



Random Forest and Logistic Regression algorithms for prediction of groundwater contamination using ammonia concentration

Ahmed Madani¹ · Mohammed Hagage² · Salwa F. Elbeih²

Received: 31 December 2021 / Accepted: 24 September 2022 / Published online: 15 October 2022
© The Author(s) 2022

Abstract

The present study aims to develop an efficient predictive model for groundwater contamination using Multivariate Logistic Regression (MLR) and Random Forest (RF) algorithms. Contamination by ammonia is recorded by many authors at Sohag Governorate, Egypt and is attributed to urban growth, agricultural, and industrial activities. Thirty-two groundwater samples representing the Quaternary aquifer are collected and analyzed for major cations (Ca, Mg, and Na), ammonia, nitrate, phosphate, and heavy metals. Lead, magnesium, iron, and zinc variables are used to test the model with ammonia which is used as an index to the groundwater contamination. Spatial distribution maps and statistical analyses show a strong correlation of ammonia with lead and magnesium variables whereas iron and zinc show less correlation. For Random Forest (RF) model, the data is divided into 70% training and 30% testing subsets. The performance of the model is evaluated using the classification reports, and the confusion matrix. Results show (1) high performance of RF model to groundwater contamination with an accuracy of 93% and (2) the MLR accuracy increased from 70 to 83% when “SOLVER” and “C” parameters are modified. The study helps to identify the contaminated zones at the study area and proved the usefulness of the machine learning models for prediction of the groundwater contamination using the ammonia concentration.

Keywords Ammonia contamination · Quaternary aquifer · Multivariate Logistic Regression · Random Forest · Sohag-Egypt

Introduction

Groundwater is an important source for many agricultural and industrial activities at Akhmim area, Sohag Governorate, Egypt. Many authors assessed the quality of groundwater and its suitability for drinking and irrigation (among them, Hagage et al. 2021; Balamurugan et al. 2020, 2021, Elbeih and El-Zeiny 2018; Ismaila and El-Rawyba 2018; Gedamy 2015; Melegy et al. 2014; Ahmed and Ali 2011). Elbeih and El-Zeiny (2018) evaluated the groundwater quality west of Sohag governorate in 2008 and 2016, based on some physicochemical characteristics

of groundwater and set of retrieved land use spectral indices. Ismaila and El-Rawyba (2018) assessed and evaluate the hydrochemical properties of groundwater resources west of Sohag, Egypt based on chemical analyses of groundwater samples collected in 2014. Melegy et al. (2014) studied the geo-chemical mobility of some heavy metals in water resources and their impact on human health in west of Sohag Governorate. The results recorded high contamination with cadmium and lead and about 50% of water samples are contaminated with iron and manganese. Ahmed and Ali (2011) reported that groundwater resources of Sohag are threatened by pollution resulting from urbanization and agricultural activities. Ammonia is found in groundwater naturally as a result of anaerobic decomposition of organic materials (Bohlke et al. 2006). It reached the groundwater through the leakage from sewage systems (Johan Lindenbaum 2012). Hagage et al. (2021) studied the suitability of the groundwater for drinking and irrigation in the Akhmim area, Egypt. They concluded that about 95% of the collected groundwater samples are highly contaminated

Responsible editor: Broder J. Merkel

✉ Ahmed Madani
aamadani@sci.cu.edu.eg

¹ Department of Geology, Faculty of Science, Cairo University, Giza, Egypt

² National Authority for Remote Sensing and Space Sciences (NARSS), Cairo, Egypt

with ammonia. This contamination resulted from urban growth and agriculture and industrial activities. The present study proposed a predictive model for the groundwater contamination at Akhmim area, Sohag Governorate, Egypt using Random Forest (RF) and the Multivariate Logistic Regression (MLR) algorithms.

Machine learning is algorithmic study of how computers simulate or implement human learning behavior. Machine learning algorithms are designed to predict accurately patterns within multivariate data (Cracknell 2014). They are classified into three main classes: supervised, unsupervised, and reinforcement algorithms (Russell and Norvig 2010). They are widely used in many applications such as pattern recognition, anomaly detection, and classification. Implementation of machine learning (ML) algorithms such as Logistic Regression and Random Forest in the prediction of water quality and groundwater contamination are tested and evaluated by many authors (Venkataraman and Uddameri 2012; Mair and El-Kadi 2013; Solanki et al. 2015; Wang et al. 2017; Muharemia et al. 2019; Vijay and Kamaraj 2019a; Rizeei et al. 2018; Hosseini et al. 2018; Aldhyani et al. 2020). Venkataraman and Uddameri (2012) utilized a Logistic Regression model to predict the exceedance of drinking-water standards of arsenic and nitrate in the Southern Ogallala aquifer. Mair and El-Kadi (2013) assessed the groundwater vulnerability to contamination in Hawaii using the Logistic Regression model. Madani and Niyazi (2015) utilized a knowledge-driven GIS model for groundwater potentiality mapping over wadi Yalamlam, Western Saudi Arabia. Solanki et al. (2015) utilized deep learning algorithms for the prediction of the water quality parameters in India. Wang et al. (2017) combined machine learning algorithms, WQI, and remote sensing spectral indices to establish a model for assessing the water quality in China. Rizeei et al. (2018) utilized a data-driven Logistic Regression model to assess the groundwater nitrate contamination hazard in a semi-arid region. Hosseini et al. (2018) presented a novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Quedraogo et al. (2018) applied the Random Forest regression in modeling groundwater contamination in Africa and compared its performance with a multiple linear regression model. Muharemia et al. (2019) applied several machine learning (ML) models to identify the anomalies in water quality time series data. Results showed that DNN, RNN, and LSTM algorithms are very vulnerable compared to SVM and LR models. Vijay and Kamaraj (2019b) investigated three ML models to predict groundwater quality. They concluded that the C5.0 classifier produced a better result with an accuracy of 96%. Venkatech et al. (2020) utilized tree-based modeling

methods to predict nitrate exceedance in the Ogallala aquifer in Texas. Aldhyani et al. (2020) utilized several ML models for water quality prediction. Results showed that SVM algorithms achieved the highest accuracy (97%) for water quality prediction. Nafouanti et al. (2021) compared the Random Forest, Logistic Regression, and artificial neural network algorithms for the fluoride contamination in groundwater at the Datong Basin, Northern China. The paper is organized as follow: (1) description of the study area, (2) geological and hydrogeological background, (3) selection of the relevant variables through spatial-statistical analyses, (4) machine learning model implementation, and (5) model evaluation through the generation of the performance metrics.

