Innovations in Geospatial Data Analysis: Applied Nonlinear Analysis, Remote Sensing, AI, and GIS for Environmental Sustainability

Dr C. Sudha[1], Ali Bostani[2], S.A.Sudha[3], Dr.T.Elangovan[4], C. Nandhini[5], Zafar Kurbonov[6]

[1]Assistant Professor, Computer science and engineering, GITAM school of Technology, GITAM deemed to be university, Hyderabad , India. Email: drcsudha17@gmail.com

[2]Assistant Professor, College of Engineering and Applied Sciences, American University of Kuwait, Salmiya, Kuwait. Eamil: abostani@auk.edu.kw

[3]Assistant Professor, Department of Computer Applications, Erode Arts And Science College, Erode, Tamilnadu, India

[4]Assistant Professor, Department of Computer Science,
Erode Arts And Science College, Erode, Tamilnadu, India
Email elangobrave@gmail.com

[5]Research Scholor, Department of Computer science,
Erode Arts And Science College, Erode, Tamilnadu, India
Email: ashonanthu@gmail.com

[6]The Department of Applied Mathematics, Karshi State University, Karshi, Uzbekistan. Email: urbonov.zm@qarshidu.uz

Article History:

Received: *05-06-2024*

Revised: *07-07-2024*

Accepted: *27-07-2024*

Abstract:
Water mapping plays a pivotal role in sustainable water resource management, particularly in the context of escalating climate change impacts. This study addresses the critical need for accurate water level predictions by introducing an innovative ensemble machine learning (ML) approach. Motivated by the increasing importance of ML and mathematical models in geospatial data analysis, the research papers from the diverse database, revealing a research gap that includes the absence of standardized methodologies and the exploration of diverse ensemble methods. Leveraging the EuroSat dataset for land use and land cover classification, the proposed ensemble model combines Principal Component Analysis, Genetic Algorithm, Gradient Boosted Decision Trees, Frequency Ratio, Deep Neural Network, and Shannon's Entropy. Demonstrating superior accuracy at 98.5%, Precision at 90.4%, and recall at 92%, EnsembleML emerges as a robust solution, emphasizing the advantage of ensemble techniques for comprehensive water mapping in the face of environmental challenges.

Keywords: Geospatial data, water, remote sensing, sustainability, machine learning, and accuracy.

1. Introduction

Advances in geospatial data analysis have recently been crucial for tackling environmental issues and advancing sustainability. How we track, examine, and manage our natural resources has changed dramatically due to the integration of remote sensing, artificial intelligence (AI), and geographic information systems (GIS). This essay examines how these technologies significantly advance environmental sustainability by facilitating resource management and knowledge-based decision-making [1]. Data collected from a distance is known as remote sensing, and it usually uses satellites, aeroplanes, or drones. Monitoring numerous environmental factors, including land cover, vegetation health, and climate change, has shown the great value of this technique. A thorough evaluation of ecosystems is made possible by high-resolution satellite photography, which also helps policymakers and researchers spot problem regions and follow environmental changes over time [2]. For instance, satellite imaging has proven invaluable in tracking natural disasters, urbanisation, and deforestation.

Artificial Intelligence has become a potent instrument in geospatial data analysis, augmenting the efficacy and Precision of environmental surveillance [3]. Computers can recognise patterns and make predictions using machine learning algorithms, a subset of artificial intelligence, and big datasets. Artificial Intelligence (AI) has applications in geospatial analysis, such as species identification, land cover classification, and environmental trend prediction. Machine learning models, for example, can use satellite imagery to identify illicit forestry activities or forecast the spread of wildfires, enabling resource allocation and prompt action. GIS, a geospatial technology, makes spatial data production, analysis, and visualisation possible [4]. It offers a framework for combining different data sources into an extensive spatial database, including field surveys, satellite photography, and demographic information. Decision-making processes are aided by the mapping and analysis of environmental data made possible by GIS. GIS can be used, for instance, to determine the best sites for renewable energy projects and to plan urban growth to minimise [5].

The combination of GIS, AI, and remote sensing produces a potent framework for handling challenging environmental problems [6]. While remote sensing offers real-time and historical data for input, artificial intelligence (AI) algorithms can be linked to GIS platforms to automate the study of massive datasets. An approach to environmental monitoring and management that is more dynamic and responsive is made possible by this integration [7]. For example, satellite imaging, machine learning, and geographic information systems (GIS) can be used in concert to track and forecast how climate change would affect biodiversity, which will aid conservation efforts. Although geographic data analysis advances have shown significant potential, there are still obstacles to overcome. It is necessary to address issues like data privacy, standardised protocol requirements, and technology accessibility in developing nations [8].

Irrespective of the context and the motive behind the instances above of environmental modelling and mappings, the studies are inevitably and understandably centred on the theme of utilization of geo-referenced data about the topology and the associated landscape. These factors are expressed as inventory and the characteristic traits of the locations in the inventory. In the case of natural resource potentiality mappings, the inventory comprises of locations that have in the past shown evidence to yielding the resource under investigation, while the characteristics for the topology would give insight into the reasons that caused or revealed that availability and extractability of the same resource [9]. For instance, in case of water resource mappings, the rainfall patterns in the region can hold clues to the water table levels in the region. Also, the soil texture in the region owing to their ability to decipher the region's porosity and water holding capacity can also explain the historical presence of water in the past inventory identified in the region [10].

