Using Auto Immune LightGBM (AI-LGBM) for prediction of Ground Water Quality in Indian Regions

Omar Michael[a]

[a] *Your Institution, District and Country*

Abstract

Groundwater quality has become a pressing global concern, as the availability of drinkable groundwater continues to decline, leading to significant challenges and consequences. In this study, we propose a novel methodology that harnesses the power of machine learning techniques to predict groundwater quality. Our model, AI-LGBM, combines the strengths of Mutual Information based Feature Selection (MIFS), Auto Immune Optimization (AIO), and LightGBM to successfully predict the quality of groundwater. Through comprehensive experiments and comparative analyses, we demonstrate that our methodology outperforms existing traditional and state-of-the-art techniques in terms of prediction accuracy. The implementation of AI-LGBM holds significant promise for addressing the challenges associated with groundwater quality assessment and management. This research contributes to the advancement of predictive modeling techniques in the field of groundwater analysis and provides a valuable tool for decision-making processes and resource management. The successful application of our methodology paves the way for future advancements in groundwater quality prediction and monitoring, ultimately leading to improved water resource sustainability and a healthier environment.

Keywords: Ground Water Quality Prediction, Optimization, LightGBM

1. Introduction

Groundwater is a vital natural resource, and its quality is essential for the sustainability of ecosystems, agriculture, and human activities. The quality of groundwater can be affected by various factors such as land use, soil characteristics, precipitation patterns, and anthropogenic activities. The prediction of groundwater quality is a crucial task for decision-makers, policymakers, and stakeholders to ensure the proper management of this valuable resource.

Groundwater refers to the water present beneath the earth's surface in the pores and crevices of soil, sand, and rocks. It is a major source of freshwater, accounting for approximately 30% of the world's freshwater resources. Groundwater plays a crucial role in sustaining human populations and

Email address: `mail id` (Omar Michael)

ecosystems, serving as a primary source of drinking water, irrigation, and industrial use.

However, groundwater quality can be affected by various factors, such as human activities, industrialization, agricultural practices, and natural phenomena. Contamination of groundwater can lead to serious health hazards, environmental degradation, and economic losses.

Given the importance of groundwater and the potential hazards associated with its contamination, accurate prediction of groundwater quality is crucial for the effective management and conservation of this vital resource. In recent years, there has been an increasing interest in the use of machine learning and artificial intelligence techniques for groundwater quality prediction, as these approaches can provide accurate and timely predictions based on large and complex datasets.

Several methods have been used for the prediction of groundwater quality, including statistical and machine-learning techniques. However, the accuracy and reliability of these methods are limited by the quality and quantity of data available. In recent years, there has been an increasing interest in using feature selection methods to improve the accuracy and efficiency of groundwater quality prediction models.

In this research paper, we propose a novel approach for groundwater quality prediction that combines the mutual information-based feature selection (MIFS) method with the LightGBM algorithm and the Auto Immune Optimization (AIO) algorithm. The proposed approach, called AI-LGBM, aims to identify the most relevant features for groundwater quality prediction, optimize the hyperparameters of the LightGBM algorithm, and improve the performance of the model by using the AIO algorithm.

The AI-LGBM model is evaluated on a real-world groundwater quality dataset, and the results show that it outperforms other state-of-the-art methods in terms of prediction accuracy and efficiency. The proposed approach can provide valuable insights into the prediction of groundwater quality and can be used as a decision-making tool for the sustainable management of groundwater resources.

2. Related Works

In this section we have provided literature survey which inspired us to identify research problems and possible solutions for the problems. In section 2.1 we have provided factors affecting ground water quality. In section 2.1.1 Non environmental factors affecting ground water quality is discussed. In section 2.1.2 we have discussed environmental and climate conditions and how they affect ground water quality. Finally in section 2.2 we have shown some of the historical works of use of machine learning to predict ground water quality.

2.1. Factors affecting Ground Water Quality

Groundwater quality prediction is a critical aspect of the management and preservation of groundwater resources. The availability of reliable and accurate groundwater quality prediction models can assist decision-makers in assessing and addressing potential contamination risks, ensuring safe water supplies, and minimising the negative environmental impacts of anthropogenic activities. A substantial body of research has been devoted to developing various predictive models for groundwater quality, ranging from traditional statistical models to more advanced machine-learning techniques. In this context, this article aims to provide an overview of the related works in groundwater quality prediction and the latest developments in the field. The section will highlight the strengths and limitations of different approaches and discuss their applicability and effectiveness for various groundwater quality prediction scenarios.

2.1.1. Non-environmental factors of Ground Water Quality

Groundwater quality is a crucial factor for both human and ecosystem health. Non-environmental factors impacting groundwater quality include human activities, urbanisation, industrialisation, and agriculture. Human activities such as construction, mining, and drilling can release contaminants into groundwater. Urbanisation can also contribute to groundwater contamination through stormwater runoff, sewage systems, and landfills. Industrial activities can introduce pollutants such as heavy metals, chemicals, and pesticides into the groundwater. At the same time, agricultural practices such as using fertilisers and pesticides can also lead to groundwater contamination. In addition to these, population growth, changing demographics, and socioeconomic factors can also impact groundwater quality by increasing water demand, leading to over-extraction and leading to an increased risk of pollution from inadequate waste management practices. It is essential to consider these non-environmental factors to understand better and manage groundwater resources.

2.1.2. Environmental factors of Ground Water Quality

Groundwater quality is affected by a variety of factors, including physical, chemical, and biological factors. Spatial factors, such as land use and land cover, soil characteristics, and hydrogeological settings, also play an important role in determining groundwater quality. Land use and land cover changes, such as urbanization, deforestation, and agricultural practices, can lead to contamination of groundwater through the introduction of pollutants such as fertilizers, pesticides, and industrial waste. Soil characteristics, such as texture, permeability, and water-holding capacity, can influence the infiltration and movement of water and contaminants through the soil and into the groundwater. Hydrogeological settings, including the depth to water table, the geological structure of the aquifer,

and the hydraulic conductivity of the aquifer, can also affect the quality of groundwater by influencing the movement and mixing of groundwater with other sources of water, such as surface water and seawater. Understanding these spatial factors and their influence on groundwater quality is essential for effective management and protection of this valuable resource.

Climate factors play an important role in the quality of groundwater. Temperature, precipitation, and evapotranspiration are some of the main climate factors that affect groundwater quality. High temperatures can increase the rate of microbial activity in the soil, which can result in increased groundwater contamination. High levels of precipitation can increase the likelihood of surface runoff, which can carry pollutants into the groundwater. In contrast, low levels of precipitation can lead to lower water tables and increased concentrations of contaminants in groundwater. Evapotranspiration can also impact groundwater quality, as it affects the amount of water that enters and exits the soil. Additionally, extreme weather events such as floods or droughts can have significant impacts on groundwater quality, as they can alter the distribution of contaminants and impact the movement of water in the subsurface. Understanding the impacts of climate factors on groundwater quality is essential for effective management and protection of this important resource.

The concentration of various elements in groundwater is a key indicator of its quality. Some of the most commonly measured elements in groundwater analysis include major cations such as calcium, magnesium, sodium, and potassium, as well as anions such as chloride, sulfate, and bicarbonate. Trace elements such as arsenic, lead, and mercury are also important to consider, as they can pose significant health risks to humans and wildlife if present in high concentrations. Additionally, nutrients such as nitrogen and phosphorus can also be present in groundwater due to agricultural and industrial activities, leading to eutrophication of aquatic ecosystems. Thus, the analysis of element concentrations in groundwater plays a crucial role in ensuring its quality and protecting public health and the environment.

2.2. Models used in Ground Water Quality Analysis

Groundwater modeling is a critical component in managing and protecting groundwater resources. Models are used to simulate the behavior of groundwater systems, assess the impact of human activities, and evaluate the effectiveness of management strategies. Several types of models are available for groundwater analysis, including analytical, numerical, and hybrid models. Analytical models are simple and easy to use, but they have limited applications. Numerical models, on the other hand, are complex but offer a high degree of accuracy and flexibility. Hybrid models combine the advantages of both analytical and numerical models and have become increasingly popular in recent years. In this review, we will explore the different types of models used in groundwater analysis, their advantages

and limitations, and current trends and future directions in this field.

2.2.1. Statistical Models

Masoud Noshadi and Amir Ghafourian (2016) [1] studied groundwater quality in Fars province, Iran, using multivariate statistical techniques. The authors used R-type factor analysis to infer a relationship between variables, variable analysis, and clustering to identify the predominant type of water in the province. The study found that the predominant type of water in the province was Ca-HCO3, which was suitable for drinking. However, water that fell in the C4-S3 category had very high salinity problems and high sodium hazard. The study was limited by its inability to effectively analyze non-linear relationships and the lack of implication of causality between factors. The authors recommended the use of more advanced algorithms to improve accuracy in future studies.

Rishi Rana, Rajiv Ganguly, and Ashok Kumar Gupta (2018) [2] aimed to assess the pollution potential of leachate generated from three non-engineered landfill sites located in the Tricity region and its effect on groundwater quality. The authors used an indexing method, water quality index (WQI), and principal component analysis (PCA) to analyze the data. The study found that the leachate generated was toxic, with high values of landfill pollution index (LPI) obtained in all the landfill sites. The authors suggested proper treatment procedures to minimize the negative effect on groundwater quality. The study was limited by its inability to effectively analyze non-linear relationships, and its simplicity in terms of analysis. The authors recommended the use of more complex predictive models to improve accuracy in future studies.

Zhou and Wang (2023) [3] evaluated surface water quality and identified pollution sources in a major city in Southeast China. The authors used water quality indexing method (WQIr) and positive matrix factorization (PMF) models to analyze the data, and machine learning models to establish a model of the relationship between environmental variables and water pollutants. The study found that the main water quality parameters of the M River that exceeded the Class III standards were TN, F. coli, Fe, and Mn. The study was limited by its inability to effectively analyze non-linear relationships and its simplicity in terms of analysis. The authors recommended major future directions to establish a model of the relationship between environmental variables and water pollutants.

2.2.2. Feature Selection Models

In their work, AO et al. [4] proposed an ensemble model of artificial neural networks (ANNs), decision trees (DTs), and support vector machines (SVMs) for surface water quality prediction. The methodology used in this study involved the creation of a hybrid ensemble model of ANNs, DTs, and SVMs to effectively handle the non-linear and complex relationship between water quality parameters.

The study's key finding was that the ensemble model showed better accuracy than the individual models in predicting surface water quality. The study was limited by the possibility of overfitting and lack of interpretability of the results. The authors recommended the use of explainable techniques in future research.

J Charles et. al [5] identified the most important features related to water quality using feature selection techniques. The authors used principal component analysis (PCA) based feature selection and SHAP (SHapley Additive exPlanations) based sensitivity analysis to reduce the impact of irrelevant or noisy features. The key finding of the study was that the combination of feature selection and weighted extreme learning machine (WELM) algorithms improved the accuracy of water quality prediction and classification models. The study's limitations included the lack of data and the interpretability of the results. The authors recommended using more data and domain transfer to address these limitations in future research.