Study area

The study area (Akhmim District) is located east of the River Nile between latitudes 26°30' and 26°44'N and longitudes 31°35' and 31°55'E (Fig. 1) about 467 km apart from Cairo. Several authors studied the hydrochemical characteristics of the groundwater of Sohag governorate and evaluated the impact of the human activities on the groundwater quality (Awad et al. 1995; Ahmed 2009; Abdel Latif and El Kashouty 2010; Youssef et al. 2011; Ahmed and Ali 2011; Melegy et al. 2014; Ismaila and El-Rawyba 2018; Elbeih and El-Zeiny 2018; Hagage et al. 2021). Youssef and Abdel Moneim (2006) studied the geo-environmental impacts of the area east of the Sohag governorate and revealed the existence of three main geo-environmental hazards. Industrial, domestic, and agricultural activities are the main groundwater contamination sources recorded in the study area. Hagage et al. (2021) studied the impacts of anthropogenic activities on the archaeological sites in the Akhmim area, Sohag Governorate, Egypt using remote sensing and GIS techniques. In this study, the authors utilized the ammonia data as an indicator for groundwater contamination because its elevated concentrations in groundwater are typically caused by anthropogenic activities. Ammonia includes the non-ionized (NH_3) and the ionized (NH_4) species. The most common nitrogen compound in groundwater is NO_3 , but in a reducing environment, ammonia is predominant. Hagage et al. (2021) identified the groundwater deterioration sources through extensive field investigation. Growing population and urbanization without proper urban planning on the study area lead to generating random sewerage systems (Hagage 2021). Lack of sewerage networks in the study area forces the inhabitants to build septic tanks and injection wells which leads to contamination of the groundwater, where more than half of the population do not have any sewerage networks (Hagage 2021).

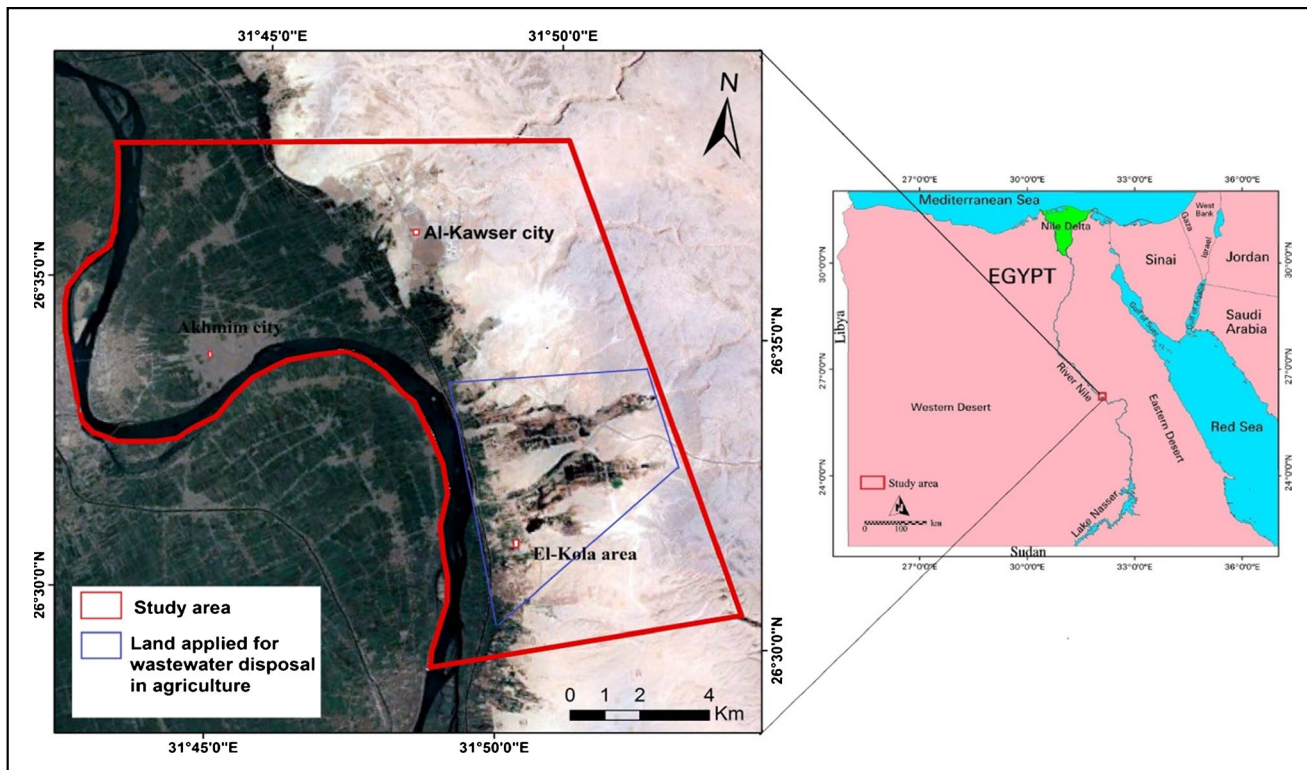


Fig. 1 Location map of the study area

Geological and hydrogeological background

The main features that characterized the study area include the cultivated Nile flood plain and the lowland desert areas along both sides of the Nile Valley (Youssef and Abdel Moneim 2006). Several authors studied the sedimentary sequence in the study area among them (Said 1960, 1981, 1990; Issawi et al. 1978; Issawi and Hinnawi 1980; Omer 1996). They revealed that it starts from the base by Lower Eocene Thebes Formation, followed by Issawia Formation, Pre-Nile Sediments, Fanglomerate, Nile silt, and the Recent Wadi deposits. Figure 2 shows a part of the geological map of the study area.

The main aquifer system in the study area is the Quaternary aquifer where the Pleistocene deposits are the major water-bearing sediments in the Akhmim area (Abdel Moneim 1999; Hagage 2021). On ancient cultivated lands, the Pleistocene aquifer is a semi-confined aquifer where the upper member consists of a clay-silt layer while in the desert fringes it is an unconfined aquifer where the clay-silt layer is replaced by desert sands. The lower boundary of the aquifer is extensive with thick deposits of Pliocene clays (Abdel Moneim 1992). The groundwater flows towards the River Nile where the groundwater level in the study area ranges from 63 m (masl) at the valley fringes to 53 m (masl) close to the River Nile. Hagage et al. (2021)

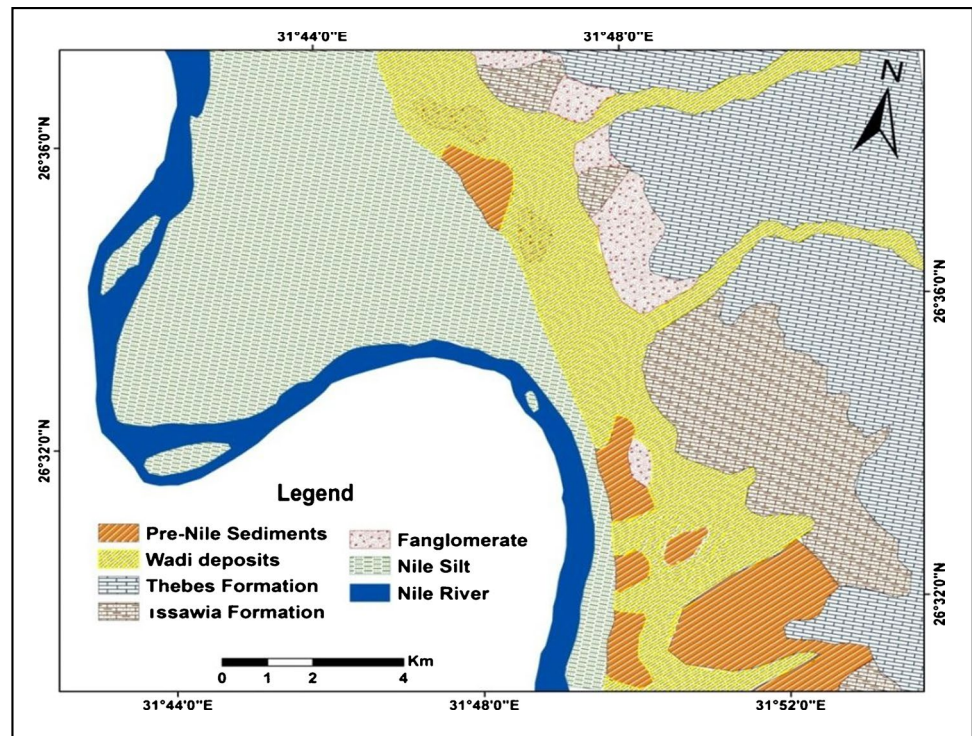
studied the groundwater quality and its suitability for drinking and irrigation in Akhmim District, Sohag, Egypt. The results of their study showed the existence of several human activities that affect the quality of groundwater and its suitability for drinking.

