# Risk Prediction of Low Bone Density in Elderly Patients with Supervised Machine Learning Algorithms

Eda Karaismailoğlu[1], Serkan Karaismailoğlu[2]

[1]Department of Medical Informatics, University of Health Sciences Türkiye, Gülhane Faculty of Medicine, Ankara, Türkiye
[2]Department of Physiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

**Background:** Low bone mineral density (BMD) is a common age-related condition that elevates the risk of fractures and mortality. Machine learning (ML) techniques offer a promising approach for early prediction using readily available clinical, biochemical, and demographic data.

**Aims:** To evaluate the predictive performance of eleven ML models in identifying low BMD and to determine the most influential risk factors using the best-performing model.

**Study Design:** Cross-sectional study.

**Methods:** Data were obtained from National Health and Nutrition Examination Survey (2005-2010, 2013-2014, and 2017-2020), focusing on individuals aged ≥ 50 years with available femoral neck or total femur BMD data. After applying exclusion criteria, 12,108 participants were included. Supervised ML algorithms were trained using 57 clinical, biochemical, demographic, and behavioral features. Model performance was assessed using accuracy, area under the curve (AUC), recall, precision, and F1 score. SHAP analysis was employed to interpret model outputs and rank predictors.

**Results:** The extra trees classifier outperformed other ML methods, achieving an accuracy of 76.7% and an AUC of 0.85. Recursive Feature Elimination with Cross-Validation identified 14 key predictors of low BMD in descending order of importance: sex, age, body mass index, race, family income-to-poverty ratio, serum uric acid, diabetes status, HDL cholesterol, urinary creatinine, alkaline phosphatase, mean cell volume, lymphocyte count, diastolic blood pressure, and glycohemoglobin.

**Conclusion:** Tree-based ML models, particularly Extra Trees, can effectively predict low BMD. The identified risk factors include both established and lesser-studied predictors. These findings support the use of ML for personalized osteoporosis and osteopenia screening and highlight its ability to capture complex, multifactorial relationships in population health data.

## INTRODUCTION

Bone mineral density (BMD) declines with age, depending on the bone mass, leading to osteopenia and, in more advanced stages, osteoporosis.[1] This condition is a major global public health concern that adversely affects the quality of life of millions worldwide. Low bone density affects approximately 200 million individuals globally, including 54 million older adults in the United States (US).[2] It is a key risk factor for hip fractures, which is one of the most serious consequences of falls in older adults, with a 1-year mortality rate of 30%.[3] Although this issue is often highlighted in postmenopausal women, it also significantly affects men, who account for nearly one-third of all hip fractures and experience poorer outcomes.[4]

Dual-energy X-ray absorptiometry (DXA) is a non-invasive imaging technique that accurately assesses bone density and strength using low radiation and is considered the gold standard for osteoporosis screening. Areal BMD measured by DXA is converted to a T-score based on the mean and standard deviation (SD) of a young adult reference group: a T-score > -1.0 is considered normal, between -1.0 and -2.5 indicates osteopenia, and ≤ -2.5 defines osteoporosis, according to the World Health Organization.[1] Early detection of low bone density, particularly among individuals with T-scores < -1.0 but not yet osteoporotic, is essential for timely intervention and fracture prevention.

In recent years, machine learning (ML) techniques have shown considerable potential in the early diagnosis and risk stratification of chronic diseases, including musculoskeletal disorders.[5-9] ML algorithms can identify complex, non-linear patterns within clinical and demographic data, enabling more precise prediction of low bone density compared with traditional statistical methods. However, the comparative performance of different ML models in predicting low bone density remains insufficiently explored.

548

Karaismailoğlu and Karaismailoğlu. Predicting Low Bone Density with Machine Learning

The aim of this study was to (1) compare the predictive performance of 11 ML algorithms in identifying individuals with low bone density using the publicly available National Health and Nutrition Examination Survey (NHANES) dataset, and (2) to identify the most influential predictors of low bone density using the best-performing model. The insights gained from this research are expected to support evidence-based clinical decision-making and guide future studies toward personalized interventions for osteoporosis prevention and management.

## MATERIALS AND METHODS

### Study population and data source

This study utilized data from selected cycles of the NHANES, a nationally representative program conducted by the National Center for Health Statistics under the Centers for Disease Control and Prevention (https://wwwn.cdc.gov/nchs/nhanes/). NHANES employed a stratified, multistage probability sampling design to obtain health-related data from the civilian, noninstitutionalized US population. Data obtained from the periods of 2005-2010, 2013-2014, and 2017-2020 were included in the study based on the availability of BMD values (g/cm$^2$) obtained by DXA using Hologic QDR 4500A fan-beam densitometers (Hologic Inc., Bedford, MA, USA).[10] The dataset also included information on the sociodemographic characteristics (e.g., age, sex, race/ethnicity, education), behavioral factors (e.g., smoking habits, physical activity, alcohol use), clinical parameters [e.g., body mass index (BMI), blood pressure], and biochemical markers (e.g., serum vitamin D, calcium, phosphorus).

Femoral neck and total femur BMD measurements were evaluated to deduce the low bone density. An individual with a femoral neck or total femur T-score value of < -1 was classified into the low bone-density group. T-scores were calculated using the following formula:

T-score = [individual BMD-mean BMD (reference population)]/SD (reference population).

The reference means and SD values were derived from the NHANES III reference data, as published by Looker et al.[11]

Individuals who met the following criteria were excluded: (1) those aged < 50 years, (2) those without BMD measurements at the femoral neck or total femur, and (3) those with a history of cancer diagnosis. Based on these criteria, 12,108 participants were enrolled in the study. The exclusion criteria applied in the study are depicted in Figure 1.

### Data preprocessing

The procedure outlined below was implemented to prepare the data for the analysis. As a first step in data preparation, variables with > 30% missing values were removed from the dataset.

**Missing value imputation:** Missing data of continuous features were imputed using the K-nearest neighbors (KNN) method, which estimates missing values based on the similarity of feature patterns among the closest observations in the dataset, and the most frequent category was used for imputing categorical variables.

**Categorical encoding:** Nominal features were converted to numerical form via one-hot encoding.

**Feature scaling:** All continuous variables were standardized using z-score normalization to ensure comparability across features. The standardized value z for a given observation x was computed was computed using the following formula:

$z_i = (x_i - \mu)/\sigma,$

where, $x_i$ represents the i$^{th}$ individual's value, $\mu$ is the mean of the feature, and $\sigma$ is the SD of the feature.

**Multicollinearity:** Features with a correlation coefficient > 0.80 or < -0.80 were identified and removed to reduce multicollinearity.

