



OPEN Enhancement of groundwater resources quality prediction by machine learning models on the basis of an improved DRASTIC method

Ali Bakhtiarizadeh¹, Mohammad Najafzadeh^{1✉} & Sedigheh Mohamadi²

Determining situation of groundwater vulnerability plays a crucial role in studying the groundwater resource management. Generally, the preparation of reliable groundwater vulnerability maps provides targeted and practical scientific measures for the protection and management of groundwater resources. In this study, in order to evaluate the groundwater vulnerability of Kerman–Baghin plain aquifer, two developed indicators including composite DRASTIC index (CD) and nitrate vulnerability index (NVI) based on DRASTIC index were considered. Soft computing methods, including Gene Expression Programming (GEP), Evolutionary Polynomial Regression (EPR), Multivariate Adaptive Regression Spline (MARS), and M5 Model Tree (MTM5) have been used to provide formulations for prediction of NVI. Soft computing techniques were fed nine input parameters: depth to water level, net recharge, aquifer environment, soil environment, topography, effect of unsaturated area, hydraulic conductivity, land use, and potential risk related to land use. After calculating the vulnerability by soft computing methods, the results showed that the EPR model with Correlation Coefficient (R) of 0.9999 and Root Mean Square Error (RMSE) = 0.2105 has the best performance in the testing stage in comparison with MARS (R = 0.9966 and RMSE = 2.408), M5MT (R = 0.9956 and RMSE = 2.988), and GEP (R = 0.9920 and RMSE = 3.491). Although the EPR and GEP models have more complex mathematical computations than other soft computing models, the MARS and MT model that have quadratic polynomial and multivariable linear structures respectively, can be considered as the best alternative. According to the MARS model, the vulnerability of the region is divided into two categories: very low vulnerability (73.06%) and low vulnerability (26.94%). Overall, the statistical results of soft computing techniques were indicative of effective formulations for evaluating the DRASTIC index.

Keywords DRASTIC, Land use, Nitrate pollution, Artificial intelligence models, Groundwater resources

As the main concern, water scarcity is accelerating due to the destruction of available water resources and an increase in environmental contamination in dry and semi-dry areas. Due to the negative impacts on human health and ecosystem services, groundwater pollution is a global problem. As a complex and long-term process, pollution of groundwater resources is irrecoverable and costly to treat due to the large volume of the reservoir, long time, and lack of physical access to the reservoir. Therefore, the prevention of aquifer pollution is a necessity for the stability and protection of groundwater resources (e.g. ^{1–11}).

Since protecting groundwater from pollution is much simpler and more logical than removing it, so identifying pollution-prone-aquifer is the first step to preventing groundwater pollution. This identification allows the area to be subdivided into sub-areas in terms of the severity of the vulnerability and relevant measures to be taken to prevent contamination of vulnerable areas. Therefore, groundwater vulnerability and risk maps are the prime cause in the management and safeguarding of groundwater resources (e.g. ^{12,13}). Due to the expensive and

¹Department of Water Engineering, Faculty of Civil and Surveying Engineering, Graduate University of Advanced Technology, Kerman 76315117, Iran. ²Department of Ecology, Institute of Environmental Sciences, Graduate University of Advanced Technology, Kerman 76315117, Iran. ✉email: moha.najafzadeh@gmail.com; m.najafzadeh@kgut.ac.ir

time-intensive nature of groundwater sampling and quality monitoring, therefore, there is an essential need for groundwater vulnerability modeling as a fast and reliable tool to investigate groundwater aquifer vulnerability³. It is accompanied by understanding the possible events of pollution and determining the extent of pollutants and describing them (e.g.^{14–17}). Models for groundwater vulnerability take into account a range of factors to evaluate the likelihood of groundwater pollution. These factors typically fall into different categories: hydrogeological elements (i.e., hydraulic conductivity, porosity, depth to water table), geological factors (i.e., nature of subsurface materials and structural features), Land Use and Land Cover (i.e., type of land use and presence of impervious surfaces), source of pollution (i.e., proximity to potential contaminant sources and depth of the contaminant source), hydraulic factors (i.e., groundwater flow velocity and direction of groundwater flow), climate and hydrological conditions (precipitation patterns and temperature), depth and construction of wells, soil type and properties, soil permeability, and unsaturated zone properties (e.g.^{18–20}).

Integrating these elements into a groundwater vulnerability model aids in pinpointing districts that are at a higher risk of contamination, aiding in the improvements of methods for groundwater protection and management. While groundwater vulnerability models employ comparable influential factors, they adopt distinct methods for integrating and analyzing data. In general, three categories of methods exist for evaluating vulnerability, namely Point Count System Models (PCSM) (or ranking and weighting methods), statistical methods, and process methods (e.g.^{18,21,22}). Index and overlap methods consider the combination of different regional maps with the assignment of a numerical index¹⁹. Among these methods, DRASTIC, SINTACS, and Susceptibility Index (SI) are introduced as the most widely used techniques due to their simplicity, the need for minimal data, and the provision of a clear description of groundwater vulnerability. DRASTIC is a numerical model first proposed by the US Environmental Protection Agency (USEPA) in order to appraise the potential for aquifer contamination while considering hydrogeological data²³.

There are a plenty of research works in which circumstances of groundwater vulnerability have been investigated for diverse parts of the world (e.g.^{24–29}). Since the main problem of this model is to apply expert opinions to rank and weight the effective factors used in it, therefore the need to apply Artificial Intelligence (AI) models so as to improve rankings and optimize weights, that have been applied in DRASTIC evaluation, is of utmost importance for enhancement of the accuracy of susceptibility results (e.g.^{24,30–35}). Table 1 summarized the explored literature review of AI models applications into vulnerability indices of groundwater resources.

Authors	Methodology
Neshat et al. ^{5,34,26}	AHP-AHP, AHP-DRASTIC,DRASTIC-AHP
Neshat et al. ^{36,37}	FR-DRASTIC, AHP-DRSTIC, SPSA-DRASTIC
Langerudi et al. ³⁸	Fuzzy DRASTIC
Barzegar et al. ¹	ANN, SFL, MLE,NF,SCMAI
Baghapour et al. ³⁰	ANN for DRASTIC-LU and DRASTIC-N
Barzegar et al. ³¹	ANN-Committee based models (ELM, MARS, SVR, M5 MT)
Nadiri et al. ³⁹	SVM-FCF
Hu et al. ⁴⁰	AHP-DRASTIC
Jesiya and Gopinath ³³	Fuzzy AHP-DRASTIC
Bordbar et al. ³²	GWO-GALDIT, GA-GALDIT, Standard-GADIT for Cl/HCO3 pollution evaluation
Torkashvand et al. ⁴¹	SWARA and GA-DRASTIC
Bordbar et al. ⁴²	GALDIT-FR and GALDIT-GA
Jahromi et al. ⁴³	ABC-SINTACS and GA-SINTACS for nitrate and sulfate pollutions
Norouzi et al. ⁴⁴	RF-DRASTIC is better than GA-DRASTIC
Elzain et al. ⁴⁵	RBF-NN,RF, and SVM for DRASTIC evaluation
Gharekhani et al. ⁴⁶	BMA-SVM, BMA-GEP, and BMA-ANN for DRASTIC evaluation
Subbarayan et al. ⁴⁷	RF, XGB, and CART models for DRASTIC due to nitrate pollution
Karimzadeh-Motlagh et al. ⁴⁸	RF, SVM, GLM
Elzain et al. ⁴⁶	KNN, ERT, and EBR models for DRASTIC due to land use

Table 1. List of related research works. SPSA single-parameter sensitivity analysis, AHP analytical hierarchy process,ANN artificial neural network, SFL sugeno fuzzy logic, MFL mamdani fuzzy logic , NF neuro-fuzzy , SCMAI supervised committee machine artificial intelligence method , NVI nitrate vulnerability index, CD composite DRASTIC index, ELM extreme learning machine, MARS multivariate adaptive regression spline , SVR support vector regression, M5MT M5 model tree , AIMF artificial intelligence multiple framework , FCF fuzzy-catastrophe framework, GWO grey wolf optimizer, GA genetic algorithm, SWARA stepwise weight assessment ratio analysis, ABC artificial bee colony, RF random forest, RBF-NN radial basis function-neural network , ERFR ensemble random forest regression, BMA Bayesian model averaging , GEP gene-expression programming, XGB extreme gradient boosting , CART classification and regression tree, GLM generalized linear model, KNN K-nearest neighborhood, ERT ensemble extremely randomized trees , EBA ensemble bagging regression.

When evaluating the prediction of the DRASTIC index for groundwater resource quality, it is important to consider the drawbacks of other AI models such as Support Vector Machines (SVM), Random Forests (RF), and Deep Learning models. SVMs, while effective in handling high-dimensional data and finding complex decision boundaries, can be challenging to tune. Their performance is highly dependent on the choice of kernel and parameters, which may require extensive cross-validation. Additionally, SVMs can struggle with large datasets, as their computational complexity increases significantly with the number of samples, leading to longer training times and potentially less interpretable results. Random Forests are known for their robustness and ability to handle nonlinear relationships. However, they can produce models that are difficult to interpret, as the ensemble approach obscures individual feature contributions. While they generally perform well in many scenarios, they may not always capture subtle interactions effectively, particularly in cases where relationships among variables are not purely additive. Moreover, Deep Learning models, particularly neural networks, have gained popularity due to their ability to model complex patterns in large datasets. However, they require large amounts of training data to generalize well, which may not be available in many environmental contexts. Additionally, deep learning models often act as "black boxes," making it difficult to extract interpretable insights, which is a significant drawback in environmental assessments where understanding the relationship between variables is crucial.

In contrast, Gene-Expression Programming (GEP), Evolutionary Polynomial Regression (EPR), Multivariate Adaptive Regression Spline (MARS), and Model Tree (MT) offer several advantages in the evaluation of groundwater quality. GEP is capable of generating interpretable models that represent relationships in a tree-like structure, making it easier to understand how inputs affect predictions. It combines the adaptability of genetic programming with a focus on evolving mathematical expressions, which can capture complex interactions effectively. EPR stands out for its ability to evolve polynomial equations that fit the data without requiring a predefined model structure. This flexibility allows for a tailored approach to modeling complex relationships, making it suitable for capturing the intricacies of groundwater pollution dynamics. MARS is particularly adept at identifying and modeling nonlinearities and interactions between variables. Its piecewise linear approach allows it to fit models that can adapt to changes in the data, enhancing predictive performance while maintaining a degree of interpretability. MTM5, as a variant of Model Trees, combines decision tree structures with linear regression, enabling it to capture both local patterns and overall trends. This dual approach enhances its interpretability and provides insights into how different factors influence the DRASTIC index.

Overall, the use of GEP, EPR, MARS, and MT in predicting the DRASTIC index offers a balanced approach that combines flexibility, interpretability, and robustness. These models are particularly well-suited for environmental applications, where understanding the underlying relationships is as important as achieving accurate predictions. In predicting the DRASTIC index for groundwater resource quality using nine input parameters—depth to water level, net recharge, aquifer environment, soil environment, topography, effect of unsaturated area, hydraulic conductivity, land use, and potential risk related to land use—it's important to evaluate the limitations of certain AI models like SVM, RF, and Deep Learning models. As a major drawbacks of other AI models, SVMs are effective for classification tasks but can face challenges with the high dimensionality of groundwater data. Tuning SVM requires careful selection of hyperparameters, and the model can become sensitive to noise in the data. Additionally, the interpretability of SVM is limited; understanding how input parameters influence the DRASTIC index can be difficult. RF, while robust and capable of handling nonlinear relationships, tend to produce complex ensemble models that lack transparency. It can be hard to ascertain the importance of individual parameters, which is critical for understanding groundwater pollution dynamics. Moreover, RF may not effectively model interactions unless they are explicitly captured in the trees. Deep Learning models excel at handling large datasets and capturing intricate patterns. However, they require significant amounts of training data, which may not be available in groundwater studies. Their "black box" nature makes it challenging to extract interpretable insights about parameter influences, reducing their utility for environmental management.

Overall, the selection of GEP, EPR, MARS, and MTM5 provides a robust framework for modeling the DRASTIC index. These models not only capture complex relationships among the input parameters but also offer greater transparency and interpretability, which are vital for effective groundwater resource management.

The novelty of the present paper lies in several key contributions that differentiate it from previously published works. Firstly, the integration of diverse machine learning methodologies—including classification, regression splines, and evolutionary computing—provides a comprehensive framework for enhancing DRASTIC index predictions. While prior studies may have focused on individual techniques, this paper offers a comparative analysis that highlights the strengths and weaknesses of each approach in the context of groundwater quality assessment. Secondly, the paper addresses specific limitations observed in the existing literature, particularly concerning the interpretability and robustness of predictions. By utilizing models like GEP, EPR, MARS, and MT, the present AI models not only improves predictive accuracy but also provides clear insights into the relationships among input parameters. This dual focus on performance and interpretability is a significant advancement over many existing models that prioritize accuracy at the expense of transparency. Moreover, the paper contributes to the field by incorporating a comprehensive set of nine input parameters that reflect the multifaceted nature of groundwater quality assessment. This nuanced approach contrasts with previous studies that may have simplified their analyses by using fewer parameters, thus potentially overlooking critical interactions. The literature review within the paper serves as a critical foundation, highlighting gaps in previous research and justifying the need for this novel approach. By systematically comparing their results with those from established studies, the authors demonstrate the superiority of their model in terms of predictive capability and practical applicability. Overall, this paper presents a significant advancement in the modeling of groundwater resource quality, bridging the gap between sophisticated machine learning techniques and practical environmental applications. Its contributions not only enhance the understanding of groundwater dynamics but also provide valuable tools for policymakers and resource managers striving for sustainable water management.

In this study, the standard DRASTIC framework is modified by land use and nitrate pollution effects. After that, four AI models (i.e., MARS, EPR, GEP, and MT) are applied to evaluate the situation of vulnerability of nitrate pollutions in the arid regions of Kerman plain, southwestern Iran. This research integrates the results with the land use map in order to foster precision degree in preparing the spatial distribution map of aquifer vulnerability risk.

Overview of case study Descriptions of under study plain

The study area of Kerman–Baghin plain, located on the edge of Lut Desert in the southeast of Iran, lies between 56.18° and 57.35° east longitude, and 29.46° and 30.32° north latitude, as seen in Fig. 1. The plain is located in the southern area of the Daranjir–Saghand basin, Kerman Province. In this basin, the cities of Kerman, Mahan, and Joupur are considered important population centers. The Kerman–Baghin plain spans an area of 5420 kilometers, of which 3200 square kilometers are alluvial surfaces and 2220 square kilometers are mountainous and foothill areas. The plain aquifer has an area of 2023.4 square kilometers. Additionally, the highest elevation is situated at 4200 m above sea level in the southern part of the plain whereas the lowest elevation in the central region of the Kerman–Baghin plain, reaching 1700 m above sea level.