Rodriguez-Galiano et. al [6] in their work evaluated different feature selection approaches for predicting nitrate pollution in groundwater. The authors used correlation-based feature selection (CFS) and ReliefF feature selection methods to identify the most impactful features related to nitrate pollution. The key finding of the study was that land use, soil type, and climatic factors were the most important features in predicting nitrate pollution. The study's limitations included the generalizability of the results and the reproduction of the results. The authors recommended extending the research to more geographic regions and investigating the transferability of the results.

2.2.3. Machine Learning Models

The study by E. Dritsas et. al [7] aims to identify water suitability for consumption using predictive models. The study uses tree-based classifiers and SMOTE for class imbalance handling. The researchers found that the stacking classification model after SMOTE with cross-validation outperforms traditional models in terms of accuracy. However, the study has limitations in terms of interpretability, overgeneralization, and noise creation. The researchers suggest using advanced preprocessing and feature engineering to overcome these limitations and using interpretable models to explain the model's predictions.

In a study by H. Ibrahim et. al [8], the researchers analyzed the groundwater quality parameters and their suitability for irrigation using ANFIS and SVM models. The researchers achieved a high level of accuracy with determination coefficients (R2) of 0.99 and 0.97 for training and 0.97 and 0.76 for testing. However, the study has limitations in terms of generalization problems and not being applicable to large datasets. The researchers recommend using novel prediction techniques to improve generalization.

Y. Zhou et al. [3] evaluated the surface water quality and identifies pollution sources in a major city in Southeast China using machine learning models. The researchers used a random forest-based WQI model for component analysis of polluted water. The study identified Mn, Fe, faecal coliform, dissolved oxygen, and total nitrogen as the top five important water quality parameters. The study has limitations in terms of scalability problems and lack of distribution-based explanation. The researchers suggest using a lightweight model and incorporating explainable models to improve scalability and explainability.

2.2.4. Deep Learning Models

Zheng (2023) [9] developed an interpretable deep learning model for predicting the quality of stream water on a large scale. The study used a feed-forward neural network with the Scaled Exponential Linear Unit (SELU) activation function to predict water quality. The study found that air temperature and the proportion of forest area were the most impactful attributes for water quality prediction. However, the study was limited by the vanishing gradient and overfitting problems of neural networks. In future studies, the author suggests the use of more advanced architectures for better performance.

Li (2020) [10] proposed a hybrid model for predicting the level of ammonia nitrogen (NH3–N) in surface water. The study used the boundary-corrected maximal overlap discrete wavelet transform (BC-MODWT) and dual-stage attention-based long short-term memory (DA-LSTM) deep learning methods to predict NH3–N levels. The proposed hybrid model was able to provide early warning when sudden high NH3–N pollution occurred. However, the study was limited by computational complexity and interpretability issues of the model. In future studies, the author suggests the use of lightweight and interpretable models.

Gorgij et. al (2020) [11] developed a deep learning model for spatiotemporal forecasting of groundwater quality for irrigation purposes. The study used the Long Short-Term Memory (LSTM) algorithm to predict the Sodium Absorption Ratio (SAR) in the Urmia aquifer and assess the saltwater intrusion possibility in hydro-chemical changes. The study found that the LSTM algorithm was able to assess the saltwater intrusion possibility in hydro-chemical changes in the Urmia aquifer. However, the study was limited by computational complexity and interpretability issues of the model. In future studies, the author suggests the use of lightweight and interpretable models.

2.3. Research issues

After examining the literature on ground water quality prediction procedures, the subsequent difficulties in conducting research on this topic have been identified:

1. Throughout the literature reviewed, there is lack of non parametric feature selection techniques

Table 1: Factors considered by few of the recent studies on crash severity analysis

Name and Year	Objective	Factors	Methodology	Key Findings	Limitations	Future Scope
Noshadi et. al [1]	Groundwater quality analysis using multivariate statistical techniques (location Iran)	Calcium Bicarbonate, Sodium	R-type factor analysis to inference a relation between variables, Variable Analysis, Clustering	predominant type of water was Ca-HCO3 in the province that is suitable for drinking. Water which lie in C4-S3 category, have very high salinity problems and high sodium hazard	Not effective for non linear relationships, lack of implications of causality between factors	More advanced algorithm use for more accuracy,
Brazdil(2022) [12]	Weather and traffic accidents in the Czech Republic	Rain, Light, Snowfall, glaze ice and rime	One-tailed t-test, scatter plots with regression lines for determination of the correlation coefficient	Rain, snowfall, glaze, ice and rime are the most important weather factors contributing to the accidents.	Data uncertainties has not been taken into account	Significant trends in selected weather categories
Rana et. al [2]	pollution potential of leachate generated from three non-engineered landfill sites located in the Tricity region	LPI, leached generate	Indexing method, WQI, PCA analysis	High values of LPI obtained in all the landfill sites indicated that the leachate generated is toxic and proper treatment procedures must be ensured	Non linear analysis is not effective, too simple	More complex predictive model use
Zhou et. al [3]	improve our understanding of surface water quality, thus providing support for the formulation of water quality management strategies	Manganese, F. Coli, Manganese	Use of WQIr and PMF models	main water quality parameters of the M River that exceeded the Class III standards were TN, F. coli, Fe, and Mn.	establish a model of the relationship between environmental variables and water pollutants	Major future direction
Al-Mukhtar et. al [13]	ensemble data-intelligence models for surface water quality prediction.	Iron, Location	Ensemble model of ANNs, DTs and SVMs	model is effective in handling the non-linear and complex relationship between water quality parameters	Possible Overfitting Lack of interpretability	Use of explainable techniques
Charles et. al [5]	to identify the most important features related to water quality	Water compositional factors Temporal factors	PCA based feature selections, SHAP based sensitivity analysis	reduction of the impact of irrelevant or noisy features, combining feature selection and WELM algorithms improves the accuracy	lack of data, lack of interpretability	use of more data domain transfer
Rodriguez-Galiano et. al [6]	evaluate different feature selection approaches for predicting nitrate pollution in groundwater.	Soil Type climate	correlation-based feature selection (CFS) and ReliefF	land use, soil type, and climatic factors as most impactful features	Generalizability, Reproduction of results	extending to more geographic regions, Investigating the transferability
Dritsas et. al [7]	identification of water suitability using predictive models	Aluminium, Arsenic, Barium	Tree based classifiers, SMOTE for class imbalance handling	Stacking classification model after SMOTE with cross validation outperforms traditional models	Interpretability, Overgeneralization, Noise creation	Using advanced preprocessing and feature engineering, use of interpretable models
Ibrahim et. al [8]	a comprehensive analysis of the groundwater quality parameters, their inter-relationship , and their suitability for irrigation	hydrochemical facies, pH, TDS, soil elements	ANFIS and SVM model applied for analysis	determination coefficient (R2)" (R2 = 0.99 and 0.97) and testing (R2 = 0.97 and 0.76).	Generalization problem, not applicable to large dataset	Improved generalization using novel prediction technique
Zhou et. al [3]	Component analysis of polution water using machine learning models	Manganese, Feacal Coliforms, Dissolved Oxygen	random forest-based WQI model	Mn, Fe, faecal coliform, dissolved oxygen, and total nitrogen were selected as the top five important water quality parameters	Scalability Problem, Lack of distribution based explanation	Use of lightweight model, incorporation of explainable models
Zheng et. al [9]	An innovative interpretable deep learning method on water quality predictions	Air Temperature, Forest Area	Multiple feed forward neural network with SELU activation	Air temperature and proportion of forest area are most impactful attributes	Vanishing gradient and Overfitting problems of neural networks	Use of more advanced architecture
Li et. al [14]	hybrid model is proposed to predict NH3–N level in surface water	Ammonia, Ground Nitrogen	Combination of Decomposition methodology of boundary corrected MODWT (BC-MODWT) and deep learning method of dual-stage attention-based LSTM (DA-LSTM)	early warning when sudden high NH3–N pollution occurred	Computational Complexity and Interpretability	Use of light weight model and interpretable model
Gorgij et. al [11]	Assessment of Saltwater Intrusion Possibility in Hydro-Chemical changes in Urmia aquifer.	Sodium, Calcuim, Potassium	LSTM used on SAR(Sodium Absorption Ratio)	Assessment of Saltwater Intrusion Possibility in Hydro-Chemical changes in Urmia aquifer.	Computational Complexity and Interpretability	Use of light weight model and interpretable model

which are efficient enough to capture the non linear relationships between features instead of just correlation based selection.

2. There is a deficiency of research works which use non-traditional optimization techniques which are meta-heuristic in nature. Meta Heuristic techniques are robust and faster in nature than both heuristic models and classical models.
3. Lack of use of interpretable models to explain the prediction of the models. Since deep learning model working procedure is difficult to comprehend, it is advisable to use explainable techniques which can increase the interpretability of the models.
4. Very few research works have proposed a standard methodology which combines both feature selection models to efficient machine learning prediction models optimised by a customised optimization technique

2.4. Research contributions

To address the issues mentioned above, our study endeavours to contribute as follows:

1. We proposed the use of Mutual Information based Feature Selection which is non-parametric in nature. Since it is a probabilistic feature selection model it captures the non-linear correlation between features and updates the importance of the features iteratively.

2. In our research we used Auto Immune optimization techniques which is metaheuristic in nature. This model is more robust and can be applied to parallelism models. Since it doesn't require gradient information for running the model it is computationally efficient.
3. We used explainable techniques like LIME and SHAP for explaining the results obtained from the predictive models. These models are model agnostic and analyse the local and global structure of the models respectively. Through the use of LIME and SHAP we can increase the interpretability of the models.
4. Our methodology combines all the techniques such as Feature selection, efficient predictive models and explainable models to create a organised analytical procedure.

3. Methodology

This section outlines the proposed approach in detail, which is represented in the flowchart in Figure 1. The methodology is divided into four primary stages:

It consists of four important phases, which are as follows:

(i)*Data collection and input*

(ii) *Data pre-processing and merging*

(iii) *Prediction and analysis*

(iv) *Explanation and Interpretation*

3.1. Preliminaries

This section aims to provide the necessary background knowledge for comprehending our proposed model. Three essential concepts have been introduced and explained, which are Schrodinger Eigenmaps, BH t-SNE, and TRSA. The subsequent subsections provide an in-depth discussion of each concept.

3.1.1. Mutual Information based Feature Selection

MIFS (Mutual Information Feature Selection) is a feature selection algorithm that uses mutual information as a measure of relevance between a target variable and each feature. Mutual information is a statistical measure that indicates the degree of dependency between two variables. In the context of feature selection, it measures the degree to which a feature and the target variable are dependent on each other. MIFS selects the k most relevant features by ranking them based on their mutual information with the target variable and also considers the redundancy among the selected features. It eliminates the redundant features iteratively until the desired number of features is selected.

