Intelligent Predictive Osteoporosis System

Walid MOUDANI
Lebanese University
Doctorate School of Sciences and Technologies
Tripoli, Lebanon

Ahmad SHAHIN
Lebanese University
Doctorate School of Sciences and Technologies
Tripoli, Lebanon

Fadi CHAKIK
Lebanese University
Doctorate School of Sciences and Technologies
Tripoli, Lebanon

Dima RAJAB
Lebanese University
Doctorate School of Sciences and Technologies
Tripoli, Lebanon

ABSTRACT

The healthcare environment is generally perceived as being information rich yet knowledge poor. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. The information technology may provide alternative approaches to Osteoporosis disease diagnosis. In this study, we examine the potential use of classification techniques on a massive volume of healthcare data, particularly in prediction of patients that may have Osteoporosis Disease (OD) through its risk factors. For this purpose, we propose to develop a new solution approach based on Random Forest (RF) decision tree to identify the osteoporosis cases. There has been no research in using the afore-mentioned algorithm for Osteoporosis patients' prediction. The reduction of the attributes consists to enumerate dynamically the optimal subsets of the reduced attributes of high interest by reducing the degree of complexity. A computer-aided system is developed for this purpose. The study population consisted of 2845 adults. The performance of the proposed model is analyzed and evaluated based on set of benchmark techniques applied in this classification problem.

Keywords

Osteoporosis Disease, Multi-Classifier Decision Trees, Prediction, features reduction.

1. INTRODUCTION

Osteoporosis is a real public health problem because of its increasing frequency over the countries. It becomes an essential index of health and economics in every country. Osteoporosis disease is a chronic complex health problem for millions of women worldwide, 80% of whom are postmenopausal, unless prevented or treated, this silent disease will continue to limit both the quantity and the quality of many older women and significantly add to health care cost for this group [1, 2]. This disease infects 30% of women after 50 years and 70% after 80 years. Osteoporosis prevention is complicated but it holds promise as the best way to decrease future fractures [4]. Looking around the world, we see that osteoporosis occurs in some areas much more than in others — just as the incidence of cancer, heart disease, and diabetes varies from one culture to another. This clarifies that the development of weak bones is not a natural artifact of aging. While the United States has one of the highest osteoporosis rates in the world, there are other areas where this disorder is relatively rare, even among the older segments of the population [9]. For example, the inhabitants of Singapore, Hong Kong, and certain sectors of former Yugoslavia, as well as the Bantu of South Africa have traditionally held extremely low rates of osteoporotic fracture. In Japan, vertebral compression fractures among women between ages 50 and 65 were so rare that many physicians doubt their existence, and the incidence of hip fractures among the elderly Japanese historically has been much less than half that of Western countries [23, 24, 25]. Africans and native peoples living traditional lifestyles have been classified as "almost immune" to osteoporosis [5]. Interestingly enough, as these less technologically advanced countries become more Westernized, their rates of osteoporotic fracture are steadily increasing [1]. We note that some Lebanese studies have showed that the mean BMD for the Lebanese female is lower than that of the European woman. Another Lebanese study showed that the hip fractures occur at a younger age in Lebanon (between 65 and 75) compared to western population (above 75) and that 60% of patients with hip fractures have osteopenia rather than osteoporosis [1, 2]. The social economic burden of osteoporosis is so large that its etiology, prevention and treatment have become an urgent issue that needs to be coped with worldwide.

Osteoporosis is a bone disease that commonly occurs among postmenopausal women. Recognizing population with high risks of osteoporosis remains a difficult challenge. Early detection and diagnosis is the key for prevention but are very difficult, without using costly diagnosing devices, due to complex factors involved and its gradual bone lose process with no obvious warning symptoms. Building an Osteoporosis prediction system using data mining techniques based on analyzing postmenopausal risk factors is the aim of this study. By discovering the osteoporosis disease warehouses for Osteoporosis, significant patterns can be extracted in order to build a robust disease prediction models that aim to guide medical decision making and provide an easier way to detect if a person can have the risk of an osteoporosis. The aim of this study is to examine the potential use of classification on a massive volume of healthcare data, particularly in prediction of patients that may have Osteoporosis Disease (OD), which unfortunately continues to increase postmenopausal in the whole world, then it will possible to prevent OD through modification of its risk factors. It enables significant knowledge, e.g. patterns, relationships between medical factors related to Osteoporosis disease, to be established.

The methodology used in this study to build the mining predictive model consists of several phases that start with medical-technical environment understanding, data understanding, data preparation, modeling, implementation and evaluation. The environment understanding phase focuses on illustrating the medical and technical parts of this research by defining the osteoporosis disease, introducing its major risk factors which will constitute the input parameters for the mining operation; and in the technical part, we will identify the role of data mining techniques and explain the classification algorithm chosen to be used in this work. Data understanding phase uses the raw of the data, proceeds to understand the data, identify its structure, gain preliminary insights, and detect interesting subsets. Data preparation phase constructs the final dataset that will be fed into the modeling tools. This includes table, record, and attribute selection as well as data cleaning and transformation. The modeling phase selects and applies various techniques, and calibrates their parameters to optimal values. The solution approach applied can predict the likelihood of patients getting with Osteoporosis risk disease while reducing the complexity of the classification process without affecting the solution quality. The implementation phase specifies the tasks that are needed to build and use the models. The results of this study should be helpful to the development of the computer-aided system in the other medical field. The performance of the proposed model is analyzed and evaluated based on set of benchmark techniques applied in classification problems.

The rest of this paper is organized as follows. Section 2 presents in details the osteoporosis disease and all related features such as: risk factors, symptoms, and prevention. Sections 3 and 4 discuss the works found in the literature related to this disease and also the computational techniques applied for solving this task. In Section 5, we present some empirical results we have obtained by applying our alternative approach to build decision trees. Section 6 shows the implementation of the tool. Finally, some conclusions and notes related to future work are given in Section 7.

2. DESCRIPTION AND ANALYSIS OF OSTEOPOROSIS DISEASE

2.1 Definition

Osteoporosis, a skeletal disease characterized by low bone mass (BMD), micro-architectural deterioration of bone tissue and an increasing risk of fracture, represent an enormous public health burden in both economic costs and human suffering (Fig 1). Osteoporosis literally leads to abnormally porous bone that is compressible, like a sponge. This disorder of the skeleton weakens the bone and results in frequent fractures (breaks) in the bones.