The recharge of the plain is primary driven by rainfall. Running water flows through the riverbeds and existing channels in the region, including Seyedi River, Tigran Sekonj River, and Chari River. The annual average rainfall between 2015 and 2018 ranged from 120–140 mm, while in recent drought events, it has decreased to 90–100 mm. Data obtained from the nearest rain gauge station in the Sirch County show that the maximum and minimum values of temperatures during the summer are 39.5°C and 14°C , respectively. In addition, the maximum temperature in winter is 27°C while the minimum temperature reaches -10°C . The average temperature in winter is 2.9°C . The amount of humidity is 40% and the amount of annual evaporation is 2690.4 million mm per year and the amount of relative humidity is 28%.

Limitations and uncertainty sources

Various sources of limitation and uncertainties can affect the measurements of key parameters used in the computation of the DRASTIC index, which assesses groundwater vulnerability. Uncertainties can arise from the accuracy of well measurements, the timing of measurements (seasonal variations), and the spatial variability of the water table. Inconsistent monitoring techniques and equipment can also contribute to inaccuracies. In this study, due to the availability of field data limitations, the incomplete database within relevant organizations has resulted in some wells being excluded from calculations due to missing information. This gap underscores the need for comprehensive data collection practices to ensure that all relevant wells are accounted for in analyses. Additionally, conducting new pumping tests with appropriate spatial distribution is crucial for accurately understanding the hydraulic properties of the aquifer and for calculating hydraulic conductivity more precisely. Moreover, the situation is exacerbated by the fact that several piezometric wells (19 in total) have either dried up or have restricted access, highlighting the necessity for drilling new wells to monitor

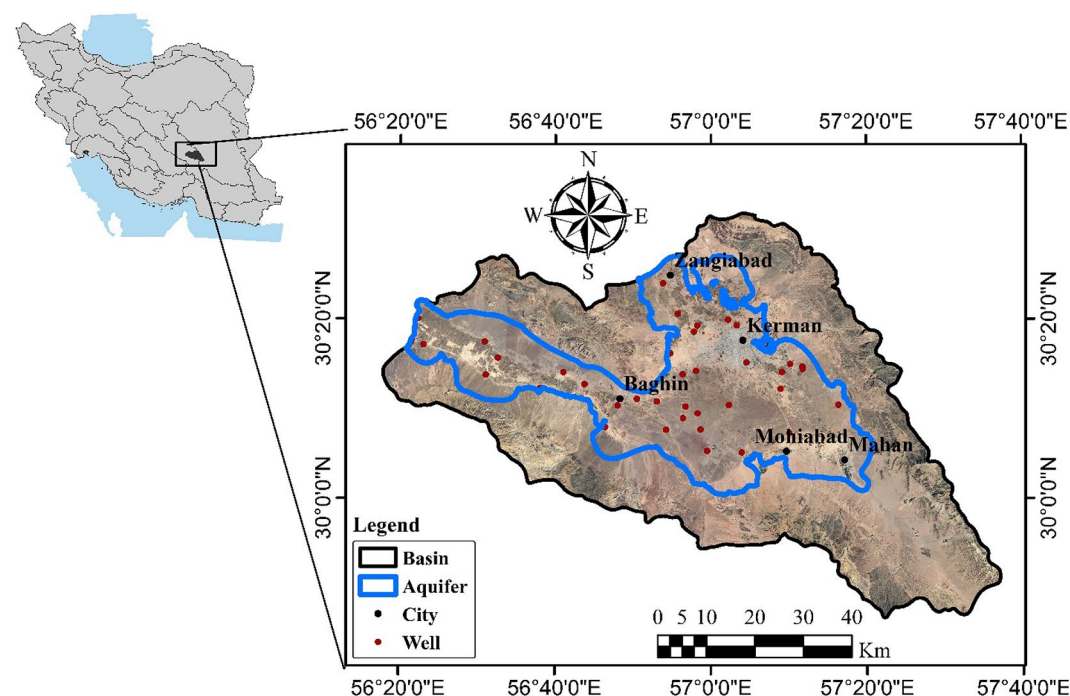


Fig. 1. Location of study area.

groundwater levels and other quality parameters. The insufficiency of exploratory wells further complicates the accurate recording of the aquifer's hydraulic characteristics. Estimating net recharge involves complexities such as variations in precipitation, evaporation, land use, and soil characteristics. Uncertainty in hydrological models used to calculate recharge, as well as limited long-term data, can further complicate accurate assessments. The geological complexity of aquifers can lead to uncertainties in characterizing their properties. Variability in rock types, fractures, and other geological features can affect how groundwater flows and is stored, complicating assessments of vulnerability. Furthermore, the limited number of drilling logs, coupled with inadequate spatial distribution relative to the area of interest, hampers the ability to assess the aquifer comprehensively. Moreover, soil properties, including texture, structure, and moisture retention capacity, can vary significantly across small distances. Inaccurate soil sampling and the use of generalized data can lead to uncertainties in determining how effectively soil can filter contaminants. On the other hand, the complexities of unsaturated zone dynamics can introduce uncertainty, particularly in estimating the thickness and properties of this zone. Variability in soil moisture content and saturation can complicate assessments of how pollutants may migrate toward the water table. Measuring hydraulic conductivity can be challenging due to spatial variability in subsurface materials. Inadequate sampling techniques or reliance on a limited number of measurements can lead to significant uncertainties in understanding how easily water and contaminants can move through the soil and aquifer. Furthermore, changes in land use practices can happen rapidly, and existing datasets may be outdated. Variability in agricultural practices, urban development, and natural land cover can all impact groundwater quality, but capturing these changes accurately in assessments can be difficult. There is also a significant lack of monthly monitoring of nitrate levels in the groundwater of the Kerman–Baghin plain, which is essential for assessing contamination risks and managing water quality effectively. Evaluating potential risks associated with land use involves uncertainty in identifying pollution sources, their proximity to groundwater, and the potential for contamination. Variability in land management practices and unforeseen events (like spills) can further complicate this assessment. In summary, the measurement of parameters integral to the DRASTIC index is fraught with uncertainties arising from spatial variability, temporal changes, limitations in data collection methods, and the inherent complexities of groundwater systems. Addressing these uncertainties through comprehensive monitoring and advanced modeling techniques is crucial for improving the reliability of DRASTIC index assessments.

Data utilized

The data used in this study include precipitation, depth to groundwater level, well logs, digital elevation model, hydraulic conductivity, soil texture, land use, and nitrate pollution. Monthly rainfall data for the years 2015 to 2018 were recorded from four rain gauge stations, while depth to groundwater level was recorded on a monthly basis from 38 piezometer wells during the same period. Drilling logs of piezometer wells, digital elevation model, and hydraulic conductivity were provided by Kerman Regional Water Company (KRWC). Soil texture was used for 117 samples prepared by Kerman Agricultural Research and Training Center. The land use map was prepared by the Forests, Rangelands and Watershed Management Organization.

DRASTIC index Composite DRASTIC index

The CD index is a modified version of the DRASTIC index, introducing a parameter (L) to specifically assess the potential risk related to the land use. The goal of this methodology is to assess the potential impact of widespread land use on aquifer quality, arising from changes in the soil matrix and unsaturated zone media over a period of time. The DRASTIC Index considers seven parameters related to hydrogeological conditions, namely Depth to water table (D), net Recharge (R), Aquifer media (A), Soil media (S), Topography (T), Impact of the vadose zone (I), and hydraulic Conductivity (C). Based on the impact of parameters on potential vulnerability, each parameter is attributed a relative numerical weight varying from 1 to 5, where 1 and 5 indicate the least and most influential, respectively. Moreover, these seven effective parameters are categorized into ranges and attributed a numerical value from 1 to 10 based on their influences on the susceptibility. Ultimately, following the collection and digitalization of hydrogeological data using GIS, the information is overlaid and integrated to generate vulnerability maps. The outcome is a new layer referred to as the DRASTIC index (Eq. 1). This index is expressed as follows,

$$\text{DRASTICindex} = D_r \cdot D_w + R_r \cdot R_w + A_r \cdot A_w + S_r \cdot S_w + T_r \cdot T_w + I_r \cdot I_w + C_r \cdot C_w \quad (1)$$

where subscripts “r” and “w” note the corresponding ratings and weights, respectively; that are summarized in Table S1, Supplementary Information (e.g.^{30,38,50,51}). To compute the Composite DRASTIC index, an extra parameter, namely land use, is incorporated into the DRASTIC index. Consequently, the CD index is determined as,

$$\text{DCindex} = D_r \cdot D_w + R_r \cdot R_w + A_r \cdot A_w + S_r \cdot S_w + T_r \cdot T_w + I_r \cdot I_w + C_r \cdot C_w + L_r \cdot L_w \quad (2)$$

In this context, L_r signifies the assigned rating for the potential risk linked with land use, L_w indicates the relative weight attributed to the potential susceptibility related to land use (as outlined in Table S2), and the remaining effective parameters are in accordance with Eq. (1). The ultimate results range from 28 to 280 and are categorized based on the classifications outlined in Table S3.

Nitrate vulnerability index

The NVI, as a modification of the DRASTIC index, designed to enhance precision in estimating the particular susceptibility to nitrate pollution. The NVI relies on the actual influence of each land use. This model aims to consolidate the risks of nitrate pollution in groundwater by taking into account land use as an expected nitrogen source. The framework considers both potential negative and positive impacts of land uses that do not contribute significantly to nitrate levels and do not lead to increased leaching over time, such as safeguarded natural zones. It relies on a framework based on multiplication that introduces a new parameter named the "potential risk associated with land use" (LU). This parameter is computed as³⁰,

$$\text{NVI} = (D_r \cdot D_w + R_r \cdot R_w + A_r \cdot A_w + S_r \cdot S_w + T_r \cdot T_w + I_r \cdot I_w + C_r \cdot C_w) \cdot \text{LU} \quad (3)$$

In this equation, LU represents the potential risk linked with land use (as specified in Table S4), while the remaining parameters remain consistent with Eq. (1). The ultimate results are categorized according to the classifications outlined in Table S3.

Preparation of vulnerability zone

Depth to water table (D)

The D parameters, as a pivotal factor in the DRASTIC framework, significantly affects the thickness of the unsaturated zone through which potential pollutants need to pass prior reaching the aquifer (e.g.^{48,52}). It imperatively represents the distance a contaminant would have to traverse to reach the water surface. This depth serves as an indicative factor for potential aquifer protection, with deeper water levels leading to longer travel times for pollutants. Generating the depth pixel map entailed interpolating existing data through the utilization of the Inverse Distance Weighting (IDW) technique, which is a spatial analysis tool in ArcGIS 10.8. The employment of this interpolation technique enabled the development of a seamless map depicting the depth to the water table throughout the research area. Following this, D values underwent reclassification, with rankings assigned on a scale from 1 to 10 (see Table S1). These rankings are essential for the subsequent assessment and computation of the NVI. It is the distance from the ground surface to the water table. In order to provide D layer, the latest information on water level of 38 piezometric wells has been applied. The locations of these wells were also presented in Fig. 1. The methodology based interpolation conception has been employed to change the mentioned point data into raster map of water level. Ultimately, this layer has been generated and then classified by the provided ranges in Table S1. Spatial variations of D values was presented in Fig. 2a. Over 90% of Kerman–Baghin plain area had $D > 30.5$ m.

Net recharge (R)

The net recharge parameter indicates the quantity of water that annually infiltrates from the Earth's surface to reach the saturated zone, which is particularly pertinent for understanding the various types of contaminations movements (convection and diffusion) from the unsaturated zone to the saturated zone (e.g.^{48,53}). In order to evaluate this parameter, the precipitation layer has been formed with aid of interpolation process on average values of the annual rainfall. The spatial variation of R map was then produced using the spatial analysis tool in ArcGIS 10.8, incorporating additional information on evaporation and runoff. The resultant map offers information regarding the average values of annual R across the entire study area. These values underwent reclassification and were assigned rankings within the range of 6–9, as illustrated in Fig. 2b and detailed in Table S1. In general, a higher R value indicates an elevated potential for contaminants to reach the water surface, given that larger volumes of infiltrating water create more pathways for the transport of pollutants. This study employs Piscopo's method⁵¹ to generate the R layer for the Kerman–Baghin Plain using Table S1 and the Eq. (4) provided below:

$$\text{Net Recharge index} = \text{Soil permeability} + \text{Rain fall} + \text{Slope} (\%) \quad (4)$$

In Eq. (4), the slope percentage was obtained from a Digital Elevation Model (DEM) created using the topographic map of the Kerman–Baghin Plain. Soil map, logarithmic observations, and exploration wells have been utilized in order to measure soil permeability. As seen in Fig. 2b, the map resulting from the spatial distribution shows that 53.2% of the plain area falls within the Class 3 range with a recharge level of 5 to 7. This class is uniformly scattered across the plain's surface.

Aquifer media (A)

The A parameter describes the characteristics of materials within the aquifer that impact the processes of pollutant attenuation, as detailed by reliable literature (e.g.^{36,37,48}). The aquifer formations, recognized using lithology and hydrogeology maps, consist of fine silt and silty sand, sand, conglomerate, and shale. According to Table S1, the spatial variations of A map has been categorized into six groups and then ranged from 1 to 9, as illustrated in Fig. 2c. In this study, information from 35 well-log profiles (Lithological wells) of the alluvial aquifer prepared in 1964 from KRWG has been employed in order to generate the A layer. A superior rank was attributed to conglomerate, indicating a highly coarse porous medium with outstanding capabilities in drainage and transmission. The spatial distribution map of the aquifer (Fig. 2c) illustrates that the aquifer environment covers the majority of the plain's surface (61.8%), particularly in its eastern region, comprising Gravel, Sand, Silt, and Clay. In the western half of the plain (37%), it mainly includes Gravel and Sand.