Materials and methods

Spatial distribution maps of index and independent variables

A field trip to the study area took place in April 2019 where 32 groundwater samples representing the Quaternary aquifer were collected. The water samples were collected in one-liter polyethylene bottles. For heavy metal analysis, 100 ml of sample was acidified with nitric acid (1%) and preserved separately and all water samples were locked carefully and labeled after collection and kept in the refrigerator until analysis. The analyses were performed in the Central Laboratory of the National Water Research Center, according to standard methods for testing water as described by the American Public Health Association (APHA 2005). The samples were chemically analyzed to determine the cations, anions, nutrients (ammonia, nitrate, and phosphate), and soluble heavy metals.

Fig. 2 Geological map for the study area (EGSMA 1983)



Index variable

In this study, ammonia is considered as the index to groundwater contamination. Ammonia concentrations in the groundwater samples range between 0.01 and 22.4 mg/l. The presence of ammonia in water is evidence of fecal pollution from wastewater and it can relatively be oxidized to nitrite and finally nitrates (Karavoltzos et al. 2008). According to the WHO (2011), the maximum

permissible ammonia concentration is 0.5 mg/l and about 95% of the groundwater collected samples are contaminated with ammonia. Figure 3a shows the histogram distribution of the ammonia concentration, whereas Figure 3b demonstrates its box-plot pattern. The figure shows the distortion of distribution with skewed values. Figure 4 shows the spatial distribution map of the ammonia concentration. The highest concentrations are recorded at the southern and middle parts of the study area whereas the

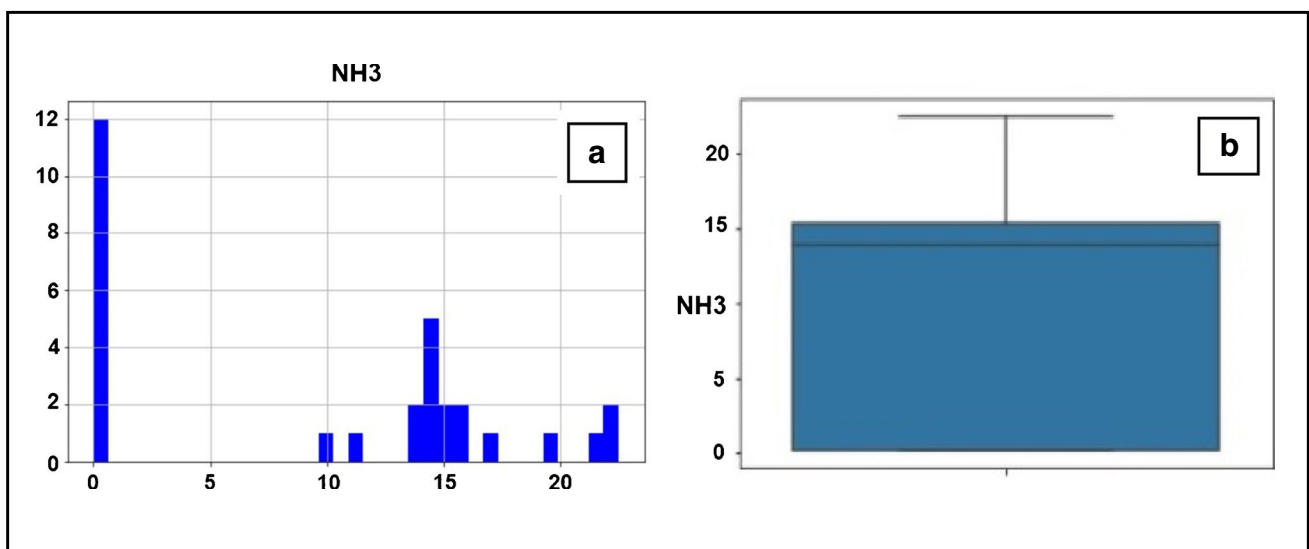
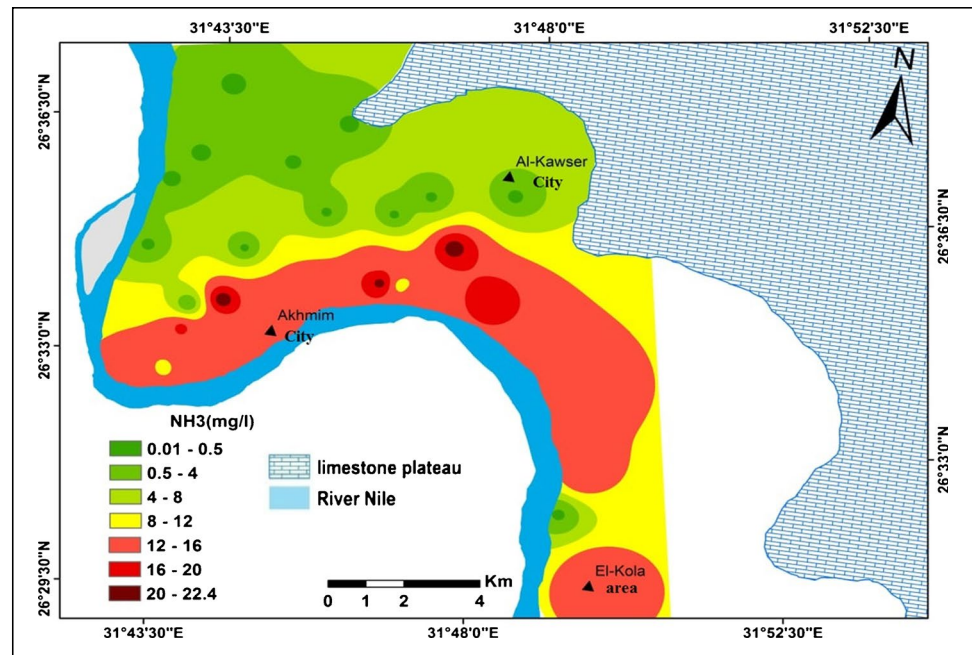


Fig. 3 a Histogram distribution of the ammonia. b Box-plot chart of the ammonia

Fig. 4 Distribution map of the ammonia concentration



northern part records values below the permissible ammonia limit. The ammonia concentrations at the southern part range between 8 and 22.4 mg/l whereas the middle parts record less than 8 to 0.5 mg/l.

