Osteoporosis International (2025) 36:811–821
https://doi.org/10.1007/s00198-025-07437-w

ORIGINAL ARTICLE

Ensemble-learning approach improves fracture prediction using genomic and phenotypic data

Qing Wu[1] · **Jongyun Jung**[1]

Received: 1 April 2024 / Accepted: 14 February 2025 / Published online: 7 March 2025

Abstract

Summary This study presents an innovative ensemble machine learning model integrating genomic and clinical data to enhance the prediction of major osteoporotic fractures in older men. The Super Learner (SL) model achieved superior performance (AUC = 0.76, accuracy = 95.6%, sensitivity = 94.5%, specificity = 96.1%) compared to individual models. Ensemble machine learning improves fracture prediction accuracy, demonstrating the potential for personalized osteoporosis management.

Purpose Existing fracture risk models have limitations in their accuracy and in integrating genomic data. This study developed and validated an innovative ensemble machine learning (ML) model that combines multiple algorithms and integrates clinical, lifestyle, skeletal, and genomic data to enhance prediction for major osteoporotic fractures (MOF) in older men.

Methods This study analyzed data from 5130 participants in the Osteoporotic Fractures in Men cohort Study. The model incorporated 1103 individual genome-wide significant variants and conventional risk factors of MOF. The participants were randomly divided into training (80%) and testing (20%) sets. Seven ML algorithms were combined using the SL ensemble method with tenfold cross-validation MOF prediction. Model performance was evaluated on the testing set using the area under the curve (AUC), the area under the precision-recall curve, calibration, accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and reclassification metrics. SL model performances were evaluated by comparison with baseline models and subgroup analyses by race.

Results The SL model demonstrated the best performance with an AUC of 0.76, accuracy of 95.6%, sensitivity of 94.5%, specificity of 96.1%, NPV of 95.1%, and PPV of 94.7%. Among the individual ML, gradient boosting performed optimally. The SL model outperformed baseline models, and it also achieved accuracies of 93.1% for Whites and 91.6% for Minorities, outperforming single ML in subgroup analysis.

Conclusion The ensemble learning approach significantly improved fracture prediction accuracy and model performance compared to individual ML. Integrating genomic and phenotypic data via the SL approach represents a promising advancement for personalized osteoporosis management.

Keywords Accuracy · Ensemble learning · Fracture · Genomics · Machine learning · Osteoporosis · Sensitivity · Specificity · Super Learner

Introduction

Osteoporotic fractures significantly impact public health, particularly among older men, where the lifetime risk can reach up to 13% [1]. As male life expectancy increases, the health burdens from these fractures are expected to rise, emphasizing the critical need for precise identification of high-risk individuals. Current models like FRAX [2], QFracture [3], and the Garvan calculator [4], despite their widespread use, fall short in predictive accuracy, particularly in their ability to account for male-specific risks and integrate genomic data, which could substantially refine risk assessments [5–7].

The limitations of current fracture risk prediction tools, such as FRAX and QFracture, are well-documented, particularly their tendency to underestimate risks in those with osteopenia or osteoporosis and the significant disparities arising from model constructs and calibration issues [8, 9]. These challenges highlight the urgent need for advanced

✉ Qing Wu
Qing.Wu@osumc.edu

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, USA

models that not only address these issues but also leverage genomic data to offer more individualized and accurate fracture risk predictions.

Machine learning (ML) applications in osteoporosis research have shown promising results in improving the detection, classification, and risk prediction of osteoporosis and related fractures. Studies have demonstrated the effectiveness of ML models, such as recurrent neural networks (RNN) and lightGBM, in predicting osteoporosis from various datasets, suggesting potential for non-invasive screening methods [10–12]. Moreover, Wu et al. have further underscored the value of ML models in fracture prediction, highlighting the significance of incorporating genetic profiling to enhance clinical risk assessments [13, 14]. However, single ML model approaches have limitations, including potential biases and overfitting, underscoring the need for ensemble learning strategies that can improve prediction robustness and accuracy by combining multiple ML models.

Given the gaps in current methodologies and the potential of ML, this study is aimed at developing an ensemble learning model for fracture prediction in older men, integrating both phenotypic and genomic data. This novel approach seeks to surpass the limitations of existing models and single ML algorithms, aiming to provide a comprehensive, accurate, and widely applicable model for predicting osteoporotic fractures. This research is aimed at significantly improving management strategies and reducing the associated health burdens by considering the multifactorial nature of osteoporotic fractures.

Materials and methods

Data source and study participants

The data utilized in this study was obtained from the Osteoporotic Fractures in Men Study (MrOS), a database stored in the Database of Genotypes and Phenotypes (dbGaP; Accession: phs000373.v1. p1). The MrOS project investigates factors related to bone health, including anthropometric, lifestyle, and medical variables, in older men residing within community settings [15, 16]. The initial participant pool, enrolled between March 2000 and April 2002, consisted of 5994 men aged 65 years or more with no history of bilateral hip replacement. Our study incorporated data from 5130 participants, selected based on genotypic and phenotypic data availability. The Ohio State University (OSU) institutional review board approved the study, with participants' informed consent acquired in the MrOS project. Data were anonymized before access for the study, and the OSU IRB waived the need for additional consent (IRB number 2022E1150).

