Computers in Biology and Medicine 192 (2025) 110289

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Enhancing osteoporosis risk prediction using machine learning: A holistic approach integrating biomarkers and clinical data

Filipe Ricardo Carvalho [a,b,c,*], Paulo Jorge Gavaia [a,b,c]

[a] Faculty of Medicine and Biomedical Sciences, University of Algarve, Faro, Portugal
[b] Centre of Marine Sciences (CCMAR/CIMAR LA), University of Algarve, Faro, Portugal
[c] University of Algarve - Campus de Gambelas, 8005-139, Faro, Portugal

ARTICLE INFO

Keywords:
Machine learning
Osteoporosis risk prediction
Biomarkers
NHANES
Stacking ensemble
Bone mineral density
Preventive medicine

ABSTRACT

Osteoporosis (OP) affects approximately 18 % of the global population, with osteoporosis-associated fractures impacting up to 37 million people annually. While dual-energy X-ray absorptiometry (DXA) remains the gold standard for diagnosis, its limitations, including restricted availability and radiation exposure, highlight the need for alternative screening methods. We developed a machine learning model to predict OP risk using routinely collected clinical data, deliberately excluding DXA measurements to ensure broad accessibility. Using data from NHANES cycles 2007–2014, we analyzed 7924 participants aged 50 years and older, identifying 1636 OP cases (20.6 %) and 6288 normal cases (79.4 %) through comprehensive criteria incorporating both WHO densitometric standards (T-scores ≤ -2.5) and anthropometric risk factors. We implemented a stacking ensemble model combining four specialized classifiers (Gradient Boosting, Random Forest, XGBoost, and LightGBM) with a logistic regression meta-classifier. The model achieved 93 % accuracy, an AUC of 0.94, and demonstrated robust performance through cross-validation (mean score: 0.929 ± 0.030). feature importance analysis revealed age (6.04 %), arm muscle circumference (5.61 %), and body weight (5.30 %) as the most influential predictors, followed by gender (3.28 %), BMI (2.71 %), and calcium intake (2.42 %). Additional significant predictors included folate (2.28 %), height (2.23 %), hand grip strength (2.21 %), and alkaline phosphatase (2.16 %). These biologically plausible relationships align with established clinical knowledge of OP risk factors. The model's strong performance metrics and reliance on readily available clinical data suggest its potential as a practical screening tool, particularly in settings with limited DXA access. All code and implementation details are openly available on GitHub, facilitating integration into existing healthcare systems. This approach offers a promising pathway for enhancing early OP detection and risk assessment across diverse healthcare settings.

1. Introduction

Osteoporosis (OP), characterized by decreased bone mineral density (BMD) and compromised bone structure, was estimated to have a prevalence worldwide of 18.3 % between 103,334,579 individuals aged 15–105 years [1]. Also, osteoporosis-associated fractures affect up to 37 million annually in individuals aged over 55 [2]. Despite women having a higher susceptibility to OP [1], men face greater adverse outcomes following fractures, particularly in old age [3]. With global populations aging [4], OP emerges as a pressing public health issue [5,6]. By 2050, hip fracture incidence worldwide is projected to surge by threefold in men and twofold in women compared to 1990 [7,8]. The considerable risk of fracture, often accompanied by increased disability and mortality, contributes significantly to the socioeconomic burden associated with bone health [9–11].

The World Health Organization (WHO) suggests dual-energy X-ray absorptiometry (DXA) as the preferred method for assessing BMD in diagnosing OP [12,13], serving as the gold standard, and the primary approach for evaluating fracture risk [14]. Since 1994, WHO established OP as a BMD 2.5 or more standard deviations below the average for young healthy adults. The National Health and Nutrition Examination Survey (NHANES) has incorporated DXA measurements since 1999, providing valuable population-level data on bone health status [15,16] While FRAX (fracture risk assessment tool) is a valuable tool in OP management and research shows that it can, without BMD data, accurately identify candidates for treatment [17], it still fails to detect

* Corresponding author. Faculty of Medicine and Biomedical Sciences, University of Algarve, Faro, Portugal
E-mail address: frcarvalho@ualg.pt (F.R. Carvalho).

https://doi.org/10.1016/j.compbiomed.2025.110289
Received 26 July 2024; Received in revised form 23 April 2025; Accepted 25 April 2025
Available online 5 May 2025

fracture risk in patients with several different subclinical conditions [18–21]. Additionally, limited access to DXA services, even in developed countries, and concerns about radiation exposure [22,23] highlight the need for alternative screening methods.

Artificial intelligence (AI) has achieved remarkable results throughout medicine, particularly in orthopedics [24–27]. Recent studies using NHANES data have demonstrated AI's potential in predicting OP risk [28]. As technological advancements continue to unfold, it is increasingly evident that AI can surpass clinicians' capabilities under specific circumstances [29] and potentially support the development of personalized treatments [30]. Despite these advancements, a notable gap persists: the absence of a reliable, cost-effective tool that can be integrated into hospital settings to detect diminished BMD without requiring additional diagnostic exams.

Utilizing machine learning techniques and the comprehensive NHANES dataset, our proposed model seeks to predict OP by leveraging routinely collected patient data such as biomarkers, anthropometric measurements, and clinical history, explicitly excluding any DXA-derived measurements to ensure broad clinical applicability. This approach is particularly valuable as NHANES provides a nationally representative sample, enhancing the generalizability of our findings. By streamlining the diagnostic process through readily available clinical data and completely independent from specialized bone density imaging, our model has the potential to enhance early detection of OP risk, particularly in settings where DXA access is limited or unavailable, democratizing osteoporosis screening across different healthcare settings.(See Fig. 1)

2. Material and methods

The data used in this study were obtained from the National Health and Nutrition Examination Survey (NHANES) cycles 2007–2008, 2009–2010, and 2013–2014. NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States, conducted by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC).

Although more recent NHANES data (2017–2020) were available, these cycles were not included in our analysis due to the absence of several crucial variables that emerged as significant predictors during our data processing and model development phase. This decision was primarily driven by the need to maintain consistency in the predictor variables across all analyzed cycles and to ensure the robustness of our predictive model.

2.1. Defining the OP target variable

Due to the complexity of OP diagnosis, we developed a comprehensive classification approach that integrates both densitometric and anthropometric criteria. According to WHO criteria, OP is traditionally defined by T-scores ≤ -2.5 at any skeletal site [31]. Our analysis incorporated T-scores from three key anatomical locations: femoral neck (FN), total femur (TF), and spine (SP) (showing differences between normal and osteoporotic groups: FN: -0.88 ± 0.90 vs. -2.07 ± 0.84; TF: -0.41 ± 0.97 vs. -1.79 ± 0.95; SP: -0.18 ± 1.16 vs. -1.54 ± 1.25, respectively, Table 1).

To enhance the diagnostic sensitivity, we created a binary classification variable (target) that identified individuals as having OP based on the following criteria (Fig. 2).

1. Densitometric criterion:
   - o T-score ≤ -2.5 at any of the measured sites (FN, TF, or SP)
2. Anthropometric criteria:
   - o For female participants only: Advanced age (subjects in the highest 15th percentile, showing 31.8 % OP prevalence (Fig. 3 A.2).
   - o Extremely low body weight (subjects in the lowest 15th percentile of weight distribution, showing 37.8 % OP prevalence) (Fig. 3 B).
   - o Severely reduced arm muscle circumference (subjects in the lowest 15th percentile, showing 36.8 % OP prevalence) (Fig. 3 C)

Table 1
Demographic and clinical characteristics of the study population. The data presents characteristics of 7924 participants exclusively aged 50 years and older from NHANES (2007–2014), categorized into normal (n = 6,288) and OP(n = 1,636) groups. Values are presented as mean ± standard deviation for continuous variables and n (%) for categorical variables. OP group shows higher mean age (69.7 vs 60.8 years), predominance of females (66.9 % vs 33.1 %), and lower anthropometric measurements compared to the normal group. T-scores across all skeletal sites were lower in the OP group.