Independent variables

About 16 physicochemical parameters are prepared and analyzed to clarify their correlation to ammonia. Only four variables are found to be relevant to the ammonia contamination. Figure 5a, b, c, and d show the spatial distribution pattern of the Pb, Mg, Fe, and Zn variables. The Pb and Mg distribution maps show a high correlation to ammonia pattern whereas Zn and Fe maps show less correlation. Figure 5 demonstrates that Akhmim and El Kola are the highly polluted sites which is due to industrial activities, urbanization, excessive use of chemical fertilizers, agricultural pesticides, and sewage leakage.

Statistical analysis

The correlation analysis is carried out to clarify the existence of a relationship between the measured variables. The correlation coefficient (CC) could be negative or positive and it ranges from -1.00 to $+1.00$. Large values of the correlation coefficient between two variables imply that they are highly correlated and this might be in the positive or negative direction. We consider the presence of a strong correlation when the CC value is greater than 0.5 and it is a weak correlation when the value is less than 0.5. The correlation values scaled between 0 (no correlation) and 1 (perfect correlation) can

be encoded with color in a 2D heat map. The R^2 values can be translated to color saturation and produced a heat map (Fig. 6) that shows the correlation scores between the independent variables and the ammonia. It confirms the presence of a strong correlation between the Pb and Mg with NH_3 and a weak correlation with the Zn and Fe variables.

Table 1 provides some descriptive statistics for the relevant variables used to predict groundwater contamination. High standard deviation is recorded by Mg compared to the standard deviation of Fe, Zn, and Pb. The maximum and minimum values of the Mg are 34 and 10.3 mg/l with a 24.12-mg/l mean value.

The maximum and minimum values of the Fe are 0.148 and 0.006 mg/l with a 0.025-mg/l mean value. The maximum and minimum values of the Pb are 0.05 and 0.004 mg/l with a 0.019-mg/l mean value. The maximum and minimum values of the Zn are 1.8 and 0.009 mg/l with a 0.18-mg/l mean value. The box plots of the Pb, Fe, and Zn variables (Fig. 7) show normal distribution whereas the box plot of the Mg shows little distortion, and the values are skewed. Because the Fe, Pb, and Zn values are not of the same magnitude as Mg values, the data are normalized before ML model implementation.

Machine learning model selection and implementation

The general methodology of machine learning includes (a) data preparation, analyses, and visualization; (b) normalization; (c) model selection and implementation; and (d) performance metrics. In this study, the following steps are

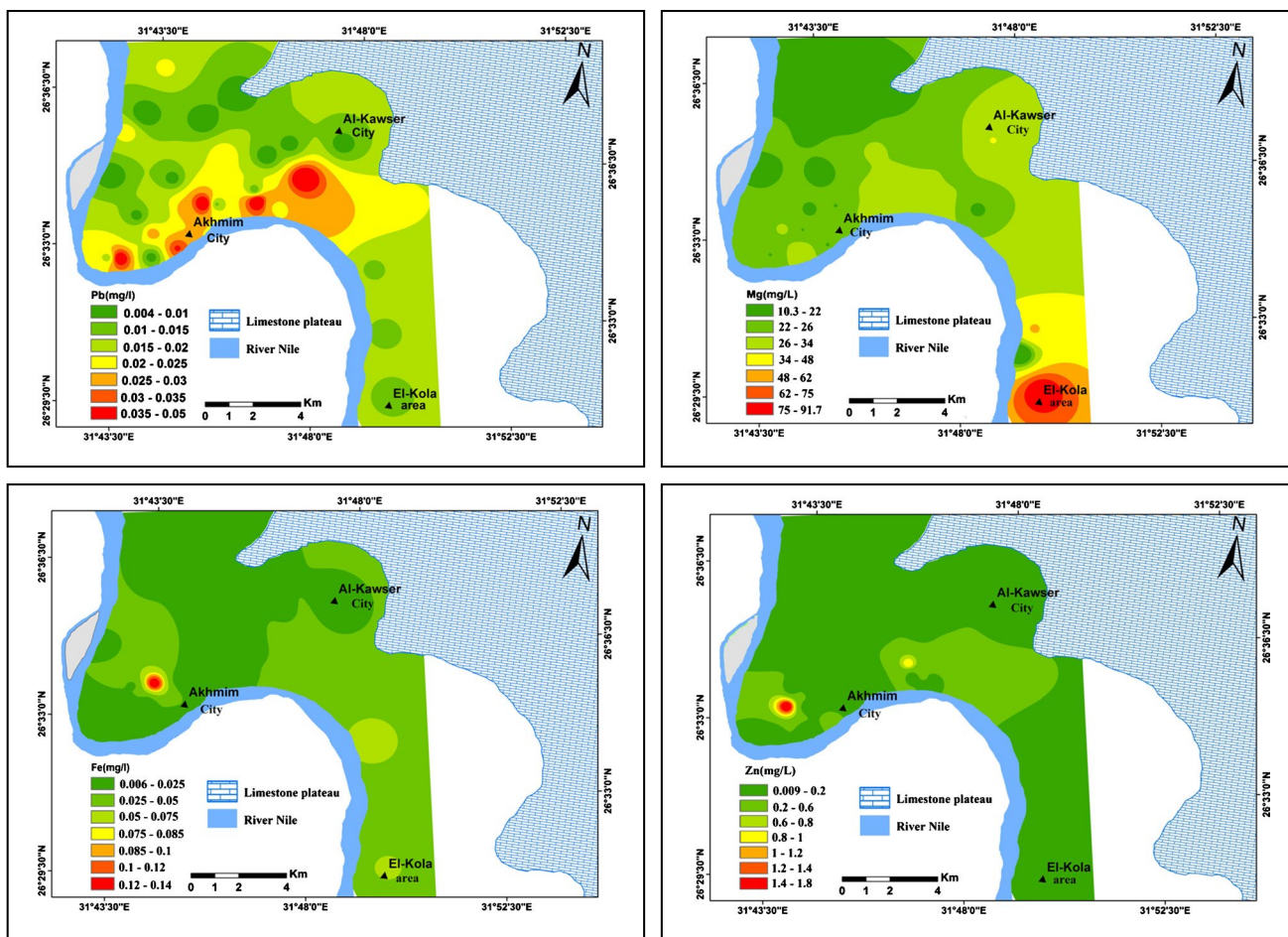


Fig. 5 Distribution maps of a lead, b magnesium, c iron, and d zinc concentrations

implemented using “Python” Code within “Anaconda Notebook”: (1) import the required libraries (NumPy, Pandas, Matplotlib, and Seaborn), (2) import the “-.csv” file containing the dataset (Mg, Fe, Pb, and Zn variables and NH₃), (3) statistical analyses and data visualization, (4) Multivariate Logistic Regression and Random Forest models, and (5) performance metrics (classification reports and confusion matrix). The following paragraphs describe each step in detail.

Figure 8 shows the pair plots of the relevant variables against NH₃. No clear relation is observed. In this case, ensemble models are the best to treat with these kinds of data. This study implemented an ensemble Random Forest model in addition to the Multivariate Logistic Regression models.

