Spatial Prediction of Groundwater Potentiality Mapping Using Machine Learning Algorithms

Sunil Saha
University of Gour Banga
Amiya Gayen
University of Culcutta
Kaustuv Mukherjee
Chandidas University
Hamid Reza Pourghasemi (hr.pourghasemi@shirazu.ac.ir)
Shiraz University https://orcid.org/0000-0003-2328-2998
M. Santosh
University of Adelaide Press: The University of Adelaide

Research Article

Keywords: Groundwater potentiality, Decision tree, Naïve Bayes classifier, Random forest, Predictive neural network

Posted Date: March 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-303246/v1

Spatial prediction of groundwater potentiality mapping using machine learning algorithms

Sunil Saha[1], Amiya Gayen[2], Kaustuv Mukherjee[3], Hamid Reza Proughasemi*[4], M. Santosh[5, 6]

[1]Department of Geography, University of Gour Banga, Malda, India; E-mail: sunilgeo.88@gmail.com

[2]Department of Geography, The University of Calcutta, Kolkata, India; E-mail: mr.amiyagayen@gmail.com

[3]Department of Geography, Chandidas Mahavidyalaya, Khujutipara, Birbhum, India; E-mail: kaustuv2000@gmail.com

[*4] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran; Email: hr.pourghasemi@shirazu.ac.ir

[5]School of Earth Sciences and Resources, China University of Geosciences Beijing, Beijing 100083, P.R. China

[6]Department of Earth Sciences, University of Adelaide, Adelaide SA 5005, Australia

*Corresponding author Email: hr.pourghasemi@shirazu.ac.ir; hamidreza.pourghasemi@yahoo.com

Tel: +98-911-120-8810

Fax: +98-71-3227-6907

Abstract

Machine learning techniques offer powerful tools for the assessment and management of groundwater resources. Here, we evaluated the groundwater potential maps (GWPMs) in Md. Bazar Block of Birbhum District, India using four GIS-based machine-learning algorithms (MLA) such as predictive neural network (PNN), decision tree (DT), Naïve Bayes classifier (NBC), and random forest (RF). We used a database of 85 dug wells and one piezometer location identified using extensive field study, and employed 12 influencing factors (elevation, slope, drainage density (DD), topographical wetness index, geomorphology, lineament density, rainfall, geology,

pond density, land use/land cover (LULC), geology, and soil texture) for evaluation through GIS. The 85 dug wells and 1 piezometer locations were sub-divided into two classes: 70:30 for training and model validation. The DT, RF, PNN, and NBC MLAs were implemented to analyse the relationship between the dug well locations and groundwater influencing factors to generate GWPMs. The results predict excellent groundwater potential areas (GPA) DT RF of 17.38%, 14.69%, 20.43%, and 13.97% of the study area, respectively. The prediction accuracy of each GWPM was determined using a receiver operating characteristic (ROC) curve. Using the 30% data sets (validation data), accuracies of 80.1%, 78.30%, 75.20%, and 69.2% were obtained for the PNN, RF, DT, and NBC models, respectively. The ROC values show that the four implemented models provide satisfactory and suitable results for GWP mapping in this region. In addition, the well-known mean decrease Gini (MDG) from the RF MLA was implemented to determine the relative importance of the variables for groundwater potentiality assessment. The MDG revealed that drainage density, lineament density, geomorphology, pond density, elevation, and stream junction frequency were the most useful determinants of GWPM. Our approach to delineate the GWPM can aid in the effective planning and management of groundwater resources in this region.

Keywords**:** Groundwater potentiality, Decision tree, Naïve Bayes classifier, Random forest, Predictive neural network

1. Introduction

Groundwater scarcity and drinking water crisis are among the severe challenges that the planet is facing in the future. Groundwater is the most valuable but diminishing resource, and proper delineation and management strategies are required. According to the World Bank report of 2012, India is a highly groundwater consuming country which uses approximately 230 km^3 of groundwater every year, which is greater than one-fourth of the global total (The World Bank

2012). If this trend continues, then in the coming 20 years, more than 60% of the Indian aquifers will be in a critical condition. The Indian aquifers are mainly of two types: those in Peninsular India, characterised by hard rock formation and low permeability and the alluvial aquifers of the Indo-Gangetic plain which are major storehouses of groundwater (Suhag 2016). Currently, groundwater demand is increasing due to the continuous evolution of primary and secondary activities such as agriculture, urbanisation, and industry (Fetter 2001; Ettazarini and Jakani 2020). Therefore, groundwater potentiality research is necessary to assess favourable circumstances for water occurrence.

Groundwater potentiality depends on the hydrogeological background of the region, especially aquifer properties, porosity, permeability, hydraulic conductivity of the aquifer materials, storage capacity, and groundwater recharge. These are broadly dependent on physical characteristics such as geology, geomorphology, drainage, rainfall, soil, and LULC (Golkarian et al., 2018). Varied natural conditions lead to differences in groundwater yield and potential. Irrigation and unhindered domestic use are rapidly destroying groundwater (GW) storage, especially in the zones of low GW potential. Presently, unscientific management and unnecessary use of groundwater have impacted groundwater recharge levels (Chaudhry et al. 2019). Therefore, groundwater potentiality assessment is an adequate management strategy in these circumstances (Chen et al. 2019a, b). Therefore, demarcation of the GW potential zone can help to identify the future prospects of GW yield, which can guide the proper management of aquifers. It will also determine the intensity and duration of GW yield in upcoming years and whether we should continue up drafting of water or not. High-potential zones should be used for GW discharge and low zones should be avoided. Hence, GW potential zone identification is the initial stage of mitigating the water scarcity problem in a wide range of areas.

Research on groundwater potentiality mapping with the help of numerus well-accepted models. Examples include the frequency ratio (Das 2019; Falah and Zeinivand 2019; Trabelsi et al. 2019), weight of evidence (Khoshtinat et al. 2019), fuzzy logic (Taheri et al. 2020), analytic hierarchy process (Chaudhry et al. 2019; Das 2019), and logistic regression (Rizeei et al. 2019; Nguyen et al. 2020). Presently, these models are not used frequently because they do not solve multi-criteria decision problems. Therefore, numerous machine learning approaches offer more effective and efficient potential decisions using geographical information systems (GIS) and remote sensing **(RS)** data (Kanevski et al. 2009; Micheletti et al. 2013). The major strength of machine learning is its capability to handle data with high dimensionality and to map those with complex classification issues. Guzman et al. (2015) used a support vector machine (SVM) and an artificial neural network (ANN) to compare and predict GW potentiality. Their study shows that the SVM model provides better results with respect to the ANN model in cases of groundwater level forecasting. Sahoo et al. (2017) applied machine learning tools for change groundwater fluctuation assessment in the productive cropping regions of the USA. Kenda et al. (2018) used machine learning methods for groundwater modelling where decision trees, gradient boosting, linear models, and random forests were used. The Gradient boosting model provided better results with $R^2$0.644, where the $R^2$of the random forest model was 0.609. Lee et al. (2018) applied ANN and SVM methods to determine groundwater potential. In this case, ANN has a greater prediction accuracy (ROC=83.57%) than the SVM model (ROC=80.83%). Khosravi et al. (2018) used machine learning techniques to map spring groundwater potentiality, whereas Shao and Campell (2002) employed machine learning methods to predict the GW level. Chen et al. (2019) applied an adaptive neuro-fuzzy inference system (ANFIS) and ensemble models of biogeography-based optimisation and teaching-learning based algorithms, but the ANFIS model showed better

prediction accuracy with an ROC of 86.6%. The various studies above confirm the potential of machine learning algorithms to predict and assess groundwater-related issues. In this study, we adopt state-of-the-art techniques in machine learning to prepare the GWPM of a highly GW-level fluctuating block in Eastern India. The DT, PNN, and NBC have not been used together in previous studies to predict GW in India. The main focus of this study was to identify and map the GWP of the Md. The Bazar Block of Birbhum District and also assesses the groundwater controlling efficiency of the predisposing factors with the MDG, which would help face the crisis of GW in the pre-monsoon season. Our study offers better results in terms of classification, zonation, and management of GW.