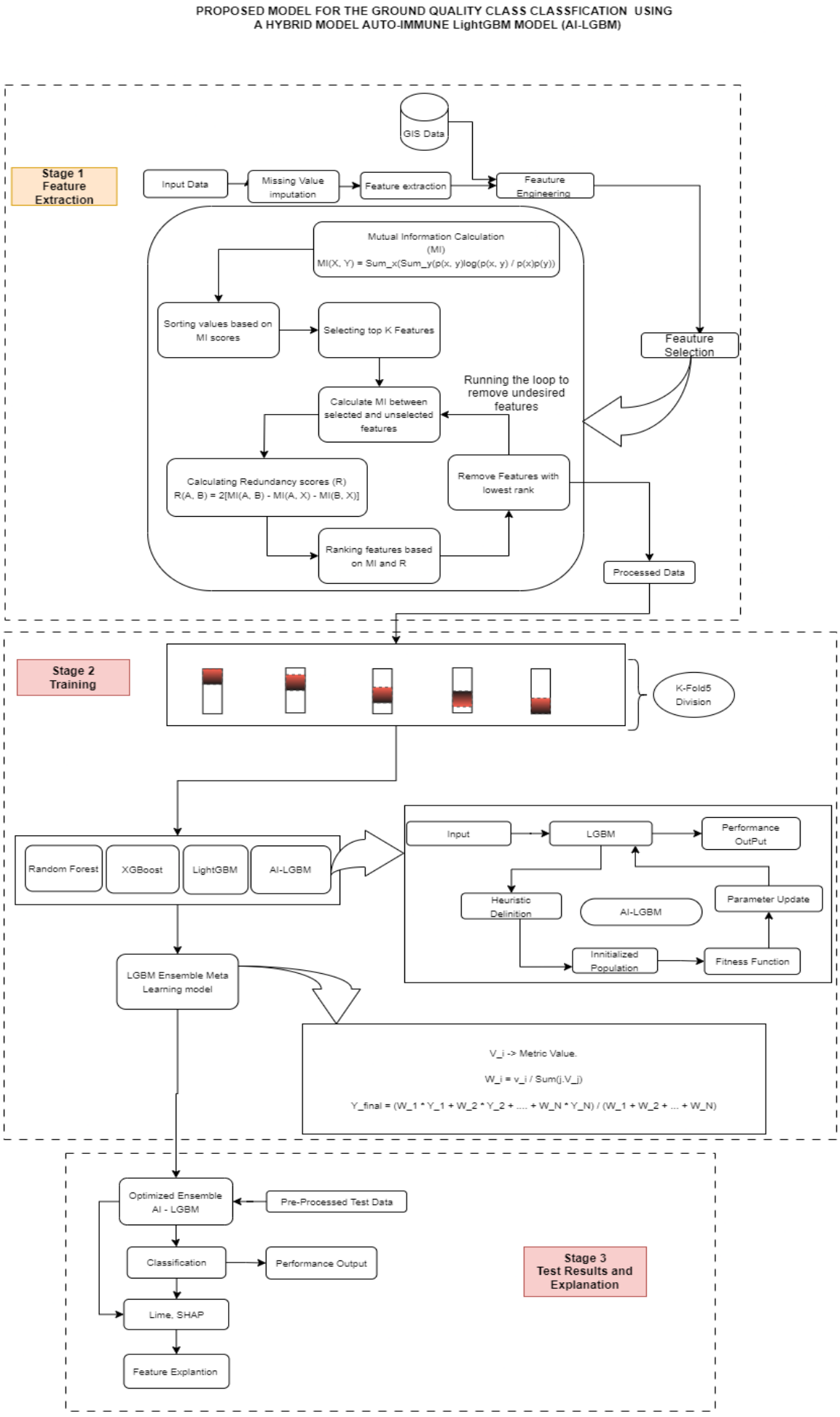


Fig. 1. Proposed methodological flowchart.

MIFS involves two main steps: relevance calculation and redundancy elimination. In the relevance calculation step, mutual information between each feature and the target variable is calculated. In the redundancy elimination step, the features are ranked based on their relevance and the most relevant feature is selected. Then, for each subsequent feature, the mutual information between the feature and all previously selected features is calculated. The feature with the highest mutual information and lowest redundancy with the previously selected features is added to the set of selected features. This process continues until k features are selected.

MIFS is a widely used feature selection algorithm and has been shown to perform well in various applications. It can be used with different machine learning models and can help to improve the performance and interpretability of the models by reducing the dimensionality of the feature space.

Sure, here are the equations for the MIFS algorithm:

Given a set of features X = X1, X2, ..., Xn and a target variable y, the goal of MIFS is to select a subset of k features F1, F2, ..., Fk that have the highest mutual information with y and the lowest redundancy with each other. The redundancy score of feature Fi with respect to feature Fj is denoted as R(Fi, Fj).

The mutual information between two features Xi and Xj is given by:

$$I(X_i; X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{1}$$

where $p(x_i, x_j)$ is the joint probability distribution of Xi and Xj, and $p(x_i)$ and $p(x_j)$ are the marginal probabilities.

The redundancy score of feature Fi with respect to feature Fj is calculated as:

$$R(F_i, F_j) = \frac{1}{2}[I(F_i; F_j) + I(F_j; F_i)] \tag{2}$$

The relevance score of feature Fi with respect to the target variable y is given by:

$$S(F_i, y) = I(F_i; y) \tag{3}$$

The final score of each feature is the sum of its relevance score with respect to y and the negative sum of its redundancy scores with respect to all other selected features:

$$Score(F_i) = S(F_i, y) - \sum_{j=1, j \neq i}^{k} R(F_i, F_j) \tag{4}$$

The algorithm selects the k features with the highest score to be included in the subset.

3.1.2. LightGBM

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is developed by Microsoft and is one of the fastest and most efficient implementations of gradient boosting available.

The key features of LightGBM are its speed, scalability, and accuracy. It is designed to handle large-scale data and can train models on datasets with billions of rows and millions of features. It also supports parallel and distributed training, which enables it to scale to clusters of machines.

LightGBM uses a technique called gradient-based one-side sampling (GOSS) to select only the informative data points for computing gradients, which significantly reduces the training time. It also uses a histogram-based approach to bin continuous features, which further speeds up the training process.

LightGBM uses a leaf-wise approach to growing trees instead of the traditional level-wise approach. This means that it grows the tree by splitting the leaf with the maximum loss reduction, which can lead to a better model accuracy. It also supports various loss functions and evaluation metrics, including binary cross-entropy, multi-class cross-entropy, and mean squared error.

In addition, LightGBM supports several advanced features such as early stopping, feature importance calculation, and GPU acceleration. It also has bindings for several programming languages, including Python, R, and Java.

Overall, LightGBM is a powerful and flexible gradient boosting framework that can be used for a wide range of machine learning tasks. Its speed and accuracy make it an attractive option for training models on large datasets, and its scalability and advanced features make it a popular choice for both research and production applications.

The objective function for training a LightGBM model is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}i) + \sum j = 1^{T} \Omega(f_j) \tag{5}$$

where θ represents the model parameters, n is the number of training samples, y_i is the true label of the ith sample, $\hat{y}_i$ is the predicted label, l is the loss function, T is the number of trees in the model, f_j is the jth tree, and Ω is the regularization term.

The gradient of the objective function with respect to the predicted values $\hat{y}$ is:

$$\frac{\partial \mathcal{L}}{\partial \hat{y}_i} = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}i} + \sum j = 1^{T} \frac{\partial \Omega(f_j)}{\partial \hat{y}_i} \tag{6}$$

The Hessian matrix of the objective function with respect to the predicted values is:

$$H = \frac{\partial^2 \mathcal{L}}{\partial \hat{y}_i^2} = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y} i^2} + \sum j = 1^T \frac{\partial^2 \Omega(f_j)}{\partial \hat{y}_i^2} \tag{7}$$

LightGBM uses a gradient-based boosting method to train the trees, and the gradients and Hessians are used to find the optimal split points for each node in the trees. The optimal split points are found by solving the following optimization problem:

$$\arg\min_{\theta} \left[\frac{1}{2} \sum_{i \in I} \frac{(\partial \mathcal{L} / \partial \hat{y} i - \theta)^2}{\sum i \in I H_{ii}} + \lambda |\theta| \right] \tag{8}$$

where I is the set of indices of the samples in the current node, λ is the regularization parameter, and θ is the optimal split point.

LightGBM also uses a histogram-based algorithm to speed up the computation of the gradients and Hessians, as well as the optimal split points. The histogram-based algorithm discretizes the feature values into bins, and computes the gradients and Hessians for each bin. The optimal split points are then found by selecting the best combination of bins.

3.1.3. Auto-Immune Optimization Algorithm

Autoimmune optimization algorithm (AIO) is a population-based optimization technique that is inspired by the biological process of the human immune system. The algorithm simulates the way the immune system identifies and eliminates foreign agents, such as viruses and bacteria, to develop a set of solutions that best fit the problem at hand.

The AIO algorithm consists of three stages:

Immune network initialization: In this stage, a set of random solutions is generated and initialized in the immune network.

Antibody selection: In this stage, the antibodies that show the highest affinity to the target antigen are selected. This is done by calculating the fitness function for each antibody, which is based on the evaluation of the solution in the problem space.

Hypermutation: In this stage, the selected antibodies are subjected to hypermutation, which simulates the process of introducing random mutations in the genetic material of the antibodies. This introduces diversity in the population and allows the algorithm to explore the solution space more effectively.

The AIO algorithm is known for its ability to handle complex and nonlinear optimization problems. It has been applied in various fields, including engineering, medicine, and finance, and has shown promising results in terms of accuracy and speed.

Here is the equations of Auto-Immune Optimization algorithm:

Objective function:

$$f(x) = \sum_{i=1}^{N} w_i g_i(x) \tag{9}$$

Antibody affinity:

$$a_i(x) = \exp\left(-\frac{\|x - x_i\|}{\beta}\right) \tag{10}$$

Antibody concentration:

$$c_i = \frac{a_i(x)}{\sum_{j=1}^{N} a_j(x)} \tag{11}$$

Antibody selection probability:

$$p_i = \frac{f(x)c_i}{\sum_{j=1}^{N} f(x)c_j} \tag{12}$$

Antibody mutation:

$$x_{new} = x_{old} + \theta(x_{best} - x_{old}) + \eta \tag{13}$$

Antibody clone:

$$x_{clone} = x + \gamma(x - x_{avg}) + \eta \tag{14}$$

Antibody mutation rate:

$$\theta = 1 - \frac{1}{1 + \exp(-\frac{a_{best}}{K})} \tag{15}$$

Antibody hypermutation rate:

$$\gamma = \frac{\gamma_{max}}{1 + \alpha c_{best}} \tag{16}$$

Antibody mutation step size:

$$\eta \sim \mathcal{N}(0, \sigma^2) \tag{17}$$

where: - x is the current solution candidate - x_i is the ith antibody in the population - N is the population size - $g_i(x)$ is the fitness function of the ith antibody - w_i is the weight of the ith antibody - β is the affinity threshold - $a_i(x)$ is the affinity of the ith antibody to the current solution candidate x - c_i is the concentration of the ith antibody - p_i is the selection probability of the ith antibody - x_{best} is the best solution found so far - θ is the mutation rate of the antibody - a_{best} is the affinity of the best antibody - K is a scaling factor for the mutation rate - x_{avg} is the average of the population - γ is the hypermutation rate of the antibody - γ_{max} is the maximum hypermutation rate - α is a scaling factor for the hypermutation rate - c_{best} is the concentration of the best antibody - η is the mutation step size - $\mathcal{N}(0, \sigma^2)$ is the normal distribution with mean 0 and variance σ^2.