	Total (n = 7924)	Normal (n = 6288)	OP (n = 1636)
Age	62.7 ± 10.9	60.8 ± 10.2	69.7 ± 10.8
Male	3916 (49.4 %)	3374 (53.7 %)	542 (33.1 %)
Female	4008 (50.6 %)	2914 (46.3 %)	1094 (66.9 %)
Weight (kg)	80.4 ± 18.8	84.7 ± 17.6	63.6 ± 13.2
Height (cm)	166.3 ± 10.2	167.8 ± 9.9	160.7 ± 9.3
Arm Circumference (cm)	32.8 ± 4.7	33.9 ± 4.2	28.5 ± 4.0
Mexican American	1151 (14.5 %)	958 (15.2 %)	193 (11.8 %)
Other Hispanic	762 (9.6 %)	609 (9.7 %)	153 (9.4 %)
Non-Hispanic White	3924 (49.5 %)	3051 (48.5 %)	873 (53.4 %)
Non-Hispanic Black	1526 (19.3 %)	1291 (20.5 %)	235 (14.4 %)
Race-Other	561 (7.1 %)	379 (6.0 %)	182 (11.1 %)
T-score Femoral Neck	−1.13 ± 1.01	−0.88 ± 0.90	−2.07 ± 0.84
T-score Total Femur	−0.70 ± 1.11	−0.41 ± 0.97	−1.79 ± 0.95
T-score Spine	−0.46 ± 1.30	−0.18 ± 1.16	−1.54 ± 1.25

This classification approach resulted in the identification of 1636 OP cases (20.6 %) and 6288 normal cases (79.4 %), suggesting an effective stratification of the study population based on established bone health indicators and recognized anthropometric risk factors.

2.2. Reference values calculation

Our T-score calculation methodology evolved through the study phases. Initially, we used established reference values from NHANES III for young adults across different demographic groups. In the second phase, to ensure more accurate T-score calculations for our expanded dataset, we developed updated reference values based on young adult (20–30 years) BMD measurements from our NHANES 2007–2014 sample. These new reference values were calculated separately for each anatomical site (femoral neck, total femur, and spine), considering both gender and ethnic variations (White, Black, and Hispanic). The calculations included mean BMD values and standard deviations for each combination of site, gender, and ethnicity, derived from our young adult reference population (code used to develop reference values can be accessed in (reference_values_t_score.ipynb, Supplementary Material Table S2). This refined approach allowed us to generate population-specific T-scores that better reflect the demographic characteristics of our study population (Supplementary Table S2).

2.3. Streamlining variable selection

In this study, our primary objective was to develop a predictive model for OP that could operate effectively with minimal and non-redundant information, deliberately excluding any DXA-derived measurements (including BMD and T-scores) from the training process to ensure the model's independence from specialized imaging equipment and its accessibility in various clinical settings. The variable selection process was conducted in two distinct phases:

In the first phase, we utilized all available variables from the NHANES 2013–2014 dataset. To determine the optimal combination of features, sampling methods, and model parameters, we implemented a comprehensive machine learning pipeline that evaluated multiple classifiers (including Gradient Boosting, Random Forest, XGBoost, and

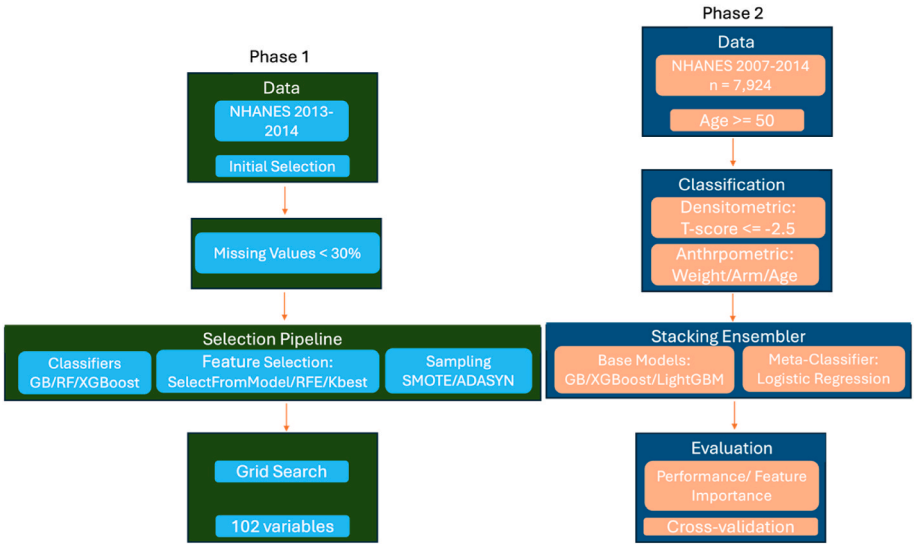

Fig. 1. Comprehensive Model Development Workflow (Phase 1). Workflow of the initial development phase using NHANES 2013–2014 dataset, implementing multiple classifiers, feature selection methods, and sampling techniques, resulting in the identification of 102 relevant variables for osteoporosis prediction. **(Phase 2).** Extended analysis utilizing NHANES 2007–2014 (n = 7924), incorporating both densitometric and anthropometric criteria for classification. The stacking ensemble combined four base models through a logistic regression meta-classifier, with comprehensive performance evaluation.

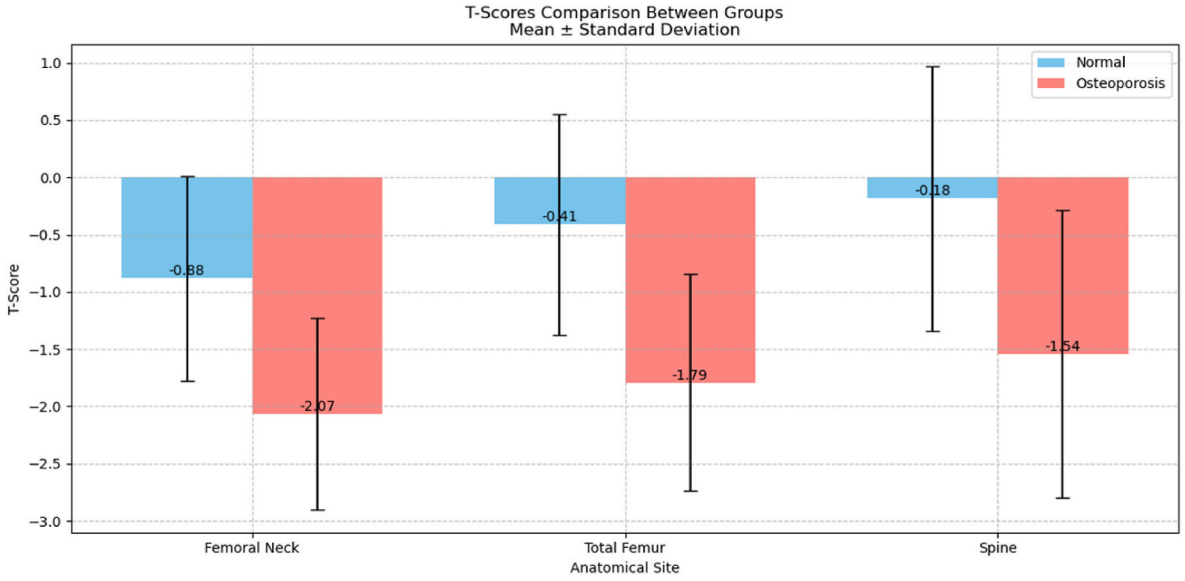

Fig. 2. Comparison of T-scores between normal and OP groups across anatomical sites. Bar graph showing mean T-scores (± standard deviation) for femoral neck (normal: −0.88 ± 0.90; osteoporosis: −2.07 ± 0.84), total femur (normal: −0.41 ± 0.97; osteoporosis: −1.79 ± 0.95), and spine (normal: −0.18 ± 1.16; osteoporosis: −1.54 ± 1.25). Results demonstrate consistently lower T-scores in the OP group across all measurement sites.