**Feature selection:** In order to optimize the model performance and decrease overfitting, feature selection was applied based on the Light Gradient Boosting Machine (LightGBM) algorithm. LightGBM calculates feature importance scores during the model training
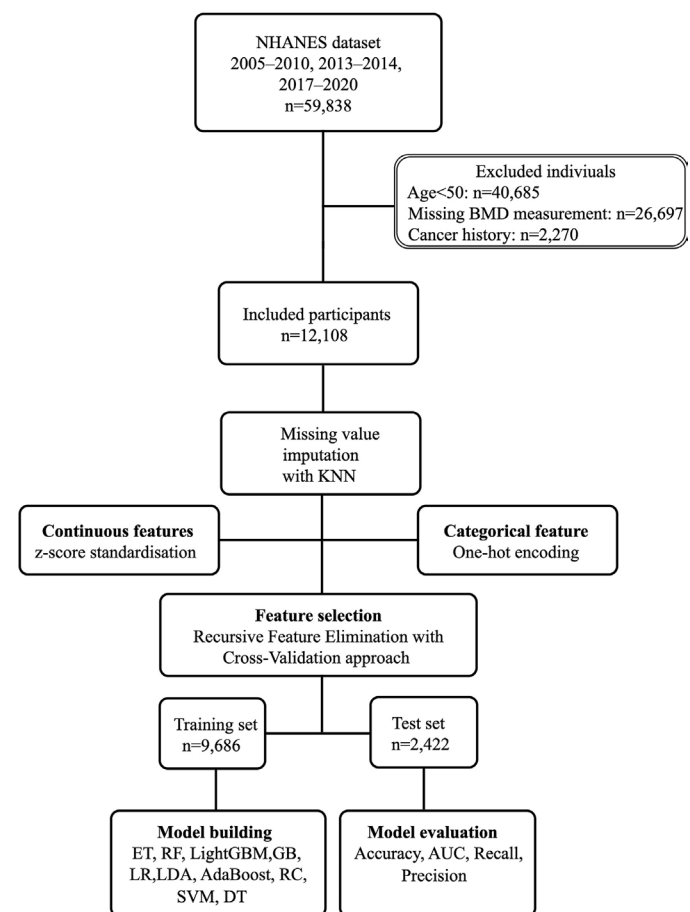


**FIG. 1.** Flowchart of the study design.

*NHANES, National Health and Nutrition Examination Survey, BMD, bone mineral density, ET, extra trees classifier; RF, random forest classifier; LightGBM, light gradient boosting machine; GB, gradient boosting classifier; LR, logistic regression; LDA, linear discriminant analysis; AdaBoost, AdaBoost classifier; RC, ridge classifier; SVM, support vector machine; DT, decision tree classifier; KNN, K-nearest neighbors; AUC, area under the receiver operating characteristic curve.*

process by evaluating each variable's contribution to information gain across decision trees. The importance scores are used to rank the input variables, and the least informative features are removed from the dataset. In this study, the threshold for selection was set to retain the top 20% of features with the highest importance scores.

**Data splitting:** The dataset was randomly partitioned into a training set (80%) and a testing set (20%). Feature selection, model fitting, and hyperparameter tuning were performed solely on the training set, and the final model performance metrics were evaluated on the testing set.

**Cross-validation:** We employed 10-fold cross-validation on the training set to assess model generalizability and guard against overfitting. The data were randomly split into ten equally sized folds, and, in each iteration, nine folds were used to fit the model and optimize hyperparameters, whereas the remaining fold was used as the validation set. Performance metrics [e.g., area under the curve (AUC), accuracy] were computed for each fold and then averaged to provide a robust estimate of model performance before the ultimate assessment on the withheld test dataset.

To avoid data leakage, all preprocessing steps (including imputation, encoding, scaling, feature selection, and feature elimination) were implemented within a strictly nested pipeline. These transformations were fitted exclusively on the training folds during cross-validation and subsequently applied to the corresponding validation fold, whereas the held-out test set was used only for the final model evaluation.

### Machine learning algorithms

A comprehensive array of supervised classification algorithms, including ensemble tree techniques, gradient-based learning systems, linear models, kernel machines, and instance-based methodologies, was used in the study.

### Extra trees classifier

The ET is an ensemble learning approach that relies on randomly generating numerous decision trees. This strategy differs from the conventional decision trees by employing entirely random threshold values at the node separation for each tree, which enhances the variety among the models. This randomization plays a role in preventing overfitting and presents robust performance on noisy datasets.[12]

### Random forest classifier

RF constructs multiple decision trees with bootstrapped subsets of the training data and integrates their predictions through majority voting. The procedure incorporates randomness in two ways: first, by sampling the data, and second, by selecting a random subset of features at each node to determine the optimal split. This approach helps reduce the variance of individual decision trees and prevents overfitting, which is common in single-tree models.[13]

### LightGBM

LightGBM is a gradient-boosted algorithm that constructs trees by prioritizing leaves, providing faster training and lower memory usage on big datasets. It leverages histogram-dependent partitioning to efficiently handle a substantial quantity of continuous features.[14]

### Gradient boosting classifier

As a sequential ensemble learning framework, the GB builds an additive model by optimizing a loss function using decision trees as weak learners. In contrast to parallel methodologies such as Random Forest, it fits trees in sequence, whereby every new tree aims to correct the residual errors of the prior ensemble of models. This approach offers a highly flexible and powerful model that can identify elaborate nonlinear relationships. Nevertheless, it is also more sensitive to overfitting.[15]

### Logistic regression

LR is a linear classification algorithm commonly used due to its simplicity, interpretability, and statistical foundations. It models the probability of a binary outcome utilizing the logistic (sigmoid) function, with the underlying assumption of a linear relationship between the log-odds of the target and the independent variables.[16]

### Linear discriminant analysis

LDA assumes normally distributed classes with equal covariance matrices and finds a linear combination of features that maximizes class separation. It is particularly effective when class distributions in scenarios by class distributions that are nearly Gaussian, and the sample sizes are within a moderate range.[17]

### AdaBoost classifier

Adaptive Boosting (AdaBoost) is an adaptive boosting algorithm that sequentially fits weak learners by reweighting incorrectly classified instances in every subsequent round. It works by sequentially training models, where each subsequent learner focuses more on the instances that were misclassified by the previous ones. However, this approach can be sensitive to noisy data and outliers.[18]

### Ridge classifier

The RC is a regularized version of linear classification that applies an L2 penalty to shrink coefficient estimates. The method reduces overfitting, especially in cases with high-dimensional or collinear data. Though it is similar to LR, this method uses a least-squares loss function rather than a log-likelihood approach.[19]

### Support vector machine

SVM is a supervised learning model used for classification tasks, especially when the feature space is high-dimensional. The linear kernel aims to identify the optimal hyperplane that maximally separates classes by maximizing the margin between the closest data points, which is known as support vectors. Although it performs well in big datasets, it may require significant computational resources.[20]

### Decision tree classifier

DT is a non-parametric, tree-based supervised learning algorithm that iteratively divides the dataset into smaller partitions, with each division determined by the feature that results in the greatest information gain. Finally, it establishes a tree structure where the

leaves represent class labels. It is easy to interpret and visualize, but prone to overfitting without proper pruning.[21]

### K-Nearest neighbors

KNN is an instance-driven, non-parametric classification algorithm that assigns class labels based on the majority vote of the "k" closest neighbors in the feature space. It stores all the training dataset and calculates distances (generally Euclidean) between a new input and existing samples during the prediction process. Despite being simple to implement, this approach may be slow during prediction and sensitive to the choice of k and the distance metric.[22]

### Performance evaluation metrics

The performance of the classification models was evaluated using accuracy, AUC, recall, precision, and F1. Accuracy measures the proportion of correctly classified instances among all observations and is defined as follows:

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

where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative values, respectively.

