Bone Reports 22 (2024) 101805

Contents lists available at ScienceDirect

Bone Reports

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

A staged approach using machine learning and uncertainty quantification to predict the risk of hip fracture

Anjum Shaik[a,1], Kristoffer Larsen[b,1], Nancy E. Lane[c], Chen Zhao[d,*], Kuan-Jui Su[e], Joyce H. Keyak[f], Qing Tian[e], Qiuying Sha[b], Hui Shen[e], Hong-Wen Deng[e], Weihua Zhou[a,g,**]

[a] *Department of Applied Computing, Michigan Technological University, 1400 Townsend Dr, Houghton, MI, 49931, United States of America*

[b] *Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA*

[c] *Department of Internal Medicine and Division of Rheumatology, UC Davis Health, Sacramento, CA 95817, USA*

[d] *Department of Computer Science, Kennesaw State University, 680 Arntson Dr, Marietta, GA 30060, USA*

[e] *Division of Biomedical Informatics and Genomics, Tulane Center of Biomedical Informatics and Genomics, Deming Department of Medicine, Tulane University, New Orleans, LA 70112, USA*

[f] *Department of Radiological Sciences, Department of Biomedical Engineering, and Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA*

[g] *Center for Biocomputing and Digital Health, Institute of Computing and Cybersystems, and Health Research Institute, Michigan Technological University, Houghton, MI 49931, USA*

ARTICLE INFO

Keywords:
Hip fracture
Dual-energy X-ray absorptiometry
Bone mineral density
Machine learning
Uncertainty quantification

ABSTRACT

Hip fractures present a significant healthcare challenge, especially within aging populations, where they are often caused by falls. These fractures lead to substantial morbidity and mortality, emphasizing the need for timely surgical intervention. Despite advancements in medical care, hip fractures impose a significant burden on individuals and healthcare systems. This paper focuses on the prediction of hip fracture risk in older and middle-aged adults, where falls and compromised bone quality are predominant factors.

The study cohort included 547 patients, with 94 experiencing hip fracture. To assess the risk of hip fracture, clinical variables and clinical variables combined with hip DXA imaging features were evaluated as predictors, followed by a novel staged approach. Hip DXA imaging features included those extracted by convolutional neural networks (CNNs), shape measurements, and texture features. Two ensemble machine learning models were evaluated: Ensemble 1 (clinical variables only) and Ensemble 2 (clinical variables and imaging features) using the logistic regression as the base classifier and bootstrapping for ensemble learning. The staged approach was developed using uncertainty quantification from Ensemble 1 which was used to decide if hip DXA imaging features were necessary to improve prediction for each subject. Ensemble 2 exhibited the highest performance, achieving an Area Under the Curve (AUC) of 0.95, an accuracy of 0.92, a sensitivity of 0.81, and a specificity of 0.94. The staged model also performed well, with an AUC of 0.85, an accuracy of 0.86, a sensitivity of 0.56, and a specificity of 0.92, outperforming Ensemble 1, which had an AUC of 0.55, an accuracy of 0.73, a sensitivity of 0.20, and a specificity of 0.83. Furthermore, the staged model suggested that 54.49 % of patients did not require DXA scanning, effectively balancing accuracy and specificity, while offering a robust solution when DXA data acquisition is not feasible. Statistical tests confirmed significant differences between the models, highlighting the advantages of advanced modeling strategies.

Our staged approach offers a cost-effective holistic view of patient health. It can identify individuals at risk of hip fracture with a high accuracy while reducing unnecessary DXA scans. This approach has great promise to guide the need for interventions to prevent hip fracture while reducing diagnostic cost and exposure to radiation.

Abbreviations: DXA, Dual-energy X-ray Absorptiometry; BMC, Bone Mineral Content; BMD, Bone Mineral Density; aBMD, areal BMD; ML, Machine Learning; CNN, Convolutional Neural Network; AUC, Area Under the Receiver Operating Characteristic Curve; CI, Confidence Interval; RFE, Recursive Feature Elimination.

* Correspondence to: C. Zhao, Department of Computer Science, Kennesaw State University, 680 Arntson Dr, Marietta, GA 30060, USA

** Correspondence to: W. Zhou, Department of Applied Computing, Michigan Technological University, 1400 Townsend Dr, Houghton, MI 49931, USA.

E-mail addresses: czhao4@kennesaw.edu (C. Zhao), whzhou@mtu.edu (W. Zhou).

[1] Anjum Shaik and Kristoffer Larsen contribute equally.

https://doi.org/10.1016/j.bonr.2024.101805
Received 30 June 2024; Received in revised form 20 August 2024; Accepted 11 September 2024
Available online 12 September 2024

1. Introduction

Hip fractures, which are often precipitated by falls, present a significant healthcare challenge, particularly among aging populations. With the global aging trend, the incidence of hip fracture is expected to rise dramatically in the coming decades. For instance, while the annual global incidence was 1.3 million in 1990, it is projected to surge to a staggering 7 to 21 million by 2050 (Gullberg et al., 1997). In the United States alone, the annual incidence per 100,000 individuals ranges between 197 and 201 for men and 511 to 553 for women, with rates increasing significantly with age (Dhanwal et al., 2011). These incidents have serious consequences on quality of life. Apart from causing morbidity and mortality, hip fractures impose a substantial economic burden. Patients often face approximately $40,000 direct medical cost within the first-year post-fracture, while the collective annual cost in the US alone surpasses $17 billion (Emmerson et al., 2024). The cost of hip fracture can be reduced by identifying patients at risk of hip fracture so they can be treated to reduce this risk.

Bone mineral density (BMD) is a key determinant of hip fracture risk. Dual-energy X-ray absorptiometry (DXA) plays a pivotal role in assessing areal BMD (aBMD) and fracture risk. DXA serves as the standard imaging modality guiding clinical decisions for the assessment of osteoporosis and hip fracture risk, initiation of treatment and follow-up of individuals at risk.

