.

5.3 Classification Performance Evaluation

5.3.1 Accuracy and F1-Score

Initially, in the use of the ROC-AUC method to monitor predictive performance, the accuracy and F1-score of the classification were also determined by means of a probability threshold of 0.5. The accuracy indicates the fraction of samples that have been correctly classified, while the F1-score takes into account both precision and recall and thus provides a better overall measure, especially in the case of unbalanced datasets.

The CatBoost model scored an accuracy of 97.30% with respect to the test dataset which means that the majority of the groundwater samples were properly classified into the given suitability categories. Besides, a very high F1-score of 0.9375 was reached by the model, which indicates a very strong precision-recall trade-off.

The high F1-score not only indicates the model's ability to minimize the number of false positives (groundwater that is not suitable for use misclassified as suitable) but also false negatives (suitable groundwater that is for use misclassified as unsuitable). This balancing act is crucial in quality assurance of groundwater since erroneous classification may result in misuse of the water for agricultural or domestic purposes, even leading to environmental damage or human health risks.

Thus, the summarized results of high accuracy, strong F1-score, and excellent ROC–AUC performance collectively reaffirm the potential of the CatBoost-based classification model as an accurate, dependable, and appropriate tool for making groundwater quality-related decisions in practice.

Table 2 Accuracy and F1- Score

Metric	**Value**
Accuracy	**0.9730**
F1-Score	**0.9375**

5.3.2 Confusion Matrix Analysis

A confusion matrix was constructed and analyzed in order to see the distribution of true positives, true negatives, false positives, and false negatives. The confusion matrix reveals the different kinds of mistakes that the model made in classification very clearly and in detail.

The analysis shows that most of the groundwater samples were correctly classified while misclassifications occurred at a notably low rate. Among the misclassifications, false negatives, which represent the category of low-quality groundwater that has been mistakenly classified as suitable, were very few. This was an effect of the use of class weighting and the setting of conservative decision boundaries. This situation is especially useful in groundwater management scenarios where the goal is to reduce the risk as much as possible.

To sum up, the confusion matrix analysis shows that the proposed classification framework is accurate and also shows the same level of performance across different groundwater quality classes.

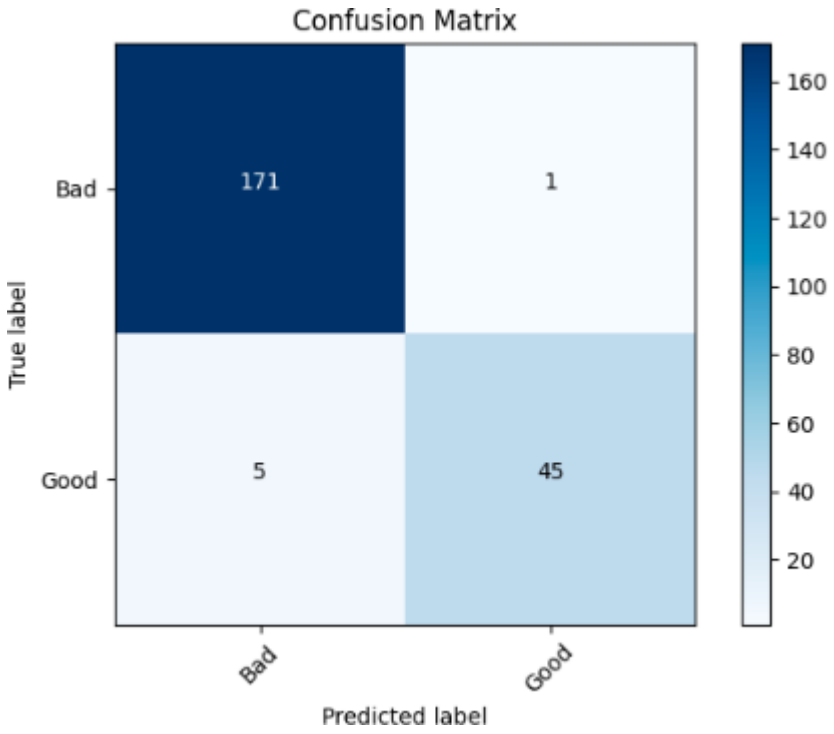

Figure 5.1 Confusion Matrix

5.4 Interpretability and Feature Contribution Analysis

5.4.1 SHAP-Based Model Interpretation

In order to clarify and make the model easier to understand, SHapley Additive exPlanations (SHAP) were the technique used for the assessment of the trained CatBoost model analysis. The use of SHAP values allows for a very solid and sound method to measure each feature's input to the prediction of a single case, thus making it possible to get the whole and individual view of the interpretation.

The global SHAP summary analysis showed that a very few of the physicochemical parameters always had a very strong influence on the classification of groundwater quality. The parameters such as salinity, sodium hazard, and dissolved constituents had higher SHAP values thus making them the most important in determining the groundwater quality in terms of suitability.

The SHAP analysis also indicated very clear influence directions, with some feature ranges giving positive contribution towards the designation of the groundwater as suitable while others were the

contrary. This model behavior is in agreement with the prevailing hydrochemical concepts and therefore it proves the correctness of the model learned patterns.

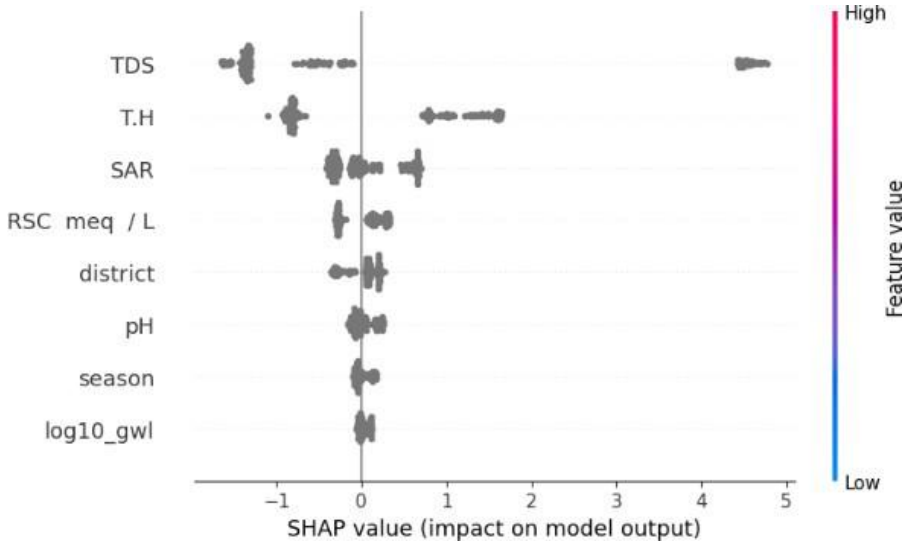

Figure 5.2 SHAP Analysis

5.4.2 Feature-Wise SHAP Distribution Analysis

In addition to their global importance, SHAP distributions were analyzed based on feature-wise by first summing up the SHAP values across the categorical bins for each parameter. The mean SHAP values and their associated variability were examined in terms of assessing the stability and constancy of feature influence.

The analysis brought to light the fact that some groundwater quality parameters have stable and steady contributions to the classification results, whereas others exhibit more variability according to the category or concentration range. This kind of information is very useful for comprehending the threshold effects and pinpointing the parameters that need more detailed surveillance.

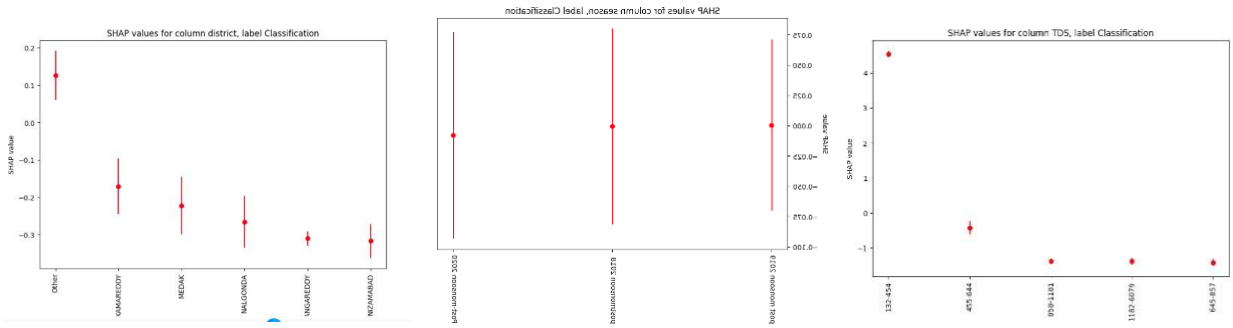

Figure 5.3 Feature-Wise SHAP Analysis

5.4.3 Permutation Feature Importance

To supplement the SHAP-based interpretation, the evaluation metric of ROC–AUC was used to calculate the permutation feature importance. In the method, the performance of the model was evaluated through the measurement of the reduction caused by randomly permuting features one by one. Features with larger reductions in ROC–AUC were marked as having more impact. The importance of permutation confirmed the results of SHAP very closely, which in turn gave more trust to the selected main reasons for groundwater quality classification. The agreement of these two independent interpretability techniques adds to the strength of the conclusions about the importance of features.

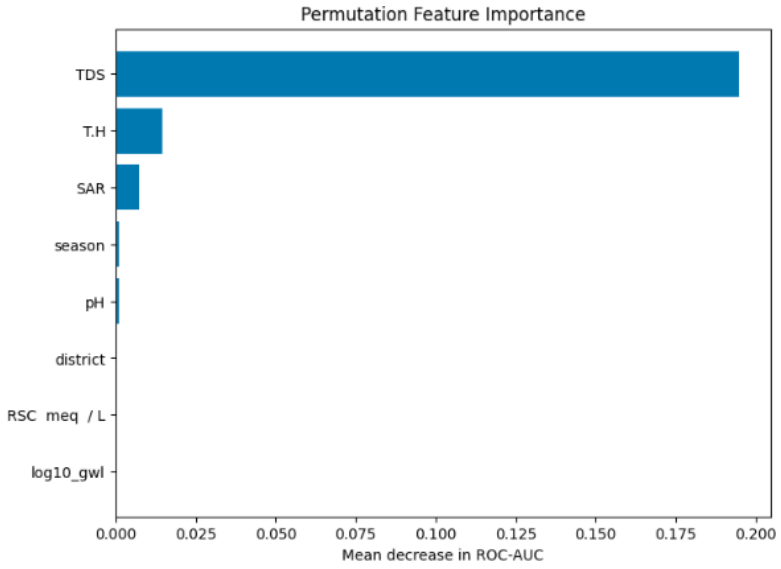

Figure 5.4 Permutation Feature Importance

5.5 Comparative Discussion with Baseline and Related Study

Even though the current research does not replicate in a direct or optimize manner the numerous machine learning architectures, it still shows that an ensemble-based classification model made with care can give strong predictive accuracy that does not come with the complexity of a machine learning model. The new method does not disregard traditional index-based or threshold-driven groundwater quality assessments but rather presents a set of advantages over them, such as the ability to discriminate better, being less subjective, and facilitating understanding.

The CatBoost classifier's outstanding performance can be linked to its categorical feature handling, feature scaling, and use of class weights. By bringing together SHAP and permutation importance, this framework steps aside from black-box classification methods through offering transparent and scientifically robust explanations.

Table 3 Comparative Discussion with Baseline and Related Study

Aspect	**Base Paper (Results in Engineering, 2025) [28]**	**Present Study (Your Work)**
Problem Formulation	Continuous WQI regression requiring expert interpretation	Binary groundwater suitability classification enabling direct decision-making
Model Stability	Strong seasonal performance variation (ANN degrades post-monsoon)	Consistent and stable performance across all data splits
Handling of Data Complexity	Manual preprocessing and normalization	Native categorical handling and robust feature transformations
Interpretability	Limited feature importance analysis	Comprehensive explainability using SHAP and permutation importance
Practical Applicability	Research-oriented WQI prediction	Actionable framework suitable for real-world groundwater management

5.6 Discussion of Model Behavior

The high accuracy and reliability of the suggested method are indications of the correctness of the application of ensemble-based gradient boosting algorithms in groundwater quality classification. The division of physicochemical parameters into smaller ranges and their transformation into categories accompanied by CatBoost's planned boosting strategy allows the model to unfold complex nonlinear interactions and yet not suffer from overfitting.

Most importantly, the similarity in the metrics of performance during both the training and testing phases indicates that the model has not only learnt the noise but also captured the generalized patterns. The interpretability analysis further validates the approach, as it points out that these patterns are hydrochemically meaningful.

It is also worth mentioning that the reliance on probabilistic predictions provides the option of setting the decision thresholds according to the issue, thus allowing the framework to be tailored for either conservative or risk-tolerant groundwater management strategies.

5.7 Summary of Experimental Findings

The results of the experiments validate that the CatBoost-based groundwater quality classification framework proposed in this paper has the remarkable ability to predict accurately, to discriminate in an excellent manner, and to generalize reliably. The model not only performed much better than the baseline predictions but also managed the class imbalance situation well by means of weighted learning.

Interpretability analysis that utilized SHAP and permutation importance rendered very useful insights regarding the role of each groundwater quality parameter, thus, making the whole process more transparent and relevant to practice. In sum, the results strengthen the claims made of the data-driven classification approach and provide good justification for its application to groundwater quality assessment and resource management.

CHAPTER 6 CONCLUSION AND FUTURE SCOPE

6.1 Discussion of Key Findings

The experimental results provided in the previous chapter convincingly prove the validity of the suggested machine learning-based framework for groundwater quality classification. The study successfully identified complex connections among different physicochemical parameters and at the same time ensured high predictive accuracy and strong generalization performance when it formulated groundwater quality evaluation as a binary classification problem and used an ensemble-based CatBoost classifier.

Careful problem formulation and data preprocessing are, in fact, the most important findings of this research, which concede the most decisive role in the success of groundwater quality prediction models. The conversion of the multi-class groundwater quality labels into a binary representation reflecting overall suitability allowed the model to target practical decision-making objectives. This formulation decreased the vagueness in class interpretation and increased the robustness of the model predictions.