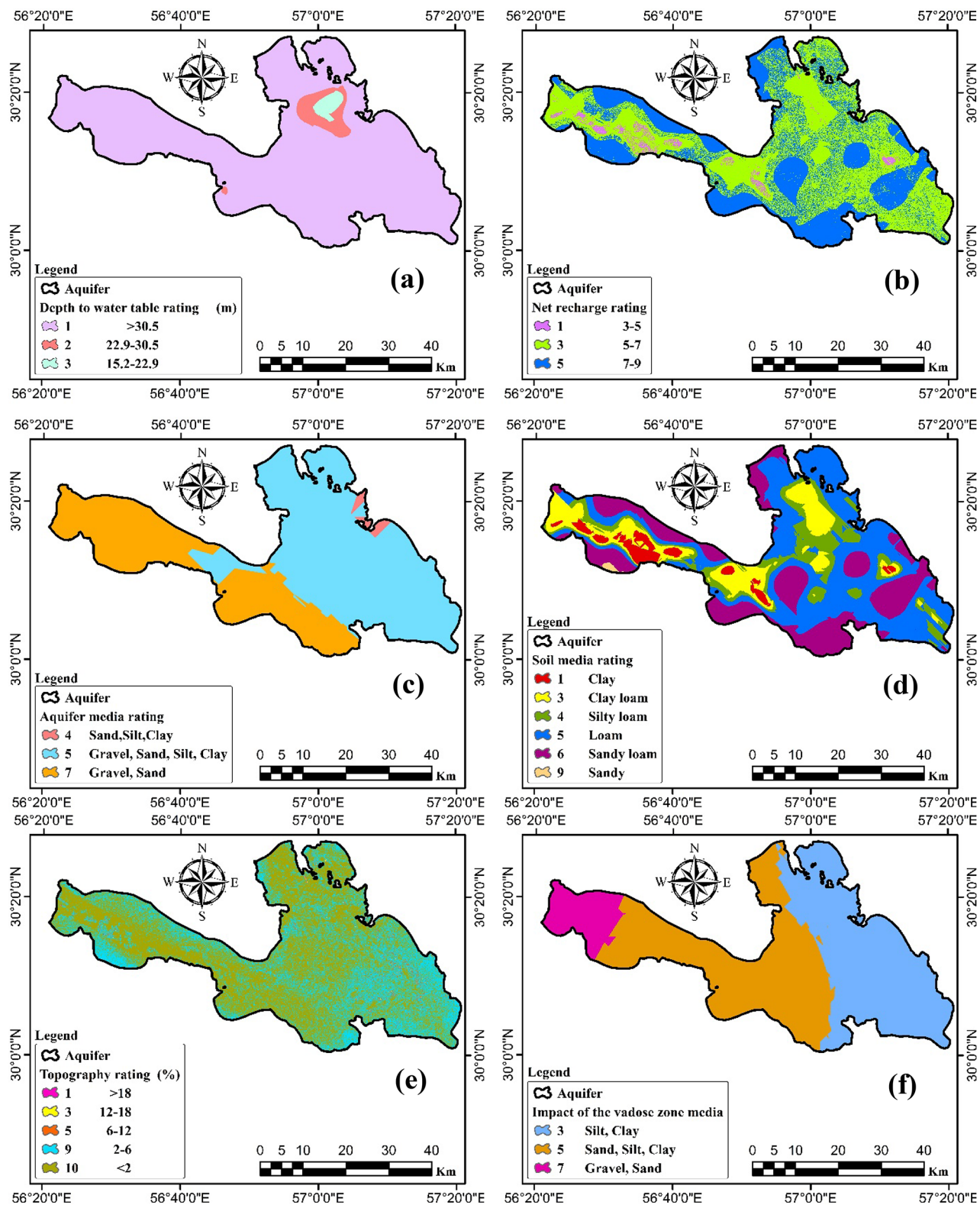


Fig. 2. Input maps for the various DRASTIC and NVI models: (a) depth to water table, (b) recharge, (c) aquifer media, (d) soil media, (e) topography, (f) impact of vadose zone, (g) hydraulic conductivity, (h) land use, and (i) spatial variation of nitrate pollution.

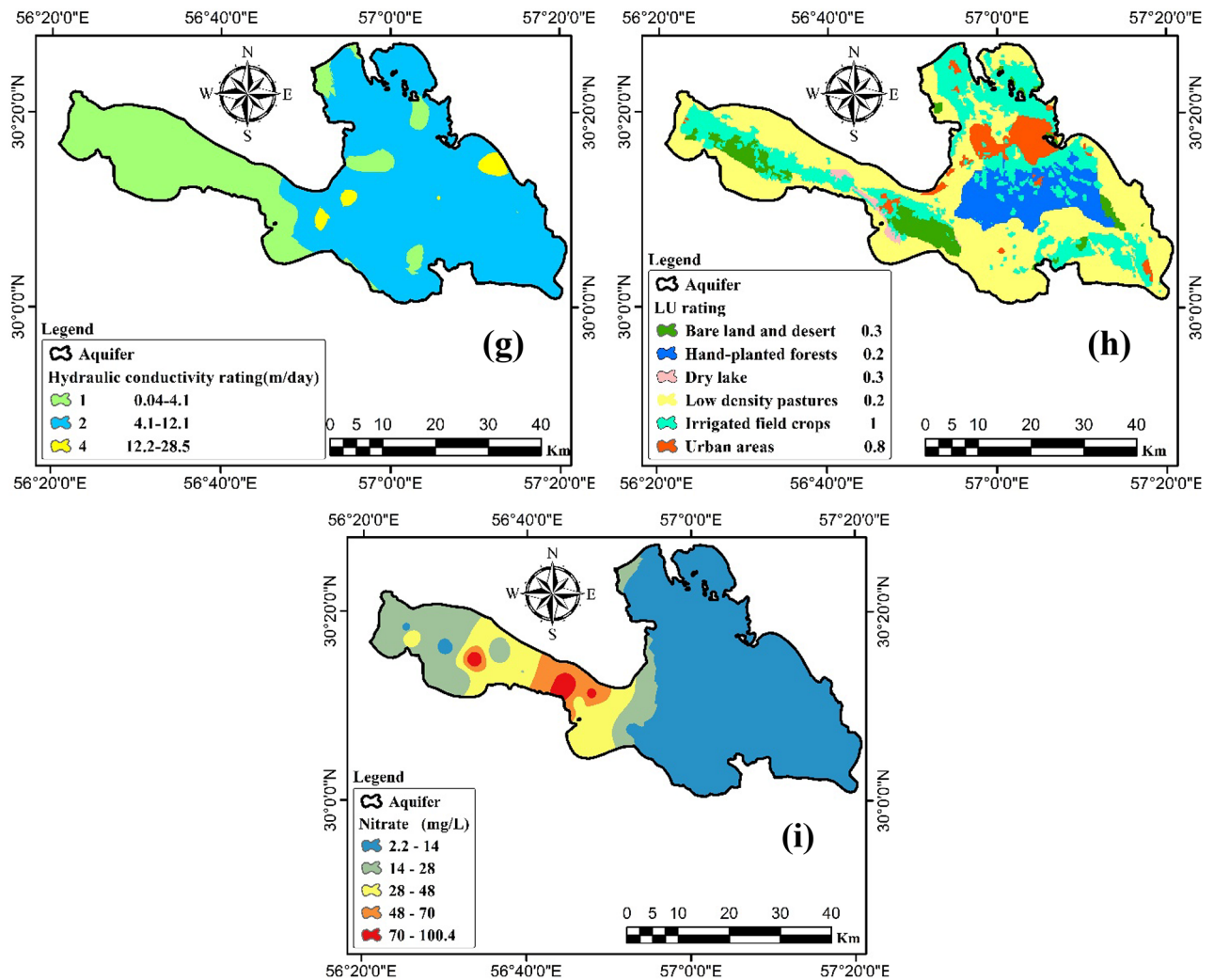


Figure 2. (continued)

Soil media (S)

The S parameter plays a critical role in characterizing the influx of nutrients and pollutants into the aquifer, influencing the purification processes, as emphasized by diverse investigations (e.g.^{1,3–5}). Additionally, soil characteristics impact the removal of pollutants. The data has been obtained from the disciplines of environmental and geological science. Rankings and weights have been designated according to the hydrological characteristics (porosity and permeability) of the soil. Table S1 present the rankings and weights related to various classes of S values and additionally Fig. 2d illustrates spatial variation map of soil media. Sandy loam soils, possessing favorable drainage that facilitates substantial water flow, were assigned a heightened ranking in terms of contaminant transmission. Conversely, clay soils hinder water flow and are less prone to groundwater contamination, leading to a lower rating. Soil data were obtained from the Kerman Institute of Agricultural Science (KIAS). Spatial variations of soil texture has been mapped by using the soil information derived from log of observation and exploration wells. The soil environment map in Fig. 2d illustrates that the majority of the plain's surface (41.2%) is characterized by loamy soil texture, and this soil exhibits dispersion throughout all parts of the plain. Sandy soil (0.3%) with Class 9 is allocated the smallest area.

Topography (T)

This parameter signifies the incline of the terrain in the studied districts and then affects the quantity of water on the capability of soil infiltration without any intermediary. A gentler slope encourages increased infiltration, leading to an increased likelihood of pollutants migrating into the aquifer (e.g.^{20,54}). To create the T layer, a digital elevation model (DEM) has been employed, and the slope function in ArcGIS 10.8 has been utilized to categorize it into five ranges. Steeper slopes are less likely to experience contamination because of the heightened potential for surface runoff. Figure 2e and Table S5 indicates the map and rankings/weights assigned to the T parameter, respectively. DEM with spatial resolution (30 m) has been utilized in order to calculate T values in ArcGIS 10.8 software. The T layer was created using the same methodology applied in generating the R layer and subsequently underwent classification. As illustrated in Fig. 2e, the slope domain map indicates that the majority

of the plain's surface (60%) has slopes less than 2%, falling into Class 10. Slopes in the ranges of 12–18% and greater than 18% constitute the smallest area (0.1%).

Impact of vadose zone (I)

The I layer is created by considering characterizations derived from hydrogeology information. The spatial variations on geological properties of the studied plain has been grouped into three ranges: Conglomerate, Marl, and calcareous shale with intercalations of limestone, and Marl, shale, sandstone, and limestone. Each category signifies unique geological units situated above the groundwater level. The vadose zone parameter map delineates the diverse geological units above the water level, as discussed by Karimzadeh-Motlagh et al.⁴⁸. The vadose zone includes the area between the Earth's surface and the aquifer, specifically referring to the unsaturated material above the water table. The dimensions and depth of the unsaturated zone have considerable importance in the computation process of DRASTIC framework. The properties of the vadose zone media have a profound impact on both water flow and the migration of pollutants. Each geological unit has been attributed to a score between 3 and 9, with permeable media such as gravel and conglomerate receiving a score of 9, while shale has been attributed to a lower rank of 3 (refer to Table S1). This research utilized lithologic data obtained from observation and exploration wells to construct the vadose zone media of the Kerman–Baghin plain. Subsequently, utilizing this information in conjunction with Table S1, we formulated the raster map for the Kerman–Baghin plain (see Fig. 2f). The unsaturated zone map on the plain's surface (Fig. 2f) indicates that the eastern half of the plain (45.7%) is characterized by Silt and Clay textures. The central to western part of the plain contains Sand, Silt, and Clay textures (44.6%), while the westernmost layer of the plain consists of Gravel and Sand (9.7%).

Hydraulic conductivity (C)

The C parameter serves as an indicator of the porous medium's ability within an aquifer to transport water, as noted by Baghapour et al.³⁰, Wang et al.⁵⁵, and Karimzadeh-Motlagh et al.⁴⁸, and it is considered as contributory factor of assessing the movement of contaminants. Elevated C values indicate a faster transportation of pollutants. Hydraulic conductivity denotes the ability of soil or rock to facilitate the movement of water, a feature affected by factors such as the proportion of pore spaces, interconnections, voids, grain size, and sorting. The C values underwent categorization, and a rating between 1 and 7 was allocated to each aquifer media type with consideration of hydraulic conductivity characterizations (refer to Table S1). We utilized data from 22 pumping tests conducted by KRWC to create the hydraulic conductivity layer. The spatial variation of C values is illustrated in Fig. 2g and additionally demonstrated that the majority of the plain's surface (67.1%) in the eastern half and central parts has hydraulic conductivity ranging from 1.12 to 1.4 m per day. The smallest area (1.7%) corresponds to hydraulic conductivity levels of 12.2–28.5 m per day, scattered in small patches in the eastern and central regions of the plain.

Land use

A land use map is a visual representation of how the land in a specific geographic area is utilized or occupied by different human activities. It illustrates the various types of land uses, such as residential, commercial, industrial, agricultural, recreational, or natural areas. Land use maps provide a comprehensive overview of the spatial distribution and patterns of human activities on the Earth's surface. The map shows the spatial distribution of land uses, indicating where specific activities or land cover types are located within the study area. In urban planning, land use maps often align with zoning regulations. Zoning boundaries define areas with specific land-use designations, such as residential, commercial, or industrial zones. Land use maps help assess the influence of human activities on the natural surroundings and ecological systems. They are valuable for conservation efforts and natural resource management. This layer is essential as it is needed for both CD and NV indices. This layer was obtained from Landsat images in 2018 and then Tables S2–S4 were utilized to assign ratings for the creation of the land use map and its corresponding risk assessment. As seen in Fig. 2h, the vast area of the Kerman–Baghin plain is red with low density pasture. The land use map of the watershed (Fig. 2h) indicates that the majority of the watershed area (49.9%) is characterized by low-density grasslands, and urban land use class constitutes the smallest area (6%).

Field measurement of nitrate concentration

Groundwater sampling was conducted at selected well locations during the summer of 2022 (from August 6 to August 10), as part of field operations. It is important to highlight that the investigation into the groundwater quality in the investigated region, characterized by an arid climate, took place during the dry season. This method aimed to evaluate the water quality over the long-term, alleviating the influence of short-term variations induced by aquifer recharge and fluctuations in water quality typically observed during the humid season, especially after severe precipitation. Dark glass containers with a volume of 250 ml were utilized for sampling, pre-washed with distilled water before collection. Each sample bottle has been marked with details including location, sampling time (date and time), sampling depth (varying from 22 to 146 m), and water temperature. Following collection, the water samples have been stored at a temperature below 4 °C and transferred to the central laboratory of the Graduate University of Advanced Technology within 6 h. Nitrate content analysis was promptly conducted using a Varian Cary 50 spectrophotometer⁵⁶. Subsequently, geostatistical analysis of the nitrate data was carried out to generate a zonation map using ArcGIS 10.4.1 and GS+ software packages. It is important to mention that the geostatistical technique, particularly the Kriging estimator, has been extensively used to assess the spatial and temporal variations of groundwater. The spatial variations of nitrate concentrations (mg/l) for Kerman–Baghin plain was illustrated in Fig. 2i. The major fraction of the plain had 2.2–14 mg/l Nitrate concentration whereas the 70–100.4 nitrate concentration had minimum coverage.

