



Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India

Shruti Sachdeva¹ · Bijendra Kumar¹

Accepted: 30 September 2020 / Published online: 6 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

In the drought prone district of Dholpur in Rajasthan, India, groundwater is a lifeline for its inhabitants. With population explosion and rapid urbanization, the groundwater is being critically over-exploited. Hence the current groundwater potential mapping study was undertaken to ascertain the areas that are more likely to yield a larger volume of groundwater against those areas that have poor groundwater potential and accordingly perpetuate the much needed damage control. Thematic layers for 14 groundwater influencing factors were considered for the study region, including elevation, slope, aspect, plan curvature, profile curvature, topographic wetness index (TWI), geology, soil, land use, normalized difference vegetation index (NDVI), surface temperature, precipitation, distance from roads, and distance from rivers. These were then subjected to an overlay operation, with the groundwater inventory which comprised of the locations of observational groundwater wells. The resulting geospatial database was then used to train two decision tree based ensemble models: gradient boosted decision trees (GBDT) and random forest (RF). The predictive performance of these models was then compared using various performance metrics such as area under curve (AUC) of receiver operating characteristics (ROC), sensitivity, accuracy, etc. It was found that GBDT (AUC: 0.79) outperformed RF (AUC: 0.71). The validated GBDT model was then used to construct the groundwater potential zonation map. The generated map showed that about 20.2% of the region has very high potential, while 22.6% has high potential to yield groundwater, and approximately 19.9–17.5% of the study region has very low to low groundwater potential.

Keywords Groundwater potential mapping · Machine learning · Ensemble models · Random forest · Gradient boosted decision trees

1 Introduction

Groundwater is one amongst the world's most over-utilized and under-appreciated natural resources. Water in the saturated zone, residing below the surface of the earth in aquifers, is referred to as groundwater (Fitts 2002). Water from surface water bodies (rivers, ponds, streams, canals, etc.) and rainfall, seeps through the earth's crust via interconnected networks of channels like crevices, fractures, cracks, crushed zones (fault zones or shear zones),

and joints and get collected in the underground reservoirs contributing to the water table (Banks et al. 2002). Most aquifers are identified by the general groundwater characteristics such as consistent temperature, extensive availability, economical extraction, and drought resilience (Jha et al. 2007). Also, due to its confined nature, the water remains shielded from most surface adversities like gas leaks, radiation exposure, etc. and hence become the major source of fresh water in emergency situations and otherwise. As a result a large portion of freshwater demands for domestic and drinking purposes is met by means of these dormant water bodies. An alarming yet an unsurprising pattern of declining water table levels, thus, have emerged owing to various unsustainable human practices. The groundwater is being extracted at a much faster pace than the rate at which it can be naturally replenished. Ever increasing population cascading to an exponential increase

✉ Shruti Sachdeva
sachdeva_shruti@yahoo.in

Bijendra Kumar
bizender@gmail.com

¹ Department of Computer Science and Engineering, Netaji Subhas University of Technology, New Delhi, India

in water demands, coupled with rapid urbanization and industrialization to support the said population, plays havoc with the fine balance of replenishment and evaporation. Rigorous agricultural practices for obtaining higher yields, with little or no regard for resources overlays a heavy cost on the dwindling water table. This stands true, especially in a country like India, where a majority of the population relies on agriculture as their only source of livelihood, thus superimposing the pressure on an already precarious resource.

The latency associated with groundwater due to the aquifers being cocooned below earth's surface, often leads to an unobserved and hence uninhibited mining for groundwater. Consequentially, many of the underground aquifers and groundwater springs have run dry, well before their lifetime. Attempts to find new springs or boring deeper into the existing ones will not only increase the chances of contamination but can also cause land subsidence. So there is a critical need for undertaking studies for delineating regions on the basis of their potential for groundwater yields. Such studies could help governments with assessment, planning, and conservation of groundwater as well as in land use planning for viable development. Aquifer location identification is a tedious task on account of their inherent physiognomy. Conventionally, test drilling and stratigraphy analysis have been relied on for this purpose, but are found unfeasible in regards of their time, cost and skilled manpower requirements (Sander et al. 1996). Geophysical survey lacks in precision and need to be validated by borehole data, thus making them expensive and inefficient (Mukherjee et al. 2012). Technologies such as Remote Sensing (RS) and Geographic Information System (GIS) have completely revolutionized the way such extensive studies were historically conducted.

With satellite images delivering accurate data at high temporal resolution and with such high spatial precision that the need for conducting field surveys has been significantly obliterated. This has resulted in a plethora of mapping studies such as landslide susceptibility mapping (Bragagnolo et al. 2020a, b; Hu et al. 2020; Sameen et al. 2020; Sansare and Mhaske 2020; Tang et al. 2020; Van Dao et al. 2020; Wang et al. 2020; Wu et al. 2020), flood susceptibility mapping (Bui et al. 2020; Chen et al. 2020a; Costache and Bui 2020; Feloni et al. 2020; Feng et al. 2020; Mishra and Sinha 2020; Pourghasemi et al. 2020; Sansare and Mhaske 2020; Sarkar and Mondal 2020), and forest fire susceptibility mapping (Abedi Gheshlaghi et al. 2020; Çolak and Sunar 2020; Rahimi et al. 2020; Sevinc et al. 2020; Venkatesh et al. 2020), mineral potential mapping (de Quadros et al. 2006) etc., employing RS for obtaining data for regions that were traditionally considered inaccessible. These studies have focused on generating zonation maps delineating the zones on the basis of

their relative potential/susceptibility/vulnerability/prone-ness using a variety of statistical techniques such as weight-of-evidence (Mastere 2020; Zaheer et al. 2020; Rahmati et al. 2016; Kayastha et al. 2012; Ozdemir 2011; Corsini et al. 2009; de Quadros et al. 2006; Lee and Choi 2004), frequency ratio (Sarkar and Mondal 2020; Rahmati et al. 2016; Naghibi et al. 2015; Tehrany et al. 2015; Ozdemir 2011; Lee and Pradhan 2007), evidential belief function (Althuwaynee et al. 2014; Nampak et al. 2014; Althuwaynee et al. 2012; Carranza and Hale 2003), analytical hierarchy process (Kaur et al. 2020; Rahmati et al. 2015; Althuwaynee et al. 2014; Pourghasemi et al. 2012; Yalcin et al. 2011; Yalcin 2008), etc. The aforementioned techniques usually rely on the knowledge base of an expert and hence are expensive in terms of cost, time and effort and are handicapped by the possibility of human error.

With the advancements in the field of data science, geospatial mapping is also experiencing a surge in the number of studies applying machine learning models like support vector machine (Tehrany et al. 2015; Pradhan 2013; Tien Bui et al. 2012; Yilmaz 2010), decision trees (Wu et al. 2020; Pradhan 2013; Tien Bui et al. 2012), artificial neural networks (Bragagnolo et al. 2020a, b; Bui et al. 2020; Chen et al. 2020a; Tang et al. 2020; Van Dao et al. 2020; Yilmaz 2010; Corsini et al. 2009), naïve bayes (Tien Bui et al. 2012; Porwal et al. 2006), k-nearest neighbors (Beaudoin et al. 2014; Gjertsen 2007), etc. These studies exhibited good results with high accuracy and generated reliable maps with a high degree of precision. Consequentially, the machine learning techniques hold promise for groundwater potential mapping studies as well. Subsequently a lot of groundwater potential mapping studies have been undertaken using machine learning and RS with the help of GIS, in various parts of the world such as Ghana (Sander et al. 1996), Turkey (Ozdemir 2011), Italy (Corsini et al. 2009), Malayasia (Nampak et al. 2014), Iran (Naghibi et al. 2015; Rahmati et al. 2015, 2016). The number of studies on the Indian subcontinent (Mukherjee et al. 2012; Jha et al. 2007) is highly disproportionate to the sheer size of its surface area and population. Although few studies have been undertaken in the last decade (Kaur et al. 2020; Pham et al. 2019), there still remains a large scope for understanding the subterranean hydrological aspect of the Indian geography. Also, a rapidly growing population and an unsustainable pace of economic growth has the country facing an acute water crisis, amplifying the pressure on the already critical groundwater resources to meet such unprecedented challenges. The situation is particularly distressing for the desert state of Rajasthan that has witnessed a 62.7% decline in groundwater from 2008 to 2018 (CGWB 2018). Considering the lack of studies in this part of the world, in conjunction with the worsening state of water resources in the region, the current study was

undertaken to investigate and get a better insight into the situation.

A new feat in the machine learning field has been the up and coming ensemble and optimization techniques such as Bagging, Boosting, Stacking and Voting (Wu et al. 2020). These have superseded the performance of the individual base models significantly by learning from their weaknesses and strengths. These hold a bright scope for providing better results in the intensive geospatial analysis of groundwater potential mapping as well and to bridge the gap in the hydrological comprehension around the world and in India in particular. The following study is one such step in this direction. Decision tree based ensemble techniques such as, Gradient Boosted Decision Trees (GBDT) and Random Forest (RF) models have found applications in a variety of geospatial mappings for instance, landslide susceptibility mappings (Arabameri et al. 2019a; Chen et al. 2017, 2018; Dou et al. 2019; Lombardo et al. 2015; Thai Pham et al. 2018), flood susceptibility mappings (Chen et al. 2020b; Choubin et al. 2019; Khosravi et al. 2018; Yariyan et al. 2020), gully erosion mappings (Arabameri et al. 2019b; Avand et al. 2019; Garosi et al. 2018; Gayen et al. 2019; Hosseinalizadeh et al. 2019; Zabihi et al. 2019) etc. The high predictive accuracies achieved by tree based ensemble in such studies are generally attributed to their numerous abilities, namely simple parameterization, adaptability to accommodate different type of predictors, flexibility in fitting the predictors, interpretability (Kuhnert et al. 2010), and computational practicality as compared to other machine learning models such as support vector machine and artificial neural networks (Rodriguez-Galiano and Chica-Olmo 2012).