Recall or sensitivity quantifies the model's capacity to identify actual positive instances and is calculated as follows:

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

Precision indicates the proportion of true positive predictions among all positive predictions.

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

The F1-score is the harmonic mean of Precision and Recall, serving as a single metric that balances the trade-off between them.

$$F1 = 2 * \frac{Precision x Recall}{Precision + Recall}$$

AUC is the probability that a randomly selected positive observation is ranked higher by the model than a randomly selected negative observation. A higher AUC reflects a better ability to distinguish between classes.

### Statistical analysis

Normality of quantitative variables was tested with the Kolmogorov-Smirnov test. An independent samples $t$ - test was implemented to compare continuous variables between low and normal bone densities, and the data were expressed as the mean $\pm$ SD. Categorical variables were shown as n (%) and compared using the Pearson $\chi^2$ test (or Fisher's exact test when expected counts were < 5). All tests were two-tailed, with $p < 0.05$ considered to indicate statistical significance. Analyses were performed in R version 4.3.2 (R Foundation for Statistical Computing, Vienna, Austria).

ML models were conducted utilizing Python (version 3.10.0) within the JupyterLab environment (version 4.3.5). Data preprocessing, feature selection, model training, and evaluation were implemented with the PyCaret (version 3.2.0) library, an open-source, low-code ML framework that provides an integrated pipeline for classification and regression tasks. Performance evaluation metrics, such as accuracy, AUC, precision, recall, and F1, were computed to assess and compare the model performance.
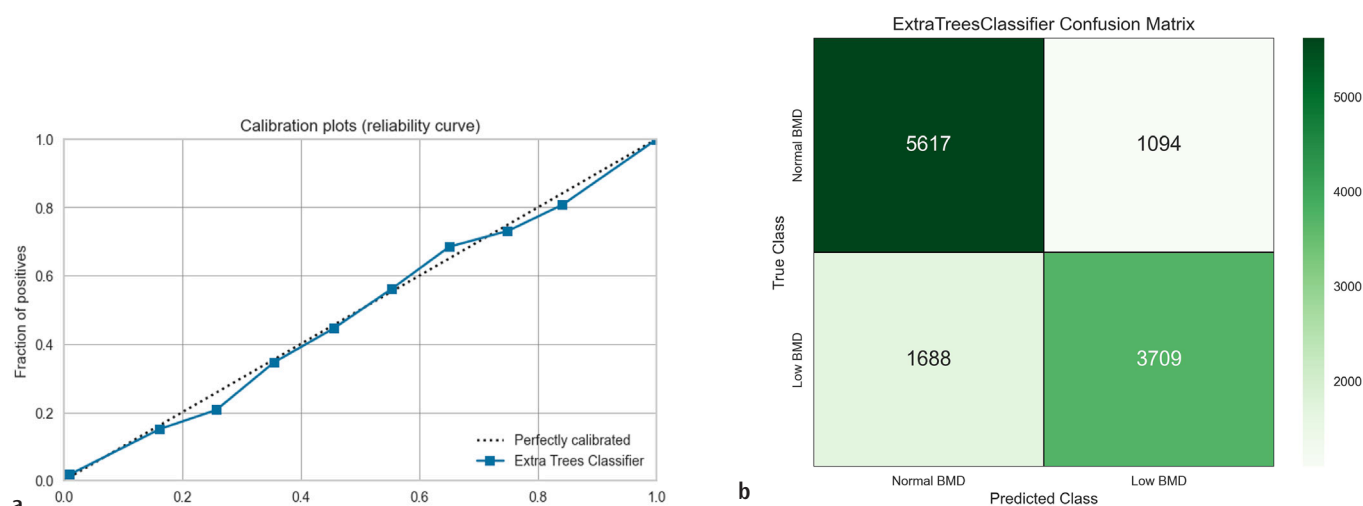
## RESULTS

A total of 12,108 individuals (female, 48%) who were ≥ 50 years of age (64 ± 9) and had complete femoral neck and total femur BMD data were included in this study. Approximately 45% of the participants were included in the low BMD group. The ML models were constructed employing 57 features, systematically categorized into seven categories to reflect their physiological and clinical relevance, as follows: 34 biochemical markers (including serum and urinary analytes), 12 hematological indices (components of the complete blood count), 4 demographic characteristics (age, sex, race/ethnicity, and income), 3 vital signs (pulse and blood pressure measurements), 2 self-reported clinical indicators (diabetes status and sleep duration), 1 anthropometric parameter (BMI), and 1 measure of physical function (walking ability).

Table 1 summarizes the performance of the 11 ML classifiers in predicting low BMD. Models were ranked by accuracy. The ET achieved the best overall performance with an accuracy of 0.7672, AUC = 0.8524, recall = 0.6873, and precision = 0.7722. The RF followed closely (accuracy = 0.7621; AUC = 0.8446). LightGBM also performed well (AUC = 0.8104; precision = 0.7329), reflecting a balanced trade-off between sensitivity and specificity. Ensemble methods such as Gradient Boosting and AdaBoost showed moderate predictive performance, whereas linear models (e.g., LR, RC) yielded slightly lower accuracy but demonstrated consistent recall and AUC values. Hyperparameter tuning of the ET model using grid search with 10-fold cross-validation did not yield improvements; thus, the default configuration was retained for the final model. The key parameters included: n_estimators = 100, criterion = "gini", max_features = "sqrt", min_samples_split = 2, min_samples_leaf = 1, and bootstrap = False. Other parameters were left at their default values, including max_depth = None, max_leaf_nodes = None, and random_state = 123 for reproducibility. The calibration of the best-performing model, the ET, was assessed to evaluate the reliability of its probability predictions. The calibration curve, as presented in Figure 2a, demonstrated that the model was well-calibrated, as the plot of its predicted probabilities closely tracks the diagonal line representing perfect calibration. This visual finding is further supported by a low Brier score of 0.12, indicating a high degree of agreement between the predicted risks and the observed outcomes. In addition, the confusion matrix indicates the model achieved a sensitivity of 69% for detecting low BMD and a specificity of 84% for identifying normal BMD (Figure 2b).