Dataset overview

From Fig. 3, Table 2 indicates fractional areas of effective DRASTIC parameters for the case study area. In this research, DRASTIC index which has been effected by NV and LU indices, is predicted by AI models. To begin with, normal effective parameters (i.e., D, R, A, S, T, I, and C) were employed to provide map of DRASTIC index and then effects of LU and nitrate pollutions were considered to provide the spatial zoning map of nitrate pollution. To develop AI models for estimating Nitrate pollution, 9 input parameters were considered: D_r , R_r , A_r , S_r , T_r , I_r , C_r , L_r , and LU. In this way, a dataset containing 100 data series were used for this purpose in which 75 and 25% of dataset have been selected to carry out the training and testing stages of the AI models, respectively. Statistical descriptions of input–output parameters were presented in Table 3

Artificial intelligence models

Multivariate adaptive regression spline

The MARS model utilizes the concepts of piecewise linear regression to formulate a linear regression expression, identifying overall patterns within the input–output system^{57,58}. The MARS method estimates the function by employing a sequence of piecewise linear regressions, accomplished through an adaptive process referred to as Basis Functions (BFs). The curve-fitting component of MARS is essentially built by a collection of Basis BFs in both the forward and backward stages. Each BF comprises a single variable (x) and a knot (K), resulting in two potential pairs: $(z - K)_+$ and $(K - z)_+$, where $(z - K)_+$ equals $(z - K)$ if $z > K$, and 0 otherwise; and $(K - z)_+$ equals $(K - z)$ if $z < K$, and 0 otherwise. The final result from the MARS model is determined through a curve-fitting equation, as described by Friedman⁵⁸.

$$NVI = \omega_0 + \sum_{i=1}^s \omega_i \cdot BF_i(D_r, R_r, A_r, S_r, T_r, I_r, C_r, LU, L_r) \quad (5)$$

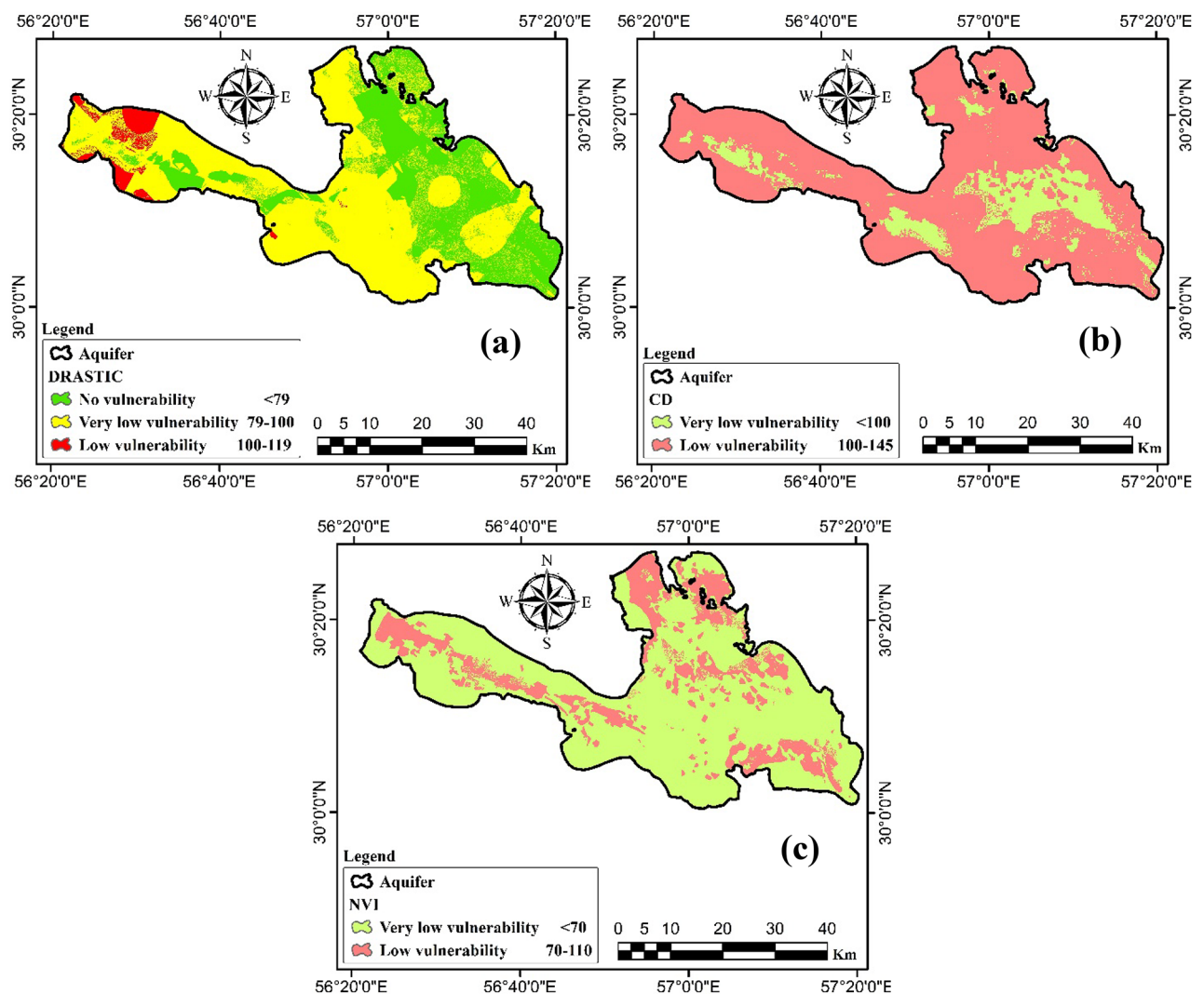


Fig. 3. Spatial variations of (a) DRASTIC, (b) DRASTIC-LU, and (c) NVI models.

Depth to water table (m)		Net recharge		Aquifer media	
Rating	Area (percentage %)	Rating	Area (percentage %)	Rating	Area (percentage %)
1	1922.2 (95)	1	57.497 (2.8)	4	24.134 (1.2)
2	73.08 (3.6)	3	1076.4 (53.2)	5	1251.4 (61.8)
3	28.113 (1.4)	5	889.49 (44)	7	747.88 (37)
Soil media		Topography (slope %)		Impact of the vadose zone	
Rating	Area (percentage %)	Rating	Area (percentage %)	Rating	Area (percentage %)
1	82.91 (4.1)	1	1.18 (0.1)	3	923.75 (45.7)
3	271.72 (13.4)	3	1.86 (0.1)	5	903.42 (44.6)
4	281.5 (13.9)	5	31.24 (1.5)	7	196.26 (9.7)
5	833.57 (41.2)	9	774.8 (38.3)		
6	547.64 (27.1)	10	1214.3 (60)		
9	6.094 (0.3)				
Hydraulic conductivity		Land use		NO ₃ ⁻ (mg/L)	
Rating	Area (percentage %)	Rating	Area (percentage %)	Rating	Area (percentage %)
1	630.4 (31.2)	Bare land and desert	143.34 (7.1)	2.2 - 14	1319.2 (65.2)
2	1358 (67.1)	Hand-planted forests	210.29 (10.4)	14 - 28	341.52 (16.9)
4	35 (1.7)	Dry lake	17.611 (0.9)	28 - 48	248.53 (12.3)
		Low density pastures	1010.3 (49.9)	48 - 70	86.247 (4.3)
		Irrigated field crops	520.58 (25.7)	70 - 100.4	27.919 (1.4)
		Urban areas	121.28 (6)		

Table 2. Fractional areas for effective DRASTIC parameters.

Parameters	Max	Min	Average	Std. dev	Skewness
D _r	3	1	1	0.41	3.6
R _r	5	1	4	1.2	-0.23
A _r	7	4	6	1.01	0.38
S _r	9	1	5	1.36	-1.2
T _r	10	5	10	0.67	-3.39
I _r	7	3	4	1.26	0.33
C _r	4	1	2	0.62	1.04
L _r	8	1	6	2.22	-1.13
LU	1	0.2	0.5	0.36	0.63
NVI	104	13.6	39.62	29.6	0.76

Table 3. Statistical descriptions of input–output parameters obtained in the Baghin Plain.

in which BF_i notes the basis function including input variables with the second-order polynomial pattern. Additionally, ω_0 , ω_i , and s note the bias term, the constant coefficient related to the typical basic functions, and the number of basis functions in the MARS model, respectively.

The MARS approach implemented using MATLAB2008a whose programming codes are freely available. In the present research, MARS produces 18 BFs using a second-order polynomial (or linear model) in order to approximate NVI values. The mathematical formulations for the BFs and their associated constant coefficients employed in NVI modeling are outlined in Table 3. Furthermore, the subsequent mathematical model establishes a correlation between NVI and nine influential parameters.

Formulation of BFs	
BF ₁	max(0, LU − 0.8)
BF ₂	max(0, 0.8 − LU)
BF ₃	max(0, LU − 0.8) × max(0, R _r − 3)
BF ₄	max(0, I _r − 5)
BF ₅	max(0, 5 − I _r)
BF ₆	max(0, 0.8 − LU) × max(0, I _r − 5)
BF ₇	max(0, 0.8 − LU) × max(0, 5 − I _r)
BF ₈	max(0, R _r − 3)
BF ₉	max(0, 3 − R _r)
BF ₁₀	max(0, C _r − 2)
BF ₁₁	max(0, 0.8 − LU) × max(0, C _r − 2)
BF ₁₂	max(0, 0.8 − LU) × max(0, 2 − C _r)
BF ₁₃	max(0, LU − 0.8) × max(0, D _r − 1)
BF ₁₄	max(0, A _r − 5)
BF ₁₅	max(0, 5 − S _r)
BF ₁₆	max(0, 0.8 − LU) × max(0, R _r − 3)
BF ₁₇	max(0, 0.8 − LU) × max(0, 3 − R _r)
BF ₁₈	max(0, 0.8 − LU) × max(0, 5 − S _r)

Table 4. List of BF developed in MARS model.

k-fold	Number of effective parameters	Number of BFs in final model	Time performance
1	43.5	18	0.65
2	38.5	16	0.69
3	28.5	12	0.61
4	38.5	16	0.63
5	43.5	18	0.66
6	46	19	0.85
7	46	19	0.52
8	43.5	18	0.56
9	28.5	12	0.31
10	41	17	0.37

Table 5. Characterization of MARS model within forward and backward passes.

$$\begin{aligned} NVI = & 73.03 + 44.331 \times BF_1 - 95.044 \times BF_2 \\ & + 8.0589 \times BF_3 + 5.4617 \times BF_4 - 5.1886 \times BF_5 \\ & - 7.5732 \times BF_6 + 7.121 \times BF_7 + 3.024 \times BF_8 \\ & - 3.8399 \times BF_9 + 4.0564 \times BF_{10} - 6.0339 \times BF_{11} \\ & - 1.418 \times BF_{12} + 25.861 \times BF_{13} + 0.81522 \times BF_{14} - 1.4806 \times BF_{15} \\ & - 3.1012 \times BF_{16} + 4.7424 \times BF_{17} + 1.8296 \times BF_{18} \end{aligned} \tag{6}$$

In the present study, the k-fold value was specified as 10, indicating that the MARS procedure was repeated 10 times to mitigate the risk of overfitting. Both forward and backward stages were executed for each k-fold setting. Furthermore, the determination of the final MARS model involved establishing the number of basis functions and identifying the count of effective parameters. The optimal model derived from the MARS approach for assessing NVI is expressed by Eq. 6, with BF₁ to BF₁₈ detailed in Table 4. The basis functions in Eq. 6 are established using input parameters LU, C_r, I_r, S_r, A_r, R_r, and D_r, while T_r and L_r are omitted. Table 5 provides insights into the features observed in 10 iterations of the MARS model, presenting outcomes across both forward and backward stages. All coefficients in Eq. 6 were determined through the Particle Swarm Optimization (PSO) algorithm, resulting in an MSE (Mean Squared Error) of 0.35661 as the optimal outcome. In Fig. 4, within the aquifer's total area of 4202.3 square kilometers, around 3147.8 square kilometers (approximately 74.73%) in

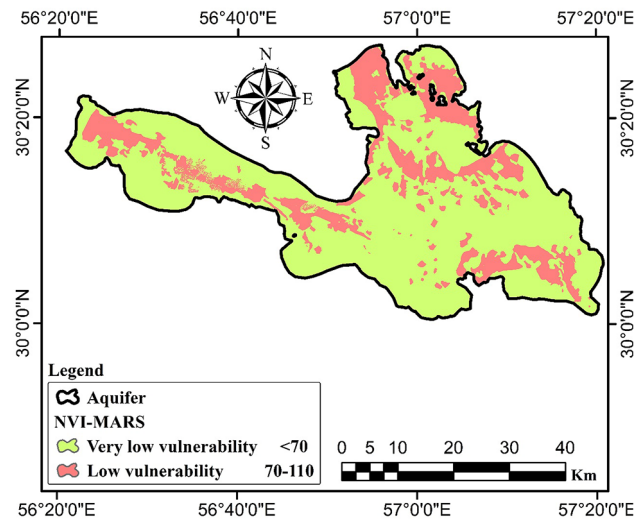


Fig. 4. Spatial variation of Nitrate vulnerability pollution by MARS.

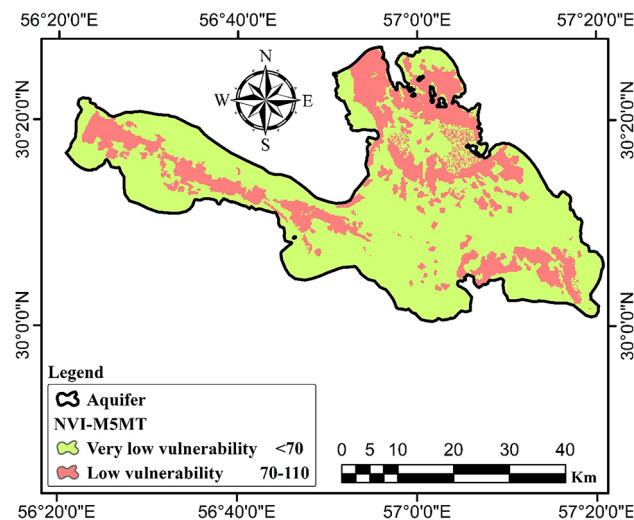


Fig. 5. Spatial variation of Nitrate vulnerability pollution by M5MT.

the region characterized by very low vulnerability exhibit an NVI value below 70. The remaining 1,545 square kilometers (25.26% of the total area) fall within the NVI range of 70–110, indicating low vulnerability.

Model tree

Quinlan⁵⁹ introduced the MT as a powerful data-mining model rooted in classification principles. A more sophisticated extension of MT, known as the M5 model, was subsequently proposed by Wang and Witten⁶⁰. The M5 Model Tree (M5MT) is capable of breaking down a complex problem into manageable domains or subspaces, with the final output being a linear summation of the input variables throughout all realms of inputs⁵⁷. The M5MT technique employs recursive methods in order to divide the search space of datasets into one or more subspaces. This involves using if–then rules to categorize input variables into one or more domains, specifying input variables to subdomains. Consequently, a set of multilinear relationships is generated within each sub-domain.

In the current study, the M5MT approach has been implemented using Weka 3.9 software. Initially, the decision tree is formed during the training phase, and subsequently, nine influential parameters (i.e., LU , C_p , I_p , S_p , A_p , R_p , D_p , T_p and L_p) were employed to acquire the multilinear regression model for classification. In order to approximate NVI, Eq. (7) is expressed as

$$NVI = \begin{cases} -9.6162 + 1.3959R_r + 0.1561S_r + 0.7123I_r + 86.5834LU & L_r \leq 7 \\ -11.6632 + 4.9336R_r + 4.9111I_r + 54.8834LU & L_r > 7 \end{cases} \quad (7)$$

In Eq. (7), the M5 tree model utilizes the input parameter L_r as a criterion to partition the input space, and the parameters LU , I_r , S_r , and R_r are incorporated into the mathematical expression. However, the input parameters C_r , T_r , A_r , and D_r are not considered. In Fig. 5, within the total area of 4,202.3 square kilometers of the aquifer, an area of approximately 3,141.5 square kilometers (approximately 94.69%) in the region characterized by very low vulnerability has an NVI value below 70. The remaining 608 square kilometers (around 5.30% of the total area) fall within the NVI range of 70 to 110, signifying a low vulnerability zone.

