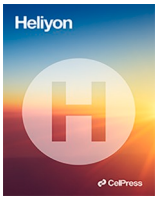
Heliyon 10 (2024) e28527

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Application of machine learning techniques to predict groundwater quality in the Nabogo Basin, Northern Ghana

Joseph Nzotiyine Apogba [a], Geophrey Kwame Anornu [a], Arthur B. Koon [a,b], Benjamin Wullobayi Dekongmen [c,d], Emmanuel Daanoba Sunkari [e,f], Obed Fiifi Fynn [g], Prosper Kpiebaya [h,i,*]

[a] *Civil Engineering Department-Regional Water and Environmental Sanitation Centre Kumasi, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*
[b] *Department of Geology, College of Engineering, University of Liberia, Fendall Campus, Liberia*
[c] *Department of Agricultural Engineering, Ho Technical University, Ho, Ghana*
[d] *Department of Civil and Environmental Engineering, University of Energy and Natural Resources, Sunyani, Ghana*
[e] *Department of Geology, Faculty of Science, University of Johannesburg, Auckland Park 2006, Kingsway Campus, P. O. Box 524, Johannesburg, South Africa*
[f] *Department of Geological Engineering, Faculty of Geosciences and Environmental Studies, University of Mines and Technology, P.O Box 237, Tarkwa, Ghana*
[g] *Water Research Institute – Council for Scientific and Industrial Research, Ghana*
[h] *Department of Agricultural Engineering, School of Engineering, University for Development Studies, P. O. Box TL 1882, Ghana*
[i] *Department of Soil Science, Faculty of Agriculture, Food and Consumer Sciences, University for Development Studies, P. O. Box TL 1882, Ghana*

ARTICLE INFO

Keywords:
Machine learning
Groundwater quality
Nabogo basin
White Volta Bain

ABSTRACT

The main objective of this study was to map the quality of groundwater for domestic use in the Nabogo Basin, a sub-catchment of the White Volta Basin in Ghana, by applying machine learning techniques. The study was conducted by applying the Random Forest (RF) machine learning algorithm to predict groundwater quality, by utilizing factors that influence groundwater occurrence and quality such as Elevation, Topographical Wetness Index (TWI), Slope length (LS), Lithology, Soil type, Normalize Different Vegetation Index (NDVI), Rainfall, Aspect, Slope, Plan Curvature (PLC), Profile Curvature (PRC), Lineament density, Distance to faults, and Drainage density. The groundwater quality of the area was predicted by building a Random Forest model based on computed Arithmetic Water Quality Indices (WQI) (as dependent variable) of existing boreholes, to serve as an indicator of the groundwater quality. The predicted WQI of groundwater in the study area shows that it ranges from 9.51 to 69.99%. This implied that 21.97 %, 74.40 %, and 3.63 % of the study area had respectively the likelihood of excellent. The models were found to perform much better with an RMSE of 23.03 and an R^2 value of 0.82. The study conducted highlighted an essential understanding of the groundwater quality in the study area, paving the way for further studies and policy development for groundwater management.

* Corresponding author. Department of Agricultural Engineering, School of Engineering, University for Development Studies. P. O. Box TL 1882, Ghana.
E-mail address: Kpiebayaprosper@gmail.com (P. Kpiebaya).

https://doi.org/10.1016/j.heliyon.2024.e28527
Received 23 September 2023; Received in revised form 20 March 2024; Accepted 20 March 2024
Available online 30 March 2024

1. Introduction

Globally, many developing communities suffer from physical water scarcity, because water is a basic human need, and its insufficientness impedes development in many places, especially in areas with little or no access to surface water. Water resources have great economic potential for agriculture, tourism, irrigation, transport, and industry [1,2]. Although water surrounds most parts of the world, however, it is not always readily available, especially potable water, which is required to meet necessities including food production, sanitation, and long-term development [3,4]. It is worth noting that only 2.5% of the water on earth is available as surface water, with the rest 1.2% being located in groundwater, glaciers, and ice caps, respectively [5].

Surface water sources include rivers, lakes, canals, runoff, and reservoirs. It is important to understand that, surface water sources can change dramatically in a short amount of time [4,6], which makes them more susceptible to chemical spills and accidental releases. Groundwater quality, on the other hand, is typically constant over time; however, changes in groundwater conditions can lead to variations in water quality over short distances. Groundwater is an adequate source of water, with little to no treatment required [7]. However, unlike surface water, groundwater is not easily accessible. Finding areas with groundwater potential requires groundwater exploration. Groundwater exploration often requires geophysical investigations which require expensive equipment for assessment. Domestic water supply accounts for about 95 % of groundwater use in Ghana, largely in rural areas and small towns [8,9]. According to Kpakpo [10], the percentage of families in Ghana who rely on groundwater for their water supply is about 41 %, with the percentage being much greater in rural areas (59%) than in urban areas (16%). However, in some parts of the Upper East and Upper West Regions of Ghana, groundwater is the main supply of water for 80% of the urban population. Less than 5 % of Ghana's groundwater is used for irrigating and watering livestock and poultry, and less than 1% of Ghana's total groundwater use due to industrial uses of groundwater (Hydrogeology of Ghana - MediaWiki, 2022).

Groundwater quality is influenced by a wide range of factors ranging from natural to anthropogenic factors. To save cost in drilling a borehole only to abandon it, because of poor water quality it would be helpful if there exists a map that shows the suitability or otherwise of the groundwater for domestic consumption. Thus, it is not enough to know only where groundwater exists, it is equally important to know if the groundwater in the area is safe for consumption. Notably, severally conventional methods have been applied to assess water resources, groundwater vulnerability and quality around the globe [11–17]. In recent years, Water Quality Index (WQI) integrated with Machine Learning (ML) and Deep Learning, as many innovative techniques [18]. Deshpande et al. [19], Goodarzi et al.

