Marine Science Center-University of Basrah

Mesopotamian Journal of Marine Sciences
Print ISSN: 2073-6428
E- ISSN: 2708-6097
www.mjms.uobasrah.edu.iq/index.php/mms

Geospatial analysis of groundwater contamination by heavy oil in the Dammam aquifer-Middle of Iraq

Huda M. Al-Mayahi and iD Alaa M. Al-Abadi*
College of Science, University of Basrah, Basrah-Iraq
**Corresponding Author: e-mail alaaatiaa@gmail.com*

Article info.
✓ *Received: 4 November 2019*
✓ *Accepted: 20 November 2019*
✓ *Published: 29December 2019*

Key Words:
Dammam aquifer
Groundwater contamination
Iraq
Machine learning

Abstract - This study aims to apply Artificial Neural Network (ANN) to map the groundwater contamination of the Dammam aquifer by heavy oil in the middle of Iraq. For this purpose, the inventory map of 139 groundwater wells (contaminated and non-contaminated with heavy oil) with the seven important factors playing a role in controlling contamination were used. The factors are distance to faults, faults density, groundwater depth, aquifer saturated thickness and hydraulic conductivity, elevation, and distance from Abu Jir fault. For the performance of ANN model, five statistical measures were used namely, accuracy, sensitivity, specificity, kappa, and the relative operating characteristic curve. Obtained results from applying the model in R statistical package indicated that ANN has a high accuracy (> 90%) in training and testing phases. The probability prediction of ANN model was categorized into five groundwater contamination classes: very low-low, moderate, high-very high. The averages of areas occupied by these zones were 5267 km^2 (65%), 488 km^2 (6%), and 2362 km^2 (29%), for very low-low, moderate, and high-very high, respectively. The contamination map developed in this study could be used to drill successful non-contaminated groundwater wells and avoid loss of many efforts in drilling contaminated wells in the study area.

التحليل المكاني لتلوث المياه الجوفية بالنفط الثقيل في خزان الدمام – وسط العراق

هدى مهدي المياحي وعلاء محسن العبادي
كلية العلوم، جامعة البصرة، البصرة-العراق

المستخلص - تهدف هذه الدراسة إلى تطبيق الشبكة العصبية الاصطناعية (ANN) لرسم خريطة تلوث المياه الجوفية لخزان الدمام بواسطة النفط الثقيل في وسط العراق. لهذا الغرض، تم استخدام خريطة الجرد التي تضم 139 بئراً للمياه الجوفية (الملوثة وغير الملوثة بالنفط الثقيل) مع العوامل السبعة التي يعتقد أنها تلعب دورًا في السيطرة على التلوث. العوامل هي المسافة إلى الفوالق، وكثافة الفوالق، وعمق المياه الجوفية، وسمك طبقة المياه الجوفية المشبعة والتوصلية الهيدروليكية، والارتفاع، والمسافة عن فالق أبو جير. بالنسبة لأداء نموذج ANN، تم استخدام خمسة مقاييس إحصائية وهي:(accuracy, sensitivity, specificity, kappa, and the relative operating characteristic curve). النتائج التي تم الحصول عليها من تطبيق النموذج في الحزمة الإحصائية R تشير إلى أن نموذج ANN لديه دقة عالية (> 90 %) في مراحل التدريب والاختبار. تم تصنيف تنبؤ الاحتمالات لنموذج ANN إلى خمس فئات تلوث للمياه الجوفية: منخفضة للغاية - منخفضة، معتدلة، مرتفعة - عالية جداً. كان معدل المساحة التي تشغلها هذه المناطق 5267 كم^2 (65 %)، 488 كم^2 (6 %)، و 2362 كم^2 (29 %)، منخفضة للغاية ومعتدلة وعالية للغاية، على التوالي. يمكن استخدام خريطة التلوث التي تم تطويرها في هذه الدراسة لحفر آبار المياه الجوفية الناجحة غير الملوثة وتجنب فقد الكثير من الجهود في حفر الآبار الملوثة في منطقة الدراسة.

الكلمات المفتاحية: تلوث المياه الجوفية ،التعلم الالي، طبقة الدمام الجوفية ،العراق.

Introduction

The Dammam aquifer in western and southern deserts of Iraq is an important aquifer due to its huge reserves and relatively good groundwater quality. In the Karbala-Najaf plateau and neighboring areas, the extracted groundwater by means of deep wells is usually failed due to groundwater contamination by heavy oil originate from the deep oil reservoirs.

