EnM
ENDOCRINOLOGY AND METABOLISM

Original Article

Endocrinol Metab 2025;40:484-497
https://doi.org/10.3803/EnM.2024.2211
pISSN 2093-596X · eISSN 2093-5978

Synthetic Data-Enhanced Classification of Prevalent Osteoporotic Fractures Using Dual-Energy X-Ray Absorptiometry-Based Geometric and Material Parameters

Luca Quagliato[1,*], Jiin Seo[1,*], Jiheun Hong[1], Taeyong Lee[1,2], Yoon-Sok Chung[3,4]

[1]Division of Mechanical and Biomedical Engineering, Ewha Womans University; [2]Division of Mechanical and Biomedical Engineering, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul; [3]Department of Endocrinology and Metabolism, Ajou University School of Medicine, Suwon; [4]Ajou Institute on Aging, Ajou University Medical Center, Suwon, Korea

Background: Bone fracture risk assessment for osteoporotic patients is essential for implementing early countermeasures and preventing discomfort and hospitalization. Current methodologies, such as Fracture Risk Assessment Tool (FRAX), provide a risk assessment over a 5- to 10-year period rather than evaluating the bone's current health status.

Methods: The database was collected by Ajou University Medical Center from 2017 to 2021. It included 9,260 patients, aged 55 to 99, comprising 242 femur fracture (FX) cases and 9,018 non-fracture (NFX) cases. To model the association of the bone's current health status with prevalent FXs, three prediction algorithms—extreme gradient boosting (XGB), support vector machine, and multilayer perceptron—were trained using two-dimensional dual-energy X-ray absorptiometry (2D-DXA) analysis results and subsequently benchmarked. The XGB classifier, which proved most effective, was then further refined using synthetic data generated by the adaptive synthetic oversampler to balance the FX and NFX classes and enhance boundary sharpness for better classification accuracy.

Results: The XGB model trained on raw data demonstrated good prediction capabilities, with an area under the curve (AUC) of 0.78 and an F1 score of 0.71 on test cases. The inclusion of synthetic data improved classification accuracy in terms of both specificity and sensitivity, resulting in an AUC of 0.99 and an F1 score of 0.98.

Conclusion: The proposed methodology demonstrates that current bone health can be assessed through post-processed results from 2D-DXA analysis. Moreover, it was also shown that synthetic data can help stabilize uneven databases by balancing majority and minority classes, thereby significantly improving classification performance.

Keywords: Femoral fractures; Osteoporosis; Machine learning

Received: 22 October 2024, **Revised:** 25 December 2024,
Accepted: 8 January 2025

Corresponding authors: Taeyong Lee
Division of Mechanical and Biomedical Engineering, Graduate Program in System Health Science and Engineering, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea
Tel: +82-2-3277-2795, **Fax:** +82-2-3277-2846, **E-mail:** tlee@ewha.ac.kr

Yoon-Sok Chung
Department of Endocrinology and Metabolism, Ajou University School of Medicine, 164 World cup-ro, Yeongtong-gu, Suwon 16499, Korea
Tel: +82-31-219-5127, **Fax:** +82-31-219-4497, **E-mail:** yschung@ajou.ac.kr

*These authors contributed equally to this work.

INTRODUCTION

Assessing bone health is crucial for preventing costly medical treatments, patient discomfort from immobilization, muscle loss, and extended rehabilitation [1,2]. With increasing life expectancy [3,4] and widespread sedentary lifestyles [5,6], skeletal health, particularly in the lower limbs, has emerged as a significant concern [7,8]. Effective prevention [9-12] and treatment strategies [13-16] are necessary for osteopenia and osteoporosis, underscoring the importance of prompt interventions. Tools such as Fracture Risk Assessment Tool (FRAX) (World Health Organization, Geneva, Switzerland) have been developed for predicting long-term fracture risks [17-19] but fall short in assessing real-time bone health and immediate fracture risk.

Conventional methods for assessing bone health focus on bone mineral density (BMD) and patient data, often overlooking the bone geometry information that can be obtained through two-dimensional dual-energy X-ray absorptiometry (2D-DXA). While computed tomography (CT) enables a detailed analysis of bone geometry, it is both expensive and time-consuming. It is crucial to implement timely interventions to maintain BMD and prevent functional loss, as delays can lead to increased risk of complications and costly post-fracture care [20,21]. Techniques such as finite element analysis [22] and statistical models [23,24] have improved predictions of fracture risk by incorporating 2D-DXA data.

Machine learning (ML) has revolutionized fracture risk assessment with its superior predictive capabilities. Early studies that employed support vector machines (SVMs) with radial basis function (RBF) kernels achieved accuracy levels of up to 90% in analyzing hip 2D-DXA scans [25,26]. However, this accuracy decreased to 76.7% when applied to fully Korean cohorts [27]. Other ML models, such as linear regression, random forest, gradient boosting, extreme gradient boosting (XGB) [27-29], and artificial/deep neural networks [27,28,30], have also been utilized to predict BMD. Nonetheless, BMD alone is insufficient for predicting fracture risk, necessitating the incorporation of additional medical data.

Recent studies have demonstrated the potential of using expanded datasets. For instance, modified gradient boosting algorithms such as CatBoost have outperformed FRAX in predicting bone fragility within a 7.5-year cohort [31]. Additionally, artificial neural network models have successfully classified osteoporotic risk levels into normal, medium, or high categories [32], with simpler architectures featuring five hidden layers showing superior performance [33]. Despite these advancements, many studies still rely on CT or 2D-DXA image-based methods [34], including the use of convolutional neural networks [35-37]. These methods often restrict the size of training datasets [38-40]. The high computational demands underscore the advantages of array-based ML models, which allow efficient training on large datasets and enable accurate modeling of complex scenarios [41].

This study addresses these limitations by developing an ML algorithm that performs binary classification of osteoporotic fracture risk based on 2D-DXA-derived geometry and material parameters. The dataset, which includes 9,260 patients from the Ajou Cohort (2017–2021, AC_17-21), was initially evaluated using three classifiers: XGB [42], SVM [27], and multilayer perceptron (MLP) [43]. The XGB algorithm demonstrated superior and more balanced performance. Subsequently, the XGB model was employed in conjunction with a balanced dataset achieved through ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning) oversampling [44,45]. This method facilitates a rapid and accurate assessment of whether the current status of the bone indicates a predisposition for osteoporotic fracture, providing physicians with real-time tools to assess bone health and implement timely preventive measures.

METHODS

Database composition

The database employed in this research, defined as Ajou Cohort 2017–2021 (AC_17-21), is based on a retrospective dataset of 2D-DXA scans of 9,260 patients, approximately 90% female and 10% male, with a mean age of 64.12±8.96 years (range, 50 to 99). To minimize potential confounding biases related to medication, only subjects who had discontinued bone health-related medications for at least 3 months were included. Additionally, environmental factors such as smoking habits, exercise patterns, and lifestyle were considered in the analysis. Among the 9,260 scans, 242 were from patients who sustained a fracture in the femur, specifically between the middle of the shaft and the head. The remaining scans were from patients who did not experience any fractures and underwent 2D-DXA as a preventive measure. For those with fractures, the 2D-DXA scans were performed post-fracture. Importantly, no patients had data that belonged to both the fracture (FX) and non-fracture (NFX) categories, referred to here as those with prevalent fractures. The FX category also included patients who either had a history of osteoporotic fractures or were diagnosed with osteoporosis at the time of the DXA examination. Conversely, patients with no