Similarly, in case of wildlife habitat suitability mappings, the distance from dense human habitation and settlements could hold vital clues to the past inventory of wildlife sightings in the past for the region under investigation. On the other hand, the natural hazard susceptibility modeling, establishes the mapping between the inventory of past and historic locations in the region that witnessed the hazard event and the influencing or the conditioning factors that probably lead to or caused or prevented the hazard in its final unfolding. For example, in case of wildfire susceptibility zonations, the surface temperature, vegetation density and distance from settlements, roads could help decipher active wildfires in the region.

An automation of such mappings require mapping between the locations in the inventory and preexisting conditions on ground. The inventory in general, comprises of locations that had evidence of historic presence of resources (coal, natural gas, petroleum, water, etc.) or evidence of historic occurrence of hazard/disaster (such as floods, landslides, wildfires, etc.). Since the inventory and the associated factors are very much locatable in the space of earth's surface, these mappings have a spacial aspect inherent in them. Any analysis that considers the space to gain an overall perspective into cognizance is referred to as spatial analysis. The connotations and applications of the field of spatial analytics are varied such as: two dimensional plane of paper (x-ray, MRI etc), to the higher dimensional human body scans (brain imaging, ultrasounds, etc.) In the present context of environmental modelling, the space refers to the earth's surface, leading to the quintessential geospatial analysis [11].

Traditionally, the aforementioned studies analyzing the ground data of inventory and associated details required huge time and capital investments since the data needed to be identified, collected and gathered by means of laborious, time consuming and expensive field surveys. Also since these were manual tasks they would require long stretches of time spent in just data collection and hence, would also inevitably be prone to human errors [12]. Additionally, in the case of natural hazard mappings, the ground work was not just tricky but would sometimes prove to be unachievable since the areas under investigation would be under tremendous stress in the aftermath of disasters and associated tragedies (failed communication and connectivity) and could even be inaccessible thus preventing any field surveys. Mapping natural resources such as oil and natural gas can also be challenging due to their location in deep reservoirs under the earth's surface, making them inaccessible. However, the science of remote sensing has come to the rescue, as it allows capturing and accessing information without the need for physical presence at the region under investigation through satellites.

2. Background and Dataset

This section explains the recent research in geospatial data analysis, research gap identified from that analysis, and dataset details.

Related Works

The delineation of potential water zones is crucial for managing freshwater resources, especially in arid and semi-arid regions. A study focused on Saveh City in Iran utilized hybrid deep learning and machine learning algorithms, including boosted tree (BT), artificial neural network (ANN), and deep learning tree (DLT). Fourteen water potential conditioning factors were considered, and the models were validated using statistical analyses. The results demonstrated the effectiveness of the models, with the deep boost (DB) model exhibiting superior performance. Altitude, rainfall, distance to fault, and soil types emerged as critical factors for water potential modeling. The study recommends the use of the Deep Boost model for its superior results in water potential mapping [13].

The increased demand for freshwater resources due to population growth necessitates optimal management of water storage. A study in the Gandheswari River Basin in India employed Analytical Hierarchy Process (AHP), Frequency Ratio (FR), and machine learning techniques (Random Forest and Naïve Bayes) to delineate water potential zones. Twelve influencing factors were considered, and the models accurately identified five water potential zones. Geomorphology, slope, rainfall, and elevation were identified as crucial factors influencing water potential. The study's results, evaluated using various metrics, highlight the effectiveness of machine learning techniques coupled with AHP and FR for delineating regional water potential areas [14].

In regions like Quetta Valley in Pakistan, where water potential zoning is lacking, this study used six machine learning algorithms to evaluate 16 water drive factors. The models, including artificial neural networks (ANN), random forest (RF), and extreme gradient boosting (XGBoost), were trained and validated, with XGBoost, RF, and ANN showing high accuracy. The study emphasizes the importance of water potential for sustainable growth and recommends machine learning algorithms for accurate water productivity potential (GWPP) mapping. The results contribute to water management by providing a basis for locating new water wells in rocky terrain [15].

Water level forecasting is vital for effective water resource planning and management. A comprehensive review article spanning from 2008 to 2020 discusses the state-of-the-art machine learning models employed for water level (GWL) modeling. The review covers various ML models, data spans, input/output parameters, and performance criteria. It identifies milestones achieved in GWL prediction and provides recommendations for future research directions to enhance accuracy. The article serves as a valuable resource for scientists and practitioners in hydrology and water resource management [16].