Ascertainment of major osteoporotic fracture and covariate assessment

The primary outcome was the occurrence of a major osteoporotic fracture (MOF), categorized as a hip, spine (clinical), wrist, or humerus fracture reported during the study follow-up period. Of the total participants ($n = 5130$), 451 men (8.8%) reported an MOF. Bone mineral density (BMD) at the total body, femur, and lumbar spine (L1 to L4) was measured using dual-energy X-ray absorptiometry (DXA). Quantitative ultrasound (QUS) parameters (including broadband ultrasonic attenuation, speed of sound, and the quantitative ultrasonic index) were measured at the right heel. Covariates, including age, race, smoking status, alcohol consumption, and mobility limitations, were collected using self-reported questionnaires and standard procedures during the baseline screening. Height and weight at the baseline were measured by an examiner using standard equipment, including a Harpenden stadiometer and balance beam scale. At baseline, walking speed was determined by timing the completion of a 6-m course performed at the participant's usual walking speed [17]. Alcohol consumption was determined via questionnaire data, later categorized into "drinking 3 units or more per day" based on any weekly alcohol intake. Grip strength was assessed with a handheld dynamometer, and impairment in Instrumental Activities of Daily Living (IADL) was ascertained by querying participants about potential difficulty with specific activities [18]. Serum calcium was measured using a Roche COBAS Integra 800.

Genotyping data

Genomic DNA extraction was performed using whole blood samples collected at the baseline, with written consent obtained for DNA utilization. Quality-controlled genotype data files were obtained from dbGaP and processed using PLINK [19]. Genotype imputation was carried out utilizing the Michigan Imputation Server [20], employing the Haplotype Reference Consortium imputation reference panel [21] and the Positional Burrows-Wheeler Transform imputing algorithm [22] to ensure the robust quality of genotype imputation. Following the findings reported in the 2019 GWAS study by Morris et al., 1103 associated single nucleotide polymorphisms (SNPs) were selected for the present investigation [23] because these SNPs demonstrated conditional independence at the threshold of genome-wide significance ($p < 6.6 \times 10^{-9}$) with estimated bone mineral density through heel quantitative ultrasound (eBMD). All 1103 SNPs underwent successful imputation within the MrOS dataset and were subsequently incorporated into the analytical framework. The imputation quality was excellent, with a mean R^2 of 0.99.

Machine learning models

Data division

The dataset ($N = 5130$) was divided into a training set (80%) and a testing set (20%) for model development and final evaluation. Stratified sampling was used to ensure that the training and testing sets maintained the same proportion of major osteoporotic fracture (MOF) cases (8.8%) and non-cases, preventing data imbalance from affecting model performance. The testing set remained unseen throughout model training and validation and was reserved for final performance evaluation. To ensure the independence of the training and testing sets, we conducted a genetic-relatedness analysis [24]. Specifically, we computed the genetic relationship matrix separately for the training and testing sets and then compared each individual in the training set to every individual in the testing set. Our analysis confirmed that all pairwise relatedness values between a training and a testing individual were consistently very low, with no values exceeding 0.05. This is well below the threshold typically used to identify first-degree (siblings) or second-degree (avuncular) relatives, where relatedness values are generally above 0.25 and 0.125, respectively [25, 26].

Base learners

The base learners included random forest, gradient boosting, support vector classification (SVC), K-nearest neighbors (KNN), deep neural networks (DNN), Gaussian Naive Bayes, and bagging classifier. Hyperparameter tuning was performed for each model using grid search within the training set, with cross-validation used to determine the optimal parameters. The bagging classifier uses random sampling to fit base classifiers on subsets of the original dataset and then aggregates individual predictions [27]. The DNN model consists of layers of connected units based on artificial neural networks, with information fed forward through the network. Gaussian Naive Bayes employs a probabilistic classifier assuming that each class's continuous values follow a Gaussian distribution [28]. Gradient boosting constructs a prediction model through a stage-wise ensemble of weak models using a differentiable loss function [29]. The KNN classifier, a non-parametric method, identifies the most frequent label among data points nearest to the query [30]. The random forest model generates multiple weak-learner trees, each built on a subset of bootstrapped samples from the original dataset. The SVC method determines the optimal hyperplane in the feature space that separates data points of different classes, maximizing the margin between them and the nearest data points [31]. The base learners were evaluated across various hyperparameter ranges, with the final hyperparameter values selected based on their performance during tenfold cross-validation.

Tenfold cross-validation

A tenfold cross-validation process was applied to the training set to avoid overfitting and ensure robust performance estimation. The training data were divided into ten equal subsets (folds). During each iteration, nine folds were used for model training, and the remaining onefold was used for validation. This process was repeated ten times, with each fold serving as the validation fold once, ensuring that all data points were used for training and validation. Performance metrics such as accuracy, AUC, sensitivity, and specificity were averaged across all ten validation folds to provide a reliable estimate of model performance. This approach minimizes the risk of overfitting, as the model is evaluated on data not seen during training in each iteration.

Super Learner

The Super Learner (SL), an ensemble learning method, was employed to combine the predictions of the base learners and create a more accurate final model (Fig. 1). The SL model was trained on the predictions generated by the base learners during cross-validation. Specifically, the predictions from the validation fold across all base learners were used as input to the SL. The SL assigns weights to each base learner based on its performance during cross-validation, optimizing the ensemble's ability to generalize to unseen data. The SL was tuned using these out-of-sample predictions, ensuring the model did not overfit the training data. By using the validation fold predictions to inform the weighting, the SL effectively captured the strengths of the base models while reducing the impact of their individual weaknesses. The SL model determines the optimal weights assigned to each base learner through cross-validation during the training phase. Specifically, the SL employs a meta-learning approach where the base learners are trained on distinct subsets of the training data, and their performance is evaluated [32]. The optimal combination of base learners and their corresponding weights is determined by minimizing the cross-validated error on the testing set. This process ensures that the SL assigns higher weights to more accurate and informative base learners, allowing it to adaptively combine the strengths of diverse models for improved predictive performance [32].

Data analysis

Missing data were imputed using the median for continuous variables, which proved an effective imputation method [33]. For categorical variables, we replaced missing values with the default value (No) applied, as recommended by the FRAX guideline [2]. The greatest degree of missingness was the speed of sound ($n = 344$, 7.7%), followed by the grip strength ($n = 104$, 2.1%) and total spine BMD ($n = 10$, 0.2%).