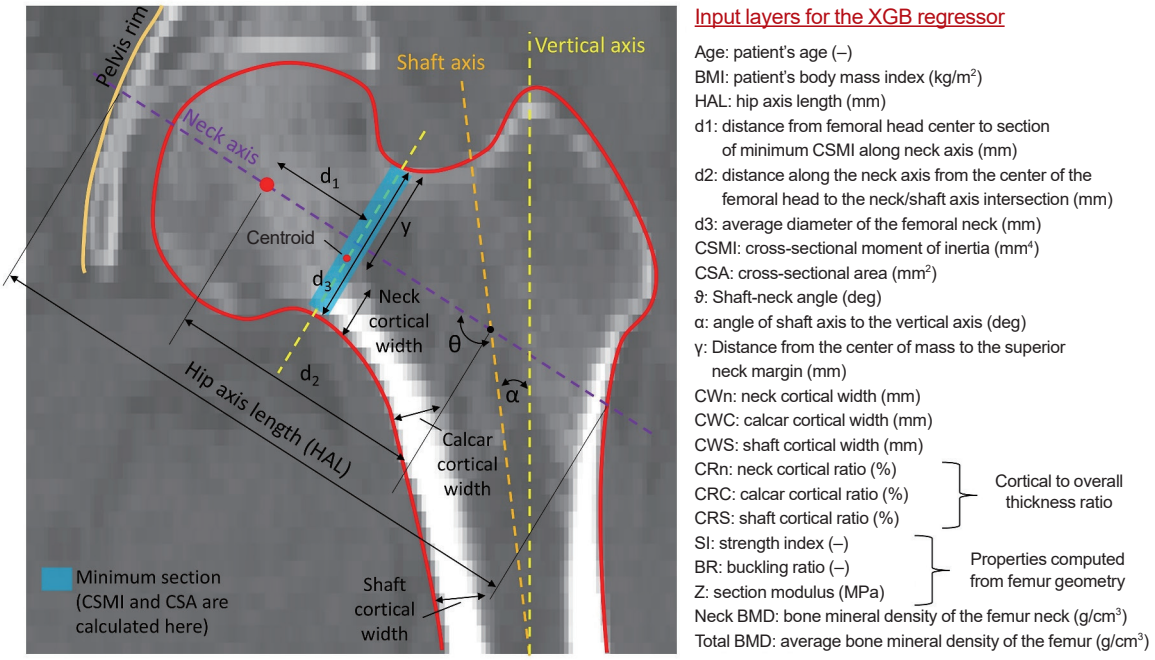

Fig. 1. Definition for the parameters in the Ajou Cohort 2017 to 2021 (*AC_17-21*) database is shown. XGB, extreme gradient boosting.

fractures or history of osteoporotic fractures at the time of examination were classified into the NFX category. Due to the privacy concerns, the dataset is not publicly available.

The data used to create the training dataset for the ML models were collected during a predefined observation period, with no further follow-up information available beyond this timeframe. As the data are limited to this specific period, any health events, such as fractures or changes in bone health that occurred after the data collection ended, are considered unobserved. Therefore, any outcomes that happened after this observation period are unknown, constituting censored data. Additionally, since the dataset was compiled by the hospital, details about patient dropout or missed follow-ups during the data collection phase were not accessible.

All patient examinations were conducted using the Prodigy BMD machine (GE Healthcare, Madison, WI, USA) at Ajou University Hospital in Suwon, South Korea. The data were then post-processed using the advanced hip assessment (AHA) software provided by the GE Lunar Company's Seoul, Korea branch. The resulting database comprises 22 scalar values, as shown in Fig. 1, and a binary output (FX or NFX). Although the sex of the patients was recorded, it was not included as a feature in the analysis. Due to its binary nature, representing sex would require conversion to a 0/1 or –1/1 format, as it is categorical data. To prevent overfitting issues that could arise from including a binary feature, the analysis was limited to bone-related features and parameters.

As shown in Table 1, the mean values and standard deviations for many of the 22 features are remarkably similar, suggesting the need to employ synthetic data to balance the classes and thus sharpen their reciprocal boundaries.

The data collection for this study received approval from the Institutional Review Board of Ajou University Hospital (Approval number: AJOUIRB-MDB-2021-299). No direct experiments were conducted on patients; instead, the raw data used in this research were derived from medical examinations performed in accordance with applicable guidelines and regulations. Informed consent was obtained from all participants or their legal guardians. Details of the modeling process described in subsequent sections of this paper are included in the supplementary material provided with this study.

Prevalent fracture risk modeling

The AC_17-21 database was used to train three supervised ML classifiers: XGB, SVM, and MLP. The training dataset, derived from the results of post-processed 2D-DXA scans and referred to as RAW DB, included 242 FX cases and an equal number of randomly selected NFX cases. No synthetic data was introduced into RAW DB, and the classes were balanced by limiting the size of the NFX class to match that of the FX class. Random

Table 1. Mean Values and Standard Deviations for the 21 Predictor Variables in the AC_17-21 Database

Parameter	NFX	FX
Age, yr	63.83±8.78	69.67±9.72
BMI, kg/m^2	24.12±3.49	24.08±3.55
HAL, mm	101.85±6.61	102.19±7.75
Strength index	1.55±0.35	1.58±0.38
Buckling ratio	3.77±1.41	3.89±1.53
Section modulus, MPa	453.44±115.82	462.17±131.67
CSMI (×10^3), mm^4	7.44±2.37	7.71±2.68
CSA, mm^2	114.42±20.93	115.44±24.18
d1, mm	15.12±2.88	14.74±3.10
d2, mm	47.66±4.45	47.30±4.80
d3, mm	31.19±2.36	31.50±2.62
y, mm	16.27±1.55	16.50±1.72
Alpha, °	–0.13±1.98	–0.12±2.17
Theta, °	124.51±3.40	124.65±5.06
Cortical width at the neck, mm	4.85±1.66	4.82±1.73
Cortical ratio at the neck, mm	16.86±5.67	16.60±5.89
Cortical width at the calcar, mm	3.53±1.21	3.60±1.36
Cortical ratio at the calcar, %	6.99±2.26	7.03±2.40
Cortical width at the shaft, %	4.80±1.02	4.80±1.20
Cortical ratio at the shaft, %	16.62±3.29	16.45±3.53
Neck BMD, g/cm^3	0.79±0.12	0.77±0.14
Total BMD, g/cm^3	0.84±0.14	0.84±0.15

Values are expressed as mean±standard deviation.
NFX, non-fracture; FX, fracture; BMI, body mass index; HAL, hip axis length; CSMI, cross-sectional moment of inertia; CSA, cross-sectional area; d1, distance from the femoral head center to section of minimum CSMI along neck axis; d2, distance along the neck axis from the center of the femoral head to the neck/shaft axis intersection; d3, average diameter of the femoral neck; y, distance from the center of mass to the superior neck margin; alpha, angle of shaft axis to the vertical axis; theta, shaft-neck angle; BMD, bone mineral density.

sampling was conducted without explicit stratification for demographic variables. However, during the resampling process, the distribution of key demographic characteristics such as age and body mass index (BMI) between the FX and NFX classes was checked to ensure representativeness. All three models—XGB, SVM, and MLP—were trained using the same database, which consisted of 22 scalar features and one binary output variable, either FX or NFX. Of the 484 cases, 30 were set aside as a test set, while the remaining 454 were divided into 90% for training and 10% for validation using a k=5 cross-validation method. The 30 test cases were not used at any stage of the training process and served as a real benchmark scenario.

To identify the best-performing hyperparameters for each model, the GridSearch method was employed. This approach involved systematically evaluating various combinations of hyperparameters to select the configuration that maximized performance metrics on the validation set.