Quantifying water recharge is essential for sustainable water resource management. A study on creating a high-resolution dataset for all of Europe utilized machine learning, including the Random Forest regressor, to map water recharge coefficients. The approach involved merging satellite-derived actual evapotranspiration and a machine learning model for estimating recharge coefficients. The resulting water recharge map provides harmonized high-resolution estimates across Europe, offering valuable information for sustainable water management. The study's dataset is available through the EuroGeoSurveys' open access European Geological Data Infrastructure (EGDI) [17].

Water potential (GWP) identification is crucial for sustainable development in arid oasis areas. A study applied ten algorithms, including shallow and hybrid models, to map GWP. The models, such as multilayer perceptron, decision tree, and gradient boosting, were evaluated based on sensitivity, specificity, and receiver operating characteristic (ROC)-area under curve (AUC) analysis. The study highlights the significance of the applied algorithms in decision support for sustainable development in oasis areas [18].

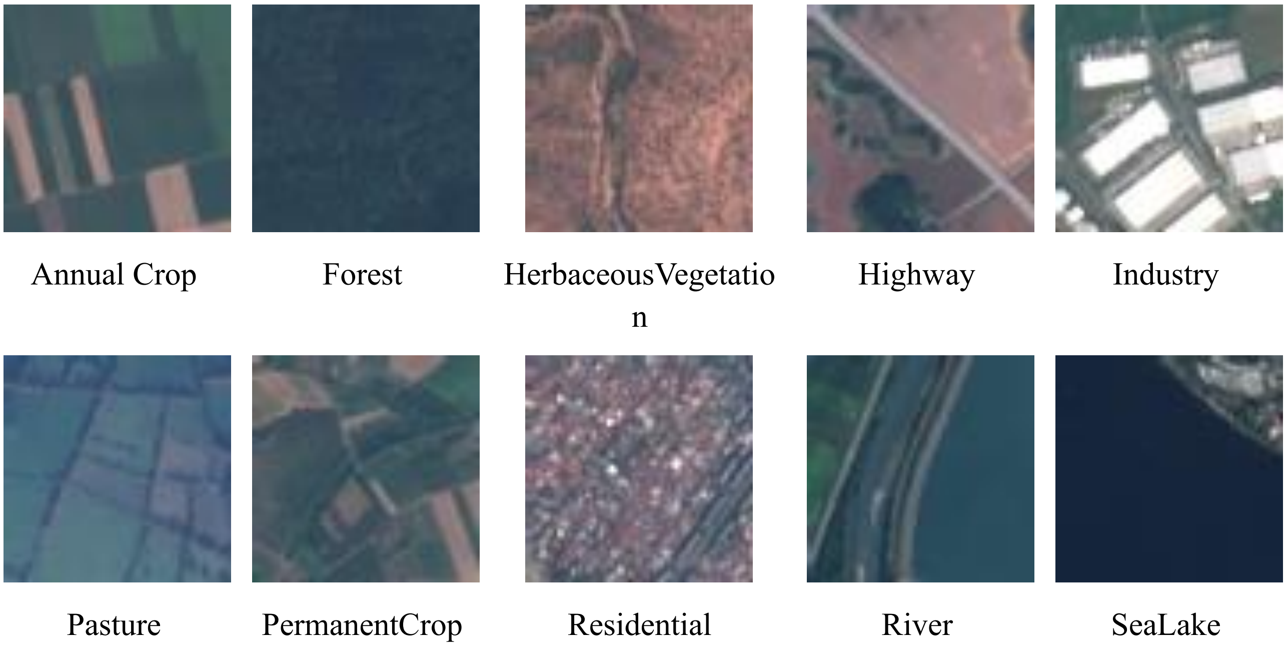
These studies collectively emphasize the importance of advanced geospatial and machine learning techniques in understanding, managing, and sustaining water resources in the face of climate change and increasing water demand. Leveraging these innovative approaches is crucial for informed decision-making and effective water resource management in diverse geographical contexts.

Research Gap

While significant progress has been made in the areas of machine learning, mathematical modelling, and geospatial data analysis for water level prediction, a number of research gaps still exist. Among these are the following: a lack of cross-disciplinary collaboration; a lack of rigorous validation processes; an inadequate consideration of spatial and temporal dynamics on a larger scale; an absence of standardised methodologies for model evaluation; a research gap in the exploration of diverse ensemble methods; an inadequate attention to model explainability and uncertainty; and a need for real-world application of developed models. Closing these gaps will improve water level prediction models' practicality, accuracy, and dependability, which will ultimately lead to more sustainable and successful water resource management in the face of environmental change.

Dataset Details

The EuroSat dataset is a widely used benchmark in the field of remote sensing and machine learning, specifically designed for land use and land cover classification using satellite imagery. Comprising 27,000 labeled Sentinel-2 satellite images, the dataset covers 13 different land use and land cover classes across Europe. These classes encompass diverse landscapes, including urban areas, farmland, forests, and water bodies. Researchers and data scientists leverage the EuroSat dataset to train and evaluate machine learning models, aiming to develop algorithms capable of automatically discerning and categorizing various land cover types from high-resolution satellite imagery. The dataset plays a crucial role in advancing the capabilities of algorithms for accurate land cover classification, which is essential for applications such as environmental monitoring, urban planning, and resource management [19]. The sample input image is given in Figure 1.



