J Med Syst (2016) 40: 61
DOI 10.1007/s10916-015-0413-1

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Automatic Estimation of Osteoporotic Fracture Cases by Using Ensemble Learning Approaches

Niyazi Kilic[1] · **Erkan Hosgormez**[2]

Received: 10 July 2015 / Accepted: 17 November 2015 / Published online: 12 December 2015

Abstract Ensemble learning methods are one of the most powerful tools for the pattern classification problems. In this paper, the effects of ensemble learning methods and some physical bone densitometry parameters on osteoporotic fracture detection were investigated. *Six* feature set models were constructed including different physical parameters and they fed into the ensemble classifiers as input features. As ensemble learning techniques, bagging, gradient boosting and random subspace (RSM) were used. Instance based learning (IBk) and random forest (RF) classifiers applied to *six* feature set models. The patients were classified into three groups such as osteoporosis, osteopenia and control (healthy), using ensemble classifiers. Total classification accuracy and *f*-measure were also used to evaluate diagnostic performance of the proposed ensemble classification system. The classification accuracy has reached to 98.85 % by the combination of model 6 (*five* BMD + *five* T-score values) using RSM-RF classifier. The findings of this paper suggest that the patients will be able to be warned before a bone fracture occurred, by just examining some physical parameters that can easily be measured without invasive operations.

This article is part of the Topical Collection on *Systems-Level Quality Improvement.*

✉ Niyazi Kilic
niyazik@istanbul.edu.tr

[1] Engineering Faculty, Electrical and Electronics Department, Istanbul University, 34320Avcilar, Istanbul, Turkey

[2] Biomedical Engineering Department, Institute of Sciences, Istanbul University, Beyazit, Istanbul, Turkey

Keywords Osteoporosis · Bone mineral density · T-score · Ensemble learning classification · IBk · Random forest

Introduction

Osteoporosis (OP) is the most common metabolic bone disease in the world. It is a major cause of morbidity and loss of work due to osteoporotic fractures [1–4]. Benhamou et al. [5], define OP as a disease "characterized by low bone mass and micro architectural alterations of bone tissue, leading to enhanced bone fragility and consequent in fracture risk". Osteoporosis disease is also widely seen in the post-menopausal woman due to a remarkable decrease in estrogen levels [6].

OP generally improves without showing any symptoms in its early phases. In the patients with OP, losses of trabecular bone and a consequent weakening of bone structure can be seen [7]. As for Osteopenia (ON), it is the first phase of OP which makes bones weak and fractures them easily. In the diagnosis of the OP or ON, bone mineral density (BMD) and T-score are vital parameters. BMD and T-score values are fundamental part of the evaluation of patients with suspicious osteoporosis. Definition osteoporosis after a World Health Organization (WHO) report published in 1994, OP is often diagnosed on the patient's T-score value which difference of BMD from young adult mean normalized to the population standard deviation. However; the assessment of BMD with in-patients is very difficult. Modern clinical methods such as QCT, single photon absorptiometry and MRI are used to measure these parameters.

Estimation of the osteoporosis could be considered as a machine learning task. Ensemble learning methods are one of the most attractive methods for data classification problems. Ensemble learning techniques consist of a combination of various classifiers to perform a classification task jointly

[8]. These techniques have preferred features which make them proper form for datasets [9, 10] The main objective of ensemble construction is to decrease the prediction error of a individual learner based classification task for the learning [11].

In this paper; bagging, gradient boosting and random subspace methods were incorporated in building IBk and RF ensemble classifiers for the classification of osteoporosis disease. *Six* different feature set models were created to examine the impact of osteoporotic parameters. Model 1 includes *twenty-one* features (5 BMD +5 T-score+5 Z-score +5 bone area and age of the patients). Model 2 consists of only *five* BMD parameters. Model 3 has only *five* T-score values. Model 4 consists of *five* Z-score values. Model 5 has only *five* bone area values. Model 6 was constructed according to a feature selection algorithm. Gain ratio attribute evaluator [12] was utilized as a feature selection method. According to this, *five* T-score, *five* BMD values (Totally 10 features) were selected and Model 6 feature set was created for these *ten* parameters.

Three hundred fifty post menopausal women participated in the study. Since osteoporosis disease is mostly seen in post-menopausal women population, post-menopausal women patients were purposefully included in the study. The participants of the study were divided into three groups as control, OP and ON. At the end of the study; total classification accuracy and *f*-measure of the real data set were calculated as performance measures of the proposed ensemble classification system.