Gene-expression programming

The GEP approach is essentially built upon the principles of Genetic Algorithm (GA), a technique capable of simplifying relationships within sophisticated systems into simpler equations. GEP integrates inherent features of GA, encompassing fixed-length and linear configurations linked to chromosomes. Moreover, GEP utilizes Expression Trees (ETs) with sub-elements of diverse sizes and shapes, including aspects such as tree depth, mathematical expressions and algebraic symbols. The GEP model is constructed through five distinct phases. In the initial stage, a fitness function (FF) is employed in order to appraise an individual program^{61–63}. In this investigation, the Mean Squared Error (MSE) is assigned as the fitness function,

$$FF = \frac{1}{1 + MSE} \times 1000$$
 (8)

The second phase commences with the specification of terminals and mathematical functions so as to provide the chromosomes. During the third stage, the overall attributes of chromosomes, such as size and shape, are configured. Following that, the fourth stage utilizes a well-established algebraic operator to construct a mathematical expression from a set of genes. Lastly, various genetic operators (as delineated in Table 6) are applied to obtain the optimal relationship. The GEP model, implemented using GeneXproTools 5 software, produced the most efficient correlation for predicting the influential parameters of NVI. In this research, the number of generations was extended to 244 until the optimal fitness function value reached 214.665 during the training phase.

$$NVI = \max (LU - (-8.70967), A_r) + \left[((LU + LU) \times L_r) \times \left(\frac{A_r + R_r}{2} \right) \right] + 1 - \left[\min \left(\left(\frac{L_r}{LU \times I_r} \right), (A_r + 0.034084) \right) - \left(\frac{T_r + A_r}{L_r + LU} \right) \right] + \frac{\exp \left(\frac{(S_r - (-5.2031)) + \exp(-9.3446)}{2} \right) + (-5.2031)}{2} + (0.891546 - R_r)$$
 (9)

In the GEP model, Eq. (9) incorporates input parameters LU , L_r , I_r , T_r , S_r , A_r , and R_r , while excluding the input parameters D_r and C_r . In Fig. 6, within the total area of 4,202.3 square kilometers of the aquifer, approximately 1,461 square kilometers (around 34.72%) in the area characterized by very low vulnerability exhibit an NVI value below 70. The remaining 562 square kilometers (approximately 65.27% of the total area) fall within the NVI range of 70 to 110, indicating low vulnerability.

Evolutionary polynomial regression

EPR, grounded in a global search technique, is capable of constructing a symbolic regression equation linking input–output vectors (or variables) so as to make systems straightforward. Conceptually, EPR generates a regression equation comprising numerous algebraic terms^{64–68}:

GEP parameters	Values
Number of chromosomes	30
Linking function	+
Mutation	0.00138
Gene recombination	0.00277
Fixed-root mutation	0.0068
Fitness function	Mean Square error
Best fitness function	510.739
Number of generation	576
Maximum depth of tree	4
Number of genes	4
Termination condition of GEP development	R-squared
Mathematical operators	Exp, LN, X ² , 3RT, Atan, Min(x ₁ , x ₂), Max(x ₁ , x ₂), Avg(x ₁ , x ₂), Tanh, NOT, Inv, +, -, *, /

Table 6. Setting parameters of GEP model.

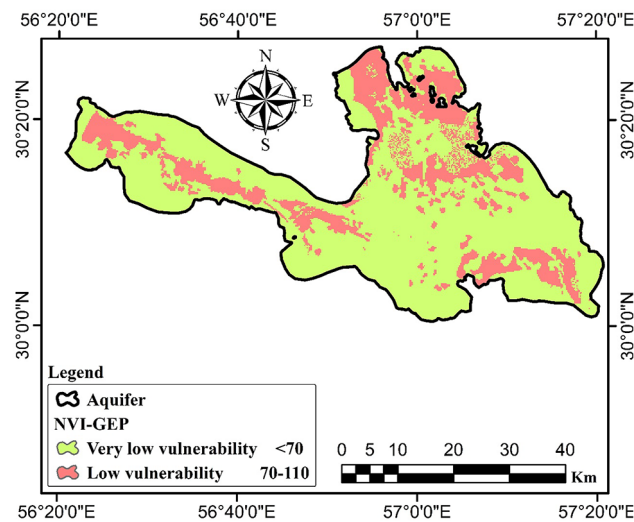


Fig. 6. Spatial variation of Nitrate vulnerability pollution by GEP.

$$NVI = \theta_0 + \sum_{g=1}^T \theta_g \cdot (D)^{EX(g,1)} \cdot (R)^{EX(g,2)} \cdot (A)^{EX(g,3)} \dots (LU_g)^{EX(g,T)} \cdot INF \left(D_1^{EX(g,G+1)} \cdot (R)^{EX(g,G+2)} \cdot (A)^{EX(g,G+3)} \dots LU_g^{EX(g,2G)} \right) \quad (10)$$

in which U , L_0 , L_u , INF , and EX note the customizable symbolic terms, an optional bias term, a group of invariable coefficients, the customizable inner function (e.g., natural logarithm, tangent hyperbolic), and customizable exponents input vectors applied in Eq. (7), respectively^{69,70}. During the development stage, a matrix is created that encompasses effective DRASTIC parameters. Following this, initial population values associated with the exponent vectors are provided. In the subsequent step, exponent vectors associated with influential DRASTIC parameters are specified; then, invariable coefficients (L_u), presented in Eq. (10), are adjusted by the least square technique. After fixing the L values and ES vectors, the values of NVI are estimated in the training phase. Subsequently, the quality provided by the EPR performance is evaluated. In this study, a logarithmic inner function (INF) is employed to deliver the most accurate estimation of the NVI , as opposed to other types of INF s such as tangent and secant hyperbolic functions and exponential function^{64–66}.

In this research, the EPR model was implemented using the EPR MOGA-XL software, which is programmed within the Excel environment. The execution of the EPR model involved the utilization of the multi-objective genetic algorithm (MOGA). The software settings were configured with an inner function (natural logarithm) defined, a power range for input variables spanning from -2 to 2 with increments of 0.5 , and a total of $19,440$ generations. Additionally, a maximum of six mathematical terms ($U=6$) was specified. Through this software, six models were generated for the evaluation of NVI , as presented in Table 7. Model#6, exhibiting the lowest mean squared error ($MSE=0.066$) during training, outperforms the other models and is selected as the most accurate model for NVI evaluation. In Model#6 (as presented in Table 7), input parameters LU , C_r , I_p , T_p , A_p , R_p , and D_r were used for modeling nitrate pollution, while S_r and L_r were excluded. In Fig. 7, out of a total area of $4,202.3$ square kilometers of the aquifer, an area of approximately $1,521$ square kilometers (about 2.75%) in the region with very low vulnerability has an NVI value less than 70 . A total of 501 square kilometers (about 79.27% of the total area) falls within the NVI range of 70 to 110 in the low vulnerability zone. These models perform a refinement in the final model by considering the impact of each input parameter on vulnerability assessment through NVI , resulting in some input parameters participating in the mathematical expression, while others are excluded.

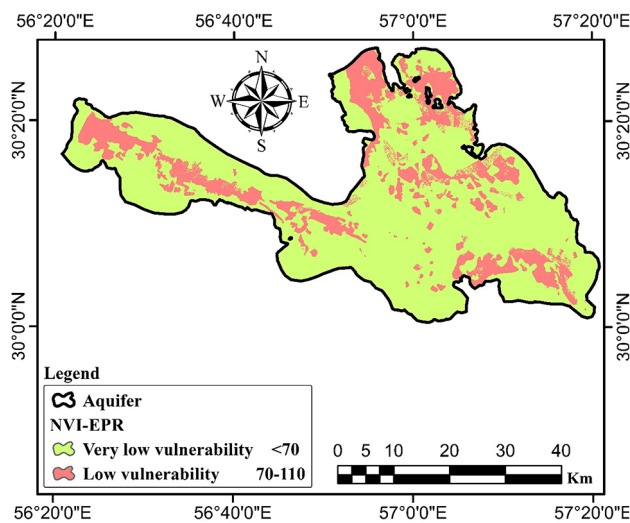
Results and discussions

Statistical performance of AI models

According to reliable relevant research works, various statistical tests have been applied in order to appraise the performance of AI models in the groundwater evaluation in terms of quality and quantity literature (e.g.^{1,24,30,31,39,63,71–76,78}). Through this method, statistical indicators such as the Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Scatter Index (SI) were employed to evaluate the effectiveness of the performance of soft computing models in two development stages (i.e., training and testing stages) related to the assessment of NVI . These statistical metrics are defined as follows:

$$R = \frac{\sum_{i=1}^N (NVI_p^i - \overline{NVI_P}) (NVI_o^i - \overline{NVI_O})}{\sqrt{\sum_{i=1}^N (NVI_o^i - \overline{NVI_O})^2 (NVI_p^i - \overline{NVI_P})^2}} \quad (11)$$

Model.no	Formulations	MSE
Model#1	$NVI = (4.2744 \times \ln(LU^{0.5})) + (7.5974C_r^{0.5}LU^{1.5})$ $+ (0.58939C_r \times \ln(C_r^{0.5})) + (3.4229S_r^{0.5}LU \times \ln(R_r^2))$ $+ (20.0097A_r^{0.5}LU \times \ln(I_r^{0.5})) + (4.6166D_r^{0.5}T_r^{0.5}LU) + 5.166$	0.123
Model#2	$NVI = (0.51714 \times \ln(C_r)) + (3.0519C_r^{0.5}LU^{1.5} \times \ln(C_r))$ $+ (6.491T_r^{0.5}LU) + (3.4285S_r^{0.5}LU \times \ln(R_r^2))$ $+ (22.1852A_r^{0.5}LU \times \ln(I_r^{0.5})) + \left(26.2349 \frac{D_r \times LU}{I_r^{1.5}}\right) + 0$	0.089
Model#3	$NVI = (0.5473 \times \ln(C_r)) + (2.411C_r^{0.5}LU^{1.5} \times \ln(I_r^{0.5} \times C_r))$ $+ (6.2099T_r^{0.5}LU) + (3.4313S_r^{0.5}LU \times \ln(R_r^2))$ $+ (21.5222A_r^{0.5}LU \times \ln(I_r^{0.5})) + \left(26.9076 \frac{D_r \times LU}{I_r^{1.5}}\right) + 0.19457$	0.080
Model#4	$NVI = (2.4808C_r^{0.5}LU^{1.5} \times \ln(I_r^{0.5} \times C_r)) + (6.1836T_r^{0.5}LU)$ $+ (3.3946S_r^{0.5}LU \times \ln(R_r^2)) + (10.7933A_r^{0.5}LU \times \ln(I_r))$ $+ (0.13577R_r \times \ln(C_r)) + \left(27.139 \frac{D_r \times LU}{I_r^{1.5}}\right) + 0.21202$	0.071
Model#5	$NVI = (2.3201C_r^{0.5}LU^{1.5} \times \ln(I_r^{0.5} \times C_r)) + (6.0645T_r^{0.5}LU)$ $+ (3.3586S_r^{0.5}LU \times \ln(R_r^2)) + (10.9199A_r^{0.5}LU \times \ln(I_r))$ $+ (0.29264R_rLU^{0.5} \times \ln(C_r)) + \left(27.4567 \frac{D_r \times LU}{I_r^{1.5}}\right) + 0.24993$	0.069
Model#6	$NVI = (2.3307C_r^{0.5}LU^{1.5} \times \ln(I_r^{0.5} \times C_r)) + (6.2175T_r^{0.5}LU)$ $+ (3.3575S_r^{0.5}LU \times \ln(R_r^2)) + (10.8407A_r^{0.5}LU \times \ln(I_r))$ $+ (0.2844R_rLU^{0.5} \times \ln(D_r^{0.5} \times C_r)) + \left(25.9049 \frac{D_r \times LU}{I_r^{1.5}}\right) + 0.24597$	0.066

Table 7. Mathematical expressions given by EPR.**Fig. 7.** Spatial variation of Nitrate vulnerability pollution by EPR.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (NVI_p^i - NVI_o^i)^2}{N}} \quad (12)$$

$$MAE = \frac{\sum_{i=1}^N |NVI_p^i - NVI_o^i|}{N} \quad (13)$$

$$SI = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N [(NVI_p^i - \overline{NVI_P}) - (NVI_o^i - \overline{NVI_O})]^2}}{\overline{NVI_O}} \tag{14}$$

In these equations, N is the number of data points, the NVI_o^i ith value represents the observed value, the NVI_p^i ith is the predicted value by the soft computing models, and the $\overline{NVI_o}$ averages refer to the mean of observed values and $\overline{NVI_P}$ is the mean of predicted values by the soft computing models.

The value of R is in the range [+1, -1] and expresses a linear (direct or inverse) relationship between parameters. When R approaches + 1, it indicates a direct correlation, and when it approaches - 1, it signifies an inverse correlation between parameters. When there is weak correlation between parameters, R approaches zero; moreover, R is dimensionless. RMSE has a value between 0 and + ∞, reflecting the spread of predictions of each model relative to observed values. The closer it is to zero, the more accurate the model is, and it has dimensions similar to the evaluated parameters (here, dimensionless). Changes in MAE range between 0 and + ∞, indicating the difference between observed and predicted values by the model. The closer it is to zero, the more accurate the model, with dimensions similar to the evaluated parameters (here, dimensionless). SI provides an indication of the relative spread of model predictions compared to the average of the observed values. A lower SI suggests that the model predictions are relatively concentrated and close to the mean of the observed values, indicating better model performance. The reliable range of SI is as same as RMSE and MAE.

Moreover, the assessment of AI model performance in both training and testing phases involves the use of receiver-operating characteristic (ROC) curves. To construct the ROC curve, it is essential to comprehend the concepts of sensitivity and specificity, which are directly employed to evaluate AI model effectiveness. Subsequently, the computation of the area under the curve (AUC) is required⁷⁷. The literature provides comprehensive explanations of ROC curves. Broadly speaking, the diverse AUC ranges are capable of characterizing AI models efficacy as follows: 0.5–0.6 (low precision), 0.6–0.7 (moderate precision), 0.7–0.8 (high accuracy), 0.8–0.9 (remarkably high precision), and 0.9–1 (excellent precision).