The CatBoost classifier showed a strong distinguishing ability, which was evident from the high ROC–AUC values on both training and testing datasets. The close alignment between training and testing performance indicates that the model learned generalized patterns effectively instead of overfitting to noise or dataset-specific artifacts. The use of class weights further facilitated model equity by tackling class imbalance and lessening the chance of misclassifying unsuitable groundwater as suitable.

Another important finding is that discretization and categorical transformation of groundwater quality parameters improved model stability and interpretability. By grouping continuous variables into meaningful bins, the framework reduced sensitivity to outliers and measurement noise while preserving relative differences among observations. This approach proved particularly effective in combination with CatBoost’s native handling of categorical features.

6.2 Interpretation of Model Behavior and Explainability Results

The SHapley Additive exPlanations (SHAP) interpretability analysis gave a profound understanding of the developed model's internal operations. As per the SHAP-based assessment,

a groundwater quality parameters subset was responsible for the classification of the model and played a strong role almost all the time. The parameters linked to salinity, sodium-related hazards, and dissolved species ranked highest among the contributors and were in agreement with the established hydrochemical knowledge of groundwater suitability.

Besides, the directional patterns seen in the SHAP values further support the model's validity. The model classified the groundwater as suitable for certain feature ranges and at the same time the other ranges contributed to its classification as unsuitable. These patterns are in line with the changing quality of groundwater and signify that the model's conclusions are derived from physically meaningful connections rather than false correlations.

Permutation feature importance evaluation backed up the SHAP conclusions and increased the strength of the recognized key parameters' robustness. The coming together of results from two different interpretability approaches not only makes the feature importance conclusions more reliable but also demonstrates the transparency of the proposed framework. Such interpretability is a major advantage in environmental applications, where decision-makers not only need predictions to be accurate but also need a very clear explanation of the factors that drive the predictions.

6.3 Comparison with Conventional and Related Approaches

The proposed framework has several benefits over traditional quality assessment methods that are based on fixed thresholds or subjective indices. Usually, the traditional techniques assess different parameters one at a time and do not take the joint influence on groundwater quality into account. The other way around, the machine learning-based classification framework joins all the currently available variables, so the model can directly learn complex nonlinear interactions from data.

Unlike previous studies that applied machine learning methods and used black-box models or regression to predict continuous indices, the current study is more like a middle way between predictive performance and interpretability. The configuration of a single, well-tuned ensemble model leads to strong accuracy for classification, while at the same time, it is not complicated unnecessarily. Additionally, this research's focus on explainability sets it apart from many others that simply try to achieve the highest accuracy.

The findings indicate that high-quality performance can be obtained without either deep hyperparameter tuning or competing models with different architectures. Rather, the coherence of the methodology—through suitable preprocessing, handling of classes, and interpretation analysis—comes to be the principal contributor in the area of successful groundwater quality modeling.

6.4 Practical Implications of the Study

The outcomes obtained from this study hold considerable significance in terms of the practical application for monitoring and managing the quality of groundwater. The recommended system offers an extremely modern approach to classifying the suitability of groundwater that is based completely on data and is even overridden if needed. It is using parameters that are usually measured such as physicochemical and thus the methodology can be recommended for use by the various stakeholders such as water resource managers, agricultural planners, and regulatory agencies.

Moreover, the framework through its probabilistic classification outputs manages to come up with very flexible decision thresholds applied according to the risk tolerance and management priorities. For example, in areas with a high-level risk of groundwater contamination that subsequently poses equally high agricultural or environmental risks, conservative thresholds can be used. The influential parameters are clearly identified and thus directly supports targeted mitigation, e.g., specific contaminants control or adjustment of the irrigation methods.

Moreover, the dependence on open-source tools and the necessity for only low-level computational resources further contributes to making the framework applicable even in resource-constrained areas. In contrast to deep learning approaches that require large datasets and high computational power, the proposed model can be implemented efficiently while still providing high accuracy and interpretation.

6.5 Conclusion

The research offered a detailed and easily understandable machine-learning framework for the classification of groundwater quality by making use of multi-year groundwater quality data. The reformulation of groundwater assessment to a binary classification problem and the use of a

CatBoost-based ensemble model led the study to successfully tackle the major drawbacks of the traditional groundwater quality evaluation methods.

The results from the experiments indicated that the framework suggested was able to obtain very accurate classification, superb generalization, and very good performance in dealing with class imbalance. The combination of SHAP and permutation feature importance gave meaningful interpretations of the model's actions which further strengthened the transparency and trust in the predictions.

In conclusion, the research shows that properly formulated data-driven classification methods can efficiently assist groundwater quality evaluation and management. The framework suggested creates a practical, scalable, and scientifically validated solution for the detection of suitable and unsuitable groundwater sources.

6.6 Future Scope

The proposed framework, though robust, still has a number of research and application improvement avenues. One of the most significant moves is to add temporal analysis to the existing framework in order to uncover the effects of seasonal and long-term variations on groundwater quality. The modelling of temporal trends could allow a better understanding of the impact of monsoon cycles, climate change, and long-term human activities on the groundwater quality.

Moreover, future research may seek to include more hydrogeological and environmental variables-such as aquifer parameters, groundwater replenishment rates, land-use patterns, and climate indicators. The incorporation of these variables has the potential of not only boosting the predictive capabilities but also offering a more holistic view of the groundwater quality dynamics.

A different, but equally promising, pathway to follow is the application of the framework to multi-class classifications which would allow for a more involved discrimination amongst the different groundwater quality categories. Such a capability would be very useful for the management of the resources where the intermediate quality classes are concerned.

One more thing the future work can focus on is GIS-based spatial analysis and visualization. Such an approach would allow for predictions, localized geographically, that could help in identifying the areas that suffer most from groundwater degradation and thus, in devising the intervention

strategies for those regions. Lastly, the framework could also be modified to suit the requirements of real-time or near-real-time monitoring systems giving the continuous assessment of groundwater quality along with the new data influx.

References

[1] Allawi, M. F., Al-ani, Y., Jalal, A. D., Malik, Z., Sherif, M., & El-shafie, A. (2024). Groundwater quality parameters prediction based on data-driven models. Engineering Applications of Computational Fluid Mechanics. https://doi.org/10.1080/19942060.2024.2364749

[2] Apogba, J. N., Anornu, G. K., Koon, A. B., Dekongmen, B. W., Sunkari, E. D., Obed Fiifi Fynn e, F., & Kpiebaya, P. (2024). Application of machine learning techniques to predict groundwater quality in the Nabogo Basin, Northern Ghana. Heliyon, 10. https://doi.org/10.1016/j.heliyon.2024.e28527

[3] Aslam, B., Maqsoom, A., Cheema, A. L. I. H., Ullah, F., Alharbi, A., & Imran, M. (2022). Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3221430

[4] Bakhtiarizadeh, A., Najafzadeh, M., & Mohamadi, S. (2024). Enhancement of groundwater resources quality prediction by machine learning models on the basis of an improved DRASTIC method. Scientific Reports.

[5] Chowdhury, T. N., Battamo, A., Nag, R., Zekker, I., & Salauddin, M. (2025). Impacts of climate change on groundwater quality: a systematic literature review of analytical models and machine learning techniques. Environmental Research Letters.

[6] Feng, F., Ghorbani, H., & Radwan, A. E. (2024). Predicting groundwater level using traditional and deep machine learning algorithms. Frontiers in Environmental Science. https://doi.org/10.3389/fenvs.2024.1291327

[7] Gad, M., Gaagai, A., Agrama, A. A., El-fiqy, W. F. M., Khadr, M., Abukhadra, M. R., Alfassam, H. E., Bellucci, S., & Ibrahim, H. (2024). Comprehensive evaluation and prediction of groundwater quality and risk indices using quantitative approaches, multivariate analysis, and machine learning models: An exploratory study. Heliyon, 10. https://doi.org/10.1016/j.heliyon.2024.e36606

[8] García, E. M., López, M. I. M., Mateo, L. F., & Quijano, M. Á. (2025). Groundwater quality prediction for drinking and irrigation uses in the Murcia region (Spain) by artificial neural networks. Applied Water Science, 15. https://doi.org/10.1007/s13201-025-02605-z

[9] Haggerty, R., Sun, J., Yu, H., & Li, Y. (2023). Application of machine learning in groundwater quality modeling - A comprehensive review. Water Research, 233. https://doi.org/10.1016/j.watres.2023.119745

[10] Halalsheh, N., Ibrahim, M., Al-shanableh, N., Al-Harahsheh, S., & Al-Mashagbah, A. (2025). Prediction of water quality in Jordanian dams using data mining algorithms. Water Science & Technology, 92(10). https://doi.org/10.2166/wst.2025.158

[11] Holami, V. G., Haleghi, M. R. K., Eimouri, M. T., & Ahour, H. S. (2023). Prediction of annual groundwater depletion: An investigation of natural and anthropogenic inCuences. Journal of Earth System Science. https://doi.org/10.1007/s12040-023-02184-0

[12] Huang, X., Yao, R., Zhang, Y., Li, X., & Yu, Z. (2025). Data-driven prediction modeling of groundwater quality using integrated machine learning in Pinggu Basin, China Xun. Journal of Hydrology: Regional Studies.

[13] Islam, R., Sinha, A., Hussain, A., Deshmukh, K., & Usama, M. (2025). Integrated groundwater quality assessment using geochemical modelling and machine learning approach in Northern India. Scientific Reports.

[14] Khan, I., Nizam, S., Bamal, A., Majed, A., Nash, S., Olbert, A. I., & Uddin, G. (2025). Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid and semi-arid regions of India. Cleaner Engineering and Technology Journal, 26.

[15] Kolli, K., & Seshadri, R. (2013). Ground Water Quality Assessment using Data Mining Techniques. International Journal of Computer Applications, 76(15).

[16] Lokman, A., Ismail, W. Z. W., & 2, N. A. A. A. (2025). A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis. Water.

[17] Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River Water Salinity Prediction Using Hybrid Machine Learning Models. Water.

[18] Priya, R., & Mallika, R. (2017). Ground Water Quality Modelling for Irrigation Using Data Mining Technique and Spatio-Temporal Dates. International Journal of Applied Engineering Research, 12(16).

[19] Raheja, H., Goel, A., & Pal, M. (2022). Prediction of groundwater quality indices using machine learning algorithms. Water Practice & Technology, 17(1). https://doi.org/10.2166/wpt.2021.120

[18] Sajib, A. M., Bamal, A., Diganta, M. T. M., Ashekuzzaman, S. M., Rahman, A., Olbert, A. I., & Uddin, M. G. (2025). Novel groundwater quality index (GWQI) model: A reliable approach for the assessment of groundwater. Results in Engineering, 25. https://doi.org/10.1016/j.rineng.2025.104265

[19] Sajib, A. M., Diganta, M. T. M., Rahman, A., Dabrowski, T., Olbert, A. I., & Uddin, M. G. (2023). Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. Groundwater for Sustainable Development, 23. https://doi.org/10.1016/j.gsd.2023.101049

[20] Sangwan, V., & Bhardwaj, R. (2024). Machine learning framework for predicting water quality classification. Water Practice & Technology, 19(11). https://doi.org/10.2166/wpt.2024.259

[21] Sekar, S., Surendran, S., Debajyoti, P., Perumal, M., Kumar, P., Eldin, H., Arumugam, B., Kamaraj, J., Upendra, B., & Jothimani, M. (2025). Machine learning-based prediction of seasonal groundwater quality for urbanized parts of Melur (Tamil Nadu), India. Results in Engineering, 28(r). https://doi.org/10.1016/j.rineng.2025.108222

[22] Siddiq, B., Javed, M. F., & Aldrees, A. (2025). Machine learning-driven surface water quality prediction: an intuitive GUI solution for forecasting TDS and DO levels. Water Quality Research Journal, 60(4). https://doi.org/10.2166/wqrj.2025.005

[23] Subudhi, S., Pati, A. K., Bose, S., Sahoo, S., Pattanaik, A., Acharya, B. M., & Thakur, R. R. (2025). Prediction of groundwater quality assessment by integrating boosted learning with DE optimizer. Scientific Reports.

[24] Tian, J., Yang, J., Liu, W., Zhang, M., & Daskalopoulou, K. (2025). Assessing groundwater quality for drinking and irrigation using hydrogeochemistry and machine learning in Northern China. Agricultural Water Management, 322. https://doi.org/10.1016/j.agwat.2025.109975

[25] Velmurugan, T., & Arunkumar, R. (2025). Quality based Analysis of Groundwater Data for the Performance of Classification Algorithms. IEEE.

[26] Yang, S., Luo, D., Tan, J., Li, S., Song, X., Xiong, R., Wang, J., Ma, C., & Xiong, H. (2024). Spatial Mapping and Prediction of Groundwater Quality Using Ensemble Learning Models and SHapley Additive exPlanations with Spatial Uncertainty Analysis. Water.

[27] Zhao, X., Chen, W., Tsangaratos, P., Ilia, I., & He, Q. (2025). Landslide spatial prediction using data-driven based statistical and hybrid computational intelligence algorithms. Geocarto International, 40(1). https://doi.org/10.1080/10106049.2025.2507919

[28] Sekar, Selvam, et al. "Machine learning-based prediction of seasonal groundwater quality for urbanized parts of Melur (Tamil Nadu), India." *Results in Engineering* (2025): 108222.

Optimizing groundwater quality.docx

by Turnitin Student

Submission date: 07-Jan-2026 06:06PM (UTC+0900)
Submission ID: 2852824836
File name: Optimizing_groundwater_quality.docx (269.25K)
Word count: 13542
Character count: 82905

TABLE OF CONTENTS

List of Tables

List of Figures

OPTIMIZING GROUNDWATER QUALITY

ABSTRACT

Groundwater quality evaluation is a prerequisite for the sustainable utilization of agricultural and drinking water in the areas where water sources are getting overexploited. The current study presents a framework that is based on machine learning for classifying groundwater suitability utilizing hydrochemical parameters that were obtained from post-monsoon groundwater samples taken in various districts of Telangana, India. The original multi-class irrigation suitability labels were reformulated into a binary classification problem to distinguish between suitable and unsuitable groundwater. A CatBoost classifier was used because of its capability to manage categorical variables as well as to synergize complex feature interactions. Various model performance measures including ROC-AUC, accuracy, F1-score, and confusion matrix analysis were utilized to evaluate the model. The model reached a test accuracy of 97.30% and an F1-score of 0.9375, meaning the predictive power and the balanced classification performance were indeed strong. Model interpretability was assessed using SHapley Additive exPlanations (SHAP) and permutation importance, which pinpointed the common hydrochemical parameters that have the greatest impact on groundwater suitability. The findings affirm that the proposed method is a reliable, interpretable, and pragmatic framework for the classification of groundwater quality and the decision support system in water resource management.