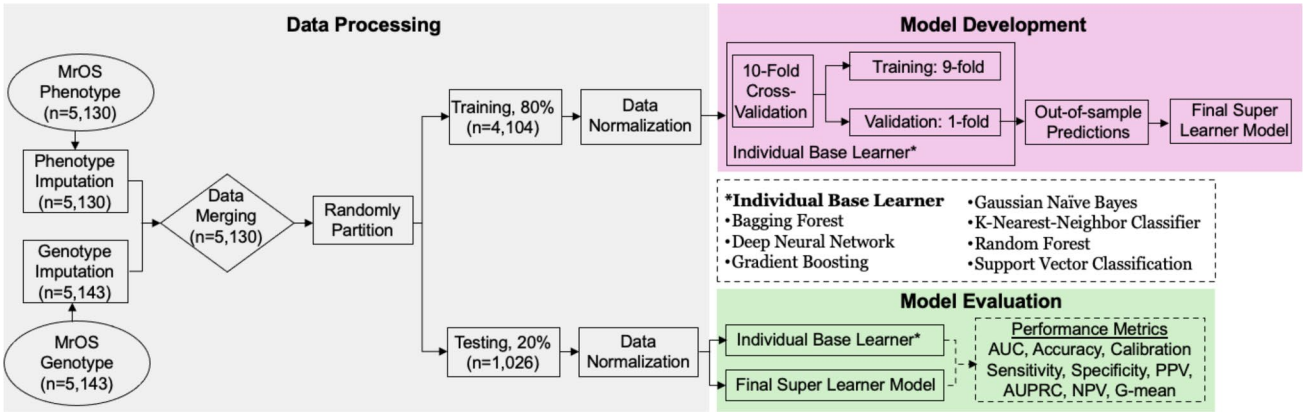

Fig. 1 Overall study flow. We randomly separated the Osteoporotic Fractures in Men Study (MrOS) data ($n=5130$) into the training (80%, $n=4104$) and testing (20%, $n=1026$). A tenfold cross-validation methodology was implemented to divide the data into ten disjoint subsets. Nine subsets were used to train each base learner, and the remaining subset was reserved for validation. The model's predictive performance was tested on an independent testing data set

We integrated phenotype ($n=5130$) and genotype ($n=5143$) data post-imputation. Our model included covariates such as age, race, body weight, height, lifestyle factors (smoking, alcohol consumption, walking speed), impairment of instrumental activities of daily living, three BMD measures (femoral neck, total hip, and total spine), mobility limitations, grip strength, serum calcium, and ultrasonic speed of sound (SOS). All predictors were adjusted for clinical location. We included multiple BMD measurements, as site-specific BMD is strongly associated with fractures at corresponding sites [34–36].

Standardization for continuous variables was performed post-split using zero-mean and unit variance, while categorical variables, including 1103 SNPs, were standardized through one-hot encoding. SNPs were created based on additive effects using Plink's –recodeA function [19]. Only SOS was included among the highly correlated QUS measurements after confirming its superior predictive performance for MOF upon centering.

Due to the low fracture rate in MrOS (8.8%), which led to data imbalance and presented challenges in model performance assessment, we utilized the synthetic minority over-sampling (SMOTE) technique [37] for data resampling in the training set. Hyperparameters for each base learner were selected via Scikit-learn's grid search cross-validation method. We specified a set of values for parameters to search over each base learner. The parameters that underwent a model selection phase in the grid search cross-validation method are shown in Table S1 with the corresponding range of values. These hyperparameters were selected based on empirical observations, existing literature, and preliminary experiments conducted on this dataset—the chosen hyperparameters aimed to optimize the performance of each algorithm in terms of accuracy and generalization capabilities.

Model evaluation

Model performance was evaluated using an independent testing set ($n=1026$). Receiver operating characteristic (ROC) curves were generated to assess the true positive rate versus the false positive rate at various thresholds. We adopted multiple evaluation metrics, including the area under the ROC curve (AUC), accuracy, sensitivity, specificity, negative predictive value, positive predictive value, and G-mean. To assess the discriminative performance of these models, precision-recall (PR) curves were generated for each model using the testing dataset ($n=1026$). The curves show the trade-off between precision (positive predictive value) and recall (sensitivity) at various classification thresholds. The area under the PR curve (AUPRC) was computed for each model. Calibration curves were also generated to evaluate the accuracy of predicted probabilities with the actual fracture risk. For each model, the predicted probabilities were divided into three quantiles, and the mean actual risk for each quantile was plotted against the predicted mean risk. The curves provide an assessment of how well the predicted probabilities align with the observed outcomes.

We evaluated the predictive performance of the SL model for fracture risk assessment in comparison to the two baseline models below. Since FRAX scores were not available in the dataset, we could not directly compute its performance due to the absence of time-to-event data and other necessary variables. Instead, we constructed the following baseline models.

1. Baseline 1: A logistic regression model was constructed using available FRAX clinical risk factors (CRFs), including age, smoking status, alcohol intake, femo-

ral neck BMD, weight, height, and sex from the MrOS dataset, as the fracture outcome was treated as a binary outcome.
2. Baseline 2: An extended logistic regression model incorporating FRAX CRFs plus additional clinical variables, including total hip and total spine BMD, walking speed, mobility limitations, IADL, grip strength, speed of sound, and serum calcium.

Due to the limitations of conventional statistical models, we could not incorporate thousands of genetic variants as predictors in logistic regression. However, the SL model was designed to integrate both genetic and clinical predictors. The SL model was trained using all CRFs from the extended model (Baseline 2) along with 1103 genetic variants. Besides AUC, reclassification metrics, including net reclassification improvement (NRI) and integrated discrimination improvement (IDI), were used to evaluate the performance of SL compared to baseline models.