Multivariate Logistic Regression model

Logistic Regression is one of the most commonly used machine learning algorithms for predicting two classes. Linear

Regression model is a linear function that demonstrates a relationship between different variables and is expressed by Eq. (1):

$$y = \beta_0 + \beta_1x + \epsilon \tag{1}$$

where y is a dependent variable, x is an independent variable, β_0 is the y -intercept, β_1 is the slope, and ϵ is a random error.

The sigmoid function is represented by Eq. (2):

$$P = 1 / 1 + e^{-y} \tag{2}$$

Apply the sigmoid function on linear regression function in Eq. (3):

$$P = 1 / 1 + e^{-(\beta_0 + \beta_1x)} \tag{3}$$

The multiple Logistic Regression model considers a set of x independent variables, which in this study are represented by the Mg, Fe, Pb, and Zn variables, to predict the likelihood of the response variable Y which is represented by the NH₃. This model is expressed as in Eq. (4):

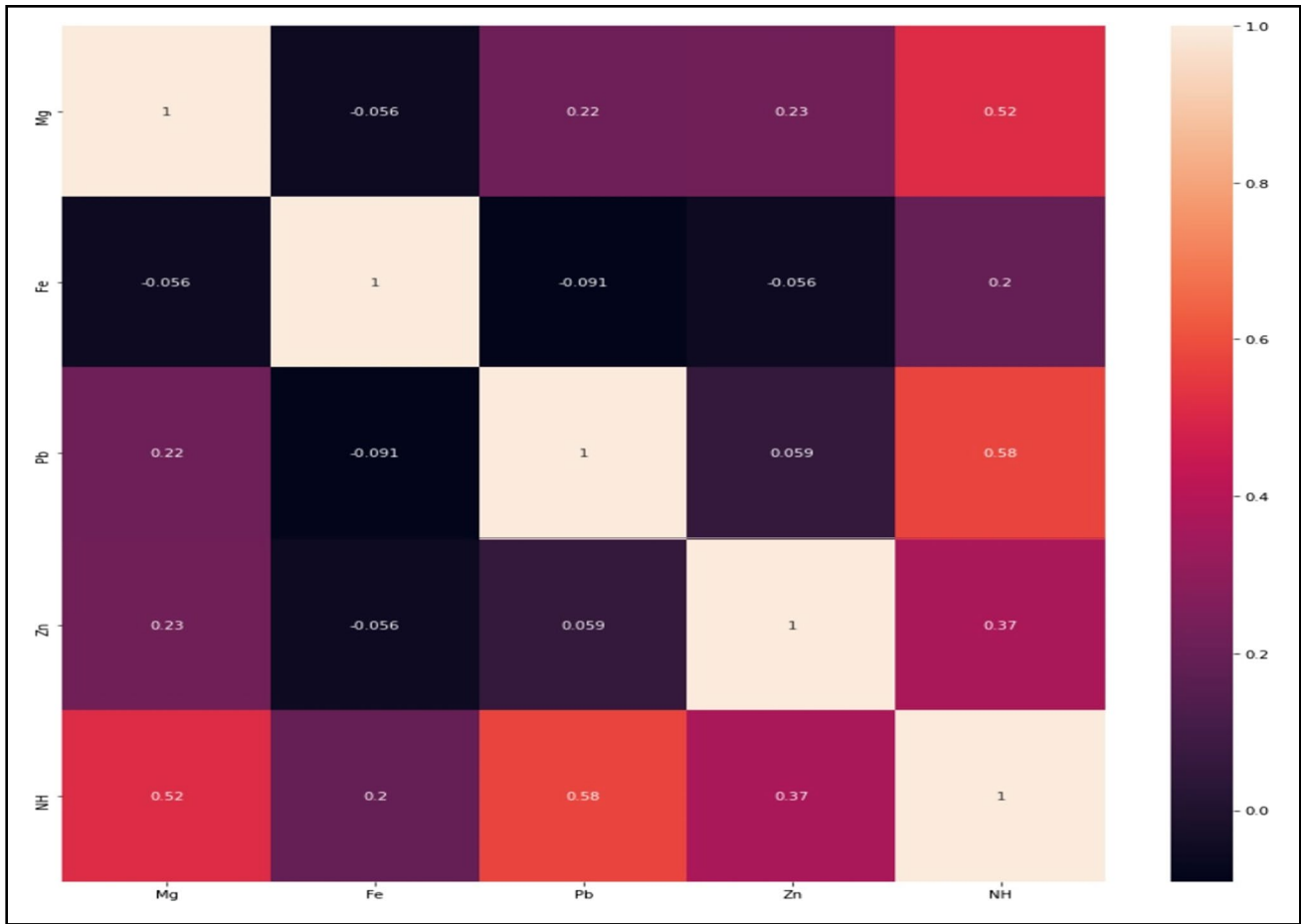


Fig. 6 Heat map of the Mg, Pb, Fe, and Zn variables correlated with the ammonia

Table 1 Descriptive statistics

	Mg	Fe	Pb	Zn	NH
count	30.000000	30.000000	30.000000	30.000000	30.000000
mean	24.124000	0.025933	0.019933	0.18140	0.600000
std	6.107463	0.026598	0.014130	0.37462	0.498273
min	10.300000	0.006000	0.004000	0.00900	0.000000
25%	21.000000	0.012500	0.007000	0.02200	0.000000
50%	23.725000	0.02500	0.016500	0.04500	1.000000
75%	29.602500	0.027250	0.027250	0.10250	1.000000
max	34.000000	0.148000	0.050000	1.81000	1.000000

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \epsilon \tag{4}$$

where Y is a dependent variable; $x_1, x_2, x_3,$ and x_4 are the independent variables; β_0 is the y intercept; $\beta_1, \beta_2, \beta_3,$ and β_4 are the slope; and ϵ is a random error.

Apply the sigmoid function in Eq. (5):

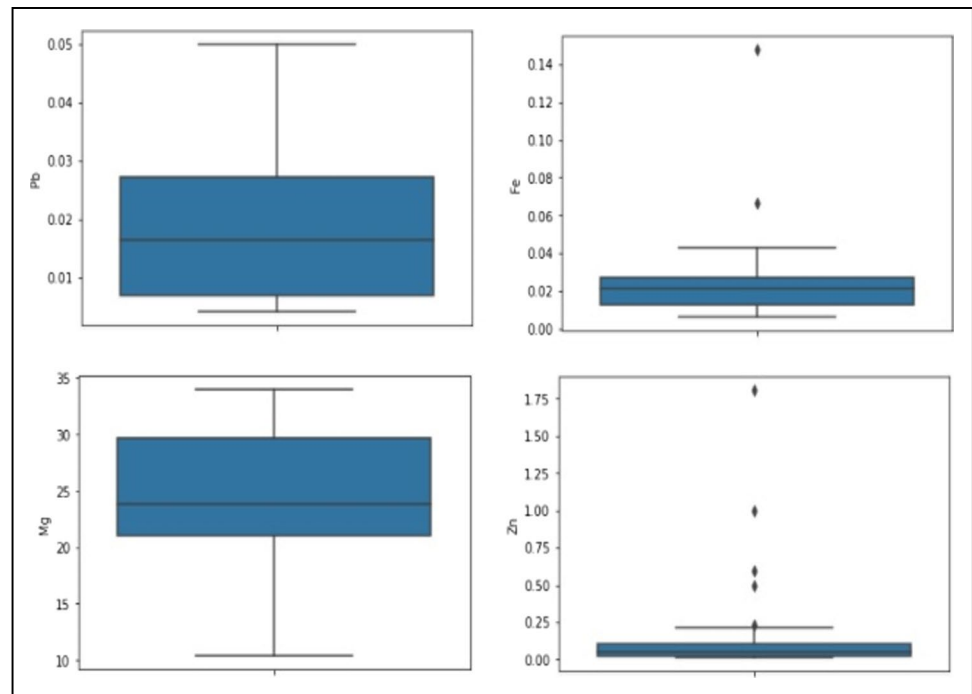
$$P = 1/1 + e^{-(\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4)} \tag{5}$$