The comparison presented in Table 8 clearly indicates that the EPR soft computing model performs the best among the soft computing models during the training phase, achieving an R of 0.9999, RMSE of 0.246, MAE of 0.191, and SI of 0.0064. Following this, in descending order of performance, are the MARS model (R=0.9998, RMSE=0.597, MAE=0.443, and SI=0.0156), M5MT model (R=0.9984, RMSE=0.679, MAE=0.189, and SI=0.0439), and GEP model, which is the least effective with an R of 0.9958, RMSE of 0.766, MAE of 0.721, and SI of 0.0724. In Table 8, a similar comparison among the soft computing models is conducted during the testing phase, and once again, the EPR model exhibits the best performance with an R of 0.9999, RMSE of 0.2105, MAE of 0.1712, and SI of 0.0040. Subsequently, in decreasing order of performance, are the MARS model (R=0.9966, RMSE=0.408, MAE=0.342, SI=0.0530), M5MT model (R=0.9956, RMSE=0.988, MAE=0.996, SI=0.0580), and the GEP model, which performs the least effectively with an R of 0.9920, RMSE of 0.491, MAE of 0.673, and SI of 0.0781.

Another way to investigate the correlation between the predicted values by soft computing models and the observed values is through scatter plots. In fact, Fig. 8a and b illustrate the qualitative performance of the soft computing models in both training and testing phases. Considering Fig. 8a, all data points during the training phase are concentrated within the ± 10% acceptable error range. For NVI in the range (30, 10), the predicted values of the M5MT model are mostly higher than the observed values, and the predicted values of the GEP model are lower. The GEP model predicts NVI values less than the observed values for the range (70, 60), higher for the range (80, 70), less for the range (90, 80), and higher for the range (110, 90). The M5MT model generally predicts higher NVI values in the range (100, 60). Throughout the plot, the predictions of the EPR and MARS models are nearly equal to the observed values. Similarly, in Fig. 8b, as in the previous figure, all data points during the testing phase fall within the acceptable error range. The predicted values of the M5MT model are higher than the observed values throughout the plot. The GEP model overestimates the predicted values in the range (80, 70) but underestimates them in the range (90, 80). In contrast to the training phase where the predictions by the MARS model align almost perfectly on the y = x line, in the testing phase, it predicts higher NVI values for the range (70, 50). The predicted values by the EPR model, like the training phase, are in line with

AI models	Training phase			
	R	RMSE	MAE	SI
GEP	0.9958	2.766	1.721	0.0724
EPR	0.9999	0.2460	0.191	0.0064
MARS	0.9998	0.5970	0.443	0.0156
M5MT	0.9984	1.6790	1.189	0.0439
AI Models	Testing phase			
	R	RMSE	MAE	SI
GEP	0.9920	3.491	2.673	0.0781
EPR	0.9999	0.210	0.171	0.0040
MARS	0.9966	2.408	1.341	0.0530
M5MT	0.9956	2.988	1.996	0.0580

Table 8. Evaluation of Statistical analysis of AI models performance in the NVI prediction.

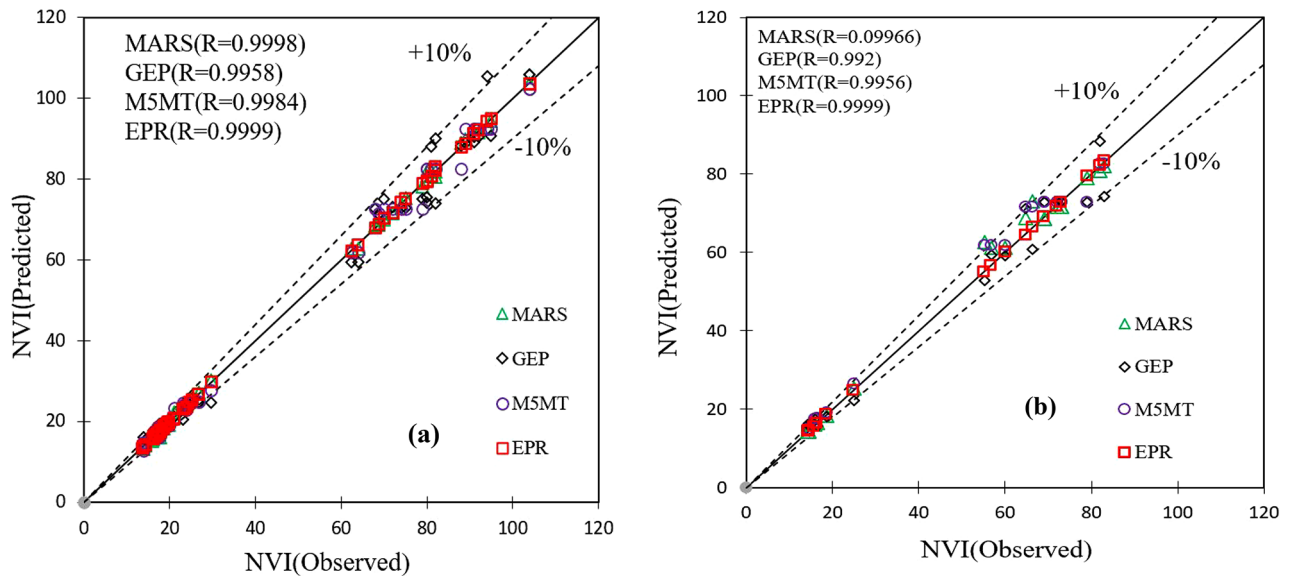


Fig. 8. Performance of AI models for both (a) training and (b) testing stages

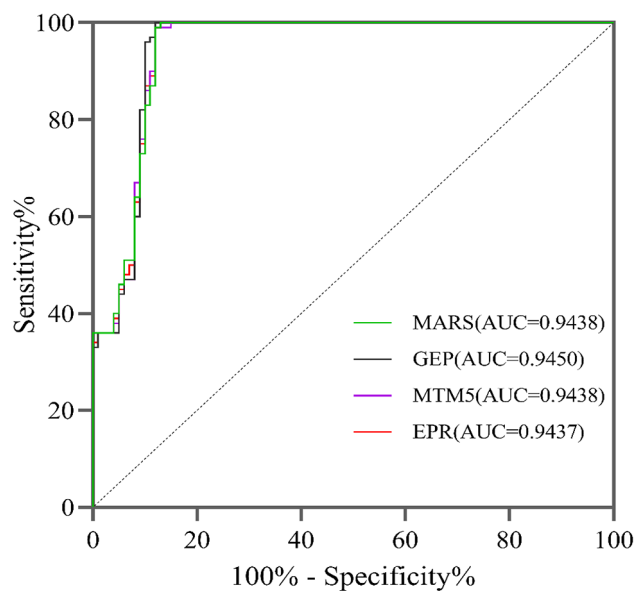


Fig. 9. ROC curve of AI models for predicting NVI.

the observed values. Regarding the complexity of the model structures, the mathematical expressions of the GEP model (Eq. 9) and the EPR model (Model#6 in Table 7) are more intricate compared to the MARS and M5MT models. The EPR model, due to the presence of six algebraic expressions, a natural logarithm internal function, and the utilization of eight input parameters, has a higher degree of mathematical complexity compared to the GEP model, which consists of three internal functions (Min, Max, Exp), four algebraic expressions, and the use of seven input parameters. Additionally, the Eq. (6) provided by the MARS model includes 18 sets of second-degree polynomials with 7 input parameters along with 10 forward and backward steps, representing more complex expressions than those generated by the M5MT (Eq. 7). As mentioned earlier, the EPR model demonstrates favorable performance compared to other soft computing models. The MARS model is well-positioned as an alternative to the EPR model among the three other models.

To draw the ROC curve and calculate the AUC, GraphPad Prism 8 software was utilized. According to Fig. 9, the GEP model (AUC=94.50%) performed excellently, followed by the MARS and M5MT models (AUC=94.38%), and finally, the EPR model (AUC=94.37%).

Comparisons of present research with relevant literature

These artificial intelligence models, based on regression and providing relationships, have found extensive application in the realm of groundwater pollution control. The mentioned AI methods can automatically identify input variables and select those with higher importance for modeling. In this study, the EPR model, using a multi-objective genetic algorithm in its structure, optimizes the number of algebraic expressions and input variables. Additionally, models such as MT and MARS, by incorporating characteristics of data mining methods like adaptive regression, the concept of standard deviation error in the structure of tree and regression models, and the inclusion of forward stages for initial regression model building and backward stages for removing algebraic expressions leading to overfitting, reduce the computational load in estimating the DRASTIC index. Moreover, the GEP model, similar to the EPR model, can easily represent the complexities in modeling the DRASTIC index (parameter uncertainties and groundwater system complexity) by simultaneously using mathematical operators. The mentioned features of the AI models in this study contribute to computational efficiency reduction. On the other hand, some AI models like ANN, ANFIS, and SVM use all input variables. The results of the AI models, based on their formulas, can be easily interpreted, and with only the values of DRASTIC input parameters (considering land use effects and nitrate pollution), the status of groundwater contamination can be examined.

In Fijani et al.²⁴ research, SCMAI was used based on Sugeno Fuzzy Logic (SFL), Mamdani Fuzzy Logic (MLF), ANN, and ANFIS models in order to assess groundwater vulnerability via DRASTIC index while considering Nitrate–N concentration for Maragheh–Bonab plain aquifer, Iran. In terms of accuracy levels, all AI models ($R=0.61, 0.82, 0.76, 0.84, 0.98$ for MFL, SFL, ANN, SCMAI, respectively) used in their study have stood at the lower precision stages compared with the present research (see Table 8). In addition, Fijani et al.²⁴ did not consider LU effects on the DRASTIC index whereas we took it into consideration for evaluation of pollution assessment of Kerman–Baghin Plain.

A notable limitation of Fijani et al.'s²⁴ study was their omission of land use (LU) effects on the DRASTIC index. This aspect is critical in understanding groundwater vulnerability, as land use changes significantly influence the dynamics of groundwater contamination. In contrast, the current research specifically considered land use effects, which allows for a more nuanced evaluation of pollution impacts in the Kerman–Baghin Plain. By integrating land use data into the analysis, the present study provides a more accurate and holistic assessment of groundwater vulnerability, reflecting the real-world complexities of how human activities interact with natural systems. Overall, the incorporation of land use considerations and the employment of more precise AI modeling techniques not only enhance the accuracy of groundwater vulnerability assessments but also address gaps identified in previous studies, such as that of Fijani et al.²⁴. This approach underscores the importance of continuously improving methodologies to ensure effective management of groundwater resources and to protect them from contamination.

In the study conducted by Barzegar et al.¹ in the Tabriz Plain aquifer, the ANN model (constructed by three hidden layers and 12 neurons with $R=0.847$) and ANFIS models (constructed by Mamdani with a correlation of 0.784 and Sugeno with a correlation of 0.784) demonstrated high accuracy in evaluating the DRASTIC index. However, the AI models ANN and ANFIS exhibited lower accuracy compared to the results of the present study (e.g., $R=0.9920, 0.9999, 0.9966$, and 0.9956 for GEP, EPR, MARS, and M5MT, respectively). Additionally, in terms of structure, the artificial intelligence models such as ANN and ANFIS act like black boxes, making the interpretation of DRASTIC data somewhat more challenging compared to the simpler AI models employed in the present research, such as M5MT and MARS, which have simpler structures represented by Eqs. (6 and 7). The developed SCMAI model, due to its simultaneous use of the inherent characteristics of ANNs and fuzzy models (Mamdani and Sugeno), achieved a correlation coefficient of 99.0%. This performance is comparable to the results of the present study, which used GEP ($R=0.9920$), EPR ($R=0.9999$), MARS ($R=0.9966$), and M5MT ($R=0.9956$). It is worth mentioning that the SCMAI model, due to its use of multiple artificial intelligence methods, has greater complexity compared to EPR and GEP models. Therefore, it can be concluded that these AI models, considering land use effects and nitrate pollutant concentrations, provides a cost-effective relationship.

Baghapour et al.³⁰ utilized an ANN model to enhance the DRASTIC method in an unrestricted aquifer in the Shiraz Plain. The outcomes indicated that the ANN was more influential in improving the DRASTIC framework when compared to the CD and NV indices. This is deduced through three plausible ways: (1) ANN, as a black-box framework, is capable of more persuasive explaining the non-linear behavior of aquifers, which are complex systems. (2) Unlike NV and CD methods, the supervised learning approach adopted by ANN has modified vulnerability, making it more suitable for training. (3) The ANN model optimizes the model's fitness using the LM optimization methodology in order to obtain a persuasive DRASTIC framework. The ANN model showed R of 0.890, which compared to the models used in the present study, exhibited a slightly lower correlation coefficient.

Barzegar et al.³¹ employed four separate machine learning models (ELM, MARS, SVR, and M5 Tree) for mapping the risk of groundwater pollution in the Marand Plain, northwest Iran. The SVR model ($R=0.8673$) demonstrated the best performance, followed by ELM ($R=0.8556$), M5 Tree ($R=0.8452$), and MARS ($R=0.8045$). Both ELM and SVR models exhibited the highest R values; thus, they outperformed MARS and M5 Tree models in assessing groundwater pollution risk. However, according to the present study, all GEP ($R=0.992$), EPR ($R=0.999$), MARS ($R=0.996$), and M5MT ($R=0.9956$) models showed better performance compared to the separate machine learning models (ELM, MARS, SVR, and M5 Tree). Additionally, the ELM model, due to its black-box nature, posed challenges in interpreting DRASTIC data compared to the simpler models MARS and M5MT. The SVR model, using the Gaussian expression as the kernel function, exhibited higher complexity than the models used in this study.

Since the constrained target-oriented genetic algorithm plays an effective role in determining the structural form of GEP ($R=0.992$) and EPR ($R=0.999$) models, it leads to the creation of complex equations (Eq. 9) and the sixth model in Table 7. However, the performance level of these models is higher compared to the

study conducted by Bordbar et al.³², who used the Grey Wolf Optimizer (GWO) algorithm to determine the weights of the GALDIT method in the coastal aquifer of the Gorgan River in northern Iran ($R=0.64$). It is worth mentioning that they considered TDS pollution in their study.

