ORIGINAL PAPER

A machine learning approach to dental fluorosis classification

Aysegul Demir Yetis[1] · Mehmet Irfan Yesilnacar[2] · Musa Atas[3]

Abstract

Fluoride in groundwater has been found to pose a severe public health threat in two villages (Karataş and Sarım) of western Sanliurfa in the southeastern Anatolia region of Turkey, where many cases of fluorosis, which detrimentally affects the teeth and bones, have been reported. Analysis of fluoride in drinking water is usually accomplished using various chemical methods, but while these techniques produce accurate and reliable results, they are expensive, labor-intensive, and cumbersome. In this study, a more cost-effective alternative, based on machine learning methods, is introduced. In this case, artificial neural network (ANN), support vector machine (SVM), and Naïve Bayes classifiers are utilized. Furthermore, a novel feature selection and ranking method known as Normalized Weighted Voting Map (NWVM) is presented. In Fisher discrimination power (FDP) scores, X-ray fluorescence (XRF) variables have higher discrimination power potential than X-Ray diffraction (XRD) attributes, the most salient feature being Zr (0.464) and CaO (219.993) from XRD and XRF, respectively. When the XRD and XRF parameters are classified separately for the effect of NWVM ranking scores on the fluoride values and dental fluoride in groundwater, CaO, SiO_2, MgO, Fe_2O_3, P_2O_5, and K_2O (for XRF) and Quartz and Zr (for XRD) present a stronger effect. In addition, when looking at the effects among themselves, the first order is the same XRF parameters and then the XRD parameters. Experiments revealed that XRF constituents including CaO, SiO_2, MgO, P_2O_5, and K_2O have higher class discrimination power than the XRD variables.

Keywords Dental fluorosis · XRD and XRF · Machine learning · Feature selection · Feature ranking

Introduction

Fluorine is the most electronegative of all chemical elements and is therefore never found in nature in its elemental form. Aggregated chemically in the form of fluorides, it ranks 17th in abundance of elements that form the earth's crust,

Responsible Editor: Amjad Kallel

✉ Aysegul Demir Yetis
guldem83@gmail.com

Mehmet Irfan Yesilnacar
iyesilnacar@gmail.com

Musa Atas
hakmesyo@gmail.com

[1] Medical Services and Techniques Department, Bitlis Eren University, Bitlis, Turkey

[2] Environmental Engineering Department, Harran University, Sanliurfa, Turkey

[3] Computer Engineering Department, Siirt University, Siirt, Turkey

representing about 0.06–0.09% of the earth's crust (Tebutt 1983; WHO 1994). Florspar, cryolite, fluorapatite, mica, and hornblende are the richest minerals. Volcanic rocks, mica minerals (sirolite, fluorite, fluorapatite), and thermal sources cause high fluoride concentrations in natural waters (Brindha and Elango 2011; Yesilnacar et al. 2013, 2016).

The fluoride level in surface waters is generally less than 1 mg/L. In deep groundwaters or in hot spring waters in contact with fluoride-rich minerals, this amount can be up to 20–53 mg/L (WHO 1994; Selinus 2005; Msonda et al. 2007; Yesilnacar et al. 2013, 2016; Yetis et al. 2019).

Fluoride occurs in rocks, soil, air, water, plants, and animals as well as in the human body. While a low fluoride intake by humans can benefit teeth (e.g., caries prevention) and bone growth, long-term excessive absorption of fluoride can lead to fluorosis of teeth and bones, and to a multitude of other health-related problems including adverse impacts on the intellectual development of children (Harvard 2015). High fluoride levels in humans result from exposure to countless sources, such as fluoride-emitting industries, coal burning, volcanic ash, excessive fluoride in brick tea (tea compressed into block form), and fluoride-contaminated drinking water. Fluorosis is

becoming a global environmental toxicological problem in a number of parts of the world, and is most commonly found in water-stressed regions (Chen et al. 2012).

Over 200 million people worldwide use drinking water that exceeds the optimal fluoride level recommended by the WHO (2006). Of this figure, 70 million are in India, 45 million are in China, and about five million are in Mexico (UNICEF 1999; Yesilnacar et al. 2016). In India, an estimated 62 million people, of whom six million are children, have dental fluorosis because of consuming fluoride-contaminated water (Brindha and Elango 2011; Gowrisankar et al. 2017). Known fluoride belts on land include one that stretches from Syria through Jordan, Egypt, Libya, Algeria, Sudan, and Kenya, and another that stretches from Turkey through Iraq, Iran, Afghanistan, India, Bengal, northern Thailand, and China. There are also similar belts in the Americas and Japan. Fluorosis has been reported in each of these areas (Selinus 2005; Dar et al. 2012; Yesilnacar et al. 2016; Thapa et al. 2018).

Fluoride found in natural drinking water is the largest source of fluorine entering the human body. Endemic fluorosis is a major public health problem seen in individuals living within geographical areas with a higher fluoride concentration than the daily optimal dose. Fluoride is taken into the human body generally by food and water consumption. Studies show that soluble fluoride in drinking water is an important source, with potable water accounting for the majority of daily fluoride intake for humans (Yesilnacar et al. 2016).

The U.S. Health and Human Services (US HHS) recommends reducing the current limit range of 0.7–1.2 mg of fluoride per liter of water to 0.7 mg (HHS 2020). In the determination of fluoride concentration in drinking water, certain chemical analytical instruments including ion chromatography, spectrophotometry, ion-selective electrode, and inductively coupled plasma are utilized. In recent years, machine learning methods have also been widely used in medical research (Groznik et al. 2013; Nakano et al. 2014; Pérez et al. 2015). Although the aforementioned techniques provide accurate and quantitative results, they are, in general, expensive, labor-intensive, and cumbersome. It is absolutely necessary to use chemical measurements or analyses to use the cost-effectiveness of machine learning. However, due to improvements in machine learning techniques, there is a reduction in these chemical analytical procedures. Thus, a cost-effectiveness can be mentioned. Therefore, the detection of fluoride in a rapid and cost-effective manner is essential, particularly for researchers, as well as for technical and medical staff.

Machine learning (ML) is a term applied for developing machines that are capable of approaching tasks in a way similar to the human brains, such as speech recognition, self-driving cars, and interpreting complex data (Sanuade et al. 2020). ANN is an information processing paradigm inspired by biological nervous systems, like our brain (Yang et al. 2015). ANN is a technique used as in the problem-solving

process of the human brain (Khashei-Siuki and Sarbazi 2015). SVM is a classification based on statistical learning theory and this method is based on two main concepts. It creates a hyper plane that best separates the two classes and maximizes the distance between them (Gasmi et al. 2016). Naïve Bayes classifier is a process that estimates the probability of a new observation belonging to a predefined category, using a probability model defined according to the theory of Bayes (Tsangaratos and Ilia 2016).

In this context, the soft computing methods of ANN, support vector machine (SVM), and Naïve Bayes algorithms have been adopted for the current study. To the best of the researchers' knowledge, very few studies (Sirisha et al. 2008; Asghari Moghaddam et al. 2010; Amini et al. 2010; Dar et al. 2012; Nadiri et al. 2013; Kumari and Pathak 2015; Barzegar et al. 2017; Charulatha et al. 2017: Kheradpisheh et al. 2018) have been conducted on ANN-based prediction of fluoride in groundwater.