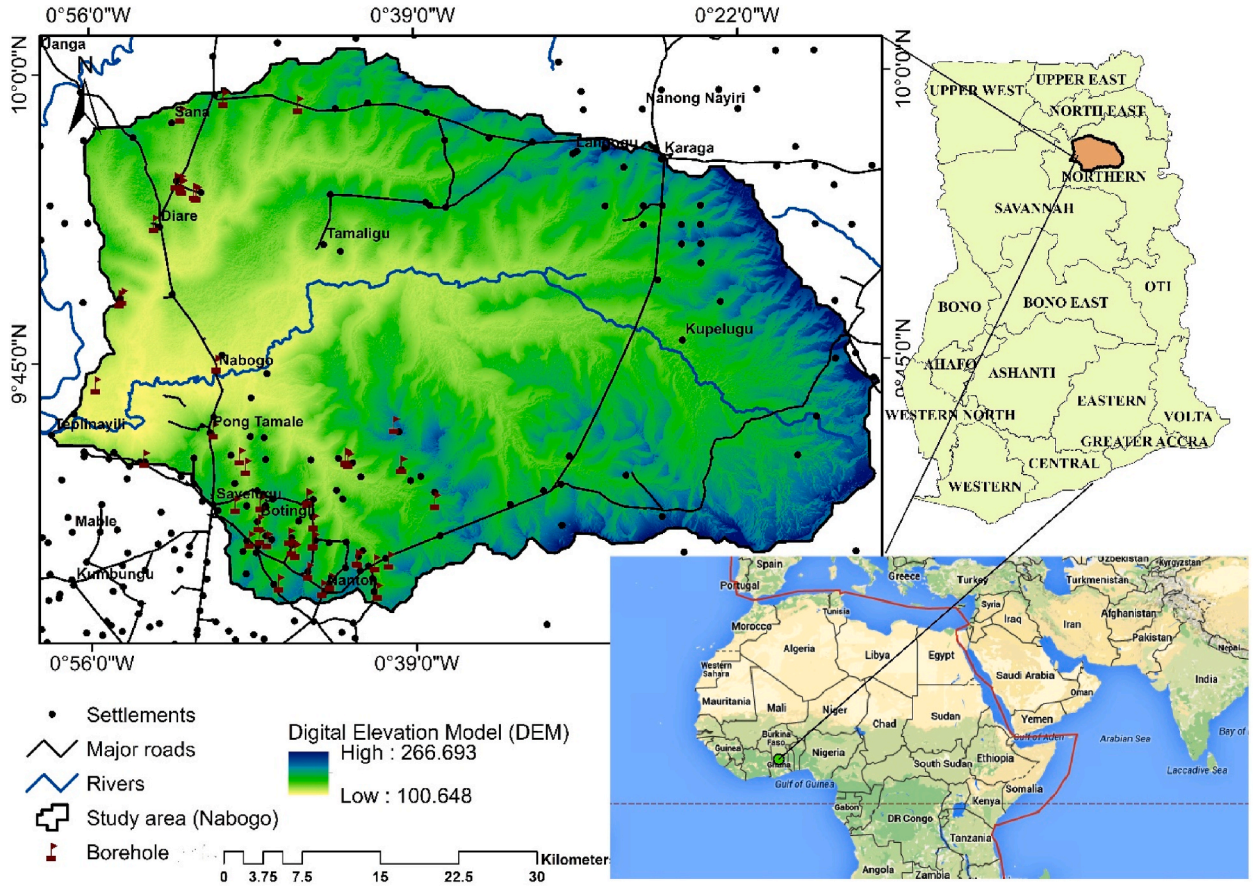


Fig. 1. Geographical location of the Study Area (Nabogo Basin).

[20], Patel et al. [21], Siriwardhana et al. [22],Abba et al. [23] and Xiong et al. [24] assessed water, and groundwater quality in their various study areas, where their findings from WQI methods proved significant for water quality management, and it recommended the replication of these methods in different territories of the world.

Therefore, the primary goal of this study was to investigate the capability of the Random Forest algorithm in RStudio, GIS, and remote sensing techniques for mapping groundwater quality in the Nabogo Basin in Ghana. Decision makers in groundwater development and management will benefit from the resulting groundwater potential map to find acceptable areas for water resource exploitation. A new and upcoming technique for groundwater prediction is the use of machine learning techniques. For groundwater mapping, machine learning algorithms like random forest have been utilized continually [25–30]. The current study aims to use machine learning to predict groundwater quality in Ghana's Nabogo Basin.

Traditional exploration of groundwater does not consider water quality but only its availability. Finding groundwater that is not of the right quality subsequently leads to either the water source being abandoned or expensive equipment being used to treat the water for domestic use. All these put a financial burden on water users. It would be prudent if one could decide beforehand whether to drill a borehole or not, based on knowledge of whether the water that will be obtained will be of good or not. Other works carried out within the study area have been to characterize groundwater resources concerning recharge, unsaturated zone process, and quality [31–35]. The processes employed by these studies are expensive and time-consuming for any interested person who wants to develop groundwater resources within the Nabogo basin. Machine learning approaches, which employ stochastic solutions to data to better constrain the processes of characterizing an area within a shorter period, are much more preferred.

Finding groundwater of the right quality for domestic use is as important as finding areas with the potential for extracting groundwater. However, the water quality of the Nabogo Basin has not yet been mapped. Most groundwater research in the study area has only mapped the availability of groundwater in the basin, without its quality. In a study by Ref. [36], which mapped groundwater availability in the Nabogo Basin, the authors recommended that further studies be carried out to map the groundwater quality in the study area. Therefore, the main objective of this study was to predict quality in the Nabogo Basin by applying machine learning techniques.

2. Study area

2.1. Location and climate

The Nabogo Basin (Fig. 1) falls under the White Volta Basin, under the Volta Basin in Ghana. The Nabogo basin encompasses the Savelugu Municipality, Nanton District, Karaga District, parts of the Gushegu Municipality, and a very small part of the Mion District in Ghana's Northern Region. The Nabogo River is formed by the convergence of several smaller tributaries before the town of Nabogo, within latitudes 9.535 N–10.022 N and longitudes 0.948 W–0.249 W. The area drained by the basin is approximately 2872 km^2. The basin's altitude above mean sea level varies from 98 to 271 m.

The basin has a single-mode rainfall system that begins in April and ends in October, with the peak months being August/September. The mean annual precipitation calculated from a 30-year historical dataset is 1094 mm. Mean minimum and maximum temperatures in the basin vary from 22.9 to 34.5 °C, respectively. Commonly, within a year, 87 % of the rain falls from May to October. However, most of the rains occurring during these months end up as surface runoff with little recharge to the groundwater [37]. The humidity levels in the rainy season vary from 70 to 81 % but fall to 28 % in the dry season. Annual evapotranspiration in the basin measured using the FAO Penman-Monteith method is 1898 mm/year. The hot temperatures within the basin cause surface waters to dry up, resulting in water scarcity and soil hardening, which might also limit infiltration and, as a result, reduce aquifer recharge.

A study by Dapaah-Siakwan & Gyau-Boakye [8], Dapaah-Siakwan & Gyau-Boakye [9] shows that Palaeozoic sedimentary rocks, locally known as Voltaian formation underlay the Nabogo Basin and consist mainly of unconsolidated mudstones, siltstones, shale, conglomerate, and sandstones. This formation has been reconsolidated over time, leading to the loss of its primary permeability. As a result, secondary porosity caused by weathering and rock deformation controls groundwater occurrence in the basin, resulting in a localized fractured-type aquifer system [38]. Aquifer properties vary across this basin due to the complex underlying geology. For example, borehole yield in the basin can range from approximately 700 l/min, while the depth to water level ranges from 2 to 30 m, with an average of 9 m from the ground level [36].

The surface water system are tributaries of the white and black volta. Borehole depths vary widely from 21.0 m to 99.0 m, with yields ranging from 0.3 m^3/h to 12.0 m^3/h. Anku et al. [39] uncovered a positive correlation between borehole depth and yield. Aquifer transmissivity among sandstones ranges from 0.1 to 52.0 m^2/d. Siltstone and mudstone aquifers have transmissivity ranging from 0.2 to 16.0 m^2/d [40,41]. In general, the Voltaian aquifers have a strong relationship between aquifer transmissivity and specific capacity [40,42]. The study area generally has groundwater suitable for domestic and agricultural use but some parameters make the quality poor [43]. This includes a low pH ranging from 3.5 to 6.0 and a high concentration of iron throughout the study area [41]. The majority of groundwater quality issues can be attributed to hydro-geochemical processes in bedrock rock aquifers and anthropogenic activities resulting in high sodium chloride concentrations.

3. Materials and methods

3.1. Laboratory analysis

Groundwater samples were obtained from One hundred and forty (140) boreholes following established sampling protocols. These

samples were collected in 100 ml acid-washed high-density linear polyethylene (HPDE) bottles. Subsequently, contaminants were removed from the samples using a Sartorius polycarbonate filtering apparatus with a 0.45-m cellulose acetate filter membrane. Upon filtration, the cation samples were promptly acidified to a pH of 2 using nitric acid, while the anion samples were left untreated.