Recent studies have explored innovative approaches, such as artificial intelligence (AI) and machine learning (ML), to enhance the accuracy of hip fracture risk prediction by leveraging DXA imaging alongside clinical data. This finding underscores the transformative impact of AI and ML technologies in augmenting the capabilities of healthcare professionals and improving patient outcomes in orthopedic care. Lex et al. (Lex et al., 2023) conducted a thorough investigation into the diagnostic accuracy of models in diagnosing hip fractures on radiographs and predicting postoperative clinical outcomes following hip fracture surgery relative to current practices. Their systematic review and meta-analysis of 39 studies revealed that AI models perform comparably to expert clinicians in diagnosing hip fractures. Cha et al. (Cha et al., 2022) systematically reviewed the use of AI and in diagnosing and classifying hip fractures, demonstrating high accuracy and effectiveness in clinical settings. Furthermore, Murphy et al. (Murphy et al., 2022) utilized two sets of radiographs: one from population without hip fractures collected as part of a bone mass study and another from those who had hip fractures from local National Hip Fracture Database audit records. Their study demonstrated that a trained neural network exhibits a remarkable 19 % increase in accuracy in classifying hip fractures compared to experienced human observers within clinical settings. Zhao et al. (Zhao et al., 2023) introduced multi-view variational autoencoder and product of expert models for predicting proximal femoral fracture loads, which are inversely associated with incident hip fracture, by integrating whole-genome sequence features and DXA-derived imaging features. Additionally, Hong et al. (Hong et al., 2021) developed a bone radiomics score using a random forest model and texture analysis of DXA hip images for predicting incident hip fractures.

Although these findings underscore the transformative impact of AI and ML technologies in diagnosing hip fractures, the problem of identifying patients at risk of hip fracture so fractures can be prevented still remains and current ML and AI approaches for predicting hip fractures have notable limitations. Some often utilize only a single modality of data, either clinical or imaging, which can lead to limited predictive accuracy. Moreover, most multi-modality ML methods require all modalities to be obtained in advance for effective prediction, increasing the cost, radiation and complexity of the diagnostic process.

To address this important limitation, we introduce a novel staged approach for predicting hip fracture which would be more efficient and economical. Unlike current methods, our approach is structured into two distinct stages. In the first stage, we focus solely on clinical characteristics. We then use uncertainty quantification of the first stage to determine if proceeding to subsequent diagnostics in the second stage is necessary. If so, this second stage expands our analysis to incorporate imaging features extracted from hip DXA images. By integrating both clinical and imaging data with ML in this second stage, which is based on uncertainty quantification in the first stage, this staged approach aims to improve prediction accuracy and adaptability to diverse clinical scenarios.

2. Materials and methods

Our study employed a staged approach to improve prediction accuracy and adaptability in predicting hip fracture risk. By integrating clinical and imaging data using ML, we aim to optimize model performance while minimizing clinical costs and procedural time. Inspired by the sequential decision-making processes commonly employed in clinical practice, our methodology incorporated advanced techniques to optimize model performance (Fig. 1).

2.1. Cohort description

The dataset utilized in this study was sourced from the UK Biobank (application ID: 61915), representing a valuable resource for investigating bone health parameters. DXA imaging, essential for evaluating BMD and morphology, was performed by trained radiographers using the GE-Lunar iDXA instrument. Regular calibration of this instrument to a manufacturer's phantom (GE-Lunar, Madison, WI) and daily quality control procedures ensured the accuracy and reliability of DXA measurements (Resource 502, 2024). This comprehensive DXA dataset covers various anatomical regions, including the whole body, lateral thoraco-lumbar spine, and bilateral hips and knees. For this study, we focused on a subset of this cohort, comprising 547 patients with DXA hip images. In addition to including DXA measurements at various sites, we focused on analysis of the DXA images of the hip. Notably, among the subset of patients with DXA hip images, 94 individuals experienced hip fractures (with 40 males) after the DXA imaging and clinical data collection, thereby representing prospective fracture events. The majority (n = 453) were non-fractured individuals (with 226 males). All individuals in our sample were of British ethnicity.

2.2. Clinical factors

Our study considered a plethora of variables crucial for understanding various aspects of participants' health profiles. These variables encompassed demographic details such as age at recruitment, sex, and genetic sex, and ethnicity. The ethnicity of all patients in our sample is British, emphasizing the homogeneity of ethnic background within our study population. Anthropometric measurements like weight were also included. Additionally, information regarding participants' average total household income before tax, and lifestyle factors including smoking and alcohol consumption statuses were considered (Supp. Table 1). We also examined dietary habits, including variations in diet and major dietary changes in the last 5 years, along with the occurrence of falls and bone fractures in the past year and 5 years, respectively. Notably, a small proportion (about 5 %) of individuals with bone fractures experienced hip fractures, underscoring the diversity in fracture types within our dataset. The regular intake of vitamin and mineral supplements was also documented. This comprehensive array of clinical data facilitated a thorough exploration of factors influencing participants' health and enabled meaningful insights into bone health and related risk factors.

2.3. DXA imaging feature analysis

In our study focused on predicting hip fracture risk, we manually annotated the left and right contours of the femur to isolate our region of interest from the raw DXA images. The raw DXA images were extracted directly from the DXA scans conducted at the UK Biobank Imaging Assessment Centre, following the procedures in their protocol (Resource 502, 2024). The images were exported as DICOM files, ensuring that they were in their original format and not derived from secondary sources such as PDF reports. To maintain measurement accuracy and consistency, these images were calibrated to bone mineral density (BMD) using the manufacturer's phantom, which is part of the standard quality control protocol. This calibration process, performed daily, ensures that all data accurately reflects the true BMD values. To ensure consistency across the dataset, DXA images were standardized to a size of 224 × 224 pixels.