Karaismailoğlu and Karaismailoğlu. Predicting Low Bone Density with Machine Learning

551

**TABLE 1.** Performance Metrics of Classification Models.

| Model | Accuracy | Recall | Precision | F1 | AUC (95% CI) |
|---|---|---|---|---|---|
| Extra trees classifier | 0.7672 | 0.6873 | 0.7722 | 0.7272 | 0.8524 (0.8415-0.8670) |
| Random forest classifier | 0.7621 | 0.6890 | 0.7697 | 0.7271 | 0.8446 (0.8276-0.8525) |
| Light gradient boosting machine | 0.7375 | 0.6762 | 0.7329 | 0.7034 | 0.8104 (0.8085-0.8278) |
| Gradient boosting classifier | 0.7230 | 0.6520 | 0.7193 | 0.6840 | 0.7960 (0.7770-0.8121) |
| Logistic regression | 0.7100 | 0.6565 | 0.6960 | 0.6756 | 0.7859 (0.7739-0.8091) |
| Linear discriminant analysis | 0.7096 | 0.6577 | 0.6948 | 0.6757 | 0.7859 (0.7637-0.7988) |
| Ada boost classifier | 0.7095 | 0.6500 | 0.6977 | 0.6730 | 0.7803 (0.7658-0.8096) |
| Ridge classifier | 0.7093 | 0.6571 | 0.6946 | 0.6753 | 0.7859 (0.7341-0.8020) |
| Support vector machine | 0.6948 | 0.6312 | 0.6837 | 0.6564 | 0.7602 (0.7498-0.7873) |
| Decision tree classifier | 0.6891 | 0.6693 | 0.6594 | 0.6732 | 0.6876 (0.6680-0.6970) |
| K neighbors classifier | 0.6664 | 0.6216 | 0.6419 | 0.6316 | 0.7220 (0.7071-0.7314) |

AUC, area under the curve, CI, confidence interval



**FIG. 2.** (a) Calibration curve demonstrating the model's reliability. (b) Confusion matrix detailing the model's classification performance. BMD, bone mineral density.

After identifying the best-performing model, we conducted feature selection to determine the most influential risk factors implicated in low BMD. The result obtained from the recursive feature elimination with cross-validation approach utilizing the ET was presented in Figure 3. The model's performance, as measured based on the cross-validated accuracy, increased with the addition of more features and then reached a plateau at approximately 14 features, with 0.767 accuracy. It demonstrates that a smaller, selected subset of variables can achieve strong predictive performance without including all available features.

Using SHAP (SHapley Additive exPlanations) method, we identified 14 key predictors, including sex, age, BMI, race, family income/poverty, serum uric acid (mg/dL), diabetes status, HDL-cholesterol (mg/dL), urinary creatinine (mg/dL), alkaline phosphatase (ALP; IU/L), mean cell volume (fL), lymphocyte (%), diastolic blood pressure (mmHg), and glycohemoglobin (%), and reported that this subset of features yielded the highest predictive accuracy for predicting low



**FIG. 3.** Optimal feature subset determination using RFECV for extra trees classifier.

BMD. The comparison of the relevant variables between the low and normal BMD groups is presented in Table 2.

The SHAP beeswarm plot illustrates both the importance and direction of each feature's impact on the model. The red points represent higher feature values, whereas the blue points indicate lower values. The x-axis reflects each feature's contribution to the model's prediction of low BMD. The top-ranked feature in the plot is the most influential in the prediction. Accordingly, sex, age, and BMI are the top three contributors to the model's prediction of low BMD. Female sex, older age, and having a lower BMI were found to be associated with a higher predicted risk of low BMD. In addition, demographic factors, race/ethnicity, and family income/poverty contributed meaningfully. Among the biochemical markers tested, lower uric acid levels were associated with increased predicted risk, whereas higher HDL-cholesterol levels were linked to higher risk (Figure 4a). The SHAP bar plot displays the average absolute impact of each feature on the model's predictions (Figure 4b). Figure 4c presents the SHAP heatmap for the test set, wherein each column represents an individual, ordered left to right by the total magnitude of the SHAP values. In this figure, rows correspond to the most influential features. Red color reflects a risk-increasing contribution to the predicted probability of low BMD, whereas blue color is associated with a decreased risk. The black trace above the heat map [f(x)] shows the model's predicted risk per individual, confirming that columns with the largest cumulative pink area correspond to the highest predicted risk. Accordingly, the plot demonstrated that female sex, older age, and having a lower BMI were the strongest predictors of high-risk outcomes, consistently elevating the predicted values. Variables such as race, uric acid level, and HDL cholesterol exhibited mixed effects depending on the individual profiles. The participants with a lower income tended to cluster in higher-risk regions.

**TABLE 2.** Comparison of Demographic and Laboratory Characteristics Between Normal and Low BMD Groups.

| Variables | Normal BMD n = 6711 | Low BMD n = 5397 | χ²/t | Effect size | p -value |
|---|---|---|---|---|---|
| Sex[†] | | | | | |
|   Male | 4311 (64.2) | 1981 (36.7) | 908.41 | 0.28 | < 0.001 |
|   Female | 2400 (35.8) | 3416 (63.3) | | | |
| Age* | 61.76 ± 8.44 | 66.38 ± 9.39 | 28.47 | 0.52 | < 0.001 |
| Body mass index (BMI)* | 30.32 ± 5.84 | 26.80 ± 5.06 | 34.97 | 0.64 | < 0.001 |
| Race/ethnicity[†] | | | | | |
|   Mexican American | 983 (14.6) | 703 (13.0) | | | < 0.001 |
|   Other Hispanic | 676 (10.1) | 537 (9.9) | | | |
|   Non-Hispanic White | 2456 (36.6) | 2663(49.3) | 528.63 | 0.21 | |
|   Non-Hispanic Black | 2006 (29.9) | 747 (13.8) | | | |
|   Other Race | 590 (8.8) | 747 (13.8) | | | |
| Family income/poverty* | 2.76 ± 1.64 | 2.54 ± 1.58 | 7.46 | 0.14 | < 0.001 |
| Uric acid (mg/dL)* | 5.83 ± 1.44 | 5.31 ± 1.40 | 20.00 | 0.37 | < 0.001 |
| Diabetes[†] | | | | | |
|   No | 4984 (74.4) | 4310 (79.9) | | | < 0.001 |
|   Borderline | 256 (3.7) | 164 (3.1) | 53.52 | 0.07 | |
|   Yes | 1469 (21.9) | 918 (17.0) | | | |
| HDL-cholesterol (mg/dL)* | 51.58 ± 15.16 | 57.67 ± 17.12 | 20.74 | 0.38 | < 0.001 |
| Creatinine, urine (mg/dL)* | 122.75 ± 74.30 | 99.59 ± 66.80 | 17.83 | 0.33 | < 0.001 |
| Alkaline phosphatase (IU/L)* | 74.83 ± 24.14 | 78.25 ± 25.73 | 7.52 | 0.14 | < 0.001 |
| Mean cell volume (fL)* | 89.33 ± 5.72 | 90.48 ± 5.47 | 11.21 | 0.21 | < 0.001 |
| Lymphocyte (%)* | 30.89 ± 8.73 | 30.13 ± 8.87 | 4.73 | 0.09 | < 0.001 |
| Diastolic blood pressure (mmHg)* | 73.15 ± 12.49 | 70.11 ± 13.46 | 12.86 | 0.23 | < 0.001 |
| Glycohemoglobin (%)* | 6.13 ± 1.25 | 5.94 ± 1.07 | 8.86 | 0.16 | < 0.001 |