It is believed that heavy oil have migrated from deep reservoir across the transversal faults distributed in the study area. For this reason, the success of any industrial or agricultural project in the region depends heavily on luck. Unfortunately, there is no study showing the spatial extent of the contamination in the Dammam aquifer, except that provided by Al-Abadi *et al.* (2018) which is limited to the Karbala-Najaf plateau.

With the advent of geographical information systems, remote sensing, and artificial intelligent techniques, a new era of geospatial models are developed that can be used for studying natural hazard and environmental related problems (Chen *et al.*, 2018). Integration of these techniques in one framework make the development of geospatial models more easily task (Zhou *et al.*, 2018; Al-Abadi, 2018; Gayen *et al.*, 2019; Lee *et al.*, 2018).

For the purpose of this study, the backpropagation artificial neural network (BPANN) was used to map the probability of groundwater contamination in the Dammam aquifer in the middle of Iraq through studying the relationship between the groundwater contamination/non-contamination well locations and a set of influential factors that belief to play a role in the Dammam aquifer contamination status.

The Considered Area:

The study area is located in the middle of Iraq (Fig. 1) and covers an area of 8117 km^2. The surface elevation ranges from -3 to 163 m with an average of 63 m. The area is arid. The average monthly temperature, relative humidity, wind speed, and sunshine hours are 24.95 °C, 47%, 2.84 m/s, and 9.29 hrs., respectively. The annual rainfall average is 101.6 mm, which mostly falls in winter. The exposed rocks include Dammam, Euphrates, Fatha, Nafyil, Injana, Dibdibba, and Quaternary deposits (Table 1). From the tectonic point of view, the study is located in the northern part of the Euphrates subzone in the Mesopotamian stable shelf (Buday and Jassim, 1987). Two major groups of faults exist in the study area located in the NE-SW (Rhaimawi-Hilla and Khanaquin-Baguba-Karbala fault) direction and NW-SE direction (Heet-Abu Jir fault) (Fig. 2).

The Dibdibba and Dammam are the main aquifers in the considered area. The Dibdibba aquifer is the shallow, top aquifer that only extends over a limited area of the Karbala-Najaf plateau, while the Dammam Formation is the confined aquifer that extends over the whole of the study area. The Dammam Formation comprises of limestone, dolomites, marls, and shales (Jassim and Goff, 2006). The groundwater depths range from 100 to 200 m and rising to 5 to 20 m in the discharge area. The value of hydraulic gradient ranges from 0.0011 to 0.00005. The total dissolved solid (TDS) of the groundwater ranges from 3 to 5 g/l and is only suitable for irrigation and livestock.

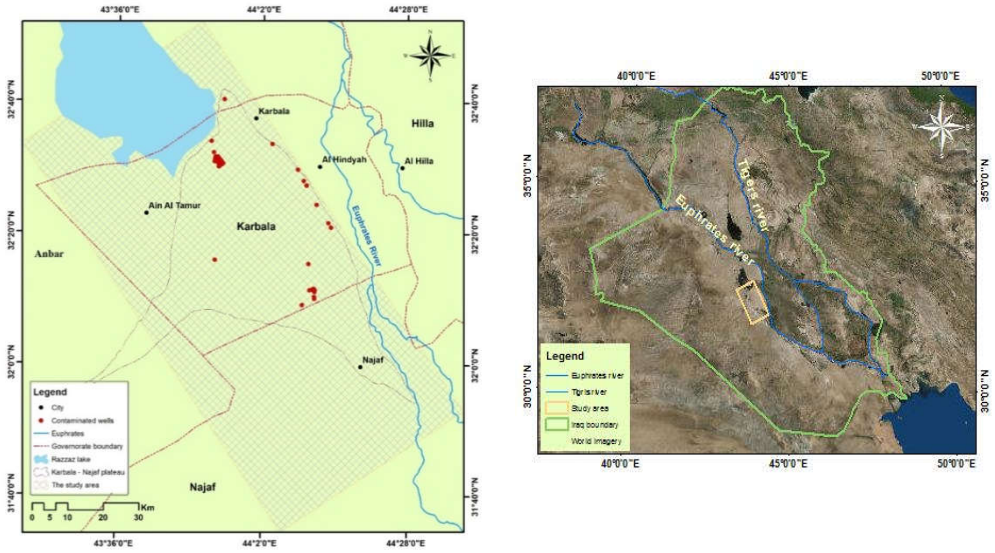

Figure 1. Location of the study area.