2. Materials and methods

2.1 Study area

Mahammad Bazar (Md. Bazar) region stretches from 23°55'N to 24°50' N and 87°25' E to 87°40' E with an area covers 313.4 km^2 (Fig. 1b) in the Jharkhand adjacent western blocks of Birbhum District (Fig.1a). This region was identified as one of the eleven drought-affected districts of West Bengal (Deb and Nambiar 2010). It is the home of 164,570 people as per the Census of India, 2011, having 12-gram panchayets (administrative village units). This region has a warm monsoon climate, where the temperature varies from 6°Cto 40°C and annual precipitation is nearly 1,200 mm (Saha 2017). The western part of the block receives slightly higher rainfall and maximum precipitation occurs during the monsoon period (July–September). The long gap in the rainy season and the over-increasing pressure of agriculture lead to continuous updraft of GW for irrigation which is one of the major issues in this region. The lack of sufficient rainfall creates meteorological drought which turns into a hydrological drought, as is quite prominent in this region, leading to the lack of proper groundwater recharge. Precise GW potentiality zoning is an

120 essential requirement for drought-affected regions. As per the data collected from CGWB's 85 dug
121 wells and one piezometer of Md. Bazar, water level has shown fluctuation from 0.7 metre to 10.47
122 metres in this block in between April 2016 to January 2017 (CGWB 2017). The average depth of
123 groundwater ranges from 30 to 100 mbgl (meter below ground level). aquifers are the main route
124 for recharging groundwater, which influences the water level. Md. Bazar area groundwater
125 recharge occurs (4-10 inches) in both anthropogenic and natural activities such as hydropower
126 dams, artificial cannels, and check dams. This study area also falls under gneisses and associated
127 rocks, older alluvium, and older alluvium with lateritic types of aquifer media. The older alluvium
128 has moderate to high yield prospects, where older alluvium with laterite rocks has adverse or
129 limited yield prospects within 100 to 700 gpd/ft^2 hydraulic conductivity in the study area (Nakhaei
130 et al., 2015; Thapa et al., 2018).

131 **2.2 Thematic layers**

132 Various geo-environmental data layers were collected from different sources to accomplish the
133 objectives of this study. Initially, several methodologies were organised for the assessment of
134 GWPZs. The SOI (Survey of India) topographical maps were employed for the development of
135 DD map of the Md. Bazar block. The ASTER digital elevation model and Landsat-8 images were
136 used for the development of elevation, slope, TWI, and LULC maps. These two data were taken
137 from the USGS Earth Explorer. Geomorphology and lithology maps were collected from the
138 National Remote Sensing Centre, Hyderabad, at a scale of 1:50,000. Delineation and identification
139 of several types of spatial thematic layers aid in highlighting the tone, size, texture, colour,
140 association, and pattern (Table 1).

141

142 **2.3 Effective groundwater determining factors**

143 Elimination of predisposing factors for GWP mapping was based on the availability of data and
144 past research (Kaliraj et al. 2014; Agarwal and Garg 2016; Al-Abadi et al. 2016). The selected
145 factor maps with their thematic data were incorporated into the ArcGIS environment. A set of 15
146 × x15 toposheet maps were applied to calculate the stream junction density, pond density, and
147 drainage density. Elevation is considered by many researchers (Golkarian et al., 2018; Sachdeva
148 and Kumar, 2020) to be an important factor in GWPM. Elevation ranges from 13m to 120m in the
149 Md. Bazar (Fig. 2a). Elevation is associated with wind direction, temperature, and moisture levels;
150 therefore, it can enhance the rate of groundwater percolation (Magesh et al. 2012; Golkarian et al.
151 2018). Generally, there is a positive relationship between height and slope. Thus, a highly elevated
152 region leads to a higher slope which enhances the velocity of moving water and does not allow
153 downward infiltration of water. Slope is also a good indicator of the groundwater recharge process
154 (Goudie, 2013; Golkarian et al., 2018).

155 The LULC is less susceptible to groundwater potentiality because it tends to initiate groundwater
156 discharge (Balamurgan et al. 2016). A LULC map of this block was generated using Landsat8-
157 OLI (9th April 2016) imagery based on a supervised classification method and the results were
158 confirmed by applying Cohen's Kappa index with 89.6% Kappa value. The study area has seven
159 LULC classes: residential area, water bodies, agricultural land, waste land, mining area, forest
160 cover, and sand cover (Fig. 2b). The regional slope angle ranged from 0% to 7.25 %. Groundwater
161 percolation spreads down slopes more quickly than upslope slopes (Arabameri et al. 2019). It also
162 reveals the slope aspect, which impacts sunlight, precipitation, wind, and soil moisture (Gayen et
163 al. 2019) and the behaviour of groundwater recharge in the study area (Fig. 2c). Rainfall intensity
164 and duration also play important roles in groundwater storage (Shekhar and Pandey 2014).

Jothibasu and Anbazhagan (2016) noted that rainfall influences GPM accuracy and moving water percolation. Climatologically, Md. The Bazar Block falls under a warm monsoon climate (Fig. 2d). Lineaments, pond, and drainage density dictate the permeability and structural characteristics of an area that influences the movement of surface and subsurface water and can significantly affect groundwater storage and movement through a hydraulic gradient. In general, locations closer to the river have greater soil moisture and more groundwater, and proximity is positively related to GWP (Fig. 2l). Lineaments define the weakness and linearly fractured zones and might include linear features such as dykes and faults in the crust (Prasad et al. 2010; Magesh et al. 2012; Goudie et al. 2013). Rahmati et al. (2015) showed that these factors have a significant influence on the groundwater potentiality assessment. The maximum fracture zones were located in the southern portion of the Md. Bazar block (Fig. 2 (e-f)). This block is mainly drained by two perennial rivers, that is, River Dwarka and River Mayurakshi and numerous non-perennial channels are also present mainly in the western part. The drainage density of this block was classified into five classes (Fig. 2g): very high, very low, high, moderate, and low. The Drainage density and groundwater permeability are inversely proportional to each other. Drainage density and ponds are inversely related to groundwater recharge because very high drainage density is an indication of more surface runoff in a region and vice versa (Fig. 2h). Junction denotes the confluence areas of the two rivers and the chances of water availability are higher at the confluence points (Fig. 2i). The TWI is used to assess the influence of topography on hydrogeomorphic processes. It is an integration of the upstream area and slope within per unit width orthogonal to the flow path. The TWI also helps to assessments soil environment and topography influence intensity of an area. The TWI has been implemented for the assessment of the GWPZ (Golkarian

et al. 2018; Arabameri et al. 2019) and to describe spatial wetness patterns (Golkarian et al. 2018). The calculation method of the TWI was proposed by Moore et al. (1991):

$$TWI = \ln(\alpha/\tan\beta), \quad (1)$$

where α is the cumulative upslope area drainage per unit width orthogonal to the flow direction and β *is the inclination of the ground surface at the point* (Fig. 2j).

Soil type is the most important predisposing factor for the assessment of the infiltration rate in any region. The study area falls under six major soil types: sandy, clay loam, loamy, sandy loam, sandy clay loam, and sandy clay. maximum area of the Md. Bazar blocks are covered by sandy loam and clay loam soil types (Fig. 2k). Generally, the potentiality of the groundwater infiltration rate is higher in sandy regions than in loamy or clayey strata. Table 2 shows the characteristics of various soil textures and their impact on the groundwater infiltration of this block. The Md. The Bazar Block is composed of two major geological formations. The western part is dominated by pink granite, whereas rocks belonging to the Vindhyan Formation occur to the east (Fig. 2m). The pisolitic and kankar ferruginous concretions were mostly found in the laterite track. Some parts of the block are covered by basaltic rocks and younger alluvium. The block falls under three primary geomorphological regions: depositional plan, anthropogenic origin, and denudational plan (Fig. 2n). The western part of this block is characterised by undulating and rugged topography; therefore, the soil erosion rate is high in this part, with the decomposed and eroded materials settling down in the eastern part of this block, generating depositional landforms.

2.4 Methods

The RF, PNN, DT, and NBC MLMs were used for the GWP. All the implemented models' algorithms were suitable with R version 4.0.0 (Zamanirad et al., 2019). first, the ArcGIS spatial analysis tool (random partition algorithm) was used for classification of 85 dug wells and the

location of the piezometer as occurrence (assigned as "1") and 24 dug well location as non-occurrence (assigned as "0"). The dug well location with periods of low water table in the well was carried out in the study area. The 12 predisposing factor raster values for each dug well location were added to the R environment to run each model's algorithms. The spatial data generated by the individual model were imported in ArcGIS 10.2, and subsequently classified using the natural-break method (North 2009; Razavi Termeh et al. 2018). For the assessment of the water quality index, 31 well locations were considered as sample data for the water chemical test (Figure 1).

2.4.1 Machine learning models for the GWPM

2.4.1.1 Decision tree (DT)

The decision tree is a hierarchical model-based decision-support method. The DT model simplifies the decision rules by accurately splitting independent factors into uniform zones (Loh and Shih 1997; Lee and Lee 2015). They split into well locations and non-wells for groundwater potentiality based on the spatial interaction between the dug well locations and each predisposing factor. The DT model is different from other statistical models because it does not make any statistical assumptions. The DT model is computationally advanced and can also trickle data illustrated on various measurement scales (Pal and Mather 2003). In this study, a data-driven model, such as DT, was adopted to obtain a more precise and reliable prediction of groundwater potentiality.