Annual Crop

Forest

HerbaceousVegetation

Highway

Industry

Pasture

PermanentCrop

Residential

River

SeaLake

Figure 1. Images used in Water Mapping

3. Methodology

Combining the advantages of several algorithms, including Principal Component Analysis (PCA), Genetic Algorithm (GA), Gradient Boosted Decision Trees (GBDT), Frequency Ratio (FR), Deep Neural Network (DNN) and Shannon's Entropy (SE), is necessary to map water prospective areas using a data-driven ensemble model. In the context of water mapping for environmental sustainability, a comprehensive ensemble model is devised by amalgamating the capabilities of Principal Component Analysis (PCA), Genetic Algorithm (GA), Gradient Boosted Decision Trees (GBDT), Frequency Ratio (FR), Deep Neural Network (DNN), and Shannon's Entropy (SE). This ensemble approach aims to harness the unique strengths of each algorithm to provide a more accurate and robust prediction of water potential. Shannon's Entropy is employed to gauge the uniformity of water potential across influencing factors, and frequency ratios are calculated for each class within these factors. These ratios, emphasizing the correlation between water locations and influencing factor values, contribute to the generation of a water potential map.

Deep Neural Network (DNN) forms a complex mathematical model, capturing intricate patterns in the data and enhancing the ensemble's predictive capabilities. Principal Component Analysis (PCA) transforms the dataset into uncorrelated eigenvectors, reducing dimensionality and aiding in efficient data handling. Gradient Boosted Decision Trees (GBDT) sequentially combines weak learners to construct a robust predictive model that captures intricate relationships within the dataset. Genetic Algorithm (GA) optimizes attribute weights obtained through PCA, ensuring an effective training process. By integrating these algorithms, the ensemble model presents a holistic and effective solution for mapping water prospective areas, contributing valuable insights to sustainable water resource management. The ensemble model-based water mapping is given as,

Shannon's Entropy (SE)

Stephan Boltzmann presented the idea of an entropy index to associate with the thermodynamic position of a model. It averages the gaps between the proportions for a sub-class to that of the whole state and hence used like a measure for the quantification of uniformity, for an evenly and equivalent distribution of classes into divisions. In the context of this case study, the associated theory's concepts are being applied to identify the impact of each influencing factor on the location water potential. Firstly, the frequency ratios were computed for all classes in each of the influencing factor's domains as defined in Equation 1. These ratios establish the individual and independent probability distribution for each of the influencing factors with the underlying assumption of the influencing factors having negligible cross effect on each other significance of water potential.

$$frequency\ ratio(\aleph) = \left(\frac{{}^{p}/_{q}}{{}^{u}/_{v}}\right)\text{---------(1)}$$

For a particular influencing factor, refers to the number of water locations or positive labeled records in the specifically chosen class. while q represents the total water locations (positive class records). On the denominator, the quantity u refers to the area coverage of the specifically chosen class for the same influencing factor in unit of pixels, while refers to the area coverage of the entire study area also expressed in number of pixels. The formula thus delivers the frequency ratio values for the specifically chosen sub-class of the same attribute. For all of the considered influencing factors, the information coefficient that represents their respective significance as a numeric quantity has been calculated using the set of equations from Equation 2 to 7.

$$\varphi_{ij} = \frac{\aleph}{\sum_{j=1}^{Mj}\aleph}\text{---------(2)}$$

$$\varepsilon_j = -\sum_{i=1}^{Mj}\varphi_{ij}log_2\varphi_{ij}, j = 1, \dots, n\text{---------(3)}$$

$$\varepsilon_{j\ max} = log_2M_j\text{---------(4)}$$

$$\alpha_j = \left(\frac{\varepsilon_{j\ max} - \varepsilon_j}{\varepsilon_{j\ max}}\right)\text{---------(5)}$$

$$\omega_j = \alpha_j \aleph_{avg}\text{---------(6)}$$

$$\aleph_{avg} = \frac{1}{M_j}\sum_{j=1}^{M_j} \aleph\text{---------(7)}$$

For a given influencing factor the number of classes the domain divides into is represented by m_j, with the probability density being referred to by φ_{ij} . Additionally, the entropy values are represented by ε_j and $\varepsilon_{j\ max}$ and the information coefficient is represented by α_j , while the mean of FR values of all subclasses within an attribute is referred to as $\aleph_{avg}$. Ultimately, the quantitative contribution of each factor on the resultant potential is referred to by ω_j .

