

Rough set–BPSO model for predicting vitamin D deficiency in apparently healthy Kuwaiti women based on hair mineral analysis

Hala S. Own¹ · Khulood O. Alyahya² · Waheeda I. Almayyan³ · Ajith Abraham⁴

Received: 3 February 2016 / Accepted: 1 July 2016
© The Natural Computing Applications Forum 2016

Abstract Vitamin D deficiency is prevalent in the Arabian Gulf region, especially among women. Recent researches show that, the vitamin D deficiency is associated with mineral status of patient. Therefore, it is important to assess the mineral status of patient to reveal the hidden mineral imbalance associated with vitamin D deficiency. A well-known test such as the red blood cells is fairly expensive, invasive, and less informative. On the other hand, a hair mineral analysis can be considered an accurate, excellent, highly informative tool to measure mineral imbalance associated with vitamin D deficiency. In this study, 120 apparently healthy Kuwaiti women were assessed for their mineral levels and vitamin D status by hair and serum samples, respectively. This information was used to build a

computerized model that would predict vitamin D deficiency based on its association with the levels and ratios of minerals. The model introduces a two-stage reduction technique based on BPSO and rough set theory as attribute reduction and rules extraction to predicting vitamin D deficiency. The results show that the proposed model (RS + BPSO), not only can effectively detect the deficiency in vitamin D, but can also provide valuable information with regard to the mineral imbalance as a cause of deficiency which should be addressed in any treatment management. To the best of our knowledge, this is the first work that predicts vitamin D deficiency based on hair minerals analysis.

Keywords Vitamin D deficiency · Women · Kuwait · Hair mineral analysis · Rough set theory · Particle swarm optimization (PSO) · Classification

✉ Hala S. Own
halaown@gmail.com

Khulood O. Alyahya
ko.alyahya@paaet.edu.kw

Waheeda I. Almayyan
wi.almayyan@paaet.edu.kw

Ajith Abraham
ajith.abraham@ieee.org

- ¹ Department of Solar and Space Research, National Research Institute of Astronomy and Geophysics, El-Marsad Street, P. O. Box 11421, Helwan, Egypt
- ² Science Department, College of Basic Education, Public Authority for Applied Education and Training, Kuwait, Kuwait
- ³ Department of Computer Information Systems, College of Business Studies, Public Authority of Applied Education and Training, Kuwait, Kuwait
- ⁴ Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, Auburn, WA, USA

1 Introduction

The Middle Eastern populations are among the lowest to obtain vitamin D levels, and vitamin D deficiency has been reported repeatedly from many parts of the region including the Arabian Gulf countries [1]. Females seem to have a higher risk of deficiency compared to males [2]. Levels below 25 nmol/L have been found common in 81 % of Saudi females [4], 78.4 % of Kuwaiti adolescent females [5], 67.6 % of Bahraini females [6], and in 51.4 % of Qatari females [7]. Limited sun exposure due to veiling, high temperature, and economic status is among the risk factors of vitamin D deficiency [5].

Numerous studies show that low serum levels of 25-hydroxyvitamin D (25OHD), the generally accepted biomarker of vitamin D status, are linked to an array of

ailments including but not limited to osteoporosis, cardiovascular disease, diabetes, obesity, cancer, and autoimmune disorders [6]. In addition, low levels of vitamin D in pregnant mothers are associated with low birth weight and higher risk of developing immune-related disorders in their newborns [7].

The evidence suggests that maintaining an adequate vitamin D status plays a critical role in the prevention of disease, and for this reason, vitamin D supplementation, both orally and intravenously, has been widely and heavily administered in vitamin D deficient patients, only to find little or no effect on the outcome of the disease. In fact, studies have not been able to confirm the benefit of supplementation with vitamin D in cases of osteoporosis, cardiovascular disease, hypertension, obesity, and autoimmune diseases [8]. In musculoskeletal studies, however, supplementation with vitamin D has shown a significant benefit in reducing falls in the elderly [9]. In conclusion to the recent scientific data, vitamin D deficiency, which was once believed to be a cause of disease, is now conceived as a sign of illness and health imbalance.

Nutrient interrelations are complex because they depend on each other to perform a single function. An imbalance in one nutrient can affect the absorption, metabolism, and accumulation of other nutrients [9]. For example, vitamin D deficiency can cause rickets and osteomalacia due to its role in calcium metabolism in the bone. Excessive vitamin D intake, on the other hand, can cause magnesium deficiency. Because magnesium is required for the activation and transportation of vitamin D throughout the body, magnesium deficiency, in turn, can cause serum vitamin D deficiency [10]. Thus, a typical deficiency in one nutrient may be secondary to the imbalance of another nutrient. This is important to consider when managing a treatment of a disease, as addressing only a single nutrient may not be effective enough. Studies have shown inconsistent results on bone mineral density when supplementing with vitamin D alone [9], but combining vitamin D with other nutrients has shown to be significantly effective in raising bone mineral density. The literature shows that vitamin D deficiency is associated with calcium, phosphorus, and magnesium, and an earlier article shows that vitamin D has a close and synergetic relation with calcium, magnesium, sodium, copper, and selenium [10]. Therefore, it is important to assess the mineral status of patients to reveal the hidden mineral imbalance associated with vitamin D deficiency.

Widespread use of medical information systems and explosive growth of medical databases require efficient mechanism methods to be coupled with clinical analysis. Data mining has been applied to the medical domain as a significant tool for knowledge discovery. Medical data mining has played a role in discovering hidden pattern,

noticing relations, and generating decision rules. Such rules have been massively applied especially in the area of medical diagnosis and treatment [11].

Several computational intelligent (CI) methods such as neural networks, fuzzy sets, decision trees, and expert systems are effectively applied in the medical field [12]. The main role of the CI tools is to tackle with potential problems in medical data such as, missing data, inconsistent information, and extraction of meaningful information from a huge set of irrelevant data. Feature selection and reduction have played a great role in deducing the hidden trends within the medical data with the purpose of identifying the most significant attributes in the dataset [13]. Therefore, feature reduction has a great impact on the effectiveness and complexity of the classification process [14].

Rough set theory (RS) is one of the most motivating areas of CI research; it become increasingly popular with medical applications. One of the major advantages of RS is that it does not require any additional information about the data such as probability distribution or grade membership [15]. RS theory also was proved to be competent in handling inconsistency which is a common problem within medical data. Such inconsistency is normally caused by the existence of an indiscernibility relation in data. Unlike traditional computational intelligent techniques, RS does not require any preprocessing to eliminate the inconsistency [16]. Moreover, RS have an efficient algorithm for dimensionality reduction.

2 Research motivation and contribution

2.1 Motivation

Vitamin D status plays a critical role in the prevention of disease therefore; improvement in predictive accuracy of vitamin D deficiency is highly required. Previous reported work in predicting vitamin D deficiency has several potential limitations that may not accurately reflect the vitamin D status.

1. Some studies used statistical methods like multiple linear regression (MLR). Though, the constructed model based on statistical methods usually depends on a predetermined model to predict the occurrence or not of an event by fitting data to a logistic curve. In addition, the computational complexities of real-time classification problems are highly nonlinear in nature. In a research published in 2013, Guo et al. [17] suggested the using of support vector regression (SVR) model to predict the serum 25(OH) D concentrations which indicates the status of vitamin D. However, the

- accuracy of the technique highly depends on choosing appropriate model parameters and kernel function which is always done in trial-and-error manner.
2. Most of published works depend on the data gathered by questionnaire surveys including self-reported sun exposure and sun protection behavior and different physical activity. However, the data collected through questionnaire may not be reliable, as the validity of the person response cannot be verified. Moreover, with questionnaire it is very difficult to detect errors, and there is a great possibility of misinterpretation of person responds. Consequently, it may lead to misclassification.
 3. Several researchers depend on blood sampling test to assessed vitamin D status (25OHD); however, on extracellular fluids such as blood, urine, and sweat may not provide a good indication of the minerals activity in the body. This is because the blood tends to balance the minerals continuously either by depositing the excessive minerals into the tissues or by forcing them out of the cells and excreting them into the urine and sweat. Measuring minerals in intracellular fluids such as the red blood cells is a well-known test, but fairly expensive, invasive, and less informative compared to the hair tissue [18].

A hair mineral analysis is a tissue biopsy that is safe, noninvasive, and relatively inexpensive. Provided the accuracy and expertise of the laboratory, the test can be considered an accurate and an important tool in diagnostics and monitoring. The hair mineral analysis is a standard test applied in routine forensic assessment, and in clinical and research evaluation of minerals and toxic metals [19]. It is also highly informative because it measures a large number of minerals and their ratios including both essential and toxic minerals just by a small amount of hair that is cut from the scalp. Hair tissue, like any other tissue in the body, contains minerals that are deposited as the hair grows [18]. Thus, the hair mineral analysis is an excellent tool in assessing the nutritional balance [20].

Recently, RS and binary particle swarm optimization (BPSO) have been useful tools in medical research. In this article, we build a computerized model that would predict vitamin D deficiency based on its association with the levels and ratios of minerals measured by hair tissue samples. We propose a hybrid model based on the integration of RS and BPSO to build a novel classification model.