Afterward, the ADASYN technique was applied to the entire database of 9,260 2D-DXA cases and used to train the XGB classifier, which demonstrated the best balance of area under the curve (AUC) and F1 scores across validation and test sets. Using ADASYN, synthetic data was generated to increase the size of the minority class (FX) until it equaled that of the majority class (NFX), thus balancing the dataset between the classes. To investigate further the impact of specific features from the 2D-DXA scans, the ADASYN-balanced database was divided into two sub-models: geometry-based features (sub-model #1) and material-based features (sub-model #2). Each sub-model associated the output class (FX or NFX) with a distinct set of features, aiming to determine which feature set most significantly influenced the prediction of osteoporotic fracture risk.

Supervised learning classifiers

In this study, we employed two well-established classifier architectures and one relatively new model to predict fracture risk associated with osteopenia and osteoporosis, using data from 2D-DXA scans. We provide a brief theoretical background for the three models used: XGB [3,4], SVM, and MLP. Equation (1) shows the XGB model objective function where each new tree $f_m(x)$ aims at reducing the difference between predicted ($\hat{y}_i$) and true (y_i) values and $\Omega(f)$ is the regularization term that penalizes complexity.

$$L(\theta)=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{m=1}^{M}\Omega(f_m) \quad \text{where} \quad \Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \qquad (1)$$

Considering a generic t-iteration, a new tree is added to the existing structure, resulting in Equation (2) where the indicator function is 1 if the input feature x belongs to the j-leaf in the t-tree and 0 if otherwise.

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t(x) \quad \text{where} \quad f_t(x)=\sum_{j=1}^{T_t} w_j^2 I\left(x\in R_{tj}\right) \qquad (2)$$

For hyperparameter optimization, colsample by tree (*col*), (0.7, 0.8, 0.9, 1.0), gamma (γ), (0.0, 0.1, 0.2), learning rate (η), (0.1, 0.2, 0.3), and max depth (*mdep*), (10, 11, 12, 13), were considered. The col parameter controls the fraction of features to consider when constructing each tree, γ is the minimum loss reduction required to make a further partition on a leaf node of the tree, η refers to the contribution of each tree to the final pre-

diction and influences the step size during the gradient descent optimization process, and *mdep* specifies the depth of each tree in the ensemble. The tree-based architecture inherently captures complex feature interactions and non-linear relationships, which are crucial for distinguishing subtle differences in fracture-related features. Additionally, it can handle imbalanced data through an integrated weighting mechanism that prioritizes the learning of minority classes.

For the SVM model, the decision function is defined in Equation (3) where $w \cdot x+b=0$ is the equation of the hyperplane that delineates the classes' separation, defined by the weight vector (w) and the bias (b). In this scenario, the objective function is defined in Equation (4) and aims at norm w, while the regularization parameter C controls the trade-off between maximizing the margin and minimizing the classification error.

$$f(x) = \text{sign}(W \cdot X + b) \tag{3}$$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{where} \quad y_i(W \cdot X_i + b) \geq 1-\xi_i,\ \xi_i \geq 0 \tag{4}$$

To map the input features from the database into the higher-dimensional space where the hyperplane is defined, the kernel function shown in Equation (5) was employed and both linear and RBF kernels were considered during the tuning process. For the SVM classifier, C, kernel type (linear and RBF), and the γ parameter of the RBF kernel were tuned.

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \begin{cases} \text{Linear}: K(x_i, x_j) = x_i \cdot x_j \\ \text{RBF}: K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \end{cases} \tag{5}$$

The third considered model, MLP, is defined as a feedforward artificial neural network, defined by an input layer with a size equal to the number of features, an L number of layers, and an output layer with a tunable neuron size, as shown in Equation (6), where $h^{(l)}$ is the activation of the l-layer, $W^{(l)}$ the weight matrix, $b^{(l)}$ the bias vector, and σ the activation function.

$$h^{(l)} = \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)}\right) \tag{6}$$

The activation function was considered as a hyperparameter, and the three formulations included in Equation (7)—namely rectified linear unit (*ReLU*), *tanh*, and *logistic*—were considered during the tuning process. Due to the nature of the binary classification problem at hand, the binary cross-entropy function of Equation (8) was considered.

$$ReLU: \sigma(x) = max(0,x)\ ;\ tanh: \sigma(x) = tanh(x)\ ;\ Logistic: \sigma(x) = \frac{1}{1+exp(-x)} \tag{7}$$

$$L(y_i, \hat{y}_i) = -\frac{1}{n}\sum_{m=1}^{M}\left[y_i\, log(\hat{y}_i) + (1-y_i)\, log(1-\hat{y}_i)\right] \tag{8}$$

In addition, adaptive moment estimation (*adam*) and stochastic gradient descent (*sgd*) as solvers, the learning rate (*constant* and *adaptive*), and the maximum number of iterations (*max_iter*) (500, 1,000, and 5,000) were considered. For the constant learning rate, we tested levels of 0.001, 0.01, and 0.1. For the adaptive learning rate, initial values were set at 0.01 and 0.1. Additionally, the maximum number of iterations (max_iter) was set at 500, 1,000, and 5,000. To manage computational demands effectively, only two network configurations were included in the tuning process: (22, 50, 100, 150, 200, 150, 100, 50, 1) and (22, 100, 100, 100, 100, 100, 100, 100, 100, 1).

ADASYN oversampler

To improve the classification capabilities of the XGB algorithm, the ADASYN technique was employed. ADASYN was chosen for this study because it dynamically adjusts the creation of synthetic samples according to the classification challenges faced

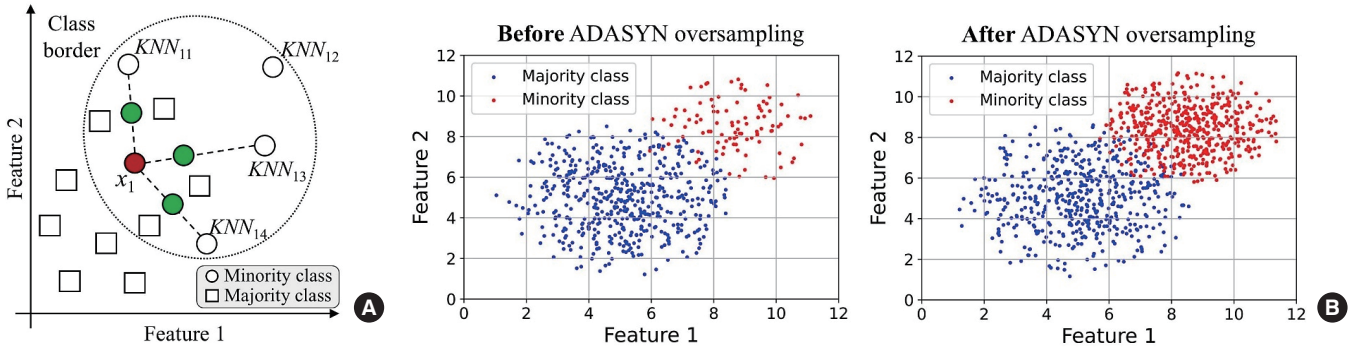

Fig. 2. Synthetic data oversampling principle of (A) the adaptive synthetic (ADASYN) sampling technique and (B) example of the synthetic data generation and resulting improved balance in binary classification cases. The dataset was enhanced using sampling techniques and validated for improvement of balance between classes. KNN, k-nearest neighbors algorithm.

by minority class instances, with a particular focus on areas near class boundaries. In contrast, methods like synthetic minority oversampling technique (SMOTE) implement uniform oversampling for all minority samples, potentially resulting in the production of redundant or less informative data [46].