others), various feature selection methods (SelectFromModel, RFE, SelectKBest), and sampling techniques (SMOTE and ADASYN) to address class imbalance (pipeline_selector_classifier, Supplementary Material Table S3). This systematic evaluation helped identify the most effective combinations of methods for our specific prediction task.

Following the pipeline analysis, we further refined our model through an extensive grid search process, which systematically explored different hyperparameter combinations for the best-performing model configuration. This grid search included optimizing parameters such as the number of estimators, learning rate, tree depth, and sampling parameters, while also fine-tuning the feature selection threshold to maintain model performance with the minimal necessary set of features (grid_search_phase_1, Supplementary Material Table S4). Through this comprehensive approach, we identified several key predictive variables including BPX_H_BPACSZ (arm circumference for blood pressure measurement), DEMO_H_RIDAGEYR (age), BMX_H_BMXARMC (arm muscle circumference), DEMO_H_RIAGENDR (gender), BMX_H_BMXWT (weight), and other anthropometric and demographic variables, along with specific clinical and lifestyle indicators such as WHQ_H_WHQ030 (waist circumference), MCQ_H_MCQ080 (general health condition), and OSQ_H_OSQ080 (OP history) (Fig. S1). The selector identified a total of 102 variables that are most relevant to OP diagnosis (link: selected_variables_NHANES_2013_2014) (link_to_final_code_1st_phase: preprocessing_and_model_NHANES_13_14).

In the second phase, we expanded our analysis by combining these identified key variables from NHANES 2013–2014 with corresponding data from earlier NHANES cycles (2007–2008 and 2009–2010). This allowed us to validate and further refine our model using data from

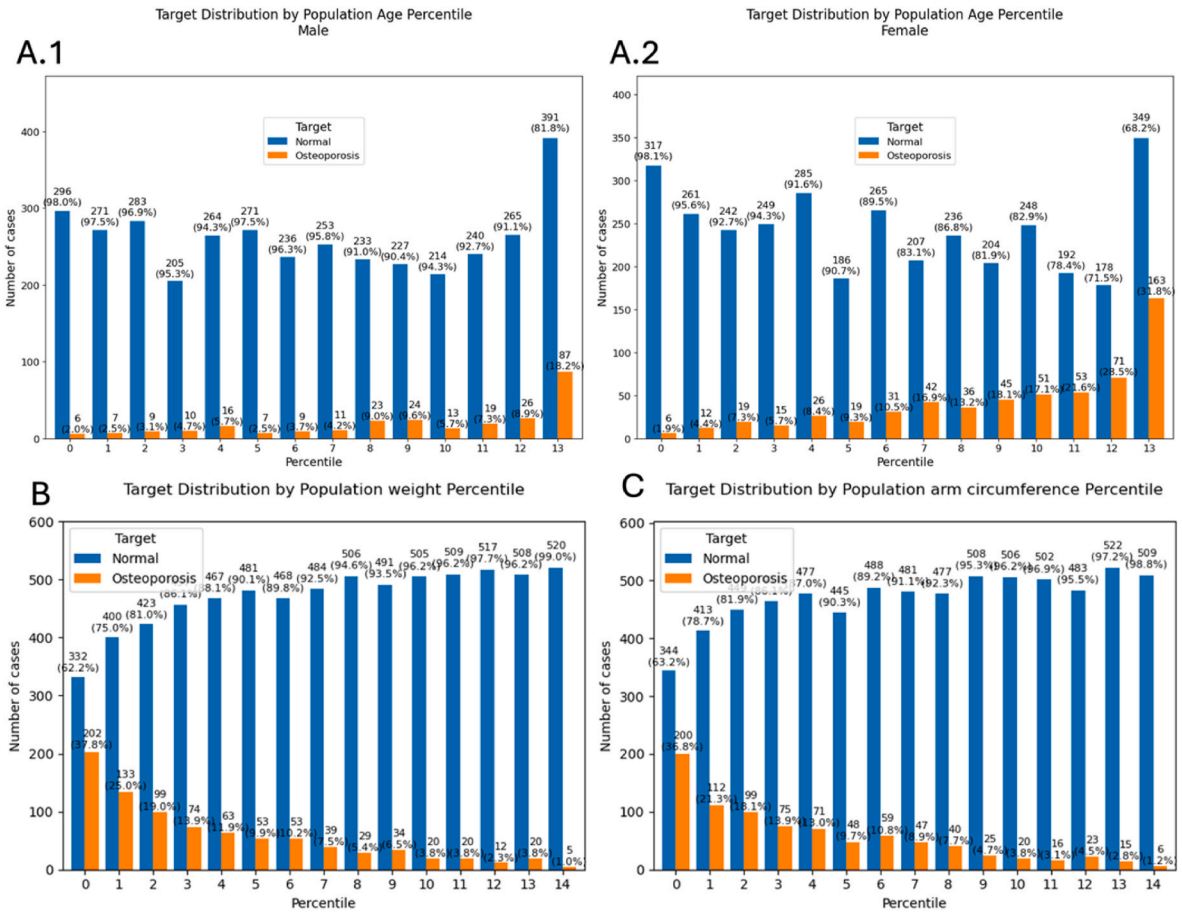

Fig. 3. Distribution of OP cases by Age and anthropometric percentiles. Bar graphs showing the percentage of OP cases across (A.1 male, A.2 female) age in years (RIAGENDR), (B) body weight (BMXWT) and (C) arm muscle circumference (BMXARMC) percentiles. The data reveals higher prevalence of OP in lower percentiles for both measurements, with notable peaks in the lowest percentile (BMXARMC: 36.8 %; BMXWT: 37.8 %) and gradual decrease towards higher percentiles. The clear gradient effect supports the validity of using these anthropometric measures as risk indicators in OP classification.

multiple NHANES cycles, enhancing its robustness and generalizability.

2.4. Data preparation and preprocessing

2.4.1. Initial data processing and feature selection

The initial NHANES 2013–2014 dataset, containing approximately 6906 variables, underwent comprehensive preprocessing to ensure data quality and suitability for machine learning analysis. Our first step involved a systematic feature selection process guided by clinical relevance to OP assessment. We retained variables with less than 30 % missing values. The following descriptions of data preparation and preprocessing steps refer to the second phase of our study, where we combined these identified variables across three NHANES cycles (2007–2008, 2009–2010, and 2013–2014 link: final_dataset) to develop and validate our final predictive model.

2.4.2. Categorical data conversion

For categorical data conversion, we employed the LabelEncoder from scikit-learn instead of one-hot encoding. The LabelEncoder transforms categorical variables into sequential integer values {0, …, n_classes-1}, where n_classes represents the number of unique categories in the variable. This choice was motivated by computational efficiency considerations and the need to prevent dimensionality expansion, particularly given our use of tree-based models in the ensemble.