High fluoride content in groundwater as a cause of dental fluorosis disease being noted in Karataş and Sarım villages, in western Şanlıurfa, Southeastern Anatolia, Turkey, was first reported by Yesilnacar et al. (2011). In their study, the geological structure, hydrogeology, and hydrobiology of these villages and their surrounding area were investigated. The acquired data enabled a comparison of the dentition of the regional population (Yesilnacar et al. 2013, 2016). However, there have been no studies estimating dental fluorosis occurrence resulting from high-fluoride groundwater using ANN and rock data obtained through XRD and XRF measurements.

In the remainder of this paper, the "Materials and methods" section provides detailed information about the area studied, the preparation, data handling, and data modeling. Outcomes of a proposed feature ranking and classification models are discussed in the "Results and discussion" section. Concluding remarks and future projections are put forward within the final section.

Materials and methods

Study area

The region between the Suruc and Bozova districts in Sanliurfa province, which is situated in Turkey's southeast Anatolia region, was selected as the study area. It covers an area of 33 km by 41 km and has the typical continental climate type of the Southeast Anatolian region. The long-term average annual precipitation is 432.50 mm, with an average temperature of 18.40 °C and an average relative humidity of 51.60%. The digital elevation model (DEM) map showing the sampling points of the study area is given in Fig. 1. The land is mostly stony, and the most common land use is dry field agriculture. Cereal and pistachio crops are the dominant agricultural production, in addition to some livestock. Drinking
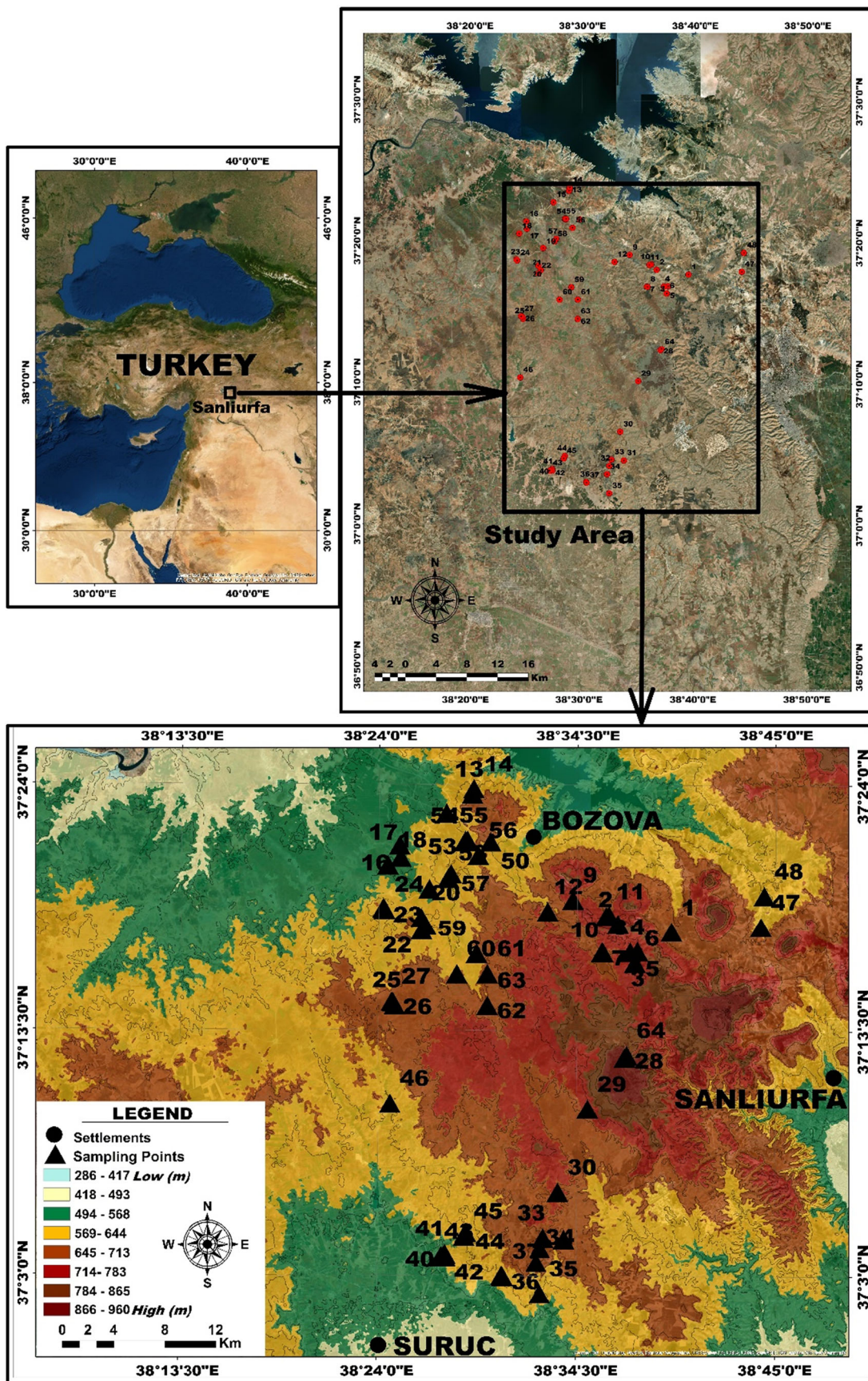
Fig. 1 DEM map showing study area sample points

water in the region is provided by groundwater. Even though there is an elementary school at each settlement (village), students are usually bussed to larger settlements for their education. The settlement populations range from 65 to 894 inhabitants (Yesilnacar et al. 2016).

There are two different types of rocks: sedimentary and volcanic rocks in the study area. Pleistocene basalts were formed during the Karacadag intrusion as a result of the magma rising to the surface through fractures and cracks with chimneys. This unit on the surface and close to the surface does not have an aquifer feature in this region. In the Siverek Karacadag region, however, basalts have been used as spring water for years. Although the Pliocene formation does not have an aquifer feature, the water quality is very low in the sections that supply very little water. Oligocene - Lower Miocene–aged formation has no aquifer feature; it can be defined hydrogeologically as aquitard. The Eocene-aged formation is generally composed of crystallized limestone. Since karstic structures are well developed, it is an important aquifer in the region. Eocene-aged formation is used in the wells drilled for drinking water in the region.

Fieldwork and data collection

The locations and routes of the study's sampling points were defined according to 1:25000 scale topographic and digital geological maps. A total of 63 village/sampling points were determined for fluoride level testing. Groundwater samples were taken seasonally for a period of 1 year. GPS was used for the coordinates of the sampling points. A portable Hach-Lange HQ40d multi-measurement device was used (in situ) to measure the fluoride concentration. Fluoride analyses of the groundwater samples were carried out using the procedures described in the EPA Method 340.2 (EPA 1993), APHA method 4500- F (APHA 1998), and ASTM D1179 - 99 (ASTM 1999) (Yesilnacar et al. 2016).

Fluorosis classifications and dental examination were undertaken by local dentists practicing within the study area. Exploration drilling was performed at two villages (Sarım and Karataş) to search for the geological structure. Samples were taken every 2 m and from 200 m deep wells in the drilling stage. The preserving and transporting of drilling samples were conducted using the procedure described in ASTM D5079 - 08 (ASTM 2008) (Yesilnacar et al. 2016).