The ADASYN algorithm [47] considers the distributions of both majority and minority classes and focuses on adding synthetic data near the class boundaries. Initially, the density distribution used to generate synthetic data is calculated, taking into account the class imbalance d_i (Fig. 2A), as shown in Equation (9).

$$r_i = d_i \left(\sum_{i=1}^{N_m} d_i \right)^{-1} \quad (9)$$

Afterward, the synthetic points are generated considering one minority instance at the time and one of the *k*-nearest neighbors of x_1, represented by the green circles in Fig. 2A, as in Equation (10). The number of synthetic samples is chosen by the user. In this research, it was set to match the size of the majority (NFX) class.

$$x_{syn} = x_i + \delta(x_{ki} - x_i) \quad with \quad \delta \in [0,1] \quad (10)$$

According to this rationale, the synthetic data points generated by the ADASYN algorithm do not introduce any new information beyond what is already present in the raw data features. However, they do facilitate clearer classification at the boundaries between classes, as illustrated in Fig. 2B. Specifically, the increased number of data points at these boundaries enables the classifier algorithm to more effectively learn and determine which combinations of features correspond to each of the two or more output classes. This enhancement in boundary definition leads to improved classification accuracy, measurable in terms of both specificity and sensitivity. In this research, most of the predictors used follow a Gaussian distribution. Consequently, as demonstrated in the 'Results' section, the ADASYN oversampling technique yielded excellent performance on the AC_17-21 database.

RESULTS

Fracture risk modeling performances on RAW DB

First, as shown in Fig. 3A, the pre-conditioned AC_17-21 database exhibits a strong correlation among its features, including both proportional and inversely proportional relationships. However, the correlation between the output class (FX/NFX) and the input features is relatively lower, highlighting the complexity of the modeling when only the 'RAW' database is used, without the addition of synthetic data. Regarding the training process, the XGB, SVM, and MLP models were trained, tuned, and subsequently benchmarked. A dataset of 484 cases, evenly split between 50% FX and 50% NFX cases, was utilized for all three models, with no synthetic data included.

For all three models, tuning was conducted using the GridSearch algorithm, exploring a total of 720, 1,160, and 1,080 parameter combinations for each of the k=5 folds in the cross-validation process. The optimal hyperparameters identified were as follows: for XGB, col=0.8, γ=0.1, η=0.1, and mdep=12; for SVM, C=0.1, γ=1, and kernel=linear; for MLP, the activation function was ReLU, α=0.1, the learning rate type is adaptive, η_0=0.1, max_iter=500, and the solver was sgd. The architecture of the layers was configured as 22, 50, 100, 150, 200, 150, 100, 50, 1.

The shapley additive explanations (SHAP) feature analysis, as shown in Fig. 3B, represents the average derived from k=5 folds, normalized from 0 to 1. It identified age as the most influential parameter, a finding that is also highlighted in the correlation heatmap of Fig. 3A. After age, BMI was identified as another critical factor, commonly used alongside age in existing algorithms for assessing fracture risk, such as FRAX® [17-19]. Regarding the receiver operating characteristic (ROC) curve and AUC, both XGB (Fig. 3C) and SVM (Fig. 3D) demonstrated promising performance across both validation and test sets. The XGB model exhibited a more balanced performance, with AUCs of 0.78 and 0.77, respectively. In contrast, the SVM model showed slightly higher accuracy on the test set than on the validation set, with AUCs of 0.74 and 0.8, respectively.

This is particularly interesting when examining the confusion matrices, which indicate flawless training for the XGB model, while the SVM model struggles with both false positive (FP) and false negative (FN) cases, despite both models undergoing similar tuning processes. Regarding the MLP model, as shown in Fig. 3E, although it has a higher structural complexity, its performance in terms of ROC and AUC is comparable to those of XGB and SVM. However, the confusion matrices reveal a tendency towards FN overfitting in the MLP model, with an FN/FP ratio close to 2, which is suboptimal for fracture risk modeling. Additionally, it is important to note that the complete training, validation, tuning, and testing phases for the XGB, SVM, and MLP models took approximately 45, 345, and 415 minutes, respectively. Furthermore, as shown in Fig. 3C-E, the F1 scores align with the trends of the AUC values, supporting the interpretation that XGB outperformed SVM and MLP in this dataset.