Keywords: *Groundwater quality, Machine learning, CatBoost, Irrigation suitability, Binary classification, SHAP.*

CHAPTER 1 INTRODUCTION

1.1 Background of Groundwater Resources

Groundwater is labeled as one of the most vital sources of freshwater. The support is not only limited to human life but also including agricultural and industrial development as well as ecological sustainability in many areas of the world. Groundwater is the water found under the Earth's crust in soil pores and rock fractures, which means that it is naturally protected and relatively stable compared to the surface water bodies like rivers, lakes, and reservoirs. Its subsurface storage means that it is less subjected to the direct impact of water loss through evaporation and it is also less affected by short-term climatic fluctuations, thus making it the region's most suitable water source, especially in the case of irregular rainfall patterns or drought situations.

On a global scale, groundwater accounts for almost half of the total drinking water needs of the human population and approximately forty percent of the entire irrigation water used in agriculture. In countries that are still developing like India, groundwater is the main source of supply for rural areas and agricultural fields. The significant socio-economic development that has been taking place in the last few decades marked by the increase in population, urbanization, and industrialization as well as the use of more agricultural inputs has increased the dependency on groundwater resources immensely. Groundwater extraction has become the primary way to meet the increasing demand as surface water becomes less available and less certain.

Groundwater resources are very important, and yet they are facing big problems both in terms of quantity and quality. The over-extraction of groundwater beyond the natural recharge rates has led to the decline of water tables in many parts of the world. This depletion not only puts the question of sustainability of the groundwater supplies at risk, but it also has a negative effect on the quality of water. One of the factors that contribute to this problem is the increase in the concentration of dissolved minerals as groundwater levels go down, and this, in turn, leads to salinity problems that affect both the quality of drinking water and the suitability of land for agricultural activities and the health of ecosystems. Groundwater pollution, unlike the surface water pollution, which is usually obvious and very quickly detectable, is a slow and hidden process. The underground characteristics of aquifers let the pollutants remain undetected for a long time, which of course,

makes it extremely hard and very expensive to carry out remedial actions after pollution has taken place. Hence, the early detection and the very reliable assessment of groundwater quality become very necessary.

The decline in the quality of groundwater is the result of a mix of both natural and human factors. The natural processes such as mineral dissolution, rock-water interactions, and geochemical evolution, among others, determine the baseline composition of groundwater. However, the human activities nowadays are significant contributors to the contamination of groundwater. Agriculture is a significant contributor to the problem with the chemical-heavy runoff that carries fertilizers and pesticides, and industrial discharges, urban sewage, landfill leachate and improper waste disposal are all responsible for drastic changes in groundwater chemistry. These pressures make it essential to develop advanced monitoring and predictive tools that will be able to support the sustainable groundwater management and informed decision-making.

1.2 Importance of Regional Groundwater Quality Assessment

One of the main factors that determine the quality of groundwater is its suitability for different uses such as irrigation and livestock consumption. In agricultural areas, groundwater is frequently applied to fields without any major treatment. The use of contaminated irrigation water can lead to the destruction of the soil, and loss of crops, and eventually the degradation of the land. In addition, the quantity and quality of groundwater determine the lifespan of the livestock. If the dissolved solids or the chemical constituents are too high, it may lead to health issues and lower productivity besides causing health disorders or lowering the productivity of the animals.

The spatial and temporal variability of the regional groundwater systems is substantial and is mainly influenced by the geological and soil characteristics, land use, and climatic conditions. This means that the different regions of groundwater quality are very complex and move gradually. In areas that are highly dependent on agriculture and groundwater supply, such variability impacts directly on crop planning, irrigation, and resource allocation. Therefore, systematic groundwater quality assessment and prediction at the regional level are necessary to achieve sustainable use.

Groundwater quality classification schemes are the main tool for categorizing water based on chemical characteristics and the potential use. Classification systems, which include salinity, sodium hazard, residual sodium carbonate, and total dissolved solids, allow for the identification

of water that is suitable, marginal, or unsuitable for irrigation and livestock use. Availability of accurate classification provides a basis for practical decision-making by farmers and water managers. Thus, they can decide on the type of crops, soil management practices, or groundwater use restrictions where necessary.

Given the increasing pressure on groundwater resources, there is a critical need for data-driven approaches that can analyze complex groundwater quality datasets and generate reliable predictions. Such approaches can assist in identifying vulnerable regions, understanding dominant influencing factors, and supporting proactive groundwater management strategies.

1.3 Limitations of Conventional Groundwater Quality Assessment

Conventional methods for groundwater quality assessment traditionally consist of taking samples for laboratory analysis of individual physicochemical parameters and subsequently comparing the results with the prescribed national or international standards. These methods, though very accurate, are still afflicted with several limitations when it comes to their application in large and complex groundwater systems because they only give measurements for specific parameters and locations at specific moments.

One of the most important drawbacks of the traditional assessment methods is their inability to represent the total impact of several parameters that interact with each other. Groundwater quality is determined by a set of factors that are quite complicated and that take non-linear forms among the constituents of the chemical stuff. Evaluating parameters one by one may not reveal the true quality of the water, especially when several factors together determine the case for irrigation or livestock use. Moreover, the availability of large datasets concerning groundwater quality is increasing, but traditional methods cannot efficiently deal with high-dimensional data because they are not designed that way.

Another limiting aspect is the lack of predictiveness. Where traditional methods give a picture of the groundwater quality at a particular time, these methods tell nothing about future conditions or trends. This characteristic makes it difficult for policymakers and water suppliers to implement preventive measures or foresee possible deterioration of water quality. Acompanying this, many classical statistical methods are based on the assumption of linear relationships among variables;

this is an unrealistic assumption in the case of groundwater systems that are subject to complex hydrogeochemical processes.

These limitations highlight the necessity for advanced analytical methods that can integrate multiple parameters, accommodate nonlinear relationships, and provide predictive insights to support sustainable groundwater management.

1.4 Data-Driven Groundwater Quality Classification

The adoption of data-driven approaches for groundwater quality evaluation has been made possible by recent advancements in data availability and computation. Data-driven classification models, as opposed to deterministic thresholds or subjective weighing schemes, utilize the observed patterns in historical data to classify different types of groundwater according to their quality.

Groundwater quality datasets usually consist of many chemical parameters, and often more than one of those parameters are interrelated and/or have nonlinear interactions. ML-based classification frameworks are the most appropriate ones for this type of data because they are able to infer complex relationships from the data without imposing any physical or chemical assumptions. Hence, data-driven models can effectively map the multidimensional quality measurements of groundwater into the quality classes that are relevant and can be applied in practice.

The use of binary classification frameworks which classify groundwater into large classes of suitable and unsuitable, is a very realistic way of handling things in the field. The classification of the water quality information is simplified thereby the essential distinctions that have to do with irrigation and livestock use are still retained. The da-ta-driven binary classification models can accurately assist groundwater management in the region by pointing out the areas where water quality is of little risk and the areas where caution or intervention is required.

1.5 Application of Machine Learning in Groundwater Quality Prediction

Environmental monitoring data becoming more and more accessible, alongside the enhancements of computational power, has brought the utilization of machine learning techniques in groundwater quality prediction to the forefront of the technology. The capability of machine learning algorithms

to unveil hidden patterns and relationships in massive datasets, so to be more fitted for the modeling of groundwater systems, which are typically defined by nonlinear behaviors and complex interactions among the variables.

The prediction of groundwater quality through machine learning models has the chemical parameters of the water integrated all at once, so that the suitability of the water can be assessed in a comprehensive manner. Besides, machine learning methods have been proven to be more efficient than traditional statistical techniques in the aspects of noisy data handling, mixed data type accommodation, and nonlinear dependencies capturing. The advantages mentioned earlier have been the driving force for the growing popularity of machine learning methods in environmental modeling and water resource studies.

Ensemble-based algorithms, among the various machine learning techniques, are the ones that have been highly appreciated due to their capability to yield robust and predictive performance in the case of structured tabular data. To reduce the variance and enhance the generalization, the model combines several decision structures. Also, the current machine learning frameworks are increasingly focusing on interpretability, which allows the researchers to understand the individual features that have contributed to the predictions of the model. Interpretability is of utmost importance in the environmental applications, where transparency and trust in the model outcomes are necessary for the policy and management decisions.

1.6 Emergence of Data-Driven and Machine Learning Approaches

The technological progress in the field of environmental monitoring and data storage made it possible to have huge, multivariate groundwater quality datasets collected from various places and times. The presence of such data has led to the use of machine learning methods for groundwater quality evaluation and forecasting. Such models can recognize complicated patterns in the past data and also work with nonlinear relationships among variables without any strict statistical assumptions.

In studies on groundwater quality, machine learning techniques have outperformed traditional statistical models significantly, especially in the case of heterogeneous and high-dimensional data. These techniques can concurrently consider a variety of chemical parameters in the process of making predictions and can even do so reliably despite the presence of noise and missing values.

Thus, machine learning has become a strong and effective tool for, among others, groundwater quality classification and decision support.

On the other hand, machine learning models have their downsides, such as the necessity of thorough data preprocessing and the possibility of loss of interpretability. In environmental models, the openness of the model is of utmost importance because the stakeholders need clear and easy to understand explanations of why some groundwater samples are labeled as good or bad. This is the reason, the groundwater quality studies of the present day more and more often resort to the machine learning models that possess both high predictive performance and good interpretability, thus allowing accurate classification and at the same time giving meaningful insight into the underlying factors that drive the process.

1.7 Need for a Robust and Interpretable Classification Framework

A proficient groundwork quality forecasting system should have the ability to convert complex multivariate data into simple and usable classifications. Instead of depending only on continuous index values or individual parameter thresholds, classification-based approaches provide a more practicable way of categorizing groundwater into the two widely suitable and unsuitable classes. Such binary classification systems become an easier option for decision-making while still keeping the necessary information concerning irrigation and livestock usage.

A strong classification system must be able to consider both the quality parameter nature and the type being either categorical or numerical. A lot of the groundwater datasets are of the mixed type which consists of both the continuous chemical concentrations and the categorical descriptors. Models that can deal with such mixed data as a whole without going through a lot of manual encoding are highly beneficial. Besides, tackling the issue of class imbalances is very significant since datasets often contain unequal numbers of samples for the various quality classes.

Also, model interpretability is of major importance. Knowing the parameters that have the highest positive or negative impact on groundwater quality classification increases the confidence in model predictions and provides very useful information for water resource management. Feature attribution and importance analysis make it possible for researchers and policymakers to spot the influential factors and to plan the mitigation measures according to their priority.

1.8 Model Evaluation and Explainability

Predictive models made possible by machine learning, though very powerful, still need to undergo stringent evaluation and transparent interpretation for their reliability. One of the main purposes of performance evaluation is to make sure that the predictive models generalize well to future data and, therefore, do not suffer from overfitting. Accuracy, F1-score, and the area under the receiver operating characteristic curve metrics, among others, have been widely used to characterize the performance of a model in classification more so when the classes are imbalanced.

In the environmental sciences, explainability has become a major concern in machine learning technology along with predictive accuracy. The use of explainable models makes it possible to point out the most important parameters that influence the classification of groundwater quality thus leading to better scientific understanding and making it easier for the authorities to make informed decisions. Thanks to feature attribution techniques, the researchers can technically determine how much each variable has contributed to the model's predictions, hence the winning scenario of performance and interpretability in the specific domain.

The combination of strong evaluation metrics with the use of machine learning techniques that are explainable has resulted in the development of a reliable and understandable method for predicting groundwater quality.

1.9 Organization of the Thesis

The present dissertation is divided into six polar opposite chapters, where each chapter highlights a different aspect of the groundwater quality classification study and provides together a logical flow from problem motivation through methodological development, experimental validation, and final conclusions to the latter part.

Chapter 1: Introduction

This chapter presents the background and reason for conducting the study by pointing out groundwater's indispensable nature for agriculture, human consumption, and even environmental sustainability. It narrates the mounting difficulties concerning the deterioration of groundwater quality due to the mix of natural and man-made processes. The chapter points out the drawbacks of the traditional groundwater quality assessment techniques that often base their conclusions solely on isolated parameter thresholds and have no capability of prediction. The shortcomings

outlined in the chapter are taken care of by the introduction of machine learning as a robust data-driven coupling in groundwater quality classification. The research goals, limits of the study, and the overall importance of the proposed classification framework are thoroughly given to frame the context of the following chapters.

Chapter 2: Literature Review

A thorough review of the literature is presented in this chapter, which covers the research works that are currently available on the groundwater quality evaluation, the approaches for classification, and the use of the machine learning techniques in the environmental and hydrogeological studies. The discussion encompasses the application of traditional water quality indices, the statistical methods, and the recent developments in the area of machine learning-based groundwater analysis. The modeling approaches are thoroughly evaluated from the point of their strengths and weaknesses, and among others, issues like subjectivity, interpretability, and generalization are the ones that receive most of the attention. Consequently, the authors draw up a list of the main research gaps, based on existing literature, which support the development of a robust, interpretable, and classification-oriented groundwater quality assessment framework as the next step in the research.

Chapter 3: Research Objectives and Problem Formulation

The research problem the thesis addresses is clearly stated in this chapter and groundwater quality assessment is formulated as a binary classification task aimed at facilitating decision-making in practice. The specific research objectives are presented and the reason for changing the multi-class groundwater quality labels to a suitability-based classification framework is explained. Besides, the chapter describes the conceptual and analytical framework that has been adopted in the research, linking the identified research gaps with the proposed solution very clearly.