According to the National Institute of Arthritis and Musculoskeletal and Skin Diseases, osteoporosis statistics show a greater burden for women in the following ways:

- 68 percent of the 44 million people with osteoporosis risk are women.
- One of every two women over age 50 will likely have an osteoporosis-related fracture in their lifetime. That’s twice the rate of fractures in men — one in four.
- 75 percent of all cases of hip osteoporosis affect women

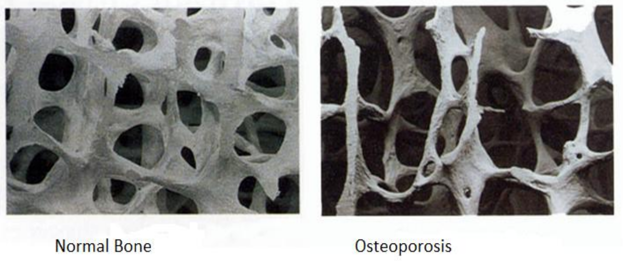


Fig 1: Difference between normal bone and bone osteoporosis

2.2 Symptoms and types of osteoporosis

Osteoporosis can be present without any symptoms for decades because osteoporosis doesn't cause symptoms until bone fractures. Moreover, some osteoporotic fractures may escape detection for years when they do not cause symptoms. Therefore, patients may not be aware of their osteoporosis until they suffer a painful fracture [2, 3]. The symptoms of osteoporosis in men are similar to the symptoms of osteoporosis in women. As the disease progresses, it may have symptoms related to weakened bones, including:

- Back pain
- Loss of height and stooped posture (Fig 2)
- A curved upper back (dowager's hump).

Fig 2: Loss of height and stooped posture caused by osteoporosis

We distinguish three types related to this disease as stated:

- Primary type 1 or postmenopausal osteoporosis: this form of osteoporosis is the most common in women after menopause.
- Primary type 2 osteoporosis or senile osteoporosis: occurs after age 75 and is seen in both females and males at a ratio of 2:1.
- Secondary osteoporosis may arise at any age and affects men and women equally. It results from chronic predisposing medical problems or disease, or prolonged use of medications such as glucocorticoids, when the disease is called steroid-or glucocorticoid-induced osteoporosis (SIOP or GIOP).

2.3 Fractures and Risk factors

Osteopenia is a condition of bone that is slightly less dense than normal bone but not to the degree of bone in osteoporosis. Normal bone is composed of protein, collagen, and calcium, all of which give bone its strength. Bones that are affected by osteoporosis can break (fracture) with relatively minor injury that normally would not cause a bone to fracture. The fracture can be either in the form of cracking (as in a hip fracture) or collapsing (as in a compression fracture of the vertebrae of the spine). The spine, hips, ribs, and wrists are common areas of bone fractures from osteoporosis although osteoporosis-related fractures can occur in almost any skeletal bone. Fragility fractures can affect many sites: Vertebral column, hip, rib and wrist. But the hip fractures are much more numerous, more severe and associated with greater mortality and morbidity.

Concerning the risk factors for osteoporotic fracture, it can be split between non-modifiable and modifiable. Each of them has a relative effect and importance. We can distinguish here two risk factors: Non-modifiable and modifiable factors [1].

Non-modifiable factors:

- Advanced age (in both men and women)
- Female gender
- Estrogen deficiency and early menopause: this deficiency is responsible for a speed increase bone remodeling (Bone remodeling (or bone metabolism) is a life-long process where mature bone tissue is removed from the skeleton (Resorption) and new bone tissue is formed (Formation) and induces an imbalance between resorption and formation, leading to net bone loss.
- Early menopause (before age 45) and any prolonged periods in which hormone levels are low and menstrual periods are absent or infrequent can cause loss of bone mass.
- Heredity: Those with a family history of fracture or osteoporosis are at an increased risk; the heritability of the fracture as well as low bone mineral density are relatively high, ranging from 25 to 80 percent.
- Previous fracture: Those who have already had a fracture are at least twice as likely to have another fracture compared to someone of the same age and sex.
- Rheumatoid disease: Those affected by rheumatoid arthritis may also have an increased risk of developing osteoporosis, a condition in which the bones become less dense and more likely to fracture.

Modifiable factors

- Excess alcohol: small amounts of alcohol do not increase osteoporosis risk but chronic heavy drinking (alcohol intake greater than 3 units/day), especially at a younger age, increases risk significantly [4].
- Tobacco smoking: tobacco smoking inhibits the activity of osteoblasts (cells responsible of formation), and is an independent risk factor for osteoporosis [6]. Smoking also results in increased breakdown of exogenous estrogen, lower body weight and earlier menopause, all of which contribute to lower bone mineral density.
- Vitamin D deficiency: Mild vitamin D insufficiency is associated with increased Parathyroid Hormone (PTH) production that increases bone resorption, leading to bone loss. Also Vitamin D is necessary to absorb calcium, while bodies can synthesize vitamin D from sunlight, in some regions where sunlight is not present for many months at a time, a supplement is necessary.
- Calcium: Calcium is the most abundant mineral in the body; the bones and teeth accounting for about 99% of the total body stores. The main function of calcium is the well-known building and renewal of the skeleton.
- Glucocorticoids: Glucocorticoids are important drugs in the treatment of variety diseases, but long-term period use can lead to various adverse effects, including osteoporosis by inhibition of osteoplastic bone formation, which results not only in decreased bone mineral density, but reduction of bone strength. The evidence suggests that daily oral glucocorticoid doses higher than 5 mg or equivalent increase the risk of fracture within 3–6 months after the start of therapy.

2.4 Osteoporosis Prevention

Effective prevention measures should include non-pharmacologic interventions and pharmacologic when necessary [21, 22, 27].