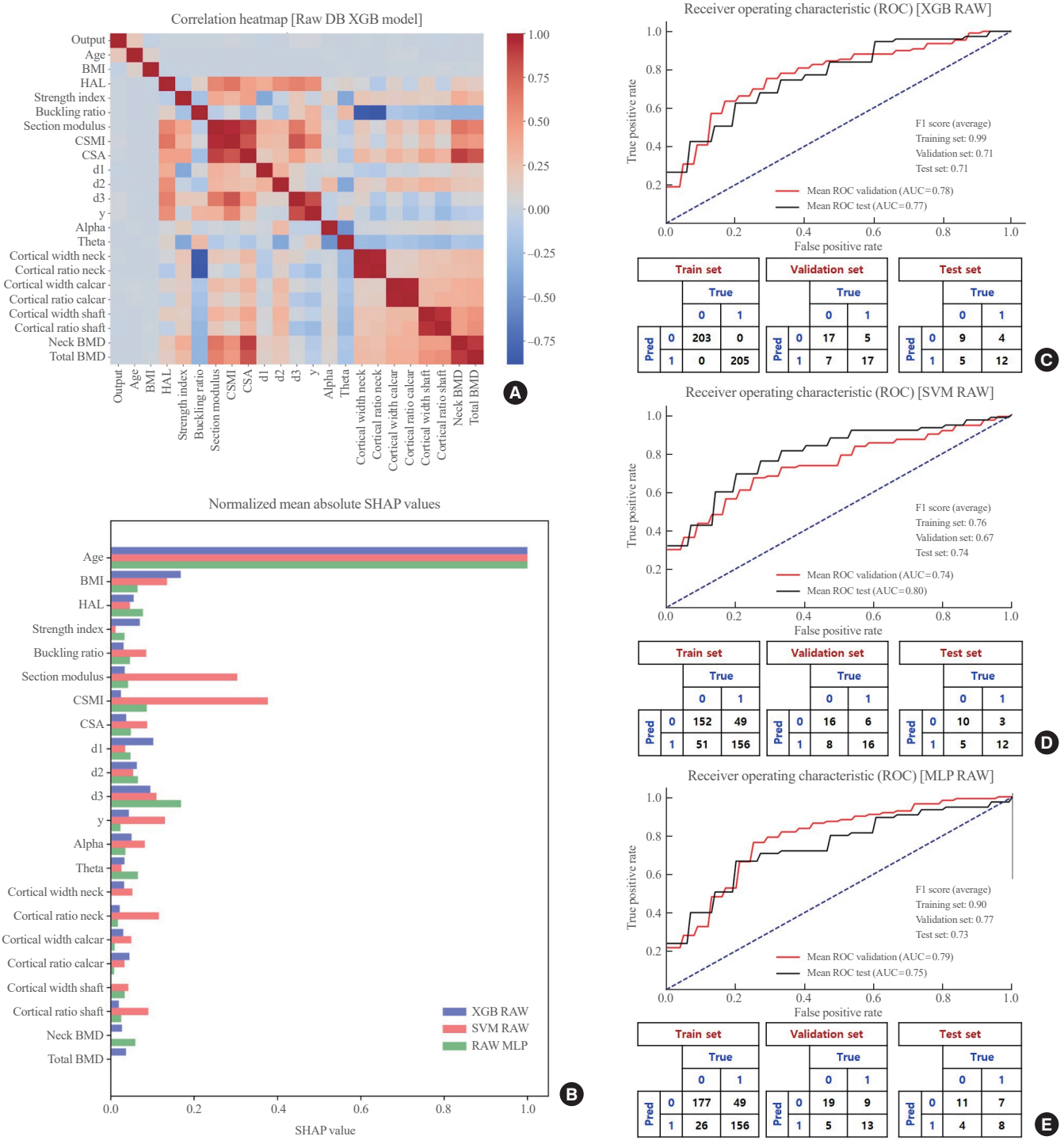

Fig. 3. Database without synthetic data (RAW DB) extreme gradient boosting (XGB) model: (A) Correlation heatmap between the output class and features and (B) normalized mean absolute shapley additive explanation (SHAP) values for the XGB, support vector machine (SVM), and multilayer perceptron (MLP) models. Receiver operating characteristic (ROC) curves and confusion matrices were drawn for training, validation, and test sets over 30 repetitions for the (C) XGB, (D) SVM, and (E) MLP models for RAW DB (242 fracture and 242 non-fracture cases). BMI, body mass index; HAL, hip axis length; CSMI, cross-sectional moment of inertia; CSA, cross-sectional area; d1, distance from the femoral head center to section of minimum CSMI along neck axis; d2, distance along the neck axis from the center of the femoral head to the neck/shaft axis intersection; d3, average diameter of the femoral neck; y, distance from the center of mass to the superior neck margin; alpha, angle of shaft axis to the vertical axis; theta, shaft-neck angle; BMD, bone mineral density; AUC, area under the curve.

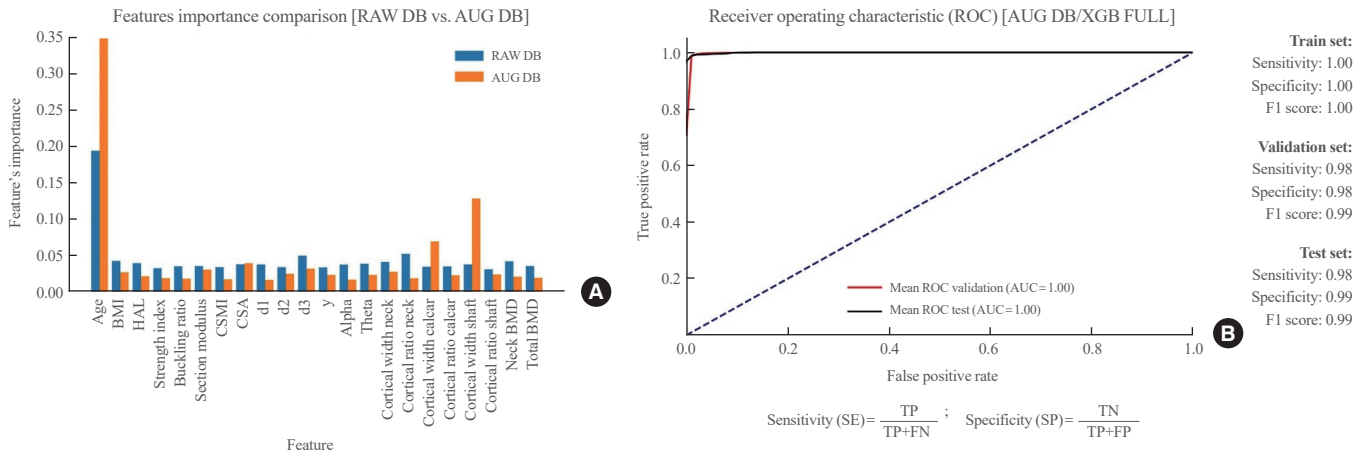

Fig. 4. (A) Feature importance comparison between database without synthetic data (RAW DB) and augmented data (AUG DB) when employed to train the extreme gradient boosting (XGB) model and (B) receiver operating characteristic (ROC) curves for validation and test datasets for the AUG-XGB model together with sensitivity, specificity, and accuracy on training, validation, and test datasets. A ROC curve of the augmented database was drawn based on the XGB model (validation set [red] and test set [blue]). BMI, body mass index; HAL, hip axis length; CSMI, cross-sectional moment of inertia; CSA, cross-sectional area; d1, distance from the femoral head center to section of minimum CSMI along neck axis; d2, distance along the neck axis from the center of the femoral head to the neck/shaft axis intersection; d3, average diameter of the femoral neck; y, distance from the center of mass to the superior neck margin; alpha, angle of shaft axis to the vertical axis; theta, shaft-neck angle; BMD, bone mineral density; AUC, area under the curve; TP, true positive; TN, true negative; FP, false positive; FN, false negative.

When the results of Fig. 3C-E are analyzed collectively, the XGB algorithm demonstrates the most balanced and promising performance compared to the other two options investigated. However, in the validation and test datasets, the AUC values are estimated at 0.78 and 0.77, respectively. These values indicate a loss in both specificity and sensitivity compared to the training set. This suggests that balancing for the minority class may lead to a loss of information, or in other words, that relying solely on the 484 RAW cases is insufficient for effective training. To address this issue, the entire AC_17-21 database was utilized, and balance was achieved by augmenting the majority class with synthetic data obtained using the ADASYN method, as detailed in the subsequent section.

ADASYN-augmented full-database XGB model

Considering the relationship between the output class (FX or NFX) and the features, the ADASYN method was employed to generate synthetic FX data to add to the RAW AC_17-21 database to match the number of NFX cases. These data were subsequently used to train the same XGB model. This was done to sharpen the boundary between FX and NFX classes while maintaining the majority-minority classes' reciprocal distribution throughout the latent space and across features. As shown in Fig. 4A, the addition of synthetic data for the minority class (FX) revealed the importance of information that was previously hidden within other features in the dataset, such as 'cortical width of calcar' and 'cortical width of shaft.'

It should be noted that the importance plot of features in Fig. 4A is derived from the tree structure of the XGB algorithm, rather than the SHAP function shown in Fig. 3B, and that 'AUG' refers to the XGB model trained with the ADASYN-augmented data, contraposed to the 'RAW' results in Fig. 3C. Although the XGB model trained only with raw data was already effective, the addition of synthetic data to balance the majority class led to a substantial improvement in its prediction capability in terms of AUC, sensitivity, and specificity, as shown in Fig. 4B. The very high scores across the training, validation, and test sets show that the ADASYN-XGB model was properly trained and cross-validated. Moreover, the results demonstrate robust performance on new data, although the new data inherently shared the same feature distribution as the training and validation sets.

ADASYN-augmented database XGB sub-models

To further investigate the influence of geometry and material predictors on training the XGB classifier, two sub-models were developed. Both models utilized the same XGB architecture but were based on two distinct databases, each augmented using the