Continuous variables were summarized as mean ± standard deviation, categorical variables were presented as n (%). *p values were calculated using independent samples t-test for continuous variables. [†]p values were calculated using Pearson χ² test for categorical variables.
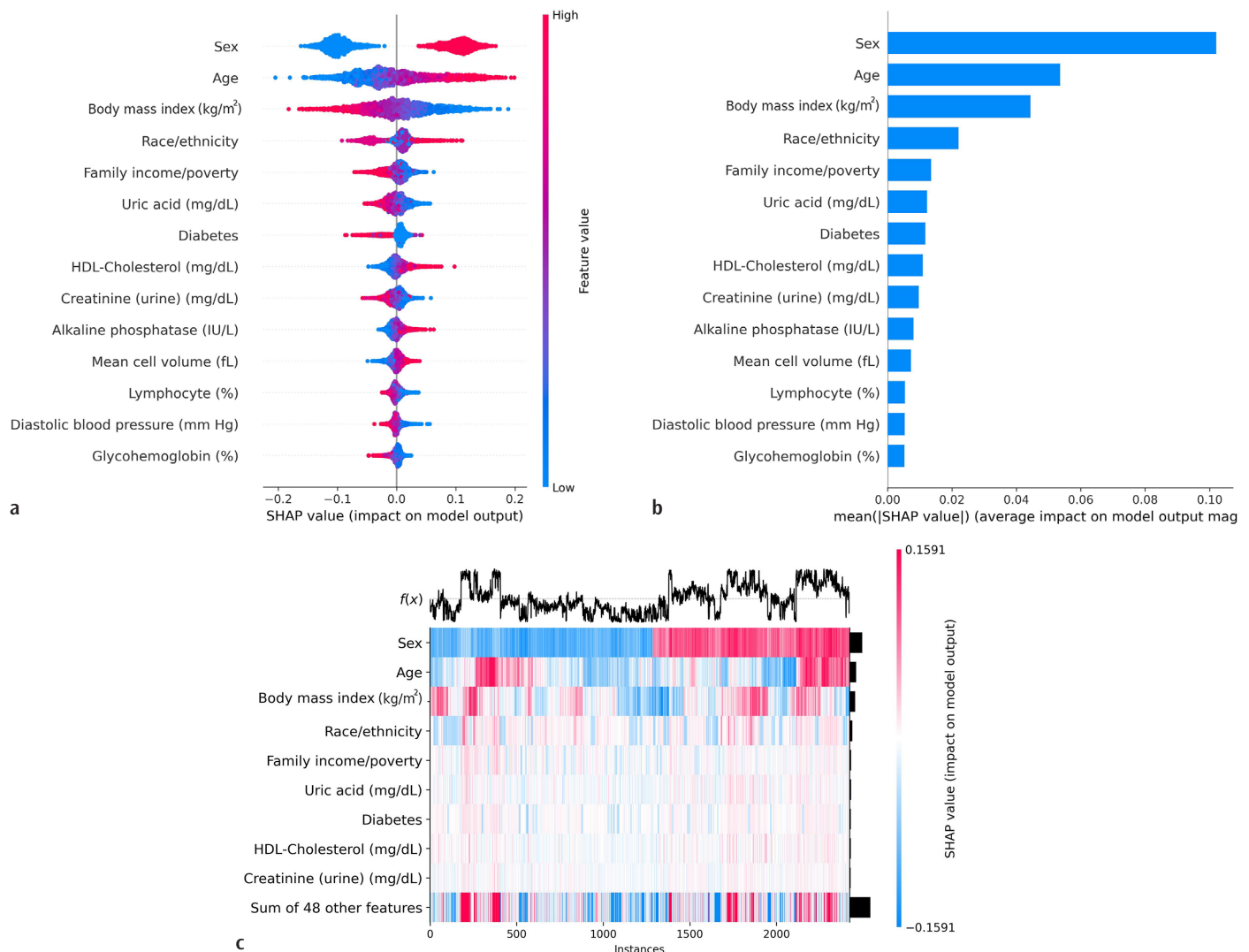BMD, bone mineral density.

Karaismailoğlu and Karaismailoğlu. Predicting Low Bone Density with Machine Learning

553



**FIG. 4.** (a) Beeswarm plot of Top 14 Features from SHAP importance analysis based on extra trees classifier. (b) Bar plot of the mean absolute SHAP values. (c) SHAP heatmap displaying participant-by-feature contributions.

Figure 5 demonstrates the model-derived influence of the 10 most important predictors on low BMD risk, visualized as SHAP-dependence plots with LOWESS smoothing. The levels of uric acid, BMI, and urinary creatinine demonstrated negative correlations; higher levels of all three were associated with a lower predicted risk of low BMD, particularly up to approximately 8 mg/dL for uric acid, 35 kg/m² for BMI, and 150 mg/dL for urinary creatinine. These relationships appear to plateau beyond these thresholds, suggesting a potential nonlinear or saturation effect in the model's predictions. The predicted risk of low BMD displayed an inverse relationship with the diabetes status. As shown in the SHAP-dependence plot, individuals with diagnosed diabetes (coded 2) had the lowest predicted risk of low BMD, followed by those with prediabetes (coded 1), compared with individuals without diabetes (coded 0). Non-Hispanic White individuals (3 coded) were most strongly associated with increased and predicted risk of low BMD, making

them the highest-risk ethnic category. The family income-to-poverty ratio showed an inverse relationship with predicted low BMD risk. In other words, a higher ratio was linked to a lower predicted risk. Higher HDL cholesterol and ALP levels were both associated with an increased predicted risk of low BMD.