Non-pharmacologic methods

- Reducing fall risk: fall prevention can help prevent osteoporosis complications. Older patients should be consistently counseled to modify the home environment to improve safety and reduce risk of fall (Removal of obstacles and loose carpets in the living environment, install railings along stairways, etc.) [2, 3, 4].
- Lifestyle: Patients should be educated to minimize their use of alcohol, caffeine and tobacco [5, 6, 22].
- Nutrition: Nutrition plays a critical role in reducing the risk of osteoporosis. An adequate calcium, vitamin D and protein intake resulted in reduced bone remodeling. Supplementation with calcium and vitamin D is a critical component of osteoporosis to improve BMD and to reduce fracture risk. The National osteoporosis foundation recommends that postmenopausal woman consume at least 1200 mg calcium per day [21, 22, 23].
- Physical exercise: A 2 year study showed that adding a physical exercise program to medication improved BMD significantly and is superior to medication alone [5, 6, 7].

Pharmacologic methods

- Estrogen replacement therapy remains a good treatment for prevention of osteoporosis but, at this time, is not recommended unless there are other indications for its use as well. There is uncertainty and controversy about whether estrogen should be recommended in women in the first decade after the menopause [5].
- Some bisphosphonates have been shown to reduce fracture risk after relatively a brief period of use. The scientists found that woman treated with Alendronate (5mg/day) had a lower relative risk for symptomatic vertebral and non-vertebral fractures within 1 year of treatment [1, 2, 22, 23].

3. LITERATURE OVEVIEW

3.1 Description of the methods applied in literature

In the past, the osteoporosis risk was usually modeled from the predominance of one factor. Adinoff and Hollister (see [21]) show that the use of oral glucocorticoids is a major determinant of fractures, while Melton et al. (see [22]) suggest that the bone mineral density (BMD) is the only factor responsible for increasing fracture risk. But recently, a great deal of research has taken place to identify factors other than BMD that contribute to fracture risk i.e. age, a previous fracture, heredity of fracture and lifestyle such as physical exercises, smoking and alcohol. In [23], a study has been validated in Asia, Europe, the United States and Latin America. The results classified the risk level into high, moderate or low. This indexation is based only on two factors: age and body weight by a series of statistical calculations: 0.2 x [(body weight in kg) – (age in years). In [24], Sen et al. have proposed one of the most important studies called the "Osteorisk" risk assessment tool. The sensitivity of this method reaches 94% and the specificity, 45%. Results given help doctors to identify patients who are at greater risk of low bone mass and request examinations of higher complexity, and even begin therapy if it is impossible to undertake such examinations or to avoid unnecessary tests for patients at low risk. A similar study on older woman [25] was based on 6252 women with 65 years or more, compares the value of FRAX models [8] that include BMD with that of parsimonious models based on age and BMD alone for prediction of fractures, also a comparison between FRAX models without BMD with simple models based on age and fracture history alone. The calculation uses the logistic regression to examine receiver operating characteristic (ROC) curves for each model across a range of sensitivities and specificities, then the area under the curve (AUC) statistics from (ROC) curve analysis were compared between FRAX tool and simple models. Since results show no difference between models and FRAX values, this suggests that both the FRAX models and simple models are limited in their ability to predict fracture in older women. To be noted that in the context of osteoporosis, there are two tools other than FRAX for fracture risk calculation, QFractur (www.qfracture.org) and the Garvan tool (www.garvan.org.au). In [18], the GLOW study shows the ability of predicting fractures using 3 algorithms: FRAX, Garvan and a simple model of age and fracture history. The analysis found that the estimation of fracture risk of postmenopausal women can be made using clinical risk factors alone, without BMD models incorporating multiple clinical risk factors including falls, were not superior to more parsimonious models in predicting future fracture in this population.

Yildirim et al. (see [27]) present the importance of osteoporosis disease in terms of medical research and pharmaceutical industry. They introduce a knowledge discovery approach regarding the treatment of osteoporosis from a historical perspective. They propose to use a freely available biomedical search engine leveraging text-mining technology to extract the drug names used in the treatment of osteoporosis from MEDLINE articles. They conclude that alendronate (Fosamax) and raloxifene (Evista) have the highest number of articles in MEDLINE and seem the dominating drugs for the treatment of osteoporosis in the last decade.

3.2 Description of the classification techniques

3.2.1 Classification based on Neural Network and ensemble data mining approach

Predicting osteoporosis in not limited on clinical factors but may be also based on intelligent models using several techniques of data mining such neural network by Chui et al. [25] where the model was developed and validated as an artificial neural network (ANN) to identify the osteoporotic subjects in the elderly. After training processes, the final best ANN was a multilayer perceptron network which determined seven input variables (gender, age, weight, height, body mass index, postmenopausal status, and coffee consumption) as significant features. The discriminatory power of ANN for test set (AUC) was excellent.

Wang and Rea (see [26]) present the research in developing an ensemble of data mining techniques for predicting the risk of osteoporosis prevalence in women. It consists to develop an intelligent decision support system based on data mining ensemble technology to assist General Practitioners in assessing patient's risk of developing osteoporosis. It focuses on investigating the methodologies for constructing effective ensembles, specifically on the measurements of diversity between individual models induced by two types of machine learning techniques, i.e. neural networks and decision trees for predicting the risk of osteoporosis. The constructed ensembles as well as their member predictors are assessed in terms of reliability, diversity and accuracy of prediction. The results indicate that the intelligently hybridized ensembles have high-level diversities and thus are able to improve their performance.

Methods have been explored in attempting to build better ensembles by trying either to generate more accurate models or to create more diverse models or both ideally. Boosting [28] and Adaboots [29] are two useful techniques to manipulate the data set by adding more weight to so called “hard” data subsets to force the models to learn the aspects represented by these weighted training data subsets. In addition, the decision fusion strategies also play important role in determining the performance of an ensemble. Averaging and simple or weighted voting, are the tow commonly used ones, pending the type of machine learning algorithms employed. For one, such as neural networks, outputs continuous value, averaging seems naturally suitable but the diversity, if not carefully handled, may have some adverse effects on the final averaged result. Voting strategy is best suitable for the modeling algorithms with categorical outputs, such as decision trees for classification problems. However, the continuous outputs can be discretized; the voting can then be applied.