Table 1. Exposed formations of the study area (summarized after Jassim and Goff (2006)).

Formation	Age		Environment	Lithological description
Dibdibba	TERTIARY	Upper Miocene Pliocene (AP11)	Fresh water environment (Delta)	Sand, pebbles, claystone, sandstone and silt
Injana		Upper Miocene (AP11)	Lagoon environment	Sandstone, siltstone and claystonc with thin limestone
Fatha		Middle Miocene (AP11)	Deposited in broad basin following a marine transgression	Mudstone, gypsum and silt, interbedded with limestone and marl.
Euphrates		Late lower Miocene (AP11)	Deposited reef and behind the reef	Basal breccia, limestone and marl
Dammam		Middle Late Eocene (AP10)	Deposited on a shallow marine shelf with high energy nummulitic shoals and deposited in a lagoonal environment in a subtropical sea.	Consists mainly of neritic shoal limestones often recrystallised and/or dolomitised, nummulitic

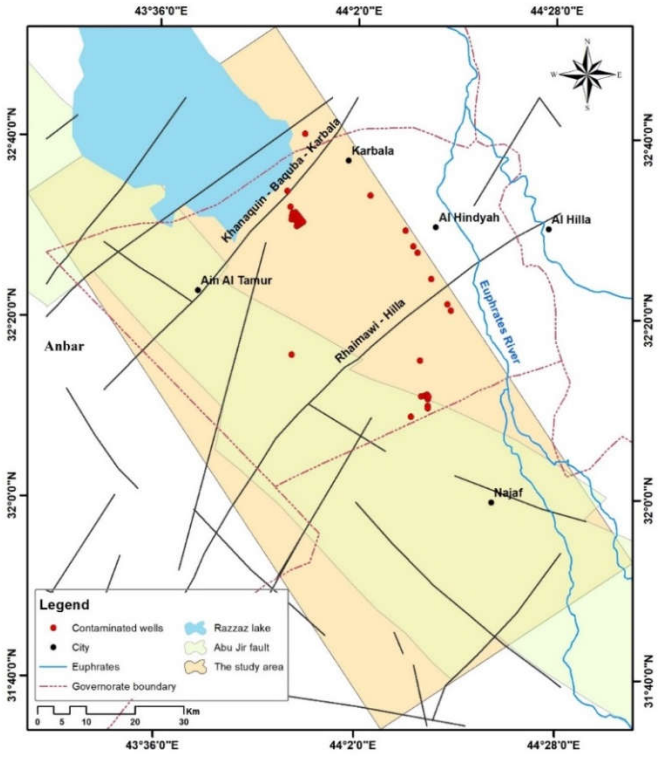

Figure 2. Structural features in the study area.

Materials and Methods

To model the probability of heavy oil contamination in the Dammam aquifer, four steps were adopted in this study: (1) preparing the inventory map of the contaminated and non-contaminated wells, (2) preparing maps of factors influence, the contamination according to the available data and site condition, (3) building the ANN model and examining the accuracy of the model using error statistics, and (4) mapping the contamination probability using the BPANN model.

Well Inventory Map:

The well inventory map was prepared using field survey and archival data obtained from the General Commission of Groundwater/Karbala branch. Figure (3) shows 139 wells that were identified and classified as contaminated (Fig. 2) and non-contaminated well. The total number of wells was partitioned into two sets with a 70/30 ratio. The training dataset has 72 non-contaminated wells and 26 contaminated wells while the testing dataset has 30 non-contaminated and 11 contaminated wells.