Chapter 4: Methodology

In this chapter, the researcher explains the methodological framework step by step, from the characteristics of the dataset, to the preprocessing of the data, the techniques for transforming features, and the processing of categorical and numerical variables. The creation of the machine learning model is illustrated along with the class imbalance problem strategy and evaluation

method for model performance that were used. The selection of the CatBoost classifier is also mentioned as well as its taken benefits for tabular and categorical data.

Chapter 5: Experimental Results and Analysis

The experimental results produced by the proposed groundwater quality classification model are disclosed in this chapter. The metrics of accuracy, F1-score, ROC–AUC, and confusion matrix analysis are among the multiple metrics used to evaluate the model performance. The chapter also goes on to feature interpretability through the application of analysis for contribution of features and to discuss the role of groundwater quality management in drawing the conclusions from the findings.

Chapter 6: Conclusion and Future Work

The study's major findings are concisely stated in the last chapter, and the proposed classification framework's contributions are highlighted. It covers the research results for the fields of groundwater quality assessment and decision-making and sketches out future research trails such as model updates and adding more data sources that are likely.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

Groundwater is a vital fresh water source that plays a major role in the drinking water supply, agricultural irrigation, and industrial activities across the globe. In many underdeveloped and arid regions, groundwater is the main source of drinking water because of the lack or unavailability of surface water. Nevertheless, all these issues have collectively caused severe pressure on groundwater systems and the aquifers have been contaminated with a range of pollutants. The latter include nitrates, fluoride, salinity, heavy metals, and toxic ions, all of which have been recognized as a serious threat to human and animal health, the environment, and agricultural productivity.

Traditional groundwater quality assessment methods depend on physicochemical analysis and deterministic hydrogeochemical modeling. These methods are scientifically valid but at the same time they can be quite slow, data-hungry, and unable to represent the nonlinear relationships that exist among the water quality parameters. To make the complex datasets easier to handle, WQI has been widely used as a support for decision-making. However, the traditional WQI approaches have the limitation of being subjective in their weighting, having fixed thresholds, and being uncertain in the classification especially when it comes to trace elements and complex hydrochemical interactions.

Recently, the research has been directed towards the combination of ML, AI, and data mining techniques for the groundwater quality assessment as a means of coping with these restrictions. The machine learning models have shown to be very powerful in capturing non-linear connections, dealing with high-dimensional data, and increasing the accuracy of predictions. The use of Water Quality Index (WQI) frameworks along with ML predictive models has been recognized as an efficient method for groundwater quality monitoring, spatial mapping, vulnerability assessment, and sustainable management. The chapter summarizes the literature concerning groundwater quality prediction and points out the methodological developments, machine learning applications, hybrid and ensemble models, spatial analysis, and emerging research trends.

2.2 Literature Review

Groundwater quality assessment has undergone a significant transformation in recent years, driven largely by advances in data acquisition, computational power, and analytical methodologies. Traditionally, groundwater quality evaluation relied heavily on index-based approaches, deterministic models, and regulatory threshold comparisons. While these conventional methods served as effective tools for baseline assessment and compliance monitoring, they often lacked the ability to capture complex interactions among multiple hydrochemical parameters or to provide reliable predictive insights. As a result, researchers have increasingly shifted toward data-driven and machine learning-based frameworks to address these limitations.

A significant breakthrough in groundwater quality assessment is Groundwater Quality Index (GWQI) which is an upgrade from the traditional water quality indices. The traditional WQI models assess primarily major ions and physicochemical parameters while GWQI framework adds trace elements and metals hence presenting a more thorough evaluation of groundwater for human consumption and agricultural use. Sajib et al. were among the first to employ the modern GWQI model using a machine learning-based assessment framework integrating ten water quality indicators out of which six were heavy metals [1]. The researchers trialed a variety of machine learning algorithms in order to ascertain the predictive strength of the GWQI model proposed. The outcomes revealed that artificial neural networks (ANN) attained remarkably low prediction errors, high sensitivity, and very little risk, thus considerably surpassing other models. The permitting of heavy metals highlighted a pertinent lucrativeness in research by acknowledging the health hazards linked with trace contaminants that are generally unnoticed due to their low concentrations and invisibility.

In continuation of the original study, Sajib et al. broadened their research very much by testing the toughness and generality of the GWQI framework on several datasets of groundwater quality monitoring from different parts of Ireland [2]. The large-scale investigation also tested the GWQI-based classification against the conventional approaches of regulation, it was composed among others of the "One Out, All Out" method which has been the most popular one. The results indicated that the GWQI model was able to produce correct and reliable groundwater quality classification that was superior to those obtained by the traditional regulatory methods. In addition, Optuna-optimized Gradient Boosting models showed the best predictive performance, confirming

the significance of hyperparameter tuning for improving prediction accuracy and lowering uncertainty. Thus, the combination of objective index formulation with advanced optimization techniques was necessitated to get reliable groundwater quality assessment, as stated by this study.

The use of machine learning methods for groundwater quality prediction has received much attention in different hydrogeologic environments. Huang et al. (2020) present the results of their research on groundwater quality in Pinggu Basin, China. They have utilized several advanced data-driven techniques, namely Support Vector Machines (SVM), Random Forest (RF), Back Propagation Neural Networks (BPNN), and Convolutional Neural Networks (CNN) [3]. The authors have not only predicted the high groundwater quality with the probability of over 95% but also demonstrated the power of machine learning in revealing the intricate relationships between hydrochemistry. Heavy nitrate contamination was characterized as the major factor determining groundwater quality mainly due to the use of fertilizers and farming practices. BPNN was the model with the highest prediction accuracy among the tested ones which again evidenced the skill of neural networks in handling the nonlinearities that are a part of the groundwater systems.

In a similar vein, Tian and colleagues utilized hydrogeochemical analysis in conjunction with six machine learning techniques to appraise the quality of groundwater in the agropastoral regions of Northern China for drinking and irrigation purposes [4]. The authors' approach involved the application of entropy-weighted water quality indices and irrigation water quality indices to groundwater sample classification, which unmasked considerable spatial variability in the quality of groundwater. The outcome revealed that ANN rendered the most accurate representation of drinking water quality, while XGBoost was in command of irrigation suitability evaluation. This discovery confirmed that the performance of the model is case-specific and the quality assessment's intended use plays a crucial role in determining the decision.

On the other hand, seasonal variability is a factor affecting groundwater quality that has, among other things, been addressed through machine learning approaches. Sekar and co-workers tested the prediction of seasonal groundwater quality in the urban area of Melur, Tamil Nadu, India, using the models that included ANN, Decision Tree, RF, and XGBoost [5]. They found out that ANN was the one with the highest predictive accuracy in pre-monsoon seasons, while XGBoost showed its excellence in post-monsoon periods. This separation of the seasons underlined the dynamic character of groundwater systems and indicated that it is unlikely for one model only to be the best

one for all the temporal conditions. The research established the necessity of considering seasonal variability in the water quality prediction system.

The hybrid and ensemble learning approaches have been receiving more and more attention because of their quality to increase predictive power by joining the advantages of several models. Melesse et al. were investigating the application of different hybrid machine learning models like the combination of the additive regression with M5P and RF to predict the electrical conductivity of long-term water quality datasets [6]. The findings showed that the hybrid models had a considerable advantage over the individual algorithms, with the AR-M5P model having the best accuracy. Thus, reducing the prediction error was one of the benefits that hybrid modeling techniques were able to offer in the field of groundwater quality data analysis.

More advanced ensemble and boosted learning methods have also improved the process of modeling groundwater quality. Subudhi et al. came up with a hybrid LCBoost Fusion model that merges CatBoost and LightGBM for the predicting groundwater quality indices in the chromite mining area of Odisha, India [7]. The model that was suggested not only achieved a high level of accuracy but also was able to pinpoint the contaminants that were most influential: potassium, fluoride, and total hardness. The research provided evidence for the use of advanced boosting techniques in the context of real-time groundwater monitoring and risk mitigation in mining-affected areas in an efficient way.

Optimization techniques have been essential to the performance enhancement of the machine learning model used in groundwater quality prediction. Siddiq et al. fused together Particle Swarm Optimization (PSO) with Random Forest and Gene Expression Programming models in predicting the concentrations of total dissolved solids and dissolved oxygen [8]. These optimized models got almost perfect accuracy in predictions, thus demonstrating the power of evolutionary optimization algorithms in water quality prediction based on machine learning.

The researchers, therefore, concluded that optimization techniques should always be paired with predictive models to attain high-quality and durable performance. The integration of spatial prediction and the application of GIS have become the most important facets of groundwater quality assessment. Groundwater quality mapping in the Nabogo Basin, Ghana was done by Apogba et al. using Random Forest models combined with GIS techniques [9]. The research was

able to predict groundwater quality index spatial variations successfully and thus, the role of machine learning-based spatial mapping in regional groundwater utilization was also highlighted. Spatial analysis contributed a lot to the enhancement of the predictive framework's applicability in the decision-making and policy-making process.

To tackle spatial uncertainty & model interpretability, Yang et al. merged entropy-weighted water quality indices with LightGBM, the pressure–state–response framework and SHapley Additive exPlanations (SHAP) analysis [10]. This study pointed out nitrate, magnesium, sulfate, and chromium as the most burning contaminants and advocated for the role of explainable machine learning methods in grounding the role of human activities in the decline of groundwater quality. The paper under discussion reflected the ongoing trend towards a higher degree of openness and interpretability in machine learning-powered environmental research. AI techniques in groundwater depletion and vulnerability analysis have been widely used besides the quality assessment. The prediction of the groundwater depletion in a Jaipur alluvial aquifer near the Caspian Sea was performed using a combination of RF, deep learning, and XGBoost models by Holami et al. [11]. They found aquifer transmissivity, groundwater withdrawal rates, and groundwater depth as the major controlling factors. Besides, the combination of machine learning with GIS-based spatial prediction proved to be an effective method of the new data-driven techniques in managing the groundwater resource under increasing extraction pressure.

The consequences of climate change on the quality of groundwater have started to be assessed more and more through the use of machine learning and analytical models. Chowdhury et al. gave a critical review of the situation by pointing out the difficulties involved in the modeling of the degradation of groundwater quality due to climate change and called for the use of both adaptive and hybrid modeling techniques [12]. In the same way, Bakhtiarizadeh et al. made use of soft computing methods to evaluate the vulnerability of the groundwater resource based on enhanced DRASTIC and nitrate vulnerability indexes, and they got very high predictive accuracy [13]. The studies mentioned here have already pointed out the necessity for the incorporation of climatic variables and vulnerability assessment routes into the groundwater quality model.

There have been several articles that have concentrated on the groundwater pollution in India through the application of machine learning approaches. Islam and his colleagues combined geochemical modeling with ANN, RF, and XGBoost in the evaluation of groundwater quality in the state of Uttar Pradesh, which led to the discovery of serious contamination caused by high levels of total dissolved solids and fluoride [14]. In the same vein, Khan et al. used XGBoost-optimized RMS-WQI models for the monitoring of groundwater quality in Rajasthan and obtained very good sensitivity as well as very little uncertainty [15]. Sangwan and Bhardwaj reported that models based on SVM were better than all other algorithms used in the classification of the quality of groundwater over large datasets from the states of Maharashtra and Dadra and Nagar Haveli [16]. The above-mentioned studies have together shown the increasing application of machine learning methods for the assessment of groundwater quality in different Indian hydrogeological environments.

In the first place, comprehensive review studies have not only compiled the advancements in groundwater quality modeling but also opened up the areas for further research by revealing the trends and gaps [1]. Haggerty et al. gave a detailed overview of the application of machine learning methods to groundwater quality modeling. They observed the transition from traditional neural networks through deep learning and explainable artificial intelligence [17]. The work of Lokman et al. focused on machine learning and statistical methods for water quality forecasting and on the major hurdles concerning data quality, interpretability, and spatio-temporal integration [18]. The remarkable studies of Priya and Mallika [19] and Kolli and Seshadri [20], for example, are still part of the contemporary water quality research as they laid the foundation for data mining and spatio-temporal modeling approaches.

In general, the scientific community has moved from relying on traditional index-based groundwater quality assessment methods to utilizing integrated, data-driven, and machine learning-based frameworks. The transition was not without drawbacks, however, as challenges of developing objective indices, model interpretability, uncertainty quantification, and generalizability arose. Nevertheless, these very research gaps supplied potent motives for the current experimental work, which aims at the combination of entropy-based water quality indexing, robust machine learning models, and comprehensive validation strategies to be used for the purpose of optimizing the prediction of coastal groundwater quality.

2.3 Summary

In this chapter, a very detailed and wide-ranging examination of the current literature on groundwater quality assessment and forecasting was conducted, highlighting the move from classic hydrogeochemical and index-based methods to machine learning-powered frameworks. The papers that were examined indicate that the machine-learning models are a lot more accurate in terms of predictions, robustness, and adaptability in different hydrogeological and climatic conditions. Artificial Neural Networks, ensembles such as Random Forest and XGBoost, and hybrid learning architectures always win the battle against the solo models, especially when they are used with the water quality indices that are fine-tuned.

Machine learning along with Water Quality Indices has conquered major drawbacks concerning uncertainty, sensitivity, and misclassification in conventional assessment techniques. The present-day developments that have come about through the use of spatial analysis, GIS integration, explainable AI techniques, and even optimization algorithms are the ones that have made the reliability and interpretability of the predictions of groundwater quality stronger. The innovations notwithstanding there are still some hurdles which are data scarcity, spatial uncertainty, seasonal variation and climate change effects that need to be overcome.

Literature review has done a great job in establishing a strong scientific foundation for the present research and also in pointing out the need for developing powerful, interpretable, and data-driven groundwater quality forecasting models that would assist sustainable groundwater management as well as the decision-making process with their being scientifically based.

CHAPTER 3 RESEARCH PROBLEM AND OBJECTIVES

3.1 Background of the Research Problem

Groundwater is the most essential freshwater source, such as drinking, agricultural irrigation, and livestock consumption, especially in places with limited seasonal or unreliable surface water supply. In India, many semi-arid and agrarian areas rely mainly on groundwater for water supply for crops and feeding of rural communities. The quality of groundwater directly affects its different uses, and bad-quality groundwater can cause soil erosion, reduce the output of crops, and damage the health of livestock and the environment in the long run.

