



Application of machine learning approaches for osteoporosis risk prediction in postmenopausal women

Jae-Geum Shim¹ · Dong Woo Kim¹ · Kyoung-Ho Ryu¹ · Eun-Ah Cho¹ · Jin-Hee Ahn¹ · Jeong-In Kim¹ · Sung Hyun Lee¹

Received: 7 May 2020 / Accepted: 29 July 2020 / Published online: 23 October 2020
© International Osteoporosis Foundation and National Osteoporosis Foundation 2020

Abstract

Summary Many predictive tools have been reported for assessing osteoporosis risk. The development and validation of osteoporosis risk prediction models were supported by machine learning.

Introduction Osteoporosis is a silent disease until it results in fragility fractures. However, early diagnosis of osteoporosis provides an opportunity to detect and prevent fractures. We aimed to develop machine learning approaches to achieve high predictive ability for osteoporosis risk that could help primary care providers identify which women are at increased risk of osteoporosis and should therefore undergo further testing with bone densitometry.

Methods We included all postmenopausal Korean women from the Korea National Health and Nutrition Examination Surveys (KNHANES V-1, V-2) conducted in 2010 and 2011. Machine learning models using methods such as the k-nearest neighbors (KNN), decision tree (DT), random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), artificial neural networks (ANN), and logistic regression (LR) were developed to predict osteoporosis risk. We analyzed the effect of applying the machine learning algorithms to the raw data and featuring the selected data only where the statistically significant variables were included as model inputs. The accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC) were used to evaluate performance among the seven models.

Results A total of 1792 patients were included in this study, of which 613 had osteoporosis. The raw data consisted of 19 variables and achieved performances (in terms of AUROCs) of 0.712, 0.684, 0.727, 0.652, 0.724, 0.741, and 0.726 for KNN, DT, RF, GBM, SVM, ANN, and LR with fivefold cross-validation, respectively. The feature selected data consisted of nine variables and achieved performances (in terms of AUROCs) of 0.713, 0.685, 0.734, 0.728, 0.728, 0.743, and 0.727 for KNN, DT, RF, GBM, SVM, ANN, and LR with fivefold cross-validation, respectively.

Conclusion In this study, we developed and compared seven machine learning models to accurately predict osteoporosis risk. The ANN model performed best when compared to the other models, having the highest AUROC value. Applying the ANN model in the clinical environment could help primary care providers stratify osteoporosis patients and improve the prevention, detection, and early treatment of osteoporosis.

Keywords Osteoporosis · Machine learning · Predict · Risk assessment

Introduction

Osteoporosis is characterized by low bone mass resulting in bone fragility fractures that occur following minimal or no trauma [1, 2]. Osteoporosis is common in postmenopausal

women but is a silent disease until the fractures occur. Fractures place a severe burden on aging individuals because they can lead to poor quality of life and increased mortality [3]. Osteoporosis should be prevented and treated before it is complicated by fractures [4].

According to World Health Organization (WHO) criteria, osteoporosis is operationally defined as a bone mineral density (BMD) that is 2.5 standard deviations or more below the mean for a young healthy adult (T score ≤ -2.5), based on the dual-energy x-ray absorptiometry (DXA) T score. [5] DXA is generally used to diagnose osteoporosis. Although the benefits of screening are apparent, as early diagnosis may help prevent

✉ Sung Hyun Lee
4321hoho@naver.com

¹ Department of Anesthesiology and Pain Medicine, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, 29, Saemooan-ro, Gonro-gu, Seoul 03181, Republic of Korea

future morbidity and decrease mortality due to fracture complications, uniform screening of the general population using DXA may not be feasible because all physicians may not have access to this equipment. Therefore, substantial research has been conducted on when and where to use DXA to screen efficiently and to avoid overdiagnosis and misdiagnosis or create a false sense of security [6–8]. Several previous studies have highlighted prescreening tools to identify women with an increased risk of osteoporosis who ought to be selected for BMD measurements. These tools are simple formulas based on risk factors of osteoporosis [9–11].

Machine learning has been shown to improve the predictive value of statistics in many areas of medicine [12–14]. Machine learning is a field of computer science that uses computer algorithms to identify patterns in large amounts of data, which can also be used as predictors for novel data [15]. Using training data with known input and output values, the machine learning algorithm is able to make data-driven predictions or decisions [15, 16]. Although machine learning models have been proposed as a tool to predict osteoporosis risk in postmenopausal Korean women, previous studies had limitations, such as only applying the ANN method or not including lifestyle factors such as smoking, physical activity, coffee, and alcohol intake [17].