2.2 Contribution

Hair mineral analysis tool is not as common as requesting a blood test to measure 25OHD, nevertheless the information

obtained with using the hair mineral analysis can assist in the prognosis of the disease, including the likelihood of vitamin D deficiency. The main focus of this study is to develop a computerized model to predict vitamin D deficiency using the minerals levels and ratios as predictive features. We claim our study is the first trial to predict vitamin D deficiency based on hair minerals samples.

Healthcare-related data mining is one of the most promising fields for discovering new facts and trends from large quantities of data. Therefore, excluding irrelevant and redundant features usually improves the understanding of the suggested computational models. Moreover, high-dimensional dataset usually worsens the classification accuracy. Consequently, feature selection has become a vital step in many biomedical data mining problems due to their ability to handle high dimensionality of input attributes.

In this study, the gathered features were reduced through BPSO algorithm firstly to construct an intermediate decision table. Then, RS was applied as a feature reduction algorithm and rules extraction technique on the intermediate decision table to perform classification. The proposed system concludes with a set of minimal minerals which has direct effect in predicting vitamin D deficiency. The results show that the proposed model, not only can effectively detect the deficiency in vitamin D, but can also provide valuable information with regard to the mineral imbalance as a cause of deficiency which should be addressed in any treatment management. To the best of our knowledge, this is the first work that predicts vitamin D deficiency based on hair minerals analysis.

Since there is considerable discussion on the serum concentration of 25OHD associated with the deficiency, two main experiments with 25 and 50 nmol/L serum concentrations were conducted to confirm the proposed model validity. The results confirm that the proposed model has greatly reduced the feature dimension while improving the classification performance as well as the outcomes concordance with the recent clinical literature. The experimental results showed how the integration approaches increased the predictive relationship between vitamin D deficiency and mineral levels extracted from the hair samples. The proposed model can be employed by specialists when deciding the clinical regimen for their patients.

The remainder of the article is organized as follows. Section 3 introduces the preliminary of RS and BPSO. Section 4 presents the suggested RS + BPSO hybrid model methodology. Section 5 describes the data gathering procedure. Sections 6 and 7 illustrate the experimental analysis and discussion. Conclusion and future work are given in Sect. 8.

3 Preliminary

3.1 Rough set theory

RS theory is a mathematical approach for handling vagueness and uncertainty in data analysis. A RS is characterized by a pair of precise concepts, called lower and upper approximations, generated using object indiscernibilities. The main advantage of rough set theory is that it does not need any preliminary or additional information about data. A detailed information about the RS can be found in [21].

3.2 PSO for feature selection

The particle swarm optimization (PSO) approach has recently gained more attention for solving optimization and feature subset selection problems. As an algorithm, the main strength of PSO is its fast convergence with few parameters, which compares favorably with many global optimization algorithms because it has strong search capability with flexible, well-balanced, and efficiently mechanism to find the set of best feature in data knowledge discovery domain. Moreover, its mathematical base is simple reasonable computational cost [22].

The particle swarm optimization (PSO) technique is a population-based stochastic optimization technique first introduced in 1995 by Kennedy and Eberhart. A detailed description with a lot of background information can be found in their textbook swarm intelligence [23]. In PSO, a possible candidate solution is encoded as a finite-length string called a particle p_i in the search space. All of the particles make use of its own memory and knowledge gained by the swarm as a whole to find the best solution. With the purpose of discovering the optimal solution, each particle adjusts its searching direction according to two features, its own best previous experience (p_{best}) and the best experience of its companions flying experience (g_{best}). Particles evolve simultaneously based on knowledge shared with neighboring particles; they make use of their own memory and knowledge gained by the swarm as a whole to find the best solution. The best previous experience of all neighbors of particle i is called g_{best} . Each particle additionally keeps a fraction of its old velocity. The particle updates its velocity and position with the following equation in continuous PSO [23]:

$$v_{pd}^{new} = \omega \times v_{pd}^{old} + C_1 \times \text{rand}_1() \times (p_{best_{pd}} - x_{pd}^{old}) + C_2 \times \text{rand}_2() \times (g_{best_{dd}} - x_{pd}^{old}) \quad (1)$$

$$x_{pd}^{new} = x_{pd}^{old} + v_{pd}^{new} \quad (2)$$

The acceleration coefficients (C_1) and (C_2) are constants represent the weighting of the stochastic acceleration terms that pull each particle toward the p_{best} and g_{best} positions. Therefore, the adjustment of these acceleration coefficients changes the amount of “tension” in the system. In the original algorithm, the value of ($C_1 + C_2$) is usually limited to 4 [24]. Particles’ velocities are restricted to a maximum velocity, V_{max} . According to Eq. 2, the particle’s new velocity is calculated according to its previous velocity and the distances of its current position from its own best experience and the group’s best experience. Afterward, the particle flies toward a new position according to Eq. 2.

3.3 Binary PSO

Kennedy and Eberhart [25] presented a discrete binary version of PSO algorithm specifically to handle discrete problems since that the initial PSO algorithm was developed for a space of continuous values, where the variable domain is finite. In the binary version of PSO, the particles are represented by binary values (0 or 1). Each particle velocity is updated according to the following equations:

$$S(V_{id}^{new}) = \frac{1}{1 + e^{-V_{id}^{new}}} \quad (3)$$

$$x_{id} = \begin{cases} 1 & \text{if } (\text{rand} < S(V_{id}^{new})) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $S(V_{id}^{new})$ denotes the particle velocity obtained from Eq. 1, function $S(V_{id}^{new})$ is a sigmoid transformation, and rand is a random number selected from a uniform distribution $U(0, 1)$.

4 Rough set–BPSO hybrid model for predicting vitamin D deficiency

The proposed model has two main stages; the feature selection stage, which integrates BPSO and RS to extract the most prominent features. The second stage employs RS rule extraction and classification capabilities to build a model for predicting vitamin D deficiency based on its association with the levels and ratios of minerals. The formal steps of the BPSO + RS method are stated in the following algorithm:

Proposed algorithm Rough set BPSO based classification

Input: information system table (*S*) the set of all conditional features *C*; *D*, the decision features.

Output: set of predictive rules

Step1

Construct a first step attribute reduction :

This step done by using BPSO. For certain number of iteration we evaluate the fitness function of the swarm and update the velocity and the position for each particle according to Eqs.1 and 2 respectively until termination condition is satisfied. The corresponding particle of the generated optima is the best combinatorial features. These features were chosen as a set of reduct feature in this stage.

Step2

Construct a second step attribute reduction :

This step done by using RS. The output features of step 1 considered as input feature in this step. We generate the reduct set of attributes $R = \{A \in P(C): \gamma(A, D) = \gamma(C, D)\}$ such that $\gamma(C, D) = \gamma(R, D)$.

Step 3

Generate the set of predictive rules :

for each reduct *R*; generate a rule such that : $(c_1 = v_1 \wedge c_2 = v_2 \dots \wedge c_n = v_n) \rightarrow d = u$

$\varphi \Rightarrow (d = v)$, where $\varphi = c_1 \wedge c_2 \wedge \dots \wedge c_q$ is a conjunction, satisfying $[\varphi]_K^+ \neq \emptyset$ and $v \in V_d, V_d$ is the set of values of *d*. If *K* is the concept, the positive cover $[C]_K^+ = [C] \cap K$ denotes the set of positive examples covered by *C*.

4.1 Feature selection stage

The entire system flow of the proposed approach is shown in Fig. 1. The individual stages detailed description is introduced in the following subsections.

In medical datasets, most of the collected features are not significant or redundant. These features may be deleterious