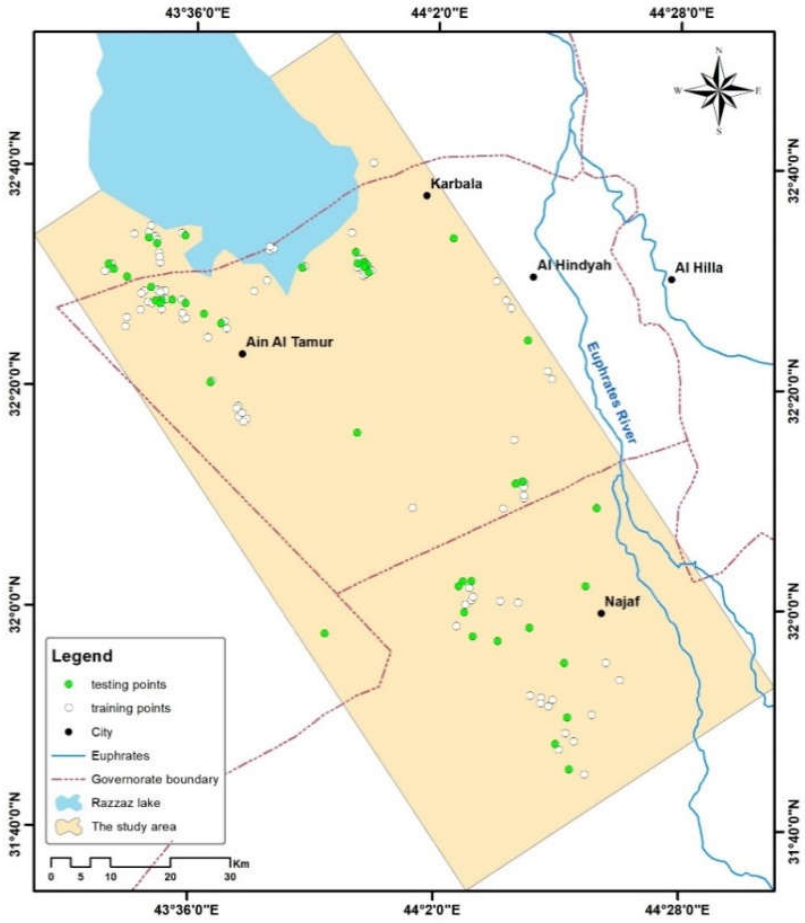

Figure 3. Training and testing dataset.

Preparing factors affecting the aquifer contamination:

Seven factors were selected to investigate the probability of groundwater contamination, the distance to faults (FDIS), fault density (FDEN), distance to Abu Jir fault (ADIS), aquifer hydraulic conductivity (k), aquifer saturated thickness (AST), depth to groundwater (GD), and surface elevation. The FDIS, FDEN, and ADIS were prepared based on the structural map of Iraq with a scale 1: 1,000,000. The faults were drawing manually and the Euclidian Distance tool in ArcGIS 10.5 was used to create the raster of the factors (Figs. 4a-4c).

For visualization purpose, the factor rasters categorized into ten equal zones. The FDEN values range from 0 to 0.0004 and they are generally very low. This may give the impression that this factor has only a minor role in controlling groundwater contamination. The aquifer parameters (k, AST, and GD) were prepared using pumping tests, analysis and drilling well records, and ordinary kriging interpolation technique (Figs. 4d-4f).

The k values increase generally from SE to NW direction and the high values are in the Razzaza lake and Ain Al-Tamur areas. Contrary, the high values of AST are in the southern part whereas the low values occupy the northern part. In the case of GD factor, the high values concentrate in the northern part and the low values distribute in the rest of the study area.

Finally, the elevation was developed using Digital Elevation Model (DEM)which is type of Shuttle Radar Topography Mission (SRTM) after preprocess of the original DEM (Fig. 4g).

Artificial Neural Networks:

ANN is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.

They are the most Machine-Learning (ML) models for solving computational problems efficiently and are used to solve a wide range of problems in different areas of artificial intelligence and Machine-Learning (ML) (Ciaburro and Venkateswaran, 2017).

The ANN comprises of large number of independent interconnected elements that are neurons and synapses (Poudyal *et al.*, 2010). Nodes are activated and send signals to the connecting units, only when individual neurons receive sufficient stimulus from preceding unit. The term network architecture is used to refer to the arrangement of the nodes in the artificial network. There are different types of ANN architecture; the most commonly used is the multi-layers perceptron (MLP).

MLP has been used extensively to solve problems via training them in a supervised manner with highly popular Backpropagation algorithm (Haykin, 2008). It has an ability to simulate non-linear systems without a prior assumption of process involved and provide to be efficient even the input data are incomplete or ambiguous (ASCE Task Committee 2004). The MLP comprises of three or more layers (input, output, and one or more hidden layers) of nonlinearly-activating nodes.

Learning happens within the perceptron by changing association weights after each piece of information is handled, based on the amount of error in the output compared to the expected result. More detail of how this algorithm works and the mathematical computation involved can be found in Haykin (2008) and Gurney (1997).