Genomic data contribution

To evaluate the contribution of genomic data to fracture risk prediction, we conducted a comparative analysis using models with and without genomic predictors (1103 SNPs). Specifically, the full model included the genomic predictors alongside clinical risk factors. In contrast, the reduced model excluded all genomic predictors while retaining the same clinical risk factors. The AUC values were computed for each ML algorithm in both scenarios to quantify the impact of genomic data on model discrimination. The performance of each model was assessed using the testing dataset ($n = 1026$) to ensure consistent evaluation and avoid overfitting. AUC values were compared to determine the improvement in discrimination attributable to the inclusion of genomic data. This analysis highlights the added predictive value of genomic information in fracture risk prediction and its contribution to enhancing the performance of various predictive algorithms.

Subgroup analysis

We performed a subgroup analysis based on race, contrasting White participants ($n = 4616$) with Minorities, encompassing individuals of African American, Asian, Hispanic, and other ethnicities ($n = 514$). We employed the McNemar statistical test to compare prediction accuracy between models in a testing set stratified by two race groups (White and Minorities) [38]. All analyses were performed in Python (v3.7.3) using the Scikit-learn package (https://www.python.org) [39]. A P-value of < 0.05 was considered statistically significant.

Results

Participant characteristics

Among the study population of 5130 participants, 8.8% ($n = 451$) experienced at least one MOF during 12 years of follow-up. Compared to participants without fractures, those with fractures were older and had lower bone mineral density at the femoral neck, total hip and spine, lower body weight, slower walking speed, more mobility limitations, weaker grip strength, and reduced speed of sound (all $P < 0.001$) (Table 1).

Stratified analyses revealed differences in baseline parameters between White and Minority participants. Among White participants with fracture ($n = 420$), significant differences vs. those without fracture were observed for age, BMD at all sites, height, weight, walking speed, instrumental activities of daily living, grip strength, and speed of sound (all $P < 0.001$) (Table S2). For Minority participants with fracture ($n = 31$), significant differences vs. those without fracture were limited to BMD at all sites and speed of sound (all $P < 0.001$) (Table S2) due to the small sample size.

Model evaluation

The SL ensemble model demonstrated superior performance compared to individual base learner models in predicting MOF in the testing dataset ($n = 1026$). The SL attained the highest area under the curve (AUC) of 0.76, surpassing the AUC of the top individual model, gradient boosting (AUC 0.72), as well as other individual models including random forest (AUC 0.71), Gaussian Naive Bayes (AUC 0.65), bagging forest (AUC 0.62), KNN (AUC 0.58), DNN (AUC 0.56), and SVC (AUC 0.55) (Fig. 2).

The area under the PR curve (AUPRC) was 0.31 with the SL model, indicating moderate model performance in identifying true fractures while minimizing false positives (Figure S1). The SL model achieved the best balance between precision and recall, maintaining higher precision across most recall values. Conversely, the SVC showed poor performance (AUPRC 0.18). Other models, such as gradient boosting (AUPRC 0.28) and random forest (AUPRC 0.27), demonstrated intermediate performance but showed a steep decline in precision at higher recall levels. These results underscore the challenges in fracture prediction within an imbalanced dataset and highlight the advantage of ensemble methods by using SL.

The SL model demonstrated the highest calibration, with predicted probabilities closely aligning with observed fracture risks across all quantiles (Fig. 3). This exceptional performance underscores the ensemble model's ability to

Table 1 Baseline descriptive statistics of 5130 participants stratified by major osteoporotic fracture (MOF) status

Variable*	Participants with MOF ($N=451$)	Participants without MOF ($N=4679$)	P-value**
Age (years), mean (SD)	75.9 (6.19)	73.6 (5.84)	< 0.001
Femoral neck BMD (g/cm^2), mean (SD)	0.71 (0.12)	0.79 (0.13)	< 0.001
Total hip BMD (g/cm^2), mean (SD)	0.87 (0.14)	0.96 (0.14)	< 0.001
Total spine BMD, mean (SD)	0.99 (0.18)	1.08 (0.19)	< 0.001
Height (cm)	174 (6.65)	174 (6.78)	0.11
Weight (kg)	80.6 (13.1)	83.4 (13.3)	< 0.001
Race (White), *n* (%)	420 (93.1%)	4196 (89.7%)	0.025
Smoking (current), *n* (%)	18 (4.0%)	153 (3.3%)	0.58
Alcohol (yes), *n* (%)	230 (51.0%)	2413 (51.6%)	0.83
Walking speed (m/s), mean (SD)	1.05 (0.25)	1.08 (0.28)	0.017
Mobility limitations, mean (SD)	0.26 (0.58)	0.19 (0.5)	0.007
Impairment of instrumental activities of daily living (score), mean (SD)	0.49 (0.99)	0.36 (0.85)	0.005
Grip strength, mean (SD)	35.5 (7.9)	38.7 (8.1)	< 0.001
Speed of sound, mean (SD)	1540 (31.8)	1560 (35.8)	< 0.001
Serum calcium (mg/Dl), mean (SD)	9.28 (0.44)	9.32 (0.39)	0.064

*Continuous variables were expressed as mean (SD, standard deviation), and categorical variables were expressed as number (%). **P-values were obtained by t-test for continuous variables and chi-square test for categorical variables. *BMD*, bone mineral density