In this study, GBDT and RF, owing to their previous good performances in similar studies (Chen et al. 2020a; Wu et al. 2020; Naghibi et al. 2016; Zabihi et al. 2016) were chosen to be applied in the groundwater potential mapping for the Dholpur district in Rajasthan, India. The main difference between the current study and the aforementioned literature from this field is that, while these works have explored decision tree based ensemble models and established their efficacy individually and against other statistical and machine learning approaches (such as naïve bayes, k-nearest neighbors, frequency ratio, etc.) in various geospatial applications and also amongst a few hydrological studies, however they have yet not been compared against each other particularly in a groundwater mapping study. This holds significance, especially given the fact that both these approaches of GBDT and RF are ensembles on the same baseline model of decision trees and a juxtaposition of their respective performances is logically relevant to the field. Also, to the best of our knowledge, no similar study has been undertaken in the study area and the state of these resources in this region exemplifies the severity of the

need for such a study. These decision tree ensembles were compared on performance metrics such as accuracy, area under curve (AUC) of receiver operating characteristic (ROC), sensitivity, etc. and the validated model was used to generate the groundwater potential zonation map for the region under study so as to get the study area delineated into different regions on the basis of their probability to yield groundwater. In all, the specific objectives of the study are to (1) explore and compare the predictive capabilities of GBDT and RF models for groundwater potential mapping, and (2) generate an accurate and reliable groundwater potential map for the region delineating it into low to high potential zones.

2 Study area

Dholpur district is situated in the eastern part of the desert state of India, Rajasthan as displayed in Fig. 1. The region under study spans over an area of 3339 sq. km and lies within the eastern longitudes 77°13'12" and 78°15'44" and the northern latitudes 26°21'44" and 26°57'25". The district is landlocked with the districts of Bharatpur and Karauli surrounding it from the northwest, south and southwest directions within the state of Rajasthan itself. While on the east it shares its borders with the states of Uttar Pradesh and Madhya Pradesh. It experiences a semi-arid type climate with the summers being very hot lasting from March to June and the winters being equally cold, sustained from November to February. In the remaining months, it receives monsoon rains with the average precipitation of 560 mm. The district has the history of witnessing the highest occurrence of mild droughts as per Central Ground Water Board (CWGB 2018).

The highest yielding aquifers in the region deliver groundwater volume in the range from 200 to 800 cubic meters per day. The lowest yielding wells tend to be the ones, taking the longest to recuperate. The depth to groundwater level varied from 5 to 40 mbgl. Southwestern, northern and northeastern parts of the district witnessed deeper water levels, greater than 20 mbgl. Also the water extracted is found fit for drinking, industrial and agricultural purposes. Although a fluoride concentration varying between 0.33 and 2.93 mg/l (exceeding the permissible limit of 1.5 mg/l), and nitrate concentration in the range of 5.1–202.5 mg/l (far exceeding the safety limit of 45 mg/l for drinking purposes) was found in isolated places (Central Ground Water Board (CGWB) 2018).



Fig. 1 Study area

3 Spatial data generation

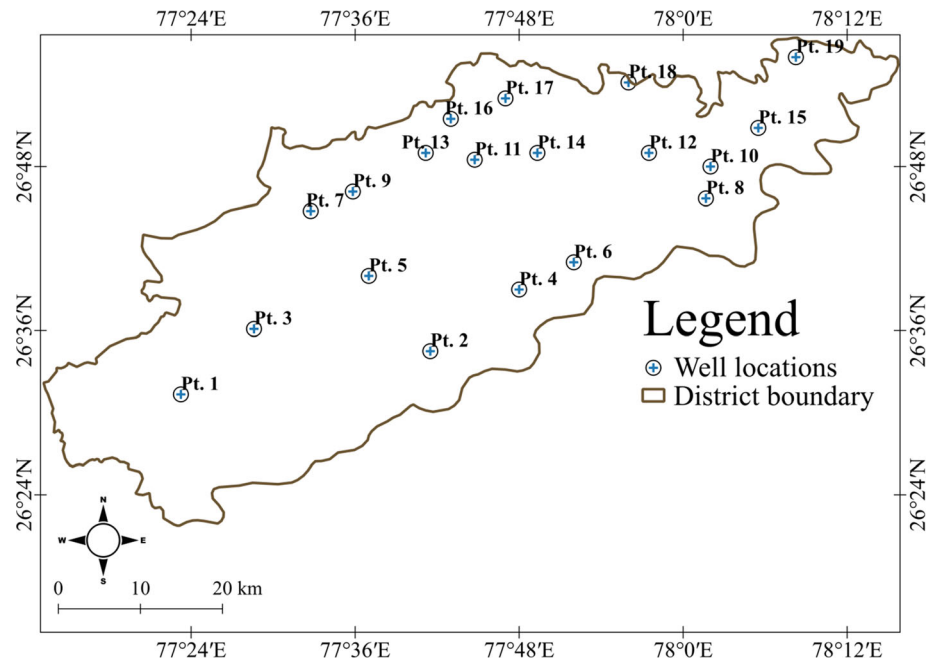
The geospatial dataset was produced for the study region by the amalgamation of the groundwater inventory with their respective topological, morphometric, hydrological, anthropogenic, and meteorological factor's values influencing the groundwater potential at those locations. These have been discussed in further details below:

3.1 Groundwater inventory

Data of a total of 19 groundwater wells was sourced from the Central Ground Water Board (CGWB) for the year 2018, as shown in Fig. 2. These wells gave an average yield of 200 cubic meters or more per day and hence were included as groundwater inventory for reliable results from the geospatial mapping. The points were mapped in QGIS and were validated by means of newspaper stories and articles in state office journals.

3.2 Groundwater influencing factors

The majority of the academicians and researchers consider the groundwater potential at a location as the probability of the location to yield an optimum volume of groundwater on underground exploration (Díaz-Alcaide and Martínez-Santos 2019). The concept of optimum volume is in itself very subjective to the region under study and varies from one to the other. The factors that predominately affect this variation are included in the ambit of groundwater influencing factors. These factors themselves exhibit an inherent variability in the studies undertaken worldwide. Based on the literature survey and data availability, a set of 14 groundwater influencing factors was considered. The influencing factors included in the study are elevation, slope, aspect, plan curvature, profile curvature, topographic wetness index (TWI), land use, normalized difference vegetation index (NDVI), geology, soil texture, surface temperature, precipitation, distance from roads and distance from rivers. The aforementioned factors and their influence on the absence or presence of groundwater have been briefly described below.

Fig. 2 Locations of the groundwater wells

Elevation at location refers to the altitude above the earth's surface. It plays an important role in determining a region's groundwater potential as the variation in elevation brings variation in the region's climatic and environmental conditions which in turns affects the region's soil and vegetation, thus influencing the region's water retention and porosity of the span of land (Al-Abadi and Shahid 2015). The elevation for the study area varies between 100 and 350 m and is displayed in Fig. 3a. The **Slope** is responsible for the gradient of the land that affects the trajectory for surface water runoff and infiltrate. The steeper the slope, greater would be the runoff, which subsequently leads to a lesser volume to seep through the surface, hence lower the recharge. It is depicted in Fig. 3b. The direction of the slope (north, northeast, east, etc.) is referred to as the **Slope Aspect**, and it indirectly affects the groundwater potential by influencing the location's exposure to sunshine, rainfall and wind. It has been deduced that south facing slopes experience greater exposure to sunshine and stronger winds and lower humidity, thus reducing the water infiltration (Zabihi et al. 2016). Slope aspect of the region is shown in Fig. 3c. **Plan curvature** as shown in Fig. 3d affects the flow convergence and divergence and is indicated by curvature of the line intersecting the surface by the horizontal plane (Lee et al. 2012). On the other hand, **Profile curvature**, is defined as the curvature of the line formed by intersecting the surface with the vertical plane. It is shown in Fig. 3e.

TWI has a direct proportionality with groundwater recharge and is calculated as in the Eq. (1).

$$TWI = \ln\left(\frac{\gamma}{\tan \delta}\right) \quad (1)$$

Here δ is the slope at the point while γ is the area of the upslope that runs off to the point. The TWI for the region is shown in Fig. 3f. **Land use** for a region demonstrates the effect and the degree of the effect that the human activities have on that region's landscape and resources. It is shown in Fig. 3g where it demarcates the study area's terrain into categories such as forests (broadleaved deciduous and needle leaved evergreen), croplands (irrigated and rain fed), urban areas, shrub lands, grasslands, bare areas, and water bodies. The **NDVI** is included in this study to account for the direct relationship found between the water table levels and the density of vegetation in the same region, i.e., the greener the region, higher are the chances for natural groundwater replenishment. The NDVI is calculated using Eq. (2) using the near infrared (NIR) and red (R) values from bands 5 and 4 respectively, of the Landsat image obtained for the region in the year 2018.

$$NDVI = \frac{NIR - R}{NIR + R} \quad (2)$$

Geology for the study area is shown in Fig. 3i and it directly influences the runoff and the infiltration properties of the terrain. The study region has a larger span of its area covered in clayey soil type and a smaller portion with loamy type as depicted in Fig. 3j. The **soil texture** defines the porosity and the coarseness which directly affects the recharge and replenishment cycle.

The water table at a location is directly affected by the meteorological conditions at that location. The