In this study, we aimed to develop and validate a selection of machine learning models using a database of 1792 patients who participated in the Korea National Health and Nutrition Examination Surveys (KNHANES) V-1 and V-2 (2010–2011) to construct an osteoporosis predictor. In databases of KNHANES, the definition of osteoporosis is based on only the *T* score for BMD assessed by DXA at the femoral neck or spine that is 2.5 standard deviations or more below the mean for a young healthy adult (T score ≤ -2.5). Low trauma hip, vertebral, proximal humerus, or pelvis fracture that could be considered clinical osteoporosis were excluded [18, 19]. The predictive model in our study is complicated and has high dimensional characteristics as it contains diet and lifestyle properties, in addition to clinical factors, that could contribute to osteoporosis [6, 20, 21]. Considering the characteristics of complex models, we compared the performances of various models, using the raw data together with the preprocessed data, wherein statistically significant features were selected in advance.

Materials and methods

Study population

We analyzed the data from 1792 postmenopausal Korean women who participated in the Korea National Health and Nutrition Examination Surveys (KNHANES) V-1 and V-2 (2010–2011). The KNHANES data are available and can be

downloaded from the KNHANES website (<https://knhanes.cdc.go.kr/>). The KNHANES is a nationwide, population-based, cross-sectional study that has been conducted periodically since 1998, which assesses the health and nutritional status of Koreans, monitors trends in health risk factors and the prevalence of major chronic diseases, and provides data for the development and evaluation of health policies and programs in Korea [22]. We excluded patients with incomplete information from our analysis. This study was approved by our institutional ethics committee (Kangbuk Samsung Hospital Institutional Review Board, Seoul, Republic of Korea; approval number: KBSMC 2020-01-007). The KNHANES received ethical approval from the Institutional Review Board of the Korea Centers for Disease Control and Prevention (IRB Nos. 2010-02CON-21-C and 2011-02CON-06-C) and complies with the Declaration of Helsinki. Informed consent was obtained from all participants for inclusion in the surveys.

Machine learning

Classification machine learning algorithms were used to predict the occurrence of osteoporosis, encoded as a binary outcome variable. The whole process was divided into five parts: (1). Data preprocessing: this included data cleaning, missing data processing, and data transformation; (2). Feature selection: the process of selecting input features for training; (3). Model building: application of the classification machine learning algorithms to achieve reasonable performance; (4). Cross-validation: a resampling procedure to evaluate the machine learning models for training and testing of raw and feature selected data; and (5). Model performance evaluation: this was conducted using area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. We plotted the AUROC curves from all the machine learning models, using the testing data.

Data preprocessing

A total of 1792 patients were included in this study, of which 613 were diagnosed with osteoporosis. Data were analyzed using R software version 3.6.2. (R Development Core Team, Vienna, Austria). Data scaling was performed using normalization and minimum-maximum scaling included in the Caret preprocessing libraries. The continuous variables were age, height, weight, body mass index (BMI), waist circumference, pregnancy, and duration of menopause. The categorical variables were estrogen therapy, hyperlipidemia, hypertension, history of fracture, osteoarthritis, rheumatoid arthritis, diabetes mellitus, smoking, alcohol, coffee, and physical activity. We used the data in the KNHANES V-1 and V-2 datasets as the training and testing data, respectively. The entire dataset was split into two categories: training and testing. For each

machine learning algorithm, the data of 1353 subjects were used for training and those of 439 for testing. The Synthetic Minority Over-Sampling Technique (SMOTE) method, which addresses class imbalance, was used to generate synthesis samples to overcome the low incidence of osteoporosis in the training set [23].

Feature selection

Numerous studies have been conducted to identify features that may potentially affect osteoporosis risk [6, 20, 21]. We found 19 potential features, including demographics and clinical variables as shown in Table 1. Feature selection is the process wherein we select those features which contribute most to our output prediction [24, 25]. In this process, the backward stepwise variable selection procedure was used to identify such variables, using the logistic regression model. To construct the machine learning model, we included only the statistically significant features in the feature selected dataset as shown in Table 2.

Model building

Except for the ANN model, all the machine learning models were imported from the Caret package containing functions for training and plotting classification and regression models (<https://CRAN.R-project.org/package=caret>). The ANN model was constructed using the Keras package designed to enable fast experimentation with deep neural networks (<https://github.com/keras-team/keras>). The machine learning approaches were developed to accurately identify patients at risk for osteoporosis. Classification models such as the k-nearest neighbors (KNN), decision tree (DT), random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), artificial neural networks (ANN), and logistic regression (LR) were used to developed prediction models.

K-nearest neighbors (KNN) is a simple algorithm that classifies unlabeled observations based on a similarity measure such as a distance function. Input values are classified by a majority vote of its neighbors by assigning them to the class most common among its k-nearest neighbors measured by a distance function [26]. Decision Trees (DT) create models in the form of a flowchart-like tree structure which represents feature at an internal node, represents a decision rule by the branch, and generates the actual prediction at the leaf nodes. The technique learns to partition the tree on the basis of the feature value in recursively manner [27, 28]. Random forest (RF) is an ensemble classification algorithm that consists of a large number of individual decision trees [29]. Gradient boosting machine (GBM) is a type of machine learning boosting. It produces an ensemble model in the form of shallow and weak successive trees with each tree learning and improving on the previous [30, 31]. Support vector machine

(SVM) split data into binary categories with a bisecting hyperplane [32]. Hyperplanes are decision boundaries that help separate the data points. The algorithm finds the hyperplane to represent the maximum distance between data points of the two categories. Input values falling on either side of the hyperplane can be assigned to different categories. Artificial neural networks (ANN) are computational models inspired by the biological neural networks that constitute animal brains [33]. It consists of input and output layers, as well as the inner hidden layers to simulate the signal transmission. Each layer comprises many nodes, and the nodes between layers are interconnected by different weights that adjust as learning proceeds. The algorithms automatically learn from the training dataset to predict output values [34]. Logistic regression (LR) is a traditional statistical method for binary classification problems, although it has been adopted as a basic machine learning model. Logistic regression predicts the probability of occurrence of a binary event utilizing the sigmoid function also called a logistic function [35].