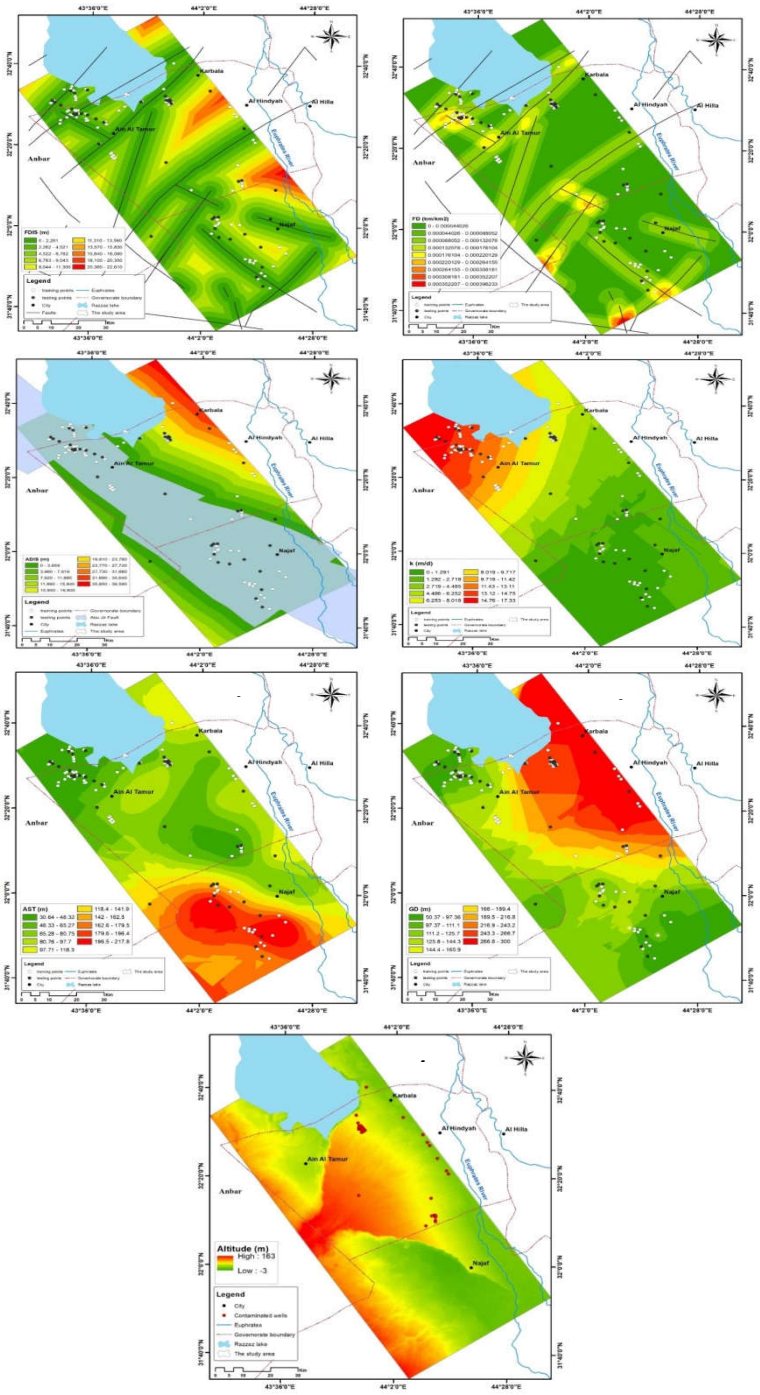

Figure 4. Factor affecting groundwater contamination (a) FDIS (m), (b) FDEN (km/km^2), (c) ADIS (m), (d) k (m/d), (e) SAT (m), (f) GD (m) and (g) Elevation (m).

Model Performance Metric Measures:

To test the model developed in this study, five evaluation performance measures were used: Accuracy, Sensitivity, Specificity, kappa, and ROC. Accuracy is the percentage of wells (contaminated and non-contaminated) that correctly predicted by the classifier. It is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where, TP is the number of contaminated wells predicted as contaminated, FN the number of contaminated wells that incorrectly predicted, TN is the number of non-correctly wells that predicted correctly, and FP is the number of contaminated wells that erroneously predicted as non-contaminated.

On the other hand, Sensitivity is the proportion of contaminated cases that are correctly predicted and is calculated using the following formula:

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

Specificity is defined as the proportion of non-contaminated cases that are correctly predicted:

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

Kappa (k) is a statistical measure to examine the agreement between two models. It is calculated as:

$$k = \frac{p_0 - p_e}{1 - p_0} \quad (4)$$

where p_0 is the observed agreement, and p_e is the expected agreement (Moosavi and Niazi 2016).

ROC, on other hand, is a performance test for classification problems at various thresholds settings. This curve is built by plotting sensitivity on y-axis versus (1-specifictiy) on x-axis. Area under ROC is between 0.5 and 1. As a rule, higher the AUC, better the model is.