X-ray diffraction (XRD) and X-ray fluorescence (XRF) modalities were analyzed on a total of 108 samples in order to determine mineralogical compositions and chemical constituents, with 62 samples from Karataş village and 46 from Sarım. The mineral composition according to depth is illustrated in Fig. 2. The chemical substances obtained from the XRF analysis were SiO_2, Al_2O_3, Fe_2O_3, MgO, CaO, K_2O, TiO_2, P_2O_5, and LOI (loss on ignition). Similarly, Calcite,

Quartz, Cu, Ni, S, Sr, Zn, and Zr substances were determined from the XRD analysis. It should be noted that in the XRD analysis, Cu, Ni, S, Sr, Zn, and Zr parameters are represented in ppm scale and the remaining are in percentage form. Some basic statistical information of these compounds and substances for the XRD and XRF modalities are listed in Table 1.

Modeling

As previously mentioned, Table 1 indicates a total of nine parameters for XRF and eight for the XRD analysis. These variables were handled as the input feature vector for the proposed simulation models. Likewise, for specifying the output variable, two aspects were considered. The first aspect is the utilization of fluoride concentrations of 2.17 and 2.63 values as an output. The second aspect is related to assigning the fluoride concentrations of 2.17 and 2.63 to dental fluorosis classes (DFC) 4 and 5, which were determined by field experts and then using those classes as the output value. Whether using fluoride concentrations or DFC as an output variable, the prediction performance of the learning algorithm actually has no effect as the problem being studied inherently has two class problems (binary classification). Therefore, DFC was selected as the output of the proposed models for the sake of simplicity.

As Table 2 depicts, in order to evaluate the classification performance of each modality (XRF and XRD) along with the input variables, various learning models were developed. Model-1 and Model-2 are based on XRF and XRD variables, respectively. Model-3, on the other hand, is a mixture of both XRF and XRD constituents. It should be noted that Model-1, Model-2, and Model-3 assess all of the variables. Conversely, Model-4, Model-5, and Model-6 try to seek the best variables from the variable pool. So, Model-4 and Model-5 were developed based on the best feature subset of XRF and XRD variables, respectively, via a global (exhaustive) search. Running a global exhaustive search on Model-6 (XRF+XRD) was not deemed to be feasible due to higher computational cost. As a result, simulated annealing (SA) meta-heuristic search, also known as suboptimal search, was applied (Kirkpatrick et al. 1983) to the Model-6 dataset.

In the current study, another significant contribution is to determine the influence of each input variable in order to rank the variables according to their saliencies. Various variable ranking methods and approaches exist in the literature. In this respect, variables can be treated individually or collaboratively (jointly) with other variables. Traditional feature selection schemes, including Fisher discrimination power (FDP), best first, forward selection and backward elimination, independent component analyses, exhaustive and genetic, search on a single variable, i.e., they evaluate the saliency of the feature/variable individually in a greedy
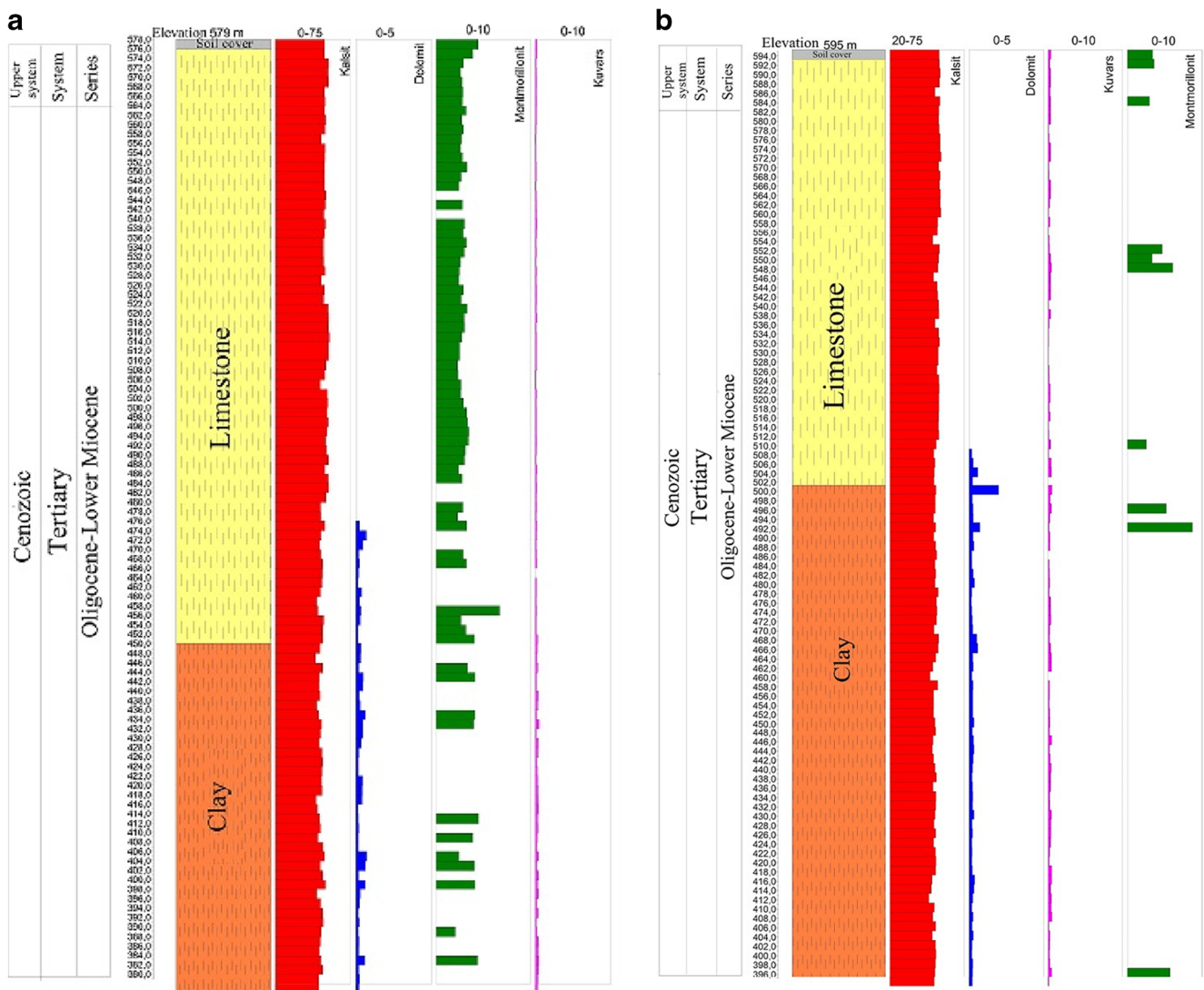
Fig. 2 Mineral composition variation of Karataş (**a**) and Sarım (**b**) villages according to depth (Yesilnacar et al. 2016)

fashion (Blachnik 2009). Likewise, global combinatorial searches such as exhaustive, genetic, and information theoretic approaches aim to find groups of variables that have better correlation to the output. As Guyon and Elisseeff (2003) indicated, there may exist circumstances where the ranking of input variables one by one cannot actually highlight the discrimination power or weakness of the attribute/variable. Figure 3 illustrates this significant inference from their work.