Using machine learning algorithms to analyze medical data to predict a disease frequently involves choosing hyperparameters. A hyperparameter can be defined as a parameter that is not tuned during the learning process through iterative optimization of an objective function. Investigators typically tune hyperparameters arbitrarily after a series of manual trials. Different model training algorithms require different hyperparameters. The optimal hyperparameters obtained in a fivefold cross-validation of the test set are summarized in Table 3.

Cross-validation

We validated the performance of all classification models using stratified k-fold cross-validation (Fig. 1). Cross-validation is a validation technique for assessing how the classification models will generalize to an unknown dataset and how accurately they will perform in practice. It is widely used in settings wherein the main goal is prediction. In this study, the dataset was randomly divided into five equal folds with approximately the same number of events. After partitioning one data sample into five subsets, one subset was selected for model validation while the remaining subsets were used to establish machine learning models. Finally, the validation results were combined to provide an estimate of the model's predictive performance.

Model performance evaluation

We evaluated diagnostic ability based on four parameters: accuracy, sensitivity, specificity, and AUROC. The AUROC is known as a strong indicator of performance for classifiers in imbalanced datasets [36, 37]. We plotted AUROC curves to compare the performances of the machine learning classification models.

Table 1 Demographic data and variable features of the included postmenopausal women

Variables	All, <i>n</i> = 1792	Training set, <i>n</i> = 1353 (KNHANES V-1)	Test set, <i>n</i> = 439 (KNHANES V-2)	<i>p</i> value*
Age (years)	62 (56–70)	62 (56–70)	62 (56–70)	0.834
Height (cm)	154 (150–158)	153 (150–157)	155 (151–158)	< 0.001*
Weight (kg)	57 (52–62)	56 (52–62)	58 (53–62)	0.070
BMI (kg/cm ²)	24.0 (22.1–26.2)	24.0 (22.1–26.2)	24.1 (22.0–26.2)	0.959
Waist circumference (cm)	82 (76–88)	82 (76–88)	83 (76–88)	0.310
Pregnancy	5 (3–6)	5 (3–6)	4 (3–6)	0.358
Duration of menopause (years)	12.0 (5.0–21.0)	13.0 (5.0–21.0)	12.0 (5.5–21.5)	0.746
Estrogen therapy (<i>n</i>)	305 (17.0)	237 (17.5)	68 (15.5)	0.363
Hyperlipidemia (<i>n</i>)	375 (20.9)	272 (20.1)	103 (23.5)	0.151
Hypertension (<i>n</i>)	745 (41.6)	567 (41.9)	178 (40.5)	0.655
History of fracture	263 (14.7)	201 (14.9)	62 (14.1)	0.765
Osteoarthritis (<i>n</i>)	566 (31.6)	434 (32.1)	132 (30.1)	0.467
Rheumatoid arthritis (<i>n</i>)	81 (4.5)	62 (4.6)	19 (4.3)	0.928
Diabetes mellitus (<i>n</i>)	217 (12.1)	160 (11.8)	57 (13.0)	0.574
Smoking (<i>n</i>)	128 (7.1)	99 (7.3)	29 (6.7)	0.692
Alcohol intake (<i>n</i>)	1197 (66.8)	889 (65.7)	308 (70.2)	0.096
Coffee intake (<i>n</i>)	1045 (58.3)	777 (57.4)	268 (61.0)	0.200
Physical activity (<i>n</i>)	591 (33.0)	440 (32.5)	151 (34.4)	0.504
Osteoporosis (<i>n</i>)	613 (34.2)	449 (33.2)	164 (37.4)	0.123

KNHANES The Korea National Health and Nutrition Examination Survey

The data are presented as median (interquartile range) or number (%). **p* value < 0.05

Statistical analysis

The continuous variables were expressed as mean ± standard deviation or median ± interquartile range, as appropriate, and analyzed by the unpaired *t* test or the Mann-Whitney *U* test. The categorical variables were presented as absolute number (*n*) and relative frequency (%) and analyzed by the chi-square test or Fisher's exact test. The machine learning classification models were constructed using R software (version 3.6.2). The performance of the classification models for osteoporosis risk assessment was measured and compared using AUROCs.

We also calculated the accuracy, sensitivity, and specificity (95% confidence interval). Differences with *p* < 0.01 were considered to be statistically significant.