Results and Discussion

To construct and train the ANN model, the caret and neurelnet packages of R software were used in this study. A MLP with feedforward back propagation algorithm was used consisting of 7 neurons (factors effecting groundwater contamination in the Dammam aquifer), one hidden layer, and one neuron in output layer. Firstly, the raw data was pre-processed using center-scaling functions, the hyper parameters of the algorithm was tuned using the grid-search approach, and the best results were obtained using the following values: (hidden layers = 4, learning rate = 0.3, momentum = 0.2). The final architecture of the ANN model was set on 7: 4: 1 (Fig. 5).

The test data was then passed to the ANN model and the results in terms of confusion matrix and performance evaluation measures were obtained (Tables 2 and 3). The obtained accuracy in this stage of analysis was 0.976. The ANN model is totally classified contaminated cases as contaminated (Sensitivity = 1.00), and is correctly classified non-contaminated cases for 20

instances and fail in one instance (Specificity = 0.967). In terms of kappa statistics, the model is almost perfect (k = 0.939). Contrary, the AUC for this model was 0.976 and is regarded as excellent model (Yesilnacar, 2005).

The ANN probability prediction for training and testing stages were interpolated to reveal the groundwater contamination hazard in the study area (Fig. 6). Values of probability prediction was 0.001 to 0.998 with an average of 0.433. These values were then classified into five classes (contamination hazard level) using NB method: very low, low, moderate, high, and very high. The very low-low classes occupy 5195 km^2 (64%) and mainly distribute in the northwestern and southeastern parts, whereas the high-very high zones encompass 2516 km^2 (31%) and occupy most of the Karbala-Najaf plateau (Table 4). The moderate zone distributes between the very low-low and high-very high zones as a strip (yellow color) and occupy only 406 km^2 (5%) from the total area and distribute spatially on the Karbala-Najaf plateau and close to the Razzaza lake.

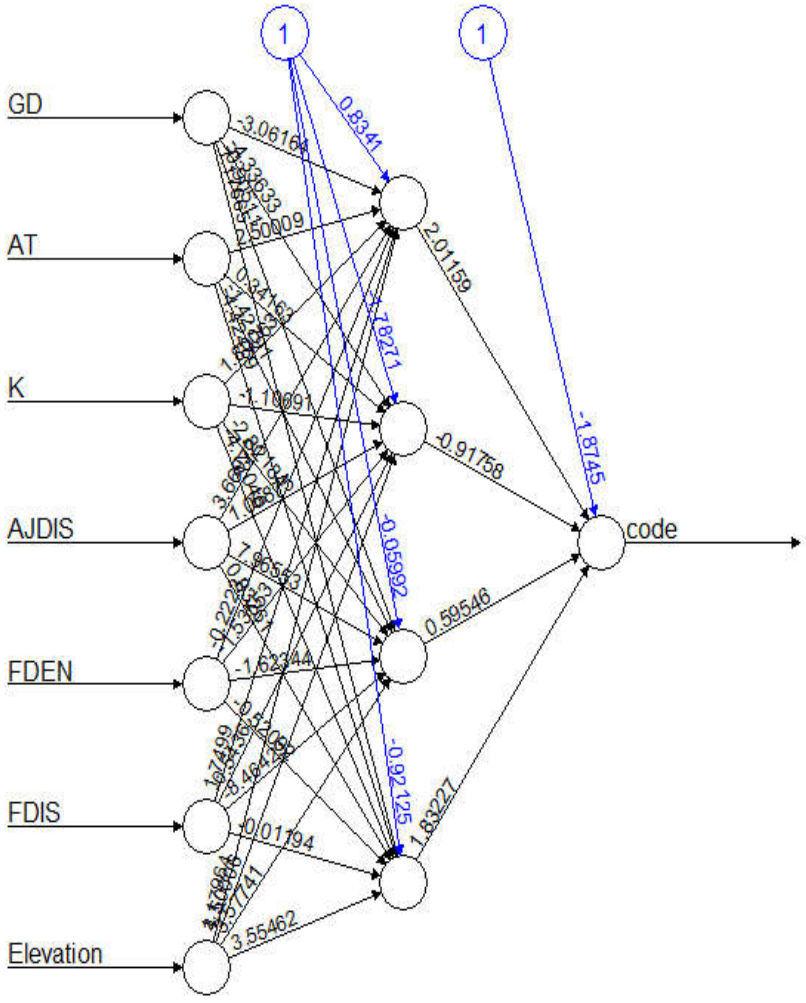

Figure 5. The architecture of ANN model used in this study.