2.4.3. Missing value imputation

Missing value handling was implemented through a two-step approach using K-Nearest Neighbors (KNN) imputation. The KNN imputation algorithm calculates missing values for each sample x using the mean value from its k nearest neighbors. For a given sample with missing values, the distance to all other samples is computed using the following Euclidean distance metric: $d(x, y) = \sqrt{(\Sigma(x_i - y_i)^2)}$. We used n_neighbors = 5, chosen to balance accuracy and computational efficiency.

2.4.4. Data integration

A crucial aspect of our methodology involved the integration of data from previous NHANES cycles (2007–2008 and 2009–2010) with our preprocessed 2013–2014 dataset. This integration required careful harmonization of variable names, verification of measurement unit consistency, and alignment of categorical variable coding across cycles. The merged dataset underwent rigorous validation to ensure population characteristics remained consistent across cycles.

2.4.5. Data Splitting and class balancing

For model development, we implemented a two-stage split approach. In the first stage, for training the base models and generating meta-features, the dataset was split into training (80 %) and validation (20 %) sets. In the second stage, for evaluating the final stacking model's performance, we used a 70/30 split ratio (training/testing). To address class imbalance, we applied SMOTE to the training data. SMOTE generates synthetic samples x_new for the minority class using: $x\_new = x + \lambda(\hat{x} - x)$ where: x is a sample from the minority class; $\hat{x}$ is one of its k-nearest neighbors; λ is a random number in [0,1]. Additionally, we

implemented class-specific weights w_i during model training: w_i = 5 for minority class (osteoporosis), w_i = 1 for majority class (normal) for the Gradient Boosting and Random Forest base models.

2.5. Stacking ensemble architecture

We developed a stacking ensemble model comprising four specialized base classifiers and a meta-classifier. The base layer consists of 1. Gradient Boosting Classifier (learning rate = 0.15, n_estimators = 170), 2. Random Forest Classifier (n_estimators = 200, max_depth = 10), 3. XGBoost Classifier (learning rate = 0.12, n_estimators = 150), 4. LightGBM Classifier (learning rate = 0.12, n_estimators = 150).

These base models were chosen for their complementary strengths in handling complex medical data. The meta-classifier, implemented as a Logistic Regression model (C = 0.5, L2 penalty), combines the predictions from the base models to produce the final classification. For the stacking ensemble probability calculation: $P(y|x) = \sigma(\beta_0 + \Sigma_i \beta_i f_i(x))$.

2.6. Model evaluation

Our evaluation strategy incorporated multiple complementary approaches.

2.6.1. Performance metrics

The model's performance was evaluated using a comprehensive set of metrics to ensure robust assessment. For primary classification assessment, we calculated accuracy, precision, recall (sensitivity), F1-score, and ROC-AUC to assess the model's discriminative ability across different probability thresholds.

Statistical validation was implemented through 5-fold stratified cross-validation. We performed class-specific analysis, evaluating performance separately for normal and osteoporotic cases, including false positive and negative rates. Performance evaluation was conducted on a held-out test set using scikit-learn implementation, with stratified sampling employed to maintain representative class distributions

Table 2
Performance metrics of the stacking ensemble model for OP prediction. Support indicates the number of samples in each class. The model shows strong overall performance with strength in identifying non-OP cases (Class 0). The high ROC AUC score of 0.9377 indicates excellent diagnostic capability. The difference between recall values (0.98 vs 0.72) suggests better sensitivity for non-OP cases, while maintaining high precision for both classes.

Class	Precision	Recall	F1-Score	Support
Non-OP	0.93	0.98	0.95	1258
OP	0.90	0.72	0.80	327
Accuracy	–	–	0.93	1585
Macro Average	0.92	0.85	0.88	1585
Weighted Average	0.92	0.93	0.92	1585
ROC AUC	–	–	0.9377	–

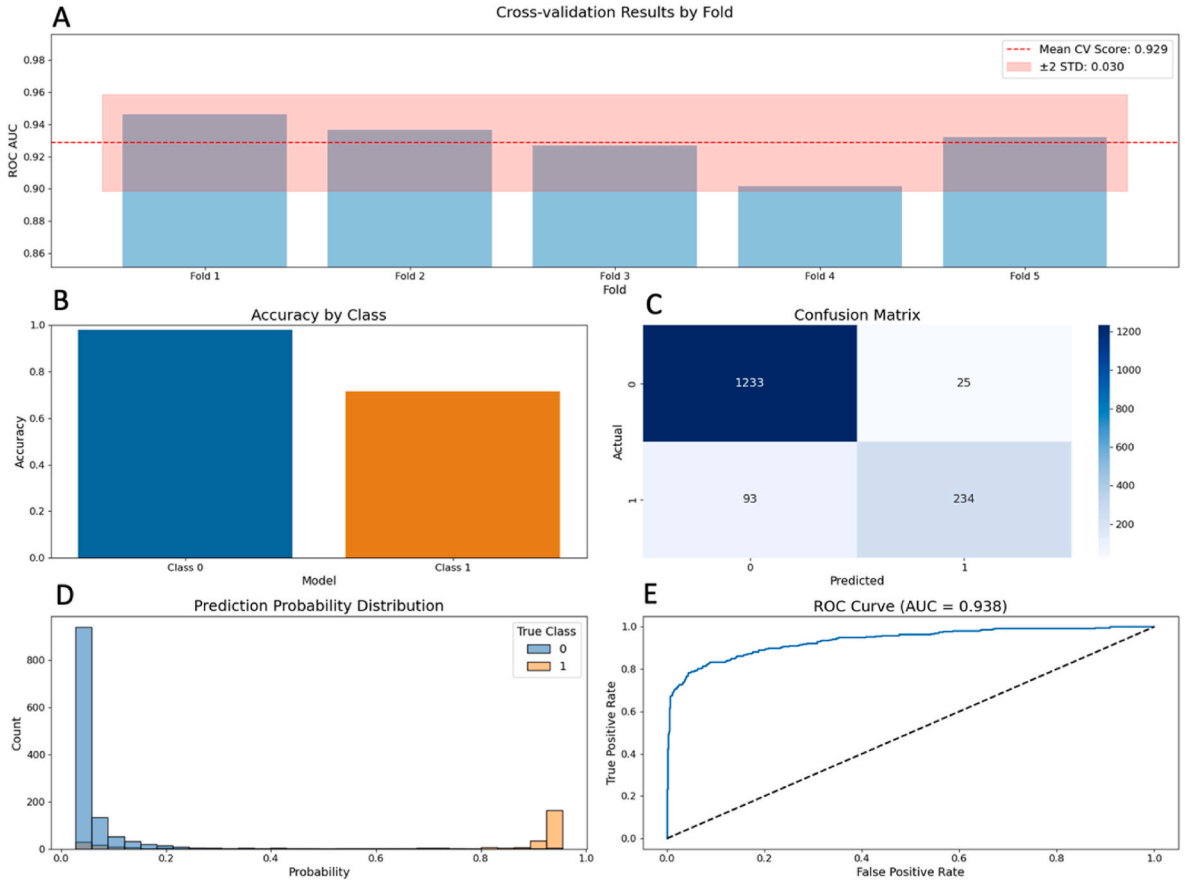

Fig. 4. Multi-panel visualization of the stacking model performance for OP prediction. (A) Cross-validation results showing ROC AUC scores across 5 folds, with scores ranging from 0.9016 to 0.9461, mean CV score of 0.9287 (±0.0298). (B) Accuracy comparison between Class 0 (non-osteoporosis) and Class 1 (osteoporosis) predictions, where Class 0 achieves 0.98 recall and 0.93 precision, while Class 1 shows 0.72 recall and 0.90 precision, indicating stronger performance in identifying non-OP cases. (C) Confusion matrix revealing detailed classification outcomes: 1232 true negatives, 91 false positives, 26 false negatives, and 236 true positives, from a total of 1585 test cases. (D) Probability distribution of predictions stratified by true class labels, showing the model's confidence level in its predictions and the separation between classes. (E) ROC curve analysis demonstrating the discrimination ability of the model with an AUC score of 0.9377, significantly above the random classifier baseline (0.5), indicating excellent diagnostic capability. The ensemble model shows strong overall performance with a weighted average F1-score of 0.92, suggesting reliable clinical applicability.