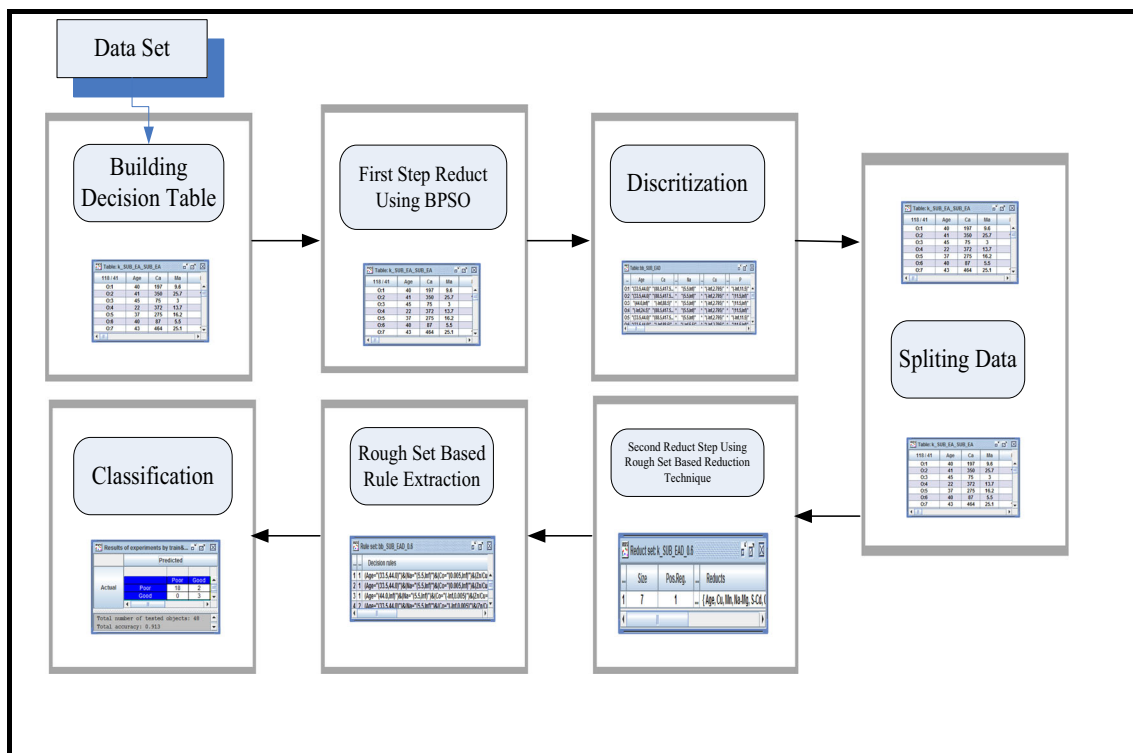


Fig. 1 The proposed RS + BPSO vitamin D deficiency predictive model

in the case of relatively small training sets, where this irrelevancy and redundancy are harder to evaluate. On the other hand, this extreme number of features carries the problem of memory usage in order to represent the dataset. Feature selection is a solution that involves finding a subset of prominent features to improve predictive accuracy and to remove the redundant features [14].

Recently, RS and PSO have been useful tools in medical research. Wang et al. [26] applied a RS-based method to predict the degree of malignancy in brain glioma. For feature selection, the researchers applied genetic algorithm PSO and RS-based feature selection (PSORSFS). Inbarani et al. [27] applied the RS two techniques of unsupervised quick reduct (USQR) unsupervised relative reduct (USRR) and the unsupervised PSO-based relative reduct (USPSO-RR) for gene expression dataset. Inbarani et al. [14] introduced a supervised feature selection methods based on hybridization of RS and PSO for the diseases diagnosis. The proposed feature selection hybridized PSO-based techniques; quick reduct and relative reduct methods. These were tested for their ability to diagnose four different medical datasets including erythemato-squamous diseases, breast tissue, prognostic, and cardiac disorders.

In this article, we introduce an efficient two-stage reduction technique for improving the classification accuracy and for detecting an optimal feature subset selection. There are three basic reasons for applying two-level reduct steps:

1. The efficiency of RS would be affected by the scale of dimensionality. Therefore, quick reduct algorithm fails to scan most of the reducts when the number of features is high [28].
2. The relative dependency method makes backward elimination of the attributes using the relative dependence parameter, and the proportion of the number of the indiscernible subsets is evaluated. One drawback of this algorithm is that in a case when the numbers of the subsets in the numerator and denominator parts are equal. The relative dependency which is measured in the lower nodes usually halts the algorithm if its value is below 1 [28].
3. PSO is a well-known tool to obtain improved results in high-dimensional search spaces by finding the optimum characteristics with the help of local and global search in an iterative manner [29]. It does not need complex operator but only require simple mathematical operators without highly time and memory consuming.

In the proposed feature selection technique, we try to combine the advantages of RS and BPSO to empower the ability for feature selection. Initially, the BPSO was applied to the vitamin D dataset to obtain the reduct information table. Due to the fact that RS theory cannot directly process continuous values of features, the obtained reduced

information table was discretized first. We adopted, in this article, RS with Boolean reasoning (RSBR) algorithm proposed by Zhong et al. [30]. The main advantage of RSBR is that it combines discretization of real-valued features and classification. The algorithm tries to discover fewer breakpoints with larger dependency degree between features. It uses a bottom-up approach which adds cuts for a given attribute one-by-one in subsequent iterations [31].

The yielded discretized information table is fed to the reduction techniques using RS. An algorithm called “dynamic reduct,” implemented in RESE [31], attempts to calculate a minimal reduct. The process requires several subtables to be examined in order to find the frequently repeating minimal subsets of features (reduct). Such dynamic reducts may be calculated for general or decision-related indiscernibility relation [31]. The output of this stage is a minimal set of features which preserve the power of the whole features.

4.2 Rule extraction and classification stage

Extracting comprehensible classification rules is the most emphasized concept in data mining. As we mentioned before, to transform a reduct into a rule, we have to bind the condition feature values of the object class from which the reduct originated to the corresponding features of the reduct. Then, to complete the rule, a decision part comprising the resulting part of the rule is added. This is done in the same way as for the condition features. To classify objects, which have never been seen before, rules generated from a training set will be used. These rules represent the actual classifier. This classifier is used to predict to which class the new objects are attached. The nearest matching rule is determined as the one whose condition part differs from the feature vector of re-object by the minimum number of features [33]. Once the rules are generated, the classification will begin using the test set to predict the actual class for each object in the test set. The output of the classification is the confusion matrix (as shown in later in Table 4), which records correctly and incorrectly recognized objects for each class in the test set.

5 Data gathering

The data used in this study was granted ethical approval by the Joint Committee for the Protection of Human Subjects in Research combined by Kuwait Institute for Medical Specialization KIMS–Ministry of Health and AbdulMihsin Al-Abdulrezzag Health Sciences Centre HSC–Health Sciences Centre–Kuwait University.

The data were collected from healthy Kuwaiti women between the age of 19–49 years old, not suffering from any

chronic illness, not pregnant or nursing, and not taking supplements for the past 6 months. Women from a previous study and from the college of Basic Education, PAAET were approached and invited to participate in the study. Those who agreed were recalled for a full examination, during which a total blood sample of 10 mL and a ≤ 3 cm hair long (~ 0.15 – 0.2 mg) cut from the scalp side of the hair was obtained. The hair was collected from 2 to 3 areas on the back of the head (mid-parietal to occipital region).

The collected serum samples were shipped on dry ice to The Doctors Laboratory (TDL) in London, UK (<http://www.tdlpathology.com/>), where they were measured for 25OHD by a radio-immunoassay, whereas the hair samples were shipped to the Trace Elements Laboratory (TEI), USA (<http://www.traceelements.com/>), to determine the hair mineral levels and ratios. Each hair sample was collected in a separate and sealed envelope.

The women in this dataset had a high prevalence of vitamin D deficiency; 81.4 % had levels <50 nmol/L, and 74.6 % had levels <25 nmol/L. These women are not restricted to a particular residence; in fact, the dataset includes women from all of the six governorates in Kuwait with no significant difference in serum 25OHD.

The hair mineral analysis shows a wide variation in the mineral status among the women, with more than 60 % of them obtaining calcium levels beyond the laboratory reference range, and 39 % of them having potassium levels below the reference range. The report of the hair mineral analysis includes nutritional minerals such as calcium, magnesium, sodium, and potassium, trace nutritional minerals such as lithium, germanium, and vanadium, and heavy toxic metals such as arsenic, lead, mercury, and cadmium (Table 1). Further information regarding method of hair collection, minerals tested, and the resulting report can be found on their website.

The TEI lab report as shown in Fig. 2 states that all mineral levels are reported in milligrams percent (milligrams per one hundred grams of hair). One mg % is equal to 10 parts per million (ppm).

The data set of 118 records, 39 predictor variables including the age attribute and 1 target variable was constructed. The target variable “Class” has two possible values “Poor” (indicate that the subject suffer from deficiency in vitamin D), “Good” (indicate the subject does not suffers from deficiency in vitamin D). Table 1 shows the 39 predictive minerals and their description, used in our article.

6 Experimental and performance analysis

There is no absolute agreement on the optimal level of serum vitamin D to date. The Institute of Medicine (IOM) [34] suggested that a sufficiency in vitamin D levels should

Table 1 List of minerals

| No. | Mineral | | Reference range (mg%)* |
|-----|--------------------|-------|------------------------|
| 1 | Calcium | Ca | 22–97 |
| 2 | Magnesium | Mg | 2–11 |
| 3 | Sodium | Na | 4–36 |
| 4 | Potassium | K | 2–24 |
| 5 | Copper | Cu | 0.9–3.9 |
| 6 | Zinc | Zn | 10–21 |
| 7 | Phosphorus | P | 11–20 |
| 8 | Iron | Fe | 0.5–1.6 |
| 9 | Manganese | Mn | 0.01–0.13 |
| 10 | Chromium | Cr | 0.02–0.08 |
| 11 | Selenium | Se | 0.03–0.18 |
| 12 | Boron | B | 0.02–0.91 |
| 13 | Cobalt | Co | 0.001–0.003 |
| 14 | Molybdenum | Mo | 0.003–0.008 |
| 15 | Sulfur | S | 3546–5336 |
| 16 | Uranium | U | ≥ 0.017 |
| 17 | Arsenic | As | ≥ 0.02 |
| 18 | Beryllium | Be | ≥ 0.001 |
| 19 | Mercury | Hg | ≥ 0.18 |
| 20 | Cadmium | Cd | ≥ 0.014 |
| 21 | Lead | Pb | ≥ 0.3 |
| 22 | Aluminum | Al | ≥ 1.8 |
| 23 | Calcium/phosphorus | Ca/P | 1.6–3.6 |
| 24 | Sodium/potassium | Na/K | 1.4–3.4 |
| 25 | Calcium/potassium | Ca/K | 2.2–6.2 |
| 26 | Zinc/copper | Zn/Cu | 4–12 |
| 27 | Sodium/magnesium | Na/Mg | 2–6 |
| 28 | Calcium/magnesium | Ca/Mg | 3–11 |
| 29 | Iron/copper | Fe/Cu | 0.2–1.6 |
| 30 | Calcium/lead | Ca/Pb | ≥ 84 |
| 31 | Iron/lead | Fe/Pb | ≥ 4.4 |
| 32 | Iron/mercury | Fe/Hg | ≥ 22 |
| 33 | Selenium/mercury | Se/Hg | ≥ 0.8 |
| 34 | Zinc/cadmium | Zn/Cd | ≥ 500 |
| 35 | Zinc/mercury | Zn/Hg | ≥ 200 |
| 36 | Sulfur/mercury | S/Hg | $\geq 28,450$ |
| 37 | Sulfur/cadmium | S/Cd | $\geq 71,126$ |
| 38 | Sulfur/lead | S/Pb | ≥ 5690 |
| 39 | Age | Years | 19–47 |