2.4.1.2 Random forest (RF)

Random forest is a well-known machine learning algorithm for both regression and classification tasks (Breiman, 2001; Naghibi et al. 2017; Mandal and Pal, 2020). For regression and classification of the training datasets, researchers will attempt to develop a decision tree (DT) for class-wise output extraction and to acquire the dependent variable (Sachdeva and Kumar, 2020). This

algorithm is structured by numerous DTs and composites to determine the relationship between the factors affecting groundwater and dug well occurrence (Kim et al. 2018). The convenience of the RF algorithm with respect to other algorithms is as follows: First, it can solve the overfitting problems of the datasets as well as mange big datasets with various dimensionality in nature. second, it does not need any hypotheses within the response variable and explanatory variables. third, it does not require any previous data to rescale and transform the datasets (Elmahdyet al., 2020). The RF classification adopted resampling methods by randomly transferring the predictive factors to enhance the diversity in every tree (Naghibi et al. 2017). The notation of the predictive variable is defined as log 2 (M + 1), where M is the total input number within the algorithm. The RF model determines the split at each node with the help of predictive variables and the number of trees (Kim et al. 2018). The average prediction of the tree is computed as:

$$S = \frac{1}{K}\sum K^{th}\, v^{response}, \quad (2)$$

where S is any groundwater prediction and K represents the separate trees in the algorithm.

2.4.1.3 Naïve Bayes Classifier (NBC)

The NBC is a well-accepted ML classifier algorithm which is developed based on Bayes' rule, and it acts through an independent of variables assumption (Soria et al. 2011). The importance of the NBC is that it is very simple for estimation schemes for modelling that do not require any kind of complex iterative parameter (We et al. 2008). Previously, the NBC model was effectively implemented for the assessment of gully erosion and landslide susceptibility (Chen et al., 2019). In this work, the NBC model was generated by using "rminer' package on R 3.0.2 software environment (Cortez 2016).

2.4.1.4 Predictive Neural Network (PNN)

The PNN is another type of neural network model which estimates the potentiality or susceptibility with the help of classified sample data (Azimi et al. 2018). This method was developed based on probability density algorithms. In fact, these PNN models solve the clustering problems generated during the implementation of Bayesian decision theory. These networks were developed through the implementation of two layers. In this implemented network, the second layer, dissimilar to other networks, targets the outcomes in the vector form with values ranging between 1 and 0 (Azimi et al. 2018; Dou et al. 2018). In this work, first, the continuous values were converted through a fuzzy membership function and newly developed values were applied as input layers in the PNN model. In this algorithm, 0.5 values denote false membership when contrast value is '0'. The degrees of negative and positive divergence ranged from 1 to 0. A sizeable fuzzy function was implemented, which was subtracted from the contrast. It is necessary to have positive values and to be efficient in dealing with classes that do not acquire any training points, as there are no chances of any contrast value (CV). In general, CV denotes the characteristics of the control layer from the output table. For the inauguration of fuzzy logic, at first the layers were first mixed through the conditional operator processes, as mentioned in the previous sections. For locate and determine the class points, the GWPI raster layer was considered as the criterion. The raster values were then converted into three different seats to demarcate the situation of the two "occurrence" and "non-occurrence" points. The newly developed data layer was implemented to demarcate the membership severity of the two individual classes (Azimi et al. 2018; Dou et al. 2018). Finally, the criterion of the potential index, the potential of groundwater, was employed into four classes.

2.5 Validation of GWPMs

model prediction assessment is an essential task for evaluating the methodological implications of a study (Obeidavi et al., 2021). Here, the researchers were used the AUC-ROC (area under the curve) to estimate the prediction efficiency of the implemented models. This is a pectoral representation of model performance through a diagnostic test (Sachdeva and Kumar, 2020). The AUC indicates the model sensitivity by plotting the actual prediction of groundwater on the vertical axis (Y) and the false prediction of groundwater on the horizontal axis (X) (Gayen et al. 2018; Gayen et al. 2019). The implemented model's prediction of non-occurrence and occurrence of groundwater locations was identified using the AUC**.** The AUC value extends from 1 to 0.5; a lower value indicates a lower performance efficiency of the models (Chen et al. 2019; Obeidavi et al., 2021).

2.6 Water quality index (WQI)

WQI is a rating of overall water quality, which is influenced by individual parameters of water quality. This is calculated in terms of the level of human consumption. The World Health Organization (WHO) proposed a standard level for drinking purposes for WQI calculations.

The entire WQI methodology was conducted in three stages. In the initial stage, 11 parameters (EC, HCO3, Ph, TH, Cl, Na, Ca, F, Mg, K, and SO4) were considered and given their weight (Wi) with respect to their degree of influence on human health (Nik and Pirohit, 2001) (Table 3). The weight was assigned from 1 to 4 based on their influence on water quality (Srinivasamoorthy et al., 2008). Subsequently, the individual parameter relative weight (Wi) was calculated using Equation 3:

$$Wi = \frac{wi}{\sum_{i=1}^{n} wi} \quad (3)$$

where wi denotes the weight and n is the total number of selected parameters. The standard values and relative weights proposed by the WHO are listed in Table 3. In the next step, Equation 4 was used to calculate the quality rating scale (qi):

$$qi = \frac{Ci}{Si} \times 100 \quad (4)$$

where Ci and Si represent the actual and WHO standard concentrations of individual chemical compositions in milligrammes per litre.

In the last stage, the WQI was calculated using the sub-index of the ith parameter (Sli). The Sli was calculated based on Equation 5 and finally, WQI was computed using Equation 6:

$$Sli = qi \times Wi \quad (5)$$

$$WQI = \sum Sli \quad (6)$$

The WQI was classified into five classes: unsuitable for human consumption, very poor, poor, good, and excellent (Sahu and Sikdar, 2008; Yidana and Yidana, 2010).

3. Results

3.1 Groundwater potentiality models

The GWPMs were developed using the four well-accepted MLAs (Figs. 3 (a-c)). The groundwater prediction probability produced by each model were divided into four classes as mentioned above with the help of natural-break statistics (Razavi Termeh et al. 2018). The GWPM developed with the RF model exhibits values ranging from 0.0 1.0. The excellent and good groundwater potential

318 zones cover 14.69% and 5.13% of the study area, respectively. Poor and moderate potentiality
319 zones make up 73.37% and 6.81% of the total area, respectively (Table 4). The PNN model
320 generated prediction values from 0.00 1.0. The results revealed that 13.97% and 4.09% of the Md.
321 Bazar area cover an excellent and good potential for groundwater. The poor and moderate GWP
322 classes cover 76.28% and 5.67% of the Md. Bazar blocks (Table 3). The DT potentiality values
323 range from 0.00 to 1.0 and classify 17.38%, 8.82%, 7.98%, and 65.81% as zones of excellent,
324 good, moderate, and poor GWP (Table 3). The NBC model generated prediction values ranges
325 from 0.00 to 1.0, and identified 20.43% and 3.77% of the study region as having excellent and
326 good potentiality to groundwater, respectively. poor and moderate GWP accounted for 75.81% of
327 the total study area. It is clear in all groundwater potential maps; the maximum probable
328 groundwater areas are found in the southern part of this block. Concurrent with the output from
329 the four models, it is revealed that the DT) model generates GWPM that is discontinuous in
330 manner, whereas the NBC, RF, and PNN models provide smooth patterns of GWP maps.

331 ***3.2 Assessment of the relative importance of the effective factors***

332 It is vital to visualise the relationships between the selection factors and dug well locations to
333 evaluate the significance of each diver in the upliftment of groundwater recharge (Fig. 4).
334 Generally, the natural recharge of groundwater is impacted by different predisposing factors
335 (Conforti et al., 2010). The MDG is implemented to assess the relative importance of the variables
336 in the implemented models. The RF-based MDG algorithm was first introduced by Breiman in
337 2001 in concert with decision trees (Zang et al. 2017). The statistical software R 3.3.3, was applied
338 for "RF" analysis (Breiman 2001). The MDG values range from 34.12 to 185.73. Drainage density
339 (185.73), lineament density (162.61), stream junction density (151.35), pond density (126.25), and
340 elevation (101.82) were the most important factors. These were followed in terms of slope (83.01),

341 geomorphology (82.85), TWI (80.26), geology (72.32), soil texture (62.96), LULC (42.16), and
342 rainfall (34.12). All 12 predisposing factors were subjected to the modelling purposes, because all
343 are contributors to groundwater potentiality occurrence. maximum area of the Md. Bazar blocks
344 fall under agricultural land, which is favourable for groundwater storage. The southwestern and
345 northeastern parts of this block are covered by forests. In the southern part of this block, the
346 probability of groundwater storage is higher because of the presence of a water reservoir.