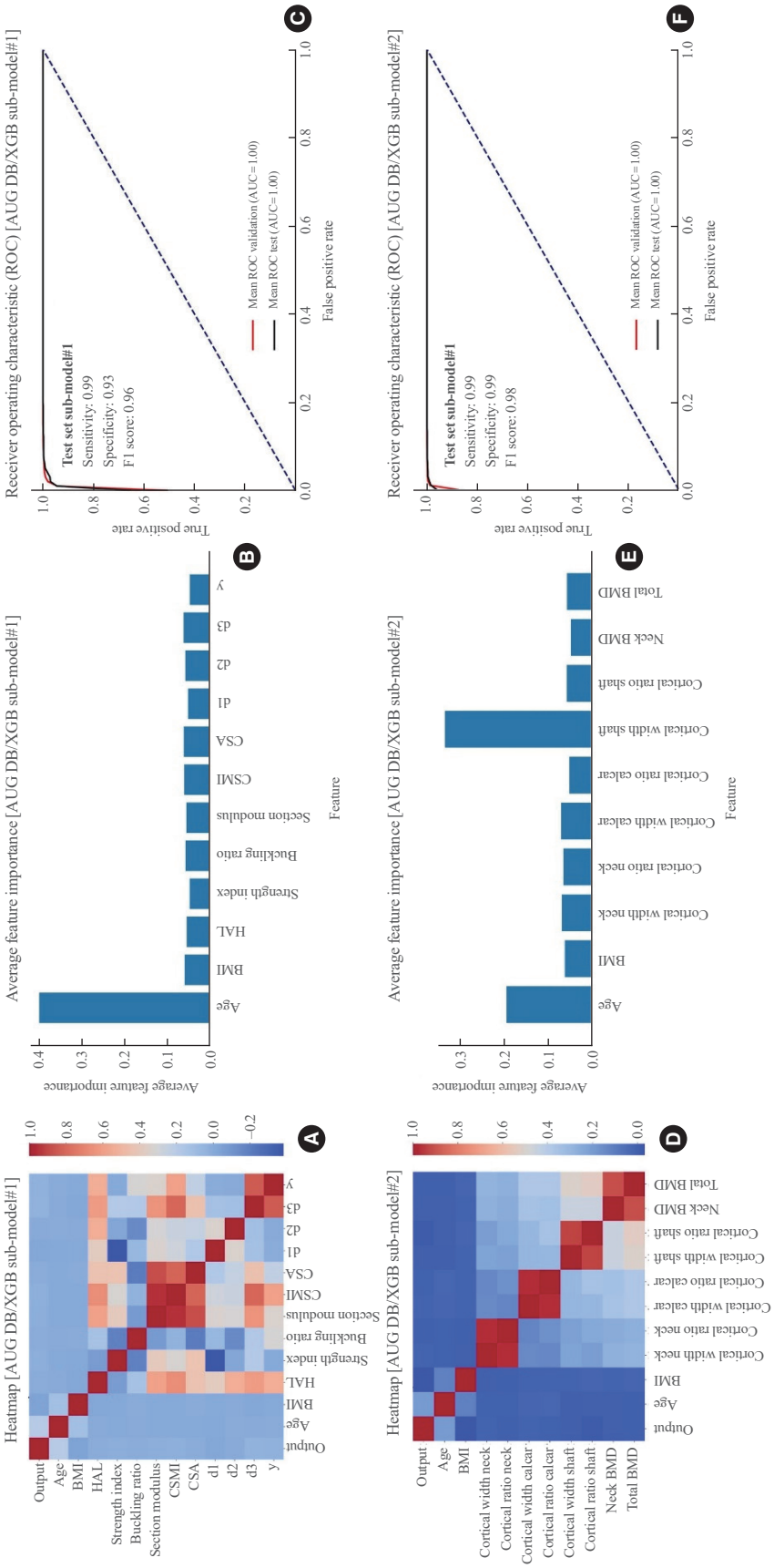

Fig. 5. Comparison between sub-model#1 and sub-model#2 trained with adaptive synthetic (ADASYN)-augmented datasets is shown. (A, D) Correlation heatmaps, (B, E) features importance charts, and (C, F) receiver operating characteristic (ROC) curves with results of sensitivity, specificity, and average accuracy for each sub-model. AUG DB, database without augmented data; XGB, extreme gradient boosting; BMI, body mass index; HAL, hip axis length; CSMI, cross-sectional moment of inertia; CSA, cross-sectional area; d1, distance from the femoral head center to section of minimum CSMI along neck axis; d2, distance along the neck axis from the center of the femoral head to the neck/shaft axis intersection; d3, average diameter of the femoral neck; y, distance from the center of mass to the superior neck margin; BMD, bone mineral density; AUC, area under the curve.

ADASYN scheme prior to training. The raw data for each database included the age and BMI features, which were paired with geometry predictors in sub-model #1 and material predictors in sub-model #2.

Sub-model #1 incorporated a range of parameters including the hip axis length, strength index, buckling ratio, section modulus, cross-sectional moment of inertia, cross-sectional area, d1, d2, d3, and the distance from the center of mass to the superior neck margin of the femur. In contrast, sub-model #2 utilized different metrics such as cortical width at the neck, cortical ratio at the neck, cortical width at the calcar, cortical ratio at the calcar, cortical width at the shaft, cortical ratio at the shaft, the BMD at the femur neck, and the average BMD of the femur in the entire 2D-DXA image analyzed. The shaft-neck angle and shaft axis angle were deliberately omitted due to their minimal relevance. Correlation heatmaps for both models are shown in Fig. 5A, D. The former displays a correlation among variables that is strikingly similar to that observed in the entire database, as indicated in Fig. 3A.

However, as reported in Fig. 5D, sub-model #2 did not exhibit any inversely proportional relationships among the variables, and most of the strong interactions between variables occur diagonally across the matrix. The importance of the two features suggests a similar interpretation to that provided for the correlation heatmap: sub-model #1 aligns with the trend of the RAW database, where the binary classification was predominantly determined by the age feature, as shown in Fig. 3B. In contrast, sub-model #2 displayed a markedly different pattern of feature importance, focusing more on cortical width at the shaft, while age became less significant, as indicated in Fig. 5E. Regarding the ROC curves in Fig. 5C, F, both sub-models—utilizing a mix of raw and synthetic data with a nearly 1:1 ratio between NFX and FX classes—demonstrated high prediction accuracy in terms of sensitivity and specificity across training, validation, and test sets. However, sub-model #1, which did not include data on material properties, exhibited approximately 6% lower specificity than sub-model #2. This suggests that sub-model #1 was more likely to classify borderline cases as positive (FX), potentially prioritizing sensitivity over specificity. The heatmaps in Fig. 5A, D show that having both positive and negative correlations leads to slightly higher ambiguity in classification. This is reflected in sub-model #1's lower specificity and F1 score compared to that of sub-model #2.

At this point, it is clear that geometry alone is insufficient to distinguish between FX and NFX classes; it must be combined with material-related features to achieve a definitive separation between the classes. Conversely, the material-related features depicted in Fig. 5D appear to contain sufficient information to independently facilitate a successful differentiation between FX and NFX cases. In this context, within the AC_17-21 database where cases are not categorized by fracture nature, the cortical to trabecular thickness ratio emerged as the key determinant, surpassing age, which was the top scoring factor in all other models, as indicated in Figs. 3B, 5B.

DISCUSSION

Considering the preliminary analysis of the RAW database and the model comparisons in Fig. 3, the 2D-DXA database alone lacks sufficient data balance for proper differentiation between FX and NFX cases. In particular, the XGB model excelled in training but showed a significant drop in the validation and test datasets, indicating overfitting. Conversely, the SVM model maintains relative stability across training, validation, and test datasets, but its lower F1 score suggests underfitting, particularly in comparison to the XGB model. The MLP model, based on a neural network architecture, allows for a high degree of complexity and generalization. However, in this scenario, the complexity of the models hampers the identification of feature boundaries, resulting in performance that is inferior to that of the XGB model [33]. In the case of the XGB model, the strong cross-influence between the target variables, the predominance of the age variable, as shown in Fig. 3B, and the relatively low correlation with the output class result, as reflected by a modest AUC of 0.77 when applied to untested data. Interestingly, balancing the dataset by randomly selecting NFX cases to match the number of FX cases led to varied performances across the three models, despite using the same data in all scenarios. Among the three tested formulations, the XGB solution demonstrated the most balanced results in terms of AUC and F1 scores on validation and test sets. Nevertheless, regardless of the model used, the predictions are relatively accurate but underscore the critical issue of relying on a small database, consisting of 484 cases, which is insufficient to model the complex interactions among the 22 features and the output classes.