* From the trace elements, Inc. hair tissue report. One mg % is equal to ten parts per million (ppm)

be marked at 50 nmol/L. Levels of 25OHD less than 50 nmol/L would be insufficient and less than 30 nmol/L are deficient. The World Health Organization (WHO) and the Nordic Council of Ministers agreed on marking sufficiency at 50 nmol/L, but defined deficiency as below 25 nmol/L [35]. Other organizations and societies found that higher than 50 nmol/L was needed for optimal health.

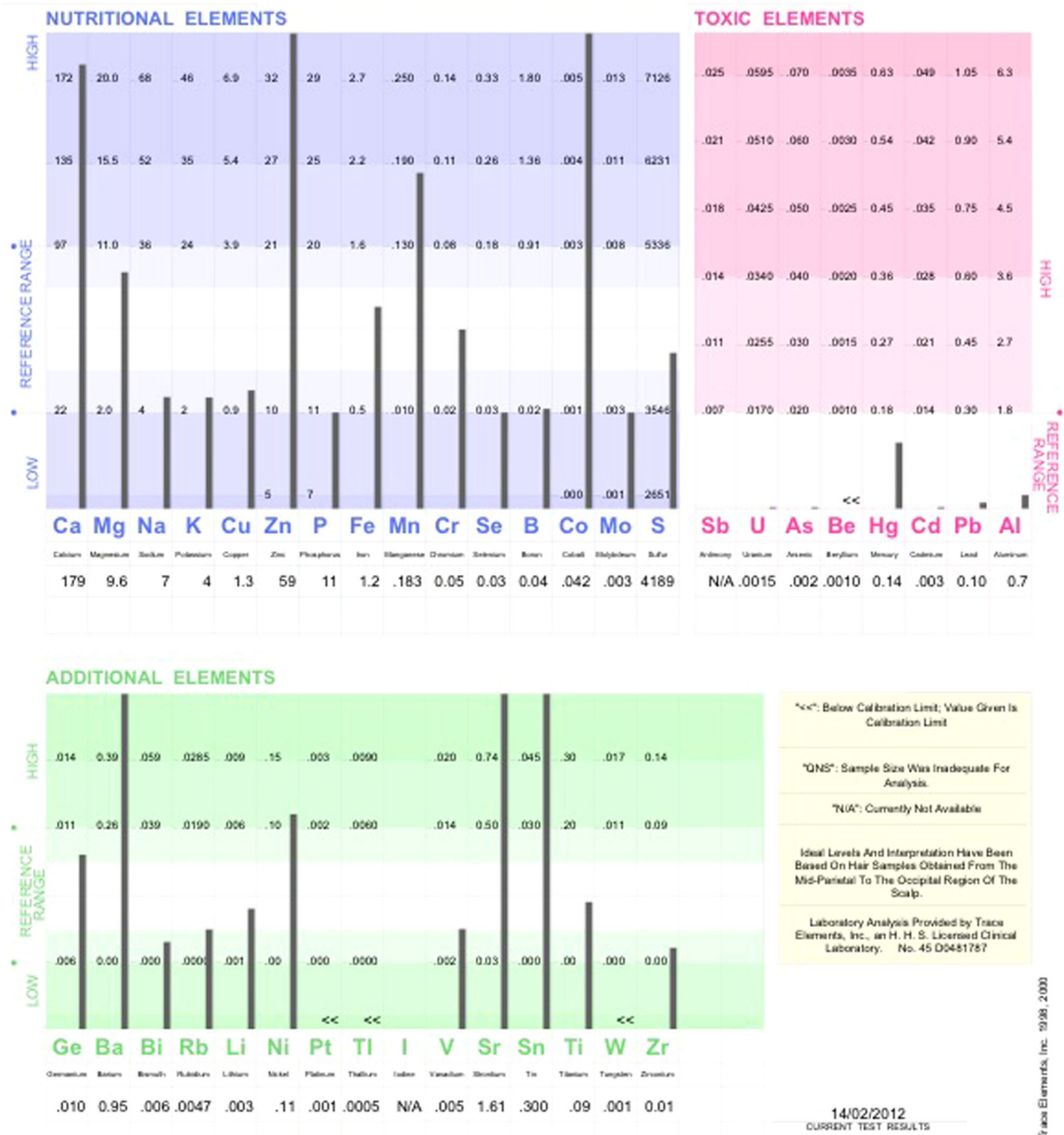


Fig. 2 A sample of TEI Lab report

Nevertheless, there is some agreement that optimal levels should be around 50 nmol/L, and due to the majority class in the 25 nmol/L cutoff being larger than the 50 nmol/L in this study, we used both 50 and 25 nmol/L cutoffs.

Two main experiments were conducted to evaluate vitamin D deficiency on apparently healthy Kuwaiti women to analyze their hair tissue samples based on its association with the levels and ratios of minerals. For all of the following experiments, a 60 % split was used for

training, and the remaining 40 % was kept for testing. The first set of experiments was conducted on a 25 nmol/L cutoff. The second set of experiments was conducted using a cutoff 50 nmol/L. The class distribution of the dataset in both cutoffs is shown in Tables 2 and 3, respectively. The computations of rules have been done only on training set. The results of computations of rules were applied to the classification of objects from the dataset kept for testing. We can notice that the class distribution of each dataset is

Table 2 25 nmol/L cutoff class distribution

| Class | Count | Weight (%) |
|-------|-------|------------|
| Poor | 30 | 25.4 |
| Good | 88 | 75.6 |

Table 3 50 nmol/L cutoff class distribution

| Class | Count | Weight (%) |
|-------|-------|------------|
| Poor | 96 | 81.3 |
| Good | 22 | 18.6 |

imbalanced. The ratio of majority class “poor” to the minority class “good” is 25.4 and 75.6 %, respectively, in the 25 nmol/L cutoff dataset, whereas the ratio is 81.3 and 18.6 % in the 50 nmol/L cutoff.

6.1 Assessment metrics

In general, the measures of the quality of classification are built from a confusion matrix (as shown in Table 4), where the performance of a classifier is evaluated based on its overall accuracy on an independent test dataset. In the confusion matrix, TP (true positive) is the number of actual positives that were correctly classified as positives. FP (false positive) is the number of actual negatives that were incorrectly classified as positives. TN (true negative) is the number of actual negatives that were correctly classified as negatives. FN (false negative) is the number of actual positives that were incorrectly classified as negatives [36].

However, the overall classification accuracy on an imbalanced dataset is mainly dominated by the majority class. For that reason accuracy is not an appropriate evaluation measure for imbalanced data [37]. Therefore, the researchers used precision, recall, and *F*-measure metrics to evaluate the performance of imbalanced dataset.

Precision is defined as the proportion of the true-positive samples out of the all samples that the classified defined as positive. On the other hand, the recall metrics measures to which extent the classifier define positive samples. The *F*-score measure conveys the balance between the precision and recall. Therefore, it is considered as a measurement of overall system performance.

Table 4 Confusion matrix

| | Predicted class | | |
|------------|-----------------|----|----|
| | Yes | No | |
| True class | Yes | TP | FN |
| | No | FP | TN |

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

$$F\text{-Measure} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

6.2 Experiments analysis

Generally, the irrelevant features in the dataset have no impact on the classification performance. However, the existence of these attributes in the dataset might affect the quality of classification. As we mentioned earlier, we applied two reduction steps. The first step is conducted using BPSO while the second step uses a RS-based reduction algorithm. Table 5 shows the reduct set for each cutoff dataset after applying our integrated reduction technique. For the 25 nmol/L cutoff and after the first reduction step, a set of 20 attributes was produced out of 39. These features were kept to the second step of reduction. The reduction percentage of this step was 51 % from the whole feature space. After applying the second step reduction using RS, the total number of features has been further reduced from 20 to 9 features which is a reduction of 55 % from the features obtained in the first reduction step. The overall feature reduction after the two-step reduction procedure, as shown in Table 5, was about 76 % for the 25 nmol/L cutoff dataset and 82 % for the cutoff of 50 dataset.