Related works

Automatic diagnosis systems to classify osteoporosis disease have attracted more attention in the last decade. Some classification methods to diagnose osteoporosis disease were reported in the past years. Sapthagirivasan et al. [13] showed a Support Vector Machine (SVM) based computer-aided diagnosis (CAD) system for osteoporotic risk detection using digital hip radiographs. They utilized *five* morphologic features extracted from digital hip radiography, *five* demographic features and *five* DXA features (totally 15 features) in order to input of the SVM classifier. Sapthagirivasan et al. [14] in their latest study, they demonstrated a new framework to automatically calculate the trabecular bone strength from femur CT images. Besides, they also extracted *three* trabecular bone features, such as solidity delta points, boundness and volume fraction in order to estimate their correlation with femoral neck BMD. Umadevi et al. [15] presented multiple classification system for fracture detection in human bone x-ray images. They used 12 features consists of texture and shape features extracted from x-ray images. As classifiers, Artificial Neural Network (ANN), k-NN and SVM classifiers were chosen. Chan et al. [16] depicted an osteoporotic classification system. They gathered 18 osteoporotic risk factors as input of the CART decision tree classifier. Lemineur et al. [7] considered both fractal and BMD parameters for inputs of ANN and they applied ANN to discriminate the osteoporotic fracture and control cases. Kim et al. [17] developed osteoporosis risk prediction system using some machine learning methods. They used some demographic (age, height, weight etc.) and clinical characteristics (pregnancy, duration of menopause, hypertension etc.) as features. They predicted osteoporosis risk with SVM, ANN, random forest (RF) and logistic regression classifiers using 15 features. Tay et al. [18] presented ensemble based regression analysis for osteopenia diagnosis. *Three* different feature sets were created. *Two* sets derived from CT scans and a set consists of physical and blood test. Totally, 18 features were utilized for regression test. Several ensemble methods (ensemble RF and ensemble ANN) were also performed. Liu et al. [19] predicted hip bone fracture using ensemble ANN technique. They used many risk factors (over the 50) for features and constructed different ensemble ANN model.

In this study, we generated *six* features sets to improve estimation of osteoporotic fracture. We are aiming to determine best feature set in order to classify osteoporotic fracture. Additionally, some ensemble learning algorithms like bagging, gradient boosting and RSM were utilized to reduce the variance of errors. As weak learners, IBk with several distance functions and RF classifiers were performed for the ensemble classification.

Materials and methods

Subjects

In the study, 350 post-menopausal women's data were analyzed. The study population was divided into three groups as follows: (1) control (n=115, mean ± SD age=55.0±5.65);

Table 1 Mean ± SD of the studied bone densitometry parameters

Patients	L1	L2	L3	L4	Total
Control	0.84±0.021	0.95±0.07	1.07±0.02	0.96±0.07	0.96±0.002
ON	0.75±0.04	0.92±0.03	0.91±0.007	0.92±0.007	0.88±0.02
OP	0.53±0.03	0.610.004	0.71±0.001	0.74±0.03	0.66±0.012

Table 2 Mean ± SD of the studied T-score parameters

Patients	L1	L2	L3	L4	Total
Control	−0.8±0	−1.6±0.56	−0.15±0.21	−1.35±0.63	−0.8±0
ON	−1.55±0.49	−0.9±0.42	−1.55±0.07	−1.75±0.07	−1.45±0.21
OP	−3.5±0.28	−1.7±0.84	−3.4±0.28	−3.4±0.28	−3.5±0.14

(2) ON (n=144, age=61.4±9.2) and (3) OP (n=91, age=62.8±12.72). Control group refers healthy people. These datasets were acquired from the hospital of Cerrahpasa Medical Faculty, Istanbul University in Turkey.

Evaluation of bone densitometry

BMD, T-score, Z-score and bone area were measured for the whole body, at the lumbar spine by dual-energy X-ray absorptiometry with a QDR 4500 densitometer (Hologic, Waltham, MA, USA).

Data analysis

In this study; age of the patient and bone densitometry parameters; L1, L2, L3 and L4 spine (BMD, area, T-score, Z-score, total BMD, total T-score and total Z-score) were analyzed. These parameters were considered as input of the osteoporotic fracture classification system. Lumbar vertebrae can be viewed differently shaped in the DXA. For example, L1, L2 and L3 have a U or Y shaped appearance whereas L4 has a block H or X shaped appearance. Furthermore, on AP DXA lumbar spine studies L1 through L4 are quantified. Besides; L1 generally has the lowest BMD value; L3 has the highest BMD value between the first four lumbar vertebrae. However; areas of the vertebrae from L1 to L4 increase [20].

Ages of patients were considered as one of the input parameters regarding the effect of age on osteoporosis. In control group, mean and standard deviation (SD) of age is as; 55±5.65. For *ON* group, age is 61.5±9.2 while it is 62.8±12.7 for *OP* group.

In bone densitometry parameters; L1, L2, L3 and L4 spine (BMD, area, T-score, Z-score, total BMD, total T-score and total Z-score) were chosen. BMD could be measured to monitor response to treatment for osteoporosis. Mean and SD values of the patients for BMD (in g/cm2) parameters are given in Table 1.

Another main bone densitometry parameter group is T-scores. T-score measures the departure of the subject's BMD value from the mean BMD for a young adult population in units of the standard deviation about the mean for the young adult age range. The young adult mean and SD are usually derived from a group of healthy subjects aged 20 to 35 years, matched for sex and race [21]. Mean and SD values of the patients for T-score parameters are given in Table 2.

One of the bone densitometry parameter groups is Z-scores. The deviation from the mean bone density of adults of the same age and gender is named Z-score. Mean and SD values of the patients for Z-score parameters are given in Table 3.