The Logistic Regression model has several optional parameters such as solver, random state, and C. The parameter “solver” is a string that decides what solver to use for fitting the model. The default is the “liblinear,” whereas the “newton-cg,” “lbfgs,” “sag,” and “saga” are other options. The “random_state” parameter is an integer that defines what pseudo-random number generator to use. The default is none. The “C” is a positive floating-point number that defines the relative strength of regularization. Smaller values indicate stronger regularization and the default is 1.0. In this study, the authors implemented the MLR model two times. The first is with the following parameters: solver is liblinear, C = 10.0, and the random_state = 0, whereas the second run has the following parameters: solver=sag, C = 80.0, and random_state = 0.

Random Forest (RF) model

Random Forest (RF) is an ensemble classification/regression method that trains several classifiers and combines the results

Fig. 7 Box-plot charts of the relevant variables to the ground-water contamination



through a voting process (Breiman et al. 1984; Breiman, 2001; Gislason et al. 2006; Pham et al. 2019). It is a method where a large number of decision trees are created and each tree is trained on the original training data and the output class is determined by a majority vote of the trees. Random Forest searches across a randomly selected subset of variables to determine a split for each node based on some metric. The type of metric is different for regression and classification tasks. In this study, the dataset has been divided into 70% training and 30% testing subsets for RF implementation. The model run under the following parameters: criterion="entropy," n_estimators = 10, and random_state = 0.

Results and discussion

Confusion matrix and classification reports are generated to evaluate the model performance. F_1 score, accuracy, precision, and recall are generated and evaluated. Precision is the number of correctly classified positive samples (TP) divided by the sum of the TP and the number of samples labeled by the system as positive (precision = $TP / (TP + FP)$) (Bottenberg & Ward, 1963). Recall is the number of correctly classified positive samples (TP) divided by the number of positive samples in the data (recall = $TP / (TP + FN)$). F_1 score is the harmonic mean of precision and recall ($F_1 = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$). The result of the model's performance is shown in Table 2. Values of the classification metrics show that the RF model scores the highest accuracy (93%) whereas the highest accuracy of MLR model scores

83%. This result proved that the ensemble RF model is the best for prediction of the groundwater contamination.

More information about the accuracy of the model can be obtained from the confusion matrix (Hamilton 2012). The confusion matrix reports the numbers of (1) true positives (TP = the number of samples classified as true while they are true), (2) true negative (TN = the number of samples classified as false while they are false), (3) false positives (FP = the number of samples classified as true while they are false), and (4) false negatives (FN = the number of samples classified as false while they are true) (Bekkar et al. 2013).

For the MLR model, among 30 samples of actual data, 18 samples are classified as true positive, 7 samples as true negative, 5 samples as false positive, and no samples for false negative. For the RF model, among 30 samples of actual data, 16 samples are classified as true positive, 12 samples as true negative, no samples as false positive, and 2 samples for false negative. Figure 9a and b show the confusion matrix of the MLR and RF models, respectively.

Results of the analyses of the groundwater samples revealed that about 95% exceeds the maximum permissible ammonia (0.5 mg/l) according to the WHO (2011). The value of ammonia in groundwater samples ranges between 0.01 and 22.4 mg/l. The reason for this contamination is due to various human activities as well as the use of wastewater for irrigation in the east of the study area (Hagage et al. 2021). Lead, magnesium, iron, and zinc content in water depend on the amount of industrial waste, fertilizers, and sewage sludge (Oluyemi et al. 2008). Lead concentration in groundwater ranges between 0.004 and 0.05 mg/l. The