The AIO algorithm has been shown to be effective in handling various types of optimization problems, including continuous, discrete, and mixed-variable problems. It has also been applied

in multi-objective optimization problems, where the algorithm can identify a set of Pareto-optimal solutions.

Overall, the AIO algorithm is a promising optimization technique that can provide efficient and effective solutions for complex optimization problems.

3.2. Ai-LGBM: The proposed model

Our proposed novel model is a combination of the previous mentioned techniques that is LightGBM model and Auto Immune Optimization technique.

Initially we have performed k fold division of the dataset. K-fold cross validation is a commonly used technique for evaluating the performance of machine learning models on a tabular dataset. The basic idea is to divide the dataset into K subsets, or "folds," of roughly equal size. Then, for each fold i in the range 1 to K, a model is trained on the union of all the other folds, and evaluated on the i-th fold. This process is repeated K times, with each fold used exactly once as the test set. The resulting K evaluation scores can then be averaged to get an estimate of the model's generalization performance. Mathematically, K-fold cross validation can be represented as follows:

Let $D = (x_1, y_1), ..., (x_n, y_n)$ be a dataset of n samples, where x_i is a vector of input features and y_i is the corresponding output label. Let f be a machine learning model that maps inputs to outputs, and let $L(y_i, f(x_i))$ be a loss function that measures the difference between the predicted output $f(x_i)$ and the true output y_i. Then, K-fold cross validation can be defined as follows:

- Divide D into K disjoint subsets, D_1, ..., D_K, of roughly equal size.
- For each fold i in the range 1 to K, train a model f_i on the union of all the other folds, i.e., D D_i.
- Evaluate the model f_i on the i-th fold, D_i, and compute the evaluation score $S_i = 1/|D_i| * sum_{(x_j,y_j)\in D_i} L(y_j, f_i(x_j))$.
- Compute the average evaluation score $S = 1/K * sum_{i=1}^{K} S_i$.

Now after dividing the dataset into K number of sets we perform our analysis on each of the cross validation sets and observe how the models are performing on each set.

Here we have defined our custom model AI-LGBM which is optimized using Auto immune algorithm after prediction is done by the LightGBM model. we have provided the mathematical details of LightGBM model in section 3.1.2.Along with application of LightGBM we optimize the model using

AUto Immune optimization. We have also provided the mathematical details of auto immune optimization technique in section 3.1.3. After application of the following techniques we obtained our AI LGBM model for the prediction of the Ground water quality.

After prediction of AI-LGBM model we combine the results of this model with several other models like XGBoost anf LightGBM to create a ensemble technique. Here we have created a meta-learning model which will take the prediction results of the baseline models and perform a classification over the observation and performance of the models. The meta learning model combines the strengths of both meta learning and LightGBM, allowing us to achieve better performance in predicting the groundwater quality.

The meta learning approach involves training multiple models on different datasets and combining their predictions to obtain better performance. In this study, we use LightGBM as the base model for meta learning due to its efficient training and prediction times. LightGBM is a gradient boosting framework that uses decision trees to model the relationships between the features and the target variable.

The proposed meta learning model is trained on K-fold cross-validation, where the dataset is randomly divided into K equally sized folds. The model takes input the outputs of models, XGBoost, LightGBM and AI-LGBM predictions. In each iteration, the model is trained on K-1 folds and validated on the remaining fold. This process is repeated K times, with each fold being used once as the validation set. The final performance of the model is then calculated as the average of the performance on each fold.

The mathematical relation for training the LightGBM model on the i-th fold can be expressed as:

$$\hat{y}_j^i = \sum_{k=1}^{N} f_k(x_j^i) + \epsilon_j^i \tag{18}$$

where $\hat{y}_j^i$ is the predicted value for the j-th instance in the i-th fold, f_k is the k-th tree function in the LightGBM model, x_j^i is the feature vector for the j-th instance in the i-th fold, and ϵ_j^i is the error term.

The proposed meta learning model is then trained on the predictions of the LightGBM model for each fold. The mathematical relation for the meta learning model can be expressed as:

$$\hat{y}_j = \sum_{i=1}^{K} \alpha_i \hat{y}_j^i \tag{19}$$

where $\hat{y}_j$ is the predicted value for the j-th instance in the test set, α_i is the weight assigned to the i-th fold, and $\hat{y}_j^i$ is the predicted value for the j-th instance in the i-th fold. Meta learning model

based on LightGBM combines the strengths of both meta learning and LightGBM to achieve better performance in predicting the groundwater quality of a tabular dataset.

After we have created the meta learning model. Further when a new data points requires prediction, we will get the predictions of each model on the data point and with this meta learning LightGBM model we can extract the ensemble prediction result which will be most efficient.

After obtaining the results of the meta learning model we have completed the prediction part of the analysis and further we will move on to the explanation of the prediction with the explainable models.

3.3. Data Collection and Input

Initially for the analysis the dataset collection for analysis is crucial. We have collected the dataset containing information about groundwater quality. This study utilizes data collected from Vietnam and India, with a focus on the eastern Indian state of Odisha. The data is sourced from the Ground Water Yearbook Report, which offers detailed analysis and insights. To obtain the necessary data, the report from 2018-20 was consulted and converted into a CSV file format. Then after collecting the data we have provided the data as input to the methodology proposed. Since the data consists of spatial and compositional information we need a methodology which can process two different type of attributes. So first we preprocess the data and make it suitable for further model application.

3.4. Data preprocessing and merging

For further application of models on the data we need to prepare the raw data. Generally raw datasets has features like missing data, unnormalized features, values of attributes in different format. Here we have preprocessed the data using the general procedure. First, we process the missing data present in the dataset. There are various techniques to process the missing data. We have adopted the mode of the attributes to impute the missing values where the attribute is categorical. Also if the attribute is continuous, we imputed the missing values using mean of the attribute. This technique is useful since the number of missing value is less. If the missing value is larger we may have to use nearest neighbor based imputation technique like using random forest.

After imputing the missing values. we have introduced some new features which are engineered from existing relations. Some of the compositional features which are continuous are transformed to binary categorical values using permitted range of the compositional elements in the ground drinking water. Using this transformation all the element concentrations have been converted to binary features. Also some features are transformed according to the permittable ranges. Some new features are also engineered using traditional informative relations of Water quality index.

A water quality index (WQI) is a tool used to assess and communicate the overall quality of water in a particular area. It is a composite measure that combines multiple water quality parameters, such as pH, dissolved oxygen, total dissolved solids, nutrients, and pollutants, into a single score that represents the overall quality of the water. The index is designed to provide a simple and standardized way to evaluate the suitability of water for different uses, such as drinking, swimming, and aquatic life support. A higher WQI score indicates better water quality, while a lower score indicates poorer water quality. WQIs can be used by government agencies, researchers, and water resource managers to track changes in water quality over time, identify areas of concern, and prioritize management actions. Water quality index is computed using the information about the concentration of each elements along with pH, hardness of the water etc.

After feature engineering is done, we need to downscale the feature range of the numerical features. That's why we normalize the data. Data normalization is a process of scaling and transforming the input data in such a way that it falls within a specific range or follows a specific distribution. The goal of data normalization is to make sure that all features in the dataset are given equal importance during the analysis. Here we have normalized the data using the following relation.

$$F'(x) = \frac{F(x) - F(\bar{x})}{\sigma(F(x))} \tag{20}$$

where,

$$F(\bar{x}) = \frac{\sum_{i=1}^{n} F(x_i)}{n} \tag{21}$$

$$\sigma(F(x)) = \frac{1}{n}\sum_{i=1}^{n}(F(x_i) - \mu)^2 \tag{22}$$

Subsequently after normalizing the features using the following relations mentioned in the Eq. 20, we use the preprocessed dataset for further analysis.

3.5. Prediction and Analysis

After we recieve the processed and engineered data we proceed with applying variant models on the data to obtain a trained predictive model. Before application of models on the data we have divided the data in k different sets of fold and performed separated model analysis on the data. Thus process is called K-fold cross validation [15]. It involves splitting the original dataset into K equal subsets (or folds) of approximately equal size. Then, the model is trained K times, where each time the model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times with each fold serving as the test set exactly once. The results from each of the K runs are then averaged to produce

a single performance estimate. The advantage of using K-fold cross-validation is that it reduces the variance of the model evaluation and provides a more accurate estimate of the model's performance on new data. It is a powerful tool for preventing overfitting and ensuring that the model is able to generalize well to new, unseen data.

Here we have used various models for our analysis comparison. These models are Logistic Regression [16], Support Vector Machines [17], KNN Classifier [18], XGBoost [19], LightGBM [20]. We have compared our proposed model AI-LGBM based ensemble model which is described in 3.2. Finally we created a ensemble model based on XGBoost, LightGBM and AI-LGBM using the voting procedure mentioned in Eq. 23.

$$Y_{final} = \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i} \tag{23}$$

Where W_i is the metric value for model i and Y_i is the model result.

3.6. Explanation and Interpretation

Following prediction of the models we move forward to explanation the results and interpreting it. The explanation of prediction results is essential in understanding how a machine learning model works and how it arrived at its predictions.

Here we use techniques like LIME, SHAP for explaining the results of the prediction. LIME (Local Interpretable Model-Agnostic Explanations) [21] is a technique for explaining the predictions of machine learning models. It is a model-agnostic method, which means it can be applied to any machine learning model regardless of the underlying algorithm or architecture. The goal of LIME is to provide local explanations for individual predictions made by the model. In other words, given an input instance and its corresponding output prediction, LIME seeks to identify which features of the input were most important in making that prediction. It does this by building a simpler, interpretable model around the input instance of interest and examining the weights assigned to each feature by this simpler model. The basic idea behind LIME is to generate a set of perturbed samples around the input instance of interest, then train a simpler, interpretable model on this set of perturbed samples. The simpler model should be both interpretable and have high accuracy on the perturbed samples. Once the simpler model is trained, the feature weights of the model can be used to explain the prediction made by the original model on the input instance of interest.

SHAP (SHapley Additive exPlanations) [22] is a method for interpreting the output of machine learning models. It provides a framework for understanding the importance of each feature in a model's output, which is particularly useful when dealing with complex models that are difficult to interpret.

SHAP builds upon the idea of the Shapley value from cooperative game theory, which measures the marginal contribution of each player to a coalition. In the context of machine learning, each feature in a dataset can be thought of as a player, and the output of a model can be thought of as the payoff of a coalition. SHAP uses a weighted linear regression model to estimate the Shapley values for each feature. The weights are determined by minimizing the error between the model's predictions and the true outcomes in a local region of the feature space. The resulting Shapley values provide a measure of the impact that each feature has on the model's output. SHAP has several advantages over other model interpretation methods. For example, it can handle complex models such as ensemble models and deep neural networks, and it can provide global as well as local explanations of model predictions. Additionally, SHAP values are theoretically grounded and have desirable properties such as consistency and additivity.

We use this models for explaining the results of the models. Providing explanations helps increase transparency and accountability, which is especially important in industries such as healthcare, finance, and law where decisions based on machine learning can have significant impacts on people's lives. It also helps build trust in the model and can improve its adoption and acceptance by stakeholders. Furthermore, explanations can reveal insights into the data and help identify biases or errors in the model, leading to improvements in future versions.