Parameters such as potential hydrogen (pH), electrical conductivity (EC), total dissolved solids (TDS), and total hardness (TH), were examined to assess groundwater quality. Major cations, including sodium (Na), potassium (K), magnesium (Mg), and calcium (Ca), as well as major anions such as chloride (Cl), sulfate (SO_4), chloride (Cl), and bicarbonate (HCO_3), were measured using appropriate methods such as atomic absorption spectrophotometry (AAS) and ultraviolet spectrophotometry (UV) respectively. Additionally, bicarbonate concentration was determined through titration, while other water quality variables were analyzed using various methodological approaches.

3.2. Data collection

This study utilized data from SRTM DEM with 30 m resolution and satellite data downloaded from the USGS EarthExplorer website, satellite imagery (Landsat 7, 30 m resolution). This data was imperative in computing the different indices. The borehole data was acquired from the Community Water and Sanitation Agency (CWSA). Table 1, shows the different data types, formats, and sources used in this study.

3.3. Methodological framework

The methodology used in this research was to obtain a DEM and collect water quality data on existing boreholes in the study area (Fig. 2). Topographic features affecting groundwater potential (i.e., elevation, TWI, LS, geology, soil, LULC, aspect, slope, PLC, PRC, lineament density, distance to faults, drainage density) were then extracted from the DEM in the GIS environment. An arithmetic Water Quality Index (WQI) was then computed for each of the boreholes using the water quality data from each borehole. WQI was determined by calculating using physiochemical properties like pH, Total Dissolved Solids (TDS), Electrical Conductivity (EC), Total Alkalinity (TA), Total Hardness (TH), Manganese (Mn), Iron (Fe), Fluoride (Fl^-), Calcium (Ca^{2+}), Magnesium (Mg^{2+}), Chloride (Cl^-), Nitrate (NO_3^-)' and Sulfate (SO_4^{2-}). The borehole WQI was used as a predictor (dependent) variable in the prediction of groundwater quality.

3.4. Machine learning model

Random Forest is a machine learning method that combines the predictions of multiple decision trees to improve performance. It offers high accuracy in predictions, reduced variance, and is less prone to overfitting compared to individual decision trees. It can be observed that the borehole data used is concentrated around the southern part of the study area and the RF model has the capacity to predict the unsampled areas with higher accuracy. It can handle missing values, and feature importance, and can be applied to both classification and regression problems. Randomness introduced during training reduces the risk of overfitting, making it a powerful choice for noisy or complex datasets. Random Forest is efficient on large datasets with numerous features and instances, making it suitable for big data applications. It does not require feature scaling, making it less sensitive to the scale of input features. Random Forest can capture complex non-linear relationships in the data, making it suitable for tasks where underlying patterns are not well-described by linear models. It is relatively easy to tune, with default hyper parameters often providing good results. However, the choice of the best model depends on the specific characteristics of the dataset and the goals of the analysis. A thorough understanding of the data is crucial for selecting the right model.

By contributing to the forecast of groundwater resources, ML approaches have the potential to promote insight into groundwater and management. This can be accomplished by facilitating the collection of large water datasets, storing these datasets in databases, and processing these datasets to obtain useful insights that water resource managers can use to: determine water quality in untested areas or depth; design monitoring programs; assist in the formulation of groundwater protective measures; and, finally, assess the viability of groundwater water supply [44].

The RF model has been used to forecast manganese removal (Bhagat, Tiyasha et al., 2020), flood vulnerability Chen et al., 2020), pollution sources in water supply networks (Grbčić et al., 2020), and water quality prediction (Chen et al., 2020). Furthermore, the XGBoost model was used to predict water quality factors [45], biological water quality monitoring Chen et al., 2018), manganese removal prediction (Bhagat, Tiyasha et al., 2020), sediment heavy metal prediction (Bhagat, Tung et al., 2021), and lead

Table 1
Data, format, and their sources.

No.	Data	Format	Source
1	Topographic Data (DEM)	TIFF	USGS
2	Lithology	Shapefile/scanned maps	Ghana Geological Survey
3	Land use land cover	Shapefile	Esri Land Cover Map
4	Soil information	Shapefile	FAO Digital Soil Map
5	Rainfall data	Excel	CHIRPS
6	Borehole water quality	Excel/PDF	CWSA

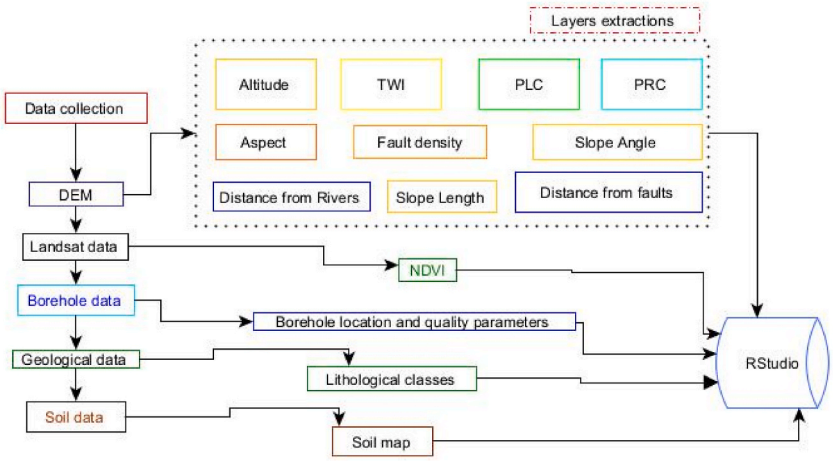


Fig. 2. Procedural framework in GIS used in this study (Authors: Source).

contamination prediction (Bhagat, Tiyasha et al., 2021), all with varying degrees of accuracy. ANN-based prediction models have been widely applied in various studies including wastewater heavy metal removal (Bhagat, Tung et al., 2020), heavy metal pollution prediction [44], flood susceptibility (Falah et al., 2019), and water level forecasting (Zhu et al., 2020).

This research makes use of the RF technique for predicting groundwater availability and quality. The RF approach employs an ensemble of classification and regression trees. Each tree is constructed using a different bootstrap sample (with replacement) from the original data [26]. In comparison to standard trees, RF adds a randomization layer to bagging, whereas in classic trees, each node is split using the optimal split among all variables. In the case of RF, only a random subset of the variables is considered when splitting a node during tree construction [27]. In comparison to other approaches, RF provides resilience to overfitting due to its random structures (Ning et al., 2022).

RStudio was used to create a stacked raster dataset (using the stack function) of the various raster layers produced from the GIS environment. These layers served as the predictor (independent) variables used in the Machine Learning algorithm (Random Forest). Thus, each pixel of the stacked raster dataset contained values for the individual raster layers. The pixel values were extracted from the stacked raster dataset for each borehole location and merged with the WQI dataset to generate a new dataset. This new dataset was used to build RF models for the prediction of groundwater quality (indicating whether or not water is suitable for domestic use). The new dataset was divided into training and testing data to build the RF model. The training data was used to train the RF model, whereas the test dataset was used to validate the model by predicting WQI in the testing dataset. The prediction of borehole WQI was done by regression in the Random Forest algorithm. An accuracy test was then carried out for the model using the OOB error in Random Forest.