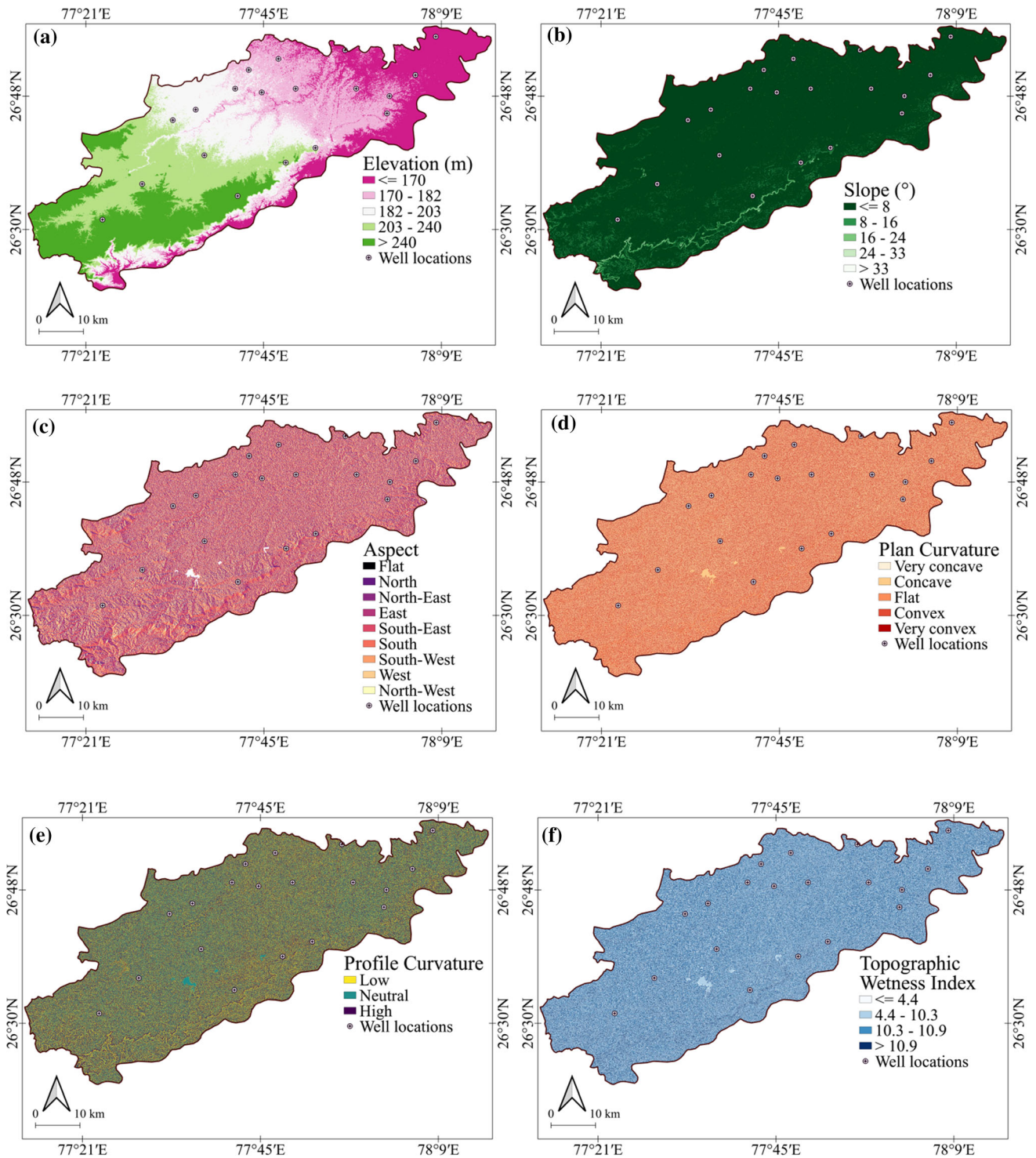


Fig. 3 a–d Groundwater influencing factor: Elevation, Slope, Aspect, and Plan curvature. **e–h** Groundwater influencing factor: Profile curvature, TWI, Land use, and NDVI. **i–l** Groundwater influencing

factor: Geology, Soil texture, temperature, and precipitation. **m, n** Groundwater influencing factor: Distance from rivers and distance from roads

temperature in the region affects the residual volume of surface water bodies left for infiltration after the surface evaporation, while the **precipitation** in the region directly affects the degree of natural replenishment from rainfall.

These are depicted in Fig. 3k and l respectively. The groundwater is directly influenced by the location’s proximity to the closest water body; hence the influencing factor of **distance from rivers** has been included in this

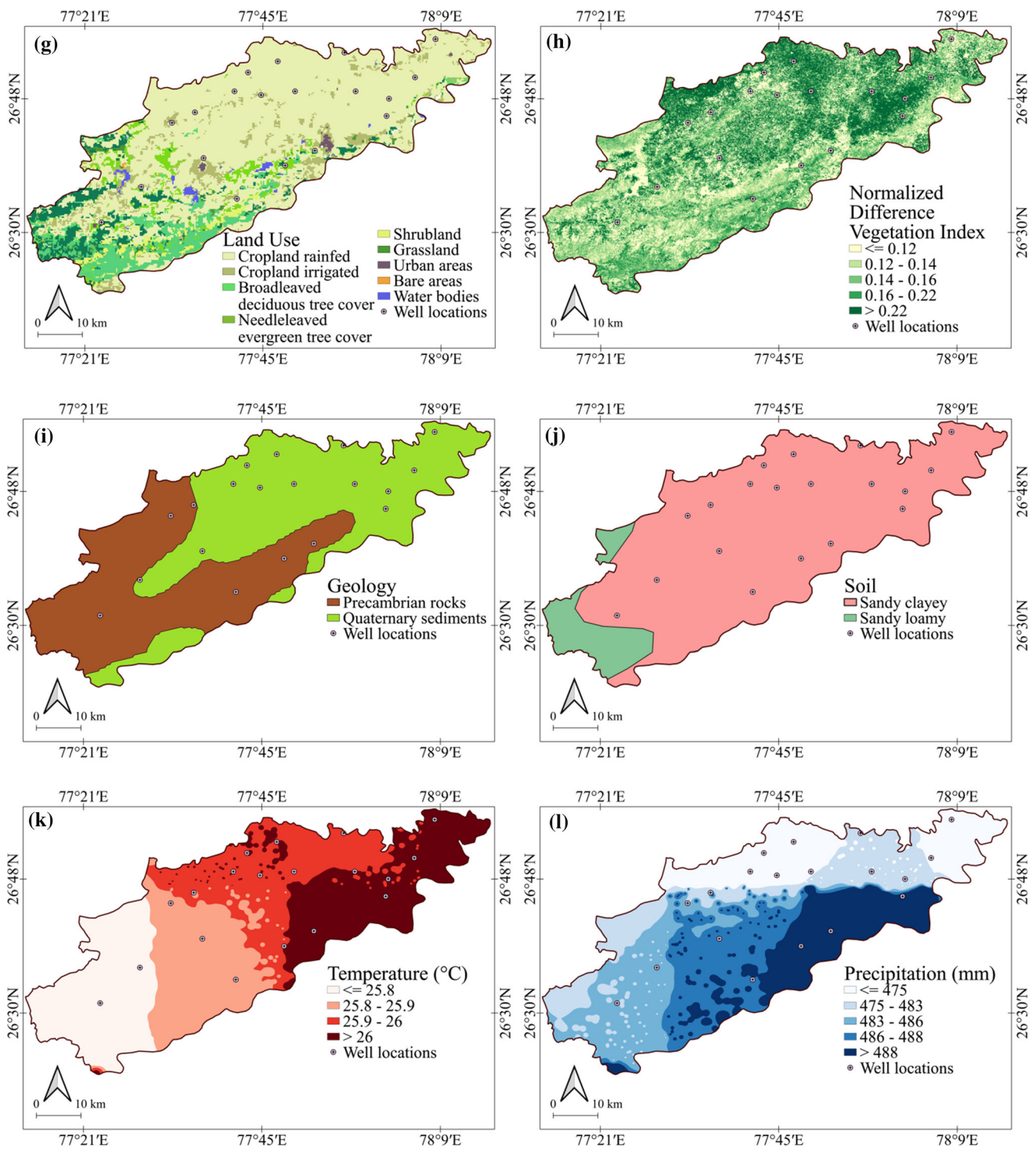


Fig. 3 continued

study as shown in Fig. 3m. Anthropological factors that account for the effect of human activities on the groundwater have been taken into consideration through the well’s **proximity to roads**. It is shown in Fig. 3n. Table 1 gives a brief summary of the aforementioned influencing factors

and their respective maps. Each influencing factor was transformed into a grid spatial database by 30×30 m size and the grid of the Dholpur area was constructed by 3507 columns and 2269 rows (3,710,000 pixels; 3339 sq.km).

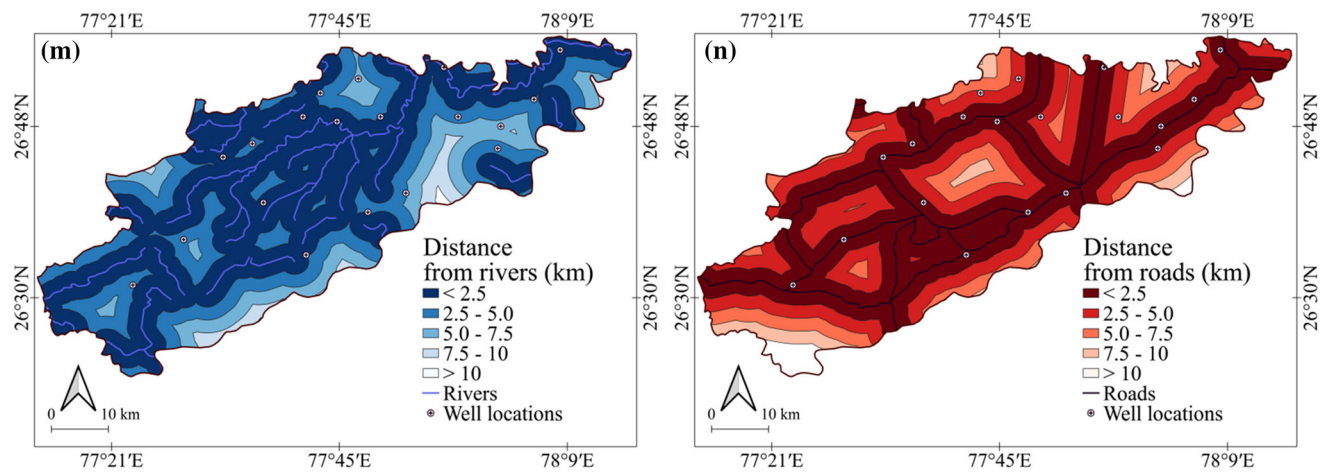


Fig. 3 continued

4 Methodology

Figure 4 displays the approach adopted in the study. Firstly, the data for the 14 groundwater influencing factors specific to the study region were collected and then were merged with the groundwater inventory that comprised of the locations of the groundwater wells in the region. The

data for an equal number of “non-groundwater locations”, as obtained from CGWB, state office reports, and newspaper articles, were also appended into the dataset to avoid bias during model development. The compiled dataset generated was then randomly split in the ratio of 80:20 for training and testing purposes. The training subset of the data set was used to train the competing ensemble models