3.2.2 Classical decision trees

It is well-known that decision trees are probably the most popular classification model [14, 15, 17]. The aim of the decision tree learning process is to build a decision tree which conveys interesting information in order to make predictions and classify previously unseen data. In order to apply classification tree, we should separate data into 2 groups: attributes (inputs or predictors) and class (output or response). Decision Trees algorithms usually assume the absence of noise in input data and they try to obtain a perfect description of data. This is usually counterproductive in real problems, where management of noisy data and uncertainty is required. The Decision Tree algorithm family includes classical algorithms, such as CLS (Concept Learning System), ID3 [17], C4.5 and CART (Classification and Regression Trees) [18], as well as more recent ones, such ART [17] and RF [10]. Some of those algorithms build binary trees, while others induce multi-way decision trees. However, when working with numerical attributes, most Decision Trees algorithms choose a threshold value in order to perform binary tests. The particular tests which are used to branch the tree depend on the heuristics used to decide which ones will potentially yield better results. Every possible test which splits the training dataset into several subsets will eventually lead to the construction of a complete decision tree, provided that at least two of the generated subsets are not empty. Each possible test must be evaluated using heuristics and, as most Decision Trees algorithms perform a one-ply look ahead heuristic search without backtracking (i.e. they are greedy), the selected heuristics plays an essential role during the learning process. For instance, most Decision Trees algorithms decide how to branch the tree using some measure of node impurity. Such heuristics, splitting rules henceforth, are devised to try to obtain the "best" decision tree according to some criterion. The objective is usually to minimize the classification error, as well as the resulting tree complexity. Several splitting rules have been proposed in the literature. CART [18] uses Gini index to measure the class diversity in the nodes of a decision tree. ID3 [17] attempts to maximize the information gain achieved through the use of a given attribute to branch the tree. C4.5 [18] normalizes this information gain criterion in order to reduce the tree branching factor and [19] adjusts C4.5 criterion to improve its performance with continuous attributes. Lopez de Mantaras [16] proposed an alternative normalization based on a distance metrics. Taylor and Silverman [20] proposed the mean posterior improvement criterion as an alternative to the Gini rule for binary trees. All the above-mentioned criteria are impurity-based functions, although there are measures which fall into other categories: some of them measure the difference among the split subsets using distances or angles, emphasizing the disparity of the subsets (on binary trees, typically), while others are statistical measures of independence between the class proportions and the split subsets, emphasizing the reliability of class predictions. Pruning techniques, used in C4.5, have proved to be really useful in order to avoid over fitting. Those branches with lower predictive power are usually pruned once the whole decision tree has been built.

3.2.3 Multi-Classifier based on decision trees

Multi-classifiers are the result of combining several individual classifiers. When individual classifiers are combined appropriately, we usually obtain a better performance in terms of classification precision and/or speed to find a better solution. Multi-classifiers differ among themselves by their diverse characteristics: the number and the type of the individual classifiers; the characteristics of the subsets used by every classifiers of the set; the consideration of the decisions; and the size and the nature of the training sets for the classifiers. In [35], Segrera divided the methods for building multi-classifiers in two groups: ensemble and hybrid methods. The first type, such as Bagging and Boosting, induces models that merge classifiers with the same learning algorithm, while introducing modifications in the training data set. The second, type such as Stacking, creates new hybrid learning techniques from different base learning algorithms. An ensemble uses the predictions of multiple base classifiers, typically through majority vote or averaged prediction, to produce a final ensemble-based decision. The ensemble-based predictions typically have lower generalization error rates than the ones obtained by a single model. The difference depends on the type of base-classifiers used, ensemble size, and the diversity or correlation between classifiers [36]. Ahn [36] indicates that, over the last few years, three ensemble-voting approaches have received attention by researchers: boosting, bagging and random subspaces. In [10], Breiman defined RF as a classifiers composed by decision trees where every tree h_t has been generated from the set of data training and a vector θ_t of random numbers identically distributed and independent from the vectors θ_1, θ_2,.., θ_{t-1} used to generate the classifiers h_1, h_2, .., h_{t-1}. Each tree provides his unitary vote for the majority class given the entry. Examples of RF are: randomization, Forest-RI (Random Input selection) and Forest-RC (random combination), double-bagging. In [37], Hamza concludes several elements such as: RFs are significantly better than Bagging, Boosting and a single tree; their error rate is smaller than the best one obtained by other methods; and they are more robust to noise than the other methods. Consequently, RF is a very good classification method with the following characteristics: it's easy to use; it does not require models, or parameters to select except for the number of predictors to choose at random at each node.

3.2.4 Description of Random Forest

Nowadays, numerous attempts in constructing ensemble of classifiers towards increasing the performance have been introduced ([10, 34]). Examples of such techniques are Adaboost, Bagging and RFs [34]. RFs have been quite successful in classification and regression tasks [33]. RF is a class of ensemble methods specifically designed for decision tree classifiers [10]. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors and with the same distribution for all trees in the forest (Fig. 3). Each decision tree is built from a random subset of the training dataset. It uses a random vector that is generated from some fixed probability distribution, where the probability distribution is varied to focus examples that are hard to classify. A random vector can be incorporated into the tree-growing process in many ways. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the data class labels. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions. There are three approaches for RFs such as: Forest-RI, Forest-RC, mixed of Forest-RI and Forest-RC. Forest-RI consists to randomly select F input features to split at each node of the decision tree. As a result, instead of examining all the available features, the decision to split a node is determined from these selected F features. The tree is then grown to its entirety without any pruning. This may help reduce the bias present in the resulting tree. Once the trees have been constructed, the predictions are combined using a majority voting scheme. The strength and correlation of RFs may depend on the size of F. if F is sufficiently small, then the trees tend to become less correlated. On the other hand, the strength of the tree classifier tends to improve with a larger number of features, F. As a tradeoff, the number of features is commonly chosen to be $F = \log_2 d + 1$, where d is the number of input features. Since only a subset of the features needs to be examined at each node, this approach helps to significantly reduce the runtime of the algorithm.

Forest-RC is used to create combination of the input features. In case the number of original features *d* is too small, then it is difficult to choose an independent set of random features for building the decision trees. One way to increase the features space is to create linear combination of the input features. Specifically, at each node, a new feature is generated by randomly selecting L of the input features. The input features are linearly combined using coefficients generated from a uniform distribution in the range of [-1, +1]. At each node, F of such randomly combined new features are generated, and the best of them is subsequently selected to split the node.