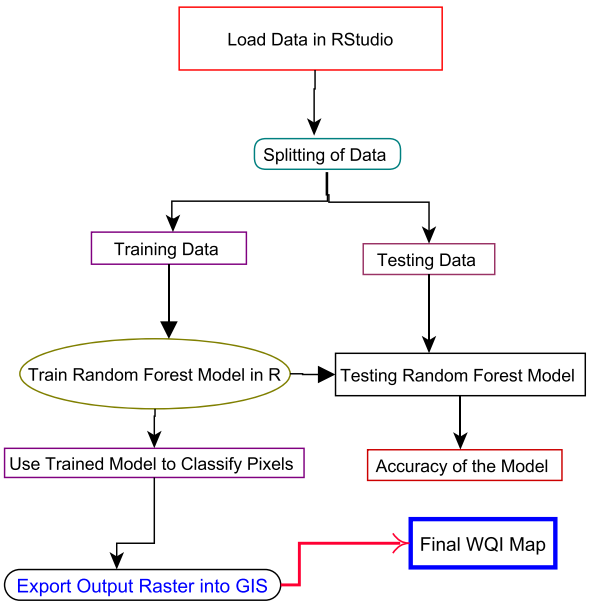


Fig. 3. Procedure in R Studio used to develop the WQI map in this study (Authors: Source).

The Random Forest algorithm in R was then used to predict the WQI for each pixel of the stacked raster dataset. The output of these classifications was saved to a new raster dataset (in TIF format). The classified raster dataset, which represents the Water Quality Index was now loaded in the GIS software. The raster dataset was converted to a vector dataset, ready to be published as a web map. The procedure is shown in Fig. 3.

3.5. Determination of water quality and water quality index

The water quality data was processed for the thirteen (13) selected physicochemical parameters and then the WQI was estimated for each water sample. The Water Quality Index is a dimensionless quantity (single-dimensional number) ranging from 0 to 100 that provides a reliable means of determining the overall quality of various parameters present in water. The physicochemical parameters were expressed in different units, and different ranges, and exhibited concentration-impact behaviour. As a result, before the development of the index, all the parameters must be converted into a single measure, and weights are then assigned based on the importance of the chemical on overall water quality.

According to the WQI, determined using the weighted arithmetic in this study, the allowable maximum value of WQI is 100, and values above 100 indicate pollution and are unsafe to drink [34,46–48]. The WQI generates a numerical measure that is used as a management tool in the assessment of water quality. The main advantage of this technique is that it integrates data from multiple water quality parameters into a mathematical equation that assigns a numerical value to water quality.

The thirteen (13) significant parameters selected for the determination of the WQI were EC, TDS, pH, TA, TH, Mn, Fe, Ca^{2+}, Mg^{2+}, Cl^{-}, F^{-}, NO^{3-}, and SO_4^{2-}. Equation (1) ([49,50];, 2020; [19–22]) was used to calculate the weighted arithmetic water quality index based on Ghana Standard Authority (GSA) water quality requirements for the different parameters.

$$WQI = \sum_{i=1}^{n} q_i W_i \tag{1}$$

Where W_i denotes the unit weight for every parameter, q_i is the 0–100 subindex rating for each variable, and n is the number of consolidated subindices. The Water Quality Index model in this research follows the five (5) steps outlined by Ref. [49].

Step 1: Parameter selection for measuring water quality.

- Thirteen (13) physicochemical parameters were selected for this study.

Step 2: Estimate the weightage of each parameter (I)

- The weightage of a parameter is inversely proportional to its allowable limits, i.e., weightage of parameter, $I = 1/S_i$, where S_i is the maximum permissible limits of the parameter [51]. The weightage of each parameter is shown in Table 2.

Step 3: Using the weightage of each parameter, estimate the unit weight of parameters (W_i).

- The unit weight (Wi) of parameters (determined using equation (2) [21]) is proportional to the weightage assigned to each parameter i.e,

$$W_i = K/S_i \tag{2}$$

And the constant of proportionality is given by equation (3) [19–22].

Table 2
Weightage of water quality parameters and the unit weight of water quality.

Parameters	GSA (S_i)	$I = 1/S_i$	$W_i = K/S_i$
pH	8.5	0.118	0.0176
EC	1000	0.001	0.0001
TDS	500	0.002	0.0003
TA	500	0.002	0.0003
TH	500	0.002	0.0003
Ca^{2+}	75	0.013	0.0020
Mg^{2+}	50	0.020	0.0030
Cl^{-}	250	0.004	0.0006
Mn	0.1	2.500	0.3739
Fe	0.3	3.333	0.4986
F^{-}	1.5	0.667	0.0997
NO^{3-}	50	0.020	0.0030
SO_4^{2-}	250	0.004	0.0006

$$K = 1 \Big/ \sum_{i=1}^{n} 1 \Big/ S_i \tag{3}$$

where K is a constant of proportionality; W_i is the unit weight of the parameter; n is the number of water quality parameters. The unit weight calculated for each parameter is shown in Table 2.

Step 4: Determining the subindex value (q_i). The subindex value is using equation (4) [19–22].

$$q_i = \frac{V_i - V_o}{S_i - V_o} \times 100 \tag{4}$$

Where;

V_i = mean concentration of the parameters in water, S_i = standard desirable or permissible concentration of the parameters in water, V_0 = Actual concentration of the parameters in pure water (generally $V_0 = 0$ for most parameters except pH).

For pH, q_i is determined using equation (5) [19–22].

$$q_i = \frac{V_i - 7}{8.5 - 7} \times 100 \tag{5}$$

1. Step 5: The overall WQI is calculated by aggregating the sub indices.

- WQI is the total of all the parameters' sub-indices (q_i) and unit weights (W_i) and was determined using equation (6) [19–22].

$$\text{Thus, } WQI = \sum_{i=1}^{n} q_i W_i \tag{6}$$

3.6. Prediction of groundwater quality

The dependent variable (predicted variable) for predicting groundwater quality was the WQI. The model was trained to predict borehole WQI in the training dataset and then validated using the test dataset. WQI is a numerical value and therefore the regression method of the Random Forest algorithm was used for the prediction. Water quality data for one hundred and forty (140) boreholes in the Nabogo Basin were used for training and validation of the machine learning model for groundwater quality.

4. Result and discussion

4.1. Variable of importance for groundwater quality prediction

The Random Forest model in RStudio was able to determine which of the independent variables were used for the prediction of groundwater availability and groundwater WQI of the study area. The following sub-sections provide details of the variables of importance for each of the predictions. For WQI prediction in this study, the variable of most importance as determined from the Random Forest algorithm shows that distance to faults is the most important, followed by rainfall and then NDVI, whereas the least important is PRC. The result of the variable of importance for the prediction of groundwater WQI in this study is shown in Fig. 4.