Table 1 Groundwater influencing factors

Influencing factor	Source	Data type	Year	Resolution/Scale	Depiction
Elevation	The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM)	Raster	2018	1 arc second (approximately 30 m)	3(a)
Slope	Derived from DEM	–	–	–	3(b)
Slope Aspect	Derived from DEM	–	–	–	3(c)
Plan Curvature	Derived from DEM	–	–	–	3(d)
Profile Curvature	Derived from DEM	–	–	–	3(e)
TWI	Derived from DEM	–	–	–	3(f)
Land Use	European Space Agency (ESA) Climate Change Initiative (CCI)	Raster	2018	300 m	3(g)
NDVI	Landsat 8 OLI	Raster	2018	30 m	3(h)
Geology	US Geological Survey (USGS) EarthExplorer	Shapefile	2005	–	3(i)
Soil Texture	Land and Water Development Division, FAO	Shapefile	2017	–	3(j)
Temperature	APHRODITE (Asian Precipitation—Highly-Resolved Observational Data Integration Towards Evaluation)	Raster	2015	0.25°	3(k)
Precipitation	APHRODITE (Asian Precipitation—Highly-Resolved Observational Data Integration Towards Evaluation)	Raster	2015	0.25°	3(l)
Distance to rivers	DIVA GIS	Shapefile	2011	–	3(m)
Distance to roads	DIVA GIS	Shapefile	2011	–	3(n)

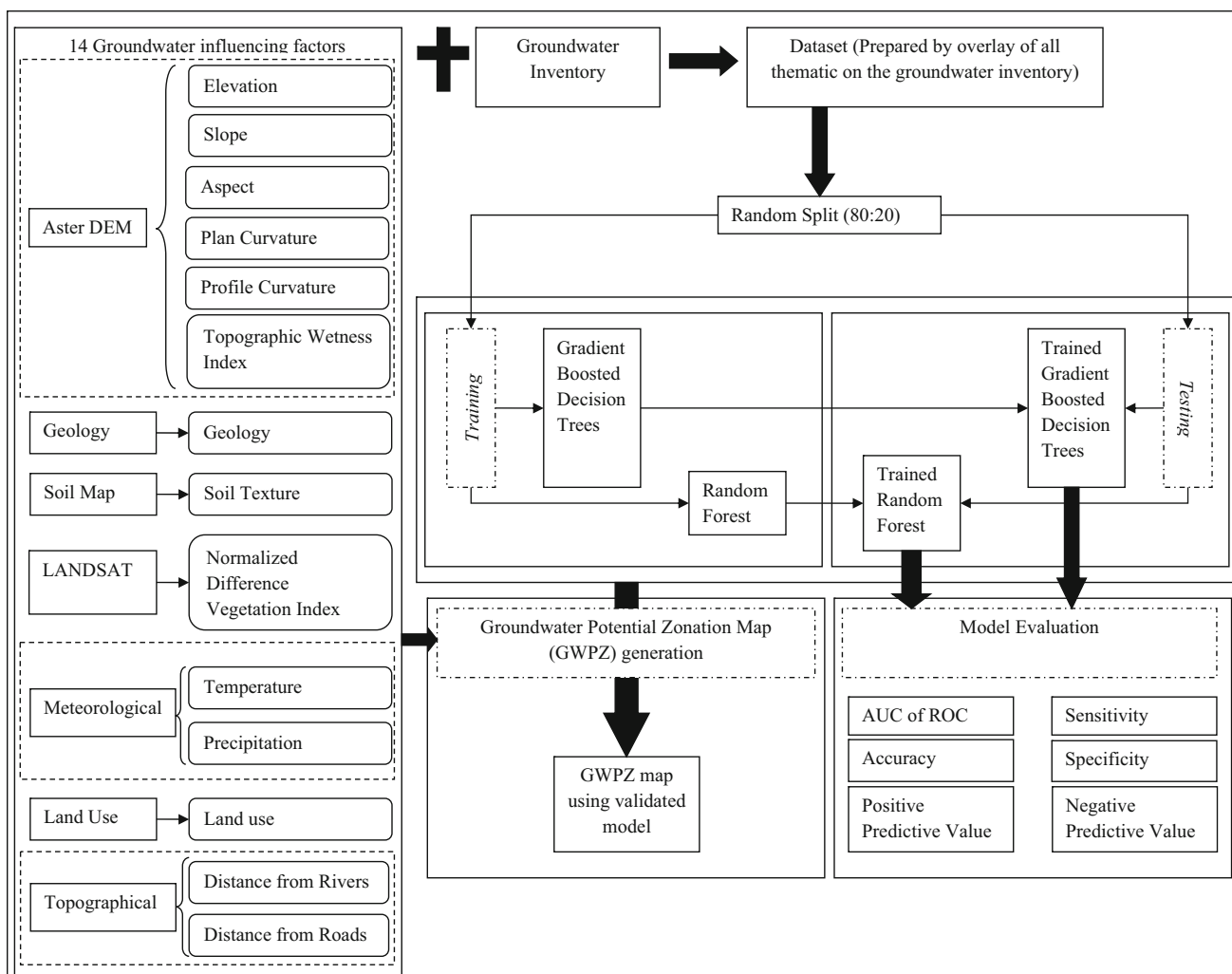


Fig. 4 Methodology

of GBDT and RF. The trained models were then used on the testing subsets to evaluate their respective performances. The split ratio of 80:20 was chosen after working with all other possible splits of 90:10, 80:20, 70:30, 60:40 and 50:50, and observing that the majority of machine learning models performed consistently well at 80:20 split whilst the other splits favored some models over the others and thus to avoid this bias, the split percentage of 80:20 was applied. The performance of GBDT and RF models at the aforementioned splits is depicted in Appendix (Table 6). The performances were compared on the performance metrics of accuracy, specificity, sensitivity, AUC etc. The validated model was then used to build the groundwater potential map for the region. Depending upon the potential assigned to a specific location in the study region by the trained model, the study area was divided into zones such as very low, low, moderate, high and very high groundwater potential zones. The GBDT and RF models as well as the performance metrics employed to

compare their performance have been described in the following sections.

4.1 Gradient boosted decision trees

Given a dataset for training the model, each training record i is of the form $\{y_i, \vec{x}_i\}$ where (y, \vec{x}) are both known, as they are part of the training set, $\vec{x} = \{x_1, x_2, x_3 \dots x_n\}$ are the attributes/features in the training set i.e. the groundwater influencing factors in our geospatial dataset, and the output or the goal is to predict the value for y based on x . This is achieved by identifying a mapping $F^*(\vec{x}) : \vec{x} \rightarrow y$ such that expected value of the loss function $\vartheta(y, F(\vec{x}))$ is minimized as shown in Eq. (3).

$$F^*(\vec{x}) = \arg \min_{F(\vec{x})} E_{y, \vec{x}} \vartheta(y, F(\vec{x})) \tag{3}$$

Here $F^*(\vec{x})$ is expanded as $F(\vec{x}) = \sum_{m=0}^M \beta_m \tau(\vec{x}; \vec{a}_m)$. Here $\tau(\vec{x}, \vec{a})$ depicts a base learner with $\vec{a} = \{a_1, a_2, \dots\}$ as

its parameters. In every stage of training, the expansion coefficients $\{\beta_m\}_0^M$ and the parameters of the base function $\{\mathbf{a}_m\}_0^M$ are simultaneously fit to make a better prediction. Initially, the training is started with a guess for $F_0(\vec{x})$ and then for $m = 1, 2, \dots, M$ we do the following steps iteratively (Eqs. 4 and 5).

$$(\beta_m, \vec{a}_m) = \arg \min_{\beta, \vec{a}} \sum_{i=1}^N \vartheta(y_i, F_{m-1}(\vec{x}_i) + \beta \tau(\vec{x}_i; \vec{a})) \tag{4}$$

and

$$F_m(\vec{x}) = F_{m-1}(\vec{x}) + \beta_m \tau(\vec{x}; \vec{a}_m) \tag{5}$$

Here the loss function $\vartheta(y, F(\vec{x}))$ is assumed to be differentiable and the function $\tau(\vec{x}; \vec{a})$ is fit by minimizing the k-class multinomial negative log likelihood (Friedman 2002). The optimal value of coefficient β_m can now be found out using the optimally fit $\tau(\vec{x}; \vec{a}_m)$ as in Eq. (6).

$$\beta_m = \arg \min_{\beta} \sum_{i=1}^N \vartheta(y_i, F_{m-1}(\vec{x}_i) + \beta \tau(\vec{x}_i; \vec{a}_m)) \tag{6}$$

The base learner $\tau(\vec{x}; \vec{a})$ is a decision tree where in each iteration m , the tree segments the input feature \vec{x} space into Z -disjoint regions $\{R_{zm}\}_{z=1}^Z$ and predicts a separate constant value in each one as in Eq. (7).

$$\tau(\vec{x}; \{R_{zm}\}_1^Z) = \sum_{z=1}^Z \ddot{y}_{zm} 1(\vec{x} \in R_{zm}) \tag{7}$$

Here \ddot{y} is the majority class predicted in each region R_{zm} i.e. the majority of the points in the R_{zm} region are predicted to be belonging to this class. Since the decision tree produces a constant value \ddot{y}_{zm} within each region R_{zm} , hence the expansion coefficients along with the base learner function's value can be deduced (Friedman 2002).

4.2 Random forest

The random forest ensemble machine learning model is actually a collection of decision trees, hence called a ‘forest’, and each decision tree is constructed from a subset of input records wherein the subset used is sampled randomly, thus justifying the ‘random’ in random forests (Breiman 2001). This works by drawing a bootstrap from the original dataset. Unlike the original decision tree model wherein, for each node of the decision tree, the splitting variable is chosen from among all the variables in the dataset, herein, the unpruned classification decision tree is built by making a choice for the splitting variable from a randomly sampled set of variables (Liaw and Wiener 2002) for the generated bootstrap. The size of the subset of variables, from which the final splitting variable for the

node is chosen from, is a parameter of the random forest classifier. N such unpruned classification trees are built, thus N is the second of the two parameters for the random forest ensemble. After having built N such trees as mentioned above, the predictions for the inputs are made by means of taking a majority vote of all classes predicted by all the generated trees. Gini index is used as the criteria for selecting the splitting variable, and it is a measure of the degree of impurity of the feature with respect to the output classes. In the training set $\vec{X}_{i=1}^N$ gini index for an input \vec{x}_i to belong to class C_i is given as:

$$\sum_{j \neq i} \left(\frac{\varphi(C_i, \vec{X})}{|N|} \right) \left(\frac{\varphi(C_j, \vec{X})}{|N|} \right) \tag{8}$$

Here $\left(\frac{\varphi(C_i, \vec{X})}{|N|} \right)$ is the probability of \vec{x}_i belonging to the class C_i .