Study design

Considering the high dimensionality of the data, which included 19 variables, we applied two different machine learning approaches, depending on where the variable reduction process was applied [38]. The first approach was to apply machine learning algorithms to the raw dataset. The second was

Table 2 Results of the stepwise logistic regression model for osteoporosis risk assessment

Variables	Beta coefficient	SE	Odds ratio	95% CI	<i>p</i> value
Hyperlipidemia	−0.452	0.181	0.636	0.444–0.904	0.0128
Diabetes mellitus	−0.543	0.216	0.581	0.378–0.883	0.0119
Estrogen	−0.472	0.201	0.624	0.417–0.918	0.0190
History of fracture	0.463	0.183	1.588	1.109–2.272	0.0114
Age	2.435	0.684	11.416	3.016–44.235	0.0004
Height	−3.443	0.568	0.032	0.010–0.096	0.0000
Waist circumference	1.271	0.886	3.565	0.631–20.421	0.1513
BMI	−4.873	0.976	0.008	0.001–0.051	0.0000
Menopause	1.876	0.630	6.527	1.904–22.560	0.0029

SE, standard error; BMI, body mass index; CI, confidence interval

The backward stepwise variable selection process was conducted to choose the statistically significant input using a cutoff *p* value of < 0.20

Table 3 Optimal hyperparameters of all machine learning models

Model	Optimal hyperparameters	
	Raw data	Feature selected
LR	Alpha = 0.1 Lambda = 0.040	Alpha = 0.1 Lambda = 0.001
KNN	$k = 20$	$k = 8$
DT	Maximum depth = 5 Criterion = Gini index	Maximum depth = 5 Criterion = Gini index
RF	mtry* = 3	mtry* = 3
GBM	Maximum depth = 2 Number of estimators = 50, Gamma = 0	Maximum depth = 2 Number of estimators = 50, Gamma = 0
SVM	Kernel = radial basis Sigma = 0.039 C = 0.25	Kernel = radial basis Sigma = 0.119 C = 0.50
ANN	Number of hidden layers = 2; number of nodes in a layer = 20, 10	Number of hidden layers = 2; number of nodes in a layer = 20, 10

LR logistic regression, KNN k-nearest neighbors, DT decision tree, RF random forest, GBM gradient boosting machine, SVM support vector machine, ANN artificial neural networks

*mtry indicates the number of variables available for splitting at each tree node

to apply logistic regression analysis to the raw dataset so as to choose only the effective variables from the training dataset variables. We identified nine variables that were significantly different between patients with osteoporosis and those without. Nonsignificant variables were removed from the algorithmic input in the feature selected dataset.

Results

Patients' characteristics

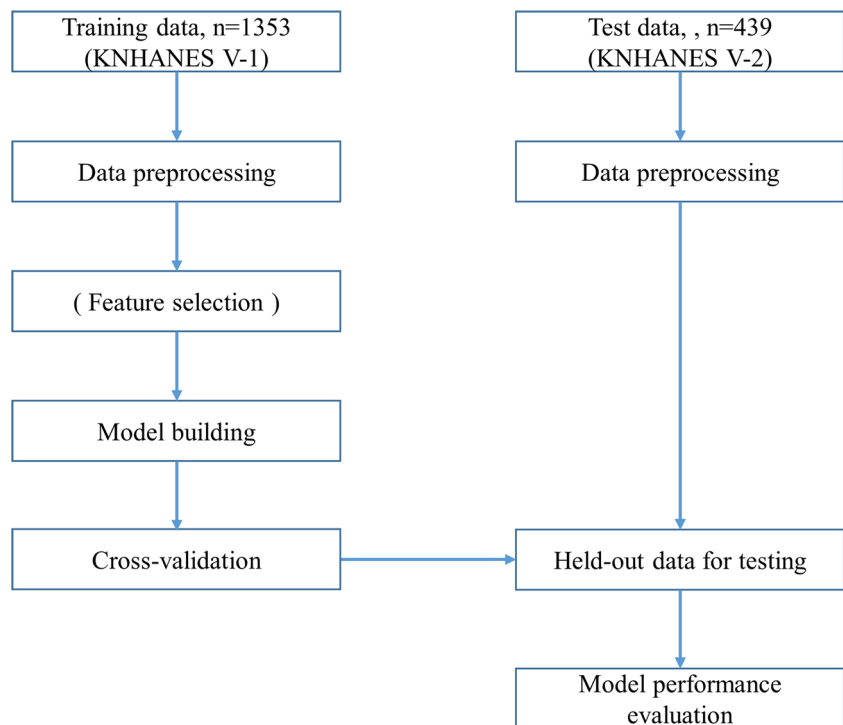
We analyzed the data of 1792 postmenopausal Korean women who participated in the KNHANES V-1 and V-2 from

January 1, 2010 to December 31, 2011. The demographic and patient characteristics are summarized in Table 1. Osteoporosis occurred in 34.2% of cases (training set, 33.2%; test set, 37.4%).