347 ***3.3 Validation of four models***

348 The results of the ROC curve indicated that the PNN, RF, DT, and NBC models had 80.1%, 78.3%,
349 75.2%, and 69.2% accuracy, respectively, with a significance level of 0.00 (Fig. 5). The ROC
350 values range from 0.69 to 0.8, demonstrating that all four models have reasonable and good GWP
351 prediction capacities. Finally, the outcome of the implemented four models, that is, PNN, RF, DT,
352 and NBC, the PNN offered more predictive accuracy. When applying the NBC algorithms with a
353 small number of variables, the training accuracy is not influenced by the increasing training data
354 size. However, such inconsistencies do not occur in the cases of RF, PNN, and DT in the cases of
355 the GWP models.

356 **3.4 Assessment of water quality**

357 Table 5 shows the locational distribution of WQI in Md. Bazar area. The WQI values ranged from
358 34.94 to 126.29. The highest and lowest WQI values were observed in the Ganpur (sample 1) and
359 Khayrakuri (sample 31) villages, respectively. The Maximum sample areas fell under good and
360 excellent water quality classes, so people can use ground water for drinking purposes. portable
361 water was found in the location of poor to moderate GWP areas and poor water was found in the
362 location of excellent GWP areas. The cause of the adverse situation may be the cause of the uneven

distribution of the net recharge. The net recharge indicates the amount of water that infiltrated into the aquifer. The infiltration route of transportation is a key route of **contaminants, even** though the penetration of groundwater dilutes the number of contaminants recharging into the aquifer (Thapa et el., 2018).

4. Discussion

The elevation and slope were very low in the southeastern portion of this Md. Bazar block. Groundwater recharge is negatively related to the elevation of the study area. Thus, locations that are situated in low-elevation areas show high groundwater potential in particular regions of the study area rather than the overall study area. Machine learning and statistical models have been implemented to produce GWP and flood susceptibility with highly satisfactory results (Avana and Moradi, 2021; Kamali et al., 2020). Four machine-learning algorithms, RF, PNN, DT, and NBC, were used to delineate GWP zones in the Md. Bazar Block. Previous studies, including those by Tsangaratos and Ilia (2016), compared the naïve Bayes and logistic regression models for landslide probability assessment. The outcome of these models revealed that the NB model had a lower prediction error than the LR model. However, in the field of groundwater potentiality, validation datasets are used for the PNN, RF, and DT models, which provide good results. The NB model did not provide high prediction accuracy with respect to the PNN, RF, and DT models. Generally, AUC is a holistic approach for accuracy assessment because it acts on false negative, true negative, false positive, and true positive (Pal and Mandal, 2019). The PNN, RF, and DT models do not require a very good improving ability of multimodal optimisation and nonlinear problems and the primary solution to start its iteration system (Dou et al. 2018). The NB also had some good characteristics for probability prediction, that is, more flexibility, permission for adequate integration, and high learning ability. Nevertheless, the NB model shows higher acceleration than

386 the PNN, RF, and DT models, meaning that the NB model works with fewer training data, although
387 in this groundwater potentiality study, the predictive capacity is not remarkable. The results are
388 discussed briefly below.

389 ***4.1 Comparison of models***

390 Diverse MLAs have been implemented to assess GWP mapping, such as multivariate adaptive
391 regression splines (MARS) (Golkarian et al., 2018), random forest (Ho 1995; Breiman 2001),
392 RBFANN (Chen et al., 1991), BPANN (Hagan et al. 1996), and ANFIS (Jang et al. 1996; Razavi-
393 Termeh et al., 2019). The ANFIS and BPANN (Satir et al. 2016) were used to assess the
394 groundwater susceptibility. In the case of nonlinear modelling, Lee and Kwak (2016) postulated
395 the potential of the RBFANN. The MARS model was also applied by West et al. (2016) for ground
396 water modelling with better results. Several studies, including the statistical model, Mamdani
397 fuzzy logic model (FLM) and advanced analytical hierarchy model, are used to model groundwater
398 potentiality (Gigovic et al. 2019). The results from these studies show that the FLM has better
399 prediction accuracy. The MARS-DFP, fuzzy AHP, and ANFIS models have also been applied to
400 forest fire susceptibility (FFS) studies (Jang et al. 1998; Razavi-Termeh et al., 2020). The ensemble
401 model and MARS-DFP models were the best models in this study. In the present work, based on
402 the ROC and AUC evaluations, the PNN was the best suited model for ground water potentiality
403 modelling among the four studied models. In other studies, such as floods, landslides and
404 assessments of spring potential, ANN and RF have also provided good results. In the sense of no
405 over-feeding of data, the PNN model is the most important model. It consists of multiple decision
406 trees with an interaction between predisposing factors and non-linearity (Catani et al., 2013; Kim
407 et al., 2018). It was also observed that RF performed well (78.3 % accuracy from AUC
408 assessments) (Fig. 5). The weaker results of the NBC are due to the data over-fitting problem

(Golkarian et al. 2018). The PNN is a promising method that can be applied to map the GWP at local to regional scales. The applied model AUC values are more than 0.75, indicating that all three models are reliable for estimating the GWPZ in the study region. The Results from the PNN and RF GWP maps provide good results, similar to previous susceptibility assessment studies with RF and PNN (of floods, gully erosion, and landslides. Excellent outputs will demark very high and high classes with less variance or more consistently (Naghibi et al. 2017). The models generated four classes of results, revealing that those generated by the PNN and RF have strong predictive capacity with the lowest percentage of variation within the implemented models. The AUC results were nearly identical, and we can conclude that the implemented models are useful for GWPZ mapping in the study area. The PNN, RF, and DT models can be implemented for various spatial probability modelling, that is, gully erosion, landslides, floods, and other endavors at small and large scales in other places.

4.2 Importance of effective factors

The importance of the twelve-groundwater controlling factors was determined by applying the MDG algorithm to the RF method. The results showed that drainage density, lineament density, stream junction density, and pond density are of the highest importance, whereas topographical wetness index, land use, and soil texture are the least important. Pourtaghi et al. (2015) concluded that soil type, land use, and geomorphology were the most significant factors that promoted groundwater potentiality. On the other hand, Hong et al. (2017) reported that average soil type, slope, and vegetation cover were factors that had the greatest influence on groundwater occurrence. Ahearn et al. (2005) established that LULC is an important and effective factor for controlling the efficiency of hydrological processes (that is, land alteration, surface runoff, infiltration, over-grazing, and evapotranspiration. Rainfall has been reason is a highly influential factor in

controlling the efficiency of soil wetness and the condition of vegetation and the inverse, dry conditions, promote desiccation of soil and vegetation and decrease GWP. Elevation and slope are also factors conducive to groundwater to control the percolation rate, local climate, and sun intensity. Generally, a larger slope angle increases the effects of gravitational forces. Furthermore, groundwater potentiality increases more rapidly down slopes and more slowly up-slopes (Kushla and Ripple 1997). Climatic factors also significantly affect groundwater conditions (Hong et al. 2018). Parisien et al. (2016) also emphasised the relationships among soil types, geological conditions, and land use types. The most significant predisposing factors included in the GWP assessment are drainage density, lineament density, and stream junction frequency because the drainage defines the depression zones where the water accumulates and lineament defines the weakness and fractured zones by which accelerated surface water recharge and infiltration in the study area.

5. Conclusion

Continuous loss of groundwater and surface water in Md. Bazar in Eastern India demands effective and suitable modelling of the GWP map. For proper analysis of groundwater condition, various models have been used around the world. After a critical evaluation of this study, several MLAs were implemented to delineate the GWP zones. Based on data from field investigations and previous studies, 12 groundwater effective factors were overlaid on a GIS platform for groundwater potentiality assessment. These factor layers were integrated with the RF, PNN, DT, and NBC models. The DT, RF, NBC, and PNN models identified 17.38%, 14.69%, 20.43%, and 13.97%, respectively, of the study area to have excellent groundwater potentiality and 65.81%, 73.37%, and 71.17%, and 76.28% poor groundwater potential in the study region. The results indicate that this region faces vulnerable situations in the future if proper water management

schemes are not implemented. The results of the models were confirmed by the ROC curve. The ROC results reveal that the PNN, RF, and DT models are more efficient than the NBC model for evaluating groundwater resources. The drainage density, lineament density, pond density, and stream junction density were the most effective factors for groundwater potentiality. The findings of this research should be used for environmental protection, groundwater resource management, groundwater contamination, and land use planning by private and local governmental agencies in the Md. Bazar Block of Birbhum District, India. further groundwater recharge processes and their management processes are required for the investigation of Md. Bazar Block.

Acknowledgment

This work was supported by College of Agriculture, Shiraz University (Grant No. 98GRC1M271143).

Ethical Approval

Not applicable.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Authors Contributions

SS, AG, KM, HRP, and MS designed experiments, run models, analyzed results, wrote and reviewed manuscript.