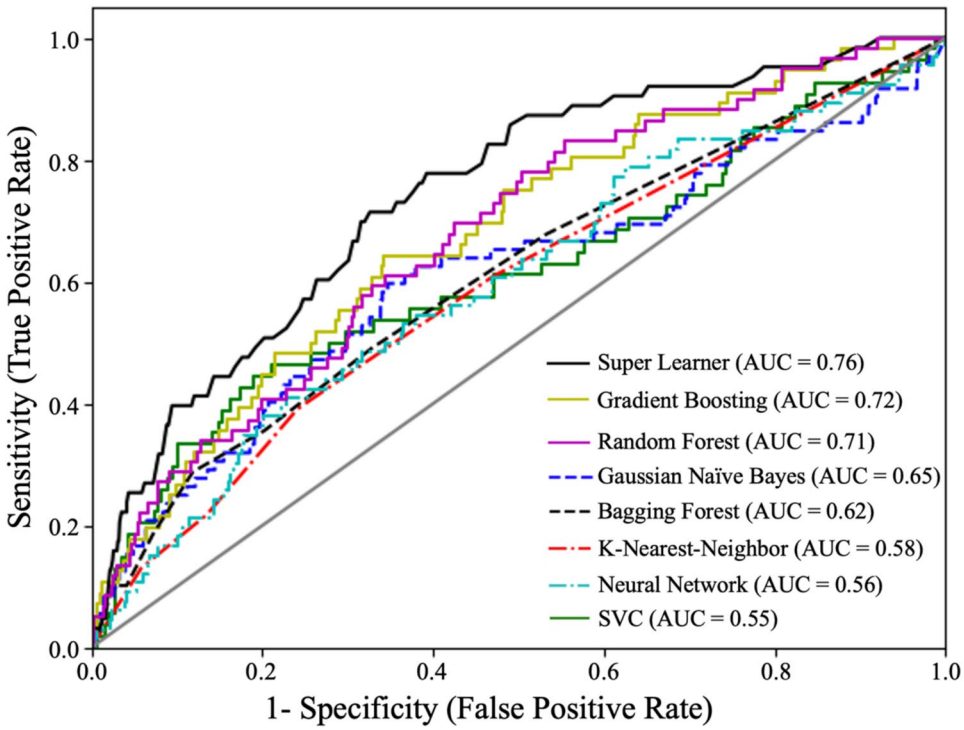

Fig. 2 Comparisons of receiver operating characteristic curve between various machine learning models for MOF prediction in the testing data set ($n=1026$). AUC, area under the curve; SVC, support vector classification

effectively combine predictions from multiple base learners, resulting in more accurate risk estimation. Among the base learners, gradient boosting showed the next best calibration performance, followed by random forest. While both models showed reasonable alignment between predicted and observed risks, their calibration curves displayed greater deviations compared to the SL model, particularly in the higher and lower risk quantiles.

The SL model also excelled across other performance metrics, including accuracy (95.6%), sensitivity (94.5%), specificity (96.1%), negative predictive value (95.1%), positive predictive value (94.7%), and G-mean (95.3%). Among

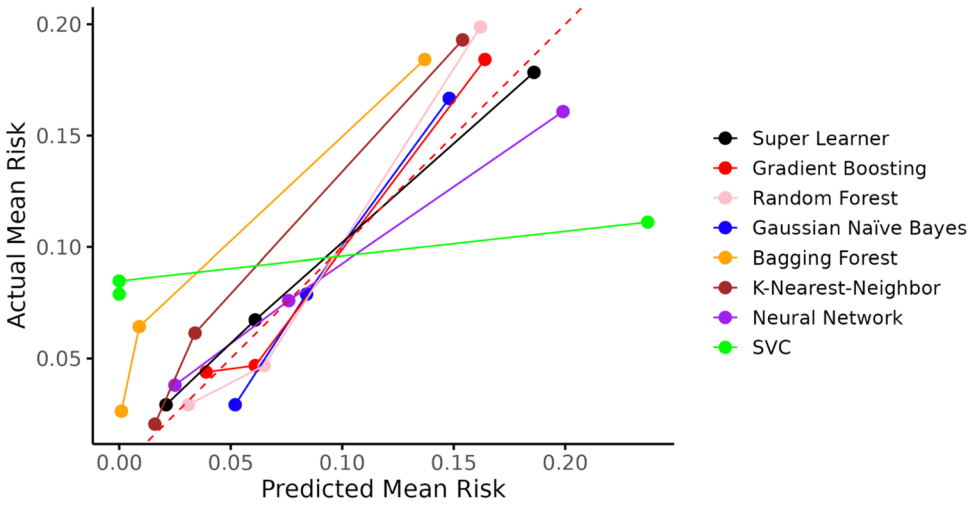

Fig. 3 Calibration curves for super learner and individual classification models. Calibration curves show the relationship between predicted and actual fracture risks in the testing dataset ($n = 1026$). Predicted probabilities are divided into three quantiles (X-axis), with curves displaying the mean actual risk for each quantile (Y-axis). The dashed red line indicates perfect calibration

individual models, gradient boosting emerged as the top performer with accuracy (92.8%), sensitivity (92.5%), specificity (92.5%), negative predictive value (92.7%), positive predictive value (91%), and G-mean (92.5%) (Fig. 4). Thus, the SL model demonstrated superior discrimination for predicting osteoporotic fractures compared to individual ML models.

The NRI values increased from 0.02 in Baseline 1 (FRAX CRFs) to 0.05 in Baseline 2 (FRAX CRFs + Additional), reaching 0.08 with the SL model. Similarly, IDI values improved from 0.03 in Baseline model 1 to 0.06 in Baseline model 2, with the SL model achieving the highest value of 0.09 (Figure S2). The SL model outperformed both baseline models with AUC increased to 0.76 from 0.66 in Baseline model 1 and 0.71 from Baseline model 2 (Figure S3). These results demonstrated the SL model has a better predictive performance than conventional models for fracture risk prediction.