A third approach for generating the random trees is to randomly select one of the F best splits at each node of the decision tree. This approach may potentially generate trees that are more correlated than Forest-RI and Forest–RC, unless F is sufficiently large. It also does not have the runtime savings of Forest-RI and Forest–RC because the algorithm must examine all the splitting features at each node of the decision tree.

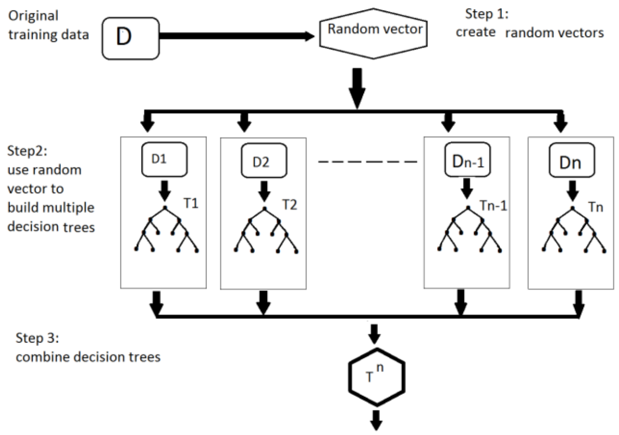

Original training data
D
Random vector
Step 1: create random vectors
Step2: use random vector to build multiple decision trees
D1
D2
Dn-1
Dn
T1
T2
Tn-1
Tn
Step 3: combine decision trees
T^n

Fig 3: Random Forest model

The use of RFs technique has provides some desirable characteristics shown such as: it is unexcelled in accuracy among current algorithms, it runs efficiently on large databases, it is relatively robust to outliers and noise; it is simple and easily parallelized; it is faster than bagging or boosting; it can handle thousands of input variables without variable deletion; it gives estimates of what variables are important in the classification; it generates an internal unbiased estimate of the generalization error as the forest building progresses, it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing, it has methods for balancing error in class population unbalanced data sets, and it computes proximities between pairs of cases that can be used in clustering.

The generalization error of RFs classifiers depends on the strength of the individual trees in the forest and the correlation between them. However, it has theoretically proven that the upper bound for generalization error of RFs converges to the following expression, when the number of trees is sufficiently large.

$$\text{Generalization error} \le \frac{\bar{\rho}\left(1-s^2\right)}{s^2} \qquad (1)$$

where $\bar{\rho}$ is the average correlation among the trees and s is a quantity that measures the strength of the tree classifiers. The strength of a set of classifiers refers to the average performance of the classifiers, where performance is measured probabilistically in terms of the classifier's margin:

$$\mathrm{marg\,in},\ M(X,Y) = P\left(\hat{Y}_{\theta} = Y\right) - \max_{Z \neq Y} P\left(\hat{Y}_{\theta} = Z\right) \qquad (2)$$

where $\hat{Y}_{\theta}$ is the predicted class of X according to a classifier built from some random vector θ. The higher the margin is, the more likely it is that the classifier correctly predicts a given example X. As the trees become more correlated or the strength of the ensemble decreases, the generalization error bound tends to increase. Randomization helps to reduce the correlation among decision trees so that the generalization error of the ensemble can be improved.

4. OSTEOPOROSIS SOLUTION APPROACH AND METHODOLOGY

In this section, we present an intelligent classification solution which is based on dynamic reduced sets `of features while preserving the solution quality. This approach is validated by using RF decision tree classification technique to identify the osteoporosis cases. The study population is composed of 2845 adults.

4.1 Description of the proposed solution

The strategy reported here can be described as a KDD (Knowledge discovery in databases) experiment. Following a typical KDD framework, where Data Mining is the core in the overall process, the experiment went through all steps of Figure 4, starting from the stage of gaining profound knowledge of the domain till the actual use of discovered knowledge. A description of database, source of data, pre-processing steps (cleaning, transformation, and integration) is given here.

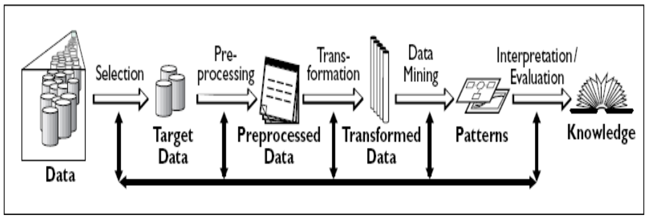


Fig. 4: Methodology roadmap of the KDD process

4.1.1 Data Source

During data collection process and after analysis based on experts' knowledge, a set of collected data related to osteoporosis information for about 2845 patients is established. All records gathered from the real cases are processed by using the FRAX tool (i.e. WHO Fracture Risk Assessment) in order to predict the appropriate risk level. FRAX is a major milestone towards helping health professionals worldwide to improve identification of patients at high risk of fracture. The FRAX algorithms give the 10-year probability of fracture. The output is a web-based calculation tool assesses the ten-year risk probability of hip fracture and the 10-year probability of a major osteoporotic fracture (clinical spine, forearm, hip or shoulder fracture). The FRAX models have been developed from studying population-based cohorts from Europe, North America, Asia and Australia. The osteoporosis risk factors for each patient are defined and saved into a .csv file representing the target dataset for our study.

4.1.2 Data Description

The study is based on a set of relevant features collected and defined after discussion with experts. Table 1 lists the description of features that are significant to osteoporosis disease. The results provided by FRAX are presented as probability values which are normalized based on experts knowledge in order to determine the set of risk level classes (table 2).