The two-dimensional projected area in cm^2 of the bones was also measured in the study. Mean and SD of the area of the bones for L1, L2, L3, L4 and total are depicted in Table 4

Ensemble learning

Ensemble learning is a machine learning technique which uses multiple base learners to increase predictive accuracy. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in several methods such as majority voting and averaging to classify new samples [22–24]. Due to the fact that combining predictions of an ensemble are often more accurate than the individual classifiers, ensemble methods were applied in the study. Ensemble learning approach could be divided into two ensemble methods as generative and non-generative. Non-generative ensemble methods mostly are based on the former feature of ensemble methods. However, generative ensemble methods mainly focus on the latter. Non-generative methods are classified as ensemble fusion (majority voting, fuzzy fusion, Meta learning etc.) and ensemble selection (forward-backward selection, test and select, clustering based selection etc.). However, generative ensembles are partitioned in Resampling, Feature selection, Mixture of experts, Output Coding, and Randomized

Table 3 Mean ± SD of the studied Z-score parameters

Patients	L1	L2	L3	L4	Total
Control	1.55±0.21	1.9±0.42	2.6±0	1.45±0.91	1.85±0.21
ON	−1.5±0.42	−0.9±0.42	−1.5±0	−1.75±0.07	−1.45±0.21
OP	−1.7±0.42	−1.7±0.84	−1.125±0.90	−1.1±0.70	−1.4±0.70

Table 4 Mean ± SD of the studied area of the bones

Patients	L1	L2	L3	L4	Total
Control	14.87±1.81	15.35±0.35	15.94±0.67	17.76±3.02	63.87±0.17
ON	11.56±0.23	12.22±0.41	14.13±0.95	18.55±0.13	56.48±1.47
OP	11.23±1.24	12.81±1.30	14.53±2.17	17.17±1.73	55.755±6.45

ensembles methods [25]. The most popular ensemble techniques are bagging, boosting, stacking and random subspace method [26].

Bagging

Bagging method proposed by Breiman in 1996, also known as boostrap aggregating is one of the most popular ensemble techniques [27]. Bagging creates separate samples of the training data set and uses a classifier or base learner for each sample. The results of these multiple classifiers are then assigned to the class based on majority voting rule. The structure of the bagging ensemble model used in the study is depicted in Fig. 1.

Gradient boosting

Gradient Boosting is an approach to learning theory by combining many weak learners. Boosting is a classification methodology which applies weighted training data to classifier algorithm, thereby taking weighted majority voting results of the sequentially modifying classifiers [28]. The main idea of the boosting algorithm is to change the model of the samples during the training depending on the error probability of selection [29]. The structure of the gradient boosting ensemble model is given in Fig. 2.

Random subspace method

Random subspace method is one of ensemble construction techniques. It was proposed by Ho in 1998. Despite the other ensemble techniques such as bagging and boosting, RSM uses modified feature space to construct ensembles of learner in order to improve the generalization error [30]. The structure of the RSM ensemble model is displayed in Fig. 3.

Instance based learning algorithms

Instance-based learning algorithms (IBk) are one of the lazy classifiers. IBk learners carry out little work when learning from the dataset, but consume more effort during the classification process of the new examples [31]. IBk algorithms are derived from nearest neighbor classifier. By saving and using only selected instance, they produce classification predictions. The advantage of IBk learners is that they can learn quickly from a very small dataset. IBk learners can also work well for numeric data [32]. IBk algorithms have several types such as IB1, IB2 and IB3.

IB1 is the simplest instance-based learning algorithm. IB1 is same to the nearest neighbour algorithm except that it normalizes its attributes' ranges, process instances incrementally, and has a simple policy for tolerating missing values [31]. IB1 uses a distance or similarity function to decide which neighbors are closest to an input vector. In this study; Euclidean, Manhattan and Chebyshev distance function are used. These functions are defined follows respectively:

$$D(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (1)$$

$$D(x,y) = \sum_{i=1}^{m}|x_i - y_i| \quad (2)$$

$$D(x,y) = \max_{i=1}|x_i - y_i| \quad (3)$$

where m is the number of input attributes, x_i and y_i are the input values for input attribute i.