4. Experiments and data

To conduct a thorough analysis of groundwater quality, it is crucial to collect accurate and extensive data. The quality of groundwater depends on various factors, including natural conditions and human activities, and it is essential to measure various parameters and attributes to evaluate its quality. Therefore, researchers use a combination of field experiments and laboratory analyses to gather data on groundwater quality. In this section, we will discuss the experiments and data used in recent research studies that employ various statistical and machine-learning models to assess groundwater quality. We will also delve into the various data preprocessing steps used to ensure the accuracy and relevance of the data. By understanding the experimental procedures and data sources used in these studies, we can gain a better understanding of the potential and limitations of statistical and machine learning models for groundwater quality assessment. In Sec. To conduct a thorough analysis of groundwater quality, it is crucial to collect accurate and extensive data. The quality of groundwater depends on various factors, including natural conditions and human activities, and it is essential to measure various parameters and attributes to evaluate its quality. Therefore, researchers use a combination of field experiments and laboratory analyses to gather data on groundwater quality.

In this section, we will discuss the experiments and data used in recent research studies that employ various statistical and machine learning models to assess groundwater quality. We will also delve into the various data preprocessing steps used to ensure the accuracy and relevance of the data. By understanding the experimental procedures and data sources used in these studies, we can gain a better understanding of the potential and limitations of statistical and machine learning models for groundwater quality assessment.4.1 we provided a description of the dataset used for analysis and in Sec 4.3 we have provided our experimental setup for the analysis.

4.1. Datasets used

Data is taken from two South East Asian country, Vietnam and India. The eastern Indian state of Odisha is the main subject of this study. The Ground Water Yearbook Report [1] examines and illustrates this data further. As a result, the data is taken from the 2018–20 published report and appropriately translated to a CSV file. There are 14 Parameters in the dataset. The data set used in this study, which includes 1241 rows and 17 columns with sample values for each, was gathered from the Indian state of Odisha. Fourteen of the seventeen columns—SL No., District, and Village—belong to physicochemical characteristics. There are 2 categories and 15 numerical variables. The dataset does not include any null values. The first (25%), second (50%) and third (75%) quartiles of data for the Carbonate column are found to be zero when a statistical distribution is performed on the dataset. Therefore, we have determined that the majority of the values in the Carbonate column are zero; if there is a value, it may be an anomaly. It is determined to remove the Carbonate column from the dataset as a result. Hence, there are now just 13 physicochemical characteristics in all.

The dataset from Vietnam was acquired through the water resource monitoring network in the Northern Delta and the North Central Coast as a component of the national water resource monitoring network in particular and the environmental resource monitoring network in general. The national underground water resource monitoring network is uniformly implemented nationwide according to the Circular 19/2013/TT-BTNMT dated July 19, 2013, of the Ministry of Natural Resources and Environment providing technical regulations on water resources monitoring. In which, underground water quality monitoring is one of the main contents of the task "National monitoring of water resources". In the case of Vietnam dataset, 12 parameters are inclusive; thus, the data is extracted from the national water resources planning and inspection centre federation for planning and investigation of water resources in the North.

Table 2: Spatial Attributes present in the dataset

Attributes	Description
Well Code	Label provided for specific well in special locations
Date_Sampling	Date of collecting the sample of water
Quarter	Quarter during which sample was collected
Type_analyzing	Kind of Analysis being performed initially
Date_analyzing	Date on which analysis is performed
Laboratory	Laboratory at which analysis was done
Number_analyzing	Number of analysis that was performed

4.2. Feature Description

Features are important to the predictions of the models. The attributes include the well code, date of sampling, quarter during which the sample was collected, type of analysis being performed, laboratory at which the analysis was done, and the number of analysis performed. In addition, the concentrations of various ions and compounds such as sodium, potassium, calcium, magnesium, ferric and ferrus ions, aluminium, ammonium, chlorine, sulphate ion, bicarbonate, cobalt ions, nitrogen dioxide, nitrogen trioxide, and phosphate ions were also measured. The attributes also include the general and temporary hardness observed in the sample, pH, reduction potential, and concentrations of various other compounds such as oxygen, carbon dioxide, and silicon dioxide. Other attributes include the color, smell, taste, and the total dissolved solids (TDS) concentrations of the samples. These attributes provide crucial information for assessing the quality of groundwater in a particular region and identifying any potential risks associated with its use.

The dataset is a collection of spatial features as well as the chemical composition of the water sampled from the groundwater. We have presented different features in two different tables. The first table 2 contains spatial features of the place from where the water is collected from. the second table 3 contains water chemical composition which is extracted from the sample inside the laboratory.

4.3. Experimental Setup

The model was implemented in Pycharm 2022.2.1. and tested on Google Colab using Python 3.9.13 and the key support modules Pandas, Numpy, Matplotlib, and Seaborn. For classification, the models like Random Forest, Logistic Regression, and Support Vector Machine are imported from the library sklearn. XGBoost model is imported from library XGBoost. The LightGBM model is imported from the library LightGBM. Mutual information-based feature selection is implemented customarily using sklearn and numpy. Auto Immune optimisation uses Numpy as the base library. Finally, an explanation using LIME and SHAP is done using libraries lime and shap, respectively for explaining the prediction.

Table 3: Composition Attributes present in the dataset

Attributes	Description
Na	Concentration of Sodium in the sample
K	Concentration of Potassium in the sample
Ca2	Concentration of Calcium in the sample
Mg2	Concentration of Magnesium in the sample
Fe3	Concentration of Ferric ion in the sample
Fe2	Concentration of Ferrus in the sample
Al3	Concentration of Aluminium in the sample
NH4	Concentration of Ammonium in the sample
Cl	Concentration of Chlorine in the sample
SO4	Concentration of Sulphate ion in the sample
HCO3	Concentration of Bicarbonate in the sample
Co3	Concentration of Cobalt ions in the sample
NO2	Concentration of Nitrogen dioxide in the sample
Hardness_general	General Hardness of the sample
NO3	Concentration of Nitrogen trioxide in the sample
PO4	Concentration of Phosphate ions in the sample
Hardness_temporal	Temporary hardness observed in the sample
Hardness_permanent	Permanent Hardness observed in the sample
pH	pH of the sample to determine acidity or basic nature
eH	Reduction Potential of the sample
CO2_free	Concentration of free Carbon dioxide in the sample
CO2_depend	Concentration of dependent Carbon dioxide in the sample
CO2_infiltrate	Concentration of Carbon dioxide in the infiltrate
Oxygen	Concentration of Oxygen in the sample
Lienhe	Lienhe for the sample
Conductivity	Conductivity of the Sample
Oxygen Dissolve	Concentration of dissolved Oxygen in the sample
SiO2	Concentration of Sillicon dioxide in the sample
Color	Color of the collected sample
Smell	Smell of the collected sample
Taste	Taste of the collected sample
TDS105	tds105 Concentration in the sample
TDS180	tds180 Concentration in the sample

Feature Name	Feature Importance Score
wqi	0.224555
cl	0.170545
hardness_general	0.163163
well_code	0.161123
k	0.138051
tatse	0.109378
hco3	0.071224
so4	0.043977
co2_infiltrate	0.042858
ca2	0.041503
co2_free	0.034894
sio2	0.016268
hardness_temporal	0.012693
fe2	0.012169
no3	0.009952
fe3	0.008096
date_sampling	0.007427
quarter	0.003557
color	0.00306
date_analyzing	0.002826
no2	0.002231
smell	0.001111
type_analyzing	0.000407
laboratory	0
al3	0
co3	0

Table 4: Feature Importance Score according MIFS

5. Results & Discussions

In this section we have presented the results obtained from our methodology. The results provide insights into the effectiveness and performance of the proposed model in predicting groundwater quality for the tabular dataset. First we have discussed about the preprocessing steps and the inferences extracted from the dataset in section 5.1. Feature Importance and other feature engineering steps analysis is presented in section 5.2. Finally Our proposed model results are presented in section 5.3.

5.1. Data pre-processing

Data preprocessing is a crucial step in any data analysis or machine learning project. It involves transforming raw data into a format that is suitable for further analysis and model training. The main objective of data preprocessing is to clean, transform, and enhance the quality of the data, ensuring that it is accurate, complete, and ready for analysis.

The data preprocessing phase typically includes several key steps. Firstly, data cleaning is performed to handle missing values, outliers, and inconsistent data entries. Missing values can be imputed

Folds	Ranfom Forest	XGBoost	LightGBM	AI-LGBM
2	0.974	0.987	0.9903	0.996
3	0.9742	0.9867	0.9914	0.9958
4	0.9784	0.9893	0.996	0.9978
5	0.9804	0.9912	0.9964	0.9992
6	0.98	0.9906	0.9958	0.9954
7	0.9764	0.9851	0.9943	0.9947
8	0.9733	0.9816	0.9937	0.9939
9	0.9703	0.98	0.9928	0.994
10	0.9679	0.9746	0.9904	0.9918

Table 5: Model Performance comparison for different number of folds

using techniques such as mean imputation, regression imputation, or multiple imputation. Outliers may be detected and treated through methods like z-score analysis or clustering-based approaches.

Here we have observed missing values in the dataset. Several features have a large number of missing values. Features like, 'number_analyzing', 'nh4', 'po4', 'eh', 'oxygen', 'lienhe', 'conductivity', 'oxygen_dissolve',, 'tds180' have significant amount of missing values. We have removed such features to avoid misdirected modelling of the data.

Also after missing value imputations we have transformed the features according to their permissible limit in ground water. So these features have converted in binary variables from continuous features. As for example if pH level is between 5.5 and 8.5 it is considered permissible ground water. So those datapoints with pH value in this region is considered as 1 and for rest of the data points it is considered as 0. Following transformation of the data continuous variables are normalized to contain the values in a specific range.

5.2. Statistical Analysis

We have performed statistical analysis on the dataset to extract high level information from the data. These analytical methods contains univariate analysis, bivariate analysis and outlier analysis.

Univariate analysis is a fundamental technique in data analysis that focuses on exploring and summarizing individual variables in a dataset. It involves examining and analyzing one variable at a time to gain insights into its distribution, central tendency, variability, and other key characteristics. Univariate analysis provides a basic understanding of the data and serves as a foundation for further statistical analysis and modeling.

Bivariate analysis is a statistical method used to examine the relationship between two variables in a dataset. It aims to explore the association, correlation, or dependency between two variables and understand how they interact with each other.