Note that for binary classification, both of the dimensions (X and Y axes) exhibit relatively weaker classification potential based on the distribution projections on the individual axes, because class distributions projected on the axes overlap. On the other hand, by taking these two features as a complementary group and projecting them onto the diagonal line, good class separation (see lower left portion) can be achieved. Furthermore, as mentioned by Ataş et al. (2012), although the

FDP and correlation coefficient of each variable with respect to target class can provide preliminary intuition about feature saliency, evaluating complementary features in the feature space can also provide significant information. For feature-ranking issues, the feature saliency score should be extracted from the list of classification performances of feature subsets. Therefore, in the current study, Normalized Weighted Voting Map (NWVM) was used, which can be considered an extension to the generic vote map algorithm used by Ataş et al. (2012).

Three different machine learning algorithms were selected as learning models. In order to investigate linearly separable characteristics of the problem, Naïve Bayes was used as a simple classifier. However, for handling the non-linearity issue, relatively complex models such as single hidden layered feed forward multi-layer perceptron (SHFF-MLP) and support vector machine (SVM)

Table 1 Basic statistical information for the variables of XRD and XRF modalities

Modality	Variable	Min	Max	Median	Mean	Std Dev.
XRF	SiO_2	0.00	15.42	2.56	2.24	1.91
	Al_2O_3	0.00	5.19	0.66	0.63	0.52
	Fe_2O_3	0.00	2.21	0.29	0.26	0.22
	MgO	0.00	2.03	0.54	0.48	0.40
	CaO	0.46	54.47	47.82	29.20	24.43
	K_2O	0.00	0.47	0.04	0.06	0.05
	TiO_2	0.00	0.31	0.03	0.03	0.03
	P_2O_5	0.00	0.17	0.04	0.05	0.04
	LOI	36.03	45.77	42.95	42.81	0.93
XRD	Calcite	28.64	83.79	57.31	57.12	8.93
	Quartz	0.01	4.30	1.72	1.88	0.73
	Cu	17.00	42.00	25.35	25.36	2.91
	Ni	5.00	54.00	31.00	31.26	4.74
	S	22.00	1681.00	108.50	200.94	168.17
	Sr	211.00	1267.00	418.00	509.07	208.34
	Zn	14.00	38.00	22.26	22.52	2.96
	Zr	7.00	95.00	64.00	59.02	17.89

classifiers were employed. All free parameters of the classifiers were set to default values and no tuning or optimization processes were carried out.

Results and discussion

Classification results

XRD and XRF features were utilized as the input parameters of the proposed models, as shown in Table 2. Field experts specified two distinct fluorosis levels based on fluoride content in the drinking water obtained from the Karataş and Sarım wells. Those levels constituted the

class labels, or output, of the proposed models. The models were evaluated by the Naïve Bayes as a linear, and ANN and SVM as non-linear classifiers with 10-fold cross-validation technique. Weka machine learning software written in Java was used for the learning algorithms (Hall et al. 2009). Classification performances of the proposed models are shown in Table 3.

The highest classification score is highlighted in italic type for each model in Table 3. Model-1, Model-2, and Model-3 basically evaluate the classification performance of the parameters as a whole. In contrast, Model-4, Model-5, and Model-6 concentrate on the classification performance of the combinatorial subset of parameters based on exhaustive search and simulated annealing suboptimal search schemes, respectively. In the subset search process, sometimes subsets having different numbers of features can yield the same results. In this case, the subset containing fewer features was preferred, which also accomplishes dimensionality reduction (admitted in machine learning community). For example, the CaO parameter was chosen as a single input feature because it could classify the fluoride problem solely by itself. These outcomes are also in accordance with Figs. 4 and 5.

Table 3 reveals that ANN outperforms SVM and Naïve Bayes classifiers in terms of the correctly classified percentage amount, and also indicates that non-linear classifiers best fit the current problem. This outcome is also in accordance with previous studies from which researchers used ANN as a machine learning algorithm (Yesilnacar et al. 2008; Sirisha et al. 2008; Asghari Moghaddam et al. 2010; Amini et al. 2010; Dar et al. 2012; Nadiri et al. 2013; Kumari and Pathak 2015).

Feature influence analysis

In order to assess the influence of each feature, FDP and Normalized Weighted Voting Map (NWVM) methods were utilized and results were then compared. The NWVM method

Table 2 List of the proposed models

Model type	Input parameters	Output parameters
Model-1	X-ray fluorescence (XRF) variables	Dental fluorosis prediction
	SiO_2, Al_2O_3, Fe_2O_3, MgO, CaO, K_2O, TiO_2, P_2O_5, LOI	
Model-2	X-ray diffraction (XRD) variables	
	Calcite, Quartz, Cu, Ni, S, Sr, Zn, Zr	
Model-3	Model-1 + Model-2	
Model-4	Combinatorial Best Feature Subset of Model-1 (exhaustive search)	
Model-5	Combinatorial Best Feature Subset of Model-2 (exhaustive search)	
Model-6	Combinatorial Best Feature Subset of Model-3 (simulated annealing search)	

Fig. 3 Projection visualization of multivariate (2 variables [*X*, *Y*]) Gaussian distribution of binary classification. Projection on diagonal line exhibits relatively good class separation
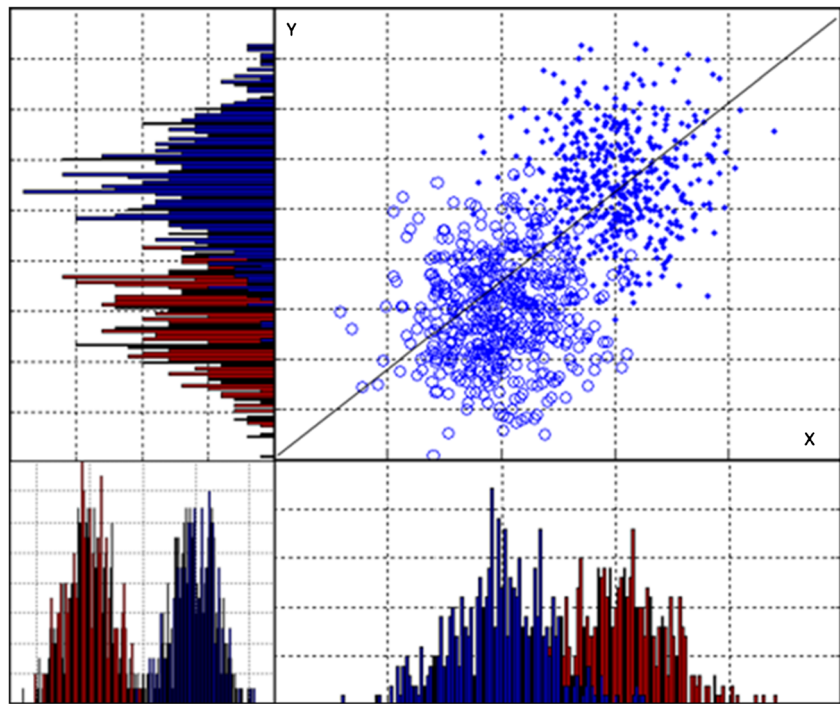