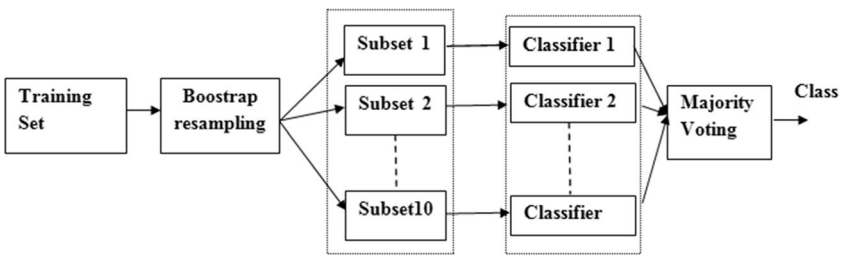

Fig. 1 The structure of the bagging ensemble model for 10 iterations

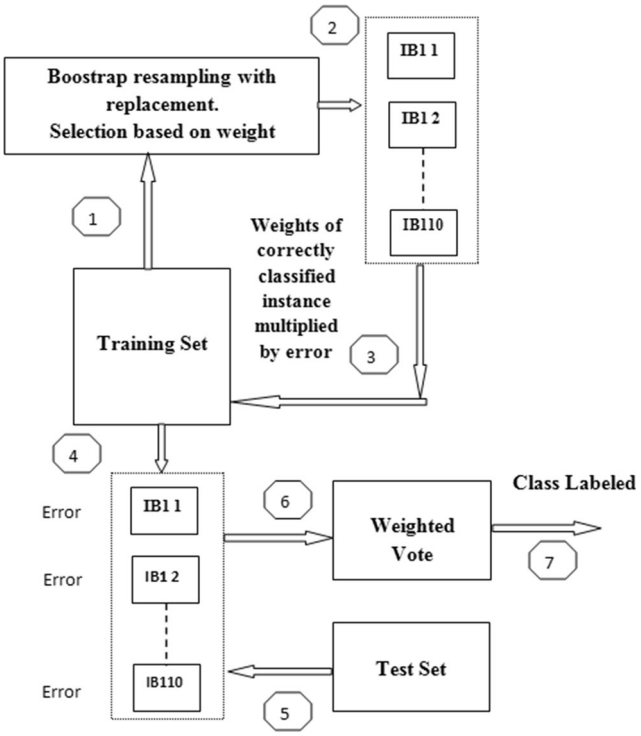

Fig. 2 The structure of the gradient boosting ensemble model

Random forest classifier

Random Forest (RF) is a tree based and fast running classifier. It is composed of a plurality of decision trees. *Random forest* is providing very good competition to ensemble techniques on various machine learning tasks. Detailed information could be given in [33, 34].

Performance measures

There are various ways to evaluate the performance of classification systems. Accuracy and *f*-measure were used to evaluate proposed ensemble classification system as performance measures. Accuracy is the common performance technique which depicts the overall performance of the classification system. It is formulated by:

$$Accuracy = \frac{True\ positives\ +\ True\ negatives}{Number\ of\ data} \tag{4}$$

f-measure is the harmonic mean of precision and recall. It utilizes both precision and the recall to compute [35]

$$F{-}Measure = \frac{\left(\beta^2+1\right) \times precision \times recall}{\beta^2 \times precision + recall} \tag{5}$$

$$precision = \frac{TP}{TP+FP} \tag{6}$$

$$recall = \frac{TP}{TP+FN} \tag{7}$$

where β is the bias value.

Experimental results

In this study, each subject contains *24* numeric attributes; age, *five* values (L1, L2, L3 L4 and Total) of BMD, T-score, Z-score, bone area and *three* class attributes (control, osteopenia, osteoporosis). *Six* feature set models were constructed as the inputs of proposed ensemble classification system in order to determine which feature group is vital to classify osteoporosis disease. In model 1; all attributes except classes were chosen. In model 2; only *five* BMD (L1, L2, L3, L4 and Total) values were used as features. In model 3; only *five T*-score (L1, L2, L3, L4 and Total) values were utilized as features. In model 4; only *five Z*-score (L1, L2, L3, L4 and Total) values were selected as features. In model 5; only *five* bone area (L1, L2, L3, L4 and Total) values were used as the inputs of the classifier. An attribute selection technique was also used to create model 6 feature set. Gain ratio attribute evaluator [12] was performed to all attributes in the data set. Importance of the features were ranked by gain ratio attribute evaluator as follows: 1-Total T-score; 2- Total BMD; 3-L3 BMD; 4-L3 T-score; 5-L2 BMD; 6-L2 T-score; 7- L4 BMD; 8-L4 T-score; 9-L1 BMD; 10- L1 T-score; 11-L3 Z-score; 12-Total Z-score; 13-L1 Z-score; 14-L4 Z-score; 15-L2 Z-score; 16-Total area; 17-L1 area; 18-L2 area; 19-L3 area; 20- L4 area and 21-Age of the patients. This ranking showed that BMD and T-score values are very important features to classify osteoporosis disease. Therefore; *five* BMD and *five* T-score parameters were taken as model 6 feature set.

Entire data set which consisted of 350 subjects was classified into *three* groups as control, OP and ON. *10*-fold cross validation procedure was used in the classification system in

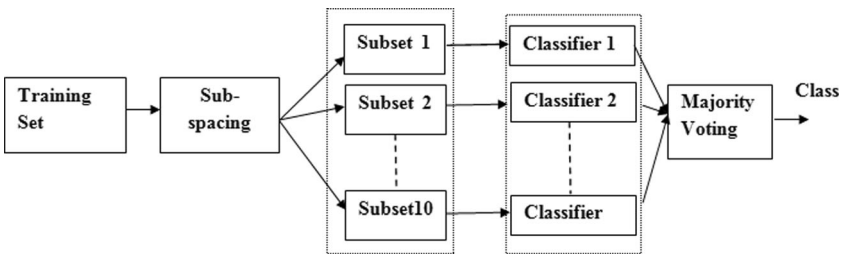

Fig. 3 The structure of the RSM ensemble model for 10 iterations

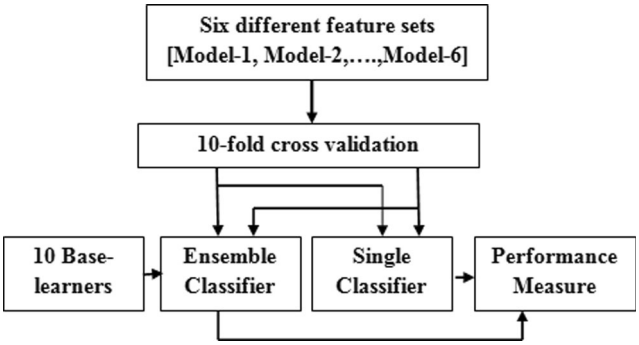

Fig. 4 Block diagram of the proposed classification system

order to obtain better network generalization. Ensemble learning techniques such as bagging, gradient boosting and RSM were applied to *six* different feature sets mentioned above. *IB1* and *RF* classifiers were utilized as the base learners of the ensemble learning techniques. The block diagram of the proposed classification system is given in Fig. 4.