The final generated sets of features, which are used to generate the list of rules for the purpose of classification both datasets are listed in Table 6. The features generated are all core features, as shown in Table 6. The reduct feature sets can give the user a good feedback on most important features which cause the vitamin D deficiency (discussed in details in next section). Using the minimal reduct set, classification rules are extracted using the Rough Set Exploration System (RSES) [32]. The total numbers of rules are 35 and 32 for cutoff 25 and 50, respectively. Furthermore, we apply filtering to the generated rules such that the less significant rule, which is supported only by one instance, can be removed from the whole generated rules to produce only the most significant one. Table 7 shows the difference in number of rules generated before and after filtering in each dataset.

Table 7 shows that the number of rules significantly decreases by 60 and 48.98 % in case of cutoff 25 and 50, respectively. Tables 8 and 9 depict some of the generated decision rules for the 25 nmol/L cutoff 25 and 50 datasets, which classify the instances depending on vitamin D

Table 5 The results of applying first and second reduction step with the overall of reduction

| | First reduct step | | Second reduct step | | Overall reduction (%) |
|------------------|-------------------|-----------------------------|--------------------|-----------------------------|-----------------------|
| | No. of features | Percentage of reduction (%) | No. of features | Percentage of reduction (%) | |
| 25 nmol/L cutoff | 20 | 51 | 9 | 55 | 76 |
| 50 nmol/L cutoff | 21 | 48 | 7 | 66 | 82 |

Table 6 The final reduct set of features

| | Final feature generated |
|------------------|---|
| 25 nmol/L cutoff | {Mg, Cu, Zn, Fe, S, Na/k, Ca/Mg, Zn/Hg, S/cd} |
| 50 nmol/L cutoff | {Age, K, Se, Co, Al, Zn/Cu, Fe/Hg} |

Table 7 Number rules generated before and after filtering

| | Number of rules before filtering | Number of rules after filtering |
|-----------|----------------------------------|---------------------------------|
| Cutoff 25 | 35 | 14 |
| Cutoff 50 | 32 | 17 |

deficiency as good or poor. As mentioned earlier, these rules will be used to assess the performance of our proposed method.

6.3 Classification performance discussion

Experiments have been carried on to compare the precision, recall, *F*-measure, and accuracy between our proposed model and other machine-learning classification paradigms to predict vitamin D deficiency. These methods include artificial neural network (MLP) statistical model (Bayes net), decision trees (*C4.5*), and ensemble classifier (bagging). Bayes net tree classifier was implemented using a hill-climbing algorithm. MLP with backpropagation was

implemented with a momentum of 0.2 and a learning rate of 0.3. The ensemble bagging was built with the *k*-NN as a based classifier with a size of 100, while the number of iterations to be performed was set to 10, and the number of execution slots used for constructing the ensemble was held at 1. The *C4.5* decision tree classifying the minimum number of instances per node was held at 2. The cross-validation folds for the testing set were held at 3 during the testing of *C4.5*, whereas confidence factor was held at 0.25. Table 10 lists the basic parameters of BPSO during the experiments.

The results are listed in Tables 11 and 12, respectively. According to classification performance, we can see that the proposed model outperformed all the classifiers in term of precision, recall, *F*-measure and accuracy. This indicates that the overall performance of our method is better in all comparisons. Moreover, this indicates that the selected feature subsets are representative and informative, and thus, it can be used instead of the complete data for classification purpose.

6.4 Measure the effectiveness of using BPSO

Other experiments are conducted to investigate the effectiveness of using BPSO as a first reduction phase. In this experiment, we use only one-step reduct using RS instead of two and repeat all the experiments again. Table 13 shows the reduct set generated for both datasets. The

Table 8 RS-based decision rules generation for 25 nmol/L cutoff dataset

| | |
|---|---|
| 1 | (Mg = "(-Inf,8.45)")&(Cu = {"(-Inf,1.75)" (1.75,Inf)})&(Zn = "(-Inf,18.5)")&(Fe = {"(Inf,0.85)" (0.85,Inf)})&(S = {"(4143.5,Inf)" (-Inf,4143.5)})&(Na/K = {"(Inf,1.97)" (1.97,Inf)})&(CA/Mg = {"(16.525,Inf)" (-Inf,16.525)})&(Zn/Hg = "(529.145,Inf)")&(S/Cd = {"(3,980,000.0,Inf)" (-Inf,3,980,000.0)}) =>(Class = {Poor[16]}) |
| 2 | (Mg = {"(-Inf,8.45)" (8.45,Inf)})&(Cu = "(-Inf,1.75)")&(Zn = {"(18.5,Inf)" (-Inf,18.5)})&(Fe = "(-Inf,0.85)")&(S = {"(-Inf,4143.5)" (4143.5,Inf)})&(Na/K = {"(1.97,Inf)" (-Inf,1.97)})&(Ca/Mg = "(-Inf,16.525)")&(ZN/Hg = {"(529.145,Inf)" (-Inf,529.145)})&(S/Cd = {"(3,980,000.0,Inf)" (-Inf,3,980,000.0)}) =>(Class = {Poor[9]}) |
| 3 | (Mg = "(-Inf,8.45)")&(Cu = "(1.75,Inf)")&(Zn = {"(-Inf,18.5)" (18.5,Inf)})&(Fe = {"(-Inf,0.85)" (0.85,Inf)})&(S = "(-Inf,4143.5)"&(Na/K = {"(1.97,Inf)" (-Inf,1.97)})&(Ca/Mg = "(16.525,Inf)")&(ZN/Hg = {"(-Inf,529.145)" (529.145,Inf)})&(S/Cd = {"(-Inf,3,980,000.0)" (3,980,000.0,Inf)}) =>(Class = {Poor[8]}) |
| 4 | (Mg = {"(8.45,Inf)" (-Inf,8.45)})&(Cu = "(-Inf,1.75)")&(Zn = {"(-Inf,18.5)" (18.5,Inf)})&(Fe = {"(0.85,Inf)" (-Inf,0.85)})&(S = "(-Inf,4143.5)"&(Na/K = "(-Inf,1.97)")&(Ca/Mg = {"(16.525,Inf)" (-Inf,16.525)})&(Zn/Hg = {"(-Inf,529.145)" (529.145,Inf)})&(S/Cd = "(-Inf,3,980,000.0)") =>(Class = {Poor[6]}) |
| 5 | (Mg = {"(8.45,Inf)" (-Inf,8.45)})&(Cu = "(1.75,Inf)")&(Zn = {"(-Inf,18.5)" (18.5,Inf)})&(Fe = "(0.85,Inf)")&(S = {"(-Inf,4143.5)" (4143.5,Inf)})&(Na/K = "(-Inf,1.97)")&(Ca/Mg = "(16.525,Inf)")&(Zn/Hg = "(-Inf,529.145)")&(S/Cd = "(-Inf,3,980,000.0)") =>(Class = {Good[5]}) |

Table 9 RS-based decision rules generation for cutoff 50 dataset

| | |
|---|---|
| 1 | (AGE = "(-Inf,33.5)")&(K = "(-Inf,5.5)")&(Se = "(0.035,Inf)")&(Co = {"(0.0065,Inf)" (-Inf,0.0065)})&(Al = {"(0.35,0.75)" (-Inf,0.35)" "(0.75,Inf)})&(Zn/Cu = {"(-Inf,9.22995)" "(9.22995,Inf)})&(Fe/Hg = {"(27.1,Inf)" (-Inf,27.1)}) =>(Class = {Poor[27]}) |
| 2 | (AGE = "(-Inf,33.5)")&(K = "(-Inf,5.5)")&(Se = {"(0.035,Inf)" (-Inf,0.035)})&(Co = "(-Inf,0.0065)")&(Al = {"(0.35,0.75)" (-Inf,0.35)" "(0.75,Inf)})&(Zn/Cu = {"(-Inf,9.22995)" "(9.22995,Inf)})&(Fe/Hg = {"(27.1,Inf)" (-Inf,27.1)}) =>(Class = {Poor[25]}) |
| 3 | (AGE = {"(-Inf,33.5)" "(33.5,Inf)})&(K = "(-Inf,5.5)")&(Se = "(0.035,Inf)")&(Co = {"(0.0065,Inf)" (-Inf,0.0065)})&(Al = "(-Inf,0.35)")&(Zn/Cu = {"(-Inf,9.22995)" "(9.22995,Inf)})&(Fe/Hg = {"(-Inf,27.1)" "(27.1,Inf)}) =>(Class = {Poor[21]}) |
| 4 | (AGE = "(-Inf,33.5)")&(K = "(-Inf,5.5)")&(Se = {"(0.035,Inf)" (-Inf,0.035)})&(Co = "(-Inf,0.0065)")&(Al = {"(0.75,Inf)" (-Inf,0.35)})&(Zn/Cu = {"(-Inf,9.22995)" "(9.22995,Inf)})&(Fe/Hg = {"(-Inf,27.1)" "(27.1,Inf)}) =>(Class = {Poor[18]}) |
| 5 | (AGE = {"(33.5,Inf)" (-Inf,33.5)})&(K = {"(-Inf,5.5)" "(5.5,Inf)})&(Se = {"(0.035,Inf)" (-Inf,0.035)})&(Co = "(-Inf,0.0065)")&(Al = {"(0.75,Inf)" (-Inf,0.35)})&(Zn/Cu = {"(-Inf,9.22995)" "(9.22995,Inf)})&(Fe/Hg = "(27.1,Inf)") =>(Class = {Poor[17]}) |
| 6 | (AGE = "(-Inf,33.5)")&(K = {"(5.5,Inf)" (-Inf,5.5)})&(Se = {"(0.035,Inf)" (-Inf,0.035)})&(Co = {"(-Inf,0.0065)" "(0.0065,Inf)})&(Al = {"(0.35,0.75)" (-Inf,0.35)" "(0.75,Inf)})&(Zn/Cu = "(9.22995,Inf)")&(Fe/Hg = "(27.1,Inf)") =>(Class = {Poor[14]}) |