Imaging feature extraction was performed using two pre-trained CNN models: VGG16 (Simonyan et al., 2024) and Xception (Chollet, 2017). These CNN models extracted rich feature representations from the preprocessed DXA images, capturing both global and fine-grained details crucial for accurate prediction. Alongside CNN-based feature extraction, 2D shape measurements and texture features from the DXA images were computed using specialized packages. Specifically, the shape measurements were computed using the IMEA package (Kroell, 2021), which assessed the 2D geometric characteristics of the femur region. Similarly, texture features were extracted using the PyRadiomics package (van Griethuysen et al., 2017), enabling the capture of detailed textural information from the DXA images. Femur parameters from medical reports of DXA images (Supp. Table 2) were also included as DXA imaging features in this study; they provided intricate measurements of BMD and bone mineral content (BMC) at various anatomical sites, shedding light on participants' overall bone health.

2.4. *Ensemble 1: Hip fracture prediction using clinical factors*

We first employed logistic regression as the base classifier and used a bootstrapping strategy to perform ensemble learning for hip fracture prediction using clinical factors. The developed model was named Ensemble 1. To mitigate dimensionality and enhance model interpretability, feature selection techniques such as univariate feature selection via near zero variance filtering and correlation filtering were employed.

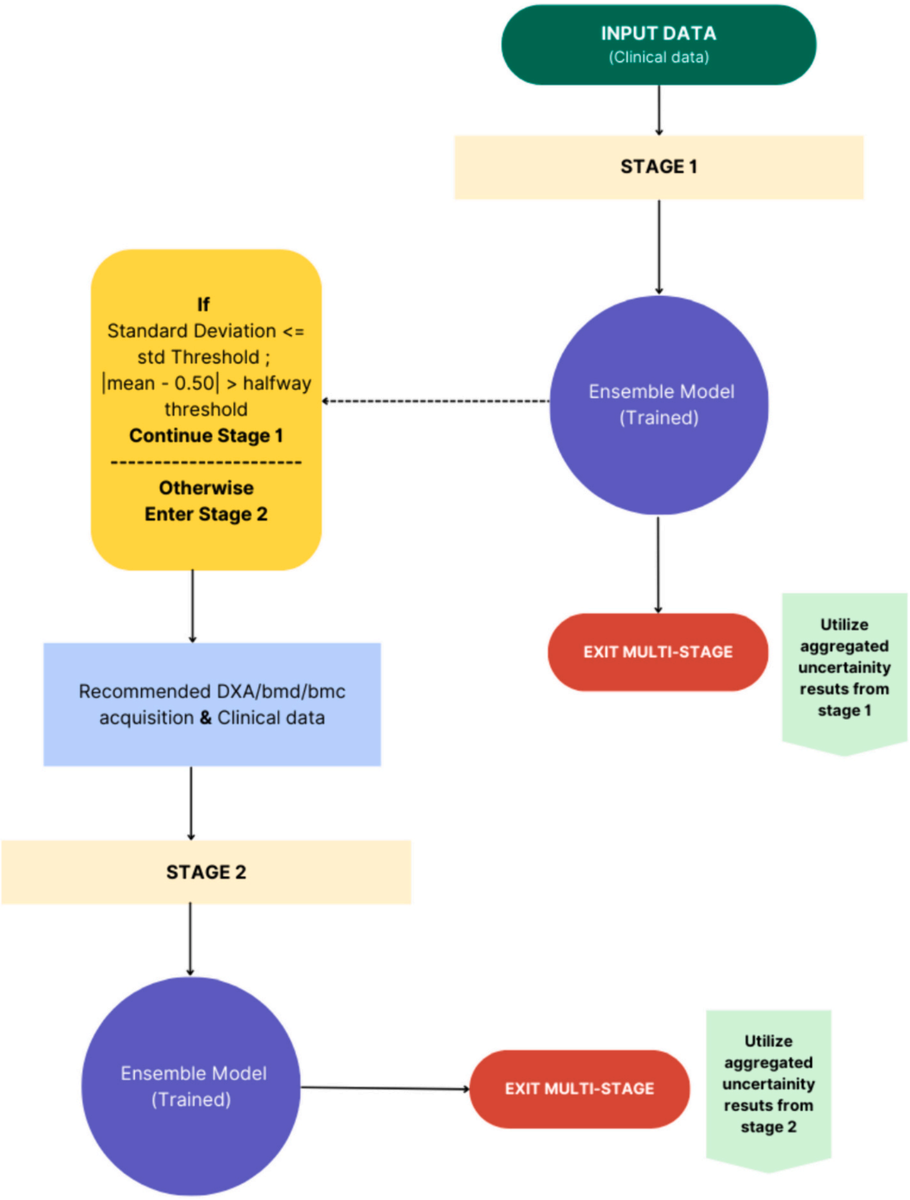


Fig. 1. The staged approach for hip fracture prediction, employing internal logic rules where the model progresses to Stage I if the standard deviation is less than or equal to the specified threshold and the absolute difference between the mean values of hip fracture probabilities and 0.50 exceeds the halfway threshold; otherwise, Stage II is initiated for comprehensive risk evaluation.

Furthermore, Recursive Feature Elimination (RFE) (Guyon et al., 2002) was used to identify the most relevant features in a multivariate fashion. These methods identified the most relevant features for predicting the target variable, ensuring that only informative features were retained for analysis. Each ensemble model was trained on a resampled subset of the data through stratified cross-validation, promoting robustness, and capturing variations within the dataset.

2.5. *Ensemble 2: Hip fracture prediction using clinical factors and DXA images*

We integrated DXA-derived imaging features with clinical factors to build a hip fracture prediction model named Ensemble 2. Similar to Ensemble 1, logistic regression was employed as the base classifier and a bootstrapping strategy was employed to train the hip fracture prediction model. The DXA derived imaging features included the extracted features from the CNN models, shape measurements and texture features.