The performance of proposed IB1 classification system was measured by assigning *k* value between from 1 to 15. The mean values of the performance measures of the proposed system were calculated in order to obtain better generalization results. *Three* different distance functions such as Euclidean, Manhattan and Chebyshev were performed for the *IB1* classifier. The number of base learner was selected as 10 to avoid over fitting for bagging, G. boosting and RSM ensemble algorithms. Overall performance measures of the IB1 classifiers with Euclidean, Manhattan and Chebyshev distance functions were shown in Tables 5, 6 and 7 respectively.

When comparing performance measures of the IB1 classifier, most suitable distance function was determined as Manhattan distance. In constrst, worst suitable distance function was found as Chebyshev distance using ensemble IB1 classifier. Furthermore; RSM ensemble technique was determined to be the most efficient to classify osteoporosis. Moreover; model 6 feature set was obtained as the best feature model among the six feature groups. Finally, combination of IB1 with Manhattan distance function, model-6 feature set (*five* BMD + *five* T-score) and RSM ensemble technique were determined as the best combination of IB1 osteoporosis classification system.

The accuracy of the best combination of *IB1* classifier has been computed for varying *k* value between 1 and 15. The comparison graph of the effect of k value on accuracy of *IB* classifier is shown in Fig. 5.

Ensemble learning algorithms usually perform better with tree based classifiers. Therefore; *RF* which is one of the tree based classifiers was used as a base learner to estimate osteoporotic fractures. In *RF* structure, number of tree was selected as 10. Furthermore; 10 *RF* base learners were utilized for ensemble RF classification system. Overall performance measures of the *RF* classifiers were depicted in Table 8.

As shown in Table 8, the best combination of the RF classifier was found as RSM ensemble technique and model-6 feature set. Accuracy of the RSM-RF classifier with model-6 feature set was calculated %98.85 and *f*-measure was found as 0.986. Confusion matrix was also presented in Table 9 for the best combination.

Discussion

In this study, *two* different ensemble classifiers (IB1, RF) and *six* different feature groups were performed together in order to determine the best combination of osteoporotic fracture classification system. At first; IB1 ensemble classifier using *three* different distance function was performed over several feature set. Comparing the results of the IB1 ensemble in Tables 5, 6 and 7, the most effective distance function was found as Manhattan distance for almost all combination of IB1 classifier. Considering the feature sets created, model-6 which consists of *five* BMD and *five* T-score values was found as the most important feature group to classify osteoporotic

Table 5 Overall performance measures of the IBK with Euclidean distance classifier

		Model type constructed according to features					
		Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
RSM IBK (Euclidean Distance)	F- Measure	0,90±0.02	0.928±0.07	0.947±0.15	0.66±0.01	0.39±0.01	0.96±0.04
	Accuracy	90.49±2.5	93.13±0.86	94.85±0.52	66.39±1.33	40.23±1.68	95.85±3.12
Bagging IBK (Euclidean Distance)	F- Measure	0.88±0.04	0.91±0.005	0.93±0.006	0.641±0.02	0.39±0.03	0.956±0.03
	Accuracy	89.35±4.32	91.71±0.57	93.8±0.7	64.35±2.43	39.55±3.5	95.78±4.4
G. Boosting IBK (Euclidean Distance)	F- Measure	0.868±0.04	0.88±0.01	0.92±0.01	0.637±0.02	0.38±0.02	0.94±0.02
	Accuracy	87.31±4.19	89.32±1.65	92.64±1.59	63.96±2.99	39.42±2.44	94.27±4.05
IBK (Euclidean Distance)	F- Measure	0.876±0.04	0.90±0.01	0.935±0.01	0.638±0.02	0.38±0.02	0.95±0.03
	Accuracy	88.13±4.49	90.33±1.13	93.74±0.85	63.99±2.96	39.43±2.44	95.06±3.72

Table 6 Overall performance measures of the IBK with Manhattan distance classifier

		Model type constructed according to features					
		Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
RSM IBK (Manhattan Distance)	F- Measure	0,937±0.03	0.932±0.04	0.95±0.06	0.665±0.02	0.426±0.02	0.961±0.02
	Accuracy	93.92±3.10	93.42±1.14	95.10±0.82	66.96±1.56	42.97±1.32	96.33±1.56
Bagging IBK (Manhattan Distance)	F- Measure	0.93±0.02	0.932±0.05	0.95±0.004	0.66±0.04	0.42±0.03	0.957±0.05
	Accuracy	93.25±3.48	93.42±1.23	95.10±1.86	66.12±2.64	42.28±1.76	95.88±3.42
G. Boosting IBK (Manhattan Distance)	F- Measure	0.921±0.04	0.918±0.02	0.93±0.03	0.648±0.03	0.41±0.03	0.943±0.02
	Accuracy	92.10±4.23	92±2.86	93.25±3.32	64.56±1.88	41.16±2.86	94.42±4.12
IBK (Manhattan Distance)	F- Measure	0.92±0.03	0.925±0.03	0.946±0.02	0.642±0.01	0.41±0.02	0.957±0.04
	Accuracy	92.6±4.12	92.71±1.85	94.85±0.72	64.42±3.23	41.23±4.12	95.63±2.85