Feature selection

The input variables used for the feature selected data included age, height, BMI, history of smoking, waist circumference, history of fracture, estrogen therapy, duration of menopause, hyperlipidemia, and diabetes mellitus. Table 1 shows the potential variables for predicting patients at risk for osteoporosis. The MASS library in the R software was used to perform stepwise backward elimination logistic regression analysis to

Fig. 1 Schematic of the machine learning pathway



obtain probability coefficients for each variable. The nine features with the greatest regression coefficient magnitudes (with $p < 0.2$) were used as input variables in classifying the machine learning models for osteoporosis risk assessment.

Model performance

The AUROCs for the test data set for all machine learning techniques for predicting osteoporosis risk are shown in Table 4. For the raw data, which included 19 variables, the ANN method achieved the best performance in terms of AUROC (0.741), followed by RF (0.727), LR (0.726), SVM (0.724), KNN (0.712), DT (0.684), and GBM (0.652). For the feature selected data, which included nine variables, the AUROCs increased slightly for all machine learning methods, with the best performance being that of ANN (0.743). Using feature selected data decreased the sensitivity for KNN (0.58) and ANN (0.72) but increased it for LR (0.79), SVM (0.73), DT (0.60), and RF (0.68). All algorithms showed better performance in terms of accuracy when using the feature selected data. The AUROCs of the seven different models are plotted in Fig. 2.

Discussion

Feature selection is an important concept in machine learning that has a huge influence on performance. Analysis and modeling with or without the feature selection process offer the opportunity to identify patients at high risk and to identify clinical factors that may increase the risk of osteoporosis. The objective of this study was to demonstrate that machine learning algorithms could accurately predict if postmenopausal women have a higher possibility of developing osteoporosis. This means that machine learning algorithms provide an alternative approach that could be useful in guiding the decision to perform DXA, considering a specific set of clinical factors. According to the United States Preventive Service Task Force (USPSTF) guidelines, the National Osteoporosis Foundation guidelines, and other guidelines, it is recommended that women aged 65 years or older, postmenopausal women starting or taking long-term (≥ 3 months) systemic glucocorticoid therapy, and perimenopausal or postmenopausal women with additional osteoporosis risk factors (low BMI, current smoker, rheumatoid arthritis, history of hip fracture in a parent, early menopause, and excessive alcohol intake) are screened for osteoporosis by BMD measurement at the hip and lumbar spine [1, 39, 40]. Considering these various factors associated with low bone density, machine learning algorithms may be supportive tools for identifying postmenopausal women at high risk for osteoporosis. In some cases, clinical efficiency can be expected through a two-step screening strategy that uses DXA testing after the use of machine learning

Table 4 Performance of all machine learning models

Model	Feature selected data							
	AUROC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
LR	0.726 (0.701–0.790)	0.758 (0.716–0.798)	0.60 (0.52–0.67)	0.85 (0.81–0.89)	0.727 (0.672–0.753)	0.749 (0.706–0.789)	0.79 (0.72–0.85)	0.66 (0.60–0.72)
KNN	0.712 (0.667–0.755)	0.729 (0.685–0.770)	0.65 (0.57–0.72)	0.78 (0.72–0.83)	0.713 (0.687–0.778)	0.747 (0.704–0.787)	0.58 (0.50–0.66)	0.85 (0.80–0.89)
DT	0.684 (0.654–0.748)	0.720 (0.675–0.761)	0.54 (0.46–0.62)	0.83 (0.78–0.87)	0.685 (0.641–0.731)	0.706 (0.661–0.748)	0.60 (0.52–0.68)	0.77 (0.71–0.82)
RF	0.727 (0.708–0.798)	0.763 (0.720–0.802)	0.59 (0.51–0.66)	0.87 (0.82–0.91)	0.734 (0.688–0.773)	0.747 (0.704–0.787)	0.68 (0.61–0.75)	0.79 (0.73–0.83)
GBM	0.652 (0.600–0.686)	0.633 (0.586–0.678)	0.73 (0.65–0.79)	0.58 (0.52–0.64)	0.728 (0.672–0.755)	0.718 (0.673–0.759)	0.77 (0.70–0.83)	0.69 (0.63–0.74)
SVM	0.724 (0.681–0.769)	0.743 (0.699–0.783)	0.65 (0.57–0.72)	0.80 (0.74–0.84)	0.728 (0.674–0.758)	0.727 (0.682–0.768)	0.73 (0.66–0.80)	0.72 (0.67–0.78)
ANN	0.741 (0.685–0.765)	0.724 (0.680–0.766)	0.80 (0.74–0.86)	0.68 (0.62–0.73)	0.743 (0.693–0.777)	0.749 (0.706–0.789)	0.72 (0.64–0.79)	0.77 (0.71–0.82)

AUROC area under the receiver operating characteristic, LR logistic regression, KNN k-nearest neighbors, DT decision tree, RF random forest, GBM gradient boosting machine, SVM support vector machine, ANN artificial neural networks