4.3 Performance metrics

The performances of the two models applied in this study were compared on the basis of the following commonly used measures. Positive predictive value (Eq. 9) is the ratio of the number of points that were correctly classified as groundwater well to the total number of points that were all predicted as being groundwater. Negative predictive value (Eq. 10) is the ratio of the points that were correctly classified as non-groundwater to the total number of points that were classified as non-groundwater. Sensitivity (Eq. 11) is the ratio of the number of points that were correctly classified as groundwater to the total number of points that were originally known to be groundwater. Specificity (Eq. 12) is the ratio of the number of points that were correctly classified as non-groundwater to the total number of points that were originally known to be non-groundwater. Accuracy (Eq. 13) is the ratio of total number of points that were correctly classified to the total number of points.

$$\text{Positive Predictive Value} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN} \tag{10}$$

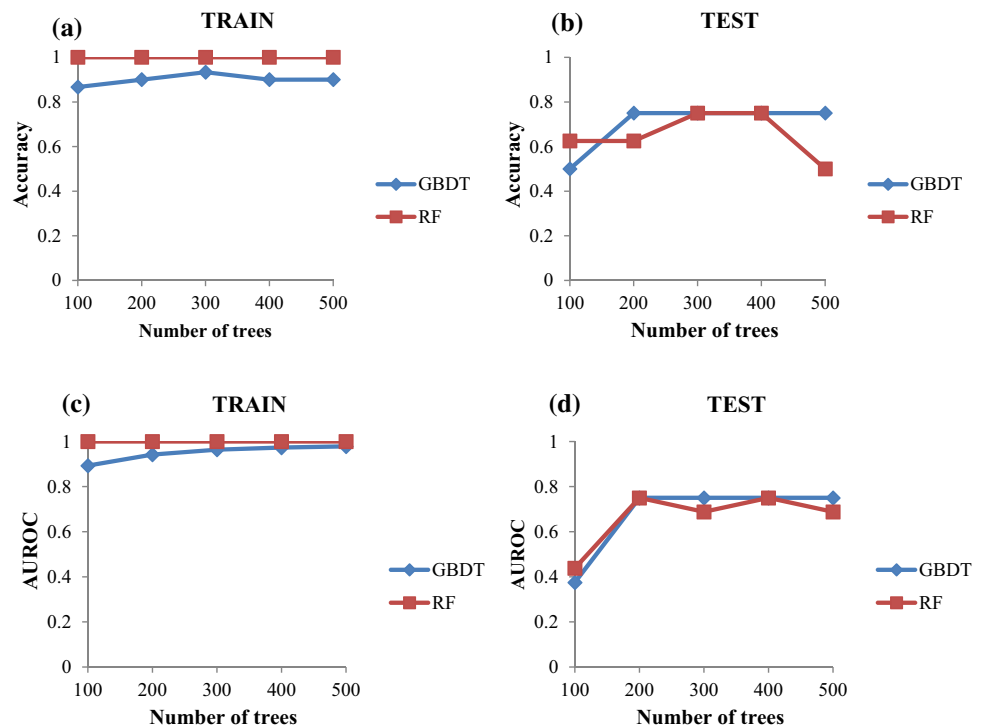
$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Here TP, TN, FP and FN are true positives, true negatives, false positives and false negatives respectively. The

Fig. 5 **a, b** Trend in the accuracy of GBDT and RF with increasing value of parameter “no. of trees” in the training and testing phase of the respective models. **c, d** Trend in the AUROC of GBDT and RF with increasing value of parameter “no. of trees” in the training and testing phase of the respective models



GBDT and RF. Thus the performance metrics for GBDT and RF were compared at values of 100, 200, 300, 400 and 500 for this parameter.

It was observed that RF achieved an almost perfect accuracy in the training phase at all values of the parameter. While GBDT had an accuracy of 0.87, when no. of trees were fixed at 100, but it improved there on after, increasing to 0.93 when no. of trees were increased to 300 as can be seen from Fig. 5a. In the testing phase (Fig. 5b), however, the GBDT performed better than RF, with RF achieving a peak 0.75 at 400 trees and there after it goes on a decline, while GBDT giving the same best accuracy of 0.75 achieved at an early peak of 200 trees and performs consistently well thereafter too. Hence, in terms of accuracy, while RF fared better than GBDT, in the training phase, the situation was reversed during the test phase, with GBDT performing better than RF. The predictive performance for both the models in the test phase declined from their respective performance levels at the time of training. Nevertheless, both models displayed satisfactory magnitudes of accuracy to be considered as effective predictors in the current study.

AUROC witnessed a trend similar to the one achieved by accuracy in the training and testing phases as can be deduced from Fig. 5c and d, with both GBDT and RF performing uniformly well at almost all values for no. of trees in the training phase by attaining an almost ideal AUC. In the testing phase, GBDT performed slightly better than RF, giving results similar to those for accuracy. With

the no. of trees fixed at 200, both GBDT and RF, produced an AUC of 0.75, and there on after, both models are witnessed retaining relative stagnancy on the AUC scale. All in all, the results corroborate both GBDT and RF as good performers in the undertaken groundwater potential analysis.

Also, it was deduced that the parameter, no. of trees tuned to the value of 300, gave optimal results, as both models achieved their peak performances by that point and retained those performances thereafter, thereby remaining consistently good predictors and not showing any more significant improvement to the increase in no. of trees. The ROC curves for GBDT and RF in the training and testing phase with parameter “no. of trees” fixed at the value 300 are depicted in Fig. 6. Generally, with machine learning models, it cannot be hypothesized with absolute certainty that the predictor with the lowest training error would also lead to the lowest error at the testing phase. In other words, it cannot be laid with high confidence that among a set of classifiers, the classifier that performs the best, during training, on the basis of any particular performance statistic such as accuracy, would inevitably turn out to be the best at the time of testing as well. This uncertainty is introduced due to bias variance trade-off where bias (error introduced due to model’s simplification of learning via various assumptions) and variance (error owing to the influence of training set on model’s learning) are the main components of the prediction error. This is especially true in case of certain models, where the model is trained in a

characteristic way so as to be a custom fit for the particular training data in question, which might lead to the model performance being worse in the testing phase as the model had not been exposed to the test data and the training set happened to be widely different from the data on which the model is being tested on.

Here specifically, the RF performed well and even better than GBDT during the training phase, but the situation was reversed at the time of testing. The reasoning for this behavior can again be chalked up to the variance bias trade-off explained as follows. Both GBDT and RF are ensemble techniques, namely gradient boosting and bagging respectively, based on decision trees (Alam et al. 2019). With GBDT, a new decision tree is generated by enhancing the performance of the tree generated in the previous iteration. Gradient boosting reduces error by mainly reducing the bias component of the model. The variance is also reduced to some extent owing to the aggregation of outputs from many previous models which themselves had relatively low variance owing to pruning of decision tree over successive iterations in gradient boosting. On the other hand, RF reduces error during training by focusing on the variance component of the error. This is possible here because the generated decision trees are deliberately made uncorrelated as they are constructed from a subset of input records wherein the subset used is sampled randomly. The decrease in correlation maximizes the reduction in variance successively. Thus, both RF and GBDT tend to decrease the variance, however, boosting also improves the bias. The impact of these error reductions is witnessed in the testing phase with the models being validated on data that was not exposed to the model during their development in training.

In order to avoid uncertainty in the results and validate the models on the entirety of the dataset for an impartial comparison of RF and GBDT models, a 5 fold cross validation was undertaken. The dataset was divided into 5 folds: with one fold being used for validation while, the remaining 4 folds were employed for training the models.

The process is repeated 5 times for achieving test results on every fold. The results are summarized in Table 3.

It can thus be deduced from Table 3, that GBDT outperforms RF by achieving an accuracy of approximately 74% and an AUC of 0.79, while RF attained an average accuracy of 59% with an AUC of 0.71.

5.2 Relative importance of influencing factors

The ranking for the groundwater conditioning factors in the order of their relative influence to the presence of groundwater by the GBDT model are listed in Table 4.

The results indicate that Profile curvature, Distance from rivers and NDVI are important factors in the current groundwater potential mapping. The influencing factors for the groundwater wells along with their respective groundwater potentials are listed in Table 5.

5.3 Generation of groundwater potential map

The trained GBDT model was used to generate the groundwater potential zonation map for the study region as shown in Fig. 7.

The zones were delineated on the basis of the range of potential they belonged to. The five classes of groundwater potential, namely very low, low, moderate, high and very high were generated using the quantile classification scheme (Tehrany et al. 2015). The region with groundwater potential in the range of 0–0.51 was classified as belonging to “very low” potential zone, while the “low” zone has potential in the range of 0.51–0.57. The regions with potential in the ranges of 0.57–0.63 and 0.63–0.66 were delineated as “moderate” and “high” potential zone respectively. Finally, “very high” potential zone was delineated for potential values greater than 0.66. It was found that approximately 20.2% of the region has very high potential for yielding groundwater, while 22.6% of the study area has high groundwater potential. 19.9 and 17.5% of the region was found to have very low to low

Fig. 6 a, b ROC curve for GBDT and RF generated during training and testing

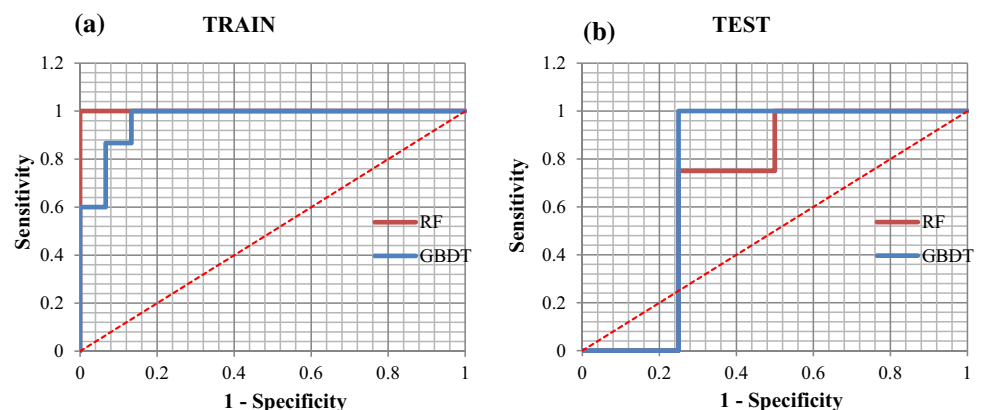


Table 3 Performance statistics for model validation using 5 fold cross-validation