Availability of data and materials

Data will send based on request.

Conflicts of Interest: "The authors declare no conflict of interest."

Funding: The study was supported by the College of Agriculture, Shiraz University (Grant No. 96GRC1M271143).

References

Agarwal, R., & Garg, P. K. (2016). Remote sensing and GIS based groundwater potential & recharge zones mapping using multi criteria decision making technique. *Water Resour. Manag. 30 (1),* 243–260.

Ahearn, D. S., Sheibley, R. W., Dahlgren, R. A., Anderson, M., Johnson, J., & Tate, K. W. (2005). Land use and land cover influence on water quality in the last free-flowing river draining the western Sierra Nevada, California. *Journal of Hydrology*, *313*(3-4), 234-247.

Al-Abadi, A. M., Al-Temmeme, A. A., & Al-Ghanimy, M. A. (2016). A GIS-based combining of frequency ratio and index of entropy approaches for mapping groundwater availability zones at Badra–Al Al-Gharbi–Teeb areas, Iraq. Sustain. *Water Resour. Manag. 2,* 265–283.

Arabameri, A., Rezaei, K., Cerda, A., Lombardo, L., & Rodrigo-Comino, J. (2019). GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Science of the Total Environment, 658,* 160–177.

Avand, M., & Moradi, H. (2021). Spatial Modeling of Flood Probability Using Geo-Environmental Variables and Machine Learning Models, Case Study: Tajan Watershed, Iran. *Advances in Space Research*.

Azimi, S., Azhdary, M. Z., & Hashemi Monfared, S. A. (2018). Prediction of annual drinking water quality reduction based on Groundwater Resource Index using the artificial neural network and fuzzy clustering. *Contaminant Hydrology,* https://doi.org/10.1016/j.jconhyd.2018.10.010

Balamurugan, G., Seshan, K., & Bera, S. (2017). Frequency ratio model for groundwater potential mapping and its sustainable management in cold desert, India. Journal of King Saud University - *Science, 29(3),* 333–347.

505 Breiman, L. (2001). Random forests. *Machine learning, 45(1),* 5-32.

506 Catani, F., Lagomarsino, D., Segoni, S. & Tofani, V. (2013). Landslide susceptibility estimation
507 by random forests technique: sensitivity and scaling issues. *Natural Hazards and Earth System*
508 *Sciences, 13(11),* 2815-2831.

509 Chaudhry, A. K., Kumar, K., & Alam, M. A. (2019). Mapping of groundwater potential zones
510 using the fuzzy analytic hierarchy process and geospatial technique. *Geocarto International*, 1-22.

511 Chen, W., Panahi, M., Tsangaratos, P., Shahabi, H., Ilia, I., Panahi, S., & Ahmad, B. B. (2019b).
512 Applying population-based evolutionary algorithms and a neuro-fuzzy system for modeling
513 landslide susceptibility. *Catena, 172,* 212-231.

514 Chen, W., Tsangaratos, P., Ilia, I., Duan, Z., & Chen, X. (2019). Groundwater spring potential
515 mapping using population-based evolutionary algorithms and data mining methods. *Science of The*
516 *Total Environment*, *684*, 31-49.

517 Conforti, M., Aucelli, P. P., Robustelli, G., Scarciglia, F. (2010). Geomorphology and GIS analysis
518 for mapping gully erosion susceptibility in the Turbolo stream catchment (Northern Calabria,
519 Italy). Nat. Hazards, 56 (3), 881–898.

520 Cortez, P. (2016). Package 'rminer'; Date/Publication 2016-09-02 22:48:18; pp. 59.

521 Das, S. (2019). Comparison among influencing factor, frequency ratio, and analytical hierarchy
522 process techniques for groundwater potential zonation in Vaitarna basin, Maharashtra,
523 India. *Groundwater for Sustainable Development*, *8*, 617-629.

524 Dou, J., Yamagishi, H., Zhu, Z., Yunus, A. P., & Chen, C. W. (2018). TXT-tool 1.081-6.1 A
525 Comparative Study of the Binary Logistic Regression (BLR) and Artificial Neural Network
526 (ANN) Models for GIS-Based Spatial Predicting Landslides at a Regional Scale. Landslide
527 Dynamics: ISDR-ICL Landslide Interactive Teaching Tools. Springer.

528 Elmahdy, S. I., Mohamed, M. M., Ali, T. A., Abdalla, J. E. D., & Abouleish, M. (2020). Land
529 subsidence and sinkholes susceptibility mapping and analysis using random forest and frequency
530 ratio models in Al Ain, UAE. *Geocarto International*, 1-17.

Falah, F., & Zeinivand, H. (2019). GIS-Based Groundwater Potential Mapping in Khorramabad in Lorestan, Iran, using Frequency Ratio (FR) and Weights of Evidence (WoE) Models. *Water Resources*, *46*(5), 679-692.

Golkarian, A., Naghibi, S. A., Kalantar, B., & Pradhan, B. (2018). Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. *Environmental Monitoring and Assessment*, *190*(3), 1-16.

Goudie, A.S. (2003). Encyclopedia of Geomorphology. Routledge, London

Guzman, S. M., Paz, J.O., Tagert, M. L. M., & Mercer, A. (2015). Artificial neural networks and support vector machines: Contrast study for groundwater level prediction. In: Paper Presented at the 2015 ASABE Annual International Meeting

Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). Neural Network Design. PWS Publishing Co., Boston, MA, USA.

Ho, T. K. (1995). Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition 271. pp. 278–282.

Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A., & Xu, C. (2018). Applying genetic algorithms to set the optimal combination of forest related variables and model forest related susceptibility based on data mining models. The case of Dayu County, China. *Science of the Total Environment, 630,* 1044-1056.

Hong, H., Tsangaratos, P., Ilia, L., Chen, W., & Xu, C. (2017). Comparing the performance of a logistic regression and a random forest model in landslide susceptibility assessments. The Case of Wuyaun Area, China. Workshop on World Landslide Forum, pp. 1043–1050.

Jang, J. S. R., Sun, C. T., Mizutani, E. (1996). Neuro-fuzzy and soft Computing-a Computational approach to learning and machine intelligence. *IEEE Transaction on Automatic Control, 42,* 1482-1484.

Jothibasu, A., & Anbazhagan, S. (2016). Modeling groundwater probability index in Ponnaiyar River basin of South India using analytic hierarchy process. *Model. Earth Syst. Environ. 2,* 109.

Kaliraj, S., Chandrasekar, N., & Magesh, N. S. (2014). Identification of potential groundwater recharge zones in vaigai upper basin, Tamil Nadu, using GIS-based analytical hierarchical process (AHP) technique. *Arab. J. Geosci. 7,* 1385–1401.

Kamali Maskooni, E., Naghibi, S. A., Hashemi, H., & Berndtsson, R. (2020). Application of advanced machine learning algorithms to assess groundwater potential using remote sensing-derived data. *Remote Sensing*, *12*(17), 2742.

Kanevski, M., Pozdnoukhov, A., & Timonin, V. (2009). Machine Learning for Spatial Environmental Data. EPFL Press, Lausanne, Switzerland, p. 392.

Khoshtinat, S., Aminnejad, B., Hassanzadeh, Y., & Ahmadi, H. (2019). Application of GIS-based models of weights of evidence, weighting factor, and statistical index in spatial modeling of groundwater. *Journal of Hydroinformatics*, *21*(5), 745-760.

Khosravi, K., Panahi, M., & Tien Bui, D. (2018). Spatial prediction of groundwater spring potential mapping based on adaptive neuro-fuzzy inference system and metaheuristic optimization. *Hydrol. Earth Syst. Sci. 22,* 4771–4792.

Kim, J. C., Lee, S., Jung, H. S. & Lee, S. (2018). Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. *Geocarto international, 33(9),* 1000-1015.

Klemen Kenda, K., Čerin, M., Bogataj, M., Senožetnik, M., Klemen, K., Pergar, P., Laspidou, C., & Mladenić, D. (2018). Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer. *Proceedings, 2,* 697.

Kushla, J. D., & Ripple, W. J. (1997). The role of terrain in a fire mosaic of a temperate coniferous forest. *Forest Ecology and Management, 95,* 97-107.

Lee, M. W., & Kwak, K. C. (2016). An incremental radial basis function network based on information granules and its application. *Comput. Intell. Neurosci. 6.*

Lee, S., & Lee, S. W. (2015). Application of Decision-Tree Model to Groundwater Productivity-Potential Mapping. *Sustainability, 7,* 13416-13432.

583 Lee, S., Hong, S. M., & Jung, H. S. (2018). GIS-based groundwater potential mapping using
584 artificial neural network and support vector machine models: the case of Boryeong city in
585 Korea. *Geocarto International, 33(8),* 847-861.

586 Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. *Stat. Sin. 7,* 815–
587 840.