(Fig. 4, Table 2).

2.6.2. Feature importance analysis

We conducted feature importance analysis using the ensemble model's built-in feature importance estimators. The analysis quantified the relative contribution of each predictor variable to the model's decisions, focusing on anthropometric measurements, demographic characteristics, biochemical markers, and clinical factors. The importance scores were normalized and expressed as percentages of total importance, allowing for direct comparison between features and providing a hierarchical understanding of the predictors' contributions to the model's performance (Fig. 5). For feature importance calculation: $I(f) = \Sigma_i(\Delta i(f)/M)$.

2.7. Implementation framework

Our model implementation was developed using Python's scientific computing ecosystem. We built a robust data processing and modeling pipeline integrating several specialized libraries: scikit-learn provided the core machine learning architecture, while imbalanced-learn enabled sophisticated class balancing through SMOTE. The ensemble model incorporated gradient boosting implementations from both LightGBM and XGBoost frameworks. Data manipulation and analysis were handled through pandas and numpy libraries, while visualization tasks were performed using matplotlib and seaborn. This comprehensive framework ensured efficient data processing, model training, and rigorous performance evaluation.

2.8. Data availability statement

All materials related to this study, including the model source code, the dataset used, and other relevant information, are available in the project page in GitHub. Repository can be accessed at https://github.com/frpcarvalho/Osteoporosis_Risk_Prediction.git.

3. Results

Our analysis followed a two-phase approach to optimize the OP prediction model. In the initial phase, using NHANES 2013–2014 data, we conducted an extensive evaluation of different feature selection methods and classifiers. However, due to modest performance metrics (overall accuracy of 0.76 and particularly low performance in identifying OP cases with F1-score of 0.47) (Supplementary Material Table S1), we recognized the need to enhance our approach. Therefore, we expanded our dataset to include NHANES cycles from 2007 to 2008 and 2009–2010, resulting in a more robust sample of 7924 participants aged 50 years and older. Additionally, we refined our OP classification criteria to incorporate both densitometric and anthropometric parameters. This comprehensive approach included not only the traditional WHO criterion (T-scores ≤ -2.5 at femoral neck, total femur, or spine) but also anthropometric risk factors such as extremely low body weight (lowest 15th percentile), severely reduced arm muscle circumference (lowest 15th percentile), and advanced age for female participants (highest 15th percentile) (Fig. 3). This enhanced methodology resulted in the identification of 1636 OP cases (20.6 %) and 6288 normal cases (79.4 %), providing a more balanced and clinically relevant stratification of the study population.

T-score analysis revealed differences between normal and osteoporotic groups across all anatomical sites. The femoral neck showed mean T-scores of -0.88 ± 0.90 for the normal group and -2.07 ± 0.84 for the osteoporotic group (mean difference: 1.19). Total femur measurements demonstrated similar differentiation (-0.41 ± 0.97 vs -1.79 ± 0.95, difference: 1.38), as did spine measurements (-0.18 ± 1.16 vs -1.54 ± 1.25, difference: 1.36). These differences provided a robust foundation for our classification approach (Table 1, Fig. 4).

Given our access to young adult BMD data within our expanded dataset, we made the methodological decision to create our own reference values rather than relying on traditional standards. This analysis revealed notable differences across demographic groups. For the femoral neck, mean BMD values ranged from 0.869 g/cm^2 (Hispanic

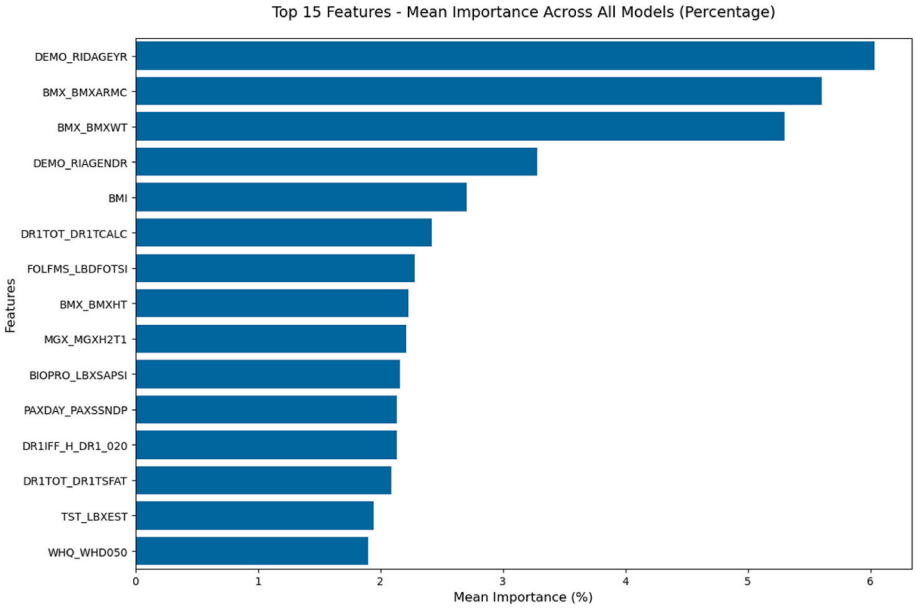

Fig. 5. Feature importance analysis. Age (DEMO_RIDAGEYR) was found the most influential predictor, accounting for 6.04 % of the model's decision-making process, followed by arm muscle circumference (BMX_BMXARMC, 5.61 %) and body weight (BMX_BMXWT, 5.30 %). Demographic factors, particularly gender (DEMO_RIAGENDR, 3.28 %), also showed significant influence. Body Mass Index (BMI, 2.71 %) and calcium intake (DR1TOT_DR1TCALC, 2.42 %) emerged as important metabolic and nutritional indicators. Additional influential features included folate (FOLFMS_LBDFOTSI, 2.28 %), height (BMX_BMXHT, 2.23 %), hand grip strength (MGX_MGXH2T1, 2.21 %), alkaline phosphatase (BIOPRO_LBXSAPSI, 2.16 %), physical activity metrics (PAXDAY_PAXSSNDP, 2.14 %), meal timing (DR1IFF_H_DR1_020, 2.14 %), total saturated fatty acids (DR1TOT_DR1TSFAT, 2.09 %), estradiol levels (TST_LBXEST, 1.95 %), and previous year weight (WHQ_WHD050, 1.90 %). This diverse set of predictors highlights the multifaceted nature of OP risk assessment, incorporating demographic, anthropometric, nutritional, and biochemical factors.