4.2. Prediction of water quality index

A Random Forest model was built and trained to predict borehole WQI in the study area. Since WQI is a numerical value, the regression method of the Random Forest algorithm was used for the prediction. Water quality data for One hundred forty-two (142)

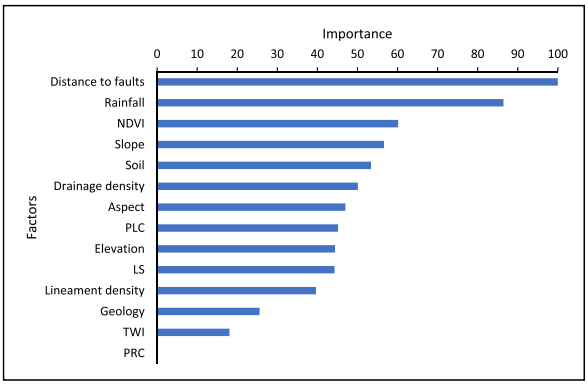


Fig. 4. Variable of importance for WQI prediction.

boreholes in the Nabogo Basin were used for training and validation of the machine learning model for groundwater quality. After extracting raster values of the 142 data points, 2 missing values were obtained for NDVI, and therefore these points were removed from the dataset, reducing the data points to 140. The 140 data points were then split into 70% for model training and 30% for model validation. This splitting was done to ensure that there remained sufficient points for validating the model. Thus, 98 points were used for model training and 42 points were used for validation. The value of mtry (randomized variables/elements to sample) affects the outcome of the accuracy (Table 3) of the RF model, as shown below (Fig. 5). The mtry value of 2 was used to select the optimal model since a mtry of 2 gave the smallest RMSE.

The predicted WQI (Fig. 6) of groundwater in the study area shows that WQI ranges from 9.51 to 69.99, and can be put in three classes according to the Arithmetic Water Quality Index method. The WQI in the study was classified as excellent (WQI up to 25), good (WQI between 25 and 50), and poor (WQI between 50 and 75) quality of groundwater sources, as seen in the water quality map (Fig. 7). It was found that the Ariel coverage for areas predicted to be excellent, good, and poor; 21.97%, 74.40%, and 3.63% respectively. The trained RF model was used to map GWQI classes across the entire region. Results showed that the Poor GWQI class is dominant in the study area, with Good GWQI found in the southwest and Very Poor GWQI observed in the north. These findings were consistent with other studies that have assessed the quality of water for different purposes using machine learning techniques [44,45, 52–56]. However, Ding et al. [25] in a study to optimize the water quality index using machine learning around the Haihe River Basin in China found that out of 178 samples evaluated the WQI in the area was classified as excellent, good, and poor quality had areas of 5.39%, 87.25%, and 7.35%. These findings were also consistent with this current study as major water quality class was found to be good. This study compares groundwater quality assessments in the Miandoab Plain Aquifer (NW Iran) and the Haihe River Basin in China due to similar environmental challenges, shared hydrogeological characteristics, method validation, global perspective, and policy implications. These areas may face similar environmental issues, such as agricultural runoff, industrial discharges, and urban development, which can affect groundwater quality. Comparing results from this region to the others can help assess the effectiveness of the random forest method, enabling a more comprehensive understanding of water quality issues worldwide.

Fundamentally, groundwater quality can be influenced by various human activities, including agricultural activities, phosphate, ammonium, bacteria, heavy metals, volatile organic compounds (VOCs), pesticides and herbicides, pH, and Total Dissolved Solids (TDS). The elevated nitrate levels in this study are linked to agricultural activities, while phosphate levels may be because of sewage discharges and detergent use. However, these parameters indicate a direct relationship to human influence around the study area. Regular monitoring is crucial to assess the potential impact of human activities and ensure water resource safety.

4.3. Correlation Matrix of Water Quality Prediction

The correlation matrix which shows the interaction between the various independent variables in the prediction of WQI (Fig. 8) showed that WQI has a positive correlation with Rainfall and NDVI (5), PLC and Elevation (12), PLC and Slope (17), PRC and LS (33), Lineament density and Drainage density (20), and Distance to fault and Elevation (39). This implied that an increase in WQI is attributable to increased Rainfall, NDVI, PLC, Elevation, Slope, Distance to faults, Limeament, and Drainage density values. The high correlation observed between PRC and LS can be mainly attributed to the topographical variation in the Nabogo basin. There was also a weak negative correlation between WQI and PLC (−14), PRC and PLC (−60), Slope and TWI (−72), Rainfall and Aspect (−21), and Drainage density and PLC (−20). Thus, locations that are well-drained are likely to have a reduced WQI. The highly negative correlation between Slope and TWI can also be attributed to the topographical variation. In summary, it can be said that groundwater quality in the Nabogo basin is mainly influenced by the topography and exacerbated by anthropogenic factors.

4.4. Effectiveness of RF for predicting WQI

The OOB error based on 10-fold cross-validation with 3 repeats for the training dataset produced an RMSE of 23.03 and an R^2 value of 0.82. The scatter plot (Fig. 5) of the prediction accuracy of the RF algorithm shows that the model was able to fairly predict the yield of 39 out of the 42 boreholes in the test dataset. The results from this work are consistent with Sami et al. (2022); Ubah et al. (2021); Wang & Ding [29] in separate studies in water quality index analysis where an R^2 of 0.98, 0.96, and 0.92 were found respectively. The difference in R^2 in the study and the other studies is attributed to the size of the training and validation dataset because several studies have a wide range of different datasets and hence would have different R^2s but the R^2 recorded in the study is considered appreciable.

Table 3
The error metrics for WQI prediction.

Mtry	RMSE	R-Squared	MAE
2	23.03	0.82	18.68
4	23.87	0.82	19.30
5	24.03	0.83	19.46
6	24.40	0.83	19.86
7	24.62	0.81	20.05
9	24.97	0.82	20.35
11	25.53	0.83	20.88
13	25.92	0.85	21.21
14	26.08	0.82	21.33

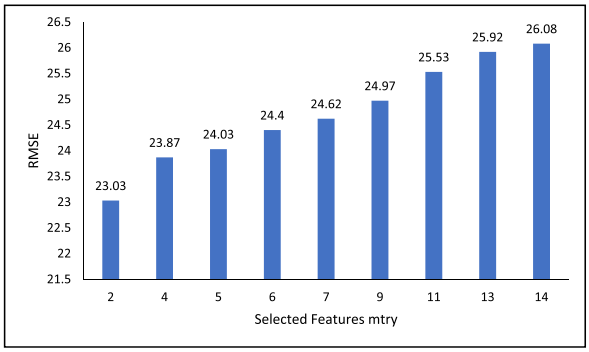


Fig. 5. Cross validation for WQI RF model.