The XGB algorithm was capable of capturing interactions among this set of variables during the training phase; however, it demonstrated poor extrapolation on both validation and test sets.

To address these challenges, data augmentation was employed using the ADASYN scheme. This approach generated synthetic data points that preserve the characteristics of the original data, based on the reciprocal distribution between the majority (NFX)

and minority (FX) classes. In this process, the entire AC_17-21 database was input into the ADASYN-XGB algorithm, which then produced synthetic FX cases to match the number of NFX cases. The synthetic data generated by ADASYN were derived from the raw data distribution of the majority and minority classes and did not introduce any new information to the feature distribution. However, their proper distribution in the latent space facilitated a balance between NFX and FX cases, particularly at the class boundaries, thereby improving both sensitivity and specificity. Employing the entire database and balancing it through the ADASYN scheme improved the AUC in both the validation and test datasets, even when the test data included cases not previously used in any stage of the training or validation process. While improving the representation of features critical for distinguishing between FX and NFX cases, the synthetic data preserve the overall significance of variables in the solution. This applied to variables such as cortical width at the calcar and cortical width at the shaft, as shown in Fig. 4A. The age feature was also emphasized. Thus, as expected, it is likely the primary factor influencing fractures, as also considered in algorithms such as FRAX [17-19].

The model training in this research is based solely on the interactions of features within the database, without any pre-conditioning to adjust the models' predictions for specific features. Notably, if the ADASYN-XGB model were adjusted for age and BMI, its predictions would likely align with those of existing tools like FRAX, which considers only a limited subset of the features examined in this study. Although it is beyond the scope of this research, an analysis of how these two features impact accuracy can be found in the supplementary material accompanying this paper.

The analysis presented in Fig. 5 underscores the importance of distinguishing between geometry and material properties during feature selection to improve our understanding of their respective associations with prevalent fractures. Bone geometry typically reflects inherited traits, whereas material properties are indicative of BMD reduction due to skeletal diseases such as osteoporosis or osteopenia. Initially, the relationship between features and output classes, as depicted in Fig. 5A, B, D, E, showed a marked difference between sub-model #1 and sub-model #2. The former closely mirrors the combined modeling approach presented in Fig. 4, particularly in the feature importance chart.

As shown in Fig. 5C, F, both sub-models performed remarkably well. However, the latter, which incorporated age along with material property variables, identified cortical width at the shaft as the predominant predictor in distinguishing between FX and NFX cases. In contrast, the former closely resembled the global modeling approach depicted in Fig. 4A. This observation underscores that, despite the well-known susceptibility of femur fractures at the neck region—the proximal end and narrowest part of the femur—it remains the most critical area for assessing potential fracture risks. This finding and the associated analysis align with the decision to avoid labeling the data based on the nature of the fracture. Consequently, the health of the bone became the sole feature class considered in estimating fracture risk. Given this focus, age emerged as the most significant factor, directly correlating with the overall weakening of the human body's musculoskeletal system.

It should be noted that this study was conducted using only a Korean database. Consequently, applying the model to databases from different ethnic backgrounds might yield varying conclusions, particularly if these databases are analyzed collectively. Additionally, the DXA data utilized in this research were exclusively sourced from the GE Lunar Prodigy system. Therefore, the applicability of our findings to other DXA systems has not been established and warrants further investigation. Moreover, due to limitations in resources and access, datasets from other institutions were not available, potentially introducing additional bias.

In this regard, it is important to remember that the inclusion of synthetic data essentially introduces a controlled bias into the training process. This approach is based on the implicit assumption that the raw data are consistent and that their distribution accurately represents the entire population. This assumption limits the generalizability of the model and underscores the need for future validation with diverse datasets.

Regarding censored data, it should be noted that the dataset used in this research is limited to a specific observation period, with no follow-up data available beyond this timeframe. Consequently, any health events, including fractures that occur after this period, are considered unobserved events and are not captured in the dataset. Furthermore, the absence of follow-up data and the unavailability of post-study health outcomes highlight the need for longer follow-up periods or continuous monitoring in future research.

The proposed methodology and the trained ADASYN-XGB model can be extended to other databases, provided that the distribution of features and their relationship with the output classes remain relatively constant. However, applying the trained model directly to cases involving different ethnicities may present limitations. Nevertheless, the ADASYN-XGB model can be

integrated with 2D-DXA post-processing software, facilitating the direct estimation of a binary fracture risk index associated with current bone health.

The dataset-balancing approach can also be applied to other 2D-DXA databases that are not specifically related to the femur bone. This allows for the expansion of fracture risk prediction to other areas of the human body and accommodates different test conditions. Furthermore, incorporating multi-institutional datasets in future studies could reduce potential biases, enhance the robustness of the research findings, and increase their clinical relevance and generalizability.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711174360, RS-2022-00164554). Prof. Luca Quagliato was supported by the RP-Grant 2024 of Ewha Womans University. The authors would also like to express their gratitude to Jung-Hyun Kim for her help in preparing the dataset.

AUTHOR CONTRIBUTIONS

Conception or design: L.Q., J.S., J.H., T.L., Y.S.C. Acquisition, analysis, or interpretation of data: L.Q., J.S., J.H., Y.S.C. Drafting the work or revising: L.Q., J.S. Final approval of the manuscript: L.Q., J.S., J.H., T.L., Y.S.C.

ORCID

Luca Quagliato *https://orcid.org/0000-0002-5379-8306*
Jiin Seo *https://orcid.org/0009-0008-8647-6427*
Taeyong Lee *https://orcid.org/0000-0002-0176-0525*
Yoon-Sok Chung *https://orcid.org/0000-0003-0179-4386*

REFERENCES

1. Cheung EY, Tan KC, Cheung CL, Kung AW. Osteoporosis in East Asia: current issues in assessment and management. Osteoporos Sarcopenia 2016;2:118-33.
2. Haseltine KN, Chukir T, Smith PJ, Jacob JT, Bilezikian JP, Farooki A. Bone mineral density: clinical relevance and quantitative assessment. J Nucl Med 2021;62:446-54.
3. Beltran-Sanchez H, Soneji S, Crimmins EM. Past, present, and future of healthy life expectancy. Cold Spring Harb Perspect Med 2015;5:a025957.
4. Crimmins EM. Recent trends and increasing differences in life expectancy present opportunities for multidisciplinary research on aging. Nat Aging 2021;1:12-3.
5. Chaput JP, Dutil C, Featherstone R, Ross R, Giangregorio L, Saunders TJ, et al. Sleep duration and health in adults: an overview of systematic reviews. Appl Physiol Nutr Metab 2020;45(10 Suppl 2):S218-31.
6. Park JH, Moon JH, Kim HJ, Kong MH, Oh YH. Sedentary lifestyle: overview of updated evidence of potential health risks. Korean J Fam Med 2020;41:365-73.
7. Rodriguez-Gomez I, Manas A, Losa-Reyna J, Rodriguez-Manas L, Chastin SF, Alegre LM, et al. Associations between sedentary time, physical activity and bone health among older people using compositional data analysis. PLoS One 2018;13:e0206013.
8. Pinheiro MB, Oliveira J, Bauman A, Fairhall N, Kwok W, Sherrington C. Evidence on physical activity and osteoporosis prevention for people aged 65+ years: a systematic review to inform the WHO guidelines on physical activity and sedentary behaviour. Int J Behav Nutr Phys Act 2020;17:150.
9. Kanis JA, Cooper C, Rizzoli R, Reginster JY; Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO) and the Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF). Correction to: European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 2020;31:801.
10. Kanis JA, Cooper C, Rizzoli R, Reginster JY; Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO) and the Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF). Executive summary of European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Aging Clin Exp Res 2019;31:15-17.
11. Gregson CL, Armstrong DJ, Bowden J, Cooper C, Edwards J, Gittoes NJ, et al. UK clinical guideline for the prevention