Frequency Ratio (FR)

In order to calculate the frequency ratios, the attributes are first divided into classes and for each class the frequency (number of) of the positive data/locations with class water are computed. The probability of occurrence of the water the class thus becomes the frequency ration of the attribute in that class. Thus, frequency ratio emphasizes the correlation that exists between the point location with class\label of water and the value of the water influencing factors for the same points. The frequency ratio for the cth class (c_{min}, c_{max}) of the i^th influencing factor (F_i), where ($1 \le i \le N$ (N is the total count of factors)) is estimated as follows in Equation 8.

$$FR_c^i = \frac{\left(\frac{No.of\ GW\ pixels\ with\ value\ f(F_i)\epsilon c\ i,e,(c_{min} \le f(F_i) \le c_{max}}{total\ number\ of\ pixcels\ in\ the\ study\ area}\right)}{\left(\frac{total\ numbefr\ of\ pixels\ with\ value\ f(F_i)\in c\ i,e,(c_{min} \le f(F_i) \le c_{max}}{total\ number\ of\ pixels\ in\ the\ study\ area}\right)}\text{---------(8)}$$

To generate the regions water potential map the frequency ratios for all the influencing factors are summed as described in Equation 9.

$$GW\ potential_{FR} = \sum_{i=1}^{N} FR^i\text{---------(9)}$$

Deep Neural Network (DNN)

A deep neural network is a complex mathematical model composed of layers of interconnected nodes. For a simple fully connected layer, it is given by:

$$a_i = \sigma\left(\sum_{j=1}^{N} w_{ij} \cdot x_i + b_i\right)\text{---------(10)}$$

where a_i is the output of node i, w_{ij} is the weight connecting node j to node i, x_i is the input from node j, b_i is the bias of node i and σ is the activation function.

Principal Component Analysis (PCA)

An un-correlated feature set of eigenvectors also called principal components or eigenvectors is obtained through an orthogonal projective transformation of the original dataset. This reduces the size of the dataset with the transformed un-correlated dataset having fewer dimensions than the original. This is achieved by having the variance among original features be captured maximally in the newer feature set, with the first eigenvector having the highest variance followed by the second eigenvector which has the second highest variance and so on and so forth. Also, the eigenvectors are orthogonal, leading to them being un-correlated. Graphically, it can visualized as the original dataset being orthogonally projected from a higher dimension space to a lower dimension space, with the aim to maximally capture the variance in the data. The dataset is described as

$$\left\{\left(X^1 = ([x_1^1, x_2^1, x_3^1, \dots\dots\dots, x_N^1], y^1)\right), \left(X^2 = ([x_1^2, x_2^2, x_3^2, \dots\dots\dots, x_N^2], y^2), \dots\dots, \left(X^M = ([x_1^M, x_2^M, x_3^M, \dots\dots\dots, x_N^M], y^M)\right)\right)\right\} \text{---------(11)}$$

where N being the count of attributes and the count of records are represented by M. W={w$_1$,w$_2$,w$_3$,…..,w$_N$} denotes the principal component for all N attributes. The projected dataset forecast is represented by {(Z^1=W^{T}x^1,y^1), (Z^2=W^{T}x^2,y^2),……, (Z^i=W^{T}x^i,y^i),……, (Z^M=W^{T}x^M,y^M)} where W^T indicates the Ws transpose. The datasets mean is transformed to $\bar{z} = \frac{1}{M}\sum_{i=1}^{M} z^i$ which is equivalent to $z = \frac{1}{M} W^T \sum_{i=1}^{M} X^i$ after the orthogonal projection. This can be written as $\bar{z} = \frac{1}{M} W^T \bar{X}$. Hence the mean of the probable data is $\bar{z} = \frac{1}{M} W^T \bar{X}$ and the variance of the projected data is $v = \frac{1}{M}\sum_{i=1}^{M}\left\{W^T x^i - W^T \bar{x}\right\}^2$. Through singular value decomposition we get Equation 12.

$$v = \frac{1}{M}\sum_{i=1}^{M}\left\{W^T x^i - W^T \bar{x}\right\}^2 = W^T S W \text{---------(12)}$$

where S is the covariance matrix determined as in Equation 13.

$$S = \frac{1}{M}\sum_{i=1}^{M}\left(X^i - \bar{X}\right)\left(X^i - \bar{X}\right)^T \text{---------(13)}$$

Maximizing the variance among the projected data, denoted as $W^T S$, in relation to W is the objective. To prevent the potential issue of W →∝ resulting from this maximization, the normalization condition $W^T W = 1$ is introduced. This normalization condition facilitates an unconstrained maximization, as indicated in Equation 14 (with λ serving as the Lagrange multiplier).

$$\max_{W}\{W^T S W + \lambda(1 - W^T W)\} \text{---------(14)}$$

Upon obtaining the expression in Equation 14, the derivation with respect to W and subsequent setting to zero leads to SW=λW. This signifies that W acts as an eigenvector for the covariance S, with λ representing the corresponding eigenvalue.

Gradient Boosted Decision Trees (GBDT)

Gradient Boosted Decision Trees (GBDT) is an ensemble learning algorithm that builds a strong predictive model by sequentially combining weak learners, typically shallow decision trees. The algorithm minimizes a specified loss function by adjusting the model predictions in the direction opposite to the gradient of the loss. The training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ consists of feature vectors xi and corresponding target variables yi. GBDT employs M weak learners hm(x) in the ensemble, with m as the iteration index. The overall model prediction is a weighted sum of the weak learners in Equation 15.

$$F_M(x) = \sum_{m=1}^{M} \alpha_m h_m\,(x) \text{ ---------(15)}$$

The training process starts with an initial model prediction F0(x), often set as the mean of the target variable. For each iteration m, the algorithm fits a weak learner hm(x) to the negative gradient of the loss function in Equation 16.

$$h_m(x) = \arg\min_h \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + h(x_i)\text{---------}(16)$$

The weight αm for the weak learner is determined based on a line search or other optimization techniques. The overall model prediction is updated as follows in Equation 17.