Table 10 Summary of BPSO parameters

| Parameters | Values |
|---|---|
| Population size | 30 |
| Maximum number of iterations (<i>G</i>) | 100 |
| Velocity limitation (<i>V_{max}</i>) | 2 |
| Inertia weight (<i>w</i>) | 0.6 |
| Acceleration constant (<i>C₁</i> and <i>C₂</i>) | <i>C₁</i> = <i>C₂</i> = 2 |

Table 11 Comparison results (cutoff 25 with BPSO)

| Classifier | Precision | Recall | <i>F</i> -measure | Accuracy |
|------------|-----------|--------|-------------------|----------|
| Bayes net | 0.719 | 0.574 | 0.613 | 0.696 |
| MLP | 0.669 | 0.681 | 0.675 | 0.716 |
| Bagging | 0.669 | 0.681 | 0.675 | 0.636 |
| C4.5 | 0.700 | 0.746 | 0.695 | 0.683 |
| BPSO + RS | 1.0 | 0.74 | 0.85 | 0.96 |

Table 12 Comparison results (cutoff 50 with BPSO)

| Classifier | Precision | Recall | <i>F</i> -measure | Accuracy |
|------------|-----------|--------|-------------------|----------|
| Bayes net | 0.662 | 0.814 | 0.730 | 0.737 |
| MLP | 0.720 | 0.771 | 0.739 | 0.814 |
| Bagging | 0.727 | 0.746 | 0.736 | 0.792 |
| C4.5 | 0.662 | 0.814 | 0.730 | 0.680 |
| BPSO + RS | 1.0 | 0.85 | 0.91 | 0.97 |

table shows that there are two different reduct sets for 25 cutoff dataset, while there are eight different reduct sets for the 50 cutoff dataset.

As illustrated in Fig. 3, the classification quality of our BPSO + RS method in terms of *F*-measure was outstanding by more than 13 % when compared with RS classification performance without using first reduct step.

Table 13 Reduct set (cutoff 50 without BPSO)

| | Final feature generated | Size of core features |
|-----------|--|-----------------------|
| Cut of 25 | {Age, Ca, Mg, K, Co, Hg, Ca/Mg} | 5 |
| | {Age, Mg, Co, Hg, Ca/Mg, Na/Mg} | |
| Cut of 50 | {Age, Na, Co, Zn/Cu, Ca/Mg, Zn/Hg} | 3 |
| | {Age, Na, Na/k, Zn/Cu, Ca/Mg, Zn/Hg} | |
| | {Age, P, Co, Zn/Cu, Ca/Mg, Zn/Cd, Zn/Hg} | |
| | {Age, Ca, Na, P, Co, Zn/Cu, Zn/Hg} | |
| | {Age, P, Na/k, Zn/Cu, Ca/Mg, Zn/Cd, Zn/Hg} | |
| | {Age, Na, Cu, P, Se, Co, Zn/Cu, Zn/Hg} | |
| | {Age, Na, P, Na/k, Zn/Cu, Zn/Hg} | |
| | {Age, Na, Cu, P, Co, Zn/Cu, Zn/Cd, Zn/Hg} | |

For the cutoff 50 dataset in Fig. 4, the *F*-measure using our proposed method was improved by more than 6 % compared to cutoff 25 dataset. Furthermore, it achieved better *F*-measure with the RS. Taking into account that RS generates more than one reduct set, whereas our proposed method generates only one reduct set for each dataset. Increasing the number of reduct consequently increase the number of generated rules.

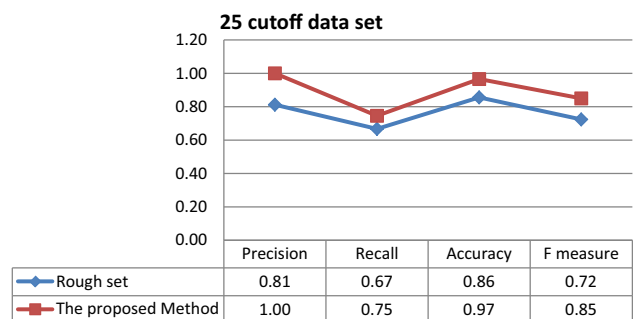


Fig. 3 Classification comparisons (cutoff 25 without BPSO)

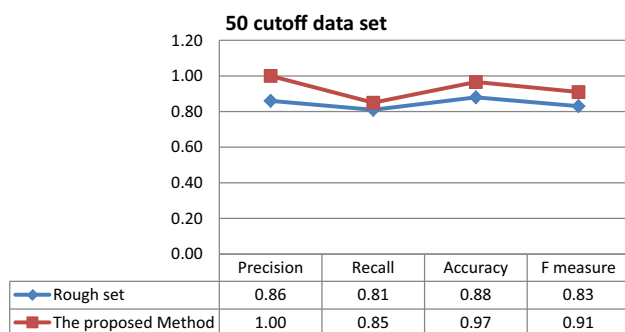


Fig. 4 Classification comparisons (cutoff 50 without BPSO)

Table 14 Classification results (cutoff 25 without BPSO)

| Classifier | Precision | Recall | <i>F</i> -measure | Accuracy |
|------------|-----------|--------|-------------------|----------|
| Bayes net | 0.719 | 0.574 | 0.613 | 0.696 |
| MLP | 0.594 | 0.489 | 0.531 | 0.675 |
| Bagging | 0.692 | 0.681 | 0.686 | 0.763 |
| C4.5 | 0.683 | 0.660 | 0.670 | 0.675 |
| BPSO + RS | 0.1 | 0.756 | 0.85 | 0.966 |

Table 15 Classification results (cutoff 50 without BPSO)

| Classifier | Precision | Recall | <i>F</i> -measure | Accuracy |
|------------|-----------|--------|-------------------|----------|
| Bayes net | 0.761 | 0.872 | 0.813 | 0.815 |
| MLP | 0.727 | 0.638 | 0.680 | 0.812 |
| Bagging | 0.829 | 0.745 | 0.777 | 0.819 |
| C4.5 | 0.709 | 0.553 | 0.621 | 0.742 |
| BPSO + RS | 1.0 | 0.85 | 0.917 | 0.974 |

Moreover, in our proposed method, the generated reduct set contains only a core features in both datasets.

Another experiment is conducted to compare the classification performance of our proposed model with the other classifiers without the first reduct step with BPSO. Table 14 shows the comparison for cutoff 25, as we notice that our method is also superior performance over the best classifier in term of precision Bayes net by 28 % approximately and the best classifier in term of *F*-measure bagging by 19 %.

Table 15 shows the comparison for cutoff 50, although the Bayes net at the cutoff 50 dataset performs better than our method, the *F*-measure achieved 91 %, whereas Bayes net achieved 81 %, which indicates that the overall performance of our method is better in all comparisons.

7 Discussion

This study introduced a new technique for dimension reduction by integrating the RST and BPSO. Although the rough set is a powerful dimensionality reduction technique,

nevertheless when the dimensionality of data increased, rough set approaches fail to find optimal reductions. Moreover, the existing hybrid techniques using rough set and BPSO tended to use the BPSO as a search technique to enhance the reduction algorithm of rough set. In our approach, we applied BPSO as a first stage reduct to reduce the dimensionality of the data then use rough set as a second stage reduction technique to combine the strength of both techniques.

The introduced predictive model for vitamin D deficiency assists in predicting as well as introduces cast light upon more clinical investigation in the association between vitamin D deficiency and a set of minerals derived from hair analysis. The proposed model compared with four different machine-learning classification paradigms and validated using four different common performance measures. We achieved improvement in all the four performance measurements in cutoff 25 dataset by more than 33, 16, 24, and 33 % in precision, recall, *F*-measure, and accuracy, respectively. Also for cutoff 50, the achievement was comparable as well.

Moreover, the performance measurements achieved without using the first stage reduction technique with BPSO in cutoff 25 shows significant improvement by using this stage with 19, 8, 19, and 13 % in precision, recall, *F*-measure, and accuracy, respectively. Further, the set of attributes generated for each reduct set in our study is in concordance with the literature that indicates the clinically relevant information of our model. When we compare our proposed model with the other classifiers without the first reduct step with BPSO, most of them had significantly worse performance compared with the proposed model. The dimensionality reduction achieved scored 76 % in cutoff 25 and 82 % in cutoff 50 confidently demonstrated that our predictive model could be used as an alternative technique that could be applicable in clinical practice. The selected features convinced that the prediction is based on reasonable justifications as will be shown in the following discussion from the expert point of view.