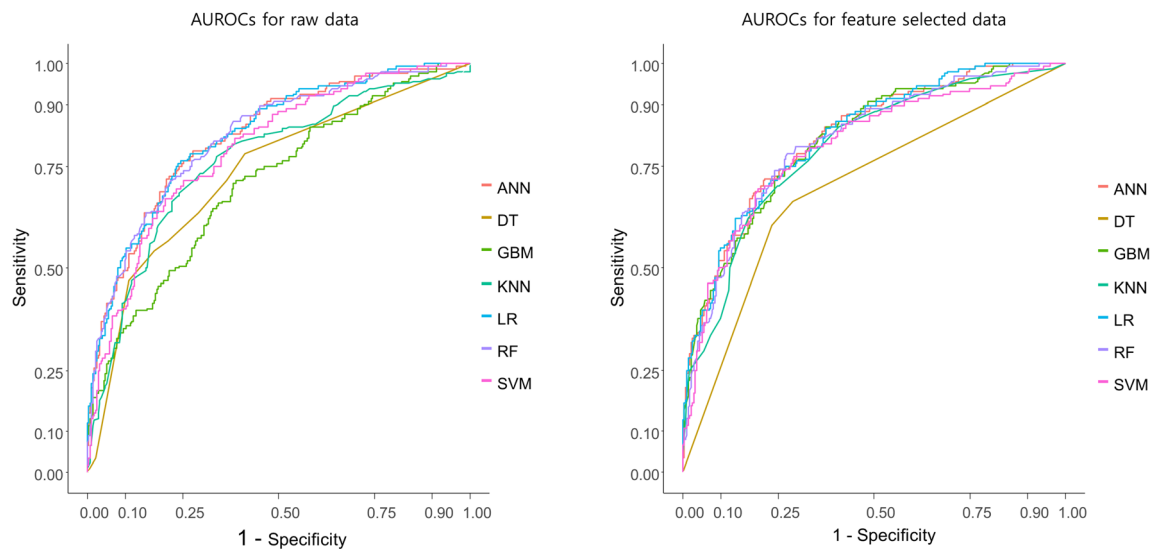


Fig. 2 Areas under the receiver operating curve for raw (left) and feature selected (right) data

algorithms. A previous randomized study, namely the Risk-Stratified Osteoporosis Strategy Evaluation (ROSE) study, investigated the effectiveness of a two-step osteoporosis screening program for women aged 65–80 years, using the Fracture Risk Assessment Tool (FRAX), a self-administered questionnaire, to select women for DXA, followed by standard osteoporosis treatment [41]. The ROSE study showed risk reduction in the group following the two-step strategy when compared to the control group; a FRAX score $\geq 15\%$ was considered to predict moderate- or high-risk of major osteoporotic fractures, hip fractures, and all fractures [42]. Effective machine learning models coupled with DXA may yield results as the ROSE study.

Machine learning algorithms have commonly been applied for classification and prediction, rather than causal inference. Our study may seek to promote health by intervening with patients at high risk of osteoporosis; this requires the ability to predict osteoporosis risk but not the need for causal inference about the effect of an input variable on that risk [43]. As for alcohol intake, there are two questions in KNHANES: life-time drinking experience and high-risk drinking frequency. We applied both of them as input variables in our models, but the performance of the predictive model was higher when life-time drinking experience was applied as input variable. Therefore, life-time drinking experience was used as a feature in this study although one time use of a drink is not a risk factor for bone loss.

We used a relaxed p value ($p < 0.20$) in a multivariable logistic regression analysis as shown in Table 2. There is no reason to worry about a relaxed p value criterion at feature selection stage because this is just a pre-selection strategy and no inference will derive from this step [44]. This relaxed p value criterion will help reduce the risk of missing important variables. In addition, it is possible to include a sufficient number of features because machine learning techniques are

relatively free of limitations of conventional statistical analysis such as multicollinearity [14].

Previous studies have employed the use of logistic regression and various machine learning models to predict patients at high risk of osteoporosis [17, 45]. However, these studies either trained the models using only the ANN method or were based on limited input features. In this study, we developed and validated our models by performing feature selection, cross-validation, and testing on completely different datasets. Thus, our findings are helpful for implementing machine learning methods in clinical settings.

We investigated the application of seven machine learning techniques to the KNHANES V-1 and V-2 databases, which involve heterogeneous clinical characteristics. Unlike previously published studies, which do not incorporate diet and lifestyle patterns, our study included these features. We demonstrated that machine learning algorithms can be applied to predict osteoporosis risk with a reasonable level of performance.

In this study, we found that the optimal ANN needed two hidden layers to predict osteoporosis risk. In the ANN model, the first and second hidden layers were composed of 20 and 10 nodes, respectively. Since no specific tool exists for obtaining the most suitable hyperparameters to construct ANN models, we obtained the optimal hyperparameters empirically. The hyperparameters found could be useful as indicators in future studies using the ANN method.