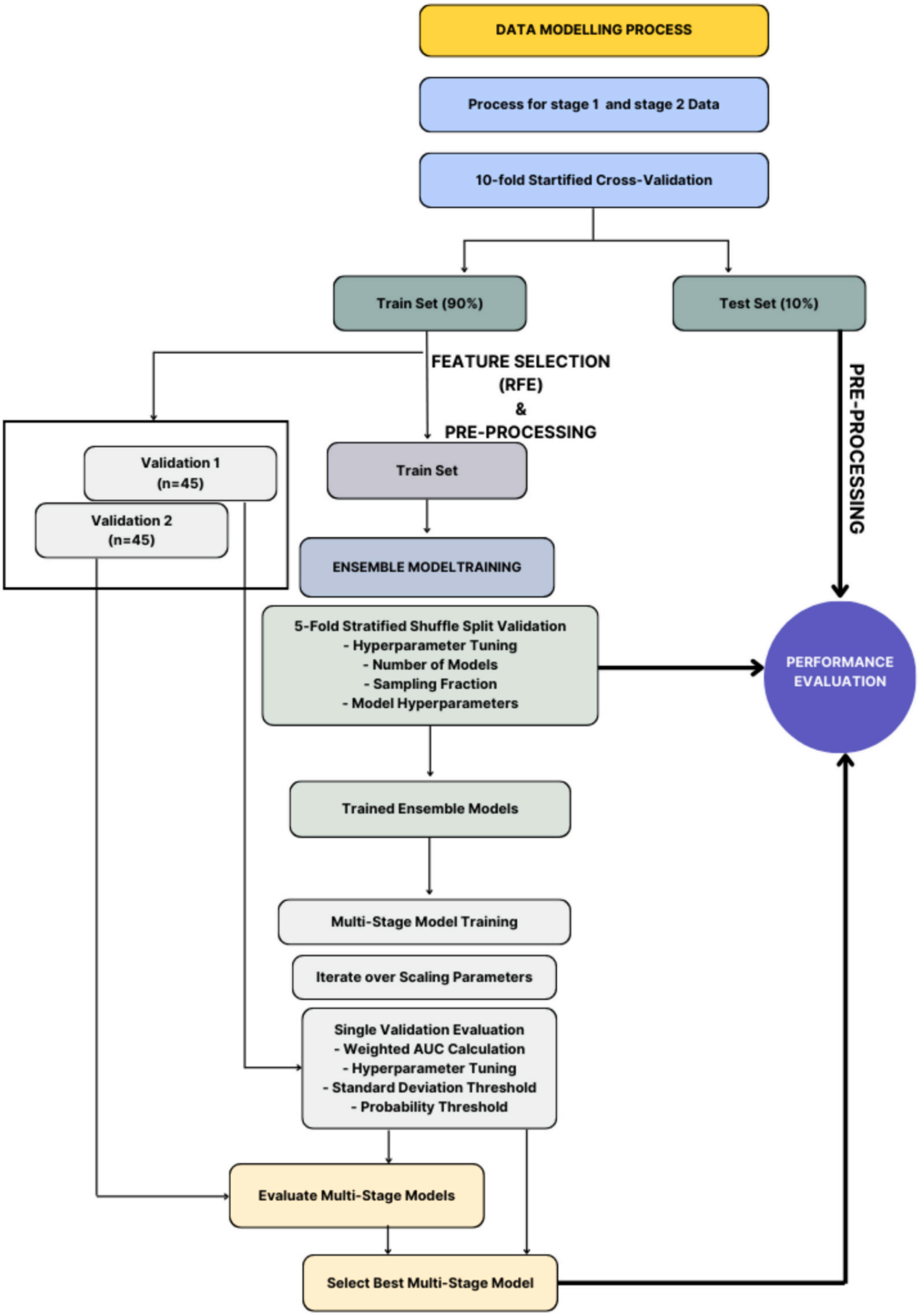

DATA MODELLING PROCESS
Process for stage 1 and stage 2 Data
10-fold Startified Cross-Validation
Train Set (90%)
Test Set (10%)
FEATURE SELECTION (RFE) & PRE-PROCESSING
PRE-PROCESSING
Validation 1 (n=45)
Validation 2 (n=45)
Train Set
ENSEMBLE MODELTRAINING
5-Fold Stratified Shuffle Split Validation
- Hyperparameter Tuning
- Number of Models
- Sampling Fraction
- Model Hyperparameters
PERFORMANCE EVALUATION
Trained Ensemble Models
Multi-Stage Model Training
Iterate over Scaling Parameters
Single Validation Evaluation
- Weighted AUC Calculation
- Hyperparameter Tuning
- Standard Deviation Threshold
- Probability Threshold
Evaluate Multi-Stage Models
Select Best Multi-Stage Model

Fig. 2. The modeling process, encompassing feature extraction, selection, and ensemble techniques, to optimize predictive performance using clinical and imaging data, followed by evaluation and validation of the resulting model.

This integrated feature set provided a holistic representation of both anatomical and clinical aspects relevant to hip fracture prediction. Inner cross-validation was conducted on the training data to optimize the hyperparameters of the base models, such as the number of base models comprised in the Ensemble 2 model, the percentage of random samples for training each base model, and the respective base models' hyperparameters.

2.6. Staged method and uncertainty analysis

To minimize clinical costs and DXA imaging exposure, a sequential, "staged" model was constructed to integrate predictions from Ensemble 1 (Stage I) and Ensemble 2 (Stage II). This sequential model leverages the strengths of the Ensembles 1 and 2, which are each comprised of multiple collections of sub-models. Our central hypothesis is that if the Ensemble 1 model can predict hip fracture for a subject with high confidence, then the subsequent stage, Ensemble 2, which uses both clinical factors and DXA imaging screening, is no longer needed. The confidence in predicting hip fracture would be based on the calculated uncertainty. If the confidence in the results from Ensemble 1 is above an established threshold (uncertainty is low), pursuing analysis via Ensemble 2 would be unnecessary, clinical costs would be reduced and DXA imaging screening would be avoided. However, if the uncertainty of hip fracture prediction is smaller than the threshold, the Ensemble 2 model would be employed to perform hip fracture risk assessment.