Table 2. Confusion matrix of the ANN model (testing stage).

Model			Predicted		Overall class error
			yes	no	
ANN	Actual	yes	11	0	0.00
		no	1	29	0.03

Table 3. Model performance using statistical evaluation metrics.

Evaluation Measure	Value
Accuracy	0.976
Sensitivity	1.000
Specificity	0.967
Kappa	0.939
AUC	0.976

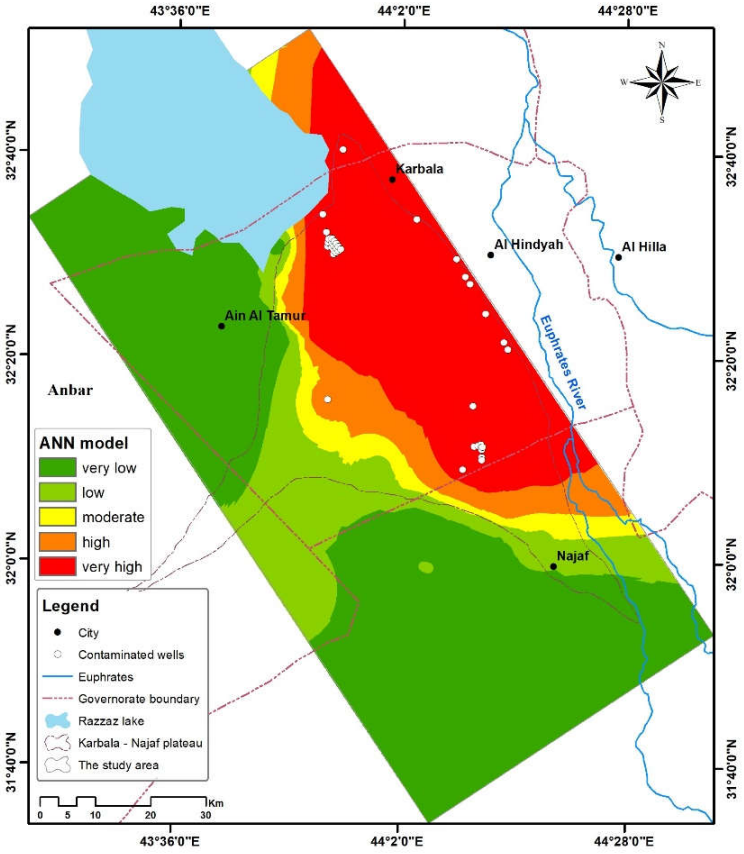

Figure 6. Probability of groundwater contamination in the Dammam aquifer.

Table 4. Areas occupied by different categories of the groundwater contamination probability values.

Model	Groundwater contamination probability zones		
	very low-low	Moderate	High-very high
ANN	64% (5195 km^2)	5% (406 km^2)	31% (2516 km^2)

Conclusions

Groundwater is an important resource for the economy of Iraq especially in the regions where the surface water is absent. The Dammam Formation is an important aquifer in the southern and western deserts of Iraq as having a huge amount of groundwater suitable for different purposes such as breeding fish, industrial, agricultural, and even drinking after a suitable treatment. In the Karbala-Najaf plateau and the adjacent areas, the drilling of groundwater wells in the Dammam aquifer is risky due to the possibility of contamination of this aquifer with hydrocarbon.

The source of the hydrocarbons is believed to be the deep oil reservoirs which rise to the Dammam aquifer via the faults that distribute through the region. The only map that shows the probable extent of contamination in this aquifer is provided by Al-Abadi (2018) and for only the Karbala-Najaf plateau and using the MCDM model. This study focuses on demarcating the probable extension of hydrocarbon contamination in the Dammam aquifer in a wider area than Karbala-Najaf plateau and using the more advanced ML models. The objective is to draw a map that show the probable extent of this contamination and thus providing a practical way for drilling non-contaminated groundwater wells. To attain the objective of this study, the relationship between the geographical locations of available contaminated-non-contaminated groundwater wells (*target* variable in the analysis) and a specific number of factors (*predictors*) that belief to control the groundwater contamination of the Dammam aquifer were investigated. These predictors were selected based on the structural setting, aquifer characteristics, and expert opinions. Seven predictors were involved seven predictors: fault density, distance to faults, distance to Abu Jir fault, groundwater depth, aquifer hydraulic conductivity and saturated thickness, and elevation.