Outliers are data points that deviate significantly from the normal patterns or distributions of

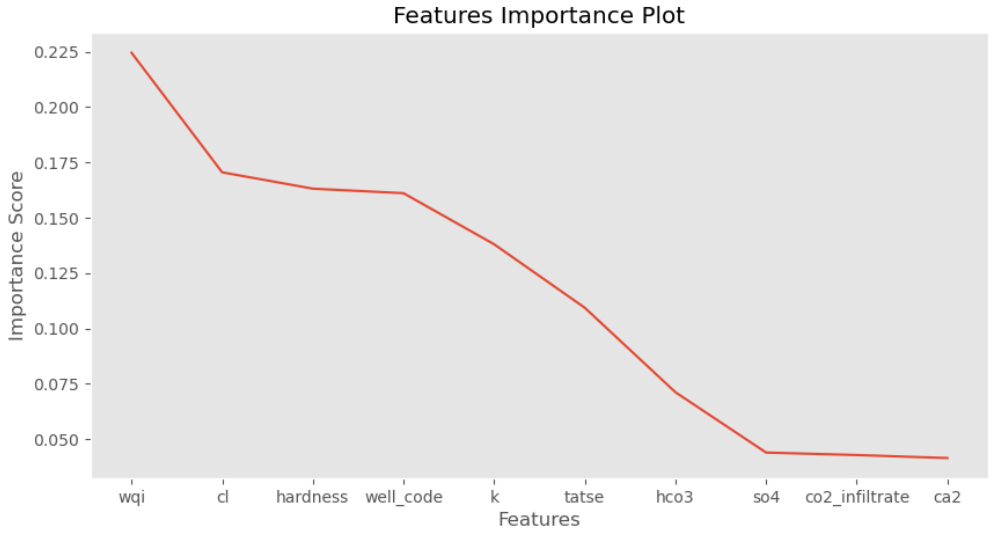


Fig. 2. Feature Importance Plot

the rest of the data. They can arise due to various reasons such as measurement errors, data entry mistakes, or genuinely unusual observations. Analyzing and handling outliers is important to ensure the integrity and reliability of data analysis and modeling.

5.3. Results of AI-LGBM

5.3.1. Feature Selection Comparative Study

In this section we have provided feature selection analysis from MIFS feature selection method. MIFS is an advanced feature selection method and is described in section3.1.1. The results are provided in Table 4. Here, we can observe that water quality index engineered as part of the feature engineering described in section 5.1. Among other features, Hardness, the well from where sample is collected, water component compositions are important to determining the ground water quality. The impact of these features are provided in the later section. We also provided a plot comparing various features according to their feature importance score in Fig. 2.

5.3.2. Optimization Results

In this section we have provided the results of the Auto Immune optimization (AIO). We have provided the mathematical basis of AIO in section 3.1.3. We have tried to optimize the depth, learning rate and number of leaves of the AI-LGBM model. The possible values of maximum depth are considered in the range of 5 to 30 at a interval of 5. Similarly for learning rate the values are considered in from 0.005 to 0.05. Also for number of leaves the values are considered in range from 10 to 50 with a interval of 10. From the optimization performed we have obtained the best hyperparameter values as maximum depth as 5, learning rate as 0.01, and number of leaves as 30.

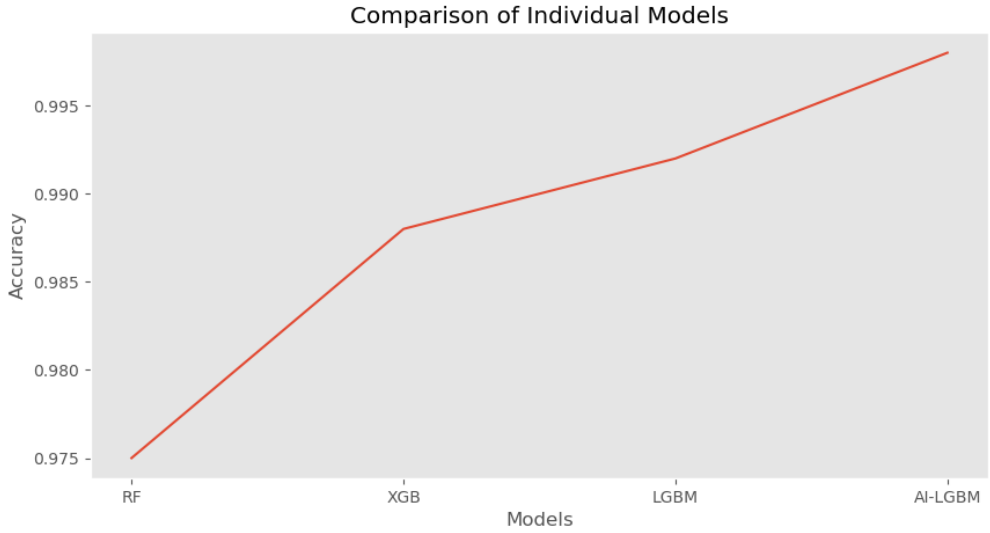


Fig. 3. Individual Models Performance Comparison

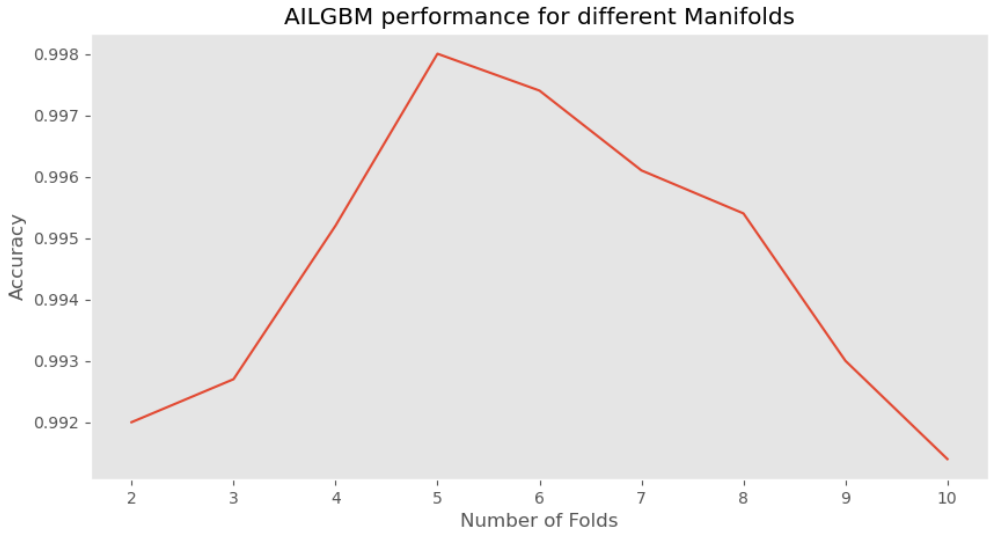


Fig. 4. AI-LGBM Performance Comparison

5.3.3. Classification Comparative Study

In this section we have provided analysis of model results. Several comparative studies have been provided. Individual model performances are compared with our proposed model AI-LGBM model. Additionaly, comparison has been done with number of folds considered for analysis. The results are provided below.

In order to evaluate and compare the performance of various prediction models, we conducted a comparative study involving Random Forest, XGBoost, LightGBM, AI-LGBM, and our proposed Ensemble model. The comparison was conducted using a 5-fold cross-validation approach, as it provided the best results in terms of model evaluation.

From Fig. 3, our observations revealed that our proposed AI-LGBM model exhibited superior

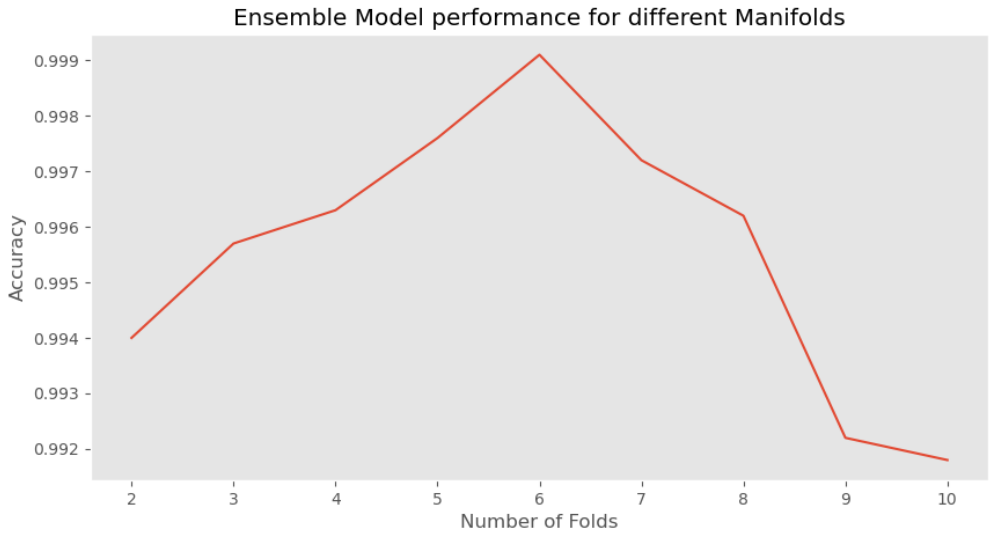


Fig. 5. Ensemble Model Performance Comparison

performance compared to the other models in the study, namely Random Forest, XGBoost, and LightGBM. The accuracy achieved by AI-LGBM on the testing set was an impressive 0.9992, showcasing its remarkable predictive capabilities.

The exceptional performance of our AI-LGBM model further substantiates its superiority over traditional and state-of-the-art prediction models. The significantly higher accuracy achieved by AI-LGBM highlights its ability to capture complex patterns and relationships within the dataset, leading to more precise and reliable predictions.

The results of this comparative study demonstrate the effectiveness and competitiveness of our proposed AI-LGBM model in the realm of prediction modeling. The exceptional performance of AI-LGBM not only enhances the predictive accuracy but also boosts the confidence in the reliability of the model's outputs. These findings reinforce the potential practical applications of AI-LGBM in various domains, including but not limited to environmental analysis, resource management, and decision-making processes.

In Fig. 4 we have presented performance comparison of AI-LGBM across different considerations of number of folds. From the plot results, it can be observed that performance accuracy in highest at 5 folds. For other consideration the performance accuracy decreases. Ensemble model performance is also compared with different number of fold consideration in Fig. 5. Similar to AI-LGBM observation it is observed that the ensemble model is providing optimum performance for 5 folds considered.