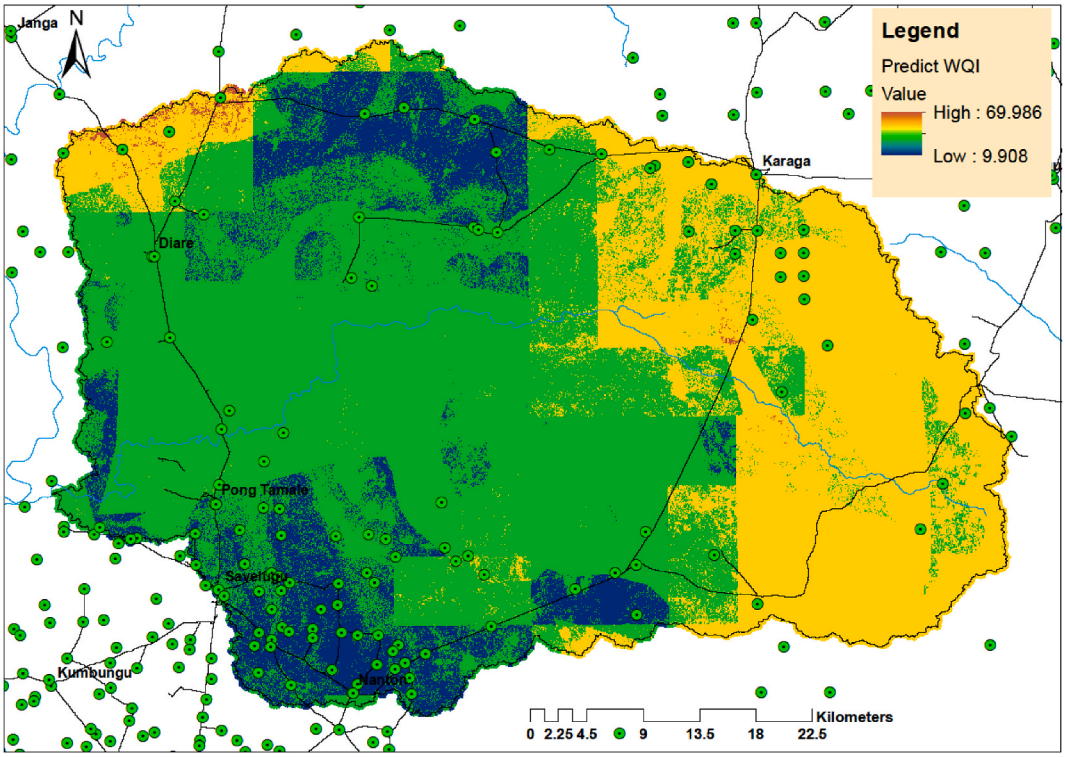


Fig. 6. Predicted WQI of the study area.

5. Conclusion and recommendation

The aim of this study was to predict groundwater quality in the Nabogo Basin by applying machine learning techniques. The results of the study suggest that.

- The resulting water quality map of the Nabogo basin showed that 21.97 %, 74.40 %, and 3.63 % of the study area had respectively the likelihood of excellent, good, and poor quality groundwater sources, based on classification using the arithmetic WQI.
- The WQI was predicted with a RMSE of 23.03 and an R^2 value of 0.82.

Similar studies should be carried out in the Nabogo Basin by applying machine learning techniques to predict quality. One of the major constraints of this study was the limited availability of data and limited spatial distribution of data in the study area. Future studies should consider a more spatially distributed dataset for the study. Secondly, the results of this study should be ground-truthed to validate the results. It is recommended that water quality tests be conducted on these newly drilled boreholes in order to verify the results obtained from this study. Other machine learning algorithms such as ANN and deep learning should also be applied in future studies in the basin to predict groundwater quality.

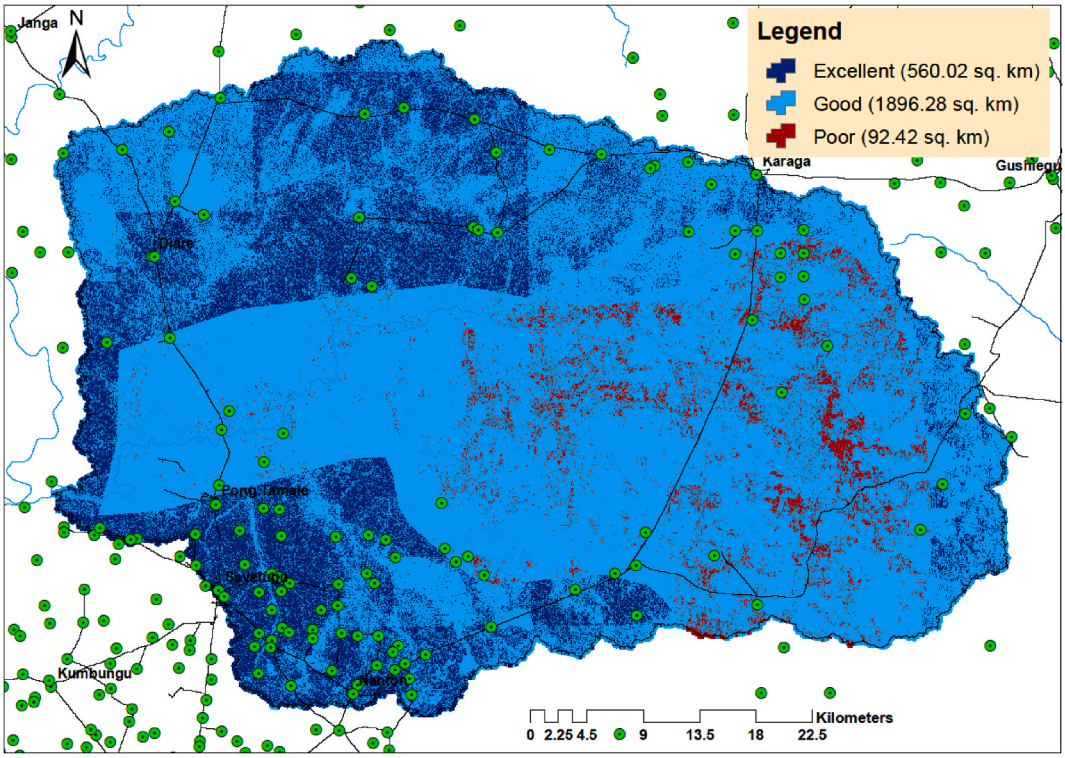


Fig. 7. Water quality of the study area.

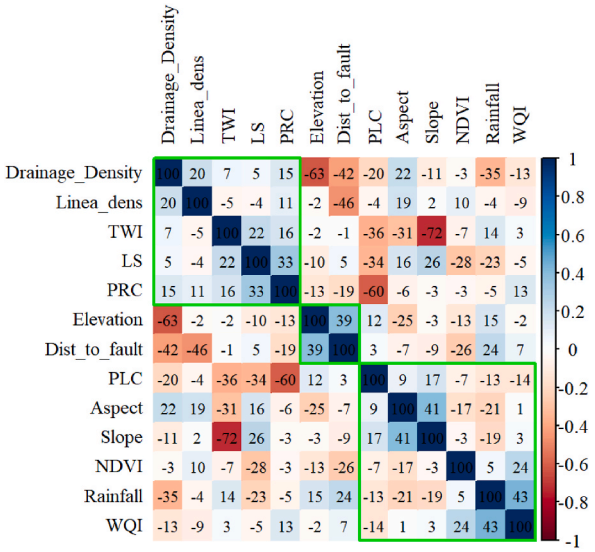


Fig. 8. Correlation matrix of water quality prediction.

Data availability statement

Data would made available on request.

CRediT authorship contribution statement

Joseph Nzotiyine Apogba: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Geophrey**

Kwame Anornu: Writing – review & editing, Supervision, Resources. **Arthur B. Koon:** Visualization, Supervision, Formal analysis, Data curation. **Benjamin Wullobayi Dekongmen:** Validation, Supervision, Resources, Investigation. **Emmanuel Daanoba Sunkari:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation. **Obed Fiifi Fynn:** Validation, Supervision, Resources, Investigation. **Prosper Kpiebaya:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Joseph Nzotiyine Apogba reports financial support and travel were provided by Regional Water and Environmental Sanitation Centre Kumasi (RWESCK). Joseph Nzotiyine Apogba reports a relationship with Regional Water and Environmental Sanitation Centre Kumasi (RWESCK) that includes: funding grants. Joseph Nzotiyine Apogba has patent pending to Joseph Nzotiyine Apogba. The authors declares no conflict of interest If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is part of the MSc thesis of the first author. It was funded by the Regional Water and Environmental Sanitation Centre Kumasi (RWESCK) at the Kwame Nkrumah University of Science and Technology (KNUST), Kumasi with funding from the Ghana Government through the World Bank under the Africa Centres of Excellence project. The views expressed in this paper do not reflect those of the World Bank, Ghana Government, and KNUST.