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x) \text{ ---------}(17)$$

GBDT minimizes a specified loss function, typically Mean Squared Error (MSE) in Equation 18.

$$L\big(y_i, F(x_i)\big) = [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]\text{---------}(18)$$

where p_i is the predicted probability.

Regularization techniques include shrinkage (v) to control the contribution of each weak learner in Equation 19.

$$F_m(x) = F_{m-1}(x) + v\alpha_m h_m(x) \text{ ---------}(19)$$

The final model prediction after M iterations is given in Equation 20.

$$F_M(x) = \sum_{m=1}^{M} v\alpha_m h_m(x)\text{---------}(20)$$

Feature importance is assessed based on the gain or improvement in the loss function brought about by each feature during training. Regularization techniques and hyperparameter tuning play crucial roles in achieving optimal performance.

Genetic Algorithm (GA)

The initial weight assignment for the attributes on the basis of train data is done through principal component analysis. For training the model, the dataset is trimmed as per the weights. So for the 1st round of training, only those attributes are retained for training the model whose assigned weights were greater than 0.5. Hence the chromosome signifying this weighing is generated with 0s and 1s, where 1 represents the inclusion of the respective attribute for training as the associated weight generated in the specific iteration was greater than the threshold limit of 0.5. Hence, the model is trained with the weighted chromosomes. The gradient boosted decision tree model was trained using a 10-fold crossvalidation approach, and the accuracy performance metric was utilized to define the fitness of the chromosome. At the conclusion of each iteration, termination conditions were examined. The training was stopped if either the objective function converged or a fixed number of iterations were completed. The training is identified to have converged if a majority of chromosome in the population has reached similar levels of fitness indicating deduction of an optimal set of features thus justifying halting the iterations.

Since convergence criteria may never be met, an upper limit on the number of iterations is placed such as 50 iterations. However, if the convergence does happen before the completion of the requisite number of iterations, the training is preemptively halted. If neither the convergence was achieved nor the number of iteration exceeded the upper limit, then in that case, the next chromosome population is generated via the processes of mutation, cross over and tournament selection over the previous populations. For this purpose, a subset population that is one-fourth the size of the original population is created from the initial population. The chromosomes within this subset population are compared to each other through various fitness comparisons tournaments, and the victor of each tournament is chosen. Selection and de-selection of the certain features randomly is referred to as mutation while an interchange among the features from 2 chromosomes is called crossover. Its through three procedures that a new population of chromosomes is generated. The termination criteria are checked after every iteration for the training to be halted.

4. Result and Discussion

The performance evaluation of the water mapping ensemble model involves assessing its predictive accuracy and effectiveness through a set of key metrics. Accuracy, Precision, and recall are fundamental performance metrics used to assess the effectiveness of predictive models across various domains. In the context of water mapping, accuracy, precision, and recall play pivotal roles in assessing the effectiveness of predictive models. Accuracy serves as a comprehensive metric, quantifying the correctness of predictions by determining the ratio of accurately predicted instances to the total instances. While accuracy provides a holistic view of a model's performance, its reliability diminishes in imbalanced datasets, where one class dominates the distribution.

Precision, however, takes a more focused approach, concentrating on the accuracy of positive predictions. For water mapping, precision is critical in situations where false positives, indicating the presence of water potential, can have significant consequences, influencing decisions related to resource management or environmental planning. Recall, or the true positive rate, gauges the model's capacity to identify all relevant instances of water potential. Calculated as the ratio of true positives to the sum of true positives and false negatives, recall becomes crucial in scenarios where missing positive instances, such as accurately identifying areas with water potential, bears a high cost. This is particularly important in applications like water resource management and environmental sustainability, where accurate identification of potential water areas is paramount for effective decision-making.

Table 1. Performance Comparison of Accuracy, Precision, and Recall

	Algorithm	100	200	300	400	500
Accuracy	ANN	96.5	97	97.5	97.8	98
	RF	95.2	96	96.5	97	97.5
	XGBoost	93.8	94.5	95	95.5	96
	EnsembleML	97	97.5	98.1	98.3	98.5
Precision	ANN	82	83.5	84	85	86
	RF	85.5	86	87	88	89
	XGBoost	88	88.5	89	89.5	90
	EnsembleML	84	86	89	90	90.4
Recall	ANN	88	88.5	89	89.5	90
	RF	87.2	87.5	88	88.5	89
	XGBoost	89	89.2	89.5	89.8	90