females) to 1.037 g/cm^2 (Black males). Total femur measurements showed similar patterns, with values ranging from 0.954 g/cm^2 (Hispanic females) to 1.148 g/cm^2 (Black males) (Supplementary Material Table S2). The creation of these study-specific reference values, derived from our own young adult population (ages 20–30), provided a more contextualized foundation for T-score calculations, better reflecting the demographic characteristics of our study population. In the second phase, we developed a stacking ensemble model incorporating multiple NHANES cycles (2007–2014) and achieved substantially improved performance. The ensemble model demonstrated excellent discrimination with an ROC-AUC score of 0.94 (Fig. 4E) and overall accuracy of 0.93. Specifically, the model achieved high precision scores of 0.93 for the normal class and 0.89 for the OP class, with corresponding recall values of 0.98 and 0.73, respectively (Fig. 4B). The F1-scores of 0.95 for normal cases and 0.80 for OP cases indicate strong balanced performance across classes (Table 2). Cross-validation analysis further confirmed the model's robustness, with ROC-AUC scores ranging from 0.902 to 0.946 across 5 folds, yielding a mean CV score of 0.929 (±0.030), demonstrating consistent and reliable performance across different subsets of the data (Fig. 4A and D). The confusion matrix further validates the model's performance, correctly identifying 1233 true negatives and 234 true positives, while minimizing misclassifications with only 25 false positives and 93 false negatives, demonstrating robust discrimination capabilities across both classes (Fig. 4C). Analysis of the probability distribution of predictions revealed clear separation between classes, with the model demonstrating high confidence in its classifications - predictions for normal cases clustered strongly towards probability values of 0, while OP cases showed distinct clustering towards probability values of 1, indicating robust discriminative capability (Fig. 4D).

The feature importance analysis revealed age as the most influential predictor (6.04 %), followed by anthropometric measurements including arm muscle circumference (5.61 %) and body weight (5.30 %). Demographic factors, particularly gender (3.28 %), also showed significant influence. Body Mass Index (2.71 %) and calcium intake (2.42 %) emerged as important metabolic and nutritional indicators. This hierarchical importance of features aligns with clinical understanding of OP risk factors.

This improved model architecture, combining multiple base classifiers through stacking, demonstrated superior performance compared to our previous approach. The results suggest that anthropometric measurements, demographic factors, and nutritional markers play crucial roles in OP risk assessment, providing valuable insights for clinical applications. The high ROC-AUC score and balanced performance metrics across classes indicate the model's potential utility as a reliable screening tool for OP risk.

4. Discussion

Our findings underscore the importance of a holistic approach to health risk assessment, considering both traditional risk factors and novel biomarkers. By incorporating feature importance analysis into predictive modeling, we enhanced our understanding of the complex interplay between various factors influencing OP risk.

Our analysis revealed that age (6.04 %), arm circumference (5.61 %), and body weight (5.30 %) were the most influential predictors. However, it's important to note that the prominence of these anthropometric variables in our model may have been partially amplified by our methodology of including extreme percentiles (lowest 15th percentile for weight and arm circumference, highest 15th percentile of age for females) in our OP classification criteria. Nevertheless, this methodological decision aligns with established clinical knowledge [31–33], as these anthropometric characteristics are well-documented risk factors for osteoporosis, and their inclusion in our classification criteria helps capture high-risk individuals who might be missed by densitometric criteria alone.

Following these anthropometric predictors, gender (3.28 %) and BMI (2.71 %) emerged as significant predictors, reflecting well-established demographic and anthropometric risk factors in OP epidemiology [34,35]. Height (2.23 %) also appeared as an influential factor, consistent with its known association with fracture risk and bone health status.

Notably, biochemical markers showed substantial predictive value in our model. Calcium intake (2.42 %) has a well-established role in bone metabolism and OPprevention [36], particularly in maintaining bone mineral density. The emergence of folate (2.28 %) as a significant predictor aligns with recent studies demonstrating its role in bone health through homocysteine metabolism and DNA methylation pathways [37]. Alkaline phosphatase levels (2.16 %) emerged as an important biochemical marker, which is consistent with its well-established role as a marker of bone turnover and formation [38]. Elevated levels of alkaline phosphatase are often associated with increased bone turnover and serve as a valuable indicator of bone metabolism activity [39].

Our model also identified several novel or less commonly considered predictors. Hand grip strength (2.21 %) emerged as a significant predictor, aligning with recent research demonstrating its value as a practical indicator of overall muscle strength and its association with bone health [40]. This finding supports the growing evidence that muscle strength assessment, particularly grip strength, could serve as a simple yet effective screening tool for OP risk [41]. Physical activity metrics (2.14 %) further reinforced the importance of musculoskeletal fitness in bone health maintenance [42].

Our analysis also revealed interesting associations with timing and behavioral patterns. Meal timing (2.14 %) emerged as a significant predictor, suggesting the potential importance of circadian rhythms and eating patterns in bone metabolism [43]. The influence of total saturated fatty acids intake (2.09 %) aligns with research on the complex relationship between dietary fat composition and bone [44].

The emergence of estradiol levels (1.95 %) as a predictor reinforces the well-established role of sex hormones in bone [45,46]. Additionally, the significance of year-over-year weight changes (1.90 %) highlights the importance of weight stability in bone health maintenance suggesting that not just current weight but weight history patterns may influence OP risk [47].

Several researchers have addressed the possibility of developing machine learning models to predict OP. Recently, researchers [48] tested different machine learning models, with XGBoost demonstrating superior predictive performance over traditional multivariate logistic regression in assessing the risk of BMD decrease in patients over 50 years old with type 2 diabetes mellitus. Also, an AI algorithm was developed for automated assessment of BMD using CT data, showing strong agreement between expert and AI measurements in patients over 50 years old [49]. In a community-based cohort study, researchers developed a fracture prediction model using CatBoost, which exhibited superior performance in predicting total fragility fractures compared to FRAX, SVM, and logistic regression models [50].

Recent strategies also include algorithm development using electronic health records to predict short-term fracture risk. The Crystal Bone algorithm achieved a ROC curve of 0.81 for patients over 50 years old, analyzing temporal patterns in patient histories [51]. More recently, Khanna and colleagues used the same NHANES dataset to predict OP through the integration of ML and explainable AI (XAI), achieving 89 % accuracy [52]. Our stacking ensemble model demonstrated superior performance with 93 % accuracy and an AUC of 0.94, with consistent performance confirmed through cross-validation (mean CV score: 0.929 ± 0.030). While both studies achieved robust results, our methodology differs in two crucial aspects: first, we explicitly excluded BMD and T-scores from feature selection and model training to develop a truly predictive model that does not rely on DXA measurements; second, we employed a standardized WHO-based definition of OP combined with anthropometric risk factors, ensuring better generalizability and reproducibility of our findings.

In our study, we have effectively engineered a robust model for

predicting OP through meticulous feature selection, a technique demonstrated to yield significant efficacy in predictive medicine [53]. By leveraging solely biomarkers, anthropometric measurements, and clinical data, our model obviates the need for supplementary diagnostic techniques like DXA, thereby suggesting its potential seamless integration into routine clinical medical check-ups or consultations without incurring additional costs. The clear separation between classes in our probability distributions and the strong performance metrics (93 % accuracy, 0.94 AUC) demonstrate the model's reliable discriminative capabilities. Our approach, along with similar strategies [54], has the potential to introduce innovative forecasting methods for OP, gradually superseding the established FRAX strategy, which alone may not consistently yield desired outcomes [55]. However, despite these promising results using the NHANES dataset, it requires validation in diverse clinical settings and populations. Future work should focus on developing and implementing an integrated clinical decision support system that can be seamlessly incorporated into hospital electronic health record systems. Such systematic implementation would not only validate our model's generalizability but also facilitate its adoption into routine clinical practice, potentially improving early OP detection and prevention strategies across diverse healthcare settings.