Table 7 Overall performance measures of the IBK with Chebyshev distance classifier

		Model Type constructed according to features					
		Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
RSM IBK (Chebyshev Distance)	F- Measure	0,87±0.04	0.903±0.02	0.938±0.05	0.636±0.02	0.393±0.03	0.941±0.02
	Accuracy	87.6±4.12	90.57±1.02	94.17±3.27	63.87±1.12	39.44±0.77	94.28±3.56
Bagging IBK (Chebyshev Distance)	F- Measure	0.83±0.02	0.895±0.04	0.925±0.01	0.63±0.03	0.39±0.03	0.94±0.04
	Accuracy	83.09±2.86	89.74±1.14	92.9±1.24	63.34±3.78	39.10±3.32	94.14±2.44
G. Boosting IBK (Chebyshev Distance)	F- Measure	0.81±0.03	0.88±0.01	0.91±0.03	0.608±0.04	0.38±0.03	0.916±0.03
	Accuracy	81.06±3.92	88±1.42	91.02±2.86	61.06±2.56	38.22±1.24	91.56±3.74
IBK (Chebyshev Distance)	F- Measure	0.81±0.03	0.89±0.02	0.917±0.03	0.61±0.02	0.38±0.02	0.93±0.03
	Accuracy	81.3±2.54	89±2.35	91.96±0.64	61.23±3.12	38.35±2.76	93.60±4.42

fracture. Besides, RSM ensemble technique was determined to be the most suitable ensemble technique for almost all combination of proposed IB1 classification system. As shown in Table 7, while the best accuracy and f-measure values were obtained from the combination of model-6 and RSM-IB1 with Manhattan distance as 96.33 %, 0.961, the worst accuracy and f-measure values were calculated from the combination of model-5 and gradient boosting as 41.16 %, 0.41, respectively.

When comparing the IB1 and RF classifiers, performance measures show that ensemble RF classifier is more successful

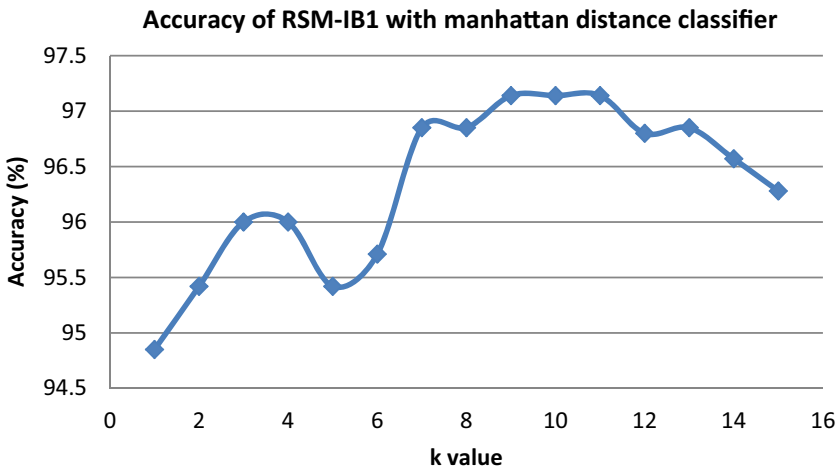

Fig. 5 The graph of accuracy of IB1based Manhattan distance classifier

Table 8 Overall performance measures of the RF classifier

		Model type constructed according to features					
		Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
RSM RF	F- Measure	0.971	0.969	0.983	0.81	0.402	0.986
	Accuracy	97.4286	97.14	98.57	81.22	40.43	98.85
Bagging RF	F- Measure	0.972	0.977	0.983	0.786	0.391	0.981
	Accuracy	97.428	98	98.57	78.57	40	98.28
G. Boosting RF	F- Measure	0.957	0.972	0.981	0.771	0.40	0.98
	Accuracy	96	97.42	98.28	77.14	40.14	98
Random Forest (RF)	F- Measure	0.929	0.974	0.98	0.799	0.379	0.974
	Accuracy	93.143	97.71	98	80	38.28	97.71

Table 9 The confusion matrices of the ensemble RF classifiers for Model-6

Actual	Predicted			Classifier types
	Control (Healthy)	ON	OP	
Control (Healthy)	113	2	0	RF
ON	2	139	3	
OP	0	1	90	
Control (Healthy)	113	1	1	G. Boosting-RF
ON	2	139	3	
OP	0	0	91	
Control (Healthy)	114	1	0	Bagging-RF
ON	2	141	1	
OP	0	0	91	
Control (Healthy)	114	1	0	RSM-RF
ON	2	141	1	
OP	0	0	91	

than ensemble IB1 classifier in the OP decision. As seen in Table 8, the best accuracy and f-measure values were obtained from combination of model-6 and RSM-RF as 98.85 % and 0.986. However, the worst results were obtained from the combination of model-5 and single RF classifier. Considering all the results, the best feature group was found to be model-6. On the other hand, these results demonstrate that combination of T-score and BMD values are vital parameters in OP decision. Besides, Z-score and bone area values were not sufficient enough to classify the osteoporosis. Hence, by the use of only *ten* physical parameters (T-score and BMD) that can easily be measured without invasion, osteoporosis patients could be classified with high accuracy as OP, ON or control group.