Genomic data contribution

To assess the contribution of genomic data to fracture risk prediction, we compared the discriminative performance of ML models with and without genomic predictors while keeping other clinical risk factors constant. Models incorporating genomic data consistently demonstrated higher AUC values across all ML algorithms. For example, the AUC for the SL model improved from 0.72 (without genomic data) to 0.76 (with genomic data). These results underscore the value of integrating genomic information to improve fracture prediction accuracy.

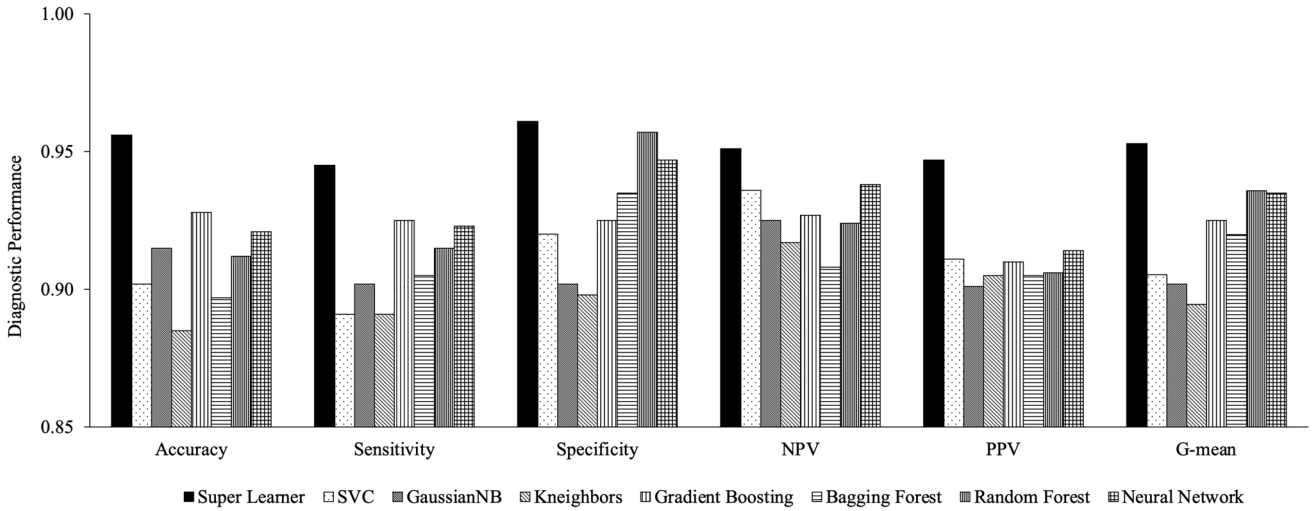

Fig. 4 Diagnostic performance between various machine learning models in predicting major osteoporotic fracture in the testing dataset ($n = 1026$) with the measurement of accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and G-mean

Subgroup analysis

Subgroup analysis examined model performance by race (White versus Minorities). Despite the smaller sample of Minorities ($n=514$), the SL model performed best in both subgroups, with an accuracy of 93.1% for Whites and 91.6% for Minorities (Figures S4-S5). Among individual models, gradient boosting had the highest performance for Whites with an accuracy of 90.5%, sensitivity of 90.3%, specificity of 91.2%, negative predictive value of 90.6%, positive predictive value of 90.4%, and G-mean 90.7%. For Minorities, gradient boosting also had the best individual model performance with an accuracy of 88.7%, sensitivity of 88.2%, specificity of 89.2%, negative 86.6%, positive 86.4%, and G-mean 88.7%. Thus, the SL model outperformed individual models in White and Minority subgroups. Gradient boosting emerged as the top-performing individual model across subgroups.

Model weights

The optimal weights of the seven base learner models in the SL model are shown in Table 2. Across the testing dataset ($n=1026$), gradient boosting had the highest weight at 0.315, followed by random forest (0.235) and DNN (0.118). A similar weight distribution was seen in White participants ($n=923$), with gradient boosting having the greatest weight at 0.329. However, the weight distribution varied in Minorities ($n=103$), with gradient boosting still the highest at 0.248 but followed closely by DNN (0.239) and random forest (0.182). Therefore, gradient boosting contributed the most weight to the SL model predictions overall and in Whites, but DNN and random forest had more contributions among Minorities. The optimal weighting differed by subgroup.

Comparison of prediction accuracy

The accuracy of the models in predicting MOF was compared using McNemar's test (Table S3). With Bonferroni correction applied ($\alpha=0.05/28=0.0018$), the difference in accuracy between each model pairwise comparison was statistically significant ($P<0.001$) among White participants. In Minorities, most pairwise comparisons were also statistically significant, with the exceptions of Gaussian Naive Bayes vs. KNN ($P=0.18$), neural network ($P=0.09$), random forest ($P=0.1$), and SVC ($P=0.1$). Additionally, KNN vs. neural network ($P=0.08$), random forest ($P=0.12$), and SVC ($P=0.15$) were non-significant. Therefore, nearly all model pairwise comparisons showed statistically significant differences in prediction accuracy among Whites. However, several comparisons did not reach statistical significance for Minorities, likely due to the smaller sample size.

Discussion

Our study highlights the benefits of the SL model's ensemble learning approach in fracture prediction. Combining multiple base learners enhances predictive performance by leveraging their individual strengths and capturing complex risk factor interactions. While ensemble learning incurs higher computational costs and complexity compared to single-based approaches, advancements in big data platforms and processing technology alleviate this constraint.