CRediT authorship contribution statement

Filipe Ricardo Carvalho: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paulo Jorge Gavaia:** Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Ethics statement

This study was conducted using publicly available data from the National Health and Nutrition Examination Survey (NHANES), which operates under protocols approved by the National Center for Health Statistics Research Ethics Review Board. All NHANES participants provided informed consent, and data were de-identified before public release. Our analysis adhered to ethical guidelines for secondary data use and was performed in compliance with the World Medical Association Declaration of Helsinki. No additional ethics approval was required as this research involved analysis of publicly available, anonymized data.

NHANES protocols ensure protection of participant privacy, appropriate informed consent processes, and ethical handling of all collected data. The study design and analytical methods were chosen to maximize scientific validity while respecting participant confidentiality and data use agreements.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study received Portuguese national funds from FCT - Foundation for Science and Technology through projects UIDB/04326/2020 (DOI:10.54499/UIDB/04326/2020), UIDP/04326/2020 (DOI:10.54499/UIDP/04326/2020) and LA/P/0101/2020 (DOI:10.54499/LA/P/0101/2020).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110289.

References

[1] A.M. Wu, C. Bisignano, S.L. James, G.G. Abady, A. Abedi, E. Abu-Gharbieh, et al., Global, regional, and national burden of bone fractures in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019, Lancet Healthy Longev 2 (9) (2021 Sep).
[2] N. Salari, H. Ghasemi, L. Mohammadi, M.H. Behzadi, E. Rabieenia, S. Shohaimi, et al., The global prevalence of osteoporosis in the world: a comprehensive systematic review and meta-analysis, J. Orthop. Surg. Res. 16 (1) (2021).
[3] T. Vilaca, R. Eastell, M. Schini, Osteoporosis in men, Lancet Diabetes Endocrinol. 10 (4) (2022).
[4] W. Lutz, W. Sanderson, S. Scherbov, The coming acceleration of global population ageing, Nature 451 (7179) (2008) 716–719.
[5] J.A. Cauley, Public health impact of osteoporosis, J Gerontol A Biol Sci Med Sci 68 (10) (2013) 1243–1251.
[6] V.H. Nguyen, Osteoporosis prevention and osteoporosis exercise in community-based public health programs, Osteoporos Sarcopenia 3 (1) (2017) 18–25.
[7] L.J. Melton, Hip fractures: a worldwide problem today and tomorrow, Bone 14 (Suppl 1) (1993).
[8] B.E. Rosengren, M.K. Karlsson, The annual number of hip fractures in Sweden will double from year 2002 to 2050, Acta Orthop. 85 (3) (2014) 234–237.
[9] S.E. Sattui, K.G. Saag, Fracture mortality: associations with epidemiology and osteoporosis treatment, Nat. Rev. Endocrinol. 10 (10) (2014) 592–602.
[10] S.K. Sandhu, G. Hampson, The pathogenesis, diagnosis, investigation and management of osteoporosis, J. Clin. Pathol. 64 (12) (2011) 1042–1050.
[11] S. Zhan, W. Xie, M. Yang, D. Zhang, B. Jiang, Incidence and risk factors of acute kidney injury after femoral neck fracture in elderly patients: a retrospective case-control study, BMC Musculoskelet. Disord. 23 (1) (2022) 177.
[12] Y. Shen, Z.M. Sardar, H. Chase, J.R. Coury, M. Cerpa, L.G. Lenke, Predicting bone health using machine learning in patients undergoing spinal reconstruction surgery, Spine 48 (2) (2023) 136–142.
[13] G. El-Hajj Fuleihan, M. Chakhtoura, J.A. Cauley, N. Chamoun, Worldwide fracture prediction, J. Clin. Densitom. 20 (3) (2017) 397–424.
[14] S. Chu, A. Jiang, L. Chen, X. Zhang, X. Shen, W. Zhou, et al., Machine learning algorithms for predicting the risk of fracture in patients with diabetes in China, Heliyon 9 (7) (2023 Jul) e18186.
[15] A.C. Looker, B. Dawson-Hughes, A.N.A. Tosteson, H. Johansson, J.A. Kanis, L. J. Melton, Hip fracture risk in older US adults by treatment eligibility status based on new National Osteoporosis Foundation guidance, Osteoporos. Int. 22 (2) (2011) 541–549.
[16] C.I. Hsieh, K. Zheng, C. Lin, L. Mei, L. Lu, W. Li, et al., Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning, Nat. Commun. 12 (1) (2021) 6772.
[17] J.A. Kanis, N.C. Harvey, H. Johansson, A. Oden, W.D. Leslie, E.V. McCloskey, FRAX and fracture prediction without bone mineral density, Climacteric 18 (Suppl 2) (2015) 2–9.
[18] J. Pepe, G. Della Grotta, R. Santori, V. De Martino, M. Occhiuto, M. Cilli, et al., Lumbar spine bone mineral density and trabecular bone score-adjusted FRAX, but not FRAX without bone mineral density, identify subclinical carotid atherosclerosis, J. Endocrinol. Investig. 44 (9) (2021) 1973–1980.
[19] K.A. Lee, H.J. Kim, H.S. Kim, Comparison of predictive value of FRAX, trabecular bone score, and bone mineral density for vertebral fractures in systemic sclerosis: a cross-sectional study, Medicine (Baltim.) 102 (2) (2023).
[20] E.M. Alfadhli, A.S. Alsharif, R.A. Alharbi, S.S. Alalawi, S.E. Darandari, S.A. Alsaedi, et al., Comparison of bone mineral density and Fracture Risk Assessment Tool in Saudi women with and without type 2 diabetes mellitus, Saudi Med. J. 43 (7) (2022) 705–712.
[21] C. Klop, F. De Vries, J.W.J. Bijlsma, H.G.M. Leufkens, P.M.J. Welsing, Predicting the 10-year risk of hip and major osteoporotic fracture in rheumatoid arthritis and in the general population: an independent validation and update of UK FRAX without bone mineral density, Ann. Rheum. Dis. 75 (12) (2016) 2095–2100.
[22] J. Damilakis, J.E. Adams, G. Guglielmi, T.M. Link, Radiation exposure in X-ray-based imaging techniques used in osteoporosis, Eur. Radiol. 20 (11) (2010) 2707–2714.
[23] G. Solomou, J. Damilakis, Radiation exposure in bone densitometry, Semin. Muscoskel. Radiol. 20 (4) (2016) 392–398.
[24] Y. Xiao, Q. Liang, L. Zhou, X. He, L. Lv, J. Chen, et al., Construction of a new automatic grading system for jaw bone mineral density level based on deep learning using cone beam computed tomography, Sci. Rep. 12 (1) (2022) 7042.
[25] Z.R. Artyukova, N.D. Kudryavtsev, A.V. Petraikin, L.R. Abuladze, A. K. Smorchkova, E.S. Akhmad, et al., Using an artificial intelligence algorithm to assess the bone mineral density of the vertebral bodies based on computed tomography data, Med Visualization 27 (2) (2023) 102–112.
[26] A.A. Shelepa, A.V. Petraikin, Z.R. Artyukova, L.R. Abuladze, N.D. Kudryavtsev, E. S. Ahmad, et al., Artificial intelligence for bone mineral density assessment: general population data, Digit Diagn 3 (1s) (2022) 26–36.
[27] R.H. Savage, M. Van Assen, S.S. Martin, P. Sahbaee, L.P. Griffith, D. Giovagnoli, et al., Utilizing artificial intelligence to determine bone mineral density via chest computed tomography, J. Thorac. Imag. 35 (Suppl 1) (2020).
[28] P. Pawar, M. Malkauthekar, Using machine learning to predict fracture risk in large U.S. population: an Analysis of NHANES 2005-2014, J. Allergy Clin. Immunol. 149 (2) (2022).
[29] M. Naghavi, K. Atlas, A. Jaberzadeh, C. Zhang, V. Manubolu, D. Li, et al., Validation of opportunistic artificial intelligence-based bone mineral density measurements in coronary artery calcium scans, J. Am. Coll. Radiol. 21 (4) (2024) 484–492.