588 Magesh, N. S., Chandrasekar, N., & Soundranayagam, J. P. (2012). Delineation of groundwater
589 potential zones in Theni district, Tamil Nadu, using remote sensing, GIS and MIF
590 techniques. *Geoscience Frontiers, 3(2),* 189-196.

591 Mandal, I., Pal, S. (2020). Modelling human health vulnerability using different machine learning
592 algorithms in stone quarrying and crushing areas of Dwarka river Basin, Eastern India. *Advances*
593 *in Space Research*, *66*(6), 1351-1371.Micheletti, N., Foresti, L., Robert, S., Leuenberger, M.,
594 Pedrazzini, A., Jaboyedoff, M., & Kanevski, M. (2013). Machine learning feature selection
595 methods for landslide susceptibility mapping. *Math. Geosci. 46 (1),* 33–57.

596 Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: a review of
597 hydrological, geomorphological, and biological applications. *Hydrological Processes, 5(1),* 3-30.

598 Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random
599 forest, and genetic algorithm optimized random forest model in groundwater potential
600 mapping. *Water Resources Management, 31(9),* 2761-2775.

601 Naik, S., & Purohit, K. M. (2001). Studies on water quality of river Brahmani in Sundargarh
602 district, Orissa. *Indian J Environ Ecoplan*, *5*(2), 397-402.

603 Nakhaei, M., Amiri, V., Rezaei, K., & Moosaei, F. (2015). An investigation of the potential
604 environmental contamination from the leachate of the Rasht waste disposal site in Iran. *Bulletin of*
605 *engineering geology and the environment*, *74*(1), 233-246.

606 Nguyen, P. T., Duong, H. H., Mohammadtaghi, A., Abolfazl, J., Huu, D. N., Nadhir, A. N., Tran,
607 V. P., et al. (2020). "Soft Computing Ensemble Models Based on Logistic Regression for
608 Groundwater Potential Mapping." *Applied Sciences* 10, no. 7 (2020): 2469.

North, M. A. (2009). A method for implementing a statistically significant number of data classes in the Jenks algorithm. In: Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on, IEEE. pp. 35–38.

Obeidavi, S., Gandomkar, M., Akbarizadeh, G., & Delfan, H. (2021). Evaluation of Groundwater Potential using Dempster-Shafer Model and Sensitivity Analysis of Effective Factors: A case study of North Khuzestan Province. *Remote Sensing Applications: Society and Environment*, 100475.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens. Environ. 86,* 554–556.

Pal, S., Mandal, I. (2019). Impacts of stone mining and crushing on environmental health in Dwarka river basin. *Geocarto International*, 1-29.

Parisien, M. A., Miller, C., Park, S. A., DeLancey, E. R., Robinne, F. N., & Flannigan, M. D. (2016). The spatially varying influence of humans on fire probability in North America. *Environmental Research Letters. 11,* 075005.

Prasad, R. K., Mondal, N. C., Banerjee, P., Nandakumar, M. V., & Singh, V. S. (2008). Deciphering potential groundwater zone in hard rock through the application of GIS. *Environmental Geology, 55(3),* 467-475.

Razavi-Termeh, S. V., Khosravi, K., Sartaj, M., Keesstra, S. D., Tsai, F. T. C., Dijksma, R., & Pham, B. T. (2019). Optimization of an adaptive neuro-fuzzy inference system for groundwater potential mapping. *Hydrogeology Journal*, *27*(7), 2511-2534.

Razavi-Termeh, S. V., Sadeghi-Niaraki, A., & Choi, S. M. (2020). Ubiquitous GIS-Based Forest Fire Susceptibility Mapping Using Artificial Intelligence Methods. *Remote Sensing*, *12*(10), 1689.

Sachdeva, S., & Kumar, B. (2020). Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. *Stochastic Environmental Research and Risk Assessment*, 1-20.

633 Saha, S. (2017). Groundwater potential mapping using analytical hierarchical process: a study on
634 Md. Bazar Block of Birbhum District, West Bengal. *Spatial Information Research, 25 (4),* 615–
635 626.

636 Sahoo, S., Russo, T. A., Elliott, J., Foster, I. (2017). Machine learning algorithms for modeling
637 groundwater level changes in agricultural regions of the US. *Water Resources Research, 53(5),*
638 3878-3895.

639 Sahu, P., & Sikdar, P. K. (2008). Hydrochemical framework of the aquifer in and around East
640 Kolkata Wetlands, West Bengal, India. *Environmental Geology*, *55*(4), 823-835.

641 Satir, O., Berberoglu, S., & Donmez, C. (2016). Mapping regional forest fire probability using
642 artificial neural network model in a Mediterranean forest ecosystem. *Geomatics Natural Hazards*
643 *and Risk, 7,* 1645–1658.

644 Shao, Q., & Campbell, N. A. (2002). Applications: modelling trends in groundwater levels by
645 segmented regression with constraints. *Aust. N. Z. J. Stat. 44 (2),* 129–141.

646 Shekhar, S., & Pandey, A.C. (2014). Delineation of groundwater potential zone in hard rock terrain
647 of India using remote sensing, geographical information system (GIS) and analytic hierarchy
648 process (AHP) techniques. *Geocarto International, 30 (4),* 402–421.

649 Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., & Ellis, I. O. (2011). A non-parametric
650 version of the naive Bayes classifier. *Knowledge-Based System, 24,* 775–784.

651 Srinivasamoorthy, K., Chidambaram, S., Prasanna, M. V., Vasanthavihar, M., Peter, J., &
652 Anandhan, P. (2008). Identification of major sources controlling groundwater chemistry from a
653 hard rock terrain—a case study from Mettur taluk, Salem district, Tamil Nadu, India. *Journal of*
654 *Earth System Science*, *117*(1), 49.

655 Taheri, F., Jafari, H., Rezaei, M., & Bagheri, R. (2020). The use of continuous fuzzy and traditional
656 classification models for groundwater potentiality mapping in areas underlain by granitic hard-
657 rock aquifers. *Environmental Earth Sciences*, *79*(5), 1-16.

Thapa, R., Gupta, S., Guin, S., & Kaur, H. (2018). Sensitivity analysis and mapping the potential groundwater vulnerability zones in Birbhum district, India: A comparative approach between vulnerability models. *Water Science*, *32*(1), 44-66.

Trabelsi, F., Lee, S., Khlifi, S., & Arfaoui, A. (2019). Frequency ratio model for mapping groundwater potential zones using GIS and remote sensing; Medjerda Watershed Tunisia. In *Advances in Sustainable and Environmental Hydrology, Hydrogeology, Hydrochemistry and Water Resources,* (pp. 341-345). Springer, Cham.

Tsangaratos, P., & Ilia, L. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of model's complexity and training dataset size. *Catena, 145,* 164-179.

West, A. M., Kumar, S., & Jarnevich, C. S. (2016). Regional modeling of large wildfires under current and potential future climates in Colorado and Wyoming, USA. *Climatic Change, 134,* 565–577.

WHO (2004) Guidelines for drinking water quality: training pack. WHO, Geneva, Switzerland

Wu, X., Kumar, V., & Ross, Q. J. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14,* 1–37.

Yidana, S. M., & Yidana, A. (2010). Assessing water quality using water quality index and multivariate analysis. Environ Earth Sci. 59, 1461–1473.

Zamanirad, M., Sarraf, A., Sedghi, H., Saremi, A., & Rezaee, P. (2019). Modeling the Influence of Groundwater Exploitation on Land Subsidence Susceptibility Using Machine Learning Algorithms. *Natural Resources Research*, 1-15.

Zang, K., Wu, X., Niu, R., Yang, K., Zhao, L. (2017). The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. Environ Earth Sci. 76 (11), 405.