In this study, we compared the performance of the SL model to baseline models of FRAX predictors. The SL model outperformed the baseline model with FRAX predictors and the baseline model with both FRAX predictors and other clinical risk factors in terms of AUC and reclassification metrics, highlighting the better performance of our SL models compared to conventional modeling approaches. Previous studies using the MrOS cohort provide a benchmark for FRAX's performance, with Ettinger et al. [40] reporting AUC values of 0.63 (without BMD) and 0.67 (with BMD) for predicting MOF and Harvey et al. [41] observing AUC values of 0.61 (without BMD) and 0.63 (with BMD) after adjusting for FRAX risk factors and falls. In comparison, the SL model outperforms these, demonstrating the added value

Table 2 Estimated optimal model weights for the Super Learner, constructed from seven base learner models. The optimal model weights were determined by minimizing the cross-validated error on the training set and computed in each participant group in the testing dataset ($n=1026$)

Base learner model	Optimal weights		
	All participants ($n=1026$)	White participants ($n=923$)	Minorities participants ($n=103$)
Bagging classifier	0.041	0.025	0.069
Deep neural network	0.118	0.126	0.239
Gradient boosting	0.315	0.362	0.248
Gaussian Naïve Bayes	0.104	0.096	0.085
K-nearest neighbor classifier	0.098	0.105	0.128
Random forest	0.235	0.228	0.182
Support vector classification	0.089	0.058	0.049

of incorporating genetic and clinical variables not considered by traditional fracture risk tools.

The SL model's ensemble learning approach has been applied across various domains for improved predictive performance [32, 42, 43]. Our study pioneers the use of this approach in osteoporotic fracture prediction, integrating structured medical records and genomic data. The SL model, based on cross-validation, combines ML algorithms to yield predictions comparable to the best input algorithm [32, 42]. Additionally, its accurate fracture prediction capability holds promise for clinical practice, facilitating early risk identification and targeted preventive interventions, thus mitigating the burden of fractures.

Several studies have explored ensemble learning for fracture prediction using various ML models. Liu et al. [44] employed ensemble learning with artificial neural networks, achieving a remarkable 93% accuracy in predicting hip fractures using clinical data from 228 patients. Chou et al. [45] developed an ensemble method for vertebral fracture prediction, achieving a notable 92% accuracy in a cohort of 7299 patients aged 60. Similarly, Li et al. [46] utilized ensemble learning with convolutional neural networks, reporting a 93% accuracy for predicting fractures in patients older than 60. However, these studies did not integrate genomic data. Our study advances this field by integrating genomic information into ensemble learning, achieving a record accuracy of 95.6%. This breakthrough highlights the potential of ensemble learning in leveraging diverse data sources and advancing personalized medicine by discerning fracture risks based on genetic predispositions.

Our study underscores the SL model's promise in predicting fractures yet acknowledges limitations like its exclusive focus on older men and reliance on additive genetic effects. Future research should diversify the participant demographic and delve into various genetic effect models, enhancing the model's clinical applicability. Critical to this endeavor is external validation across diverse populations to ensure the model's reliability and generalizability in different clinical settings.

In conclusion, our study highlights the effectiveness of the SL ensemble model for fracture prediction, showcasing superior performance through integrating phenotypic and genomic information. This approach not only enhances adaptability and robustness but also holds significant promise for clinical applications, facilitating the timely identification of high-risk individuals and implementing preventive measures.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00198-025-07437-w.

Acknowledgements We sincerely thank the original MrOS study investigators and the invaluable participants for their pivotal contributions to advancing men's health research. We also express our gratitude to the National Institutes of Health (NIH) and the dbGaP for granting access to analyze the MrOS data. This work reflects our independent analysis and interpretation and does not represent the views of other parties associated with the MrOS study. We sincerely appreciate the collective efforts and contributions of all institutions, collaborators, and teams involved in the MrOS study.

Funding The research and analysis described in the current publication were supported by a grant (R21MD013681 to QW) from the National Institute on Minority Health and Health Disparities and a grant (R01AG080017 to QW) from the National Institute on Aging. The funders had no role in study design, data collection and analysis, publication decisions, or manuscript preparation. Funding support for the original MrOS study was provided by the National Institutes of Health, including the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS); the National Institute on Aging (NIA); the National Center for Research Resources (NCRR); and the National Heart, Lung, and Blood Institute (NHLBI).

Data availability The data/analyses presented in the current publication are based on study data downloaded from the database of Genotypes and Phenotypes (dbGaP) website under phs000373.v1.p1.

Declarations

Conflicts of interest None.

References

1. Nguyen ND, Ahlborg HG, Center JR, Eisman JA, Nguyen TV (2007) Residual lifetime risk of fractures in women and men. J Bone Miner Res 22:781–788
2. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E (2008) FRAX™ and the assessment of fracture probability in men and women from the UK. Osteoporos Int 19:385–397
3. Hippisley-Cox J, Coupland C (2009) Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. The BMJ 339:b4229
4. Nguyen ND, Frost SA, Center JR, Eisman JA, Nguyen TV (2008) Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks. Osteoporos Int 19:1431–1444
5. Force UPST, Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB et al (2018) Screening for osteoporosis to prevent fractures: US Preventive Services Task Force recommendation statement. JAMA 319:2521–2531
6. Trajanoska K, Morris JA, Oei L, Zheng HF, Evans DM, Kiel DP et al (2018) Assessment of the genetic and clinical determinants of fracture risk: genome wide association and Mendelian randomisation study. The BMJ 362:1–14