[30] A. Coronato, M. Naeem, G. De Pietro, G. Paragliola, Reinforcement learning for intelligent healthcare applications: a survey, Artif. Intell. Med. 109 (2020) 101964.
[31] World Health Organization, WHO Scientific Group on the Assessment of Osteoporosis at Primary Health Care Level, WHO, Geneva, 2004.
[32] E.M. Lewiecki, N.B. Watts, M.R. McClung, S.M. Petak, L.K. Bachrach, J. A. Shepherd, et al., Official positions of the international society for clinical densitometry, J. Clin. Endocrinol. Metab. 89 (8) (2004) 3651–3655.
[33] N.M. Cummins, E.K. Poku, M.R. Towler, O.M. O'Driscoll, S.H. Ralston, Clinical risk factors for osteoporosis in Ireland and the UK: a comparison of FRAX and QFractureScores, Calcif. Tissue Int. 89 (2) (2011) 172–177.
[34] J. Compston, C. Bowring, A. Cooper, C. Cooper, C. Davies, R. Francis, et al., Diagnosis and management of osteoporosis in postmenopausal women and older men in the UK: national Osteoporosis Guideline Group (NOGG) update 2013, Maturitas 75 (4) (2013) 392–396.
[35] J.A. Kanis, E.V. McCloskey, H. Johansson, C. Cooper, R. Rizzoli, J.Y. Reginster, European guidance for the diagnosis and management of osteoporosis in postmenopausal women, Osteoporos. Int. 24 (1) (2013) 23–57.
[36] M.J. Bolland, W. Leung, V. Tai, S. Bastin, G.D. Gamble, A. Grey, et al., Calcium intake and risk of fracture: systematic review, Br. Med. J. 351 (2015).
[37] Z. Zheng, H. Luo, W. Xu, Q. Xue, Association between dietary folate intake and bone mineral density in a diverse population: a cross-sectional study, J. Orthop. Surg. Res. 18 (1) (2023) 189.
[38] J. Shu, A. Tan, Y. Li, H. Huang, J. Yang, The correlation between serum total alkaline phosphatase and bone mineral density in young adults, BMC Musculoskelet. Disord. 23 (1) (2022) 494.
[39] K.E. Naylor, E.V. McCloskey, R.M. Jacques, N.F.A. Peel, M.A. Paggiosi, F. Gossiel, et al., Clinical utility of bone turnover markers in monitoring the withdrawal of treatment with oral bisphosphonates in postmenopausal osteoporosis, Osteoporos. Int. 30 (4) (2019) 917–922.
[40] Y. Luo, K. Jiang, M. He, Association between grip strength and bone mineral density in general US population of NHANES 2013–2014, Arch. Osteoporosis 15 (1) (2020) 47.
[41] Y.H. Lin, H.C. Chen, N.W. Hsu, P. Chou, M.M.H. Teng, Hand grip strength in predicting the risk of osteoporosis in Asian adults, J. Bone Miner. Metabol. 39 (2) (2021) 269–277.
[42] J. Ji, Y. Hou, Z. Li, Y. Zhou, H. Xue, T. Wen, et al., Association between physical activity and bone mineral density in postmenopausal women: a cross-sectional study from the NHANES 2007–2018, J. Orthop. Surg. Res. 18 (1) (2023) 28.
[43] L.A. Riley, K.A. Esser, The role of the molecular clock in skeletal muscle and what it is teaching us about muscle-bone crosstalk, Curr. Osteoporos. Rep. 15 (3) (2017) 222–230.
[44] Z.B. Fang, G.X. Wang, G.Z. Cai, P.X. Zhang, D.L. Liu, S.F. Chu, et al., Association between fatty acids intake and bone mineral density in adults aged 20–59: NHANES 2011–2018, Front. Nutr. 10 (2023) 1129897.
[45] D. Schmitz, W.E. Ek, E. Berggren, J. Höglund, T. Karlsson, Å. Johansson, Genome-wide association study of estradiol levels and the causal effect of estradiol on bone mineral density, J. Clin. Endocrinol. Metab. 106 (11) (2021).
[46] G. Huitrón-Bravo, E. Denova-Gutiérrez, J.O. Talavera, C. Moran-Villota, J. Tamayo, A. Omaña-Covarrubias, et al., Levels of serum estradiol and lifestyle factors related with bone mineral density in premenopausal Mexican women: a cross-sectional analysis, BMC Musculoskelet. Disord. 17 (1) (2016) 15.
[47] C.J. Crandall, V.O. Yildiz, J. Wactawski-Wende, K.C. Johnson, Z. Chen, S.B. Going, et al., Postmenopausal weight change and incidence of fracture: post hoc findings from women's health initiative observational study and clinical trials, Br. Med. J. 350 (2015).
[48] J. Zhang, Z. Xu, Y. Fu, L. Chen, Prediction of the risk of bone mineral density decrease in type 2 diabetes mellitus patients based on traditional multivariate logistic regression and machine learning: a preliminary study, Diabetes Metab Syndr Obes 16 (2023) 2523–2534.
[49] T. Ling, L. Jake, J. Adams, K. Osinski, X. Liu, D. Friedland, Interpretable machine learning text classification for clinical computed tomography reports – a case study of temporal bone fracture, Comput Methods Programs Biomed Update 3 (2023) 100078.
[50] S.H. Kong, D. Ahn, B. Kim, K. Srinivasan, S. Ram, H. Kim, et al., A novel fracture prediction model using machine learning in a community-based cohort, JBMR Plus 4 (3) (2020).
[51] Y.A. Almog, A. Rai, P. Zhang, A. Moulaison, R. Powell, A. Mishra, et al., Deep learning with electronic health records for short-term fracture risk identification: crystal bone algorithm development and validation, J. Med. Internet Res. 22 (10) (2020).
[52] V.V. Khanna, K. Chadaga, N. Sampathila, R. Chadaga, S. Prabhu, S.S. K, et al., A decision support system for osteoporosis risk prediction using machine learning and explainable artificial intelligence, Heliyon 9 (12) (2023) e22456.
[53] N. Gutowski, D. Schang, O. Camp, P. Abraham, A novel multi-objective medical feature selection compass method for binary classification, Artif. Intell. Med. 127 (2022) 102290.
[54] G. Mazziotti, W. Vena, R. Pedersini, S. Piccini, E. Morenghi, D. Cosentini, et al., Prediction of vertebral fractures in cancer patients undergoing hormone deprivation therapies: reliability of who fracture risk assessment tool (frax) and bone mineral density in real-life clinical practice, J. Bone Oncol. 33 (2022) 100418.
[55] D. Santra, S. Goswami, J.K. Mandal, S.K. Basu, Low back pain expert systems: clinical resolution through probabilistic considerations and poset, Artif. Intell. Med. 120 (2021) 102152.