Figure 6 shows sex-stratified SHAP importance profiles. In both graphs, age and BMI acted as the primary influencing factors, albeit the magnitude of their effects differed. For instance, higher BMI was associated with a greater reduction in the predicted risk in women (SHAP range: -0.30 to 0.00) compared to men (-0.15 to 0.00). In addition to the similar predictors, sex-specific differences were recorded among other influential features in the top 10, with urinary creatinine and ALP levels having a greater impact in men, whereas the serum globulin and C-reactive protein (CRP) levels were more influential in women.
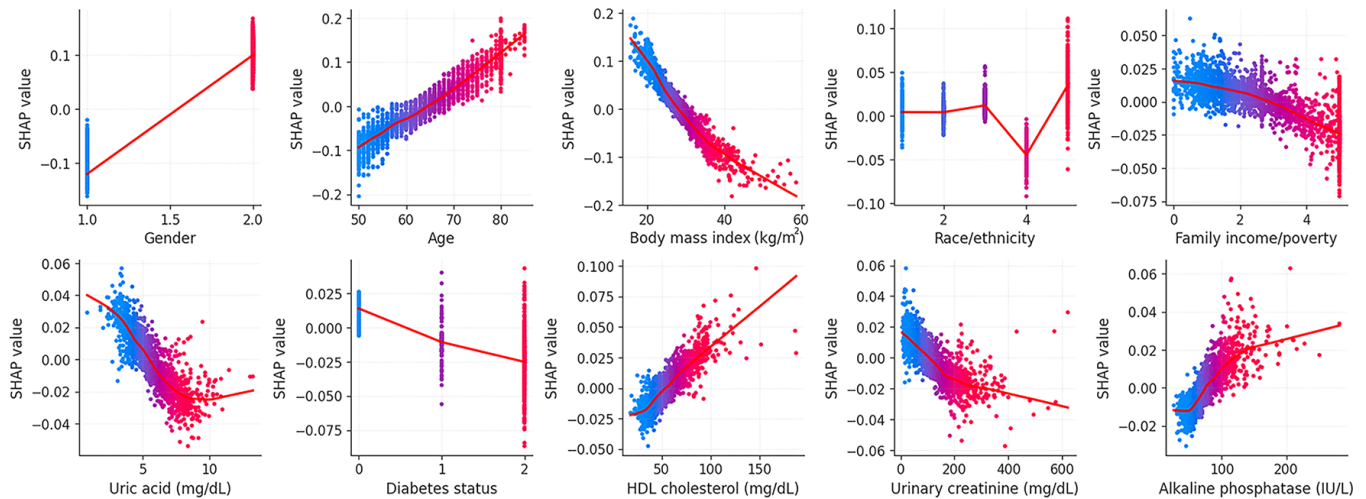
**FIG. 5.** SHAP dependence plots with LOWESS trends, Top 10 features of low BMD risk.
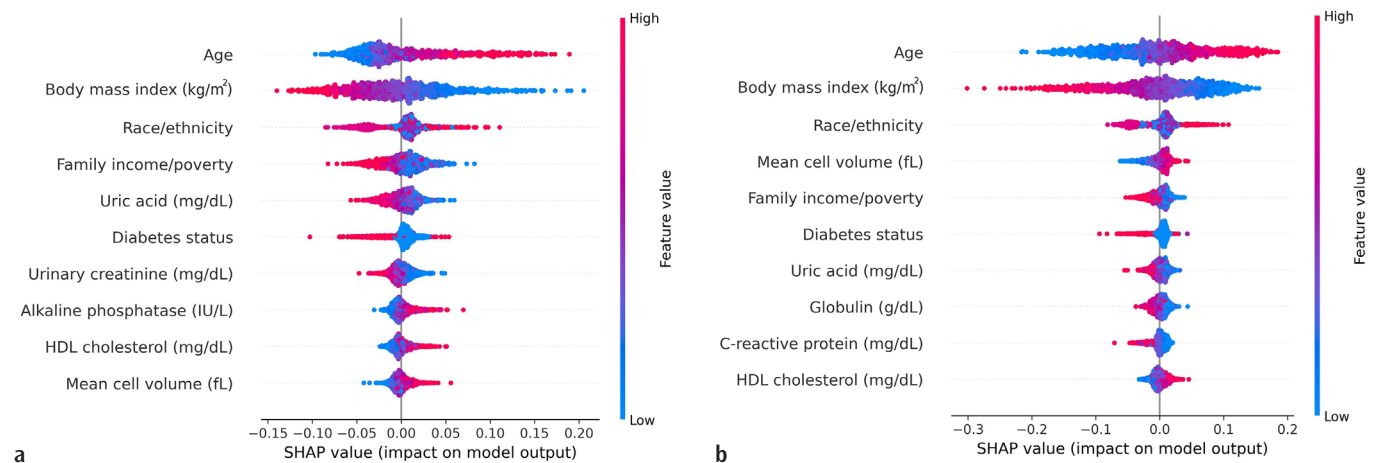*BMD, bone mineral density.*



**FIG. 6.** Sex stratified SHAP summaries showing the top 10 predictors of low BMD risk in (a). men and (b). women.
*BMD, bone mineral density.*

## DISCUSSION

In this study, we applied ML approaches to predict low BMD using demographic, clinical, and biochemical data from the NHANES dataset. Among eleven classifiers, ensemble tree-based models, particularly the ET, highlights the capability of ML to capture complex, nonlinear interactions. Similar findings have been reported in previous studies, where tree-based ML algorithms, such as Random Forest, XGBoost, and LightGBM, achieved superior performance in predicting osteoporosis or low BMD risk.[23-25] The Extra Trees algorithm has also exhibited strong predictive accuracy in other biomedical domains, including coronary artery disease, Hodgkin lymphoma, Parkinson's disease, cervical cancer, and *Helicobacter pylori*. These results emphasize its robustness and adaptability across complex and diverse patterns and datasets.[26-30]

Our findings reaffirmed well-established predictors such as age, sex, and BMI, which consistently show strong associations with osteoporosis risk.[31,32] Older age, female sex, and lower BMI were significantly associated with reduced BMD, consistent with prior evidence linking age-related bone loss and postmenopausal hormonal decline to reduced bone mass.[31-33]

Although univariate analyses (Table 2) revealed statistically significant differences for all variables, such results must be interpreted cautiously, given the large sample size, where small absolute differences may appear significant. For instance, variations in family income/poverty or lymphocyte percentage may hold limited clinical importance individually. However, this finding highlights the strength and rationale of the ML approach. Notably, the variables presented in Table 2 were identified through SHAP values as the top contributors in the ML models. Unlike the

Karaismailoğlu and Karaismailoğlu. Predicting Low Bone Density with Machine Learning

555

marginal comparisons shown in Table 2, the SHAP values quantify the conditional contribution of each feature, considering the interactions with all other predictors in the model. Consequently, a variable exhibited a small univariate effect, but continued to play a crucial role in the model's predictive power owing to its cumulative and interactive effects. This dual perspective emphasized the necessity of integrating traditional statistical comparisons with ML-based feature importance for a comprehensive interpretation of the predictors of low BMD.