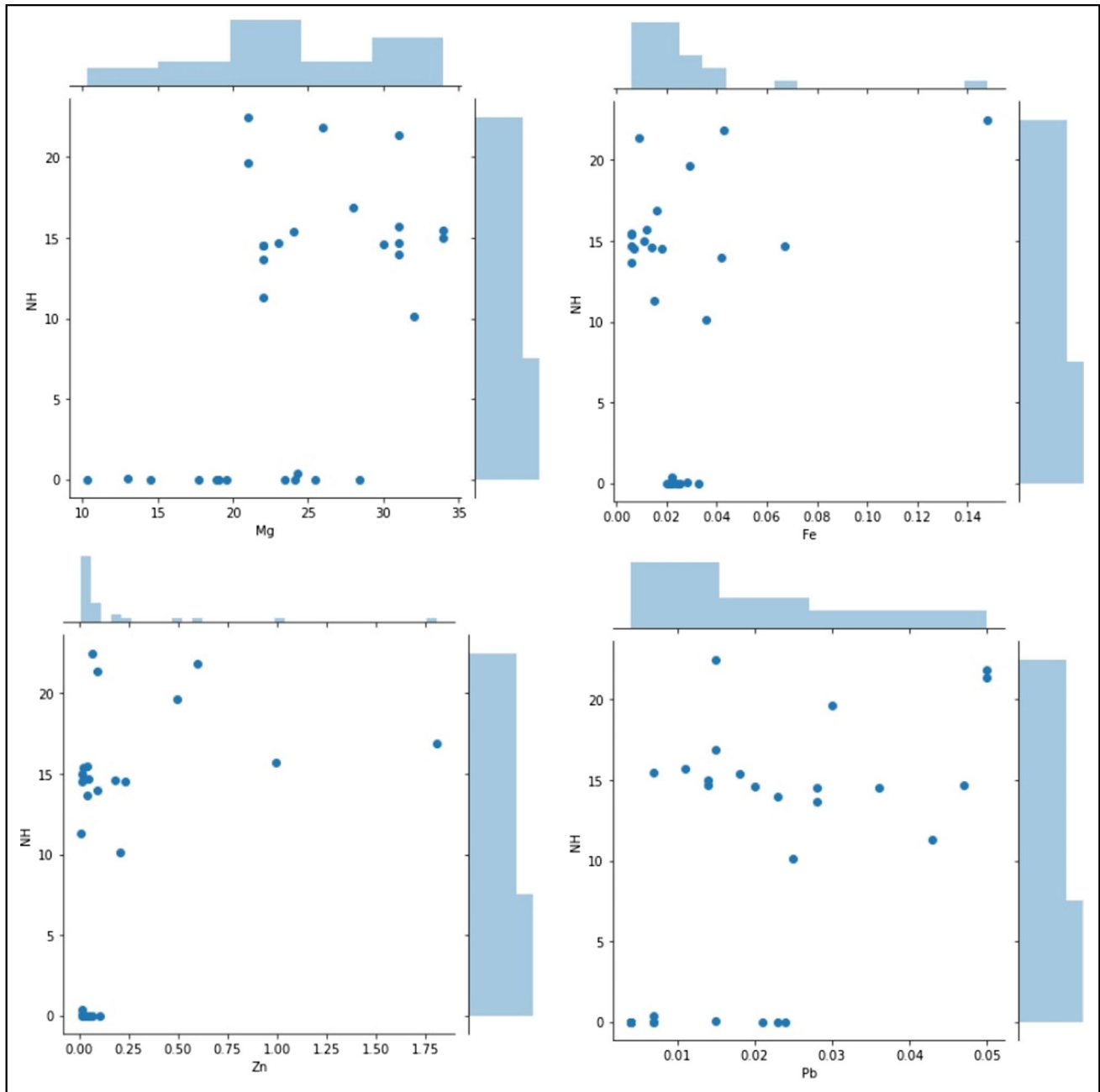


Fig. 8 Pair plot of the Mg, Fe, Pb, and Zn variables against the NH₃

Table 2 Classification metric results

Machine Learning Model	Model Parameters	Accuracy (%)	Precision	Recall	F1-score
MLR	solver="liblinear", C=10.0,random_state=0	70%	0.74	0.78	0.76
	solver="liblinear", C=3.0,random_state=0	83%	0.78	0.58	0.88
RF	criterion="entropy",n_estimators=10,random_state=0	93%	0.86	0.89	0.94

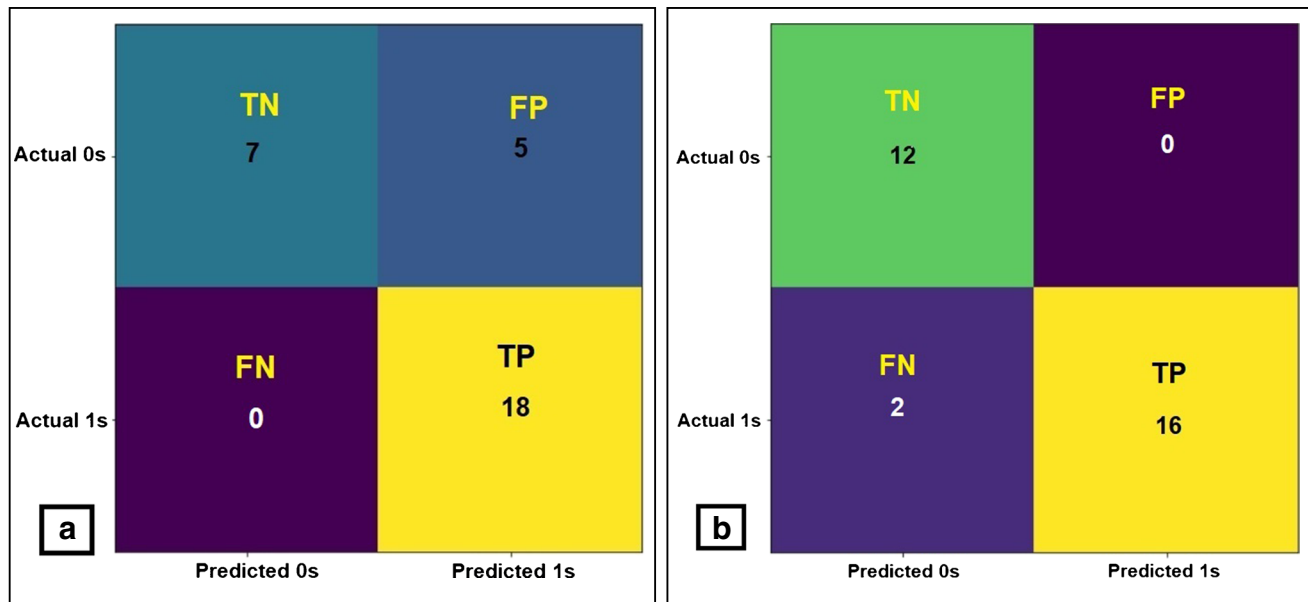


Fig. 9 Confusion matrix: **a** the MLR model and **b** the RF model

high pollution of lead in groundwater is due to industrial activities, urbanization, excessive use of chemical fertilizers, agricultural pesticides, and sewage leakage (Hagage et al. 2021; Krishna and Kurakalva 2014).

Conclusions

In general, groundwater contamination by ammonia is a significant issue in Sohag Governorate, Egypt and is attributed mainly to urban growth. The present study developed a predictive model for groundwater contamination using ensemble RF and MLR models. Results of the performance of these models are evaluated using classification metrics and confusion matrix. The study concluded the following:

- 1- Performance of the RF model is better than the MLR model. It scores high accuracy (93%) compared to the (83%) recorded by the MLR model.
- 2- A strong relation is observed between the urban expansion and the high ammonia concentration. Lack of sewerage networks in the study area forced the inhabitants to build sewage rooms and injection wells which leads to high contamination of the groundwater.
- 3- Akhmim and El Kola are highly polluted sites as demonstrated by the spatial distribution maps.
- 4- The study proved the usefulness of the ML models for predicting groundwater contamination using the ammonia index and its relevant variables.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

References

- Abdel Latif A, El Kashouty M (2010) Groundwater investigation in Awlad Salameh, Southern Sohag, Upper Egypt. *Earth Sci Res J* 14(1):63–75
- Abdel Moneim AA (1992) Numerical simulation and groundwater management of the Sohag aquifer. In: *The Nile Valley, Upper Egypt*. Ph.D. thesis. University of Strathclyde, Glasgow, Scotland
- Abdel Moneim AA (1999) Geoelectrical and hydrogeological investigations of the groundwater resources on the area to the west of the cultivated land at Sohag, Upper Egypt. *Egypt J Geol* 43(2):253–268
- Ahmed A (2009) Using generic and pesticide DRASTIC GIS-based models for vulnerability assessment of the Quaternary aquifer at Sohag, Egypt. *Hydrogeol J* 17:1203–1217