and other software modules were developed in Java using NetBeans 8.0 integrated development environment. As mentioned previously, FDP seeks class separation capability of each feature individually and yields limited class separation on feature subsets. Another significant difference between the FDP and NWVM method is that FDP relies on the filter method and therefore it is invariant to classifiers, whereas NWVM is based on the wrapper approach and is strongly related to the classification algorithm used. Fisher discriminant (FD) was first proposed by Fisher in 1936 (Fisher 1939), and tries to project data from *n*-dimensional space to a one-dimensional space (i.e., line or axis) from where between-class scatter is at its maximum and within-class scatter is at its minimum (Ataş et al. 2012). FD can be computed as:

$$FDP = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \tag{1}$$

where *FDP* is Fisher discrimination power, and μ and σ depict the mean and variance of each class, respectively. Figures 4 and 6 highlight the influence of each variable based on between-class scatter/distribution as a box-plot visualization and Fisher scores. In order to unify distribution ranges of parameters, *z*-score normalization was applied. It should be noted that in Fig. 6, the notches represent medians and edges of the boxes lie between the 25th and 75th percentiles. In addition, "+" pointers are used to indicate the outliers in the distribution. Class separation quality is determined by the distance between notches and its level of overlap. The highest gap between notches and the smallest overlap represents good class separation power of the feature. In Fig. 4, the features having relatively higher discrimination powers have higher FDP scores, and from this perspective, both Figs. 4 and 6 are in agreement. It can therefore be inferred from these figures that the XRF variables have higher discrimination power potential than the XRD attributes, and that the most salient feature is Zr and CaO from XRD and XRF, respectively.

Table 3 Classification performance of the proposed models

Models	Input parameters and applied search strategies	Correctly classified %		
		Naïve Bayes	ANN	SVM
Model-1	XRF parameters (whole)	99.07	100.00	100.00
Model-2	XRD parameters (whole)	76.85	77.78	77.22
Model-3	XRD+XRF parameters (whole)	99.07	100.00	100.00
Model-4	CaO (exhaustive search result)	100.00	100.00	100.00
Model-5	Quartz, Cu, Ni, Sr, Zn (exhaustive search result)	65.19	73.15	71.55
Model-6	CaO (simulated annealing search result)	100.00	100.00	100.00

a

b
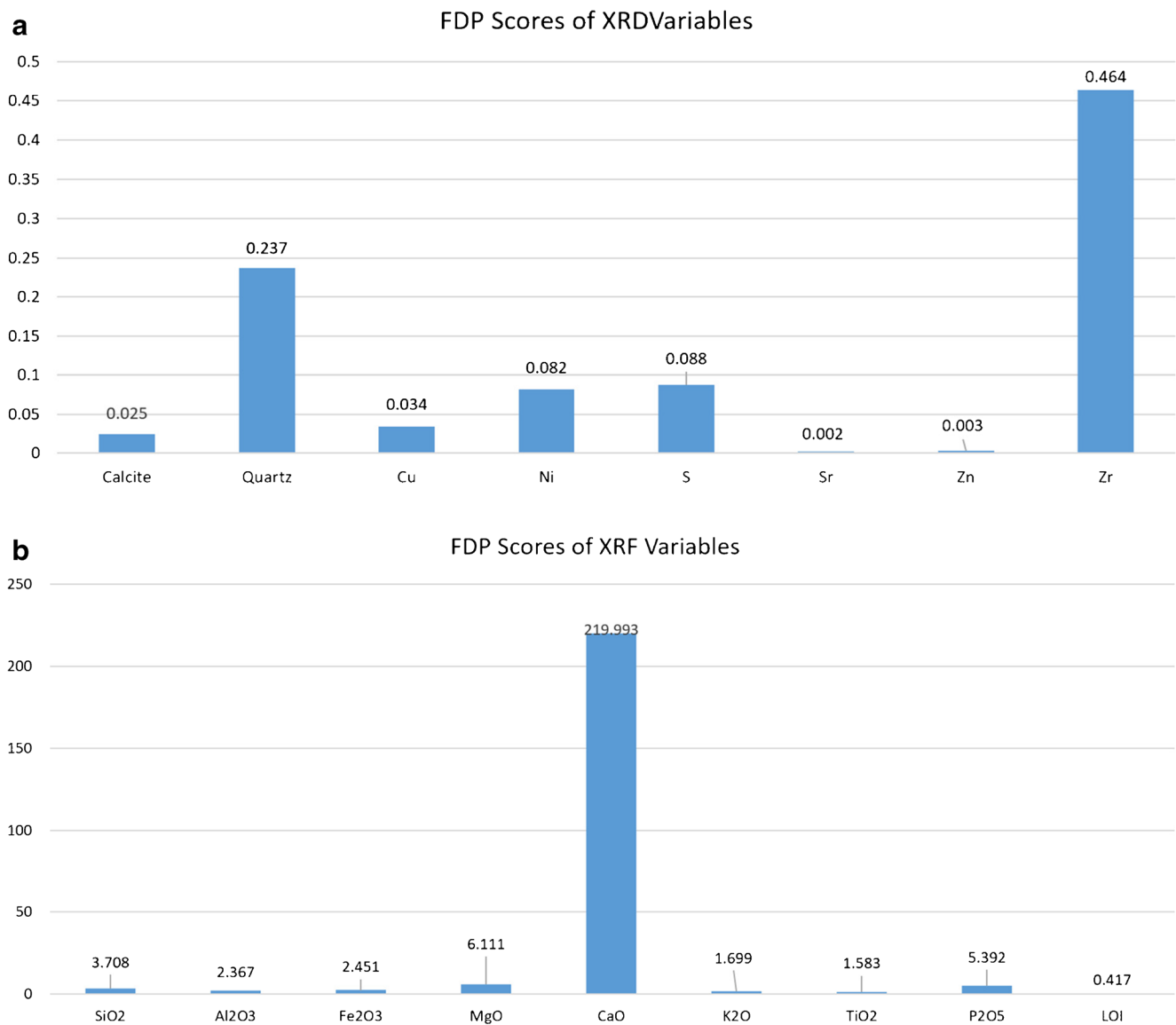
Fig. 4 FDP scores of XRD (**a**) and XRF (**b**) variables

Although the FDP approach provides good preliminary intuition about feature saliency/influence, it is inadequate in terms of feature complementarity. As Fig. 3 shows, sometimes weak features expose a stronger class separation power when they are combined. Thus, a new feature-ranking method was utilized, Normalized Weighted Voting Map (NWVM) based on the voting approach. The proposed NWVM method can be considered an extension of the classical voting map. That is, in the normal voting method, while a voting is made by calculating the contribution of each component equally, in the NVWM method, there is a weighted contribution for each component and the sum of the weights is one. The main idea of NWVM is demonstrated as shown in Fig. 7. If there are N number of variables, then theoretically there are 2^N-1 subsets available. Basically, the objective is to find the best subset (complementary feature group) within the subset pool.

Figure 7 indicates that all combinatorial groups of variables were evaluated, and that the classification accuracy of each trial was recorded in the dynamic list structure.