Aside from comparison of number of folds, we have used meta learning models to ensemble the results of different models using a predictive model LightGBM. The results of meta learning is provided in Fig. 6. The meta learning model performance is compared for different fold consideration. It is

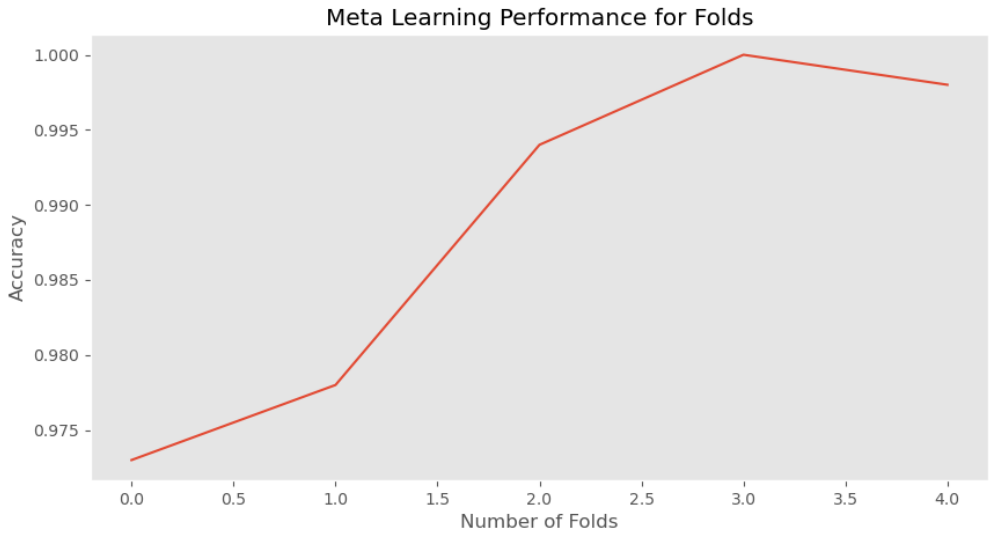


Fig. 6. Meta Learning Performance Comparison

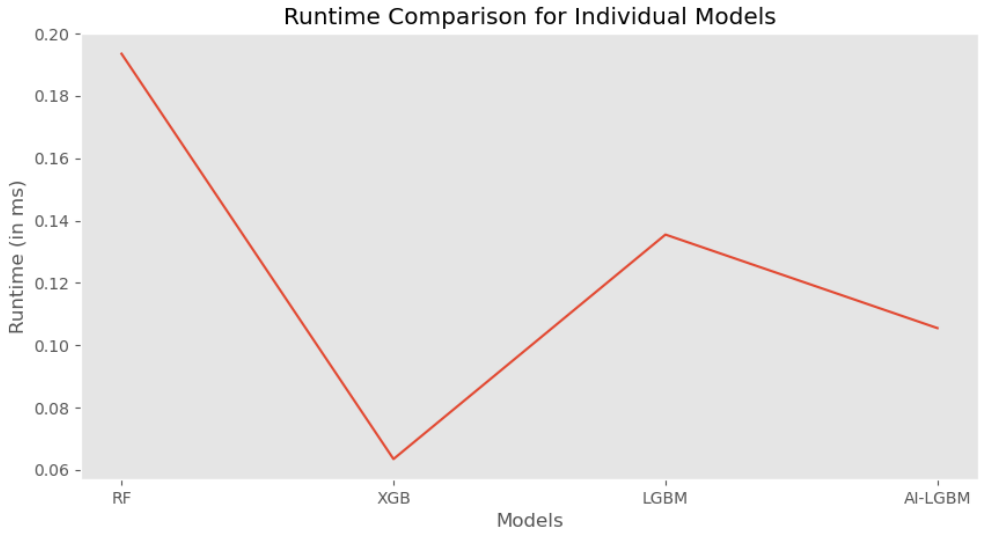


Fig. 7. Individual Models Runtime Comparison

observed that the model is providing highest performance for third fold.

So from these results and comparative study we can infer that our proposed methodology outperforms existing traditional and state of the art analysis techniques and confidently predict the ground water quality.

5.3.4. Runtime comparison

We have performed runtime comparison among the considered models and also different number of folds considered. It involves evaluating the efficiency of an algorithm by measuring the amount of time it takes to execute as a function of the input size. Understanding the importance of runtime analysis is crucial for several reasons. Firstly, runtime analysis helps in comparing different algorithms and selecting the most efficient one for a given problem. By analyzing the runtime complexities of different

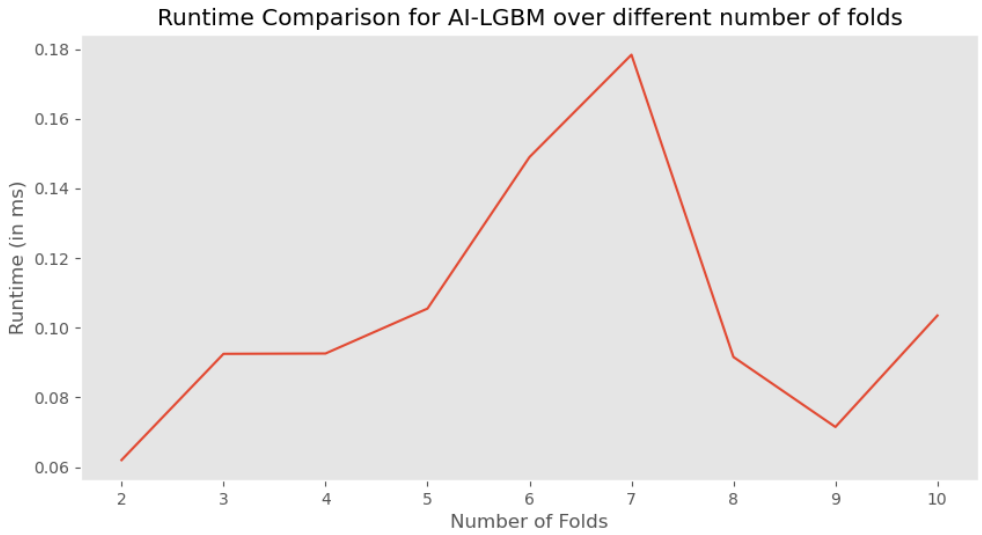


Fig. 8. AI-LGBM Runtime Comparison on number of Folds

algorithms, we can identify which algorithm has a better performance in terms of time efficiency. This information is particularly valuable when dealing with large-scale or time-sensitive applications where even small improvements in runtime can have significant impacts. Secondly, runtime analysis allows us to predict the scalability of an algorithm. As the input size increases, some algorithms may experience a significant increase in runtime, making them impractical for large datasets or real-time applications. By analyzing the runtime complexity of an algorithm, we can estimate its performance for various input sizes and assess its scalability. This information helps in making informed decisions about algorithm selection and system design.

The comparison of runtime among different models are presented in Fig. 7. From the plot, we can infer that xgboost has the lowest runtime of all the models. Also AI-LGBM model has a lower runtime than LightGBM and Random Forest model. Also comparison of runtime of AI-LGBM model over different number of folds are present in Fig. 8. From this plot we can observe that, runtime is lowest when number of folds are 2. Also, it is also observed that when number of folds is 9 runtime is lower than other considerations. Runtime is highest when considered number of folds are 7. Additionally, another comparison is also performed for runtime comparison for different fold of analysis. It is observed that our model is performing optimally when the considered number of folds are 5. The comparison results are provided in Fig. 9. It is observed that the model is taking lowest time of first fold and highest time of second fold.

5.4. Explanation and Interpretation

In this section we have provided an detailed analytical explanations of the prediction results obtained. The importance of explanations of prediction results cannot be overstated in the realm of

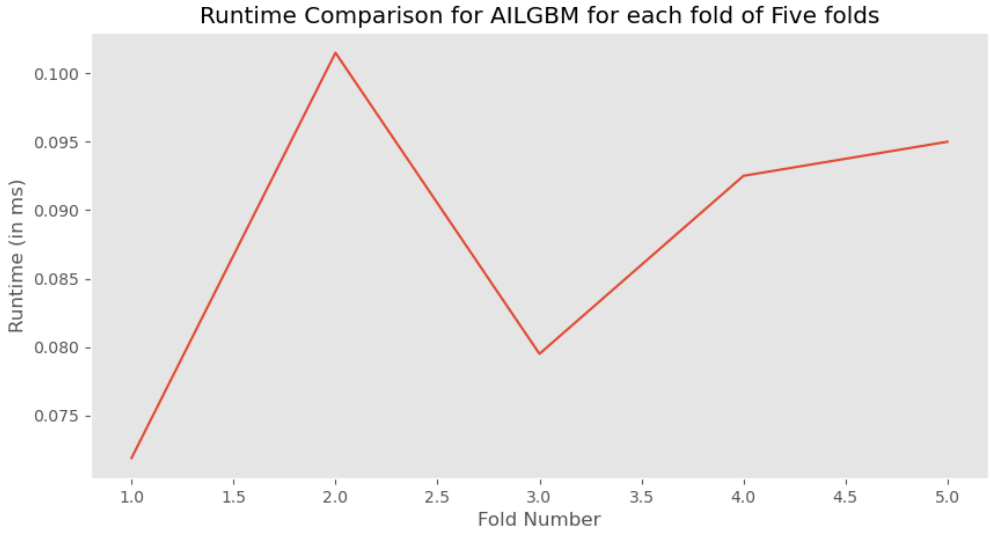
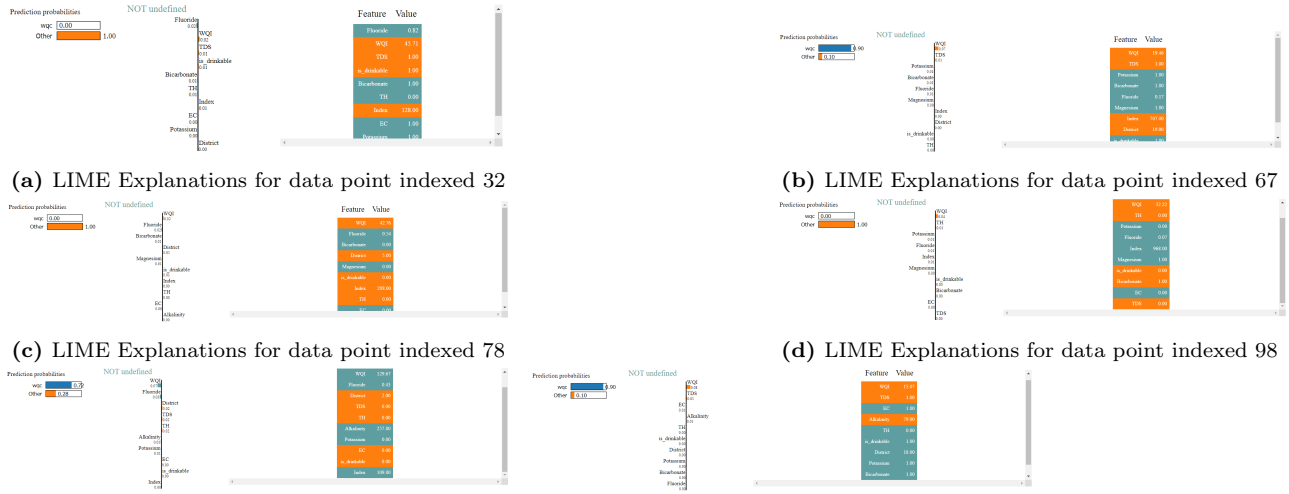


Fig. 9. AI-LGBM Runtime Comparison on different Folds



(a) LIME Explanations for data point indexed 32

(b) LIME Explanations for data point indexed 67

(c) LIME Explanations for data point indexed 78

(d) LIME Explanations for data point indexed 98

(e) LIME Explanations for data point indexed 135 **(f)** LIME Explanations for data point indexed 149

Fig. 10. LIME explanation Plots

data analysis and machine learning. While accurate predictions are valuable, understanding the underlying reasons and factors contributing to those predictions is equally crucial. Explanations provide transparency, interpretability, and insights into the decision-making process of a model, allowing stakeholders to trust and effectively utilize the predictions. One key advantage of explanations is the ability to gain insights into the features or variables that have the most significant influence on the predictions. By understanding which features are driving the predictions, we can identify important patterns, relationships, and causal factors in the data. This knowledge can be invaluable in various domains, such as healthcare, finance, and marketing, where understanding the key drivers behind predictions can guide decision-making, improve processes, and identify areas for intervention or improvement.