Table 1: Osteoporosis factors including in study

Attribute	Type	Description
Age	Numeric	Between 40 and 90 years.
BMI= weight/(height)^2	Numeric	a.Weight:Є[34kg-110kg] b.Height: Є[139cm-185cm]
Previous fracture	Boolean	
Osteoporosis Heredity	Boolean	
Smoking	Boolean	
Glucocorticoids	Boolean	Treatment for more than 3 months at a dose of 5 mg daily or more.
Rheumatoid	Boolean	
Secondary osteoporosis	Boolean	Premature menopause (before 45 years), chronic malnutrition, or malabsorption and chronic liver disease.
Excess alcohol	Boolean	3 unit/day or more
Estrogen	Numeric	premenopausal: 30 to 400 pg/mL; postmenopausal: 0 to 30 pg/mL
Calcium	Numeric	[8.5mg/dl- 10.2 mg/dl].
Vitamin D	Numeric	[30.0 ng/ml - 74.0 ng/ml]
BMD value	Numeric	Normal bone: T-score better than -1 Osteopenia: T-score between -1 and -2.5 Osteoporosis: T-score less than 2.5
Excess caffeine	Boolean	
Immobilization	Boolean	Ex: long immobilization after a fracture

Table 2: Osteoporosis risk level

Class name	Range
No risk	< 5%
Low risk	[5%-20%[
Moderate risk	[20%-40%[
High risk	[40%-50%[
Severe risk	>50%

4.1.3 Data processing

Data are transformed and normalized in order to fit with the requirements of the classification techniques used in our case. Moreover, some data collected to build the database requests to be cleaned, integrated, and normalized to realize the process.

4.2 Classification using RF

In this section, we describe the proposed methodology which is based RF technique in order to predict osteoporosis patients. We examine also the features considered initially in the prediction process and the reduction of these features, leading to generate dynamic equivalence subsets of features without affecting the solution quality. The main concept of the proposed approach is the built of the effective RF multi-classifier decision trees. The accuracy of the prediction can be improved gradually depending on the relevant features. The highest accuracy is somehow associated to acceptable reduced subset(s) of features.

5. RESULTS ANALYSIS AND DISCUSSION

The results of the classification techniques applied in this study are now processed and analyzed in order to compare the relative performance followed by an interpretation, validation and discussion. We have concluded some features stated as below:

- Age and BMI are the most effective attributes that lead to Osteoporosis prediction. The risk level is proportional to the age, inversely proportional to BMI;
- The influence of previous fracture and heredity is highly important, especially when both are available;
- Alcohol or smoking alone has no effects on this disease;
- The risk level 'No Risk' is not available when the age factor is above 80 years.

Moreover, the performance evaluation of the prediction system is based on some parameters such as: Attributes reduction, Misclassification rate, and Accuracy. The results issued by applying several decision trees techniques are summarized in table 3. The variation of misclassification rate between the different techniques shows that the rate of the incorrectly classified instance (0.0007%) is the lowest by using Forest-RC comparing to the other set of techniques presented. ID3 decision tree produces the highest rate of misclassification (0.1543). Therefore, the accuracy rate of Forest-RI is the best among the different techniques presented in table 3. The classification results using RFs are obtained from ten-fold cross-validation. However, we conclude that the initial number of attributes has been reduced while using RFs multi-classifier technique from 16 to 9. The relevant features are only taken into consideration which leads to enhance the complexity of the proposed model by focusing the study based on reduced features. This number is somehow great when using ID3 and J48 techniques.

After analyzing the values of different parameter, the performance of the classification process generates a highly precision while using RF multi-classifier decision trees, especially, when using the variant Forest-RC which provides the highest accuracy rate. For this reason, the proposed Predictive Osteoporosis System (POS) is built based on RFs multi-classifier decision trees in order to build a high accuracy and robustness solution.

Table 3: Results of classification techniques parameters

Classification Methods	Reduced # of attributes (Initial, Reduced)	Incorrectly Classified Instances	Error rate	Accuracy
J48	(16, 11)	15	0.0052	0.9947
ID3	(16, 13)	439	0.1543	0.8456
Forest-RC	(16, 9)	2	0.0007	0.9992
Forest-RI	(16, 9)	4	0.0017	0.9982

In table 4, we show a prototype of instances as they are fed into the Predictive Osteoporosis System (POS). It displays the classification provided by both POS and FRAX tools. As mentioned in the table 2, FRAX classification is normalized based on experts' knowledge in order to define the appropriate risk level. The obtained results show high accuracy prediction using POS. The data used in this study has been obtained while processing several patients' information that covers almost the whole cases validated by experts'. Also, all the results given by POS are validated by physicians, so the output of the model is reliable but this will not exclude some error maybe occur.

Table 4: Comparison of predicting level risk between POS and FRAX

Inst	Input	Output by POS	Output by FRAX	Match
1	48 23.8 0 0 0 0 1 0 0 …	4	14%	Yes
2	55 19.5 1 1 0 0 0 0 0 …	4	6.4%	Yes
3	79 18 1 1 1 0 0 1 0 …	4	12%	Yes
4	82 41.3 0 0 1 0 1 0 1 …	5	20%	Yes
5	42 34.2 0 1 0 1 1 1 0 …	3	3.3%	Yes
6	90 23.6 0 0 1 0 0 0 1 …	7	51%	Yes
7	56 21.3 1 1 1 1 0 1 0 …	3	19%	No
8	86 18.9 0 0 1 0 1 1 1 …	6	46%	Yes
9	63 33.5 0 1 1 0 0 1 0 …	5	13%	No
10	40 21 0 0 1 1 1 0 1 …	4	8.5%	Yes
11	77 15.8 1 0 0 0 1 0 0 …	5	29%	Yes
12	45 23.2 0 0 0 0 1 1 1 …	4	6.3%	Yes
13	87 21.6 0 0 0 0 1 1 1 …	6	46%	Yes
14	76 29 0 0 1 1 1 1 1 …	4	13%	Yes
15	51 17 0 1 0 1 0 1 1 …	3	4.6%	Yes
16	78 19.8 0 0 1 1 1 1 1 …	4	17%	Yes
17	60 23.7 0 0 0 0 0 0 0 …	5	27%	Yes
18	66 28.3 1 1 0 0 0 0 0 …	4	9.4%	Yes
19	42 20.9 0 1 1 1 1 0 0 …	3	4.4%	Yes
20	81 24.1 0 0 0 1 1 1 1 …	5	23%	Yes
…	….	…	…	…

6. IMPLEMENTATION OF POS TOOL

In order to allow the non-technical persons or users of the system to utilize this tool, we have implemented an Osteoporosis Risk Prediction Web based system. It has been developed in such a manner that it can be used easily and in comfortable way without the request to have the support of a technical person. In fact, we have simplified as most as we can the user interface of the system and the way that its user manipulate it. It includes some features to be used such as: predicting the risk level of patients, presenting statistics, showing osteoporosis factors and the prevention. POS is a web application, having the system's engine built based on RFs classification algorithms. It has a simple user interface and allows an access to database in order to store the patients' data and their osteoporosis risk prediction. The interface illustrates a questionnaire that allow patient to use this tool, by filling the questions in order to be informed about his osteoporosis risk level (Fig. 5). This questionnaire resumes all necessary data including in our database.