FOLD	RF																										
	GBDT										RF																
	TRAIN					TEST					TRAIN					TEST											
1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	μ		
TP	15	15	14	15	14	-	4	2	3	3	3	3	3	3	3	15	16	15	16	14	-	3	2	4	2	2	-
TN	11	12	13	15	16	-	1	3	4	3	2	2	3	3	2	15	15	15	15	16	-	1	3	1	3	1	-
FP	4	3	2	0	0	-	3	1	0	1	1	1	0	0	0	0	0	0	0	0	0	3	1	3	1	2	-
FN	0	1	1	1	0	-	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	1	1	0	1	3	-
PPV	0.79	0.83	0.88	1.00	1.00	0.90	0.57	0.67	1.00	0.75	0.75	0.75	0.75	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.67	0.57	0.67	0.50	0.58
NPV	1.00	0.92	0.93	0.94	1.00	0.96	1.00	0.75	0.80	1.00	0.50	0.81	0.81	0.81	0.81	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.75	1.00	0.75	0.25	0.65
SST	1.00	0.94	0.93	0.94	1.00	0.96	1.00	0.67	0.75	1.00	0.60	0.80	0.80	0.80	0.80	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.67	1.00	0.67	1.00	0.70
SPC	0.73	0.80	0.87	1.00	1.00	0.88	0.25	0.75	1.00	0.75	0.67	0.68	0.68	0.68	0.68	1.00	1.00	1.00	1.00	1.00	1.00	0.25	0.75	0.25	0.75	0.33	0.47
ACC	0.87	0.87	0.90	0.97	1.00	0.92	0.63	0.71	0.88	0.86	0.63	0.74	0.74	0.74	0.74	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.71	0.63	0.71	0.38	0.59
AUC	0.94	0.94	0.97	0.98	1	0.97	0.62	0.92	0.94	1	0.47	0.79	0.79	0.79	0.79	1	1	1	1	1	1	0.44	0.92	0.87	0.83	0.47	0.71

TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; PPV: positive predictive value; NPV: negative predictive value; SST: sensitivity; SPC: specificity; ACC: accuracy; AUC: area under curve of receiver operating characteristics (AUROC); μ: average

Table 4 Importance of groundwater influencing factors by GBDT

Influencing factor	Variable importance
Profile curvature	13.3
Distance from rivers	12.3
NDVI	11.2
TWI	10.2
Plan curvature	9.8
Temperature	7.2
LULC	6.9
Precipitation	6.0
Elevation	5.9
Slope	5.1
Geology	4.3
Distance from roads	4.1
Aspect	3.3
Soil	0.4

groundwater potential. It can be observed that low to very low potential zones are concentrated around the southern, southwestern and central parts of the region. Such evaluations can help augment efforts for groundwater management through measured abstractions, land use planning, site identification for artificial recharge, etc.

5.4 Comparison with related work

A variety of groundwater potential delineation mappings have been carried out around the world, particularly in the arid and semi-arid climate zones that are frequently at the receiving end of recurring droughts threatening the local water, food and energy security of the region (Moghaddam et al. 2020). For instance, Rahmati et al. (2015) employed the integrated analytical hierarchy process for delineating the groundwater resource potential zones in the Kurdistan plain of Iran (Rahmati et al. 2015). They employed five groundwater conditioning factors, namely slope, rainfall, lithology, drainage and lineament density and upon model validation they achieved an AUC of 0.7366 and hence showed reasonably good accuracy in predicting the groundwater potential. Another similar study that employed the conventional geospatial mapping techniques was undertaken at the Varamin Plain, Tehran province, Iran (Razandi et al. 2015). In this study, the standard approaches of analytical hierarchy process, frequency ratio, and certainty factor models were applied for mapping the groundwater potential in the aforementioned region that also lies in an arid climate zone. On quantitative validation, the work derived the AUC for all the three applied models as frequency ratio (77.55%), analytical hierarchy process (73.47%), and certainty factor (65.08%).

Table 5 Groundwater influencing factors and potentials for the groundwater wells in the inventory

Latitude (N)	Longitude (E)	DEM (m)	Slope (°)	Aspect	Profile curvature	Plan curvature	TWI	Geology	Soil	Precipitation (mm)	Temperature (°C)	NDVI	LULC	Distance from rivers (m)	Distance from roads (m)	GWP
26° 56' 00"	78° 08' 15"	157	1.05	89.4/E	- 0.0014	0.0000	4.68	QS	SC	470.92	26.04	0.2146	CR	1655.92	872.51	0.94
26° 41' 00"	77° 52' 00"	189	2.30	153.9/SE	0.0005	0.0195	12.41	PCR	SC	499.16	26.07	0.3175	CR	5298.64	301.92	0.90
26° 54' 10"	77° 56' 00"	163	2.01	179.9/S	- 0.0026	0.0001	11.99	QS	SC	474.78	26.01	0.1133	CR	215.46	711.04	0.88
26° 48' 00"	78° 02' 00"	173	1.44	314.8/NW	0.0013	0.0525	2.51	QS	SC	478.64	26.02	0.2727	CR	6251.98	404.45	0.87
26° 31' 20"	77° 23' 15"	249	2.31	332.7/NW	- 0.0001	- 0.0338	11.85	PCR	SC	483.84	25.82	0.2544	CR	3251.27	227.87	0.86
26° 53' 00"	77° 47' 00"	176	2.95	45.9/NE	- 0.0025	- 0.0067	10.94	QS	SC	457.95	26.01	0.1887	CR	6455.86	4053.43	0.82
26° 48' 30"	77° 44' 45"	182	3.35	341.6/N	0.0007	- 0.0117	6.85	QS	SC	463.11	26.00	0.1164	CR	69.63	955.18	0.81
26° 34' 30"	77° 41' 30"	291	7.08	116.6/SE	- 0.0024	- 0.0133	10.82	PCR	SC	488.61	25.91	0.1571	CR	2445.45	492.82	0.81
26° 49' 00"	77° 49' 20"	178	1.49	44.7/NE	- 0.0011	- 0.0652	11.52	QS	SC	474.19	26.01	0.2056	CR	2411.84	3274.20	0.79
26° 49' 00"	77° 41' 10"	187	1.50	44.9/NE	- 0.0004	0.0133	1.94	QS	SC	460.44	26.00	0.1390	CR	1574.95	1502.62	0.79
26° 50' 50"	78° 05' 30"	166	2.98	314.9/NW	- 0.0014	0.0260	11.06	QS	SC	473.67	26.03	0.1425	CR	2899.00	1264.98	0.77
26° 44' 45"	77° 32' 45"	201	0.01	67.5/E	0.0000	0.0001	11.31	PCR	SC	483.63	25.91	0.1157	CR	1212.92	251.33	0.76
26° 45' 40"	78° 01' 40"	175	4.24	270.0/W	0.0000	0.0000	4.41	QS	SC	494.94	26.07	0.1853	CR	3617.28	2324.78	0.73
26° 49' 00"	77° 57' 30"	172	5.04	180.0/S	- 0.0017	0.0000	3.49	QS	SC	477.81	26.01	0.2599	CR	3739.68	4808.01	0.73
26° 46' 10"	77° 35' 50"	196	1.06	360.0/N	0.0014	0.0000	2.41	PCR	SC	487.97	25.90	0.1680	CR	3505.71	1862.21	0.64
26° 51' 30"	77° 43' 00"	180	0.69	180.6/S	- 0.0009	0.0007	6.68	QS	SC	455.20	26.01	0.1422	CR	2905.99	4510.42	0.63
26° 39' 00"	77° 48' 00"	235	2.12	270.0/W	0.0000	0.0741	10.26	PCR	SC	494.63	26.01	0.1318	CR	2663.02	291.81	0.63
26° 40' 00"	77° 37' 00"	210	3.82	56.3/NE	0.0027	0.0616	5.19	QS	SC	486.58	25.91	0.1435	CR	1466.86	1762.05	0.61
26° 36' 07"	77° 28' 36"	227	0.99	0.58/N	- 0.0013	0.0000	3.83	QS	SC	484.72	25.84	0.1360	CR	4137.03	170.70	0.57

CR: Cropland rain fed; SC: Sandy clayey; QS: Quaternary sediments; PCR: Precambrian rocks; GWP: Groundwater potential

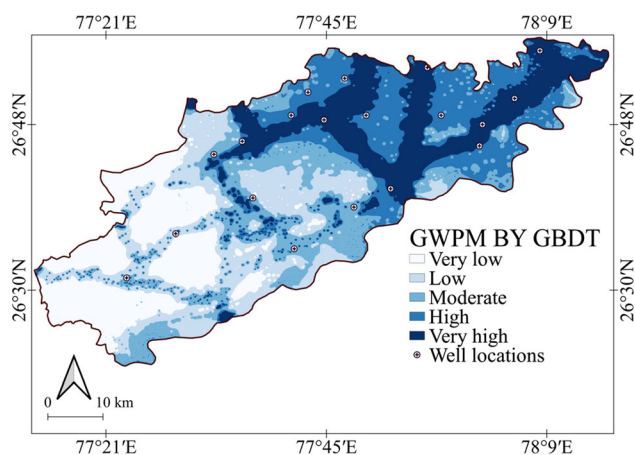


Fig. 7 Groundwater potential zonation map constructed by GBDT

While the above mentioned studies showed acceptable prediction accuracies, they also divulged a great scope for improvement. The current study indicated that the GBDT achieved a superior performance to RF by attaining an AUC of 0.79 to RF's AUC of 0.71. These results clearly indicate the superiority of GBDT's performance over aforementioned approaches of analytical hierarchy process (AUC: 0.7366 (Rahmati et al. 2015); AUC: 0.7347 (Razandi et al. 2015)), frequency ratio (AUC: 0.77), certainty factor (AUC: 0.65) etc. These evaluations reinforce the competency of tree based ensembles over the traditionally employed conventional statistical techniques. These results can be generalized well over to the competency of machine learning models over their statistical counterparts.