Bordbar et al.⁴², in another study, employed frequency ratio (FR) methods and a genetic algorithm to enhance the rate and weight of the GALDIT model in the coastal aquifer of the Gorgan River. The correlation (R) between the combined models GALDIT-FR and GALDIT-GA was obtained as 0.69 and 0.61, respectively. However, this correlation increased to 0.76 after combining the adjusted rates using the statistical FR method and the optimized weights of the genetic algorithm. In any case, the correlation established between the GALDIT model and the combined models is lower than the GEP ($R=0.992$), EPR ($R=0.999$), MARS ($R=0.996$), and M5MT ($R=0.995$) models. Barzegar et al.⁷¹ used two resampling methods [e.g., Bootstrap Aggregating (BA) and Disjoint Aggregating (DA)] in order to improve performance of ML models (i.e., XGBoost [$R=0.613$], Light Gradient Boosting Machine [LGBM], AdaBoost, Categorical Boosting [CatBoost], and RF) for prediction of the GALDIT groundwater vulnerability in the Shabestar Plain aquifer, Iran. From their study, although ML models used were fast-to-learn with lower number of setting parameters, the efficacy of the all coupled ML models was not efficient in the evaluation of groundwater vulnerability (BA-XGBoost [$R=0.6598$], BA-LGBM [0.525], BA-AdaBoost [$R=0.591$], BA-RF [$R=0.591$], BA-CatBoost [$R=0.584$], DA-XGBoost [$R=0.562$], DA-LGBM [0.571], DA-AdaBoost [$R=0.587$], DA-RF [$R=0.616$], DA-CatBoost [$R=0.571$]) than those yielded in the present study. Furthermore, Norouzi et al.⁴⁴ utilized the Random Forest (RF) method to optimize the DRASTIC framework in the unconfined aquifer of the Miyandouab Plain, northwest Iran. The raster layer data of 7 parameters and the output of the DRASTIC model were used as input and the target variable for the RF model, respectively. According to the ROC-AUC performance criterion, the RF method has a high AUC ($AUC=0.977$), and the weight of the DRASTIC model has been optimized with greater reliability. In the present study, the GEP model ($AUC=0.9450$) performed first, followed by the MARS and M5MT ($AUC=0.9438$), and finally, the EPR model ($AUC=0.9437$), all demonstrating excellent performance compared to the BDF-RF method with less reliable optimization.

Elzain et al.⁴⁵ used successfully RFR ($R=0.96$ and $AUC=0.97$) so as to ascertain the most precise performance during the assessment of adjusted vulnerability index (AVI) related to DRASTIC index, LU parameter, and nitrate values for Miryang City (located South Korea) when compared with SVR ($R=0.866$ and $AUC=0.89$) and RBNN ($R=0.781$ and $AUC=0.76$). From their study, the performance of RFR model was the best accurate as well as the results given by this research: EPR ($R=0.9999$ and $AUC=0.9437$), M5MT ($R=0.9956$ and $AUC=0.9438$), GEP ($R=0.9920$ and $AUC=0.9450$), and MARS ($R=0.9966$ and $AUC=0.9438$). However, the general structure of RFR with 100 trees and SVM with four polynomial kernel functions are more complex than M5MT and MARS developed by two rules and 18 BF, respectively.

Later, Elzain et al.⁴⁹ concluded that the performance of combination of KNN, BA, and ERT models in order to evaluate DRASTIC index by considering Nitrate pollution and LU parameters, demonstrated the most accurate predictions ($R=0.974$) as well as the present investigation. In this regards, ERT model developed by Elzain et al.⁴⁹ research, has potential disadvantages of pollution vulnerability predictions when compared to M5MT and MARS given by the present investigation: (i) Interpretability: Extra Trees (ERT) tend to be less interpretable compared to M5MT and MARS models. The randomness in feature selection and splitting decisions may make it harder to understand the specific relationships between features and the target variable. (ii) Overfitting: while ERT are designed to be more robust to overfitting than conventional RFs, they can still be prone to overfitting, especially when dealing with noisy datasets or datasets with a small number of samples. (iii) Less control over splitting: the randomness in feature selection and threshold determination in ERT model means less control over how splits are made. In some cases, this lack of control may result in suboptimal splits, particularly when dealing with certain types of pollution data patterns. Recent investigations, carried out by Karimzadeh-Motlagh et al.⁴⁸, concluded that RF ($AUC=0.987$) had the best performance in the approximation of DRASTIC index than GLM ($AUC=0.788$) and SVM (0.78) for Najafabad Plain, Iran. Additionally, RF model was classified DRASTIC index as well as AI models in the present investigation (e.g., $AUC=0.9450$, 0.9438 , and 0.9437 for GEP, MARS, and EPR, respectively).

According to the above-mentioned comparisons, the advantages of MT, EPR, MARS, and GEP in assessing the DRASTIC index for groundwater quality can be particularly pronounced when compared to boosting learning models such as AdaBoost, XGBoost, LightGBM, and CatBoost. One key advantage of the aforementioned models lies in their structural complexity. Model Trees, for instance, combine the interpretability of decision trees with regression capabilities, allowing for a clear visualization of how input variables influence the output. This is crucial when assessing factors (ecological and geochemical factors) that affect groundwater vulnerability, as stakeholders can easily comprehend the relationships and impacts of various parameters like land use and nitrate concentrations. Similarly, MARS employs piecewise linear splines, which allow for flexibility in modeling non-linear relationships while maintaining interpretability. This stands in contrast to boosting models, which often function as black boxes, making it challenging to discern how specific features contribute to predictions. EPR offers a unique approach by evolving polynomial expressions that can capture complex relationships among variables while still being interpretable. This is especially beneficial for groundwater studies where multiple interacting factors must be considered. GEP also presents a clear advantage through its genetic programming framework, allowing the discovery of mathematical models that succinctly represent relationships in the data, facilitating easier understanding and communication of results. When it comes to handling uncertainty in parameter settings, MT, EPR, MARS, and GEP provide greater robustness and flexibility. The present AI models are generally less sensitive to hyperparameter tuning compared to boosting algorithms, which often require meticulous adjustment to achieve optimal performance. For example, boosting models can be quite susceptible to overfitting if parameters like learning rate and tree depth are not carefully calibrated. In contrast, the other models tend to perform well across a wider range of parameter settings, reducing the burden on researchers to

fine-tune models extensively. Furthermore, the simplicity in the application of these models allows researchers and practitioners to focus more on understanding the implications of their findings rather than getting lost in complex tuning processes. This is particularly advantageous in the context of groundwater management, where clear communication of results to non-expert stakeholders is crucial for effective decision-making. Additionally, the ability of these models to incorporate domain knowledge into their structures enhances their applicability in environmental studies. For instance, the polynomial expressions generated by EPR can directly reflect scientific understanding of hydrological processes, while MARS can effectively model interactions between multiple environmental factors. This incorporation of expert knowledge helps in developing models that are not only accurate but also relevant to real-world scenarios. In summary, the structural advantages and robustness of MT, EPR, MARS, and GEP models present compelling benefits over boosting learning models. Their interpretability, flexibility in handling uncertainties, and ability to incorporate domain knowledge make them particularly suitable for assessing the DRASTIC index in groundwater quality studies. This ultimately leads to more reliable and understandable assessments, facilitating better groundwater management practices.

Conclusion

This research aimed to investigate susceptibility state of groundwater resources by the DRASTIC framework. Subsequently, two indices, CD and NVI, based on the same DRASTIC model, were employed to evaluate vulnerability. Due to uncertainties in hydrogeological parameters and complexities of the groundwater system, AI methods such as GEP, EPR, MARS, and MTM5 were applied for the susceptibility evaluation. In this way, the ranges of DRASTIC index values were classified into three categories: no vulnerability (35.7%), very low vulnerability (60.2%), and low vulnerability (4.1%). The CD index indicated that the lowest classification of vulnerability was associated with very low vulnerability (17.4%), while the highest was related to low vulnerability (82.6%). In contrast to the other two indices, the NVI index showed that 75.52% of the study area falls into the category of very low vulnerability, and 24.48% falls into the category of low vulnerability. All soft computing methods used in both the testing and training phases have high correlation coefficients. However, the EPR model ranked first, followed by the MARS, MTM5, and GEP models. The order of these AI models in terms of precision levels and expected error is consistent. According to the ROC-AUC method, all models demonstrated excellent performance. The EPR model has a higher level of complexity when compared to other AI models, so the MARS model can be a suitable alternative. Vulnerability classification through the MARS model reveals that approximately 73.06% of the studied plain is in the very low vulnerability zone, while 26.94% is in the low vulnerability zone, covering an area of about 1478.3 square kilometers and 545.1 square kilometers, respectively, in the Kerman–Baghin aquifer. Furthermore, the outcomes of the present investigation were absolutely comparable with those reported in reliable literature. This means that the AI models used in this study had superiority in the prediction of DRASTIC index, influenced by nitrate pollution and land use factors, over similar investigations.

The practical implications for groundwater management using the DRASTIC index for agricultural purposes are significant and multifaceted. First, by identifying areas of varying vulnerability to contamination, farmers and agricultural planners can implement targeted practices to protect groundwater quality. For instance, regions classified with high vulnerability can benefit from best management practices (BMPs) that minimize the use of fertilizers and pesticides, thus reducing nitrate leaching into groundwater. Additionally, the DRASTIC index provides a framework for assessing the potential impacts of different land use practices on groundwater quality. Agricultural managers can use this information to make informed decisions about crop selection, irrigation methods, and land management strategies that align with the vulnerability classifications. For example, crops that require less water or are less nutrient-intensive could be favored in more vulnerable areas to mitigate contamination risks. The research also emphasizes the need for regular monitoring and assessment of groundwater quality in agricultural regions. Implementing a monitoring program that tracks changes in the DRASTIC index over time can help identify emerging risks and facilitate timely interventions. This ongoing assessment can guide adaptive management strategies that respond to changing environmental conditions and agricultural practices. Moreover, collaboration with local communities and stakeholders is crucial. Educating farmers about the implications of groundwater vulnerability and encouraging sustainable practices can foster a collective effort to protect these vital resources. Initiatives that promote community engagement in groundwater management can enhance compliance with protective measures and contribute to better overall water quality. Finally, the findings can support policy development aimed at regulating agricultural practices in sensitive areas. Policymakers can use the DRASTIC index to prioritize regions for conservation efforts, funding, and technical assistance, ensuring that resources are allocated efficiently to safeguard groundwater quality. In summary, leveraging the DRASTIC index for agricultural purposes can lead to improved groundwater management through targeted practices, informed decision-making, regular monitoring, community engagement, and effective policy development, ultimately promoting sustainable agricultural practices and protecting vital water resources.

The research on groundwater resource susceptibility using the DRASTIC framework reveals several current challenges that warrant attention. One significant challenge is the inherent uncertainty in hydrogeological parameters, which complicates the accurate assessment of groundwater vulnerability. These uncertainties arise from variations in data quality, availability, and the complexity of the groundwater systems themselves. Additionally, the interaction of multiple environmental factors, such as land use changes and climate impacts, adds another layer of complexity, making it difficult to develop universally applicable models. Another challenge is the interpretability of AI models. While the EPR model demonstrated high predictive accuracy, its complexity can make it difficult for stakeholders and policymakers to understand the underlying relationships between input parameters and the DRASTIC index. This lack of clarity can hinder effective decision-making and limit the practical application of the results in groundwater management.

In light of these challenges, several future recommendations can enhance the robustness of groundwater vulnerability assessments. Firstly, there is a need for the integration of diverse data sources, including remote sensing and advanced geospatial analytics, to provide a more comprehensive understanding of hydrogeological conditions. This integration could improve the accuracy and reliability of the models used in vulnerability assessments. Furthermore, future research should focus on refining AI models to balance complexity and interpretability. Exploring ensemble approaches that combine the strengths of various models may yield more robust predictions while maintaining clarity in the results. Conducting longitudinal studies to monitor changes in groundwater quality over time will also provide valuable insights into how dynamic environmental factors influence vulnerability. Strengthening community engagement and raising public awareness about groundwater conservation is crucial for effective resource management. Educating local stakeholders about the implications of land use and pollution on groundwater quality can foster greater involvement in protection efforts. Lastly, addressing the challenges of data availability and quality remains vital. Ensuring access to reliable hydrogeological data will enhance the accuracy of assessments and facilitate ongoing research in the field. By focusing on these future directions and overcoming current challenges, the research can contribute significantly to the sustainable management of groundwater resources.

Data availability

The data are not publicly available due to restrictions such their containing information that could compromise the privacy of research participants. Contact the corresponding author to request data.

Received: 17 July 2024; Accepted: 4 November 2024

Published online: 02 December 2024

References

- Barzegar, R., Moghaddam, A. A. & Baghban, H. A. Supervised committee machine artificial intelligent for improving DRASTIC method to assess groundwater contamination risk: A case study from Tabriz plain aquifer, Iran. *Stochastic Environ. Res. Risk Assess.* **30**, 883–899 (2015).
- Demiroglu, M. & Dowd, J. The Utility of vulnerability maps and GIS in groundwater management: A case study. *Turk. J. Earth Sci.* **23**, 80–90 (2014).
- Jang, W. S., Engel, B., Harbor, J. & Theller, L. Aquifer vulnerability assessment for sustainable groundwater management using DRASTIC. *Water* **9**(10), 792 (2017).
- Li, P., Karunanidhi, D., Subramani, T. & Srinivasamoorthy, K. Sources and Consequences of Groundwater contamination. *Arch. Environ. Contam. Toxicol.* **80**, 1–10 (2021).
- Neshat, A., Pradhan, B., Pirasteh, S. & Shafri, H. Z. M. Estimating groundwater vulnerability to pollution using a modified DRASTIC model in the Kerman agricultural area, Iran. *Environ. Earth Sci.* **71**(7), 3119–3131 (2014).
- Maghrebi, M., Noori, R., Partani, S., Araghi, A., Barati, R., Farnoush, H. & Torabi-Haghighi, A. *Iran's groundwater hydrochemistry*. *Earth Sp. Sci.* e2021EA001793 (2021).
- Noori, R. et al. Iran is draining its aquifers dry. *Nature* **594**, 476 (2021).
- Noori, R., Maghrebi, M., Mirchi, A., Tang, Q., Bhattarai, R., Sadegh, M., Noury, M., Torabi-Haghighi, A., Kløve, B. & Madani, K. *Proceedings of the National Academy of Sciences* **118**(25), e2024221118 (2021).
- Noori, R. et al. Decline in Iran's groundwater recharge. *Nat. Commun.* **14**(1), 6674 (2023).
- Noori, R., Ghahremanzadeh, H., Kløve, B., Adamowski, J. F. & Baghvand, A. Modified-DRASTIC, modified-SINTACS and SI methods for groundwater vulnerability assessment in the southern Tehran aquifer. *J. Environ. Sci. Health Part A* **54**(1), 89–100 (2019).
- Nobre, R. C. M., Rotunno Filho, O. C., Mansur, W. J., Nobre, M. M. M. & Cosenza, C. A. N. Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool. *J. Contam. Hydrol.* **94**, 277–292 (2007).
- Kazakis, N. & Voudouris, K. S. Groundwater vulnerability and pollution risk assessment of porous aquifers to nitrate: Modifying the drastic method using quantitative parameters. *J. Hydrol.* **525**, 13–25 (2015).
- Rezaei, F., Safavi, H. R. & Ahmadi, A. Groundwater vulnerability assessment using fuzzy logic: A case study in the Zayandehrood aquifers, Iran. *Environ. Manag.* **51**, 267–277 (2013).
- Abunada, Z., Kishawi, Y., Alslaibi, T. M., Kuhail, N. & Mittelstet, A. The application of SWAT-GIS tool to improve the recharge factor in the DRASTIC framework: Case study. *J. Hydrol.* **592**, 125613 (2021).
- Mohammadi, K., Niknam, R. & Majd, V. J. Aquifer vulnerability assessment using GIS and fuzzy system: A case study in Tehran-Karaj aquifer Iran. *Environ. Geol.* **58**, 437–446 (2009).
- Panagopoulos, G. P., Antonakos, A. K. & Lambrakis, N. J. Optimization of the DRASTIC method for groundwater vulnerability assessment via the use of simple statistical method and GIS. *Hydrogeol. J.* **14**(6), 894–911 (2006).
- Vrba, J. & Zopporozec, A. *Guidebook on Mapping Groundwater Vulnerability International Contributions to Hydrogeology* (Verlag Heinz Heise GmbH and Co. KG, 1994).
- Koycegiz, C. & Buyukyildiz, M. Calibration of SWAT and two data-driven models for a data-scarce mountainous headwater in Semi-Arid Konya Closed basin. *Water* **11**(1), 147 (2019).
- Neshat, A. & Pradhan, B. Evaluation of groundwater vulnerability to pollution using DRASTIC framework and GIS. *Arab J Geosci.* **10**, 501 (2017).
- Yu, C., Yao, Y., Hayes, G., Zhang, B. & Zheng, C. Quantitative assessment of groundwater vulnerability using index system and transport simulation, Huangshuihe catchment. *China. Sci. Total Environ.* **408**, 6108–6116 (2010).
- Umar, R., Ahmed, I. & Alam, F. Mapping groundwater vulnerable zones using modified DRASTIC approach of an alluvial aquifer in parts of central Ganga plain, western Uttar Pradesh. *J. Geol. Soc. Ind.* **73**, 193–201 (2009).
- National Research Council. *Ground Water Vulnerability Assessment: Predicting Relative Contamination Potential Under Conditions of Uncertainty* (The National Academies Press, 1993). <https://doi.org/10.17226/2050>.
- Knox, R. C., Sabatini, D. A. & Canter, L. W. *Subsurface Transport and Fate Processes* (Lewis Publishers, 1993).
- Fijani, E., Nadiri, A. A., Moghaddam, A. A., Tsai, F. T. & Dixon, B. Optimization of DRASTIC Method by supervised committee machine artificial intelligence to assess groundwater vulnerability for Maragheh-Bonab plain aquifer, Iran. *J. Hydrol.* **503**, 89–100 (2013).
- Jafarzadeh, F., Moghaddam, A. A., Razzagh, S., Barzegar, R., Cloutier, V. & Rosa, E. A meta-ensemble machine learning strategy to assess groundwater holistic vulnerability in coastal aquifers. *Groundw. Sustain. Dev.* 101296 (2024).
- Neshat, A., Pradhan, B. & Dadras, M. Groundwater vulnerability assessment using an improved DRASTIC method in GIS. *Resour. Conserv. Recycl.* **86**, 74–86 (2014).