We have utilised two different techniques for explaining and interpreting the prediction of the results. These techniques are LIME and SHAP. The mathematical details of LIME and SHAP are

(a) SHAP Force Plot of First Variable for data point indexed 32

(b) SHAP Force Plot of Second Variable for data point indexed 32

(c) SHAP Force Plot of First Variable for data point indexed 67

(d) SHAP Force Plot of Second Variable for data point indexed 67

(e) SHAP Force Plot of First Variable for data point indexed 78

(f) SHAP Force Plot of Second Variable for data point indexed 78

(g) SHAP Force Plot of First Variable for data point indexed 98

(h) SHAP Force Plot of Second Variable for data point indexed 98

(i) SHAP Force Plot of First Variable for data point indexed 135

(j) SHAP Force Plot of Second Variable for data point indexed 135

(k) SHAP Force Plot of First Variable for data point indexed 149

(l) SHAP Force Plot of Second Variable for data point indexed 149

Fig. 11. SHAP Force Plots

provided in section 3.1. We have performed prediction using ensemble model consisting of our novel defined model AI-LGBM. After performing prediction we provided the trained modelas input to these models. LIME perform local explanations where SHAP perform global explanations. The explanations from LIME for some of the local data points are provided in Fig. 10. Here we can observe some sample explanations of prediction on different data points. Here, for an example, if we consider the data point indexed as 32, the WQI value of 45.71 falls on non drinkable side. Since it is the most important feature inferenced from section 5.2. That's why it is predicted as Non drinkable.

Here we have also provided SHAP explanation results. This model is able to perform explanation both locally and globally. The results of SHAP application is presented in Fig. 11. Several data points are explained to understand the explanations. For example, for 32 indexed data point Fig. 11[a] shows the important features which are contributing to the prediction value becoming 0. And Fig 11[b] show for same data points the features which are contributing to prediction value becoming 1. In this way we can explain other instances also. In Fig. 12 shows global distribution of WQI for all sample instances.

6. Conclusions

In this section we have compiled all the analysis and observations obtained from the analysis, First we have provided an overview of the analysis in section 6.1. Following providing the overview,

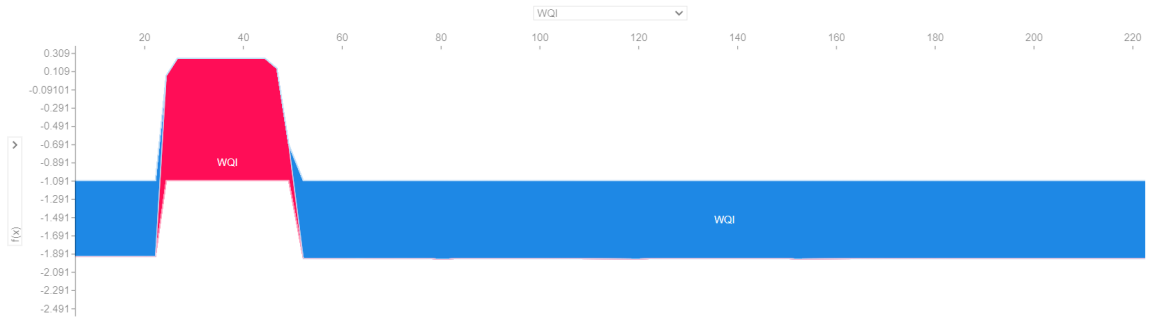


Fig. 12. Force Plot of WQI using SHAP

contribution of the proposed work to the theory is discussed in section 6.2. Also contribution of the proposed work to the practice is also discussed in section 6.3. Finally, limitations and future scopes of the work is discussed respectively.

6.1. Overview of the Analysis

In this study, we conducted a comprehensive analysis of a tabular dataset focusing on ground water quality prediction. The research involved multiple stages, beginning with a basic statistical analysis to gain initial insights into the data. Subsequently, we employed our proposed ensemble model, AI-LGBM, to predict the ground water quality. Notably, our results demonstrated that AI-LGBM outperformed both traditional and state-of-the-art techniques in terms of predictive accuracy. Furthermore, we assessed the runtime performance of our model and compared it with other existing techniques, highlighting the efficiency and effectiveness of AI-LGBM. Finally, we emphasized the importance of interpretability and understanding by employing explainable models to elucidate the rationale behind our predictions. The comprehensive analysis and superior performance of AI-LGBM underscore its potential as a valuable tool for ground water quality prediction. Overall, our findings highlight the significance of advanced machine learning techniques, such as AI-LGBM, in addressing the challenges and complexities of ground water quality analysis, while providing transparent and interpretable results for informed decision-making in water resource management and environmental protection.

6.2. Contribution to the Theory

Our research makes significant contributions to the theory of data analysis and modeling by introducing several novel techniques and methodologies. Firstly, we proposed the utilization of Mutual Information based Feature Selection, which is a non-parametric approach. By incorporating probabilistic feature selection, this model effectively captures the intricate non-linear correlations among features and iteratively updates their importance. This advancement enables a more comprehensive understanding of the underlying data structure.

Furthermore, we integrated Auto Immune optimization techniques into our research, which exhibit metaheuristic properties. This approach enhances the robustness of the models and allows for their application in parallel computing environments. By eliminating the need for gradient information, our methodology achieves higher computational efficiency, enabling faster and more scalable analyses. To enhance the interpretability of our predictive models, we incorporated explainable techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME provides localized insights into the models' decision-making process, while SHAP analyzes the global structure of the models. The combination of these explainable models empowers researchers to gain a deeper understanding of the factors influencing predictions and increases the transparency and interpretability of the overall analysis. Lastly, our methodology presents an integrated analytical framework that encompasses all the aforementioned techniques, including feature selection, efficient predictive models, and explainable models. By combining these components, we establish a cohesive and organized approach to data analysis, allowing for comprehensive exploration and understanding of complex datasets.

Overall, our contributions advance the theory of data analysis by introducing novel techniques that address the limitations of traditional approaches. Through the integration of non-parametric feature selection, metaheuristic optimization, and explainable models, we enhance the interpretability and accuracy of our analyses, providing valuable insights into complex data patterns.

6.3. Contribution to the Practice

Our research work makes significant contributions to the practical implementation of water quality prediction and analysis. In the context of Indian villages facing water scarcity, our research holds great significance. By employing our methodology, we can identify key factors and variables that contribute to the drinkability of water sources. This knowledge can aid in understanding the reasons behind the presence of drinkable water in certain areas, despite water scarcity. Consequently, our work can inform policymakers, water resource management authorities, and local communities about the potential sources of safe drinking water, enabling more targeted and effective interventions to address water scarcity issues. Furthermore, our research has practical implications for water irrigation systems. Accurate prediction and analysis of water quality parameters are essential for efficient and sustainable irrigation practices. By utilizing our proposed methodology, water irrigation systems can be better informed about the quality of the water source. This information can help optimize irrigation strategies, prevent potential water contamination risks, and improve overall agricultural productivity.

Overall, our research contributes to the practical implementation of water quality analysis by providing an organized analytical framework. The ability to predict water quality and explain factors

contributing to drinkable water in water-scarce regions, as well as the impact on water irrigation systems, offers valuable insights for real-world applications. These practical implications can guide decision-makers, policymakers, and stakeholders in making informed decisions regarding water resource management and addressing water scarcity challenges in Indian villages and similar contexts.

We have utilised limited data resources for training our model which may bring somr drawback to our methodology. In the future we want to extend our work in more domains water resources. Also application of AI-LGBM in other domains present another possible scope for the future.

References

[1] M. Noshadi, A. Ghafourian, Groundwater quality analysis using multivariate statistical techniques (case study: Fars province, iran), Environmental monitoring and assessment 188 (2016) 1–13.

[2] R. Rana, R. Ganguly, A. K. Gupta, Indexing method for assessment of pollution potential of leachate from non-engineered landfill sites and its effect on ground water quality, Environmental monitoring and assessment 190 (1) (2018) 46.

[3] Y. Zhou, X. Wang, W. Li, S. Zhou, L. Jiang, Water quality evaluation and pollution source apportionment of surface water in a major city in southeast china using multi-statistical analyses and machine learning models, International Journal of Environmental Research and Public Health 20 (1) (2023) 881.

[4] S. K. S. Al-Doori, Y. S. Taspinar, M. Koklu, Distracted driving detection with machine learning methods by cnn based feature extraction, International Journal of Applied Mathematics Electronics and Computers 9 (4) (2021) 116–121.

[5] J. Charles, G. Vinodhini, R. Nagarajan, An efficient feature selection with weighted extreme learning machine for water quality prediction and classification model, Annals of the Romanian Society for Cell Biology (2021) 1969–1994.

[6] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, M. P. Mendes, Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods, Science of the total environment 624 (2018) 661–672.

[7] E. Dritsas, M. Trigka, Efficient data-driven machine learning models for water quality prediction, Computation 11 (2) (2023) 16.

[8] H. Ibrahim, Z. M. Yaseen, M. Scholz, M. Ali, M. Gad, S. Elsayed, M. Khadr, H. Hussein, H. H. Ibrahim, M. H. Eid, et al., Evaluation and prediction of groundwater quality for irrigation using

an integrated water quality indices, machine learning models and gis approaches: A representative case study, Water 15 (4) (2023) 694.

[9] H. Zheng, Y. Liu, W. Wan, J. Zhao, G. Xie, Large-scale prediction of stream water quality using an interpretable deep learning approach, Journal of Environmental Management 331 (2023) 117309.

[10] Y. Li, R. Li, Predicting ammonia nitrogen in surface water by a new attention-based deep learning hybrid model, Environmental Research 216 (2023) 114723.

[11] A. D. Gorgij, G. Askari, A. Taghipour, M. Jami, M. Mirfardi, Spatiotemporal forecasting of the groundwater quality for irrigation purposes, using deep learning method: Long short-term memory (lstm), Agricultural Water Management 277 (2023) 108088.

[12] R. Brázdil, K. Chromá, P. Zahradníček, P. Dobrovolnỳ, L. Dolák, Weather and traffic accidents in the czech republic, 1979–2020, Theoretical and Applied Climatology (2022) 1–15.

[13] A. O. Al-Sulttani, M. Al-Mukhtar, A. B. Roomi, A. A. Farooque, K. M. Khedher, Z. M. Yaseen, Proposition of new ensemble data-intelligence models for surface water quality prediction, IEEE Access 9 (2021) 108527–108541.

[14] B. Li, J. Chen, Z. Huang, H. Wang, J. Lv, J. Xi, J. Zhang, Z. Wu, A new unsupervised deep learning algorithm for fine-grained detection of driver distraction, IEEE Transactions on Intelligent Transportation Systems (2022).

[15] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, S. Ridella, The'k'in k-fold cross validation., in: ESANN, 2012, pp. 441–446.

[16] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, Logistic regression, Springer, 2002.

[17] W. S. Noble, What is a support vector machine?, Nature biotechnology 24 (12) (2006) 1565–1567.

[18] L. E. Peterson, K-nearest neighbor, Scholarpedia 4 (2) (2009) 1883.

[19] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).

[21] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[22] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).