In detail, by training 50 individual logistic regression classifiers on different bootstrapped samples of our training dataset, we created a robust ensemble capable of delivering accurate hip fracture risk prediction. For each subject in our test dataset, we generated predictions from all 50 individual logistic regression classifiers from Ensemble 1, and then calculated the mean and standard deviation of these predictions. The mean provided our final predicted risk, while the standard deviation served as a measure of uncertainty. In addition, if the standard deviation is less than or equal to a standard deviation threshold, indicating low variability and high reliability of the predictions, and the mean prediction deviates significantly from 0.50 beyond a set halfway threshold (midway threshold in Fig. 1), the predicted risk in Stage I is accepted. This means the prediction is deemed reliable and no DXA testing is required. Conversely, if either condition is not met, indicating higher uncertainty or insufficient confidence in the prediction, the process moves to Stage II, necessitating additional DXA testing to ensure accuracy. This approach optimizes the diagnostic process by balancing prediction reliability with the need for further stages.

For model training, nested cross-validation structure was employed (Fig. 2). Initially, the dataset was divided into an outer training fold comprising 492 samples and an outer test fold with 55 samples. From the training fold, two separate validation sets were extracted, each containing 45 samples. To preprocess the data and mitigate outliers, centering/scaling and spatial sign transformations were applied. Next, two validation sets were sliced from the original training fold to fine-tune hyperparameters for the multi-stage bridge between the Ensemble models, which include standard deviation thresholds and midway thresholds that govern the transition from Stage I to Stage II, i.e., whether evaluation of the patient will require DXA images for an evaluation using Ensemble 2. These two thresholds are optimized to achieve a balanced trade-off between predictions retained from Ensemble 1 and those cascading into Ensemble 2, using a scaled weighted Area Under the Curve (AUC) metric. Finally, the best-performing staged model is identified using the outer test set to ensure its robustness and generalization.

2.7. Statistical analysis

In this study, comprehensive statistical analyses were performed to evaluate model performance and feature associations with hip fracture risk. The DeLong test was used to compare AUC curves and McNemar's test assessed sensitivity and specificity variations. Chi-square test and Fisher's exact tests revealed the associations between categorical variables and fracture risk, while *t*-tests highlighted differences in continuous variables between fracture groups.

3. Results

The study cohort comprised 547 patients, with 94 individuals who experienced hip fractures. An initial assessment revealed that 54.49 % of the patients did not require analysis via Stage II, i.e. Ensemble 2, which involved DXA scanning, while 45.52 % did. Patients who did not require DXA scanning lacked significant risk factors such as younger age, no history of prior fractures, absence of clinical risk factors for osteoporosis (e.g., history of smoking, excessive alcohol consumption), and initial clinical assessments indicating low risk. The distribution of patients not requiring DXA was characterized by the following percentiles: 25th percentile at 35.45 %, 50th percentile at 46.78 %, and 75th percentile at 59.55 %.

Table 1 summarizes the performance metrics of the models employed in this study. Notably, Ensemble 2 emerged as the frontrunner with the highest AUC of 0.95 (95 % CI: 0.87–1.00), followed closely by the staged model at 0.85 (95 % CI: 0.78–0.92). Ensemble 1 exhibited a comparatively lower AUC of 0.70 (95 % CI: 0.55–0.85). These findings underline the superior predictive performance of Ensemble 2 and the staged model in fracture risk assessment. Diving deeper into accuracy and specificity, the staged model showcased superior performance, with accuracy reaching 86.11 % and specificity peaking at 92.49 %. Additionally, fracture risk assessment tool (FRAX) with aBMD (left hip) and FRAX without aBMD yielded AUC scores of 0.76 and 0.62, respectively, lower than those for the Ensemble 2 and the Staged models, highlighting the marked improvements of our ML approaches over FRAX models for fracture risk assessment.

Our analysis utilized rigorous statistical testing, DeLong tests and McNemar's sensitivity and specificity test, to reveal significant differences between the models. Confidence intervals (CIs) for the DeLong tests were computed, indicating AUC 95 % CIs for the staged model of AUC = 0.81–0.89, Ensemble 1 of 0.49–0.61, Ensemble 2 of 0.94–0.98, FRAX with aBMD (left) of 0.70–0.81, and FRAX without aBMD of 0.55–0.69. Additionally, the DeLong tests yielded *p*-values, indicating the significance of the differences in AUC between different model pairs: Staged vs. Ensemble 1 $p < 0.001$, and Staged vs. FRAX with BMD (left) $p = 0.004$. McNemar's sensitivity and specificity test also provided insights, with p-values indicating the significance of differences in sensitivity and specificity between model pairs, such as between Staged and Boot1 (sensitivity: <0.0001, specificity: <0.0001).

Additionally, the study identified significant associations between categorical variables (Table 2) like alcohol consumption and average household income with fracture risk, as well as notable differences in continuous variables such as age and various BMD measurements among patient groups (Table 3). Baseline statistics including *p*-values from chi-square, Fisher, and *t*-tests for categorical (Table 4) and continuous (Table 5) variables were also calculated. These findings underscore the potential of ensemble learning and staged modeling in enhancing hip-fracture risk assessment, offering insights for clinical decision-making and preventive strategies.

To visually encapsulate the findings, AUCs of the ensemble stage I (Fig. 3A), ensemble stage II (Fig. 3B) and staged (Fig. 3C) models are presented. Ensemble 2 emerged as the standout performer, consistently surpassing its counterparts. However, no significant disparities were observed between the staged model and either Ensemble 1 or 2, underscoring the robustness of the staged approach and its effectiveness in reducing medical costs. Figs. 4A and 4B further enrich our understanding by highlighting the importance of various features. Ensemble models underscored age, weight, and dietary changes as significant predictors (Fig. 4A). Conversely, Ensemble 2 prioritized DXA parameters, such as convex area and projection area, accentuating their role in

Table 1
The AVG performance metrics, such as AUC, accuracy, sensitivity, and specificity, for the various models. It includes STD values to indicate metric variability across evaluations. **AVG**: Average, STD: Standard Deviation.