Since the contribution of each variable in the subset was unknown, equal scores of classification accuracy were assigned to each feature. Next, the scores gained for each variable were summed for all of the subsets. Then, weights of the variables were normalized to within the 20–100 range through min-max normalization, and finally the ranking procedure was performed. The researchers selected the 20–100 range because differences of influence score among the variables could be more easily visualized within that range. It should be noted that, as the XRD and XRF modalities each have only a small number of variables, NWVM computation was deemed to be straightforward due to its low computational cost. However, if these modalities were combined, as in the
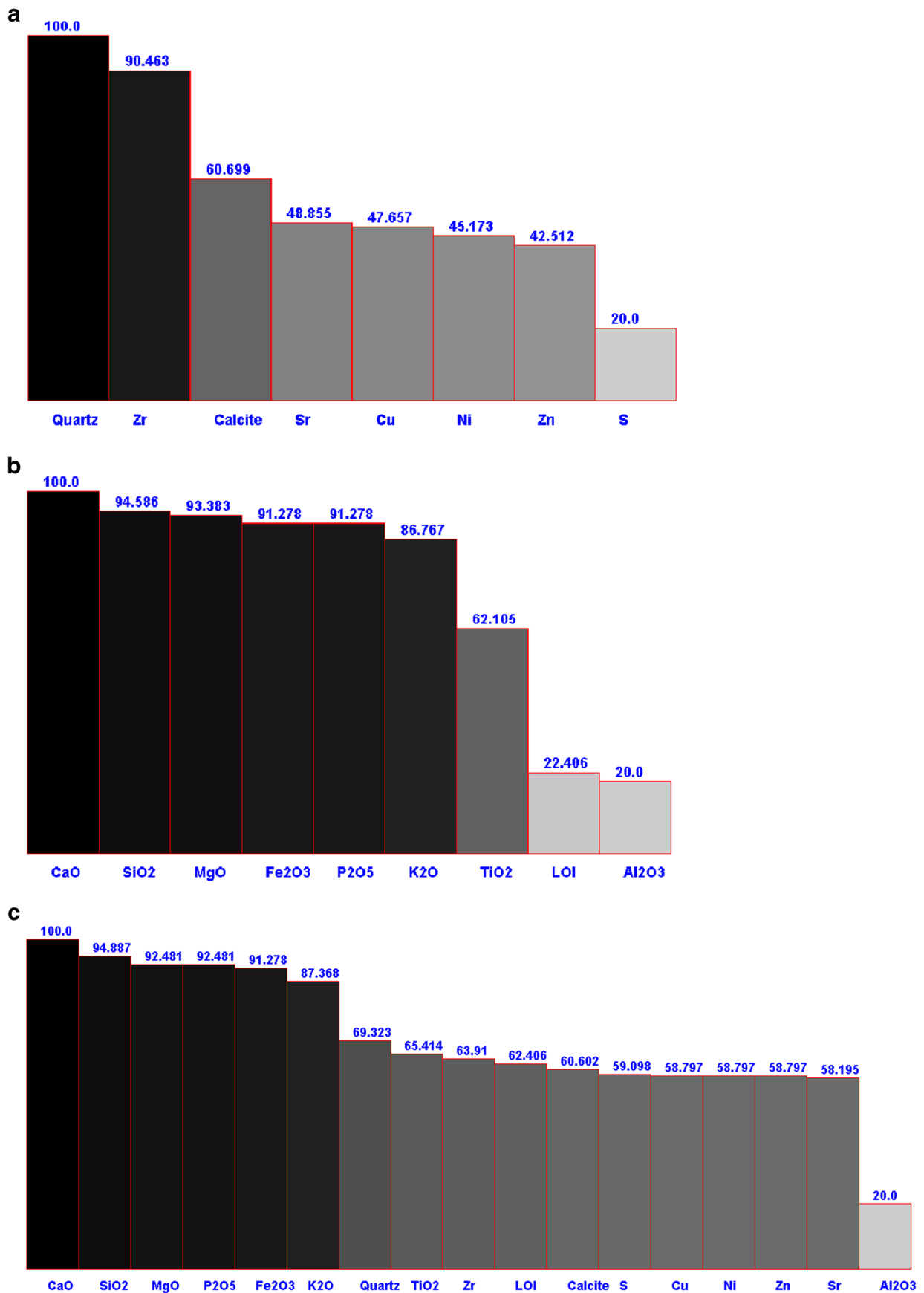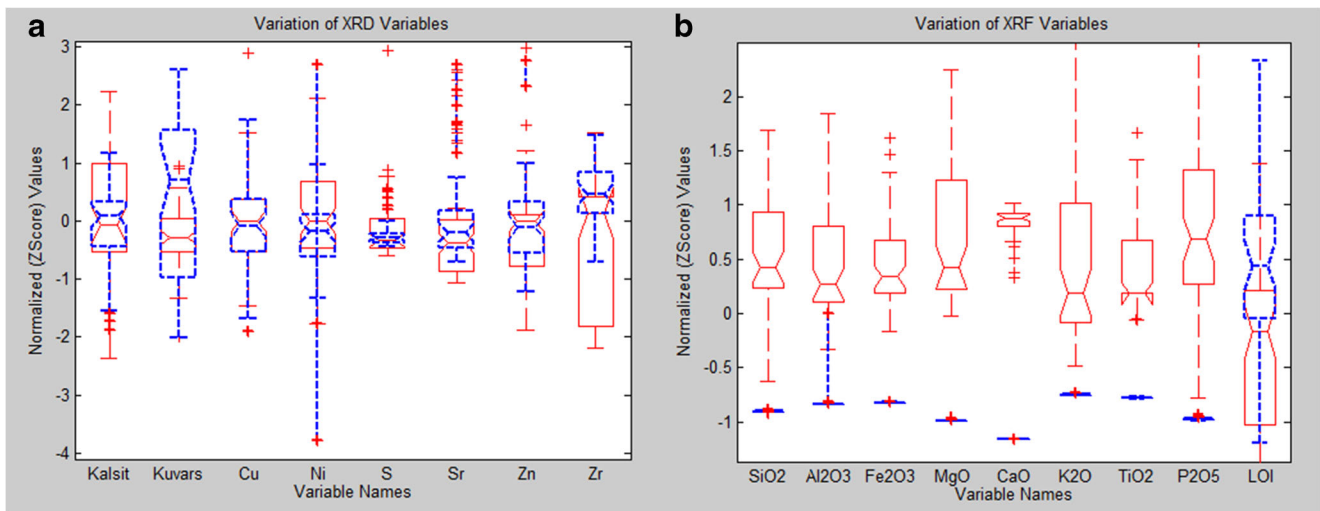
Fig. 5 NWVM-based ranked variables graph of XRD (**a**), XRF (**b**), and joint, (**c**) modalities

Fig. 6 Visualization of normalized distribution of XRD (**a**) and XRF (**b**) modalities as boxplots

case of Model-6, the computational cost would increase exponentially and only the suboptimal search would be feasible. Simulated annealing seeks to find suboptimal results for the relatively larger search space where global search is infeasible. Figure 5 illustrates the NWVM graph for (a) XRD, (b) XRF, and (c) a combination of both XRD and XRF.

The influence of the attribute can be monitored both by the score located on the top of the bars and the gray color gradients. Light color indicates a weak influence, and dark color a strong influence of the features. It should also be noted that NWVM ranking scores do not represent actual scores; rather they exhibit relative influences among the features.