The ANN model was then used to examine the relationship between *target* and *predictors*. The performance of the ANN model were then investigated in the test stage using five evaluation statistics: Accuracy, Sensitivity, Specificity, kappa, and AUC. The probability prediction of these ANN model in training and testing stages were then exported to the ArcGIS 10.3 software, interpolated, and then categorized into five groundwater contamination levels: very low-low, moderate, high-very high. The average areas occupied between these zones were distributed as: 5267 km^2 (65%), 488 km^2 (6%), and 2362 km^2 (29%), for very low-low, moderate, and high-very high, respectively. With respect to the spatial distribution of these zones, the high-very high encompass most part of the Karbala-Najaf plateau, moderate zone form a strip and mainly distributed in the Karbala-Najaf plateau and the area close to the east side of Razzaza Lake. Contrary, the low-very low zones are distributed in the northwestern and southeastern parts and occupy the desert area. As a conclusion, to drill successful, non-contaminated groundwater wells, it is better to exclude the center and the east parts of the Karbala-Najaf plateau and the areas close to the east of Razzaza Lake.

References

Al-Abadi, A.M. 2018. Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: A comparative study. Arab. J. Geosci., 11: 218. https://doi.org/10.1007/s12517-018-3584-5

Al-Abadi, A.M., Al-Kubisis, Q.Y. and Al-Ghanimy, M.A. 2018. Mapping groundwater zones contaminated by hydrocarbons in the Dammam aquifer in the Karbala-Najaf plateau, Iraq. Environ. Earth Sci., 77: 633. https://doi.org/10.1007/s12665-018-7827-2

ASCE Task Committee 2000. Artificial neural networks in hydrology I: preliminary concepts. Journal of Hydrologic Engineering, 5(2): 115-123.https://doi.org/10.1007/s12665-018-7827-2

Buday, T. and Jassim, S.Z. 1987. The regional geology of Iraq: Tectonism, Magmatism, and Metamorphism. Edited by Kassab, I.I. and Abbas, M.J., Baghdad, Vol. 2, 352 p https://scholar.google.com/scholar?cluster=8673684302455840 74&hl=ar&as_sdt=2005&sciodt=0,5

Chen, W., Han, H., Huang, B., Huang, Q. and Fu, X. 2018. A data-driven approach for landslide susceptibility mapping: A case study of Shennongjia Forestry District, China. Geomatic, Natural Hazards and Risk, 9(1): 720-736. https://doi.org/10.1080/19475705.2018.1472144

Ciaburro, G. and Venkateswaran, B. 2017. Neural Networks with R. Packet Publishing, UK., 314p.URL

Cortes, C. and Vapnik, V. 1995. Support-Vector Networks. Machine Learning, 20(3): 273-297. https://doi.org/10.1023/A:1022627411411

Gayen, A., Pourghasemi, H.R., Sunil, S., Keesstra, S. and Bai, S. 2019. Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. Science of the Total Environment, 998: 124-138. https://doi.org/10.1016/j.scitotenv.2019.02.436

Gurney, K. 1997. An Introduction to Neural Networks. UCL press. London, 148p.URL

Haykin, S. 2008. Neural Networks and Learning Machines. Prentice Hall 3rd , 937p.URL

Jassim, S.Z. and Goff, J.C. 2006. Geology of Iraq. Dolin, Prague and Moravian Museum, Brno, 431p. URL

Lee, S., Hong, S-H. and Jung, H-S. 2018. GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. Geocarto International, 33(8): 847-861. https://doi.org/10.1080/10106049.2017.1303091

Moosavi, V. and Niazi, Y. 2016. Development of hybrid wavelet packet-statistical models (WP-SM) for landslide susceptibility mapping. Landslides, 13(1): 97-114. https://doi.org/10.1007/s10346-014-0547-0

Poudyal, C.P., Chang, C., Oh, H-J. and Lee, S. 2010. Landslide susceptibility maps comparing frequency ratio and artificial neural networks: a case study from the Nepal Himalaya. Environ Earth Sci., 61(5): 1049-1064. https://doi.org/10.1007/s12665-009-0426-5

Yesilnacar, E.K. 2005. The application of computational intelligence to landslide susceptibility mapping in Turkey. Ph.D. Thesis, Department of Geomatics, University of Melbourne, 423 pp.URL

Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F. and Pourghasemi, H.R. 2018. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China. Computers and Geosciences, 112: 23-37. https://doi.org/10.1016/j.cageo.2017.11.019