	EnsembleML	89	89.7	90.98	91.21	92

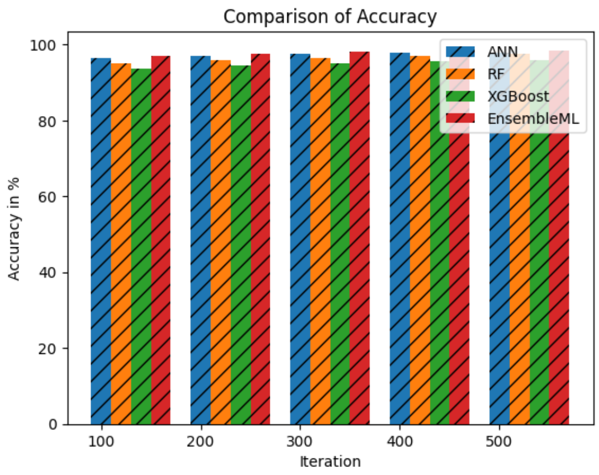


Figure 2. Comparison of Accuracy

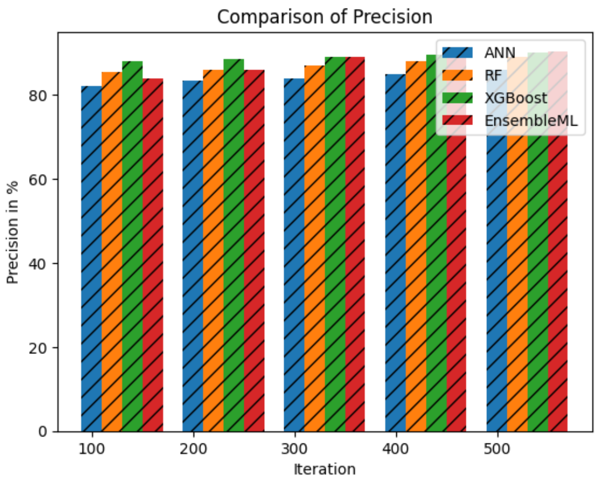


Figure 3. Comparison of Precision

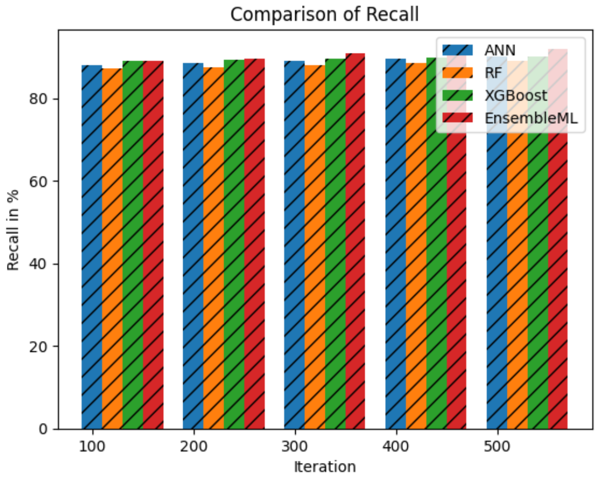


Figure 4. Comparison of Recall

The performance metrics for Accuracy, Precision, and Recall across different algorithms—Artificial Neural Network (ANN), Random Forest (RF), XGBoost, and EnsembleML—reveal insightful trends. The ANN displays a consistent increase in accuracy, ranging from 96.5% to 98%, showcasing its robust learning capabilities. Random Forest demonstrates steady accuracy improvement, reaching 97.5%, while XGBoost exhibits similar positive trends, achieving up to 96% accuracy. Notably, EnsembleML outperforms its counterparts, achieving the highest accuracy at 98.5%, underscoring the efficacy of ensemble methods. Precision, reflecting the ability to identify positive instances correctly, shows improvement across thresholds for all algorithms. EnsembleML excels in Precision, reaching 90.4%, reinforcing the advantage of combining diverse models. In terms of Recall, the ANN, RF, and XGBoost exhibit incremental improvements, with EnsembleML consistently leading and achieving a peak Recall of 92%. In summary, EnsembleML emerges as the top-performing algorithm, showcasing superior Accuracy, Precision, and Recall, reaffirming the effectiveness of ensemble approaches in comprehensive and high-performance model development.

5. Conclusion

This research work presents an advanced ensemble machine learning (ML) approach, EnsembleML, for water mapping, showcasing its remarkable accuracy, Precision, and recall rates of 98.5%, 90.4%, and 92%, respectively. The research addresses a significant research gap by combining diverse ML algorithms, providing a comprehensive and effective solution for predicting water levels. The advantage of EnsembleML over individual algorithms underscores the efficacy of ensemble techniques in handling the complex relationships within geospatial datasets. Future enhancements could involve refining the ensemble model by incorporating additional influencing factors, exploring new ensemble methods, and incorporating real-time data for improved predictions. Cross-disciplinary collaborations and standardized methodologies should be prioritized to further enhance the practicality and reliability of water level prediction models. Additionally, focusing on model explainability and uncertainty analysis will contribute to these models' broader acceptance and application in real-world scenarios, ensuring sustainable water resource management in the face of evolving environmental dynamics.