	AUC	Accuracy	Sensitivity	Specificity
Ensemble Model 1	0.5548 ± 0.1367	0.7239 ± 0.0645	0.1956 ± 0.1486	0.9342 ± 0.0511
Ensemble Model 2	0.9541 ± 0.0358	0.9195 ± 0.0274	0.8078 ± 0.1336	0.9427 ± 0.0224
STAGED Model	0.8486 ± 0.0918	0.8611 ± 0.0463	0.5578 ± 0.2338	0.9249 ± 0.0394

Table 2
Baseline statistics of categorical features related to hip fracture. It compares the information on different factors related to hip fractures between individuals who experienced hip fractures (“Hip Fracture (Yes)”) and those who didn’t (“Hip Fracture (No)”). Each row represents a specific feature mentioned in the study. The numbers in the table represent percentages and counts within each group.

Feature	Hip fracture (No)	Hip fracture (Yes)
Alcohol consumption (Never Consumed)	1 (0.2 %)	4 (4.3 %)
Alcohol consumption (previous)	3 (0.7 %)	0 (0 %)
Alcohol consumption (current)	449 (99.1 %)	90 (95.7 %)
Average Household Income(Do not Know)	11 (2.4 %)	2 (2.1 %)
Average Household Income (Prefer not to answer)	24 (5.3 %)	7 (7.4 %)
Average Household Income (<18,000£)	63 (13.9 %)	23 (24.5 %)
Average Household Income (18,000£ to 30,999£)	130 (28.7 %)	22 (23.4 %)
Average Household Income (31,000£ to 51,999£)	130 (28.7 %)	20 (21.3 %)
Average Household Income (52,000£ to 100,000£)	76 (16.8 %)	17 (18.1 %)
Average Household Income (>100,000£)	19 (4.2 %)	3 (3.2 %)
Variation in diet (Never/Rarely)	153 (33.8 %)	34 (36.2 %)
Variation in diet (Sometimes)	258 (57.0 %)	51 (54.3 %)
Variation in diet (Often)	42 (9.3 %)	9 (9.6 %)
Falls in last year (Prefer not to answer)	1 (0.2 %)	1 (1.1 %)
Falls in last year (None)	356 (78.6 %)	74 (78.7 %)
Falls in last year (Only one)	65 (14.3 %)	13 (13.8 %)
Falls in last year (More than one)	31 (6.8 %)	6 (6.4 %)
Fracture/broken bones in last 5 years (Do not know)	1 (0.2 %)	0 (0 %)
Fracture/broken bones in last 5 years (Prefer to know)	1 (0.2 %)	0 (0 %)
Fracture/broken bones in last 5 years (None)	420 (92.7 %)	86 (91.5 %)
Fracture/broken bones in last 5 years (yes)	31 (6.8 %)	8 (8.5 %)
Genetic sex (female)	227 (50.1 %)	54 (57.4 %)
Genetic sex (male)	226 (49.9 %)	40 (42.6 %)
Major change in diet IN LAST 5 YEARS(Not due to illness)	283 (62.5 %)	65 (69.1 %)
Major change in diet IN LAST 5 YEARS (due to illness)	29 (6.4 %)	7 (7.4 %)
Major change in diet IN LAST 5 YEARS (other reasons)	141 (31.1 %)	22 (23.4 %)
Sex (female)	227 (50.1 %)	54 (57.4 %)
Sex (male)	226 (49.9 %)	40 (42.6 %)
Smoking (Prefer not to answer)	3 (0.7 %)	0 (0 %)
Smoking (Never)	247 (54.5 %)	52 (55.3 %)
Smoking (Previous)	187 (41.3 %)	36 (38.3 %)
Smoking (Current Smoker)	16 (3.5 %)	6 (6.4 %)
Vitamin Supplement (None)	377 (83.2 %)	79 (84.0 %)
Vitamin Supplement (Yes)	76 (16.8 %)	15 (16.0 %)

Table 3
Baseline statistics of continuous features related to hip fracture. It compares measurements and statistics between two groups: individuals who experienced hip fractures and those who did not. Each row represents a specific feature. The columns show the average value (mean) and the variation (standard deviation) within each group.

Feature	Mean		Standard deviation	
	Hip fracture (No)	Hip fracture (Yes)	Hip fracture (No)	Hip fracture (Yes)
Age	57.70	59.70	7.05	7.51
Femur neck BMC (left)	4.96	4.48	0.99	1.12
Femur neck BMD (left)	0.94	0.82	0.14	0.13
Femur total BMD (left)	1.01	0.86	0.16	0.13
Femur Total BMD T-score(left)	−0.32	−1.47	1.15	0.92
Femur troch BMD (left)	0.85	0.72	0.16	0.14
Femur troch BMD T-score(left)	−0.10	−1.21	1.25	1.06
Femur wards BMD (left)	0.73	0.60	0.14	0.12
Femur wards BMD T-score(left)	−1.59	−2.51	1.08	0.89
Pelvis BMC	334.96	299.12	79.90	73.75
Femur neck BMC (right)	4.99	4.43	1.00	0.89
Femur neck BMD (right)	0.94	0.82	0.14	0.11
Femur total BMD (right)	1.00	0.86	0.16	0.13
Femur neck BMD T-score(right)	−0.34	−1.43	1.14	0.94
Femur troch BMD (right)	0.84	0.72	0.16	0.14
Femur troch BMD T-score(right)	−0.15	−1.22	1.23	1.09
Femur wards BMD (right)	0.73	0.60	0.15	0.15
Femur wards BMD T-score(right)	−1.58	−2.55	1.13	0.85
Weight	77.27	74.48	14.39	15.05

fracture risk assessment (Fig. 5).

4. Discussion

In our study, we developed a staged based ML model to predict hip fracture, utilizing data from 547 patients, including 94 individuals with a history of hip fracture from the UK Biobank dataset. Ensemble 1 included only clinical features while Ensemble 2 included DXA imaging features along with clinical features. The Staged model performed comparable to Ensemble 2 with an AUC of 0.85 compared to 0.95, accuracy of 0.86 compared to 0.92, while requiring DXA data only 45.52 % of the subjects. The inclusion of imaging features in Stage II, reduced uncertainty for both hip fracture and non-fracture subjects (Fig. 6), improving sensitivity and specificity (0.81 and 0.94, respectively).