Beyond conventional risk factors, SHAP analysis revealed additional, biologically plausible contributors. The present study's results align with the latest population-based analysis demonstrating an inverse (non-linear) association between the serum uric acid levels and BMD among Chinese and American cohorts, suggesting that higher uric acid may offer a protective effect on the bone mass up to a threshold level.[34] Considering the damaging consequences of oxidative stress on bone metabolism, the antioxidant properties of uric acid may represent a potential protective mechanism against low BMD.[35] Although HDL cholesterol is regarded as a protective factor in the context of cardiovascular diseases, recent studies have suggested that high HDL levels may be associated with an increased risk of osteoporosis.[36] Preclinical studies have indicated that HDL leads to reduced BMD by adversely affecting the osteoblast number and function.[37] Consistent with these findings, our study recorded that higher HDL levels, particularly above 65-70 mg/dL, were associated with a higher predicted risk of low BMD. Despite diabetes appearing as a risk factor in the low BMD group, the SHAP values exhibited an overall positive influence, indicating a higher BMD among diabetic individuals. This finding is consistent with those of several studies, particularly in relation to type-2 diabetes that report normal or even higher BMD values.[38,39] In addition, glycohemoglobin level reflects the average level of glucose in the blood over the past 2-3 months, and it serves to diagnose prediabetes. Similarly, our findings indicated that higher glycohemoglobin levels were inversely associated with lower BMD, as has been documented across several previous studies.[40,41]

Another notable finding was a positive association between urinary creatinine and BMD, which has most commonly been explored through ratio-based indices, such as urinary calcium/creatinine or ALP/creatinine, with emerging evidence supporting a direct association. For instance, Kim et al.[42] reported that, in patients with chronic kidney disease, higher urinary creatinine levels were positively correlated with BMD, implying that urinary creatinine could indicate maintained muscle mass and metabolic health, both of which are beneficial for skeletal integrity. Similarly, Schwaderer et al.[43] reported that children with low BMD exhibited significantly lower urinary creatinine levels. These findings are consistent with our results that identified urinary creatinine as a positive predictor of BMD, particularly in men, highlighting its potential utility as a noninvasive marker in osteoporosis risk assessment.

ALP displayed a nonlinear association with BMD, where levels up to 130 U/L correlated with sharp declines in BMD, followed by a plateau. This pattern is consistent with findings from Cheng and Zhao[44], who reported similar nonlinearity between ALP and BMD

using generalized additive models. However, in our sex-stratified analysis, ALP ranked higher among males, whereas its impact in females was relatively weaker, likely due to the dominance of hormonal and inflammatory factors.

Socioeconomic status, as quantified by the family income/poverty, demonstrated an inverse association with a low BMD risk, which is consistent with past reports linking lower income to poorer bone health outcomes as a result of reduced access to nutrition, healthcare, and physical activity.[45] The presence of such a gradient highlights the importance of incorporating social determinants into predictive modeling for osteoporosis. Similarly, racial and ethnic backgrounds have emerged as influential factors, reflecting disparities in bone health potentially stemming from genetic differences, cultural behaviors, or unequal access to healthcare resources.[46]

Sex-specific analyses revealed that age and BMI remained the strongest predictors in men and women, though the protective effect of BMI was more pronounced in men, while age exerted a stronger influence in women. Urinary creatinine and ALP ranked higher among men, whereas globulin and CRP were stronger predictors among women. Elevated CRP, a marker of inflammation, has been widely associated with lower BMD, particularly in women,[48] while globulin-related measures (e.g., albumin-to-globulin ratio) may reflect nutritional and inflammatory status relevant to osteoporosis risk.[47] Despite its strength, this study has several limitations. The cross-sectional design of NHANES restricts causal inference. Key variables such as genetic data, menopausal status, and physical activity were excluded due to missing data. Although SHAP values enhance interpretability, the biological implications of some nonlinear relationships warrant further investigation. Additionally, our analyses were conducted on an unweighted NHANES sample, and external validation in independent populations would strengthen generalizability. Nonetheless, the primary goal, to compare ML algorithms and identify influential predictors, was effectively achieved.

In conclusion, this study demonstrates the feasibility and interpretability of ML-based models, particularly tree-based algorithms, for accurate prediction of low BMD using routinely collected health data. The ET showed the highest predictive performance. By integrating SHAP-based interpretability, we uncovered biologically and socioeconomically meaningful predictors that could inform personalized screening strategies and public health interventions aimed at reducing the burden of osteoporosis and osteopenia.

# REFERENCES

1. Kanis JA. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. WHO study group. *Osteoporos Int.* 1994;4:368-381. [CrossRef]

2. Cotts KG, Cifu AS. Treatment of osteoporosis. *JAMA.* 2018;319:1040-1041. [CrossRef]

3. Lisk R, Yeong K. Reducing mortality from hip fractures: a systematic quality improvement programme. *BMJ Qual Improv Rep.* 2014;3:u205006.w2103. [CrossRef]

4. Adler RA. Osteoporosis in men: a review. *Bone Res.* 2014;2:14001. [CrossRef]

5. Si Z, Zhang D, Wang H, Zheng X. PrOsteoporosis: predicting osteoporosis risk using NHANES data and machine learning approach. *BMC Res Notes.* 2025;18(1):108. [CrossRef]

6. Xu R, Chen Y, Yao Z, et al. Application of machine learning algorithms to identify people with low bone density. *Front Public Health.* 2024;12:1347219. [CrossRef]

7. Nabavi A, Safari F, Faramarzi A, et al. Machine learning analysis of cardiovascular risk factors and their associations with hearing loss. *Sci Rep.* 2025;15(1):9944. [CrossRef]

8. Guo K, Ni W, Du L, Zhou Y, Cheng L, Zhou H. Environmental chemical exposures and a machine learning-based model for predicting hypertension in NHANES 2003-2016. *BMC Cardiovasc Disord.* 2024;24(1):544. [CrossRef]

9. Liu Y, Li K, Zhang Y, et al. Impact of phthalate exposure and blood lipids on breast cancer risk: machine learning prediction. *Environ Sci Eur.* 2025;37:39. [CrossRef]

10. Hinton BJ, Fan B, Ng BK, Shepherd JA. Dual energy X-ray absorptiometry body composition reference values of limbs and trunk from NHANES 1999-2004 with additional visualization methods. *PLoS One.* 2017;12:e0174180. [CrossRef]

11. Looker AC, Wahner HW, Dunn WL, et al. Updated data on proximal femur bone mineral levels of US adults. *Osteoporos Int.* 1998;8(5):468-489. [CrossRef]

12. Zoppi T, Gazzini S, Ceccarelli A. Anomaly-based error and intrusion detection in tabular data: no DNN outperforms tree-based classifiers. *Future Generation Computer Systems.* 2024;160:951–965. [CrossRef]

13. Nakagomi A, Kondo K, Shiba K. Heterogeneity in the association between internet use and dementia among older adults: a machine-learning analysis. *Arch Gerontol Geriatr.* 2025;136:105912. [CrossRef]

14. Sai MJ, Chettri P, Panigrahi R, Garg A, Bhoi AK, Barsocchi P. An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes. *Int J Comput Intell Syst.* 2023;16:14. [CrossRef]

15. J M SL, P S. Unveiling the potential of machine learning approaches in predicting the emergence of stroke at its onset: a predicting framework. *Sci Rep.* 2024;14(1):20053. [CrossRef]

16. Anghele AD, Marina V, Dragomir L, Moscu CA, Anghele M, Anghel C. Predicting deep venous thrombosis using artificial ıntelligence: a clinical data approach. *Bioengineering (Basel).* 2024;11:1067. [CrossRef]

17. Zhao S, Zhang B, Yang J, Zhou J, Xu Y. Linear discriminant analysis. *Nat Rev Methods Primers.* 2024;4. [CrossRef]

18. Li N, Peng E, Liu F. Prediction of lymph node metastasis in cervical cancer patients using AdaBoost machine learning model: analysis of risk factors. *Am J Cancer Res.* 2025;15:1158-1173. [CrossRef]