and treatment of osteoporosis. Arch Osteoporos 2022;17:58.
12. Albergaria BH, Chalem M, Clark P, Messina OD, Pereira RM, Vidal LF. Consensus statement: osteoporosis prevention and treatment in Latin America: current structure and future directions. Arch Osteoporos 2018;13:90.
13. LeBoff MS, Greenspan SL, Insogna KL, Lewiecki EM, Saag KG, Singer AJ, et al. The clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int 2022;33:2049-102.
14. Curtis EM, Woolford S, Holmes C, Cooper C, Harvey NC. General and specific considerations as to why osteoporosis-related care is often suboptimal. Curr Osteoporos Rep 2020;18:38-46.
15. Liu J, Curtis EM, Cooper C, Harvey NC. State of the art in osteoporosis risk assessment and treatment. J Endocrinol Invest 2019;42:1149-64.
16. Jiang X, Gruner M, Tremollieres F, Pluskiewicz W, Sornay-Rendu E, Adamczyk P, et al. Diagnostic accuracy of FRAX in predicting the 10-year risk of osteoporotic fractures using the USA treatment thresholds: a systematic review and meta-analysis. Bone 2017;99:20-5.
17. McCloskey EV, Harvey NC, Johansson H, Lorentzon M, Liu E, Vandenput L, et al. Fracture risk assessment by the FRAX model. Climacteric 2022;25:22-8.
18. Nguyen TV. Individualized fracture risk assessment: state-of-the-art and room for improvement. Osteoporos Sarcopenia 2018;4:2-10.
19. Kanis JA, Harvey NC, McCloskey E, Bruyere O, Veronese N, Lorentzon M, et al. Algorithm for the management of patients at low, high and very high risk of osteoporotic fractures. Osteoporos Int 2020;31:1-12.
20. Adami G, Rossini M, Fassio A, Viapiana O, Gatti D. Comments on Kanis et al.: Algorithm for the management of patients at low, high, and very high risk of osteoporotic fractures. Osteoporos Int 2020;31:1015.
21. Kanis JA, McCloskey EV, Harvey NC, Cooper C, Rizzoli R, Dawson-Hughes B, et al. The need to distinguish intervention thresholds and diagnostic thresholds in the management of osteoporosis. Osteoporos Int 2023;34:1-9.
22. Grassi L, Vaananen SP, Jehpsson L, Ljunggren O, Rosengren BE, Karlsson MK, et al. 3D finite element models reconstructed from 2D dual-energy X-ray absorptiometry (DXA) images improve hip fracture prediction compared to areal BMD in Osteoporotic Fractures in Men (MrOS) Sweden Cohort. J Bone Miner Res 2023;38:1258-67.
23. Aldieri A, Paggiosi M, Eastell R, Bignardi C, Audenino AL, Bhattacharya P, et al. DXA-based statistical models of shape and intensity outperform aBMD hip fracture prediction: a retrospective study. Bone 2024;182:117051.
24. Aldieri A, Bhattacharya P, Paggiosi M, Eastell R, Audenino AL, Bignardi C, et al. Improving the hip fracture risk prediction with a statistical shape-and-intensity model of the proximal femur. Ann Biomed Eng 2022;50:211-21.
25. Wu Q, Nasoz F, Jung J, Bhattarai B, Han MV, Greenes RA, et al. Machine learning approaches for the prediction of bone mineral density by using genomic and phenotypic data of 5130 older men. Sci Rep 2021;11:4482.
26. Sapthagirivasan V, Anburajan M. Diagnosis of osteoporosis by extraction of trabecular features from hip radiographs using support vector machine: an investigation panorama with DXA. Comput Biol Med 2013;43:1910-9.
27. Yoo TK, Kim SK, Kim DW, Choi JY, Lee WH, Oh E, et al. Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. Yonsei Med J 2013;54:1321-30.
28. Prakash UM, Kottursamy K, Cengiz K, Kose U, Hung BT. 4x-expert systems for early prediction of osteoporosis using multi-model algorithms. Measurement 2021;180:109543.
29. Sebro R, De la Garza-Ramos C. Support vector machines are superior to principal components analysis for selecting the optimal bones' CT attenuations for opportunistic screening for osteoporosis using CT scans of the foot or ankle. Osteoporos Sarcopenia 2022;8:112-22.
30. Sato Y, Yamamoto N, Inagaki N, Iesaki Y, Asamoto T, Suzuki T, et al. Deep learning for bone mineral density and T-score prediction from chest X-rays: a multicenter study. Biomedicines 2022;10:2323.
31. Kong SH, Ahn D, Kim BR, Srinivasan K, Ram S, Kim H, et al. A novel fracture prediction model using machine learning in a community-based cohort. JBMR Plus 2020;4:e10337.
32. Yu X, Ye C, Xiang L. Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing 2016;214:376-81.
33. Iliou T, Anagnostopoulos CN, Stephanakis IM, Anastassopoulos G. A novel data preprocessing method for boosting neural network performance: a case study in osteoporosis prediction. Inf Sci 2017;380:92-100.
34. Smets J, Shevroja E, Hugle T, Leslie WD, Hans D. Machine learning solutions for osteoporosis: a review. J Bone Miner Res 2021;36:833-51.
35. Ho CS, Chen YP, Fan TY, Kuo CF, Yen TY, Liu YC, et al. Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography. Arch

Osteoporos 2021;16:153.
36. Su R, Liu T, Sun C, Jin Q, Jennane R, Wei L. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 2020;385:300-9.
37. Kirilov N, Kirilova E, Krastev E. Using machine learning to predict bone mineral density from dual-energy X-ray absorptiometry images of the lumbar spine. CEUR Workshop Proc 2021;2933:220-6.
38. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell 2022;15:1-22.
39. Senjyu T, Mahalle PN, Perumal T, Joshi A. Information and Communication Technology for Intelligent Systems (Proceedings of ICTIS 2020). Singapore: Springer; 2021. Chapter 3, Convolutional neural networks: an overview and its applications in pattern recognition; p. 21-30.
40. Lee S, Quagliato L, Park D, Kwon I, Sun J, Kim N. A new approach to preform design in metal forging processes based on the convolution neural network. Appl Sci 2021;11:7948.
41. Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci Rep 2022;12:15889.
42. Lee S, Park J, Kim N, Lee T, Quagliato L. Extreme gradient boosting-inspired process optimization algorithm for manufacturing engineering applications. Mater Des 2023;226:111625.
43. Joo Y, Namgung E, Jeong H, Kang I, Kim J, Oh S, et al. Brain age prediction using combined deep convolutional neural network and multi-layer perceptron algorithms. Sci Rep 2023;13:22388.
44. Moldagulova A, Sulaiman RB. Using KNN algorithm for classification of textual documents. Proceedings of the 8th International Conference on Information Technology (ICIT); 2017 May 17-18; Amman, Jordan. New York: IEEE; 2017. p. 665-71.
45. Garcia EK, Feldman S, Gupta MR, Srivastava S. Completely lazy learning. IEEE Trans Knowl Data Eng 2009;22:1274-85.
46. Letteri I, Di Cecco A, Dyoub A, Della Penna G. A novel resampling technique for imbalanced dataset optimization. arXiv 2020 Dec 30 [Preprint]. https://doi.org/10.48550/arXiv.2012.15231.
47. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008 Jun 1-8; Hong Kong, China. New York: IEEE; 2008. p. 1322-8.