7. CONCLUSIONS AND PERSPECTIVES

Osteoporosis related data are voluminous in nature and are issued from several sources with not entirely appropriate in structure or quality. Nowadays, the exploitation of knowledge, based on the experience of specialists and the clinical screening data of patients, have been widely recognized. In this paper, we have presented an efficient approach for extracting significant patterns from the osteoporosis disease data warehouses for the efficient prediction of osteoporosis risk. This work has described the research of an effective algorithm to construct a model that can be used in order to predict the risk level of osteoporosis disease when it attacks any woman. We can resume the most important steps as:

- The results issued from POS has not focuses only on informing about the presence of a risk or not, but also it provides the level of the osteoporosis risk for the patient.
- The proposed tool, POS, contributes in managing osteoporosis by reducing the risk of fractures, identifying early the patients, assessing accurately the risk, and improving the patient's perception of that risk.
- The key step is the compilation of representative and expressive data that will cover the large number of cases in order to generalize and extract rules by determining the effect of each attribute.
- Enhancing the complexity by defining the optimal number of relevant attributes that can be used in order to build the model without affecting the solution quality.
- Building the model for prediction of the osteoporosis risk level using multi-classifier decision trees instead of one decision tree.
- Performing an evaluation of the performance of the proposed model based on set of benchmark techniques applied in classification problems such as: RFs and its variants, ID3, J48.

In perspective, we suggest to integrate a set of positive features in order to improve the knowledge quality services, such as:

- Studying the relation with the surrounding countries known as the highest osteoporosis rates in the world in order to compare it with the Lebanese population and to get relevant knowledge.
- Improving the quality of the prediction by applying new classification techniques.
- Applying Text Mining to mine the vast amount of unstructured data available in Osteoporosis databases.
- Providing greater accessibility in order to help physicians make informed about the treatment decisions. POS may become accessible as a real online questionnaire in clinical settings using the Internet.
- Ensuring that high-risk individuals are identified and ultimately leading to the more effective management of patients with osteoporosis.
- Providing fast and portable physician access to the risk calculator by proposing an iPhone POS application aiming to make the diagnostic tool more accessible for patients.

8. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the proposed intelligent osteoporosis prediction tool especially to Dr Wassim DANDACHI and Dr Mountasser AL ASSAAD.

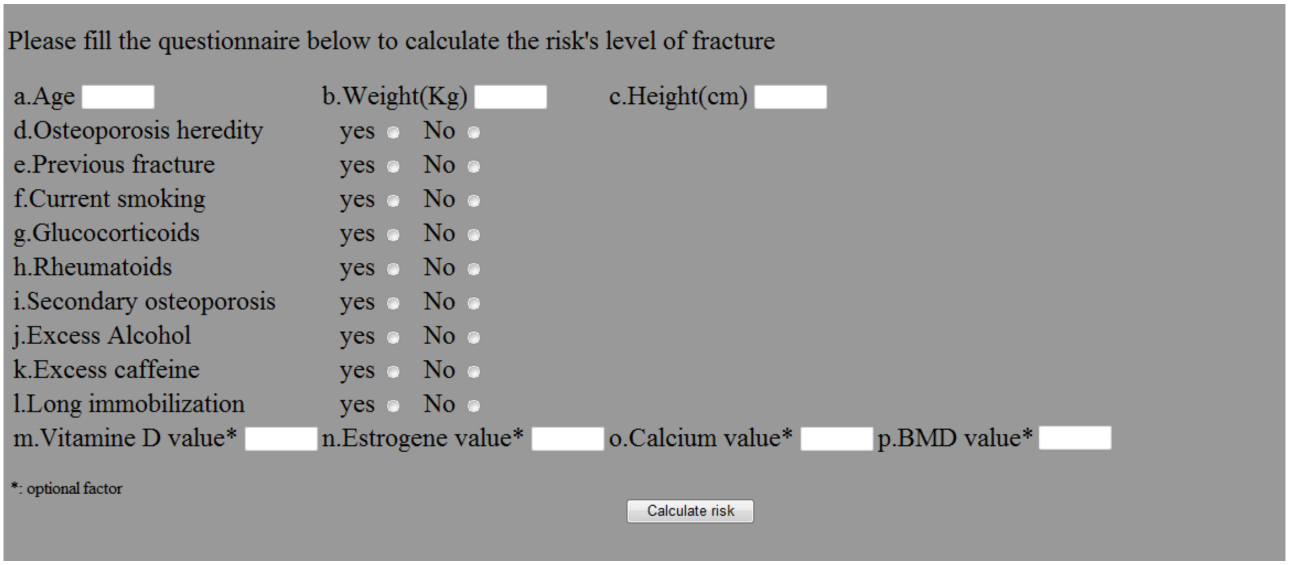
9. REFERENCES

[1] Taylor BC, Schreiner PJ, Stone KL, et al., 2004. Long-term prediction of incident hip fracture risk in elderly white women: study of osteoporotic fractures. J Am Geriatr Soc. 2004;52:1479–1486. [PubMed]

[2] Kanis JA. 2002. Diagnosis of osteoporosis and assessment of fracture risk. Lancet. 2002;359:1929–1936. [PubMed]

[3] Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. 2008. FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int. 2008;19:385–397. [PMC free article][PubMed]

[4] Kanis JA, Johansson H, Johnell O, et al., 2005. Alcohol intake as a risk factor for fracture. Osteoporos Int.2005;16:737–742. [PubMed]