References

[1] M.M. Mekonnen, A.Y. Hoekstra, Sustainability: Four billion people facing severe water scarcity, Sci. Adv. 2 (2) (2016) 1–7, https://doi.org/10.1126/sciadv.1500323.
[2] F.N.W. Nsubuga, E.N. Namutebi, M. Nsubuga-Ssenfuma, Water resources of Uganda: an assessment and review, J. Water Resour. Protect. 6 (14) (2014) 1297–1315, https://doi.org/10.4236/jwarp.2014.614120.
[3] P.A. Owusu, S. Asumadu-Sarkodie, P. Ameyo, A review of Ghana's water resource management and the future prospect, Cogent Engineering 3 (1) (2016) 1164275.
[4] M.T.H. Van Vliet, E.R. Jones, M. Flörke, W.H.P. Franssen, N. Hanasaki, Global Water Scarcity Including Surface Water Quality and Expansions of Clean Water Technologies OPEN ACCESS Global Water Scarcity Including Surface Water Quality and Expansions of Clean Water Technologies, 2021.
[5] USGS, Where is Earth's Water? | U.S. Geological Survey (2018). June 2, https://www.usgs.gov/special-topics/water-science-school/science/where-earths-water.
[6] D. Askenaizer, Drinking Water Quality and Treatment, 2001.
[7] L.P. Chegbeleh, B.A. Akurugu, S.M. Yidana, Assessment of Groundwater Quality in the Talensi District, Northern Ghana, Scientific World Journal, 2020, https://doi.org/10.1155/2020/8450860, 2020.
[8] S. Dapaah-Siakwan, P. Gyau-Boakye, Hydrogeologic framework and borehole yields in Ghana, Hydrogeol. J. 8 (4) (2000) 405–416, https://doi.org/10.1007/PL00010976.
[9] S. Dapaah-Siakwan, P. Gyau-Boakye, Hydrogeologic framework and borehole yields in Ghana, Hydrogeol. J. 8 (4) (2000) 405–416, https://doi.org/10.1007/PL00010976.
[10] Kpakpo, GHANA LIVING STANDARDS SURVEY REPORT OF THE FIFTH ROUND (GLSS 5, Ghana Statistical Service, 2008.
[11] B. Snapir, D.M. Simms, T.W. Waine, 'Mapping the expansion of galamsey gold mines in the cocoa growing area of Ghana using optical Remote Sensing' (2017) 1–18, https://doi.org/10.1016/j.jag.2017.02.009.
[12] R.A. Kuffour, B.M. Tiimub, D. Agyapong, Impacts of Illegal mining (Galamsey) on the environment (water and soil) at Bontefufuo area in the Amansie west district, Environment and Earth Science 8 (7) (2018) 98–107.
[13] P. Li, Mine water Problems and solutions in China, Mine Water Environ. 37 (2018) 217–221, https://doi.org/10.1007/s10230-018-0543-z.
[14] E.E. Kwaansa-ansah, E.K. Armah, Assessment of total Mercury in Hair , urine and Fingernails of small – scale Gold miners in the amansie west district , Ghana, Health & Pollution 9 (21) (2019) 1–9, https://doi.org/10.5696/2156-9614-9.21.190306.
[15] A. Duncan, The dangerous couple: illegal mining and water pollution — a case study in fena river in the ashanti region of Ghana, J. Chem. 2020 (2020) 1–9, https://doi.org/10.1155/2020/2378560.
[16] P. Kpiebaya, A. Shaibu, E.E.Y. Amuah, R.W. Kazapoe, E. Salifu, B.W. Dekongmen, Impact of surficial factors on groundwater quality for irrigation using spatial techniques: emerging evidence from the northeast region of Ghana, H2Open Journal 6 (3) (2023) 387–402, https://doi.org/10.2166/h2oj.2023.156.
[17] A.B. Koon, G.K. Anornu, B.W. Dekongmen, E.D. Sunkari, A. Agyare, C. Gyamfi, Evaluation of groundwater vulnerability using GIS-based DRASTIC model in Greater Monrovia , Montserrado, Urban Clim. 48 (August 2022) (2023) 101427, https://doi.org/10.1016/j.uclim.2023.101427.
[18] A.M. Sajib, M.T.M. Diganta, A. Rahman, T. Dabrowski, A.I. Olbert, M.G. Uddin, Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach, Groundwater for Sustainable Development 23 (101049) (2023) 1–17, https://doi.org/10.1016/j.gsd.2023.101049.
[19] S.M. Deshpande, U.S. Bhagwat, K.R. Aher, Mathematical computation of weighted arithmetic water quality index of jui dam of jalna district, Maharashtra, Bull. Pure Appl. Sci. Geol. 40f (2) (2021) 219–226, https://doi.org/10.5958/2320-3234.2021.00019.6.
[20] M.R. Goodarzi, A.R.R. Niknam, A. Barzkar, M. Niazkar, Y.Z. Mehrjerdi, M.J. Abedi, M.H. Pour, Water quality index estimations using machine learning algorithms: a case study of yazd-ardakan plain, Iran, Water (Switzerland) 15 (10) (2023) 1–24, https://doi.org/10.3390/w15101876.
[21] D.D. Patel, D.J. Mehta, H.M. Azamathulla, M.M. Shaikh, S. Jha, U. Rathnayake, Application of the weighted arithmetic water quality index in assessing groundwater quality: a case study of the South Gujarat region, Water (Switzerland) 15 (19) (2023) 1–10, https://doi.org/10.3390/w15193512.
[22] K.D. Siriwardhana, D.I. Jayaneththi, R.D. Herath, R.K. Makumbura, H. Jayasinghe, M.B. Gunathilake, H.M. Azamathulla, K. Tota-Maharaj, U. Rathnayake, A Simplified equation for calculating the water quality index (WQI), Kalu river, Sri Lanka, Sustainability 15 (15) (2023) 1–15, https://doi.org/10.3390/su151512012.
[23] S.I. Abba, M.A. Yassin, A.S. Mubarak, S.M.H. Shah, J. Usman, A.Y. Oudah, S.R. Naganna, I.H. Aljundi, Drinking water resources suitability assessment based on pollution index of groundwater using improved Explainable Artificial Intelligence, Sustainability 15 (15655) (2023) 1–21, https://doi.org/10.3390/su152115655.
[24] Y. Xiong, T. Zhang, X. Sun, W. Yuan, M. Gao, J. Wu, Z. Han, Groundwater quality assessment based on the random forest water quality index—taking karamay city as an example, Sustainability 15 (14477) (2023) 1–18, https://doi.org/10.3390/su151914477.