Figures

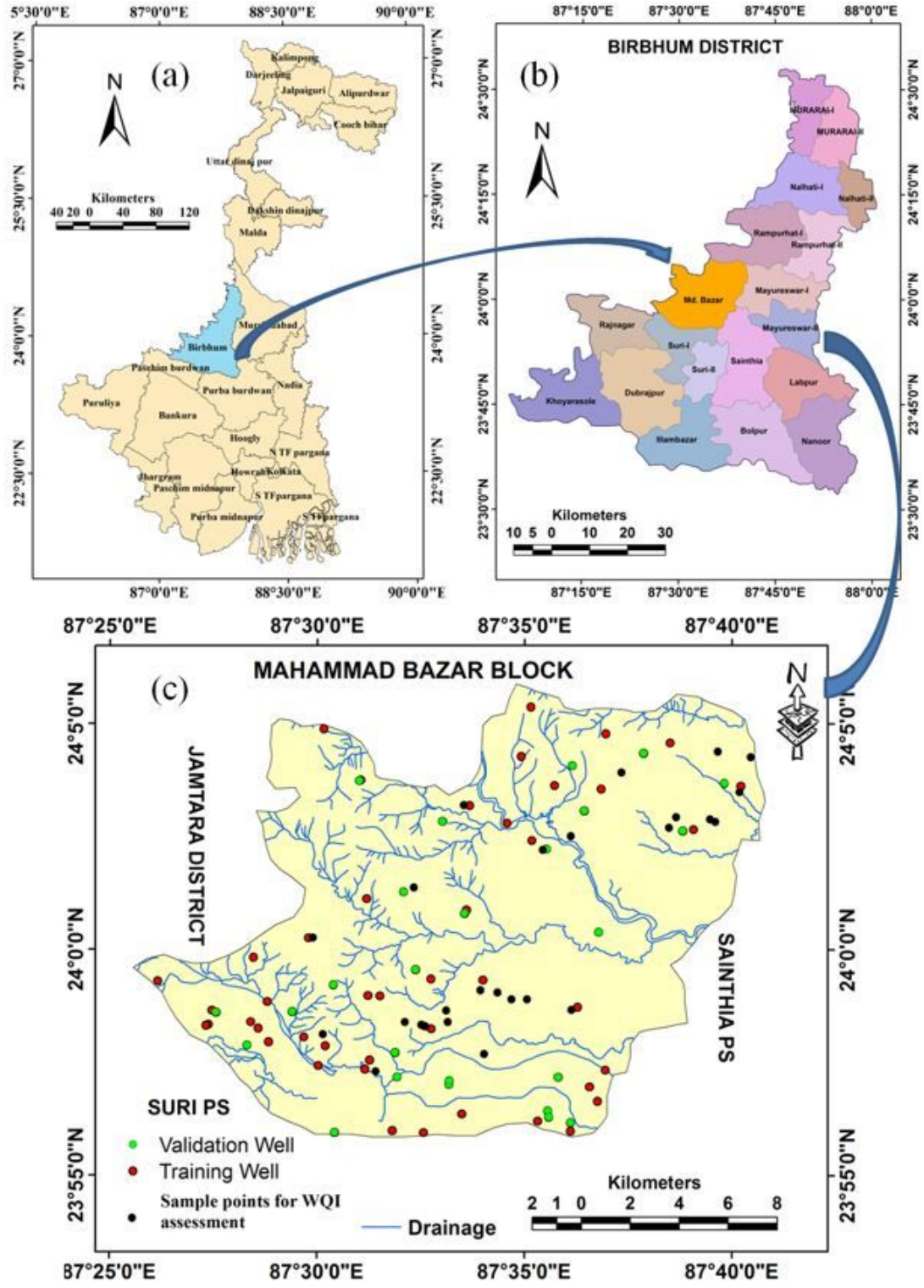

Figure 1

Location map of the study area; (a) West Bengal (b) Birbhum District & (c) Mahammad Bazar Block Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country,

territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

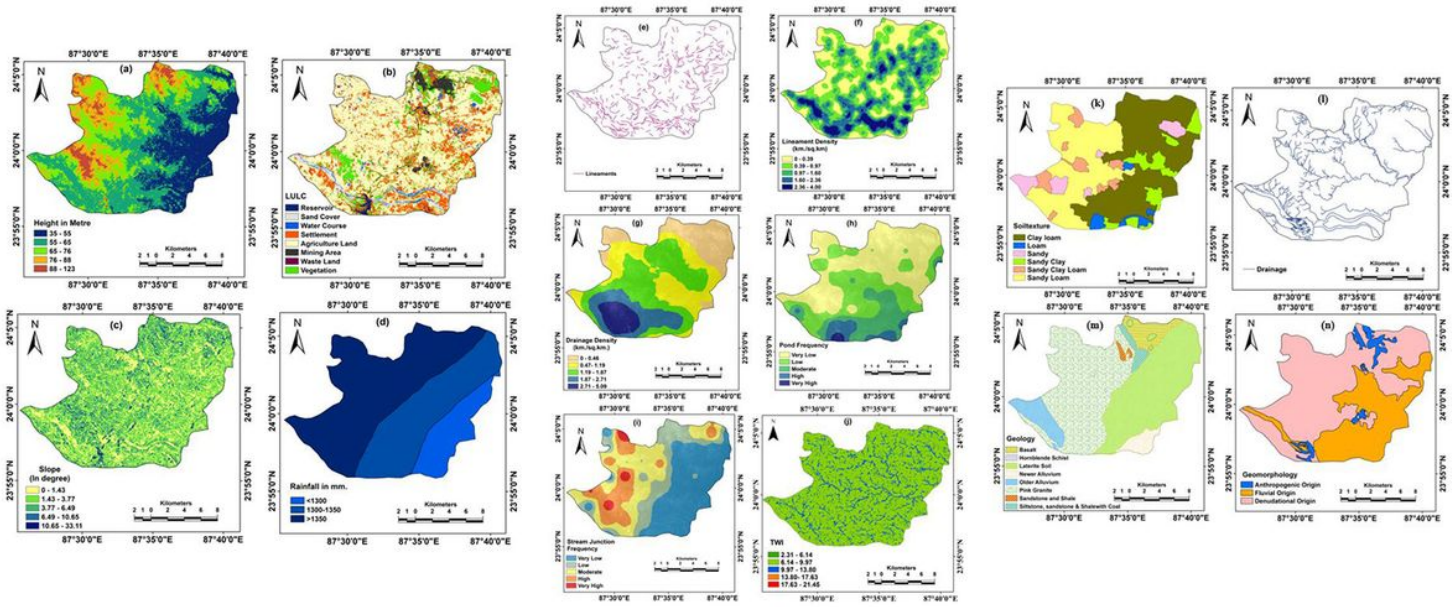

Figure 2

Thematic data layers representing groundwater potentiality factors- (a) Relief map (b) Land use map (c) Slope map (d) Spatial rainfall map Thematic data layers representing groundwater potentiality factors- (e) Lineament map (f) Lineament density (g) Drainage density (h) Pond Frequency (i) Stream junction Frequency (j) Topographic Wetness Index Physical attributes of the study area (k) Soil map (l) Drainage map (m) Geological Map (n) Geomorphological map Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

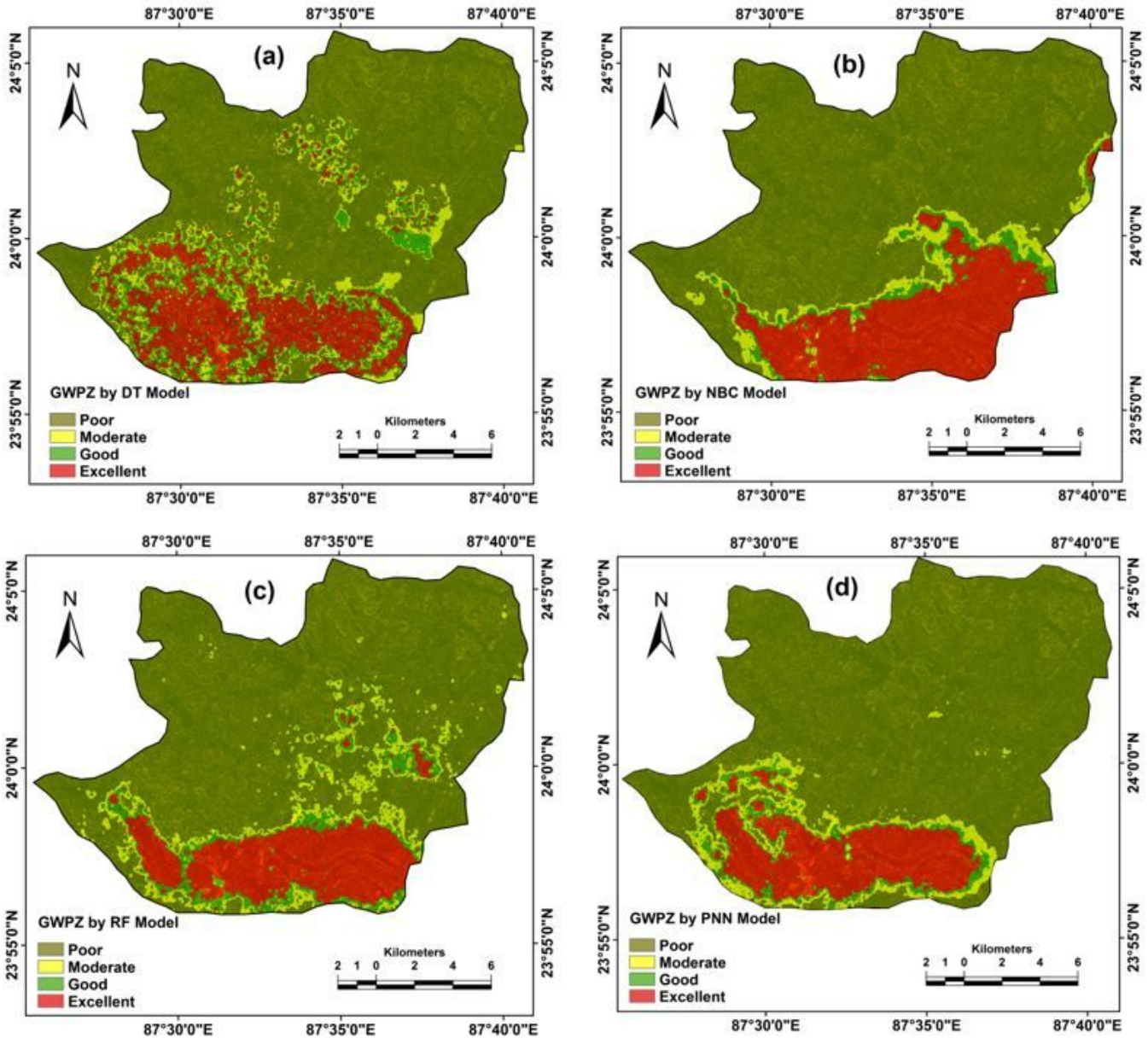

Figure 3

Thematic layers showing the groundwater potential zones (a) DT (b) NBC (c) RF, and (d) PNN Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