27. Raisa, S. S., Sarkar, S. K. & Sadiq, M. A. Advancing groundwater vulnerability assessment in Bangladesh: A comprehensive machine learning approach. *Groundw. Sustain. Dev.* **25**, 101128 (2024).
28. Samani, S. Unraveling aquifer dynamics: Time series evaluation for informed groundwater management. *Groundw. Sustain. Dev.* **25**, 101174 (2024).
29. Samani, S., Vadiati, M., Kisi, O., Ghasemi, L. & Farajzadeh, R. Qanat discharge prediction using a comparative analysis of machine learning methods. *Earth Sci. Inf.* 1–22 (2024).
30. Baghagpour, M. A. et al. Optimization of DRASTIC method by artificial neural network, nitrate vulnerability index, and composite DRASTIC models to assess groundwater vulnerability for unconfined aquifer of Shiraz Plain, Iran. *J. Environ. Health Sci.* **14**, 1–16 (2016).
31. Barzegar, R., Moghaddam, A. A., Deo, R., Fijani, E. & Tziritis, E. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. *Sci. Total Environ.* **621**, 697–712 (2018).
32. Bordbar, M., Neshat, A., Javadi, S., Pradhan, B. & Aghamohammadi, H. Meta-heuristic algorithms in optimizing GALDIT framework: A comparative study for coastal aquifer vulnerability assessment. *J. Hydrol.* **585**, 124768 (2020).
33. Jesiya, N. P. & Gopinath, G. A customized FuzzyAHP–GIS based DRASTIC–L model for intrinsic groundwater vulnerability assessment of urban and peri urban phreatic aquifer clusters. *Groundw. Sustain. Dev.* **8**, 654–666 (2019).
34. Neshat, A., Pradhan, B. & Shafri, H. Z. M. An integrated GIS based statistical model to compute groundwater vulnerability index for decision maker in agricultural area. *J. Indian Soc. Remote Sens.* **42**(4), 777–788 (2014).
35. Voutchkova, D. D., Schullehner, J., Rasmussen, P. & Hansen, B. A high-resolution nitrate vulnerability assessment of sandy aquifers (DRASTIC-N). *J. Environ. Manag.* **277**, 111330 (2021).
36. Neshat, A. & Pradhan, B. Risk assessment of groundwater pollution with a new methodological framework: Application of Dempster–Shafer theory and GIS. *Nat. Hazards* **78**(3), 1565–1585 (2015).
37. Neshat, A. & Pradhan, B. An integrated DRASTIC model using frequency ratio and two new hybrid methods for groundwater vulnerability assessment. *Nat. Hazards* **76**, 543–563 (2015).
38. Langrudi, M. A., Siuki, A. K., Javadi, S. & Hashemi, S. R. Evaluation of vulnerability of aquifers by improved fuzzy drastic method: case study: Aastane Kochesfahan plain in Iran. *Ain Shams Eng. J.* **7**, 11–20 (2016).
39. Nadiri, A. A., Sedghi, Z., Khatibi, R. & Sadeghfam, S. Mapping specific vulnerability of multiple confined and unconfined aquifers by using artificial intelligence to learn from multiple DRASTIC frameworks. *J. Environ. Manag.* **227**, 415–428 (2018).
40. Hu, X., Ma, C., Qi, H. & Guo, X. Groundwater vulnerability assessment using the GALDIT model and the improved DRASTIC model: a case in Weibei Plain, China. *Environ Sci Pollut Res* **25**, 32524–32539 (2018).
41. Torkashvand, M., Neshat, A., Javadi, S. & Yousefi, H. DRASTIC framework improvement using stepwise weight assessment ratio analysis (SWARA) and combination of genetic algorithm and entropy. *Environ. Sci. Pollut. Res* **28**, 46704–46724 (2021).
42. Bordbar, M., Neshat, A., Javadi, S. & Shahdany, S. M. H. A Hybrid approach Based on statistical method and meta-heuristic optimization algorithm for coastal aquifer vulnerability assessment. *Environ. Model. Assess.* **26**(3), 325–338 (2021).
43. Jahromi, M. N., Gomeh, Z., Busico, G. et al. Developing a SINTACS-based method to map groundwater multi-pollutant vulnerability using evolutionary algorithms. *Environ Sci Pollut Res* **28**, 7854–7869 (2021).
44. Norouzi, H., Moghaddam, A. A., Celico, F. & Shiri, J. Assessment of groundwater vulnerability using genetic algorithm and random forest methods (case study: Miandoab plain, NW of Iran). *Environ. Sci. Pollut. Res.* **28**(29), 39598–39613 (2021).
45. Elzain, H. E. et al. Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination. *Ecotoxicol. Environ. Saf.* **229**, 113061 (2022).
46. Gharekhani, M., Nadiri, A. A., Khatibi, R., Sadeghfam, S. & Asghar Moghaddam, A. A study of uncertainties in groundwater vulnerability modelling using Bayesian model averaging (BMA). *J. of Environmental Management* **303**, 114168 (2022).
47. Subbarayan, S., Thiagarajan, S., Karuppanan, S. & Panneerselvam, B. Enhancing groundwater vulnerability assessment: Comparative study of three machine learning models and five classification schemes for Cuddalore district. *Environ Research*, 117769 (2023).
48. Karimzadeh-Motlagh, Z., Derakhshani, R. & Sayadi, M. H. Groundwater vulnerability assessment in central Iran: Integration of GIS-based DRASTIC model and a machine learning approach. *Groundw. Sustain. Dev.* **23**, 101037 (2023).
49. Elzain, H. E. et al. Novel machine learning algorithms to predict the groundwater vulnerability index to nitrate pollution at two levels of modeling. *Chemosphere* **314**, 137671 (2023).
50. Al-Adamat, R. A. N., Foster, I. D. L. & Baban, S. M. J. Groundwater vulnerability and risk mapping for the basaltic aquifer of the Azraq basin of Jordan using GIS, remote sensing and DRASTIC. *Appl. Geogr.* **23**, 303–324 (2003).
51. Piscopo, G. *Groundwater Vulnerability Map, Explanatory Notes, Castlereagh Catchment*, NSW. Department of Land and Water Conservation, Australia, Found at: http://www.dlwc.nsw.gov.au/care/water/groundwater/reports/pdfs/castlereagh_map_notes.pdf (2001).
52. Awawdeh, M. & Jaradat, R. Evaluation of Aquifers vulnerability to contamination in the Yarmouk River Basin, Jordan, based on DRASTIC method. *Arab. J. Geosci.* **3**(3), 273–282 (2010).
53. Liggett, J. & Talwar, S. Groundwater Vulnerability assessments and integrated water resource management. *Streamline Watershed Manag. Bull.* **13**(1), 18–29 (2009).
54. Oh, H. J., Kim, Y. S., Choi, J. K., Park, E. & Lee, S. GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. *J. Hydrol.* **399**, 158–172 (2011).
55. Wang, J., He, J. & Chen, H. Assessment of groundwater contamination risk using hazard quantification, a modified DRASTIC model and groundwater value, Beijing Plain, China. *Sci. Total Environ.* **432**, 216–226 (2021).
56. Mohamadi, S., Honarmand, M., Ghazanfari, S. & Hassanzadeh, R. Hotspot and accumulated hotspot analysis for assessment of groundwater quality and pollution indices using GIS in the arid region of Iran. *Environ. Sci. Pollut. Res.* **30**, 69955–69976 (2023).
57. Adnan, R. M. et al. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. *J. Hydrol.* **586**, 124371 (2020).
58. Friedman, J. H. Multivariate adaptive regression splines. *Ann. Stat.* **19**, 1–67 (1991).
59. Quinlan, J. R. Learning with continuous classes. In *Proceedings of Australian Joint Conference on Artificial Intelligence* (Hobart, Australia) 343–348 (1992).
60. Wang, Y. & Witten, I. H. Induction of model trees for predicting continuous classes. In *Proceedings of the poster papers of the European Conference on Machine Learning, University of Economics and Statistics* (Prague, 1997).
61. Ferreira, C. Gene expression programming: A New adaptive algorithm for solving problems. *Complex Syst.* **13**(2), 87–129 (2001).
62. Ferreira, C. Gene expression programming : Mathematical modeling by an artificial intelligence. In *Springer eBooks*, Vol. 21 (2006).
63. Nadiri, A. A., Gharekhani, M., Khatibi, R. & Sadeghfam, S. A. A. M. Groundwater vulnerability indices conditioned by supervised intelligence committee machine (SICM). *Sci. Total Environ.* **691**(706), 574 (2017).
64. Ahangar-Asr, A., Faramarzi, A., Javadi, A. A. & Giustolisi, O. Modelling mechanical behaviour of rubber concrete using EPR. *Eng. Comput.* **28**(4), 492–507 (2011).
65. Doglioni, A., Mancarella, D., Simeone, V. & Giustolisi, O. Inferring groundwater system dynamics from hydrological time-series data. *Hydrol. Sci. J.* **55**(4), 593–608 (2010).
66. Giustolisi, O., Berardi, L. & Walski, T. M. Some explicit formulations of Colebrook–White friction factor considering accuracy vs. computational speed. *J. Hydroinformatics* **13**(3), 401–418 (2011).
67. Laucelli, D. & Giustolisi, O. Scour depth modelling by a multi-objective evolutionary paradigm. *Env. Mod. Soft.* **26**(4), 498–509 (2011).

68. Markus, M., Hejazi, M., Bajcsy, P., Giustolisi, O. & Savic, D. A. Prediction of weekly nitrate-N Fluctuations in a small agricultural watershed in illinois. *J. Hydroinformatics* **12**(3), 251–261 (2010).
69. Giustolisi, O. & Savic, D. A. A symbolic data-driven technique based on EPR. *J. Hydroinformatics* **22**(8), 207–222 (2006).
70. Savic, D. A. et al. Modelling sewer failure by evolutionary computing. *Water Manag.* **159**(2), 111–118 (2006).
71. Barzegar, R. et al. Improving GALDIT-based groundwater vulnerability predictive mapping using coupled resampling algorithms and machine learning models. *J. Hydrol.* **598**, 126370 (2021).
72. Kayhomayoon, Z., Babaeian, F., Ghordoyee-Milan, S., Arya-Azar, N. & Berndtsson, R. A. Combination of metaheuristic optimization algorithms and machine learning methods improves the prediction of groundwater level. *Water* **14**, 751 (2022).
73. Kayhomayoon, Z., Ghordoyee Milan, S., Arya-Azar, N. & Kardan-Moghadam, H. A new approach for regional groundwater level simulation: clustering, simulation, and optimization. *Nat. Resour. Res.* **30**, 4165–4185 (2021).
74. Kardan-Moghaddam, H., Ghordoyee-Milan, S., Kayhomayoon, Z., Rahimzadeh-Kivi, Z. & Arya-Azar, N. The prediction of aquifer groundwater level based on spatial clustering approach using machine learning. *Environ. Monit. Assess.* **193**, 173 (2021).
75. Nadiri, A. A., Sedghi, Z., Khatibi, R. & Gharekhani, M. Mapping vulnerability of multiple aquifers using multiple models and fuzzy logic to objectively derive model structures. *Sci. Total Environ.* **593**, 75–90 (2017).
76. Nadiri, A. A., Gharekhani, M., Khatibi, R. & Moghaddam, A. A. Assessment of groundwater vulnerability using supervised commit-tee to combine fuzzy logic models. *Environ. Sci. Pollut. Res.* **24**, 8562–8577 (2017).
77. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**(8), 861–874 (2006).
78. Khosravi, K., Sartaj, M., Karimi, M., Levison, J. & Lotfi, A. A GIS-based groundwater pollution potential using DRASTIC, modified DRASTIC, and bivariate statistical models. *Environ. Sci. Pollut. Res.* **28**, 50525–50541 (2021).

Author contributions

B.A., N.M., M.S. Conceptualization; B.A., N.M., M.S. Providing literature review, performance of machine learning models; B.A., N.M., M.S. Conducting field study and collecting datasets; B.A. M.S., Development of AI models and results analysis; B.A., N.M. Statistical analysis and comparisons; B.A., N.M., M.S. Writing-Original Draft Preparation; N.M., M.S. Writing-Review and Editing; N.M., M.S. Visualization; B.A., N.M., M.S. Supervision. All authors reviewed the manuscript.

Funding

This research received no external funding.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-78812-6>.

Correspondence and requests for materials should be addressed to M.N.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024