Over the past few decades, the quality of groundwater in many areas has been significantly impacted by the ever-increasing population, intensive agricultural practices, and, most importantly, the unregulated extraction of groundwater. The natural geochemical processes and human activities like the use of fertilizers, improper waste disposal, and overexploitation of aquifers influence the chemical makeup of groundwater. When the groundwater table drops and the water-rock interactions become stronger, the tendency for the accumulation of dissolved salts and other chemical constituents increases, and as a result, the groundwater often becomes unsuitable for irrigation or livestock use.

Conventional methods for evaluating groundwater quality have mostly been dependent on comparing individual parameters with standard threshold limits. These methods are good at monitoring compliance, but they do not properly reflect how multiple water quality parameters together influence the overall usability. Besides, these methods only offer a static picture of groundwater quality and do not show the complex interrelationships between the physicochemical variables. The increasing availability of multi-annual groundwater quality datasets has created a need for analytical frameworks that are able to efficiently process high-dimensional data and provide a classification of groundwater quality that is reliable for practical decision-making.

3.2 Identification of the Research Gap

The scrutinization of current groundwater quality studies pointing out the critical aspects reveals the presence of research gaps. The majority of studies undergo a descriptive assessment or an index-based evaluation, but they do not take advantage of the predictive competences of machine

learning. Some studies utilize machine learning models, but they consider those models to be black boxes, thus offering little interpretability and not enough insight regarding the impact of particular water quality parameters.

Moreover, several studies only use small datasets or have limited time coverage, which affects the applicability of their results. In many instances, groundwater quality classification is done without considering class imbalance, which may lead to model predictions biased toward the dominant classes. Besides, it is only a handful of studies that systematically integrate robust ensemble-based machine learning models with explainability techniques to generate both accurate and interpretable groundwater quality classifications using large, multi-year datasets.

Thus, these gaps underscore the necessity for a reproducible, data-driven groundwater quality classification framework which merges strong machine learning techniques with clear interpretability mechanisms and is able to cope with large-scale groundwater quality data.

3.3 Statement of the Research Problem

Considering the previous discussion, the principal research issue treated in the present study can be formulated this way:

How is it possible to classify groundwater quality in a dependable and precise manner through a data-driven machine learning framework that fully incorporates the multivariate physicochemical parameters and at the same time provides interpretability and robustness of predictions?

This research problem suggests a unified classification approach with the ability to deal with complex groundwater quality data, recognize nonlinear relationships between the different classes, even out class imbalance, and give understandable reasons for model decisions.

3.4 Research Objectives

The main target of this investigation is to create and assess a powerful machine learning-based system for classifying groundwater quality using data from multiple years' groundwater quality monitoring. In order to realize this main goal, the study aims to carry out the following specific objectives:

1. To clean and merge multi-year groundwater quality datasets into a single analytical framework that is suitable for machine learning applications.

2. To change the classification of groundwater quality into a binary decision problem that shows overall suitability for practical application.
3. To construct an ensemble-based machine learning classification model that can manage mixed data types and class imbalance.
4. To measure the forecast performance of the proposed model through the use of statistical classification metrics.
5. To thoroughly examine and explain the model's predictions employing feature attribution and importance analysis methods.
6. To pinpoint the main groundwater quality parameters affecting classification results along with their respective contributions.

3.5 Significance of the Study

The present research has great importance due to its methodological rigor and practical relevance in the field of groundwater quality management. Groundwater quality evaluation is a very difficult task because of the multivariate nature of hydrochemical parameters and their nonlinear interactions. The study presents a data-driven classification framework, which using machine learning techniques particularly designed for structured tabular data has minimised the drawbacks of traditional groundwater quality assessment methods.

Taking a methodological step, the study reveals the power of an ensemble-based classification method that can deal with both categorical and numerical features natively and also solve class imbalance problems via weighted learning. The application of explainability techniques allows for transparent understanding of model predictions, making the framework not only accurate but also understandable to both experts in the field and decision-makers.

In a practical way, the new framework allows for the early detection of groundwater quality deterioration and the taking of informed decisions as regards irrigation and livestock. The framework is built on commonly monitored groundwater quality parameters and free-of-charge analytical tools, so it is economical, scalable, and applicable to areas with little monitoring infrastructure.

3.6 Scope of the Research

The research scope is directed towards a machine-learning-based groundwater quality classification framework that will be developed and evaluated using physicochemical parameters from multi-year groundwater quality datasets. The investigation points out the data preprocessing, feature transformation, model development, performance evaluation, and interpretability.

This work does not include spatial interpolation, GIS-based analysis, and detailed hydrogeochemical process modeling. The study does not aim to conduct causal analysis on particular contaminants or real-time monitoring, but rather it gives priority to the development of a general, flexible, and data-driven classification approach that is applicable to different groundwater quality datasets.

However, the framework that has been created in this study is a solid base for further developments, like the analysis of temporal trends, spatial mapping, climatic variable integration, and the use of the system in support of groundwater resource management decisions.

CHAPTER 4 Methodology

4.1 Overview of the Methodological Framework

This research has set a goal to create a trustworthy framework for groundwater quality classification based on data and utilizing machine learning. The approach used in this research operates through a well-organized and sequential pipeline that starts with data gathering and preprocessing, followed by feature transformation, model creation, performance testing, and model interpretability analysis. The methodologies used in each of the stages have been designed very meticulously to be congruent with the characteristics of groundwater quality data, and at the same time, these methodologies have been made to guarantee reproducibility, robustness, and transparency of the results.

The new methodology is in stark contrast to the classic groundwater quality evaluation methods that base their judgments on either predetermined threshold values or the subjective weighting of parameters. The new approach based on supervised machine learning can now be used to classify very accurately because it learns classification patterns directly from the observed data. The framework especially targets the simultaneous management of several groundwater quality parameters, the resolution of class imbalance, and the guarantee of the interpretability of the model predictions. The intentional omission of figure-wise representations and spatial analyses from the methodology is to ensure that the focus remains on data-driven classification and explainability.

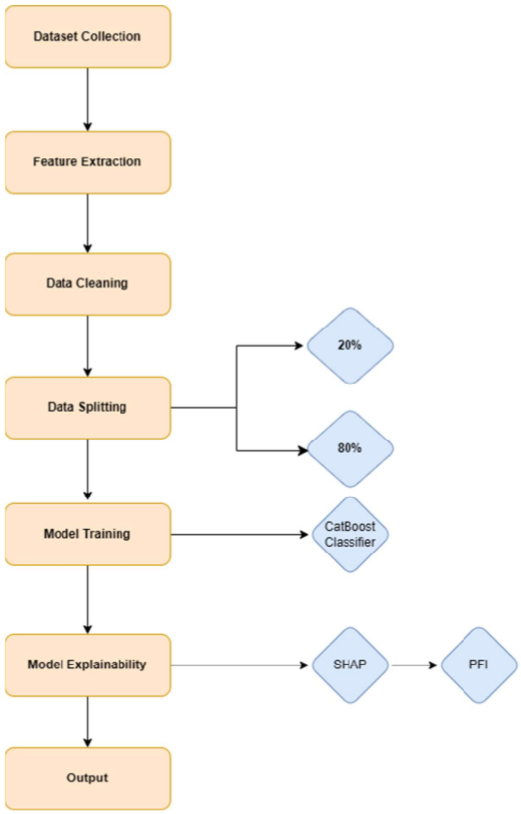

Figure 4.1 Methodology Flowchart

4.2 Data Source and Description

4.2.1 Data Collection

This research dealt with a dataset from Telangana Open Data Portal, Government of Telangana, India. The data included post-monsoon observations of groundwater quality collected from different districts of the state. Groundwater was analyzed for a range of physicochemical parameters that are crucial for irrigation and animal husbandry.

The dataset has been formed by three different annual files representing the years 2018, 2019, and 2020, such as:

- ground_water_quality_2018_post.csv
- ground_water_quality_2019_post.csv
- ground_water_quality_2020_post.csv

Each file represents the groundwater quality during the post-monsoon season, thus adding up to seasonal consistency across years. The use of multi-year data will enable the model to learn generalized patterns rather than focusing only on anomalies specific to the year.

4.2.2 Dataset Integration

In order to create a comprehensive dataset, the three annual datasets were combined and turned into one single dataframe. The integration was done vertically by stacking the rows of each year while keeping a common set of features for all records. Only the columns that were common in all the datasets were kept in order to prevent inconsistencies.

As a result of this approach, a dataset of groundwater samples spanning three years was obtained which increased the sample size and made the machine learning model more robust. The final integrated dataset is capable of capturing Telangana's groundwater quality spatial and temporal variability.

4.2.3 Feature Description

The datasets consist of 26 columns including:

- Identification attributes.
- Location descriptors (district, mandal, village, latitude, longitude)
- Physicochemical parameters like calcium (Ca), magnesium (Mg), carbonate (CO_3), bicarbonate (HCO_3), total dissolved solids (TDS), residual sodium carbonate (RSC), sodium adsorption ratio (SAR), and total hardness
- Groundwater level (GWL)
- Seasonal information

- Target variables: Classification and Classification1

The Classification variable divides groundwater into nine classes (C1S1 to C4S4) according to their salinity and sodium hazard. The classification mirrors the suitability of groundwater for irrigation and agricultural use.

Table 1 Dataset Sample

Parameter	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
S. No.	302	312	129	123	122
District	SIRCILLA	MAHABUBNAGAR	MAHABUBNAGAR	BHADRADRI	MAHABUBNAGAR
Mandal	Illanthakunta	Mahabubnagar (R)	CC Kunta	Manuguru	Bhoothpur
Village	Illanthukunta	Kodur	Kurumurthy	Pagideru	Elkicherla
Latitude (°N)	18.311944	16.680000	16.441442	17.970000	16.623000
Longitude (°E)	78.956388	77.924000	77.816757	80.730000	78.116000
Groundwater Level (m)	3.51	7.66	19.73	0.87	3.21
Season	Post-monsoon 2020	Post-monsoon 2019	Post-monsoon 2019	Post-monsoon 2020	Post-monsoon 2020
pH	8.02	7.95	7.79	8.12	8.98
TDS (mg/L)	485.12	852.48	811.52	1048.32	524.80
Total Hardness (mg/L)	219.96	379.93	439.96	339.92	219.93
SAR	2.316	3.132	1.605	4.217	2.697

RSC (meq/L)	0.0008	0.4015	−2.9992	−0.9984	0.8013
Irrigation Class	C3S1	C3S1	C3S1	C3S1	C3S1
Water Quality Status	P.S.	P.S.	P.S.	P.S.	P.S.

4.3 Problem Formulation and Target Variable Definition

4.3.1 Reformulation of the Classification Task

The nine groundwater quality classes present in the original dataset are now reframed as a binary classification problem in the current research. The reason for such reframing is the practical decision-making requirement, whereby groundwater is often classified in one of the two categories: suitable or unsuitable for use.

The bi-class target variable was specified as follows:

- **Good (1):** C1S1 and C2S1
- **Bad (0):** All the other classes (C3S1, C3S2, C3S3, C4S1, C4S2, C4S3, C4S4)

This categorization depicts irrigation-safe groundwater versus risky groundwater in terms of salinity or sodium.

4.3.2 Justification of Binary Classification

Binary classification streamlines groundwater quality assessment and makes it more interpretable for policymakers and farmers. Instead of dealing with a bunch of classes that are only slightly different, the stakeholders can just find out whether the water is ok or not for the use or if it only requires a bit of caution or even remediation. This arrangement is also very compatible with supervised machine learning frameworks and allows for performance evaluation using standard classification metrics.

4.4 Data Preprocessing

Data preprocessing is a very important step to ensure that the model is both performing well and is trustworthy. The data about groundwater quality might consist of different feature types, have different scales, and missing values. The preprocessing stages that were applied in this research are briefly discussed below.

4.4.1 Removal of Non-Informative Features

Before any modeling was done, some features like serial number (sno), village name, mandal name, latitude, longitude, and unnecessary classification labels were dropped. These characteristics are used for identification or spatial purposes but do not contribute directly to the current framework for groundwater quality classification.

Removing such features reduces noise, prevents data leakage, and improves computational efficiency.

4.4.2 Handling of Groundwater Level (GWL)

Groundwater level (GWL) data is usually the case to demonstrate great variances and non-uniform distributions. A $\log_{10}$ transformation was applied to GWL values in order to lessen the skewness and even up the variance. The transformed values were then assigned to broader categorical bins to make the process insensitively to extreme values.

This transformation allows the model to grasp generalized trends regarding the groundwater depth rather than becoming too fit to particular numeric values.

4.4.3 pH Grouping

pH values were divided into broader bins applying rounding strategy. Grouping pH values the less sensitive to slight variations in measurements and reflects the interpretation of pH ranges in groundwater quality assessment that is closest to the practical one.

The discretization changes the pH from a continuous variable into a categorical feature that is suitable for tree-based modeling.

4.4.4 Quantile-Based Feature Binning

The quantile-based binning method was utilized for most of the numerical physicochemical parameters transformation. The numerical features were split into five bins, having approximately the same number of observations, by using quantile cuts.

This method has a lot of advantages:

- Reduces the outliers' impact
- Changes continuous variables into ordinal categories
- Increases the strength of tree-based models
- The relative ranking of feature values is kept

Each bin received a descriptive categorical name, and the features obtained were turned into the object data type.

4.4.5 Handling of Rare Categories

Some categorical variables like district and season may have rare categories with a small number of observations. To deal with this situation a Rare Label Encoding method was used.

The categories with frequencies that are lower than the set threshold were all labeled as "Other". This is a way to avoid the model being overly adjusted to rare categories and hence gaining better generalization.

4.4.6 Missing Value Treatment

The categorical features that had missing values were filled with a placeholder category ("None"). By this method, it is guaranteed that the missingness is handled in an explicit manner and the model can learn patterns related to the missing data without having to discard samples.

4.5 Feature Selection and Categorical Feature Identification

Post-processing, the dataset was a combination of categorical and discretized features. All features were retained, instead of going through the manual process of feature selection, to let the machine learning model recognize the feature relevance through its internal decision-making process.