7. Holloway-Kew KL, Zhang Y, Betson AG, Anderson KB, Hans D, Hyde NK et al (2019) How well do the FRAX (Australia) and Garvan calculators predict incident fractures? Data from the Geelong Osteoporosis Study. Osteoporos Int 30:2129–2139
8. Silverman SL, Calderon AD (2010) The utility and limitations of FRAX: a US perspective. Curr Osteoporos Rep 8:192–197
9. Kanis JA, Harvey NC, Johansson H, Odén A, McCloskey EV, Leslie WD (2017) Overview of fracture prediction tools. J Clin Densitom 20:444–450
10. Sivasakthi B, Selvanayagi D (2020) A comparison of machine learning algorithms for osteoporosis prediction. 2022 First Int Conf Electr Electron Inf Commun Technol ICEEICT [Internet]. [cited 2024 Feb 12]. p. 1–6. Available from: https://ieeexplore.ieee.org/abstract/document/9768568
11. Shim J-G, Kim DW, Ryu K-H, Cho E-A, Ahn J-H, Kim J-I et al (2020) Application of machine learning approaches for osteoporosis risk prediction in postmenopausal women. Arch Osteoporos 15:169
12. Inui A, Nishimoto H, Mifune Y, Yoshikawa T, Shinohara I, Furukawa T et al (2023) Screening for osteoporosis from blood test data in elderly women using a machine learning approach. Bioengineering 10:277
13. Wu Q, Nasoz F, Jung J, Bhattarai B, Han MV (2020) Machine learning approaches for fracture risk assessment: a comparative analysis of genomic and phenotypic data in 5130 older men. Calcif Tissue Int 107:353–61
14. Wu Q, Nasoz F, Jung J, Bhattarai B, Han MV, Greenes RA et al (2021) Machine learning approaches for the prediction bone mineral density by using genomic and phenotypic data of 5,130 older men. Sci Rep 11:4482
15. Orwoll E, Blank JB, Barrett-Connor E, Cauley J, Cummings S, Ensrud K et al (2005) Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study - a large observational study of the determinants of fracture in older men. Contemp Clin Trials 26:569–585
16. Blank JB, Cawthon PM, Carrion-Petersen ML, Harper L, Johnson JP, Mitson E et al (2005) Overview of recruitment for the osteoporotic fractures in men study (MrOS). Contemp Clin Trials 26:557–568
17. Orwoll E, Blank JB, Barrett-Connor E, Cauley J, Cummings S, Ensrud K et al (2005) Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — a large observational study of the determinants of fracture in older men. Contemp Clin Trials 26:569–585
18. Pincus T, Summey JA, Soraci JRSA, Wallston KA, Hummon NP (1983) Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 26:1346–1353
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
20. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A et al (2016) Next-generation genotype imputation service and methods. Nat Genet 48:1284–1287
21. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK et al (2016) Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48:1443–1448
22. Durbin R (2014) Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics 30:1266–1272
23. Morris JA, Kemp JP, Youlten SE, Laurent L, Logan JG, Chai RC et al (2019) An atlas of genetic influences on osteoporosis in humans and mice. Nat Genet 51:258–266
24. Weir BS, Anderson AD, Hepler AB (2006) Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet 710(7):771–80
25. Conomos MP, Reiner AP, Weir BS, Thornton TA (2016) Model-free estimation of recent genetic relatedness. Am J Hum Genet 98:127–148
26. Truong VQ, Woerner JA, Cherlin TA, Bradford Y, Lucas AM, Okeh CC et al (2022) Quality control procedures for genome-wide association studies. Curr Protoc 2:e603
27. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
28. Murphy KP (2003) Naive Bayes classifiers. Univ Br Columbia 18:1–8
29. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
30. Peterson LE (2009) K-nearest neighbor. Scholarpedia 4:1883
31. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
32. MJ V der L, EC P, AE H (2007) Super Learner. Stat Appl Genet Mol Biol 6:1–23
33. Acuña E, Rodriguez C (2004) The treatment of missing values and its effect on classifier accuracy. In: Banks D, McMorris FR, Arabie P, Gaul W (eds) Classif Clust Data Min Appl. Springer, Berlin, Heidelberg, pp 639–647
34. Melton LJ III, Atkinson EJ, O'Fallon WM, Wahner HW, Riggs BL (1993) Long-term fracture prediction by bone mineral assessed at different skeletal sites. J Bone Miner Res 8:1227–1233
35. Kanis JA, Borgstrom F, De Laet C, Johansson H, Johnell O, Jonsson B et al (2005) Assessment of fracture risk. Osteoporos Int 16:581–589
36. Stone KL, Seeley DG, Lui L-Y, Cauley JA, Ensrud K, Browner WS et al (2003) BMD at multiple sites and risk of fracture of multiple types: long-term results from the Study of Osteoporotic Fractures. J Bone Miner Res 18:1947–1954
37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
38. Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–30
40. Ettinger B, Ensrud KE, Blackwell T, Curtis JR, Lapidus JA, Orwoll ES (2013) Performance of FRAX in a cohort of community-dwelling, ambulatory older men: the Osteoporotic Fractures in Men (MrOS) study. Osteoporos Int 24:1185–1193
41. Harvey NC, Odén A, Orwoll E, Lapidus J, Kwok T, Karlsson MK et al (2018) Falls predict fractures independently of FRAX probability: a meta-analysis of the Osteoporotic Fractures in Men (MrOS) study. J Bone Miner Res 33:510–516
42. Polley EC, van der Mark JL (2010) Super learner in prediction. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 266. https://biostats.bepress.com/ucbbiostat/paper266. Accessed 14 June 2023
43. Ju C, Combs M, Lendle SD, Franklin JM, Wyss R, Schneeweiss S et al (2019) Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. J Appl Stat 46:2216–2236
44. Liu Q, Cui X, Chou YC, Abbod MF, Lin J, Shieh JS (2015) Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders. Biomed Signal Process Control 21:146–156
45. Chou PH, Jou TH, Wu HH, Yao YC, Lin HH, Chang MC et al (2022) Ground truth generalizability affects performance of the artificial intelligence model in automated vertebral fracture

detection on plain lateral radiographs of the spine. Spine J 22:511–523
46. Li YC, Chen HH, Horng-Shing LuH, Hondar Wu HT, Chang MC, Chou PH (2021) Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists? Clin Orthop Relat Res 479:1598–1612