Fig. 7 Schematic diagram for NWVM method
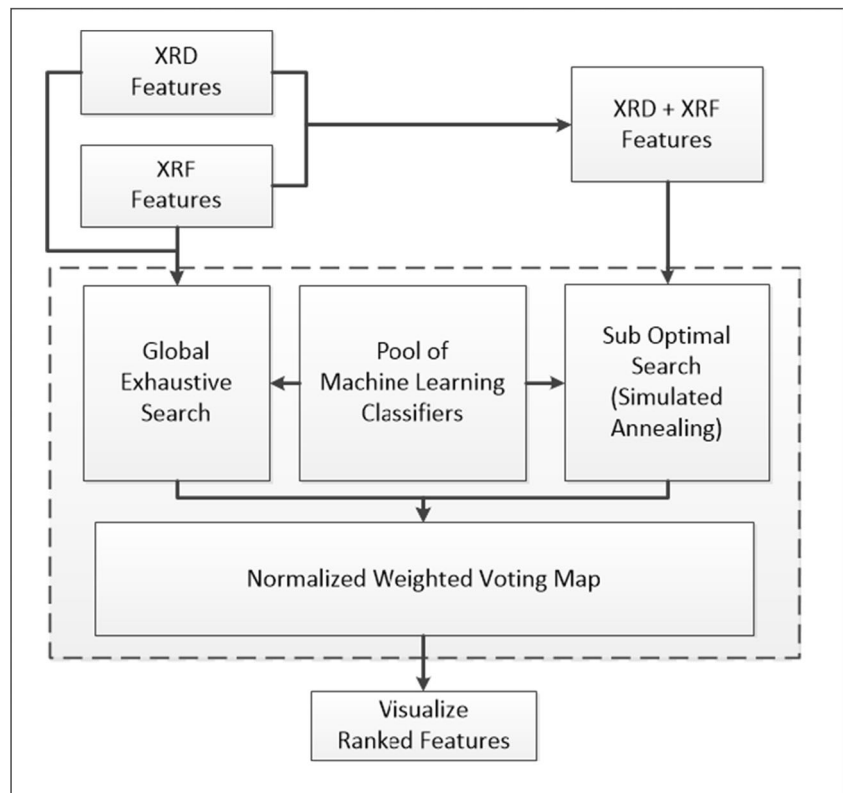
Conclusion

The interaction of rock chemistry on the occurrence of high fluoride concentrations in the groundwater, and as a result of this issue, dental and skeletal fluorosis cases, has been widely studied in many parts of the world (Barzegar et al. 2017; Sracek et al. 2015; Li et al. 2015; He et al. 2013; Fekri and Kasmaei 2013; Singaraja et al. 2013; Srinivasamoorthy et al. 2012; Naseem et al. 2010). However, there have been no studies estimating dental fluorosis occurrence resulting from high-fluoride groundwater using machine learning techniques and rock data obtained through XRD and XRF measurements.

Detection of fluoride content in groundwater is commonly conducted via chemical approaches. Although these have been shown to provide accurate, consistent, and reliable results, they are acknowledged to be labor-extensive, cumbersome, and costly processes. An alternative and more subtle approach is proposed in the current study. In this context, ANN, SVM, and Naïve Bayes classifiers were utilized and a novel feature selection and ranking method known as NWVM was presented. According to the results of the current study, in FDP scores, XRF variables have higher discrimination power potential than XRD attributes and most salient feature is Zr (0.464) and CaO (219.993) from XRD and XRF, respectively. When the XRD and XRF parameters are classified separately for the effect of NWVM ranking scores on the fluoride values and dental fluoride in groundwater, CaO, SiO_2, MgO, Fe_2O_3, P_2O_5, and K_2O (for XRF) and Quartz and Zr (for XRD) offer a stronger effect. In addition, when looking at the effects among themselves, the first order is the same XRF parameters and then the XRD parameters. Experiments reveal that X-ray fluorescence (XRF) constituents including CaO, SiO_2, MgO, P_2O_5, and K_2O have higher class discrimination power than the X-ray diffraction (XRD) variables.

The proposed NWVM feature-ranking method provides a simple yet powerful tool that assesses the influence of the variables, based on the complementarity/grouping potential, and therefore can be used as a generic ranking tool for analysis of multivariate problems. Influence score of Quartz in XRD modality based on FDP method was reported as 0.237 as the second salient variable. However, according to the NWVM approach, it is the most significant feature because it contributes to the classification performance of most feature subsets. Al_2O_3 is the least salient feature in the XRF parameters, considering both the NWVM and FDP methods, and as a result should be removed from the parameter list. Experimental results have shown that the classification of dental fluorosis stemming from groundwater and rocks by machine learning techniques is possible, and that the proposed NWVM method provided promising results for feature influence analysis.

References

Amini M, Abbaspour KC, Johnson CA (2010) A comparison of different rule - based statistical models for modeling geogenic groundwater contamination. Environ Model Softw 25:1650–1657

APHA, AWWA, WEF (1998) Standard methods for the examination of water and wastewater (20th edn), American Public Health Association, Washington, DC, USA

Asghari Moghaddam A, Nadiri AA, Fijani E (2010) Spatial prediction of fluoride concentration using artificial neural networks and geostatic models. Water and Soil Science (Agricultural Science) 19.1(2):129–145

ASTM International (ASTM D1179-99B) (1999) Standard test methods for fluoride ion in water. 100 Barr Harbor Drive, West Conshohocken, PA

ASTM International (ASTM D5079–08) (2008) Standard practices for preserving and transporting rock core samples, West Conshohocken, PA

Ataş M, Yardimci Y, Temizel A (2012) A new approach to aflatoxin detection in chili pepper by machine vision. Comput Electron Agric 87:129–141

Barzegar R, Moghaddam AA, Adamowski J, Fijani E (2017) Comparison of machine learning models for predicting fluoride contamination in groundwater. Stoch Environ Res Risk Assess 31:2705–2718

Blachnik, M (2009) Comparison of various feature selection methods in application to prototype best rules. In Computer recognition systems 3, vol. 57. Springer Berlin Heidelberg pp. 257–264

Brindha K, Elango L (2011) Fluoride in groundwater: causes, implications and mitigation measures. In: Monroy, S.D. (Ed.), Fluoride properties, applications and environmental management pp. 111-136

Charulatha G, Srinivasalu S, Uma Maheswari O, Venugopal T, Giridharan L (2017) Evaluation of ground water quality contaminants using linear regression and artificial neural network models. Arab J Geosci 10:128

Chen H, Yan M, Yang X, Chen Z, Wang G, Schmidt-Vogt D, Xu Y, Xu J (2012) Spatial distribution and temporal variation of high fluoride contents in groundwater and prevalence of fluorosis in humans in Yuanmou County, Southwest China. Journal of Hazardous Materials 15(235–236):201–209

Dar IA, Sankar K, Dar MA, Majumder M (2012) Fluoride contamination - artificial neural network modeling and inverse distance weighting approach. Journal of Water Science 25(2):165–182

EPA (1993) Environmental Sciences Section, Inorganic Chemistry Unit, Wisconsin State Lab of Hygiene. ESS Method 340.2: total suspended solids, mass balance (dried at 103–105 °C) volatile suspended solids (ignited at 550 °C). Wisconsin State Lab of Hygiene, Madison, WI, USA