Upon comparing the ensemble learning techniques, RSM ensemble technique emerged as the most effective technique for the decision of OP among the others. Additionally; in the study, RSM and bagging ensemble techniques were found to be more effective than gradient boosting and individual IB1 or RF classifiers to diagnose osteoporosis disease. In addition, this study has demonstrated that ensemble learning techniques confirms a relation between individual densitometry results and the outcome of investigation for osteoporosis case.

The comparison of this study with previous studies, in terms of the methodology, number of features and accuracy was reported in Table 10. It is difficult to make a fair comparison of the effectiveness of previous studies because their feature selection, validation procedure and classifier techniques are different. Besides, most of studies given in Table 10 have two-class classification problem, but this study has *3*-class classification problem which makes it difficult to

Table 10 Comparison of proposed study with previous studies

Authors	Methodology	The number of features	Accuracy (%)
Sapthagirivasan et al.	SVM	15	90
Umadevi et al.	ANN, k-NN and SVM	12	91.89
Chan et al.	ANN, decision tree	18	65
Lemineur at al.	ANN	6	81.66
Kim et al.	SVM, ANN and RF	15	77.8
Tay et al.	Ensemble RF and Ensemble ANN	18	0.946 (AUROC)
Liu et al.	Ensemble ANN	Over 50	85
Proposed study	RSM-IB1 with Manhattan distance	10	96.33
	RSM-RF	10	98.85

obtain better accuracy score. However, the results of this study compare favorably to the others in total accuracy as 98.85 % using combination of model-6 and RSM-RF ensemble classifier.

Conclusion

In this study, the effects of *six* different osteoporotic features model and ensemble learning methods on osteoporosis disease decision support system were investigated. In order to carry out the study, *six* feature set models were considered as inputs to ensemble classifiers (gradient boosting, bagging and RSM). By using model-6 feature sets, high diagnosis accuracy was obtained with RSM ensemble techniques. These results illustrate that both T-score and BMD values are very important parameters to estimate osteoporosis disease. Otherwise, the accuracy and f-measure rates dramatically decreased by the use of model-4 and model-5 features for the all classifiers. Thus, these results show that bone area and Z-score values were less effective parameters to classify osteoporosis disease. Likewise, this study also emphasizes that RSM-RF ensemble classifier is the most effective method to classify osteoporosis disease.

IBk Instance based learning, *RF* Random forest, *RSM* Random subspace method, *BMD* Bone mineral density, *OP* Osteoporosis, *ON* Osteopenia, *QCT* Quantitative computed tomography, *MRI* Magnetic resonance imaging, *SD* Standard deviation, *TP* True positive, *FP* False positive, *FN* False negative, *ANN* Artificial neural network, *SVM* Support vector machine, *CAD* Computer aided diagnosis, *k-NN* k- nearest neighbor, *WHO* World health organization.

Acknowledgments This work was partially supported by The Research Fund of Istanbul University. Project Number: UDP-7098 and YADOP-36785. We would like to extend our appreciation to all medical staff who participated in this study for their invaluable support, especially Dr. Sait Sager, MD