4.1. Staged modeling for hip fracture risk prediction

Our staged approach for hip fracture risk prediction represents a novel methodology aimed at enhancing the accuracy and reliability of fracture risk assessment. Unlike traditional single-stage models, which often rely on a singular set of features for prediction, our approach systematically integrates multiple stages, each tailored to leverage specific types of data. In the first stage of our staged approach, we focus on utilizing clinical variables to build a foundational understanding of each patient’s health profile. This initial stage incorporates demographic details, medical history, lifestyle factors, and other relevant clinical indicators to establish a comprehensive baseline for fracture risk assessment. Following the initial clinical assessment, our approach progresses to the Ensemble Stage II, where imaging features extracted from hip

Table 4
P-values from both the chi-square test and Fisher test for categorical features. The tests evaluate associations between categorical variables and the outcome.

Feature	Chi-square	Fisher's
Falls in Last Year (Prefer Not to Answer)	0.7691	0.3144
Falls in Last Year (None)	0.9999	1.0000
Falls in Last Year (Only One)	0.9999	1.0000
Falls in Last Year (More Than One)	0.9999	1.0000
Major Change in Diet in Last 5 Years (Not Due to Illness)	0.2684	0.2402
Major Change in Diet (Due to Illness)	0.8860	0.6522
Major Change in Diet (Other Reasons)	0.1721	0.1725
Variation in Diet (Never/Rarely)	0.7444	0.7201
Variation in Diet (Sometimes)	0.7144	0.6488
Variation in Diet (Often)	0.9999	0.8480
Alcohol Consumption (Never Consumed)	0.0017	0.0036
Alcohol Consumption (Previous)	0.9810	1
Alcohol Consumption (Current)	0.0448	0.0329
Smoking (Prefer Not to Answer)	0.9810	1
Smoking (Never)	0.9786	0.9098
Smoking (Previous)	0.6744	0.6452
Smoking (Current Smoker)	0.3213	0.2428
Average Household Income (Do Not Know)	0.9999	1.0000
Average Household Income (Prefer Not to Answer)	0.5654	0.4599
Average Household Income (<£18,000)	0.0162	0.0185
Average Household Income (£18,000 to £30,999)	0.3596	0.3148
Average Household Income (£31,000 to £51,999)	0.1800	0.1628
Average Household Income (£52,000 to £100,000)	0.8757	0.7635
Average Household Income (>£100,000)	0.8714	1.0000
Vitamin Supplement	0.9665	1.0000
Genetic Sex	0.2373	0.2132
Fracture/Broken Bones in Last 5 Years	0.8658	0.6689
Sex	0.2373	0.2132

Table 5
P-values obtained from t-tests for categorical features used in the study. The t-tests evaluate differences in means between groups for each variable.

Feature	p-value
Age	0.0188
Weight	0.1012
Right Femur Neck BMD	<0.0001
Right Femur Neck BMC	<0.0001
Right Femur Total BMD	<0.0001
Right Femur Total BMD T-score	<0.0001
Right Trochanter BMD	<0.0001
Right Trochanter BMD T-score	<0.0001
Right Wards BMD	<0.0001
Right Wards BMD T-score	<0.0001
Left Femur Neck BMD	<0.0001
Left Femur Neck BMC	<0.001
Left Femur Total BMD	<0.0001
Left Femur Total BMD T-score	<0.0001
Left Trochanter BMD	<0.0001
Left Trochanter BMD T-score	<0.0001
Left Wards BMD	<0.0001
Left Wards BMD T-score	<0.0001
Pelvis BMC	<0.0001

DXA images are included. By incorporating this additional layer of data, we aim to enrich the predictive capabilities of our model, capturing subtle nuances and anatomical insights that may not be discernible from clinical variables alone. Ensemble 2 emerged as the top-performing model, achieving a high AUC with strong accuracy, sensitivity, and specificity. In assessing the performance of the stage II model within our staged framework, we scrutinized its AUC alongside corresponding CIs relative to standard deviation percentiles (Fig. 5). As our analysis progressed from left to right along these percentiles which results in a smaller and smaller subset of the data whose patients have higher uncertainty, we notice the performance of the model in terms of AUC decreases. This approach allows us to delve into predictions with higher uncertainty, showcasing that increased uncertainty leads to decreased performance. Moreover, this behavior potentially creates the opportunity to add a third stage of analysis, e.g. genetic data (Nethander et al., 2022) or quantitative computed tomography images (Awal and Faisal, 2024) which are more costly and less available than DXA but can provide more nuanced and complementary information in such a sequential approach.

One of the key strengths of our staged approach lies in its adaptability and flexibility. The use of internal logic rules allows for dynamic decision-making, determining whether the acquisition of DXA data is necessary based on the information gathered in the initial clinical stage. This ensures that resources are allocated efficiently, with additional imaging studies being performed only when deemed essential for accurate risk assessment. Moreover, our staged approach offers enhanced interpretability compared to complex AI-driven models. By breaking down the prediction process into distinct stages, clinicians can better understand the rationale behind each decision, facilitating trust and confidence in the model's outputs.

4.2. Comparison with BMD T-scores

Our study compared the performance of our staged approach for hip fracture risk prediction with left femur BMD T-scores, a widely used metric for diagnosing osteoporosis (Faulkner, 2005). The classification of bone density categories by the World Health Organization (WHO) Study Group is determined by T-scores. According to their criteria, bone density is considered normal when the T-score is −1 or greater, while osteopenia is characterized by T-scores ranging between −1 and − 2.5, and osteoporosis is diagnosed when the T-score falls at −2.5 or below (Cosman et al., 2014). Using the T-score to predict hip fracture risk, individuals with normal bone density had a prediction accuracy of 0.37, sensitivity of 0.26, and specificity of 0.40. Those with osteoporosis showed higher accuracy (0.83) and specificity (0.98) but lower sensitivity (0.13). Osteopenia had intermediate with accuracy, sensitivity, and specificity of 0.62. In our study, 52 % of individuals with normal bone density, 56 % with osteopenia and 64 % with osteoporosis had low uncertainty in Stage I, suggesting that DXA scans were often unnecessary for assessing fracture risk. Our staged approach with an average AUC of 0.85, an accuracy of 0.86, a sensitivity of 0.56, and a specificity of 0.92, effectively identified high-risk individuals and demonstrated strong predictive performance, supporting its potential utility in clinical practice.