Please fill the questionnaire below to calculate the risk's level of fracture

a.Age □ b.Weight(Kg) □ c.Height(cm) □

d.Osteoporosis heredity yes ○ No ○

e.Previous fracture yes ○ No ○

f.Current smoking yes ○ No ○

g.Glucocorticoids yes ○ No ○

h.Rheumatoids yes ○ No ○

i.Secondary osteoporosis yes ○ No ○

j.Excess Alcohol yes ○ No ○

k.Excess caffeine yes ○ No ○

l.Long immobilization yes ○ No ○

m.Vitamine D value* □ n.Estrogene value* □ o.Calcium value* □ p.BMD value* □

*: optional factor

Fig 5: POS Interface for calculating Osteoporosis Risk Level

[5] De Laet C, Kanis JA, Oden A, et al. 2005. Body mass index as a predictor of fracture risk: a meta-analysis. Osteoporos Int. 2005;16:1330–1338. [PubMed]

[6] Kanis JA, Johnell O, Oden A, et al., 2005. Smoking and fracture risk: a meta-analysis. Osteoporos Int.2005;16:155–162. [PubMed]

[7] Kanis JA, Oden A, Johnell O, et al., 2007. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. Osteoporos Int.2007;18:1033–1046. [PubMed]

[8] FRAX® WHO fracture risk assessment tool [8-13-2008]. http://www.shef.ac.uk/FRAX/

[9] Dawson-Hughes B, Tosteson AN, Melton LJ, III, et al., 2008. Implications of absolute fracture risk assessment for osteoporosis practice guidelines in the USA. Osteoporos Int. 2008;19:449–458.[PubMed]

[10] Leo Brieman. Random Forests. In: Machine Learning, Kluwer Academic Publisher, 45(1), 2001

[11] K. Thangavel, Q. Shen, and A. Pethalakshmi, 2006. “Application of Clustering for Feature selection based on rough set theory approach”, AIML Journal, Vol. 6 (1), pp.19-27.

[12] Moudani W., Shahin A., Chakik F., and Mora-Camino, F., 2011. Dynamic Rough Sets Features Reduction, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9(4).

[13] R.E. Bellman, 1957. Dynamic Programming, Princeton University Press.

[14] F. Berzal, J.C. Cubero, N. Marın, D. Sanchez, Building multi-way decision trees with numerical attributes, Technical report available upon request, 2001.

[15] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, California, USA, 1984, ISBN 0-534-98054-6.

[16] R. Lopez de Mantaras, A distance-based attribute selection measure for decision tree induction, Mach. Learn. 6 (1991).

[17] J.R. Quinlan, Learning decision tree classifiers, ACM Comput. Surveys 28 (1) (1986) 71–72.

[18] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993, ISBN 1-55860-238-0.

[19] J.R. Quinlan, Improved use of continuous attributes in C4.5, J. Artif. Intell. Res. 4 (1996) 77– 90.

[20] P.C. Taylor, B.W. Silverman, Block diagrams and splitting criteria for classification trees, Statist. Comput. 3 (4) (1993) 147–161.

[21] Adinoff AD & Hollister JR. Steroid induced fractures and bone loss in patients with asthma. N. Engl. J. Med. 1983*;* 309*:* 265–8.

[22] Melton LJ 3rd, Kan SH, Wahner HW, Riggs BL (1988) Lifetime fracture risk: an approach to hip fracture risk assessment based on bone mineral density and age. J Clin Epidemiol 41:985–994

[23] Koh LK, Sendrine WB, Torralba TP, et al. A simple tool to identify asian women at increased risk of osteoporosis. Osteoporos Int. 2001;12(8):699-705.

[24] Sen SS, Rives VP, Messina OD, et al. A risk assessment tool (OsteoRisk) for identifying Latin American women with osteoporosis. J Gen Intern Med. 2005;20(3):245-50.

[25] Chiu JS, Li YC, Yu FC, Wang YF, Applying an artificial neural network to predict osteoporosis in the elderly. Osteoporos Int. 2006;124:609-14. [PubMed]

[26] Wang W. and Rea S. Intelligent ensemble system aids osteoporosis early detection. EC'05 Proceedings of the 6th World Scientific and Engineering Academy and Society (WSEAS) international conference on Evolutionary computing, 2005.

[27] Yildirim P., Çeken Ç, Hassanpour R., Esmelioglu S. and Tolun M.R. Mining MEDLINE for the Treatment of Osteoporosis, JOURNAL OF MEDICAL SYSTEMS, DOI: 10.1007/s10916-011-9701-6, April 2011.

[28] Schapire et al (1997): Boosting the margin: A new explanation for the effectiveness of voting methods. In Fisher, D.(Ed) Machine Learning: Proceedings of 14th Int. Conference, Morgan Kaufmann.

[29] Freund, Y. & Schapire R.E. (1996): Experiments with a new boosting algorithm, in L. Saitta, ed., Machine Learning: Proceedings of the 13th national conference, Morgan Kaufmann. pp148-156

[30] A. Hedar, J. Wangy, and M. Fukushima, "Tabu search for attribute reduction in rough set theory", Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications, Springer-Verlag Berlin, Heidelberg, 2008, Vol. 12 (9).

[31] F. Glover and M. Laguna, "Tabu Search", Kluwer Academic Publishers, Boston, MA, USA, 1997.

[32] A. Hedar and M. Fukushima, "Tabu search directed by direct search methods for nonlinear global optimization", European Journal of Operational Research, 2006, Vol. 170, pp. 329–349.

[33] Leo Breiman. Bagging predictors. Machine Learning Journal, 26(2):123140, 1996.

[34] T.K. Ho. 1998. The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(8):832844.

[35] S. Segrera, M. Moreno. Multiclassifiers: applications, methods and architectures. Proc. of InternationalWorkshop on Practical Applications of Agents and Multiagents Systems, IWPAAMS05, 263–271, 2005.

[36] H. Ahn, H. Moon, M.J. Fazzari, N. Lim, J.J. Chen and R.L. Kodell. Classification by ensembles from random partitions of high dimensional data. Computational Statistics and Data Analysis, 51:6166–6179, 2007.

[37] M. Hamza and D. Larocque. An empirical comparison of ensemble methods based on classification trees. Statistical Computation & Simulation 75(8):629-643, 2005.