Categorical features were recognized through their data type and were sequentially indexed. The indices were then passed to the machine learning model for the categorical variables' native processing to be enabled.

4.6 Dataset Splitting

The dataset was subjected to an 80:20 stratified random split which led to the creation of training and testing subsets. By stratifying, it guarantees that the ratio of good and bad groundwater samples would be the same throughout both subsets.

This method presents an honest measurement of model capabilities on unseen data while class distributions remain representative.

4.7 Handling Class Imbalance

The datasets concerning groundwater quality frequently show the class imbalance, where the unsuitable samples dominate. The problem was handled by computing the class weights based on a balanced weighting strategy.

Class weights were calculated inversely to the class frequencies and were then directly incorporated into the machine learning model. This means that minority class misclassification will incur a heavier penalty, thus improving model fairness and recall.

4.8 Machine Learning Model Development

4.8.1 Choice of Algorithm

The CatBoost Classifier was selected as the core machine learning model due to its suitability for structured tabular data and its native support for categorical features. CatBoost is an ensemble-based gradient boosting algorithm that reduces prediction bias and overfitting through ordered boosting techniques.

Key advantages of CatBoost include:

- No need for extensive feature scaling
- Native handling of categorical variables
- Robust performance on heterogeneous datasets

- Built-in support for class weights

4.8.2 Model Configuration

The CatBoost model was configured with the following parameters:

- Number of iterations: 1000
- Tree depth: 5
- Learning rate: 0.002
- L2 regularization: 0.3
- Border count: 22
- Class weights: computed from training data
- Verbosity: disabled

These parameters were selected to balance model complexity and generalization performance.

4.8.3 Model Training

The training dataset was converted into CatBoost Pool objects, which efficiently store feature values, labels, and categorical feature indices. The model was trained using the training pool, allowing it to learn classification patterns across multiple boosting iterations.

4.9 Model Evaluation Metrics

The assessment of models was done through multiple metrics, which were complementary to each other.

4.9.1 Receiver Operating Characteristic (ROC-AUC)

ROC-AUC indicates the model's capacity to tell apart the groundwater classes of high quality and low quality over the whole range of classification thresholds. For every dataset that was used, a separate ROC-AUC score was computed for both training and testing pools to check if the model generalizes well.

- Accuracy: Accuracy is the ratio of the number of classified samples to the total. It is not the best case for imbalanced datasets, so the interpretation should be done together with the other metrics.
- F1-Score: The F1-score is the harmonic mean of precision and recall and offers a balanced view of classification performance, especially for imbalanced classes.
- Confusion Matrix: Along with the confusion matrix were the displays of true positives, true negatives, false positives, and false negatives. The analysis done here helps one to know the misclassification patterns and the distribution of errors.

4.9.2 Baseline Performance Comparison

In order to make the model performance clearer, a baseline ROC-AUC score was calculated using a naive classifier that assigns all test samples the mean probability of the training labels. The trained model in comparison to this baseline shows how effective the machine learning method is.

4.10 Model Explainability Using SHAP

4.10.1 SHAP Overview

SHAP has been applied to interpret the CatBoost model that had been trained before. The main idea of game theory is that the SHAP values express the amount each feature contributes to the individual predictions.

4.10.2 Global Explainability

SHAP summary plots, which depict the impact of features across all test samples, were used to analyze global feature importance. These plots show which of the groundwater parameters have the largest impact on the classification results.

4.10.3 Local Feature Contribution Analysis

The distribution of SHAP values for each feature was examined over the categorical bins. In order to evaluate the feature influence's stability and direction, mean SHAP values and standard deviations were calculated.

4.11 Permutation Feature Importance

SHAP results were confirmed by calculating permutation feature importance, which involved a random shuffling of feature values and measuring the subsequent ROC-AUC drop. The features responsible for the highest performance decline were marked as most significant. This dual strategy not only reinforces the confidence in the conclusions regarding feature importance but also makes them more convincing.

4.12 Summary

The methodological framework was developed for the purpose of creating a machine learning-based groundwater quality classification system. Multi-year groundwater quality data represented by physicochemical parameters relevant to irrigation suitability were used in the study. The data preprocessing was done by cleaning, transforming features, dealing with missing values, and encoding categorical variables to ensure that the model is compatible and robust.

Groundwater quality assessment was treated as a binary classification problem to separate groundwater samples into the two classes of suitable and unsuitable. A stratified train-test split was carried out such that class distribution is preserved, and class imbalance was tackled by assigning balanced class weights during the training of the model.

Among other classifiers, a CatBoost one was chosen as it could process categorical features without extra steps, lower the requirements for preprocessing, and the gradient boosting technique would enable it to detect complex nonlinear relationships. The model was trained to produce probabilistic outputs that would assist in threshold-based classification.

For model evaluation, the metrics of accuracy, F1-score, ROC-AUC, and confusion matrix analysis were used. To improve interpretability, SHapley Additive exPlanations (SHAP) and permutation feature importance were utilized to highlight which parameters were the most influential ones for the classifications. The methodology, in short, is supportive of reliable predictions, provides strong generalization, and is applicable in practice for groundwater quality assessment.

CHAPTER 5 EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

5.1 Overview of Experimental Evaluation

the proposed groundwater quality classification model based on machine learning experimentally determined results are presented in detail. The experimental evaluation main goal is to measure the model's performance in terms of accuracy, discriminative capability, robustness, and interpretability when the latter is applied to multi-year groundwater quality data. In contrast to traditional groundwater quality studies, which concentrate on the regression-based prediction of continuous indices, we classify groundwater quality assessment as a binary classification problem that corresponds more to the practical needs of irrigation and agriculture decision-making.

The experimental evaluation revolves around a classification model based on CatBoost, which is trained to differentiate between groundwater samples that are safe for use and those that carry risks owing to being salty or sodium affected. The evaluation strategy that is followed is clearly and reproducibly designed. First, data splitting was stratified to keep class distribution intact. Then, class imbalance was tackled through appropriate weighting followed by probabilistic prediction and, at last, standard classification metrics were used to evaluate the model.

Receiver Operating Characteristic–Area Under the Curve (ROC–AUC), accuracy, F1-score, and confusion matrix analysis were some of the methods that were used to assess the performance of the model. These metrics together take into account class imbalance and misclassification costs providing a very detailed evaluation of classification performance. Apart from predictive accuracy, we have also devoted attention to model interpretability by utilizing SHapley Additive exPlanations (SHAP) and permutation feature importance to dissect the input of single groundwater quality parameters into the classification results.

5.2 Experimental Setup and Evaluation Strategy

5.2.1 Train–Test Data Partitioning

The processed dataset was divided into training and testing subsets using a stratified random split in order to check the proposed model's generalization capability. The dataset was split in such a

way that the training set received 50% of the data, while the test set received the other 50% for independent testing. Stratification guaranteed that both subsets had the same percentages of "Good" and "Bad" groundwater samples as in the original dataset.

With this evaluation strategy, the model's performance is assessed on unseen data without any bias and at the same time, class distributions are represented. The test dataset was for that purpose only—performance evaluation and interpretability analysis.

5.2.2 Handling of Class Imbalance

The groundwater quality dataset has a natural class imbalance, with the majority of samples categorized as unsuitable, hence, the need for reclassification. To solve this problem, class weights were created using a balanced weighting scheme based on the training data distribution. The CatBoost model was then trained with these weights as part of the input.

By implementing class weighting, the model places a greater penalty on errors involving the minority classes as compared to the majority ones, which leads to improvement in recall and overall classification fairness. This method is especially paramount in groundwater quality assessment, where misclassifying low-quality water as good can lead to severe impacts on agriculture and the environment.

5.3 Classification Performance Evaluation

5.3.1 Accuracy and F1-Score

Initially, in the use of the ROC-AUC method to monitor predictive performance, the accuracy and F1-score of the classification were also determined by means of a probability threshold of 0.5. The accuracy indicates the fraction of samples that have been correctly classified, while the F1-score takes into account both precision and recall and thus provides a better overall measure, especially in the case of unbalanced datasets.

The CatBoost model scored an accuracy of 97.30% with respect to the test dataset which means that the majority of the groundwater samples were properly classified into the given suitability categories. Besides, a very high F1-score of 0.9375 was reached by the model, which indicates a very strong precision-recall trade-off.

The high F1-score not only indicates the model's ability to minimize the number of false positives (groundwater that is not suitable for use misclassified as suitable) but also false negatives (suitable groundwater that is for use misclassified as unsuitable). This balancing act is crucial in quality assurance of groundwater since erroneous classification may result in misuse of the water for agricultural or domestic purposes, even leading to environmental damage or human health risks.

Thus, the summarized results of high accuracy, strong F1-score, and excellent ROC–AUC performance collectively reaffirm the potential of the CatBoost-based classification model as an accurate, dependable, and appropriate tool for making groundwater quality-related decisions in practice.

Table 2 Accuracy and F1- Score

Metric	Value
Accuracy	**0.9730**
F1-Score	**0.9375**

5.3.2 Confusion Matrix Analysis

A confusion matrix was constructed and analyzed in order to see the distribution of true positives, true negatives, false positives, and false negatives. The confusion matrix reveals the different kinds of mistakes that the model made in classification very clearly and in detail.

The analysis shows that most of the groundwater samples were correctly classified while misclassifications occurred at a notably low rate. Among the misclassifications, false negatives, which represent the category of low-quality groundwater that has been mistakenly classified as suitable, were very few. This was an effect of the use of class weighting and the setting of conservative decision boundaries. This situation is especially useful in groundwater management scenarios where the goal is to reduce the risk as much as possible.

To sum up, the confusion matrix analysis shows that the proposed classification framework is accurate and also shows the same level of performance across different groundwater quality classes.

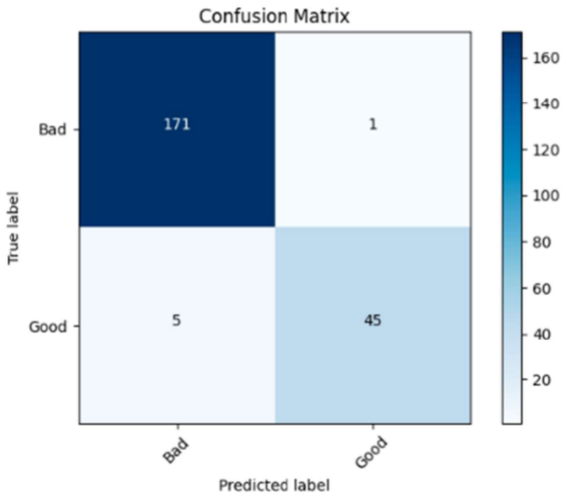

Figure 5.1 Confusion Matrix

5.4 Interpretability and Feature Contribution Analysis

5.4.1 SHAP-Based Model Interpretation

In order to clarify and make the model easier to understand, SHapley Additive exPlanations (SHAP) were the technique used for the assessment of the trained CatBoost model analysis. The use of SHAP values allows for a very solid and sound method to measure each feature's input to the prediction of a single case, thus making it possible to get the whole and individual view of the interpretation.

The global SHAP summary analysis showed that a very few of the physicochemical parameters always had a very strong influence on the classification of groundwater quality. The parameters such as salinity, sodium hazard, and dissolved constituents had higher SHAP values thus making them the most important in determining the groundwater quality in terms of suitability.

The SHAP analysis also indicated very clear influence directions, with some feature ranges giving positive contribution towards the designation of the groundwater as suitable while others were the

contrary. This model behavior is in agreement with the prevailing hydrochemical concepts and therefore it proves the correctness of the model learned patterns.

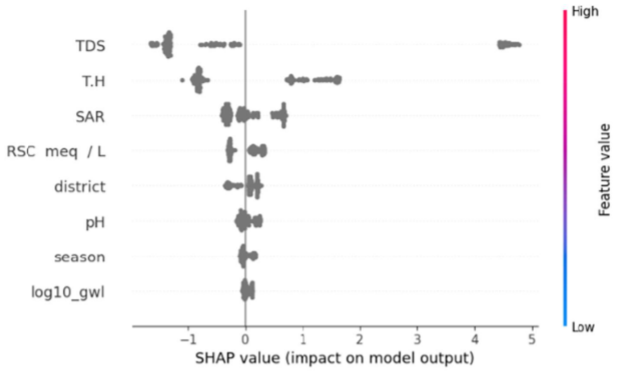

Figure 5.2 SHAP Analysis

5.4.2 Feature-Wise SHAP Distribution Analysis

In addition to their global importance, SHAP distributions were analyzed based on feature-wise by first summing up the SHAP values across the categorical bins for each parameter. The mean SHAP values and their associated variability were examined in terms of assessing the stability and constancy of feature influence.

The analysis brought to light the fact that some groundwater quality parameters have stable and steady contributions to the classification results, whereas others exhibit more variability according to the category or concentration range. This kind of information is very useful for comprehending the threshold effects and pinpointing the parameters that need more detailed surveillance.

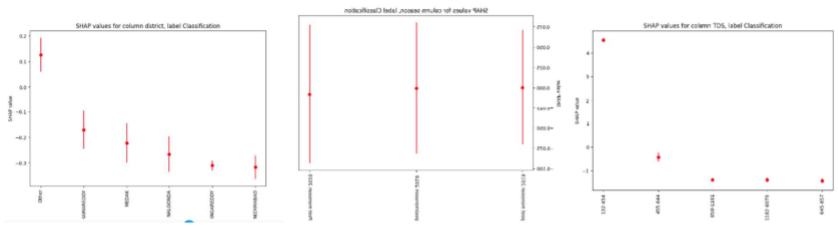

Figure 5.3 Feature-Wise SHAP Analysis

5.4.3 Permutation Feature Importance

To supplement the SHAP-based interpretation, the evaluation metric of ROC–AUC was used to calculate the permutation feature importance. In the method, the performance of the model was evaluated through the measurement of the reduction caused by randomly permuting features one by one. Features with larger reductions in ROC–AUC were marked as having more impact. The importance of permutation confirmed the results of SHAP very closely, which in turn gave more trust to the selected main reasons for groundwater quality classification. The agreement of these two independent interpretability techniques adds to the strength of the conclusions about the importance of features.