Naghibi et al. (2017, 2018), Zabihi et al. (2016), Miraki et al. (2019) stated the decision tree based ensembles such as random forest and boosted trees are very useful in complex decision making problems such as resource potential mappings. Naghibi et al. 2017 proved that RF and a genetically optimized RF (AUC: 85.6%) are better predictors than support vector machine models paired with 4 different kernels (linear, polynomial, sigmoid, and radial based). On the other hand, Miraki et al. 2019 concluded the superiority of RF models over Naïve Bayes and Logistic regression models. Zabihi et al. 2016 compared the models of Multivariate adaptive regression splines and RF and concluded them to be both equally efficient (AUC: 0.79) in groundwater potential mapping undertaken in Iran. The results of the current study are in alignment with the results of (Naghibi et al.2016). They used the machine learning

models: boosted regression trees, decision trees and RF to produce the potential maps for the Koohrang watershed, Iran. The boosted regression tree model performed better than the other two and achieved an AUC of 0.8103. The findings from the current study, thus concur with those mentioned above, over the fact that tree based ensembles in general and GBDT and RF in particular are equally and often times more competent than other machine learning models such as logistic regression, support vector machine, etc.

Few of the geospatial mappings around the world have also compared the performances of GBDT and RF and found varying verdicts about their predictions such as mentioned forth. In contrast to current study's finding, Lee et al. 2017, on comparing GBDT and RF for a flood susceptibility mapping in Seoul, Korea, found RF's (AUC: 0.79) performance, taking the lead over that of GBDT's (AUC: 0.77) although with a smaller margin. On the other hand, for a landslide susceptibility mapping undertaken at Pyeong-Chang, Korea by Kim et al. 2018, displayed GBDT (AUC: 0.85) having an upper hand over RF (AUC: 0.79). Since geospatial mappings are localized and the results from each case study are region specific hence, it is difficult to make generalizations over the absolute supremacy of a model's performance over all its counterparts especially when the margin between their respective metrics is not significant. Additionally, the scalability aspect of all models as probable candidates should be explored further in research studies over datasets larger than the one employed in the current study. However, the assessments from this study coupled with those of the literature, it can be stated with significant confidence that GBDT is amongst the most competent tree based ensemble models and should henceforth be used as a benchmark for evaluating predictive aptitude in the field of geospatial analytics.

6 Conclusion

Water in its one of its most pristine form is naturally provided to us by earth's own underground reservoirs as confined and unconfined aquifers. It's not coincidental that the majority of human needs for drinking water and other activities are met by this natural resource. Although, groundwater is naturally replenishable, however, due to increasing population, and with the increase in the unsustainable activities of the said population, the nature is

unable to replenish the resources at the pace at which they are being consumed. This has put our groundwater springs and wells under tremendous strain. The desert state of India, Rajasthan has become a continuing witness of such circumstances. This study was undertaken in the district of Dholpur with the aim to develop a reliable groundwater potential map for the region. For this purpose, the data for 14 groundwater influencing factors specific to the region were collected and then were merged with the groundwater inventory in the region. The compiled dataset generated was then randomly split in the ratio of 80:20 for training and testing purposes. The training subset of the data set was used to train the competing ensemble models of GBDT and RF. The trained models were then used on the testing subsets to evaluate and compare their respective performances. It was found through the study that RF performed better at the training phase, but GBDT fared better at testing. Also, it was found that both RF and GBDT reached their peak performance when the parameter “number of tree” was tuned to a value of 300. On a 5 fold cross-validation, GBDT achieved an average accuracy of 74% and an AUROC of 0.79, while RF attained an accuracy of 59% and AUROC of 0.71. The validated GBDT model was used to generate the region’s groundwater potential map, which revealed that approximately 19.9% of the study region has low groundwater potential, while 20.2 and 22.6% span of the area fell into the categories of very high and high potential respectively. Such a study holds promise for further such endeavors in this field and prepares government agencies and other concerned parties for necessary action that need to be taken to counteract the damage that has already been done and prevent any more in the future.

Appendix

The performances of RF and GBDT models at various dataset splits of (50/50, 60/40, 70/30, 80/20, and 90/10) are summarized in Table 6.

From Table 6, it can be observed that during testing, the best performance for both the models is delivered when 80% of the dataset were being used for training, i.e. developing the model, while the rest 20% of the dataset is used for testing purpose. Since a classifier’s true performance is gauged by how well the model performs on unseen data, hence the accuracy attained by the model at testing phase is taken as the true metric for determining the ratio for splitting the dataset. Hence, the ratio of 80:20 was employed for generating training and testing data sets.

Table 6 Performance statistics for various training–testing splits

Split	50:50				60:40				70:30				80:20				90:10				
	Train		Test		Train		Test		Train		Test		Train		Test		Train		Test		
	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	GBDT	RF	
TP	5	10	2	6	9	11	4	6	13	13	13	3	4	13	15	1	3	17	17	2	2
TN	10	10	8	1	8	11	2	0	11	13	3	3	0	13	15	3	2	15	17	0	0
FP	0	0	1	8	3	0	6	8	2	0	3	3	6	2	0	1	2	2	0	2	2
FN	5	0	7	3	2	0	4	2	0	0	3	3	2	2	0	3	1	0	0	0	0
PPV	1	1	0.67	0.43	0.75	1	0.4	0.43	0.87	1	0.5	0.5	0.4	0.87	1	0.5	0.6	0.89	1	0.5	0.5
NPV	0.67	1	0.53	0.25	0.8	1	0.33	0	1	1	0.5	0.5	0	0.87	1	0.5	0.67	1	1	1	–
SST	0.5	1	0.22	0.67	0.82	1	0.5	0.75	1	1	0.5	0.5	0.67	0.87	1	0.25	0.75	1	1	1	1
SPC	1	1	0.89	0.11	0.73	1	0.25	0	0.85	1	0.5	0.5	0	0.87	1	0.75	0.5	0.88	1	0	0
ACC	0.75	1	0.56	0.39	0.77	1	0.375	0.375	0.92	1	0.5	0.5	0.33	0.87	1	0.5	0.625	0.94	1	0.5	0.5

TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; PPV: positive predictive value; NPV: negative predictive value; SST: sensitivity; SPC: specificity; ACC: accuracy

References

- Abedi Gheshlaghi H, Feizizadeh B, Blaschke T (2020) GIS-based forest fire risk mapping using the analytical network process and fuzzy logic. *J Environ Plan Manag* 63(3):481–499
- Al-Abadi AM, Shahid S (2015) A comparison between index of entropy and catastrophe theory methods for mapping groundwater potential in an arid region. *Environ Monit Assess* 187(9):576
- Alam MZ, Rahman MS, Rahman MS (2019) A Random Forest based predictor for medical data classification using feature ranking. *Inform Med Unlocked* 15:100180
- Althuwaynee OF, Pradhan B, Lee S (2012) Application of an evidential belief function model in landslide susceptibility mapping. *Comput Geosci* 44:120–135
- Althuwaynee OF, Pradhan B, Park HJ, Lee JH (2014) A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. *CATENA* 114:21–36
- Arabameri A, Pradhan B, Rezaei K, Sohrabi M, Kalantari Z (2019a) GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms. *J Mt Sci* 16(3):595–618
- Arabameri A, Pradhan B, Lombardo L (2019b) Comparative assessment using boosted regression trees, binary logistic regression, frequency ratio and numerical risk factor for gully erosion susceptibility modelling. *CATENA* 183:104223
- Avand M, Janizadeh S, Naghibi SA, Pourghasemi HR, Khosrobeigi Bozchaloei S, Blaschke T (2019) A comparative assessment of Random Forest and k-Nearest Neighbor classifiers for gully erosion susceptibility mapping. *Water* 11(10):2076
- Banks D, Robins N, Robins N (2002) An introduction to groundwater in crystalline bedrock. *Norges geologiske undersøkelse, Trondheim*
- Beaudoin A, Bernier PY, Guindon L, Villemaire P, Guo XJ, Stinson G, Hall RJ (2014) Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. *Can J For Res* 44(5):521–532
- Bragagnolo L, da Silva RV, Grzybowski JMV (2020a) Artificial neural network ensembles applied to the mapping of landslide susceptibility. *CATENA* 184:104240
- Bragagnolo L, da Silva RV, Grzybowski JMV (2020b) Landslide susceptibility mapping with r landslides: a free open-source GIS-integrated tool based on Artificial Neural Networks. *Environ Model Softw* 123:104565
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Bui QT, Nguyen QH, Nguyen XL, Pham VD, Nguyen HD, Pham VM (2020) Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. *J Hydrol* 581:124379
- Carranza EJM, Hale M (2003) Evidential belief functions for data-driven geologically constrained mapping of gold potential, Baguio district, Philippines. *Ore Geol Rev* 22(1–2):117–132
- Central Ground Water Board (CGWB), Ministry of Jal Shakti, Department of Water Resources, River Development and Ganga Rejuvenation, Government of India, Assessment of Ground Water (2018). <http://cgwb.gov.in/>. Accessed 18 Jan 2020
- Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Ma J (2017) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA* 151:147–160
- Chen W, Zhang S, Li R, Shahabi H (2018) Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Sci Total Environ* 644:1006–1018
- Chen J, Li Q, Wang H, Deng M (2020a) A machine learning ensemble approach based on random forest and radial basis function neural network for risk evaluation of regional flood disaster: a case study of the Yangtze River Delta, China. *Int J Environ Res Public Health* 17(1):49
- Chen W, Li Y, Xue W, Shahabi H, Li S, Hong H, Ahmad BB (2020b) Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. *Sci Total Environ* 701:134979
- Choubin B, Moradi E, Golshan M, Adamowski J, Sajedi-Hosseini F, Mosavi A (2019) An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Sci Total Environ* 651:2087–2096
- Çolak E, Sunar F (2020) Evaluation of forest fire risk in the Mediterranean Turkish forests: a case study of Menderes region, Izmir. *Int J Disaster Risk Reduct* 45:101479
- Corsini A, Cervi F, Ronchetti F (2009) Weight of evidence and artificial neural networks for potential groundwater spring mapping: an application to the Mt. Modino area (Northern Apennines, Italy). *Geomorphology* 111(1–2):79–87
- Costache R, Bui DT (2020) Identification of areas prone to flash-flood phenomena using multiple-criteria decision-making, bivariate statistics, machine learning and their ensembles. *Sci Total Environ* 712:136492
- de Quadros TF, Koppe JC, Strieder AJ, Costa JF (2006) Mineral-potential mapping: a comparison of weights-of-evidence and fuzzy methods. *Nat Resour Res* 15(1):49–65
- Díaz-Alcaide S, Martínez-Santos P (2019) Advances in groundwater potential mapping. *Hydrogeol J* 27(7):2307–2324
- Dou J, Yunus AP, Bui DT, Merghadi A, Sahana M, Zhu Z, Pham BT (2019) Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci Total Environ* 662:332–346
- Feloni E, Mousadis I, Baltas E (2020) Flood vulnerability assessment using a GIS-based multi-criteria approach—the case of Attica region. *J Flood Risk Manag* 13:e12563
- Feng B, Wang J, Zhang Y, Hall B, Zeng C (2020) Urban flood hazard mapping using a hydraulic-GIS combined model. *Nat Hazards* 100:1089–1104
- Fitts CR (2002) *Groundwater science*. Elsevier, Amsterdam
- Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data Anal* 38(4):367–378
- Garosi Y, Shekhabadi M, Pourghasemi HR, Besaltpour AA, Conoscenti C, Van Oost K (2018) Comparison of differences in resolution and sources of controlling factors for gully erosion susceptibility mapping. *Geoderma* 330:65–78
- Gayen A, Pourghasemi HR, Saha S, Keesstra S, Bai S (2019) Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Sci Total Environ* 668:124–138
- Gjertsen AK (2007) Accuracy of forest mapping based on Landsat TM data and a kNN-based method. *Remote Sens Environ* 110(4):420–430
- Hosseinalizadeh M, Kariminejad N, Chen W, Pourghasemi HR, Alinejad M, Behbahani AM, Tiefenbacher JP (2019) Gully headcut susceptibility modeling using functional trees, naïve Bayes tree, and random forest models. *Geoderma* 342:1–11
- Hu Q, Zhou Y, Wang S, Wang F (2020) Machine learning and fractal theory models for landslide susceptibility mapping: case study from the Jinsha River Basin. *Geomorphology* 351:106975
- Jha MK, Chowdhury A, Chowdary VM, Peiffer S (2007) Groundwater management and development by integrated remote