- Ahmed A, Ali M (2011) Hydrochemical evolution and variation of groundwater and its environmental impact at Sohag, Egypt. *Arab J Geosci* 4(3):339–352
- Aldhyani T, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artificial intelligence algorithms. *Appl Bionics Biomech* 2020:6659314, 12 pages. <https://doi.org/10.1155/2020/6659314>
- APHA (2005) Standard methods for the examination of water and wastewater, 21st edn. American Public Health Association, Washington, DC
- Awad MA, Nada AA, Hamza MS, Froehlich K (1995) Chemical and isotopic investigation of groundwater in Tahta region, Sohag-Egypt. *Environ Geochem Health* 17:147–153
- Balamurugan P, Kumar PS, Shankar K, Nagavinothini R, Pauline Selvaraj P (2020) A GIS-based evaluation of hydrochemical characterization of groundwater in hard rock region, South Tamil Nadu, India. *Arab J Geosci* 13:837
- Balamurugan P, Karuppannan S, Muniraj K (2021) Evaluation of drinking and irrigation suitability of groundwater with special emphasizing the health risk posed by nitrate contamination using nitrate pollution index (NPI) and human health risk assessment (HHRA). *Human Ecol Risk Assess: An Int J* 27:5
- Bekkar M, Akrouf Alitouche Taklit (2013) Imbalanced data learning approaches review. *Int J Data Min Knowl Manag Process (IJDKP)* 03(04):15–33
- Böhlke JK, Smith RL, Miller DN (2006) Ammonium transport and reaction in contaminated groundwater: application of isotope tracers and isotope fractionation studies. *Water Resour Res* 42:W05411. <https://doi.org/10.1029/2005WR004349>
- Bottenberg RA, Ward JH (1963) Applied multiple linear regression, PRL-TDR-63-6, AD-413 128, lackland AFB, TX: 6570 Personnel Research Laboratory, Aerospace Medical Division
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman & Hall/CRC, London
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Cracknell MJ (2014) Machine learning for geological mapping: algorithms and applications. Ph.D. University of Tasmania, Tasmania
- EGSMA (1983) Geological map of Egypt (1:250000)
- Elbeih SF, El-Zeiny AM (2018) Qualitative assessment of groundwater quality based on land use spectral retrieved indices: case study Sohag Governorate, Egypt. *Remote Sens Appl: Soc Environ* 10:82–92
- Gedamy YR (2015) Hydrochemical characteristics and pollution potential of groundwater in the reclaimed lands at the desert fringes, West of Sohag Governorate – Egypt. *Curr Sci Int* 4:288–312
- Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. *Pattern Recognit Lett* 27 (4):294–300
- Hagage M (2021) Impacts of anthropogenic activities on the deterioration of groundwater and archaeological sites in Akhmim area, Sohag Governorate, Egypt: remote sensing and GIS applications, M.Sc. Cairo University Egypt, Cairo
- Hagage M, Madani A, Elbeih S, Faid A, El-Kammar A (2021) Groundwater quality and its suitability for drinking and irrigation in Akhmim District, Sohag Governorate, Egypt. *Ann Geol Survey* 38(7):118–133
- Hamilton H (2012). Confusion matrix. *Knowledge Discovery in Databases*
- Hosseini FS, Malekian A, Choubin B, Rahmati O, Cipullo S, Coulon F, Pradhan BA (2018) Novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Sci Total Environ* 644:954–962
- Ismaila E, El-Rawyba M (2018) Assessment of groundwater quality in West Sohag, Egypt. *Desalin Water Treatment* 5:1–8
- Issawi B, Hinnawi M (1980) Contribution to the geology of the plain west of the Nile between Aswan and Kom Ombo. In: Close AE (ed) *Loaves and fishes*. Southern Methodist University Press, Texas, pp 311–330
- Issawi B, Hassan MW, Osman R (1978) Geological studies in the area of Kom Ombo, Eastern Desert, Egypt. *Ann Geol Survey* 8:187–235
- Karavoltos S, Sakellari A, Mihopoulos N, Dassenakis M, Scoullou MJ (2008) Evaluation of the quality of drinking water in regions of Greece. *Desalination* 224(1-3):317–329
- Krishna K, Kurakalva RM (2014) Risk assessment of heavy metals and their source distribution in waters of a contaminated industrial site. *Environ Sci Pollut Res* 21:3653–3669
- Lindenbaum J (2012) Identification of sources of ammonium in groundwater using stable nitrogen and boron isotopes in Nam Du, Hanoi. M.Sc
- Madani AA, Niyazi B (2015) Groundwater potential mapping using remote sensing techniques and weights of evidence GIS model: a case study from Wadi Yalamlam basin, Makkah Province, Western Saudi Arabia. *Environ Earth Sci* 74(6):5129–5142
- Mair A, El-Kadi AI (2013) Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. *J Contam Hydrol* 153:1–23
- Melegy AA, Shaban AM, Hassaan MM, Salman S (2014) Geochemical mobilization of some heavy metals in water resources and their impact on human health in Sohag Governorate, Egypt. *Arab J Geosci* 7:4541–4552
- Muharema F, Logofătua D, Leonb F (2019) Machine learning approaches for anomaly detection of water quality on a real-world data set. *J Inform Telecommun* 3(3):294–307. <https://doi.org/10.1080/24751839.2019.1565653>
- Nafouanti MB, Li J, Mustapha NA, Uwamungu P, Dalal AA (2021) Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: comparison of random forest, logistic regression and artificial neural network. *Appl Geochem* 132:105054
- Oluyemi EA, Feuyit G, Oyekunle JA, Ogunfowokan AO (2008) Seasonal variations in heavy metal concentrations in soil and some selected crops at a landfill in Nigeria. *Afr J Environ Sci Technol* 2(5):89–96
- Omer AA (1996) Geological, mineralogical and geochemical studies on the Neogene and Quaternary Nile Basin deposits, Qena-Assiut Stretch, Egypt. PhD Thesis. South Valley University, Qena
- Pham BT, Prakash I (2019) A novel hybrid model of Bagging-based Naïve Bayes Trees for landslide susceptibility assessment. *Bull Eng Geol Environ* 78(3):1911–1925
- Rizeei HM, Azeez OS, Pradhan B, Khamees HH (2018) Assessment of groundwater nitrate contamination hazard in a semi-arid region by using integrated parametric IPNOA and data-driven logistic regression models. *Environ Monit Assess* 190:633
- Russell S, Norvig P (2010) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall, Upper Saddle River
- Said R (1960) Planktonic foraminifera from the Thebes Formation, Luxor, Egypt. *Micropaleontology* 16:227–286
- Said R (1981) The geological evaluation of the River Nile. Springer-Verlag, New York
- Said R (1990) The geology of Egypt. A.A, Balkema, Rotterdam/Brookfield
- Solanki A, Agrawal H, Khare K (2015) Predictive analysis of water quality parameters using deep learning. *Int J Comput Appl* 125:0975–8887
- Venkataraman K, Uddameri V (2012) Modeling simultaneous exceedance of drinking-water standards of arsenic and nitrate in the Southern Ogallala aquifer using multinomial logistic regression. *J Hydrol* 458:16–27
- Vijay S, Kamaraj K (2019a) Ground water quality prediction using machine learning algorithms in R. *Int J Res Anal Rev* 743:6
- Vijay S, Kamaraj K (2019b) A novel approach on various machine learning algorithms for predicting ground water quality. *JETIR* 6(4):37–40

- Wang X, Zhang F, Jianli Ding J (2017) Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Sci Rep* 7:12858. <https://doi.org/10.1038/s41598-017-12853-y>
- WHO (2011) Guidelines for drinking-water quality, vol 564, 4th edn. World Health Organization, Geneva
- Youssef M, Abdel Moneim AA (2006) Evaluation of the geoenvironmental hazards in relation to the future development using the geographical information systems, East Sohag Governorate. The third international conference for development and the environment in the Arab world Assiut University, Assiut, pp 673–692
- Youssef AM, Omer AA, Ibrahim MS, Ali MH, Cawfield JD (2011) Geotechnical investigation of sewage wastewater disposal sites and use of GIS land use maps to assess environmental hazards: Sohag, Upper Egypt. *Arab J Geosci* 4:719–733