The set of attributes generated for each reduct set of vitamin D in our study is in concordance with the literature. The selected attributes of the two vitamin D reduct set reflects the associations between serum vitamin D and the minerals shown as attributes. The attributes Zn, Cu, Fe, and Hg appeared in sets 25 and 50 nmol/L cutoffs which may indicate a stronger relation between vitamin D and these minerals. However, relations of vitamin D with the other minerals selected as attributes have also been reported in the literature. Earlier findings have shown that at least one aspect of this relation is that vitamin D has a positive effect on the intestinal absorption of a number of minerals such as calcium, magnesium, iron, zinc, cadmium, aluminum, and cobalt [38]. Although further investigation is needed to explore the full picture of the interconnections vitamin D

posses with other nutrients, the literature, with no doubt, shows that many ailments are associated with deficiencies in a number of nutrients. A study on women with primary ovarian insufficiency showed that most of them were deficient in vitamin D, zinc, and copper, all of which are needed for a normal immune function [39]. Another study showed that children with ADHD are more likely to be deficient in vitamin D, magnesium, iron, and zinc [40]. The following subheadings explain in detail the interrelationships between vitamin D and the reduct attributes.

7.1 Vitamin D, calcium, and magnesium

The appearance of calcium and magnesium as attributes was strongly expected due to the strong association between vitamin D and these two minerals. Calcium is found mainly in our bones where it is stored and provides density to the bone. However, it is also essential for other functions as it regulates cell membrane permeability to control nerve impulse transmission and muscle contraction. It promotes blood clotting, and it regulates hormonal secretion and cell division [19]. Vitamin D can indirectly increase bone density by increasing the intestinal absorption of calcium and magnesium [41]. However, a combination of both vitamin D and calcium supplementation has been associated with a higher bone mineral density and decreased risk of hip fractures [42].

Magnesium, on the other hand, is a cofactor in over 500 enzymes that regulate sugar metabolism, energy production, cell membrane permeability, muscle relaxation, and nerve conduction [18]. It also plays an essential role in the synthesis and metabolism of vitamin D. It is required to transport vitamin D by binding it to its transport protein and to activate vitamin D into its biological form in the liver and the kidneys [43]. Magnesium deficiency may activate the catabolism of the active vitamin D in the kidney resulting in the loss of the biologically active form of vitamin D [43]. Dietary magnesium intake has been shown to reduce risks of vitamin D deficiency and sufficiency in a cohort of the National Health and Nutrition Examination Survey (NHANES) 2001–2006 [44]. Their results show that the association between cardiovascular mortality and serum vitamin D deficiency and insufficiency can be modified by the intake of magnesium.

Further, the calcium magnesium ratio is a better indicator of the level of calcium and magnesium. This is because they both cooperate in bone mineral deposition, nerve impulse, and muscle movement, to name a few. This also means that the insufficiently available mineral can prevent the action of the sufficient mineral. This finding is consistent with the outcome of this research results as Ca/Mg appeared in the reduct set.

7.2 Vitamin D, zinc, and copper

Both zinc and copper appeared as attributes in the 25 nmol/L reduct set and their ratio in the 50 nmol/L set, which reflects the association between these two minerals and vitamin D. Studies have shown that the zinc to copper ratio and vitamin D are indicative of the immune status.

Zinc is required in over 200 enzymes in the human body. It plays a role in the immune function, wound healing, protein synthesis, and cell division. Copper is required to build healthy skin, arteries, pigments of the skin and hair, blood formation, bone formation, estrogen synthesis, and energy production for neurotransmitters and is an antimicrobial [18].

Vitamin D increases zinc absorption [45] and copper absorption [38] via the intestines. Several studies suggest a strong relation between vitamin D, zinc and copper in a way that an imbalance in one could affect the function of the others [46]. A study on women with primary ovarian insufficiency showed that most of them were deficient in vitamin D, zinc, and copper [39].

Zinc and copper have antagonistic properties; therefore, their ratio (Zn/Cu) it is worth determining when investigating the health consequences of their imbalance [18].

7.3 Vitamin D, iron, and cobalt

Iron appeared in both of the reduct datasets, on its own in one set and as a ratio with mercury in the other. Iron is required in the synthesis of hemoglobin, the oxygen transporting protein in the blood. Iron is also required for detoxification and energy production in the cells. Cobalt, on the other hand, is essential for life as part of the vitamin B12 molecule. Vitamin B12 is required for blood formation and the nervous system. Deficiencies in both iron and vitamin B12 cause anemia [18].

In chicks, vitamin D has been found to increase the intestinal absorption of iron and cobalt only when dietary calcium was low [47].

7.4 Vitamin D, potassium, and sodium

Potassium appeared in both reduct sets, alone in the 50 nmol/L and as a ratio with sodium in the 25 nmol/L set. Both sodium and potassium are essential electrolytes which maintain proper osmotic pressure of body fluids. Sodium tends to increase water retention, while potassium decreases it [48]. Because these minerals work together, fluid retention is dependent on the ratio of sodium to potassium (Na/K). A higher ratio favors fluid retention.

7.5 Vitamin D, selenium, and sulfur

Selenium was one of the reduct attributes in the 50 nmol/L, while sulfur was an attribute of the 25 nmol/L reduct set. Selenium is required for thyroid activity, protein synthesis, and antioxidant production, namely (glutathione peroxidase). It is essential for detoxification, especially from cadmium and mercury. It is also required for the development of certain higher brain centers [18]. Vitamin D has been found to increase the intestinal absorption of selenium [38].

7.6 Vitamin D, aluminum, cadmium, and mercury

The appearance of these toxic metals in both reduct sets indicates an association between vitamin D and heavy metal contamination, the latter having a profound effect on the state of health and degree of illness.

Aluminum is associated with memory impairment and dementias such as Alzheimer's disease. Aluminum is widely used in beverage cans, aluminum foils, antiperspirants, antacids, and aluminum cookware. Table salt and baking agents may contain aluminum.

Cadmium can cause high blood pressure, heart disease, cancer, fatigue, arthritis, violence, infections, back pain, and other conditions.

Mercury can contribute to hypothyroidism, impaired immune system, and digestive problems such as yeast infections, emotional difficulties, learning disabilities, ADHD, and other nervous and developmental conditions. There is a two-way direction relation between vitamin D and toxic metals. Earlier studies have shown that higher serum vitamin D levels can increase the intestinal absorption of toxic metals like aluminum, cadmium, and lead [41]. Aluminum and cadmium, in turn, can interfere with the metabolism of vitamin D by blocking the renal synthesis of the activated form of vitamin D [1, 25(OH)D] [41]. The intestinal absorption of mercury, however, was not found to increase with vitamin D intake [47].

It is worth noting that much of the studies on toxic metals monitoring have been measured in the blood and urine. These toxicants, however, accumulate in tissues such as the bone, brain, and liver [49]. Thus, such studies may not show the accurate and absolute relation. This is why using a tissue like hair is a better indicative of the actual burden.

7.7 Vitamin D and age

Age is commonly associated with serum 25OHD [3]. This association in our group of healthy women was strong enough to appear as an attribute in the 50 nmol/L cutoff dataset, but not the 25 nmol/L, possibly because higher

levels of serum 25OHD were obtained by older women than younger.

The photosynthesis of vitamin D from its precursor, cholesterol, is affected by the skin's capacity to convert the precursor in the presence of ultraviolet-B. This efficiency declines with age, and therefore, elderly individuals require more vitamin D international units; i.e., 800 IU for ages <70 years [50], presumably because their skin cannot generate enough. Therefore, it is expected that vitamin D deficiency is higher among the elderly, and a number of studies have confirmed this [3]. In our study, however, the age group was limited to 47 years old. The older women in our study were found to be relatively health conscious compared to the younger girls in their early 20s. This may explain why lower levels of vitamin D were obtained by the younger girls and that vitamin D levels were associated positively with age in our study group.

8 Conclusion and future work

In medical field, it is important to understand from a wider perspective, the etiology of a disease as it has a profound effect on the outcome of the treatment. Taking into consideration the interrelationships between nutrients that are associated with a health problem can accelerate the treatment and enhance the quality of life.

This article proposed a classification model that combines the strength of RS theory and PSO to tackle the problem of predicting vitamin D deficiency in apparently healthy Kuwaiti women based on hair mineral analysis. This combination has greatly reduced the feature dimension while improving the classification performance as well as the outcomes concordance with the recent clinical literature. The proposed classification approach has been applied on a dataset gathered from apparently healthy Kuwaiti women by analyzing their hair samples to determine their intracellular mineral levels as an indicator of vitamin D deficiency. The experimental results showed how the integration approaches increased the predictive relationship between vitamin D deficiency and mineral levels extracted from the hair samples. The proposed approach can be employed by specialists when deciding the clinical regimen for their patients. As a future work, we will use association rules mining techniques to figure out the relations between the different minerals within the dataset.

A possible future research investigation could be the use of other feature selection techniques as a first stage reduction technique like GA. Moreover, we will investigate the association rules dependency between different minerals.