Fekri M, Kasmaei LS (2013) Fluoride pollution in soils and waters of Koohbanan region, southeastern Iran. Arab J Geosci 6:157–161. https://doi.org/10.1007/s12517-011-0336-1

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7:179–188

Gasmi A, Gomez C, Zouari H, Masse A, Ducrot D (2016) PCA and SVM as geo-computational methods for geological mapping in the southern of Tunisia, using ASTER remote sensing data set. Arab J Geosci 9:753

Gowrisankar G, Jagadeshan G, Elango L (2017) Managed aquifer recharge by a check dam to improve the quality of fluoride-rich

groundwater: a case study from southern India. Environ Monit Assess 189: 200

Groznik V, Guid M, Sadikov A, Mozina M, Georgiev D, Kragelj V, Ribaric S, Pirtosek Z, Bratkoa I (2013) Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med 57:133–144

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update, vol. 11, SIGKDD Explorations; (1), pp 10–18

Harvard (2015) http://www.hsph.harvard.edu/news/features/fluoride-childrens-health-grandjean-choi/. Accessed 4 Feb 2015

He X, Ma T, Wang Y, Shan H, Deng Y (2013) Hydrogeochemistry of high fluoride groundwater in shallow aquifers, Hangjinhouqi, Hetao Plain. J Geochem Explor 135:63–70

HHS (2020) EPA ve U.S. Department of Health & Human Services (HHS). https://archive-it.org/organizations/745. Accessed 14 Dec 2020

Khashei-Siuki A, Sarbazi M (2015) Evaluation of ANFIS, ANN, and geostatistical models to spatial distribution of groundwater quality (case study: Mashhad plain in Iran). Arab J Geosci 8:903–912

Kheradpisheh Z, Mahvi HA, Mirzaei M, Mokhtari M, Azizi R, Fallahzadeh H, Ehrampoush MH (2018) Correlation between drinking water fluoride and TSH hormone by ANNs and ANFIS. Journal of Environmental Health Science and Engineering 16(1):11–18

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Kumari N, Pathak G (2015) Modelling fluoride pollution of groundwater using artificial neural network in the western parts of Jharkhand. World Academy of Science, Engineering and Technology, vol. 3, (2)

Li C, Gao X, Wang Y (2015) Hydrogeochemistry of high-fluoride groundwater at Yuncheng Basin, northern China. Sci Total Environ 508:155–165

Msonda KWM, Masamba WRL, Fabiano E (2007) A study of fluoride groundwater occurrence in Nathenje, Lilongwe, Malawi. Phys Chem Earth 32:1178–1184

Nadiri AA, Fijani E, Tsai FT-C, Asghar Moghaddam A (2013) Supervised committee machine with artificial intelligence for prediction of fluoride concentration. J Hydroinf 15(4):1474–1490

Nakano Y, Takeshita T, Kamio N, Shiota S, Shibata SN, Yoneda M, Hirofuji T, Yoshihisa Y (2014) Supervised machine learning-based classification of oral malodorbased on the microbiota in saliva samples. Artif Intell Med 60:97–101

Naseem S, Rafique T, Bashir E, Bhanger MI, Laghari A, Usmani TH (2010) Lithological influences on occurrence of high-fluoride groundwater in Nagar Parkar area, Thar Desert, Pakistan. Chemosphere 78:1313–1321

Pérez NP, López MAG, Silva A, Ramos I (2015) Improving the Mann–Whitney statistical test for feature selection: an approach in breast cancer diagnosis on mammography. Artif Intell Med 63:19–31

Sanuade OA, Hassan AM, Akanji AO, Olaojo AA, Oladunjoye MA, Abdulraheem A (2020) New empirical equation to estimate the soil moisture content based on thermal properties using machine learning techniques. Arab J Geosci 13:377

Selinus O (2005) Essentials of medical geology: impacts of the natural environment on public health, Elsevier Academic Press

Singaraja C, Chidambaram S, Anandhan P, Prasanna MV, Thivya C, Thilagavathi R (2013) A study on the status of fluoride ion in groundwater of coastal hard rock aquifers of south India. Arab J Geosci 6:4167–4177. https://doi.org/10.1007/s12517-012-0675-6

Sirisha P, Sravanti KN, Ramakrishna V (2008) Application of artificial neural networks for water quality prediction. Int J Sys Technol 1(2): 115–123

Sracek O, Wanke H, Ndakunda NN, Mihaljevič M, Buzek F (2015) Geochemistry and fluoride levels of geothermal springs in Namibia. J Geochem Explor 148:96–104

Srinivasamoorthy K, Vijayaraghavan K, Vasanthavigar M, Sarma S (2012) Assessment of groundwater quality with special emphasis on fluoride contamination in crystalline bed rock aquifers of Mettur region, Tamilnadu, India. Arab J Geosci 5:83–94. https://doi.org/10.1007/s12517-010-0162-x

Tebutt THY (1983) Principles of quality control. Pergamon, England, p 235

Thapa R, Gupta S, Gupta A, Reddy DV, Kaur H (2018) Geochemical and geostatistical appraisal of fluoride contamination: An insight into the Quaternary aquifer. Sci of The Total Environ 640–641:406–418

Tsangaratos P, Ilia I (2016) Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. Catena 145:164–179

UNICEF (1999) States of the art report on the extent of fluoride in drinking water and the resulting endemicity in India. Report by Fluorosis and Rural Development Foundation for UNICEF, New Delhi

World Health Organization (WHO) (1994) Fluorides and oral health, (technical report series no. 846), World Health Organization, Geneva

World Health Organization (WHO) (2006) Fluoride in drinking water, UK: IWA publishing, London, p 144

Yang Q, Hou Z, Wang Y, Delgado J (2015) A comparative study of shallow groundwater level simulation with WA–ANN and ITS model in a coastal island of south China. Arab J Geosci 8:6583–6593

Yesilnacar MI, Rastgeldi Dogan T, Kumral M (2011) The effect of high fluoride in groundwater on human health: initial findings and recommendations of a case study from Sanliurfa, Turkey. ISEB-International Symposium on Environmental Biogeochemistry, Istanbul, Turkey

Yesilnacar MI, Sahinkaya E, Naz M, Ozkaya B (2008) Neural network prediction of nitrate in groundwater of Harran Plain, Turkey. Environ Geol 56(1):19–25

Yeşilnacar MI, Atasoy AD, Kumral M, Dülgergil ÇT, Aydoğdu M (2013) Yeraltı Sularında Yüksek Florürün Oluşumu, Dağılımı, İnsan Sağlığına Etkileri ve Florür Gideriminin Araştırılması, KB Şanlıurfa. TUBİTAK Project No: 110Y234. Sanliurfa, Turkey (in Turkish)

Yeşilnacar M, Demir Yetiş A, Dülgergil ÇT, Kumral M, Atasoy AD, Rastgeldi Doğan T, Tekiner Sİ, Bayhan İ, Aydoğdu M (2016) Geomedical assessment of an area having high-fuoride groundwater in southeastern Turkey. Environ Earth Sci 75:162

Yetis R, Demir Yetis A, Atasoy AD, Yesilnacar MI (2019) Hydrogeochemical characteristics and quality assessment of groundwater in Balikligol Basin, Sanliurfa, Turkey. Environ Earth Sci 78:331