Reference

[1] Bai, Z., Liu, Q., & Liu, Y. (2022). Water potential mapping in hubei region of china using machine learning, ensemble learning, deep learning and automl methods. *Natural Resources Research*, *31*(5), 2549-2569.

[2] Hasanuzzaman, M., Mandal, M. H., Hasnine, M., & Shit, P. K. (2022). Water potential mapping using multi-criteria decision, bivariate statistic and machine learning algorithms: evidence from Chota Nagpur Plateau, India. *Applied Water Science*, *12*(4), 58.

[3] Yariyan, P., Avand, M., Omidvar, E., Pham, Q. B., Linh, N. T. T., & Tiefenbacher, J. P. (2022). Optimization of statistical and machine learning hybrid models for water potential mapping. *Geocarto International*, *37*(13), 3877-3911.

[4] Gómez-Escalonilla, V., Martínez-Santos, P., & Martín-Loeches, M. (2022). Preprocessing approaches in machine-learning-based water potential mapping: an application to the Koulikoro and Bamako regions, Mali. *Hydrology and Earth System Sciences*, *26*(2), 221-243.

[5] Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Kim, S., ... & Yaseen, Z. M. (2022). Water level prediction using machine learning models: A comprehensive review. *Neurocomputing*, *489*, 271-308.

[6] Liu, R., Li, G., Wei, L., Xu, Y., Gou, X., Luo, S., & Yang, X. (2022). Spatial prediction of water potentiality using machine learning methods with Grey Wolf and Sparrow Search Algorithms. *Journal of Hydrology*, *610*, 127977.

[7] Bertalan, L., Holb, I., Pataki, A., Négyesi, G., Szabó, G., Szalóki, A. K., & Szabó, S. (2022). UAV-based multispectral and thermal cameras to predict soil water content–A machine learning approach. *Computers and Electronics in Agriculture*, *200*, 107262.

[8] Podgorski, J., Araya, D., & Berg, M. (2022). Geogenic manganese and iron in water of Southeast Asia and Bangladesh–machine learning spatial prediction modeling and comparison with arsenic. *Science of The Total Environment*, *833*, 155131.

[9] Eid, M. H., Elbagory, M., Tamma, A. A., Gad, M., Elsayed, S., Hussein, H., ... & Péter, S. (2023). Evaluation of water quality for irrigation in deep aquifers using multiple graphical and indexing approaches supported with machine learning models and GIS techniques, Souf Valley, Algeria. *Water*, *15*(1), 182.

[10] Elzain, H. E., Chung, S. Y., Senapathi, V., Sekar, S., Lee, S. Y., Roy, P. D., ... & Sabarathinam, C. (2022). Comparative study of machine learning models for evaluating water vulnerability to nitrate contamination. *Ecotoxicology and Environmental Safety*, *229*, 113061.

[11] Wang, D., Qian, J., Ma, L., Zhao, W., Gao, D., Hou, X., & Ma, H. (2022). Characterizing water distribution potential using GIS-based machine learning model in Chihe River basin, China. *Environmental Earth Sciences*, *81*(12), 324.

[12] Kumar, S., & Pati, J. (2022). Assessment of water arsenic contamination level in Jharkhand, India using machine learning. *Journal of Computational Science*, *63*, 101779.

[13] Chen, Y., Chen, W., Chandra Pal, S., Saha, A., Chowdhuri, I., Adeli, B., ... & Mosavi, A. (2022). Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting water potential. *Geocarto International*, *37*(19), 5564-5584.

[14] Hasanuzzaman, M., Mandal, M. H., Hasnine, M., & Shit, P. K. (2022). Water potential mapping using multi-criteria decision, bivariate statistic and machine learning algorithms: evidence from Chota Nagpur Plateau, India. *Applied Water Science*, *12*(4), 58.

[15] Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Kim, S., ... & Yaseen, Z. M. (2022). Water level prediction using machine learning models: A comprehensive review. *Neurocomputing*, *489*, 271-308.

[16] Martinsen, G., Bessiere, H., Caballero, Y., Koch, J., Collados-Lara, A. J., Mansour, M., ... & Stisen, S. (2022). Developing a pan-European high-resolution water recharge map–Combining satellite data and national survey data using machine learning. *Science of the Total Environment*, *822*, 153464.

[17] Ouali, L., Kabiri, L., Namous, M., Hssaisoune, M., Abdelrahman, K., Fnais, M. S., ... & Bouchaou, L. (2023). Spatial Prediction of Water Withdrawal Potential Using Shallow, Hybrid, and Deep Learning Algorithms in the Toudgha Oasis, Southeast Morocco. *Sustainability*, *15*(5), 3874.

[18] Madani, A., & Niyazi, B. (2023). Water Potential Mapping Using Remote Sensing and Random Forest Machine Learning Model: A Case Study from Lower Part of Wadi Yalamlam, Western Saudi Arabia. *Sustainability*, *15*(3), 2772.

[19] https://www.kaggle.com/datasets/apollo2506/eurosat-dataset