This study has several limitations. First, the study used a cross-sectional survey that captured a population at a single point in time that is not guaranteed to be representative. Second, the prediction model in our study was based on Korean women. Thus, it may be difficult to generalize our study to a more diversified population. Third, there is some ambiguity in the survey at the KNHANES. More specifically,

a survey of pregnancy history was assessed through questionnaires for each individual. The questionnaire contained the following questions: Have you had any pregnancy experience (currently during pregnancy, natural abortion, artificial abortion, ectopic pregnancy, etc.)? and if you answered yes to the question, how many times have you had total pregnancy? Unfortunately, there are some shortcomings that are not clear whether same occurrences in one individual or different individuals. Furthermore, our study could not predict the occurrence of osteopenia and osteoporosis using a multi-classification algorithm to reduce the risk of osteoporosis before the occurrence of fractures. In our database, osteoporosis, which classified osteoporosis only according to an operational definition, was not considered as another clinical standard such as low trauma fracture.

In conclusion, this study is important because it promotes the identification of patients at high risk of osteoporosis in a population of postmenopausal Korean women. The findings of this study show that the ANN model is the best machine learning classification model for predicting osteoporosis risk using a feature selected dataset. We made two different observations regarding osteoporosis risk assessment using machine learning models. First, input variables composed of clinical and diet and lifestyle factors, such as coffee intake, alcohol intake, and physical activity, were used to train our machine learning model. Second, we used two entirely different datasets, KNHANES V-1 and V-2, as the training and testing datasets, respectively. This means that the dataset used to train our classifier models (KNHANES V-1) was not of relevance to the dataset used for testing (KNHANES V-2). However, careful attention is required for practical clinical application of our study findings, as our study was limited to postmenopausal Korean women and had a limited data size.

Compliance with ethical standards

Conflict of interest None.

Ethics approval This study was approved by the Kangbuk Samsung Hospital Institutional Review Board. The KNHANES received ethical approval from the Institutional Review Board of the Korea Centers for Disease Control and Prevention. Informed consent was obtained from all participants for inclusion in the surveys.

References

- Cosman F, de Beur SJ, LeBoff MS, Lewiecki EM, Tanner B, Randall S, Lindsay R (2014) Clinician's guide to prevention and treatment of osteoporosis. *Osteoporos Int* 25:2359–2381
- Kim SY, Ok HG, Birkenmaier C, Kim KH (2017) Can denosumab be a substitute, competitor, or complement to bisphosphonates? *Korean J Pain* 30:86–92
- Black DM, Rosen CJ (2016) Clinical Practice. Postmenopausal Osteoporosis. *N Engl J Med* 374:254–262
- Diab DL, Watts NB (2013) Postmenopausal osteoporosis. *Curr Opin Endocrinol Diabetes Obes* 20:501–509
- Kanis JA (1994) Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Report of a WHO Study Group. *World Health Organ Tech Rep Ser* 843:1–129
- NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy (2001) Osteoporosis prevention, diagnosis, and therapy. *Jama* 285:785–795
- Gallagher JC (2018) Advances in osteoporosis from 1970 to 2018. *Menopause* 25:1403–1417
- Yedavally-Yellayi S, Ho AM, Patalinghug EM (2019) Update on osteoporosis. *Prim Care* 46:175–190
- Cadarette SM, Jaglal SB, Murray TM, McIsaac WJ, Joseph L, Brown JP (2001) Evaluation of decision rules for referring women for bone densitometry by dual-energy x-ray absorptiometry. *Jama* 286:57–63
- Ma Z, Yang Y, Lin J, Zhang X, Meng Q, Wang B, Fei Q (2016) BFH-OST, a new predictive screening tool for identifying osteoporosis in postmenopausal Han Chinese women. *Clin Interv Aging* 11:1051–1059
- Toh LS, Lai PSM, Wu DB, Bell BG, Dang CPL, Low BY, Wong KT, Guglielmi G, Anderson C (2019) A comparison of 6 osteoporosis risk assessment tools among postmenopausal women in Kuala Lumpur, Malaysia. *Osteoporos Sarcopenia* 5:87–93
- Kim JS, Merrill RK, Arvind V, Kaji D, Pasik SD, Nwachukwu CC, Vargas L, Osman NS, Oermann EK, Caridi JM, Cho SK (2018) Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. *Spine (Phila Pa 1976)* 43:853–860
- Lee HC, Yoon HK, Nam K, Cho YJ, Kim TK, Kim WH, Bahk JH (2018) Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. *J. Clin. Med.* 7(10):322
- Lee HC, Yoon SB, Yang SM, Kim WH, Ryu HG, Jung CW, Suh KS, Lee KH (2018) *J Clin Med* 7(11):428
- Motwani M, Dey D, Berman DS, Germano G, Achenbach S, al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ, Chinnaiyan K, Chow BJ, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtnner G, Kaufmann PA, Kim YJ, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R, Shaw LJ, Stehli J, Villines TC, Dunning A, Min JK, Slomka PJ (2017) Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J* 38:500–507
- Obermeyer Z, Emanuel EJ (2016) Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med* 375:1216–1219
- Yoo TK, Kim SK, Kim DW, Choi JY, Lee WH, Oh E, Park EC (2013) Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. *Yonsei Med J* 54:1321–1330
- Kanis JA, Cooper C, Rizzoli R, Reginster JY (2019) European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporos Int* 30:3–44
- Siris ES, Adler R, Bilezikian J, Bolognese M, Dawson-Hughes B, Favus MJ, Harris ST, Jan de Beur SM, Khosla S, Lane NE, Lindsay R, Nana AD, Orwoll ES, Saag K, Silverman S, Watts NB (2014) The clinical diagnosis of osteoporosis: a position statement from the National Bone Health Alliance Working Group. *Osteoporos Int* 25:1439–1443
- Bijelic R, Milicevic S, Balaban J (2017) Risk factors for osteoporosis in postmenopausal women. *Med Arch* 71:25–28
- Schnatz PF, Marakovits KA, O'Sullivan DM (2010) Assessment of postmenopausal women and significant risk factors for osteoporosis. *Obstet Gynecol Surv* 65:591–596