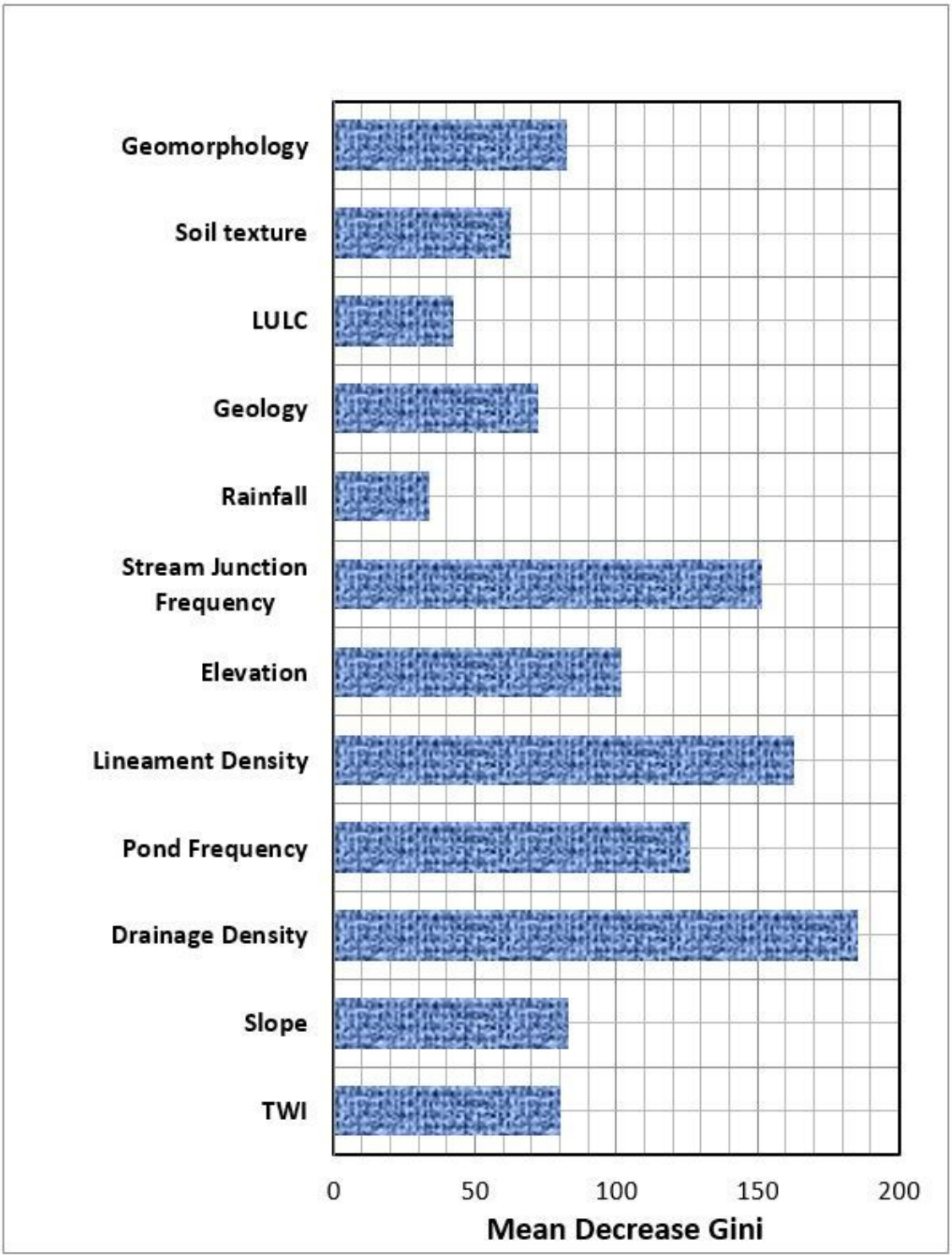

Figure 4

Bar diagram showing the variables important based on Mean Decrease Gini

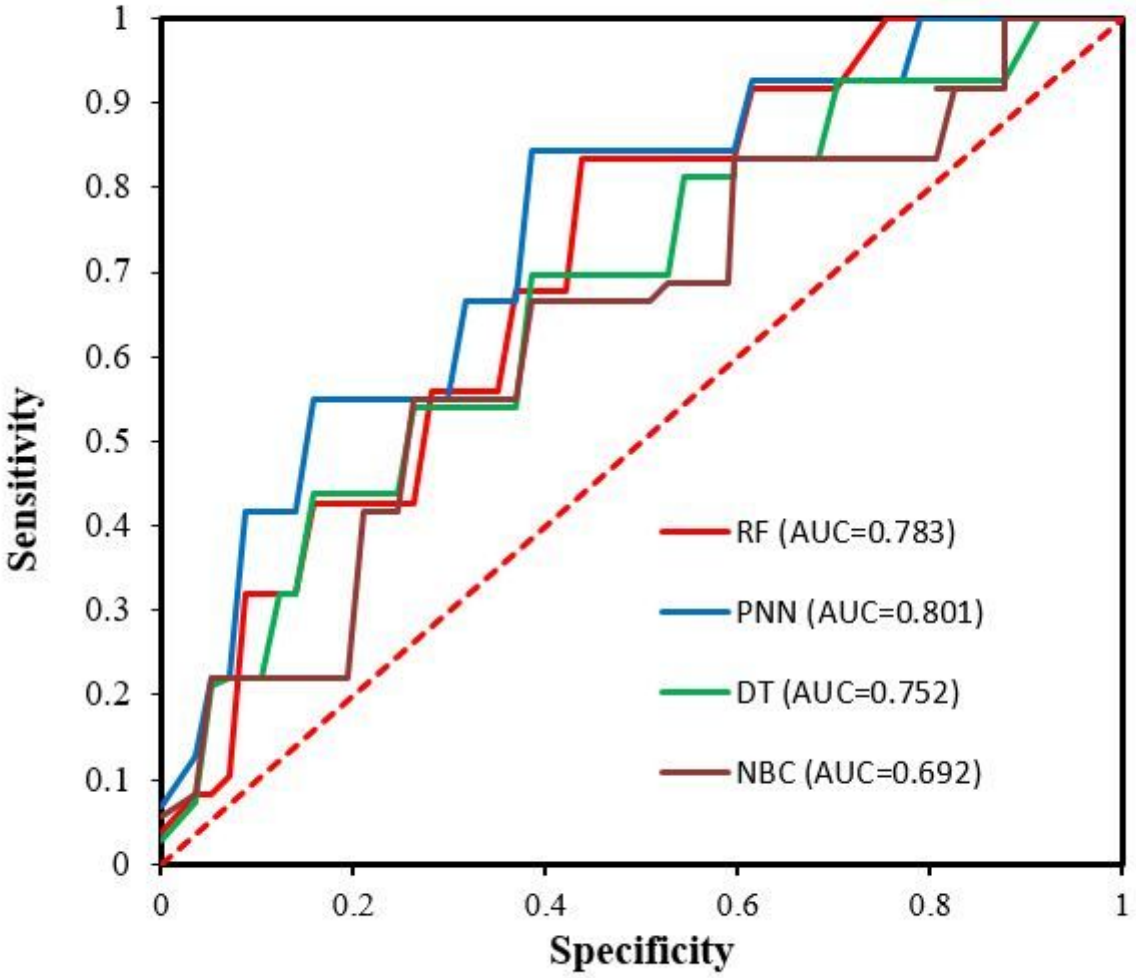

Figure 5

Receiver operating characteristics (ROC) curve of the ground water potentiality models