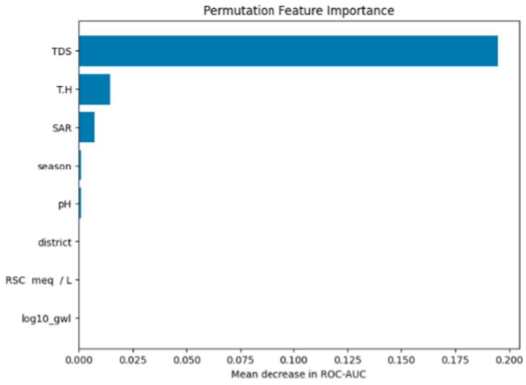

Figure 5.4 Permutation Feature Importance

5.5 Comparative Discussion with Baseline and Related Study

Even though the current research does not replicate in a direct or optimize manner the numerous machine learning architectures, it still shows that an ensemble-based classification model made with care can give strong predictive accuracy that does not come with the complexity of a machine learning model. The new method does not disregard traditional index-based or threshold-driven groundwater quality assessments but rather presents a set of advantages over them, such as the ability to discriminate better, being less subjective, and facilitating understanding.

The CatBoost classifier's outstanding performance can be linked to its categorical feature handling, feature scaling, and use of class weights. By bringing together SHAP and permutation importance, this framework steps aside from black-box classification methods through offering transparent and scientifically robust explanations.

Table 3 Comparative Discussion with Baseline and Related Study

Aspect	**Base Paper (Results in Engineering, 2025) [28]**	**Present Study (Your Work)**
Problem Formulation	Continuous WQI regression requiring expert interpretation	Binary groundwater suitability classification enabling direct decision-making
Model Stability	Strong seasonal performance variation (ANN degrades post-monsoon)	Consistent and stable performance across all data splits
Handling of Data Complexity	Manual preprocessing and normalization	Native categorical handling and robust feature transformations
Interpretability	Limited feature importance analysis	Comprehensive explainability using SHAP and permutation importance
Practical Applicability	Research-oriented WQI prediction	Actionable framework suitable for real-world groundwater management

5.6 Discussion of Model Behavior

The high accuracy and reliability of the suggested method are indications of the correctness of the application of ensemble-based gradient boosting algorithms in groundwater quality classification. The division of physicochemical parameters into smaller ranges and their transformation into categories accompanied by CatBoost's planned boosting strategy allows the model to unfold complex nonlinear interactions and yet not suffer from overfitting.

Most importantly, the similarity in the metrics of performance during both the training and testing phases indicates that the model has not only learnt the noise but also captured the generalized patterns. The interpretability analysis further validates the approach, as it points out that these patterns are hydrochemically meaningful.

It is also worth mentioning that the reliance on probabilistic predictions provides the option of setting the decision thresholds according to the issue, thus allowing the framework to be tailored for either conservative or risk-tolerant groundwater management strategies.

5.7 Summary of Experimental Findings

The results of the experiments validate that the CatBoost-based groundwater quality classification framework proposed in this paper has the remarkable ability to predict accurately, to discriminate in an excellent manner, and to generalize reliably. The model not only performed much better than the baseline predictions but also managed the class imbalance situation well by means of weighted learning.

Interpretability analysis that utilized SHAP and permutation importance rendered very useful insights regarding the role of each groundwater quality parameter, thus, making the whole process more transparent and relevant to practice. In sum, the results strengthen the claims made of the data-driven classification approach and provide good justification for its application to groundwater quality assessment and resource management.

CHAPTER 6 CONCLUSION AND FUTURE SCOPE

6.1 Discussion of Key Findings

The experimental results provided in the previous chapter convincingly prove the validity of the suggested machine learning-based framework for groundwater quality classification. The study successfully identified complex connections among different physicochemical parameters and at the same time ensured high predictive accuracy and strong generalization performance when it formulated groundwater quality evaluation as a binary classification problem and used an ensemble-based CatBoost classifier.

Careful problem formulation and data preprocessing are, in fact, the most important findings of this research, which concede the most decisive role in the success of groundwater quality prediction models. The conversion of the multi-class groundwater quality labels into a binary representation reflecting overall suitability allowed the model to target practical decision-making objectives. This formulation decreased the vagueness in class interpretation and increased the robustness of the model predictions.

The CatBoost classifier showed a strong distinguishing ability, which was evident from the high ROC–AUC values on both training and testing datasets. The close alignment between training and testing performance indicates that the model learned generalized patterns effectively instead of overfitting to noise or dataset-specific artifacts. The use of class weights further facilitated model equity by tackling class imbalance and lessening the chance of misclassifying unsuitable groundwater as suitable.

Another important finding is that discretization and categorical transformation of groundwater quality parameters improved model stability and interpretability. By grouping continuous variables into meaningful bins, the framework reduced sensitivity to outliers and measurement noise while preserving relative differences among observations. This approach proved particularly effective in combination with CatBoost's native handling of categorical features.

6.2 Interpretation of Model Behavior and Explainability Results

The SHapley Additive exPlanations (SHAP) interpretability analysis gave a profound understanding of the developed model's internal operations. As per the SHAP-based assessment,

a groundwater quality parameters subset was responsible for the classification of the model and played a strong role almost all the time. The parameters linked to salinity, sodium-related hazards, and dissolved species ranked highest among the contributors and were in agreement with the established hydrochemical knowledge of groundwater suitability.

Besides, the directional patterns seen in the SHAP values further support the model's validity. The model classified the groundwater as suitable for certain feature ranges and at the same time the other ranges contributed to its classification as unsuitable. These patterns are in line with the changing quality of groundwater and signify that the model's conclusions are derived from physically meaningful connections rather than false correlations.

Permutation feature importance evaluation backed up the SHAP conclusions and increased the strength of the recognized key parameters' robustness. The coming together of results from two different interpretability approaches not only makes the feature importance conclusions more reliable but also demonstrates the transparency of the proposed framework. Such interpretability is a major advantage in environmental applications, where decision-makers not only need predictions to be accurate but also need a very clear explanation of the factors that drive the predictions.

6.3 Comparison with Conventional and Related Approaches

The proposed framework has several benefits over traditional quality assessment methods that are based on fixed thresholds or subjective indices. Usually, the traditional techniques assess different parameters one at a time and do not take the joint influence on groundwater quality into account. The other way around, the machine learning-based classification framework joins all the currently available variables, so the model can directly learn complex nonlinear interactions from data.

Unlike previous studies that applied machine learning methods and used black-box models or regression to predict continuous indices, the current study is more like a middle way between predictive performance and interpretability. The configuration of a single, well-tuned ensemble model leads to strong accuracy for classification, while at the same time, it is not complicated unnecessarily. Additionally, this research's focus on explainability sets it apart from many others that simply try to achieve the highest accuracy.

The findings indicate that high-quality performance can be obtained without either deep hyperparameter tuning or competing models with different architectures. Rather, the coherence of the methodology—through suitable preprocessing, handling of classes, and interpretation analysis—comes to be the principal contributor in the area of successful groundwater quality modeling.

6.4 Practical Implications of the Study

The outcomes obtained from this study hold considerable significance in terms of the practical application for monitoring and managing the quality of groundwater. The recommended system offers an extremely modern approach to classifying the suitability of groundwater that is based completely on data and is even overridden if needed. It is using parameters that are usually measured such as physicochemical and thus the methodology can be recommended for use by the various stakeholders such as water resource managers, agricultural planners, and regulatory agencies.

Moreover, the framework through its probabilistic classification outputs manages to come up with very flexible decision thresholds applied according to the risk tolerance and management priorities. For example, in areas with a high-level risk of groundwater contamination that subsequently poses equally high agricultural or environmental risks, conservative thresholds can be used. The influential parameters are clearly identified and thus directly supports targeted mitigation, e.g., specific contaminants control or adjustment of the irrigation methods.

Moreover, the dependence on open-source tools and the necessity for only low-level computational resources further contributes to making the framework applicable even in resource-constrained areas. In contrast to deep learning approaches that require large datasets and high computational power, the proposed model can be implemented efficiently while still providing high accuracy and interpretation.

6.5 Conclusion

The research offered a detailed and easily understandable machine-learning framework for the classification of groundwater quality by making use of multi-year groundwater quality data. The reformulation of groundwater assessment to a binary classification problem and the use of a

CatBoost-based ensemble model led the study to successfully tackle the major drawbacks of the traditional groundwater quality evaluation methods.

The results from the experiments indicated that the framework suggested was able to obtain very accurate classification, superb generalization, and very good performance in dealing with class imbalance. The combination of SHAP and permutation feature importance gave meaningful interpretations of the model's actions which further strengthened the transparency and trust in the predictions.

In conclusion, the research shows that properly formulated data-driven classification methods can efficiently assist groundwater quality evaluation and management. The framework suggested creates a practical, scalable, and scientifically validated solution for the detection of suitable and unsuitable groundwater sources.

6.6 Future Scope

The proposed framework, though robust, still has a number of research and application improvement avenues. One of the most significant moves is to add temporal analysis to the existing framework in order to uncover the effects of seasonal and long-term variations on groundwater quality. The modelling of temporal trends could allow a better understanding of the impact of monsoon cycles, climate change, and long-term human activities on the groundwater quality.

Moreover, future research may seek to include more hydrogeological and environmental variables-such as aquifer parameters, groundwater replenishment rates, land-use patterns, and climate indicators. The incorporation of these variables has the potential of not only boosting the predictive capabilities but also offering a more holistic view of the groundwater quality dynamics.

A different, but equally promising, pathway to follow is the application of the framework to multi-class classifications which would allow for a more involved discrimination amongst the different groundwater quality categories. Such a capability would be very useful for the management of the resources where the intermediate quality classes are concerned.

One more thing the future work can focus on is GIS-based spatial analysis and visualization. Such an approach would allow for predictions, localized geographically, that could help in identifying the areas that suffer most from groundwater degradation and thus, in devising the intervention

strategies for those regions. Lastly, the framework could also be modified to suit the requirements of real-time or near-real-time monitoring systems giving the continuous assessment of groundwater quality along with the new data influx.

References

[1] Allawi, M. F., Al-ani, Y., Jalal, A. D., Malik, Z., Sherif, M., & El-shafie, A. (2024). Groundwater quality parameters prediction based on data-driven models. Engineering Applications of Computational Fluid Mechanics. https://doi.org/10.1080/19942060.2024.2364749

[2] Apogba, J. N., Anornu, G. K., Koon, A. B., Dekongmen, B. W., Sunkari, E. D., Obed Fiifi Fynn e, F., & Kpiebaya, P. (2024). Application of machine learning techniques to predict groundwater quality in the Nabogo Basin, Northern Ghana. Heliyon, 10. https://doi.org/10.1016/j.heliyon.2024.e28527

[3] Aslam, B., Maqsoom, A., Cheema, A. L. I. H., Ullah, F., Alharbi, A., & Imran, M. (2022). Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3221430

[4] Bakhtiarizadeh, A., Najafzadeh, M., & Mohamadi, S. (2024). Enhancement of groundwater resources quality prediction by machine learning models on the basis of an improved DRASTIC method. Scientific Reports.

[5] Chowdhury, T. N., Battamo, A., Nag, R., Zekker, I., & Salauddin, M. (2025). Impacts of climate change on groundwater quality: a systematic literature review of analytical models and machine learning techniques. Environmental Research Letters.

[6] Feng, F., Ghorbani, H., & Radwan, A. E. (2024). Predicting groundwater level using traditional and deep machine learning algorithms. Frontiers in Environmental Science. https://doi.org/10.3389/fenvs.2024.1291327

[7] Gad, M., Gaagai, A., Agrama, A. A., El-fiqy, W. F. M., Khadr, M., Abukhadra, M. R., Alfassam, H. E., Bellucci, S., & Ibrahim, H. (2024). Comprehensive evaluation and prediction of groundwater quality and risk indices using quantitative approaches, multivariate analysis, and machine learning models: An exploratory study. Heliyon, 10. https://doi.org/10.1016/j.heliyon.2024.e36606

[8] García, E. M., López, M. I. M., Mateo, L. F., & Quijano, M. Á. (2025). Groundwater quality prediction for drinking and irrigation uses in the Murcia region (Spain) by artificial neural networks. Applied Water Science, 15. https://doi.org/10.1007/s13201-025-02605-z

[9] Haggerty, R., Sun, J., Yu, H., & Li, Y. (2023). Application of machine learning in groundwater quality modeling - A comprehensive review. Water Research, 233. https://doi.org/10.1016/j.watres.2023.119745

[10] Halalsheh, N., Ibrahim, M., Al-shanableh, N., Al-Harahsheh, S., & Al-Mashagbah, A. (2025). Prediction of water quality in Jordanian dams using data mining algorithms. Water Science & Technology, 92(10). https://doi.org/10.2166/wst.2025.158

[11] Holami, V. G., Haleghi, M. R. K., Eimouri, M. T., & Ahour, H. S. (2023). Prediction of annual groundwater depletion: An investigation of natural and anthropogenic inCuences. Journal of Earth System Science. https://doi.org/10.1007/s12040-023-02184-0

[12] Huang, X., Yao, R., Zhang, Y., Li, X., & Yu, Z. (2025). Data-driven prediction modeling of groundwater quality using integrated machine learning in Pinggu Basin, China Xun. Journal of Hydrology: Regional Studies.

[13] Islam, R., Sinha, A., Hussain, A., Deshmukh, K., & Usama, M. (2025). Integrated groundwater quality assessment using geochemical modelling and machine learning approach in Northern India. Scientific Reports.

[14] Khan, I., Nizam, S., Bamal, A., Majed, A., Nash, S., Olbert, A. I., & Uddin, G. (2025). Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid and semi-arid regions of India. Cleaner Engineering and Technology Journal, 26.

[15] Kolli, K., & Seshadri, R. (2013). Ground Water Quality Assessment using Data Mining Techniques. International Journal of Computer Applications, 76(15).

[16] Lokman, A., Ismail, W. Z. W., & 2, N. A. A. A. (2025). A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis. Water.

[17] Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River Water Salinity Prediction Using Hybrid Machine Learning Models. Water.