[25] F. Ding, W. Zhang, S. Cao, S. Hao, L. Chen, X. Xie, W. Li, M. Jiang, Optimization of water quality index models using machine learning approaches, Water Res. 243 (2023), https://doi.org/10.1016/j.watres.2023.120337.
[26] E.A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula, M. Vaccari, Groundwater prediction using machine-learning tools, Algorithms 13 (11) (2020) 1–16, https://doi.org/10.3390/a13110300.
[27] E.A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula, M. Vaccari, Groundwater prediction using machine-learning tools, Algorithms 13 (11) (2020) 300, https://doi.org/10.3390/A13110300, 2020.
[28] S.A. Naghibi, M. Dolatkordestani, A. Rezaei, P. Amouzegari, M.T. Heravi, B. Kalantar, B. Pradhan, Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential, Environ. Monit. Assess. 191 (4) (2019), https://doi.org/10.1007/s10661-019-7362-y.
[29] X. Wang, F. Zhang, J. Ding, Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/S41598-017-12853-Y.
[30] B.F. Ziyad Sami, S.D. Latif, A.N. Ahmed, M.F. Chow, M.A. Murti, A. Suhendi, B.H. Ziyad Sami, J.K. Wong, A.H. Birima, A. El-Shafie, Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan, Sci. Rep. 12 (1) (2022), https://doi.org/10.1038/s41598-022-06969-z.
[31] British Geological Survey, Groundwater Quality: Ghana, National Environment Research Council, 2000, pp. 1–4.
[32] O.F. Fynn, S.M. Yidana, L.P. Chegbeleh, G.B. Yiran, Evaluating groundwater recharge processes using stable isotope signatures—the Nabogo catchment of the White Volta, Ghana, Arabian J. Geosci. 9 (4) (2016) 279, https://doi.org/10.1007/s12517-015-2299-0.
[33] H. Haghnazar, K.H. Johannesson, R. González-Pinzón, M. Pourakbar, E. Aghayani, A. Rajabi, A.A. Hashemi, Groundwater geochemistry, quality, and pollution of the largest lake basin in the Middle East: comparison of PMF and PCA-MLR receptor models and application of the source-oriented HHRA approach, Chemosphere 288 (2022), https://doi.org/10.1016/j.chemosphere.2021.132489.
[34] A. Ram, S.K. Tiwari, H.K. Pandey, A.K. Chaurasia, S. Singh, Y.V. Singh, Groundwater quality assessment using water quality index (WQI) under GIS framework, Appl. Water Sci. 11 (2) (2021) 1–20, https://doi.org/10.1007/s13201-021-01376-7.
[35] S.M. Yidana, M.O. Addai, L.-P. Chegbeleh, D. Adomako, B. Banoeng-Yakubo, Groundwater recharge processes in the Nasia sub-catchment of the White Volta Basin: analysis of porewater characteristics in the unsaturated zone, J. Afr. Earth Sci. 122 (2016) 4–14, https://doi.org/10.1016/j.jafrearsci.2015.04.006.
[36] E. Nsiah, E.K. Appiah-Adjei, K.A. Adjei, Hydrogeological delineation of groundwater potential zones in the Nabogo basin, Ghana, J. Afr. Earth Sci. 143 (2018) 1–9, https://doi.org/10.1016/j.jafrearsci.2018.03.016.
[37] M.L. Krautstrunk, An estimate of groundwater recharge in the Nabogo River Basin, Ghana using water table fluctuation method and chloride mass balance (2012).
[38] A.B. Adam, E.K. Appiah-Adjei, Groundwater potential for irrigation in the Nabogo basin, northern region of Ghana, Groundwater for Sustainable Development 9 (2019) 100274, https://doi.org/10.1016/J.GSD.2019.100274.
[39] Y.S. Anku, B. Banoeng-Yakubo, D.K. Asiedu, S.M. Yidana, Water quality analysis of groundwater in crystalline basement rocks, Northern Ghana, Environmental Geology 58 (5) (2009) 989–997, https://doi.org/10.1007/s00254-008-1578-4.
[40] G.K. Anornu, B.K. Kortatsi, Z.M. Saeed, Evaluation of groundwater resources potential in the Ejisu-Juaben district of Ghana, Afr. J. Environ. Sci. Technol. 3 (10) (2009) 332–340, https://doi.org/10.5897/AJEST09.048.
[41] M.S. Zango, M. Anim-Gyampo, B. Ampadu, Assessment of groundwater sustainability in the Bawku east municipality of Ghana, J. Sustain. Dev. 7 (3) (2014) 59–70, https://doi.org/10.5539/jsd.v7n3p59.
[42] S. Abdul-Ganiyu, K. Prosper, Estimating the groundwater storage for future irrigation schemes, Water Supply (2021) 1–15, https://doi.org/10.2166/ws.2021.041.
[43] Y.S.A. Loh, B.A. Akurugu, E. Manu, A.S. Aliou, Assessment of groundwater quality and the main controls on its hydrochemistry in some Voltaian and basement aquifers, northern Ghana, Groundwater for Sustainable Development 10 (2020) 100296, https://doi.org/10.1016/j.gsd.2019.100296. November 2019.
[44] S. Singha, S. Pasupuleti, S.S. Singha, R. Singh, S. Kumar, Prediction of groundwater quality using efficient machine learning technique, Chemosphere 276 (2021) 130265, https://doi.org/10.1016/J.CHEMOSPHERE.2021.130265.
[45] K. Joslyn, Water quality factor prediction using supervised machine learning, REU Final Reports (2018). https://pdxscholar.library.pdx.edu/reu_reports/6.
[46] R. Gupta, A.N. Singh, A. Singhal, Application of ANN for water quality index, International Journal of Machine Learning and Computing 9 (5) (2019) 688–693, https://doi.org/10.18178/IJMLC.2019.9.5.859.
[47] H.K. Pandey, V. Tiwari, S. Kumar, A. Yadav, S.K. Srivastava, Groundwater quality assessment of Allahabad smart city using GIS and water quality index, Sustainable Water Resources Management 6 (2) (2020), https://doi.org/10.1007/s40899-020-00375-x.
[48] K.N. Rao, P.S. Latha, Groundwater quality assessment using water quality index with a special focus on vulnerable tribal region of Eastern Ghats hard rock terrain, Southern India, Arabian J. Geosci. 12 (8) (2019), https://doi.org/10.1007/s12517-019-4440-y.
[49] K.A. Shah, G.S. Joshi, Evaluation of water quality index for River Sabarmati, Gujarat, India, Appl. Water Sci. 7 (3) (2017) 1349–1358, https://doi.org/10.1007/s13201-015-0318-7.
[50] F.M. Kizar, A comparison between weighted arithmetic and Canadian methods for a drinking water quality index at selected locations in shatt al-kufa, IOP Conf. Ser. Mater. Sci. Eng. 433 (1) (2018), https://doi.org/10.1088/1757-899X/433/1/012026.
[51] S. Deepak, N.U. Singh, Water quality index for ground water (GWQI) of Dhar town MP, India, International Research Journal of Environment Sciences 2 (11) (2013) 72–77.
[52] U. Ahmed, R. Mumtaz, H. Anwar, A.A. Shah, R. Irfan, J. García-Nieto, Efficient water quality prediction using supervised machine learning, Water (Switzerland) 11 (11) (2019), https://doi.org/10.3390/W11112210.
[53] A. Arabameri, S.C. Pal, F. Rezaie, O.A. Nalivan, I. Chowdhuri, A. Saha, S. Lee, H. Moayedi, Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques, J. Hydrol.: Reg. Stud. 36 (February) (2021) 100848, https://doi.org/10.1016/j.ejrh.2021.100848.
[54] D.T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, N. Kazakis, Improving prediction of water quality indices using novel hybrid machine-learning algorithms, Sci. Total Environ. 721 (2020), https://doi.org/10.1016/j.scitotenv.2020.137612.
[55] J. Kim, H. Han, L.E. Johnson, S. Lim, R. Cifelli, Hybrid machine learning framework for hydrological assessment, J. Hydrol. 577 (2019), https://doi.org/10.1016/j.jhydrol.2019.123913.
[56] E. Rozos, Machine learning, urban water resources management and operating policy, Resources 8 (4) (2019), https://doi.org/10.3390/RESOURCES8040173.