Acknowledgments The authors are gratefully thanks the Public Authority for Applied Education and Training (PAAET), Kuwait for providing the dataset used in this study.

References

- Bassil D, Rahme M, Hoteit M, Fuleihan GE-H (2013) Hypovitaminosis D in the Middle East and North Africa. *Derm Endocrinol* 5(2):274–298
- Mishal A (2001) Effects of different dress styles on vitamin D levels in healthy young Jordanian women. *Osteoporos Int* 12(11):931–935
- Siddiqui A, Kamfar H (2007) Prevalence of vitamin D deficiency rickets in adolescent school girls in Western region, Saudi Arabia. *Saudi Med J* 28(3):441–444
- Alyahya K, Lee WTK, Al-Mazidi Z, Morgan J, Lanham-New S (2014) Risk factors of low vitamin D status in adolescent females in Kuwait: implications for high peak bone mass attainment. *Arch Osteoporosis* 9:1–11
- Holick MF (2007) Vitamin D deficiency. *N Engl J Med* 357(3):266–281
- Javaid M, Crozier S, Harvey N, Gale C, Dennison E, Boucher B, Arden N, Godfrey K, Cooper C (2006) Maternal vitamin D status during pregnancy and childhood bone mass at age 9 years: a longitudinal study. *Lancet* 367(9504):36–43
- Autier P, Boniol M, Pizot C, Mullie P (2014) Vitamin D status and ill health: a systematic review. *Lancet Diabetes Endocrinol* 2:76–89
- Kulie T, Groff A, Redmer J, Hounshell J, Schragar S (2009) Vitamin D: an evidence-based review. *JABFM* 22(6):698–706
- Reid IR, Bolland MJ, Grey A (2014) Effects of vitamin D supplements on bone mineral density: a systematic review and meta-analysis. *Lancet* 383:146–155
- Watts D (1990) Nutrient interrelationships minerals–vitamins–endocrines. *J Orthomol Med* 5(1):11–19
- Nguyen T, Khosravi A, Creighton D, Nahavandi S (2015) Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Syst Appl* 42(4):2184–2197
- Srimani PK, Koti MS (2011) The impact of rough set approach on medical diagnosis for cost effective feature selection. *IJCR* 3(12):175–178
- Inbarania HH, Azarb AT, Jothi G (2014) Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Comput Methods Progr Biomed* 113:175–185
- Macas M, Bhondekar AP, Kumar R, Kaur R, Kuzilek J, Gerla V, Lhotská L, Kapur P (2014) Binary social impact theory based optimization and its applications in pattern recognition. *Neurocomputing* 132:85–96
- Kumar SU, Inbarani HH (2015) A novel neighborhood rough set based classification approach for medical diagnosis. *Proc Comput Sci* 47:351–359
- Rzepiński T (2014) Randomized controlled trials versus rough set analysis: two competing approaches for evaluating clinical data. *Theor Med Bioeth* 35(4):271–288
- Guo S, Lucas RM, Ponsonby A-L, The Ausimmune Investigator Group (2013) A novel approach for prediction of vitamin D status using support vector regression. *PLoS One* 8(11):e79970
- Wilson L (2010) Nutritional balancing and hair mineral analysis, 4th edn. LD Wilson Consultants, Inc, Prescott
- Ahmad G, Kuhl H, Mohit A (2013) A review hair tissue analysis: an analytical method for determining essential elements, toxic elements, hormones and drug use and abuse. *Int Res J Appl Basic Sci* 4(11):3675–3688
- Qayyum MA, Shah MH (2014) Comparative study of trace elements in blood, scalp hair and nails of prostate cancer patients in relation to healthy donors. *Biol Trace Elem Res* 162:46–57
- Pawlak Z (1982) Rough sets. *Int J Comput Inf Sci* 11:341–356
- Talbi E-G, Jourdan L, Garcia-Nieto J, Alba E (2008) Comparison of population based metaheuristics for feature selection: application to microarray data classification. In: *Proceedings of AICCSA 2008*
- Kennedy J, Eberhart RC (1999) The particle swarm: social adaptation in information processing systems. In: Corne D, Dorigo M, Glover F (eds) *New ideas in optimization*. McGraw-Hill, London
- Shi Y, Eberhart RC (1998) Parameter selection in particle swarm optimization. In: *Evolutionary programming VII: proceedings of the 7th annual conference on evolutionary programming*, New York, pp 591–600
- Kennedy J, Eberhart RC (1997) A discrete binary version of the particle swarm algorithm. In: *Proceedings of the world multi conference on systemics, cybernetics and informatics*, Piscataway, pp 4104–4109
- Wang X, Yang JXT, Xia W, Jensen R (2007) Feature selection based on rough sets and particle swarm optimization. *Pattern Recogn Lett* 28:459–471
- Inbarani HH, Banu PKN, Andrews S (2012) Unsupervised hybrid PSO-quick reduct approach for feature reduction. In: *Proceedings of international conference on recent trends in information technology*, pp 11–16
- Zuhtuogullari K, Allahverdi N, Arıkan N (2013) Genetic algorithm and rough sets based hybrid approach for reduction of the input attributes in medical systems. *Int J Innov Comput Inf Control* 9(7):3015–3037
- Saini S, Rambli DRA, Zakaria N, Sulaiman S (2014) A review on particle swarm optimization algorithm and its variants to human motion tracking. *Math Probl Eng*. doi:10.1155/2014/704861
- Zhong N, Skowron A (2001) Rough sets based knowledge discovery process. *Int J Appl Math Comput Sci* 11(3):603–619
- Bazan JG, Szczuka M (2005) The rough set exploration system, transactions on rough sets III. Springer, Berlin, pp 37–56
- Group of logic in poland. RSES. <http://logic.mimuw.edu.pl/~rses>
- Own HS, Hassanien AE (2009) Rough sets: a medical knowledge discovery (RS-MKD) scheme for prostate cancer data analysis. *Int J Comput Sci Softw Technol (IJCSST)* 2(2):999–1005. <http://iom.nationalacademies.org/>
- Lamberg-Allardt C, Brustad M, Meyer H, Steingrimsdottir L (2013) Vitamin D—a systematic literature review for the 5th edition of the nordic nutrition recommendations. *Food Nutr Res* 5:57
- Wang Q (2014) A hybrid sampling SVM approach to imbalanced data classification, abstract and applied analysis. Hindawi Publishing Corporation, Cairo. doi:10.1155/2014/972786 (SD-008)
- Own HS, Abraham A (2012) A new weighted rough set framework based classification for egyptian neonatal jaundice. *Appl Soft Comput Elsevier* 12(3):999–1005
- Schwalfenberg GK, Genius SJ (2015) Vitamin D, essential minerals, and toxic elements: exploring interactions between nutrients and toxicants in clinical medicine. *Sci World J*. doi:10.1155/2015/318595
- Kebapcilar A, Kulaksizoglu M, Kebapcilar L, Gonen M, Ünlü A, Topcu A, Demirci F, Taner C (2013) Is there a link between premature ovarian failure and serum concentrations of vitamin D, zinc, and copper? *Menopause* 20(1):94–99
- Villagomez A, Ramtekkar U (2014) Iron, magnesium, vitamin D, and zinc deficiencies in children presenting with symptoms of attention-deficit/hyperactivity disorder. *Children* 1(3):261 (1994)

41. Moon J (1994) The role of vitamin D in toxic metal absorption: a review. *J Am Coll Nutr* 13(6):559–564
42. Rodriguez-Martinez MA, Garcia-Cohen EC (2002) Role of Ca(2+) and vitamin D in the prevention and treatment of osteoporosis. *Pharmacol Therapeutics* 93(1):37–49
43. Zittermann A (2013) Magnesium deficit—overlooked cause of low vitamin D status? *BMC Med* 11:229
44. Deng X, Song Y, Manson JE, Signorello LB, Zhang SM, Shrubsole MJ, Ness RM, Seidner DL, Dai Q (2013) Magnesium, vitamin D status and mortality: results from US National Health and Nutrition Examination Survey (NHANES) 2001 to 2006 and NHANES III. *BMC Med* 11(187):1741–7015
45. Worker NA, Migicovsky BB (1961) Effect of vitamin D on the utilization of zinc, cadmium and mercury in the chick. *J Nutr* 75:222–224
46. Kechrid Z, Hamdi M, Naziroglu M, Flores-Arce M (2012) Vitamin D supplementation modulates blood and tissue zinc, liver glutathione and blood biochemical parameters in diabetic rats on a zinc-deficient diet. *Biol Trace Elem Res* 148(3):371–377
47. Masuhara T, Migicovsky BB (1963) Vitamin D and the intestinal absorption of iron and cobalt. *J Nutr* 80:332–336
48. Terry J (1994) The major electrolytes: sodium, potassium, and chloride. *J Intraven Nurs* 17(5):240–247
49. Archibeque-Engle SL, Tessari JD, Winn DT, Keefe TJ, Nett TM, Zheng T (1997) Comparison of organochlorine pesticide and polychlorinated biphenyl residues in human breast adipose tissue and serum. *J Toxicol Environ Health* 52:285–293
50. Ross AC (2011) The 2011 report on dietary reference intakes for calcium and vitamin D. *Public Health Nutr* 14(5):938–939