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Özdemir, Z. T., Acar, A., and Karabulut, L., Osteoporosis And Vertebral Fractures In Inflammatory Bowel Disease. *Bozok Med. J.* 4(1):48–54, 2014.
2. Yildirim, P., Ceken, C., Hassanpour, R., Esmelioglu, S., and Tolun, M. R., Mining MEDLINE for the Treatment of Osteoporosis. *J. Med. Syst.* 36:2339–2347, 2012.
3. Iliou, T., Anagnostopoulos, C. N., and Anastassopoulos, G., osteoporosis detection using machine learning techniques and feature selection. *Int. J. Artif. Intell. T.* 23(05), 9 pages, 2014.
4. Akgundogdu, A., Jennane, R., Aufort, G., Benhamou, C. L., and Ucan, O. N., 3D image analysis and artificial intelligence for bone disease classification. *J. Med. Syst.* 34(5):815–828, 2010.
5. Consensus Development Conference: Diagnosis, prophylaxis and treatment of osteoporosis. *Am. J. Med.* 94:646–650, 1993.
6. Ordóñez, C., Matías, J. M., de Cos Juez, C. F., and García, P. J., Machine learning techniques applied to the determination of osteoporosis incidence in post-menopausal women. *Math. Comput. Model.* 50(5-6):673–679, 2009.
7. Lemineur, G., Harba, R., Kilic, N., Ucan, ON., Osman, O., Benhamou, L., Efficient estimation of osteoporosis using artificial neural networks. 33rd Annual Conference of the IEEE: Taipei, pp 3039–3044, 2007.
8. Benhamou, C. L., Poupon, S., Lespessailles, E., Loiseau, S., Jennane, R., Siroux, V., Ohley, W., and Pothuaud, L., Fractal analysis of radiographic trabecular bone texture and bone mineral density: two complementary parameters related to osteoporotic fractures. *J. Bone Miner. Res.* 16:697–704, 2001.
9. Tartar, A., Kilic, N., Akan, A., Classification of pulmonary nodules by using hybrid features. *Comput. Math. Methods. Med.* Article ID 148363, pages 11.
10. Erdal, H. I., Karakurt, O., and Namli, E., High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform. *Eng. Appl. Artif.* 26(4): 1246–1254, 2013.
11. Tartar, A., Akan, A., Kilic, N., A novel approach to malignant-benign classification of pulmonary nodules by using ensemble learning classifiers. 36th Annual Conference on Engineering in Medicine and Biology Society (EMBC), Chicago.
12. Panwar, S. S., and Raiwani, Y. P., Data reduction techniques to anayze NSL-KDD dataset. *Int. J. Comput. Eng. Technol.* 5(10): 21–31, 2014.
13. Sapthagirivasan, V., and Anburajan, M., Diagnosis of osteoporosis by extraction of trabecular features from hip radiographs using support vector machine: An investigation panorama with DXA. *Comput. Biol. Med.* 43(11):1910–1919, 2013.
14. Sapthagirivasan, V., Anburajan, M., and Janarthanam, S., Extraction of 3D Femur Neck Trabecular Bone Architecture from Clinical CT Images in Osteoporotic Evaluation: a Novel Framework. *J. Med. Syst.* 39(8):1–11, 2015.
15. Umadevi, N., Geethalakshmi, S.N., Third International Conference on Computing Communication & Networking Technologies (ICCCNT), Coimbatore, pages1–8, 2012.
16. Chan, Y. T., Miller, P. D., Barret-Conner, E., Weiss, T. W., Sajjan, S. G., and Siris, E. S., An approach for identifying postmenopausal women age 50–64 years of increased short-term risk for osteoporortic fracture. *Osteoporos. Int.* 18:1287–1296, 2007.
17. Kim, S.K., Yoo, T.K., Kim, D.W., Osteoporosis risk prediction using machine learning and conventional methods. In Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE, Osaka, pp. 188–191.
18. Tay, W. L., Chui, C. K., Ong, S. H., and Ng, A. C. M., Ensemble-based regression analysis of multimodal medical data for osteopenia diagnosis. *Expert Syst. Appl.* 40(2):811–819, 2013.
19. Liu, Q., Cui, X., Chou, Y. C., Abbod, M. F., Lin, J., and Shieh, J. S., Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders. *Biomed Signal. Process.* 21: 146–156, 2015.
20. Bonnick, S.L., Lewis, L.A., *Bone densitometry for technologists. Humana Press*, 2006.
21. Blake, G., and Fogelman, M., Interpretation of bone densitometry studies. *Semin. Nucl. Med.* 27(3):248–260, 1997.

22. Cunha, P., Moura, D. C., López, M. A. G., Guerra, C., Pinto, D., and Ramos, I., Impact of Ensemble Learning in the Assessment of Skeletal Maturity. *J. Med. Syst.* 38(9):1–10, 2014.
23. Dietterich, T.G., Ensemble methods in machine learning. In: Proceedings of Conference on Multiple Classifier Systems. 1857, 1–15, 2000.
24. de Pinho Valente, C.T.M., A tool for text mining in molecular biology domains (Doctoral dissertation, Universidade do Porto), 2013.
25. Re, M., Valentini, G., Ensemble methods: a review. Data Mining and Machine Learning for Astronomical Applications, Data Mining and Knowledge Discovery Series, Chapman & Hall, pp 563–594, 2012.
26. Mert, A., Kilic, N., and Akan, A., Evaluation of bagging ensemble method with time domain feature for diagnosing of arrhythmia beats. *Neural. Comput. Applic.* 24(2):317–326, 2014.
27. Breiman, L., Bagging predictors. *Mach. Learn.* 24(2):123–140, 1996.
28. Freund, Y., and Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55(1):119–139, 1997.
29. Niwas, S. I., Lin, W., Bai, X., Kwoh, C. K., Sng, C. C., Aquino, M. C., and Chew, P. T. K., Reliable Feature Selection for Automated Angle Closure Glaucoma Mechanism Detection. *J. Med. Syst.* 39(3):1–10, 2015.
30. Ho, T. K., The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(8): 832–844, 1998.
31. Martin, B., Instance based learning: nearest neighbour with generalisation. Master of Science Thesis in University of Waikato, New Zealand, 1995.
32. Aha, D. W., Kibler, D., and Albert, M. K., Instance-based learning algorithms. *Mach. Learn.* 6(1):37–66, 1991.
33. Breiman, L., Random forests, Tech. Rep., Statistics Department, University of California, Berkeley, Calif, USA, 1999.
34. Breiman, L., Random forests. *Mach. Learn.* 45(1):5–32, 2001.
35. Hripcsak, G., and Rothschild, A. S., Agreement, the f-measure, and reliability in information retrieval. *J. Am. Med. Inform. Assoc.* 12(3):296–298, 2005.