19. Peng C, Cheng Q. Discriminative Ridge Machine: A Classifier for High-Dimensional Data or Imbalanced Data. *IEEE Trans Neural Netw Learn Syst.* 2021;32:2595-2609. [CrossRef]

20. Du KL, Jiang B, Lu J, Hua J, Swamy MNS. Exploring kernel machines and support vector machines: principles, techniques, and future directions. *Mathematics.* 2024;12:3935. [CrossRef]

21. Hou R, Wang X, Zhang W, et al. Decision tree extraction for clinical decision support system with ıf-else pseudocode and planselect strategy. *IEEE J Biomed Health Inform.* 2025;29:3642-3653. [CrossRef]

22. Boddu AS, Jan A. A systematic review of machine learning algorithms for breast cancer detection. *Tissue Cell.* 2025;95:102929. [CrossRef]

23. He W, Chen S, Fu X, Xu L, Xie J, Wan J. Development and evaluation of interpretable machine learning regressors for predicting femoral neck bone mineral density in elderly men using NHANES data. *Biomol Biomed.* 2025;25:375-390. [CrossRef]

24. Jin W, Xu L, Yue C, et al. Development and validation of explainable machine learning models for female hip osteoporosis using electronic health records. *Int J Med Inform.* 2025;199:105889. [CrossRef]

25. Shi H, Fang Y, Ma X. Application of machine learning algorithms in osteoporosis analysis based on cardiovascular health assessed by life's essential 8: a cross-sectional study. *J Health Popul Nutr.* 2025;44:180. [CrossRef]

26. Lyu Y, Wu HM, Yan HX, et al. Classification of coronary artery disease using radial artery pulse wave analysis via machine learning. *BMC Med Inform Decis Mak.* 2024;24:256. [CrossRef]

27. Albano D, Cuocolo R, Patti C, et al. Whole-body MRI radiomics model to predict relapsed/refractory Hodgkin Lymphoma: a preliminary study. *Magn Reson Imaging.* 2022;86:55-60. [CrossRef]

28. Rojas-Velazquez D, Kidwai S, Liu TC, et al. Understanding Parkinson's: the microbiome and machine learning approach. *Maturitas.* 2025;193:108185. [CrossRef]

29. Wang Q, Liang T, Li Y, Zhou P, Liu X. Machine learning for prediction of Helicobacter pylori infection based on basic health examination data in adults: a retrospective study. *Front Med (Lausanne).* 2025;12:1587540. [CrossRef]

30. Arage FG, Tadese ZB, Taye EA, Tsegaw TK, Abate TG, Alemu EA. Cervical cancer screening uptake and its associated factor in Sub-Sharan Africa: a machine learning approach. BMC Med *Inform Decis Mak.* 2025;25:197. [CrossRef]

31. Wang J, Shu B, Tang DZ, et al. The prevalence of osteoporosis in China, a community based cohort study of osteoporosis. *Front Public Health.* 2023;11:1084005. [CrossRef]

32. Piñar-Gutierrez A, García-Fontana C, García-Fontana B, Muñoz-Torres M. Obesity and Bone Health: A Complex Relationship. *Int J Mol Sci.* 2022;23:8303. [CrossRef]

33. Kanto A, Kotani Y, Murakami K, et al. Risk factors for future osteoporosis in perimenopausal Japanese women. *Menopause.* 2022;29:1176-1183. [CrossRef]

34. Li X, Peng Y, Chen K, Zhou Y, Luo W. Association between serum uric acid levels and bone mineral density in Chinese and American: a cross-sectional study. *Sci Rep.* 2025;15:8304. [CrossRef]

35. Kimball JS, Johnson JP, Carlson DA. Oxidative stress and osteoporosis. *J Bone Joint Surg Am.* 2021;103:1451-1461. [CrossRef]

36. Tang Y, Wang S, Yi Q, Xia Y, Geng B. High-density lipoprotein cholesterol ıs negatively correlated with bone mineral density and has potential predictive value for bone loss. *Lipids Health Dis.* 2021;20:75. [CrossRef]

37. Hussain SM, Ebeling PR, Barker AL, Beilin LJ, Tonkin AM, McNeil JJ. Association of plasma high-density lipoprotein cholesterol level with risk of fractures in healthy older adults. *JAMA Cardiol.* 2023;8:268-272. [CrossRef]

38. Zoulakis M, Johansson L, Litsne H, Axelsson K, Lorentzon M. Type 2 diabetes and fracture risk in older women. *JAMA Netw Open.* 2024;7:e2425106. [CrossRef]

39. Oei L, Zillikens MC, Dehghan A, et al. High bone mineral density and fracture risk in type 2 diabetes as skeletal complications of inadequate glucose control: the Rotterdam Study. *Diabetes Care.* 2013;36:1619-1628.[CrossRef]

40. Ji X, Hong J, Qu Z, et al. HemoglobinA1c ıs a risk factor for changes of bone mineral density: a mendelian randomization study. *Front Endocrinol (Lausanne).* 2022;13:942878. [CrossRef]

41. Gao L, Liu Y, Li M, Wang Y, Zhang W. Based on HbA1c analysis: bone mineral density and osteoporosis risk in postmenopausal female with T2DM. *J Clin Densitom.* 2024;27:101442. [CrossRef]

42. Kim SE, Jung SH, Yang J, et al. Association between urine creatinine excretion and bone mineral density in chronic kidney disease: Results from the KNOW-CKD study. *J Nephrol.* 2024;38:189–196. [CrossRef]

43. Schwaderer AL, Cronin R, Mahan JD, Bates CM. Low bone density in children with hypercalciuria and/or nephrolithiasis. *Pediatr Nephrol.* 2008;23:2209-2214. [CrossRef]

44. Cheng X, Zhao C. The correlation between serum levels of alkaline phosphatase and bone mineral density in adults aged 20 to 59 years. *Medicine (Baltimore).* 2023;102:e34755. [CrossRef]

45. Gough Courtney M, Roberts J, Godde K. Structural ınequity and socioeconomic status link to osteoporosis diagnosis in a population-based cohort of middle-older-age Americans. *Inquiry.* 2023;60:469580231155719. [CrossRef]

46. Noel SE, Santos MP, Wright NC. Racial and ethnic disparities in bone health and outcomes in the United States. *J Bone Miner Res.* 2021;36:1881-1905. [CrossRef]

47. Yan H, Yang J, Dai Y, Li Y, Zhang P, Li L. The association between the albumin to globulin ratio and thoracic spine bone mineral density in adolescents: NHANES 2011-2016. *Front Nutr.* 2025;12:1560013. [CrossRef]

48. Ilesanmi-Oyelere BL, Schollum L, Kuhn-Sherlock B, et al. Inflammatory markers and bone health in postmenopausal women: a cross-sectional overview. *Immun Ageing.* 2019;16:15. [CrossRef]