22. Kweon S, Kim Y, Jang MJ, Kim Y, Kim K, Choi S, Chun C, Khang YH, Oh K (2014) Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). *Int J Epidemiol* 43:69–77
23. Wang Q, Luo Z, Huang J, Feng Y, Liu Z (2017) A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM. *Comput Intell Neurosci* 2017:1827016
24. Wu CC, Hsu WD, Islam MM, Poly TN, Yang HC, Nguyen PA, Wang YC, Li YJ (2019) An artificial intelligence approach to early predict non-ST-elevation myocardial infarction patients with chest pain. *Comput Methods Prog Biomed* 173:109–117
25. Wu CC, Yeh WC, Hsu WD, Islam MM, Nguyen PAA, Poly TN, Wang YC, Yang HC, Jack Li YC (2019) Prediction of fatty liver disease using machine learning algorithms. *Comput Methods Prog Biomed* 170:23–29
26. Zhang Z (2016) Introduction to machine learning: k-nearest neighbors. *Ann Transl Med* 4:218
27. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgeman RN, van Berkel VH, Frieboes HB (2017) Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform* 108:1–8
28. Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: an overview and their use in medicine. *J Med Syst* 26:445–463
29. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
30. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29:1189–1232
31. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O (2019) Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med* 7:152
32. Cortes C, Vapnik V (1995) Support-vector networks. In: *Machine learning*, vol 20. Kluwer Academic Publisher, Boston, pp 237–297
33. Papadopoulos MC, Abel PM, Agranoff D, Stich A, Tarelli E, Bell BA, Planche T, Loosemore A, Saadoun S, Wilkins P, Krishna S (2004) A novel and accurate diagnostic test for human African trypanosomiasis. *Lancet* 363:1358–1363
34. Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 65:386–408
35. Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 35:352–359
36. Buda M, Maki A, Mazurowski MA (2018) A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw* 106:249–259
37. Mehmood A, Maqsood M, Bashir M, Shuyuan Y (2020) A deep Siamese convolution neural network for multi-class classification of Alzheimer disease. *Brain Sci* 10(2):84
38. Panesar SS, D'Souza RN, Yeh FC, Fernandez-Miranda JC (2019) Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous Glioma database. *World Neurosurg* X 2:100012
39. The Board of Trustees of The North American Menopause Society (2010) Management of osteoporosis in postmenopausal women: 2010 position statement of The North American Menopause Society. *Menopause* 17(1):25–54
40. Rossini M, Adami S, Bertoldo F, Diacinti D, Gatti D, Giannini S, Giusti A, Malavolta N, Minisola S, Osella G, Pedrazzoni M, Sinigaglia L, Viapiana O, Isaia GC (2016) Guidelines for the diagnosis, prevention and management of osteoporosis. *Reumatismo* 68:1–39
41. Rubin KH, Holmberg T, Rothmann MJ, Høiberg M, Barkmann R, Gram J, Hermann AP, Bech M, Rasmussen O, Glüer CC, Brixen K (2015) The risk-stratified osteoporosis strategy evaluation study (ROSE): a randomized prospective population-based study. Design and baseline characteristics. *Calcif Tissue Int* 96:167–179
42. Rubin KH, Rothmann MJ, Holmberg T, Høiberg M, Möller S, Barkmann R, Glüer CC, Hermann AP, Bech M, Gram J, Brixen K (2018) Effectiveness of a two-step population-based osteoporosis screening program using FRAX: the randomized risk-stratified osteoporosis strategy evaluation (ROSE) study. *Osteoporos Int* 29:567–578
43. Crown WH (2019) Real-world evidence, causal inference, and machine learning. *Value Health* 22:587–592
44. Sperandei S (2014) Understanding logistic regression analysis. *Biochem Med (Zagreb)* 24:12–18
45. Meng J, Sun N, Chen Y, Li Z, Cui X, Fan J, Cao H, Zheng W, Jin Q, Jiang L, Zhu W (2019) Artificial neural network optimizes self-examination of osteoporosis risk in women. *J Int Med Res* 47:3088–3098

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.