- sensing and geographic information systems: prospects and constraints. *Water Resour Manag* 21(2):427–467
- Kaur L, Rishi MS, Singh G, Thakur SN (2020) Groundwater potential assessment of an alluvial aquifer in Yamuna sub-basin (Panipat region) using remote sensing and GIS techniques in conjunction with analytical hierarchy process (AHP) and catastrophe theory (CT). *Ecol Ind* 110:105850
- Kayastha P, Dhital MR, De Smedt F (2012) Landslide susceptibility mapping using the weight of evidence method in the Tinau watershed, Nepal. *Nat Hazards* 63(2):479–498
- Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Bui DT (2018) A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Sci Total Environ* 627:744–755
- Kim JC, Lee S, Jung HS, Lee S (2018) Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. *Geocarto Int* 33(9):1000–1015
- Kuhnert PM, Henderson AK, Bartley R, Herr A (2010) Incorporating uncertainty in gully erosion calculations using the random forests modelling approach. *Environmetrics* 21(5):493–509
- Lee S, Choi J (2004) Landslide susceptibility mapping using GIS and the weight-of-evidence model. *Int J Geogr Inf Sci* 18(8):789–814
- Lee S, Pradhan B (2007) Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* 4(1):33–41
- Lee S, Song KY, Kim Y, Park I (2012) Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model. *Hydrogeol J* 20(8):1511–1527
- Lee S, Kim JC, Jung HS, Lee MJ, Lee S (2017) Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomat Nat Hazards Risk* 8(2):1185–1203
- Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2(3):18–22
- Lombardo L, Cama M, Conoscenti C, Märker M, Rotigliano EJNH (2015) Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). *Nat Hazards* 79(3):1621–1648
- Mastere M (2020) Mass movement hazard assessment at a medium scale using weight of evidence model and neo-predictive variables creation. In: *Mapping and spatial analysis of socio-economic and environmental indicators for sustainable development*, pp 73–85. Springer, Cham
- Miraki S, Zanganeh SH, Chapi K, Singh VP, Shirzadi A, Shahabi H, Pham BT (2019) Mapping groundwater potential using a novel hybrid intelligence approach. *Water Resour Manag* 33(1):281–302
- Mishra K, Sinha R (2020) Flood risk assessment in the Kosi megafan using multi-criteria decision analysis: a hydro-geomorphic approach. *Geomorphology* 350:106861
- Moghaddam DD, Rahmati O, Panahi M, Tiefenbacher J, Darabi H, Haghizadeh A, Bui DT (2020) The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *CATENA* 187:104421
- Mukherjee P, Singh CK, Mukherjee S (2012) Delineation of groundwater potential zones in arid region of India—a remote sensing and GIS approach. *Water Resour Manag* 26(9):2643–2672
- Naghibi SA, Pourghasemi HR, Pourtaghi ZS, Rezaei A (2015) Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed, Iran. *Earth Sci Inform* 8(1):171–186
- Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ Monit Assess* 188(1):44
- Naghibi SA, Ahmadi K, Daneshi A (2017) Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resour Manag* 31(9):2761–2775
- Naghibi SA, Pourghasemi HR, Abbaspour K (2018) A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. *Theoret Appl Climatol* 131(3–4):967–984
- Nampak H, Pradhan B, Manap MA (2014) Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *J Hydrol* 513:283–300
- Ozdemir A (2011) GIS-based groundwater spring potential mapping in the Sultan Mountains (Konya, Turkey) using frequency ratio, weights of evidence and logistic regression methods and their comparison. *J Hydrol* 411(3–4):290–308
- Pham BT, Jaafari A, Prakash I, Singh SK, Quoc NK, Bui DT (2019) Hybrid computational intelligence models for groundwater potential mapping. *CATENA* 182:104101
- Porwal A, Carranza EJM, Hale M (2006) Bayesian network classifiers for mineral potential mapping. *Comput Geosci* 32(1):1–16
- Pourghasemi HR, Pradhan B, Gokceoglu C (2012) Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. *Nat Hazards* 63(2):965–996
- Pourghasemi HR, Termeh SVR, Kariminejad N, Hong H, Chen W (2020) An assessment of metaheuristic approaches for flood assessment. *J Hydrol* 582:124536
- Pradhan B (2013) A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput Geosci* 51:350–365
- Rahimi I, Azeez SN, Ahmed IH (2020) Mapping forest-fire potentiality using remote sensing and GIS, case study: Kurdistan Region-Iraq. In: *Environmental remote sensing and GIS in Iraq*, pp 499–513. Springer, Cham
- Rahmati O, Samani AN, Mahdavi M, Pourghasemi HR, Zeinivand H (2015) Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arab J Geosci* 8(9):7059–7071
- Rahmati O, Pourghasemi HR, Zeinivand H (2016) Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golestan Province, Iran. *Geocarto Int* 31(1):42–70
- Razandi Y, Pourghasemi HR, Neisani NS, Rahmati O (2015) Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Sci Inf* 8(4):867–883
- Rodriguez-Galiano V, Chica-Olmo M (2012) Land cover change analysis of a Mediterranean area in Spain using different sources of data: multi-seasonal Landsat images, land surface temperature, digital terrain models and texture. *Appl Geogr* 35(1–2):208–218
- Sameen MI, Sarkar R, Pradhan B, Drukpa D, Alamri AM, Park HJ (2020) Landslide spatial modelling using unsupervised factor optimisation and regularised greedy forests. *Comput Geosci* 134:104336
- Sander P, Chesley MM, Minor TB (1996) Groundwater assessment using remote sensing and GIS in a rural groundwater project in Ghana: lessons learned. *Hydrogeol J* 4(3):40–49
- Sansare DA, Mhaske SY (2020) Natural hazard assessment and mapping using remote sensing and QGIS tools for Mumbai city, India. *Nat Hazards* 100:1117–1136

- Sarkar D, Mondal P (2020) Flood vulnerability mapping using frequency ratio (FR) model: a case study on Kulik river basin, Indo-Bangladesh Barind region. *Appl Water Sci* 10(1):17
- Sevinc V, Kucuk O, Goltas M (2020) A Bayesian network model for prediction and analysis of possible forest fire causes. *For Ecol Manag* 457:117723
- Tang RX, Kulatilake PH, Yan EC, Cai JS (2020) Evaluating landslide susceptibility based on cluster analysis, probabilistic methods, and artificial neural networks. *Bull Eng Geol Environ* 79:2235–2254. <https://doi.org/10.1007/s10064-019-01684-y>
- Tehrany MS, Pradhan B, Jebur MN (2015) Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stoch Environ Res Risk Assess* 29(4):1149–1165
- Thai Pham B, Tien Bui D, Prakash I (2018) Landslide susceptibility modelling using different advanced decision trees methods. *Civ Eng Environ Syst* 35(1–4):139–157
- Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012) Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes Models. *Math Probl Eng* 2012:974638. <https://doi.org/10.1155/2012/974638>
- Van Dao D, Jaafari A, Bayat M, Mafi-Gholami D, Qi C, Moayedi H, Luu C (2020) A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *CATENA* 188:104451
- Venkatesh K, Preethi K, Ramesh H (2020) Evaluating the effects of forest fire on water balance using fire susceptibility maps. *Ecol Ind* 110:105856
- Wang Y, Feng L, Li S, Ren F, Du Q (2020) A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. *CATENA* 188:104425
- Wu Y, Ke Y, Chen Z, Liang S, Zhao H, Hong H (2020) Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *CATENA* 187:104396
- Yalcin A (2008) GIS-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in Ardesen (Turkey): comparisons of results and confirmations. *CATENA* 72(1):1–12
- Yalcin A, Reis S, Aydinoglu AC, Yomralioglu T (2011) A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey. *CATENA* 85(3):274–287
- Yariyan P, Janizadeh S, Van Phong T, Nguyen HD, Costache R, Van Le H, Tiefenbacher JP (2020) Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. *Water Resour Manag* 34:3037–3053
- Yilmaz I (2010) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environ Earth Sci* 61(4):821–836
- Zabihi M, Pourghasemi HR, Pourtaghi ZS, Behzadfar M (2016) GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environ Earth Sci* 75(8):665
- Zabihi M, Pourghasemi HR, Motevalli A, Zakeri MA (2019) Gully erosion modeling using GIS-based data mining techniques in Northern Iran: a comparison between boosted regression tree and multivariate adaptive regression spline. In: *Natural hazards GIS-based spatial modeling using data mining techniques*, pp. 1–26. Springer, Cham
- Zaheer M, Zaheer A, Hamza A (2020) Use of geoinformatics for landslide susceptibility mapping: a case study of Murree, Northern Area, Pakistan. In: *Transportation soil engineering in cold regions*, vol 2, pp 191–199. Springer, Singapore

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.