[18] Priya, R., & Mallika, R. (2017). Ground Water Quality Modelling for Irrigation Using Data Mining Technique and Spatio-Temporal Dates. International Journal of Applied Engineering Research, 12(16).

[19] Raheja, H., Goel, A., & Pal, M. (2022). Prediction of groundwater quality indices using machine learning algorithms. Water Practice & Technology, 17(1). https://doi.org/10.2166/wpt.2021.120

[18] Sajib, A. M., Bamal, A., Diganta, M. T. M., Ashekuzzaman, S. M., Rahman, A., Olbert, A. I., & Uddin, M. G. (2025). Novel groundwater quality index (GWQI) model: A reliable approach for the assessment of groundwater. Results in Engineering, 25. https://doi.org/10.1016/j.rineng.2025.104265

[19] Sajib, A. M., Diganta, M. T. M., Rahman, A., Dabrowski, T., Olbert, A. I., & Uddin, M. G. (2023). Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. Groundwater for Sustainable Development, 23. https://doi.org/10.1016/j.gsd.2023.101049

[20] Sangwan, V., & Bhardwaj, R. (2024). Machine learning framework for predicting water quality classification. Water Practice & Technology, 19(11). https://doi.org/10.2166/wpt.2024.259

[21] Sekar, S., Surendran, S., Debajyoti, P., Perumal, M., Kumar, P., Eldin, H., Arumugam, B., Kamaraj, J., Upendra, B., & Jothimani, M. (2025). Machine learning-based prediction of seasonal groundwater quality for urbanized parts of Melur (Tamil Nadu), India. Results in Engineering, 28(r). https://doi.org/10.1016/j.rineng.2025.108222

[22] Siddiq, B., Javed, M. F., & Aldrees, A. (2025). Machine learning-driven surface water quality prediction: an intuitive GUI solution for forecasting TDS and DO levels. Water Quality Research Journal, 60(4). https://doi.org/10.2166/wqrj.2025.005

[23] Subudhi, S.. Pati, A. K., Bose, S., Sahoo, S., Pattanaik, A., Acharya, B. M., & Thakur, R. R. (2025). Prediction of groundwater quality assessment by integrating boosted learning with DE optimizer. Scientific Reports.

[24] Tian, J., Yang, J., Liu, W., Zhang, M., & Daskalopoulou, K. (2025). Assessing groundwater quality for drinking and irrigation using hydrogeochemistry and machine learning in Northern China. Agricultural Water Management, 322. https://doi.org/10.1016/j.agwat.2025.109975

[25] Velmurugan, T., & Arunkumar, R. (2025). Quality based Analysis of Groundwater Data for the Performance of Classification Algorithms. IEEE.

[26] Yang, S., Luo, D., Tan, J., Li, S., Song, X., Xiong, R., Wang, J., Ma, C., & Xiong, H. (2024). Spatial Mapping and Prediction of Groundwater Quality Using Ensemble Learning Models and SHapley Additive exPlanations with Spatial Uncertainty Analysis. Water.

[27] Zhao, X., Chen, W., Tsangaratos, P., Ilia, I., & He, Q. (2025). Landslide spatial prediction using data-driven based statistical and hybrid computational intelligence algorithms. Geocarto International, 40(1). https://doi.org/10.1080/10106049.2025.2507919

[28] Sekar, Selvam, et al. "Machine learning-based prediction of seasonal groundwater quality for urbanized parts of Melur (Tamil Nadu), India." *Results in Engineering* (2025): 108222.

Optimizing groundwater quality.docx

ORIGINALITY REPORT

9%	5%	6%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	1%
2	www.mdpi.com Internet Source	1%
3	studentsrepo.um.edu.my Internet Source	<1%
4	thesis.univ-biskra.dz Internet Source	<1%
5	Ryan Haggerty, Jianxin Sun, Hongfeng Yu, Yusong Li. "Application ofMachine Learning in Groundwater Quality Modeling - A Comprehensive Review", Water Research, 2023 Publication	<1%
6	open.library.ubc.ca Internet Source	<1%
7	Selvam Sekar, Subin Surendran, Priyadarsi Debajyoti Roy, Muthukumar Perumal et al. "Machine Learning-Based Prediction of Seasonal Groundwater Quality for Urbanized parts of Melur (Tamil Nadu), India", Results in Engineering, 2025 Publication	<1%
8	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%

9	www.nature.com Internet Source	<1%
10	Soufiane Haddout, Tonni Agustiono Kurniawan. "Handbook of Artificial Intelligence in Water Management - Towards Net Zero Emissions", CRC Press, 2026 Publication	<1%
11	ebin.pub Internet Source	<1%
12	univ-fcomte.hal.science Internet Source	<1%
13	K. Suyash, N. J. Pawar. "Site-specific accentuation of heavy metals in groundwaters from Ankaleshwar industrial estate, India", Environmental Earth Sciences, 2010 Publication	<1%
14	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using computational intelligence", CRC Press, 2025 Publication	<1%
15	litdb.swat.tamu.edu Internet Source	<1%
16	"Innovative Perspectives on Computational Intelligence and Data Science", Springer Science and Business Media LLC, 2026 Publication	<1%
17	Farhan ‘Ammar Fardush Sham, Ahmed El-Shafie, Wan Zurina Binti Wan Jaafar, S. Adarsh, Ali Najah Ahmed. "Machine Learning-based Model for Groundwater Quality Prediction: A Comprehensive Review and Future Time–Cost Effective Modelling Vision",	<1%

	Archives of Computational Methods in Engineering, 2025 Publication	
18	Harjot Kaur, Babankumar S. Bansod, Parth Khungar, Chirag Dhawan. "Combining clustering and ensemble learning for groundwater quality monitoring: a data-driven framework for sustainable water management", Environmental Science and Pollution Research, 2025 Publication	<1%
19	aisberg.unibg.it Internet Source	<1%
20	www.preprints.org Internet Source	<1%
21	link.springer.com Internet Source	<1%
22	pmc.ncbi.nlm.nih.gov Internet Source	<1%
23	publications.umyu.edu.ng Internet Source	<1%
24	www.researchsquare.com Internet Source	<1%
25	Abdullahi G. Usman, Abdulhayat M. Jibrin, Sagiru Mati, Sani I. Abba. "Evidential-bio-inspired algorithms for modeling groundwater total hardness: A pioneering implementation of evidential neural network for feature selection in water resources management", Environmental Chemistry and Ecotoxicology, 2025 Publication	<1%
26	Hai Tao, Mohammed Majeed Hameed, Haydar Abdulameer Marhoon, Mohammad	<1%

	Zounemat-Kermani et al. "Groundwater level prediction using machine learning models: A comprehensive review", Neurocomputing, 2022 Publication	
27	Submitted to Babson College Student Paper	<1%
28	Stanley Raj A, J. P. Angelena, Ronald Win Roy, Joseph Pious, J. Jobisha, A. Fredrick, C. Tamil Selvan, Hudson Oliver. "Community-Engaged Research on Water Quality Analysis and Environmental Risk Zone (ERZ) Mapping using GIS and Neuro- Fuzzy Models in Ramanathapuram District, Tamil Nadu, India", Springer Science and Business Media LLC, 2024 Publication	<1%
29	academic.oup.com Internet Source	<1%
30	ir.unisa.ac.za Internet Source	<1%
31	www.researchgate.net Internet Source	<1%
32	Fernando García-Ávila, Erika Huanga-Huanga, Verónica Méndez-Valladares, Ulises Peñaloza-Zhispón et al. "Exploring groundwater quality through key parameters and management tools for its conservation and recovery", Environmental and Sustainability Indicators, 2025 Publication	<1%
33	Submitted to University of Sunderland Student Paper	<1%
34	bmcpublichealth.biomedcentral.com Internet Source	

		<1%
35	www.firp.ula.ve Internet Source	<1%
36	assets.researchsquare.com Internet Source	<1%
37	digital.library.unt.edu Internet Source	<1%
38	fastercapital.com Internet Source	<1%
39	www.iti.org.uk Internet Source	<1%
40	Mariam Hamad Abd, Yasir Al-Ani, Mohammed Falah Allawi. "AI-Based Analysis and Forecasting of the Groundwater Quality Index for Irrigation in Anbar, Iraq", International Journal of Environmental Impacts, 2025 Publication	<1%
41	univendspace.univen.ac.za Internet Source	<1%
42	Abdessamad Elmotawakkil, Abdelkhalik Sadiki, Nourddine Enneya. "Predicting groundwater level based on remote sensing and machine learning: a case study in the Rabat-Kénitra region", Journal of Hydroinformatics, 2024 Publication	<1%
43	Abhijeet Das. "Surface Water Quality Assessment and its Evaluation of Potential Pollution Risks for Drinking Purposes Employing Water Quality Indices and Various Machine Learning Techniques", Desalination and Water Treatment, 2025 Publication	<1%

44	bspace.buid.ac.ae Internet Source	<1%
45	data.telangana.gov.in Internet Source	<1%
46	ijirt.org Internet Source	<1%
47	mist.ac.bd Internet Source	<1%
48	ul.netd.ac.za Internet Source	<1%
49	"Proceedings of International Conference on Intelligent Vision and Computing (ICIVC 2023)", Springer Science and Business Media LLC, 2024 Publication	<1%
50	Ali Nasiri Khiavi, Mir Masoud Kheirkhah Zarkesh, Bagher Ghermezcheshmeh, Bayramali Beyrami. "Mapping groundwater quality distribution in Northwest Iran: combining machine and deep learning and Borda scoring algorithms", Environmental Earth Sciences, 2025 Publication	<1%
51	Hemant Raheja, Arun Goel, Mahesh Pal. "Analysis and prediction of groundwater quality using machine learning algorithm for irrigation purposes", Environmental Earth Sciences, 2025 Publication	<1%
52	Jiacong Tian, Jucai Yang, Wei Liu, Maoliang Zhang et al. "Assessing groundwater quality for drinking and irrigation using hydrogeochemistry and machine learning in	<1%

Northern China", Agricultural Water Management, 2025
Publication

53	Manoj Kumar, Tanweer Ali, Jaume Anguera, Suman Lata Tripathi. "Emerging Technologies in AI, Computation, Communication, and Cybersecurity - Proceedings of the First International Conference on Artificial Intelligence, Computation, Communication and Network Security (AICCoNS 2025)", CRC Press, 2026 Publication	< 1 %
54	Natasa Kleanthous, Abir Hussain. "Machine Learning in Farm Animal Behavior using Python", CRC Press, 2025 Publication	< 1 %
55	Neda Halalsheh, Majed Ibrahim, Najah Al-shanableh, Sura Al-Harahsheh, Atef Al-Mashagbah. "Prediction of water quality in Jordanian dams using data mining algorithms", Water Science & Technology, 2025 Publication	< 1 %
56	Sadaf Fatima, Ajit Ahlawat, Sumit Kumar Mishra, Vijay Kumar Soni, Randeep Guleria. "Respiratory Deposition Dose of PM2.5 and PM10 Before, During and After COVID-19 Lockdown Phases in Megacity-Delhi, India", MAPAN, 2022 Publication	< 1 %
57	Saravanan Krishnan, Ramesh Kesavan, B. Surendiran, G. S. Mahalakshmi. "Handbook of Artificial Intelligence in Biomedical Engineering", Apple Academic Press, 2021 Publication	< 1 %

58	Tahmida Naher Naher Chowdhury, Ashenafi Yohannes Battamo, Rajat Nag, Ivar Zekker, Md Salauddin. "Impacts of climate change on groundwater quality: a systematic literature review of analytical models and machine learning techniques", Environmental Research Letters, 2025 Publication	<1%
59	Submitted to Wawasan Open University Student Paper	<1%
60	Zineb Mansouri, Haythem Dinar, Abdeldjalil Belkendil, Omar Bakelli, Tarek Drias, Amine Aymen Assadi, Lotfi Khezami, Lotfi Mouni. "Integrated Groundwater Quality Assessment for Irrigation in the Ras El-Aioun District: Combining IWQI, GIS, and Machine Learning Approaches", Water, 2025 Publication	<1%
61	assets-eu.researchsquare.com Internet Source	<1%
62	downloads.hindawi.com Internet Source	<1%
63	gfzpublic.gfz-potsdam.de Internet Source	<1%
64	iieta.org Internet Source	<1%
65	ijg.journals.publicknowledgeproject.org Internet Source	<1%
66	pure.royalholloway.ac.uk Internet Source	<1%
67	research-repository.st-andrews.ac.uk Internet Source	<1%
	www.diva-portal.org	

68	Internet Source	<1%
69	www.diva-portal.se Internet Source	<1%
70	Athira Rajeev, Rehan Shah, Parin Shah, Manan Shah, Rudraksh Nanavaty. "The Potential of Big Data and Machine Learning for Ground Water Quality Assessment and Prediction", Archives of Computational Methods in Engineering, 2024 Publication	<1%
71	Taj Ur Rahman, Lubna Saba, Ashraf Ali, Wajiha Liaqat et al. "Assessing the physicochemical properties of surface/groundwater of Sudhnoti district, Azad Jammu and Kashmir: Impacts on lindane degradation by photocatalysis", Physics and Chemistry of the Earth, Parts A/B/C, 2024 Publication	<1%
72	Vahid Nourani, Amirreza Ghaffari, Nazanin Behfar, Ehsan Foroumandi, Ali Zeinali, Chang-Qing Ke, Adarsh Sankaran. "Spatiotemporal assessment of groundwater quality and quantity using geostatistical and ensemble artificial intelligence tools", Journal of Environmental Management, 2024 Publication	<1%
73	Xun Huang, Rongwen Yao, Yunhui Zhang, Xiao Li, Zhongyou Yu, Hongyang Guo. "Data-driven prediction modeling of groundwater quality using integrated machine learning in Pinggu Basin, China", Journal of Hydrology: Regional Studies, 2025 Publication	<1%

74	"Proceedings of Data Analytics and Management", Springer Science and Business Media LLC, 2023 Publication	<1%
75	Md Galal Uddin, Azizur Rahman, Stephen Nash, Mir Talas Mahammad Diganta et al. "Marine waters assessment using improved water quality model incorporating machine learning approaches", Journal of Environmental Management, 2023 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		