4.3. Comparison with FRAX

FRAX predictions were calculated using clinical and DXA information extracted from the UK Biobank dataset. We utilized the official FRAX tool from the University of Sheffield, incorporating both BMD and non-BMD methods. For BMD calculations, femur neck BMD measurements from the UK Biobank were used, and the tool was set to reflect the use of Lunar equipment. In the absence of BMD data, the FRAX without BMD option was applied. Missing data were managed by excluding participants with incomplete information from the analysis, ensuring that only complete data were used for FRAX calculations. This approach maintained the accuracy of the predictions, with the final 10-year probability of hip fracture derived from a systematic web automation process. In comparing the performance of our staged approach for hip fracture risk prediction with the FRAX tool, we observe notable differences in predictive accuracy. Our approach, leveraging a combination of clinical variables and imaging features, achieved an AUC of 0.85 with an accuracy of 0.86, sensitivity of 0.56, and specificity of 0.93, which is far superior to FRAX with BMD (left femur), AUC of 0.76 (95 % CI of 0.70–0.82), an accuracy of 0.71, sensitivity of 0.69 and specificity of 0.72. The higher AUC value of our approach indicates enhanced sensitivity and specificity in identifying individuals at risk of hip fractures, thereby improving the overall predictive performance. Similarly, when comparing our approach to FRAX without BMD, which yielded an AUC

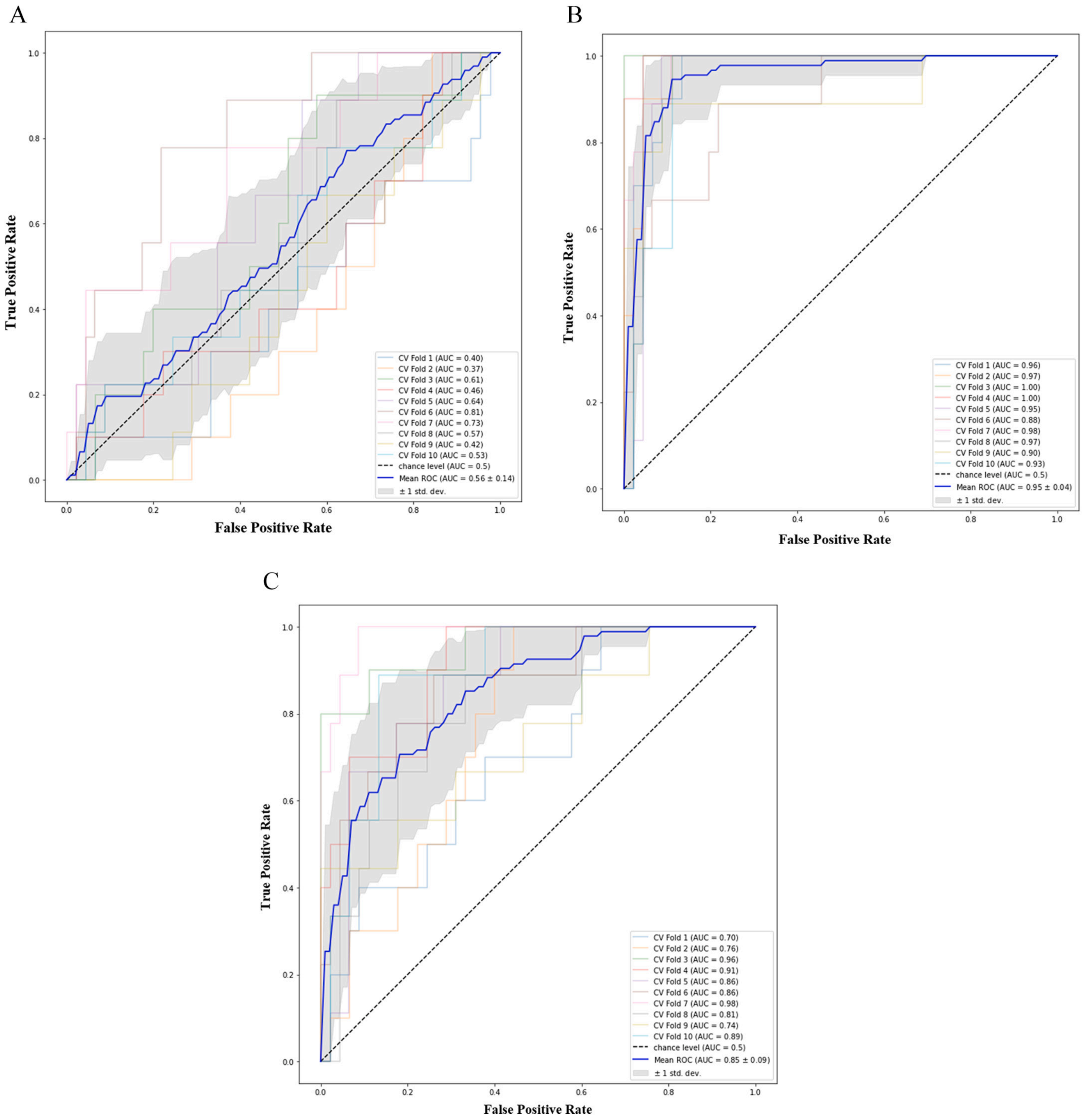


Fig. 3A. ROC curves for ensemble stage I

Fig. 3A plot depicts the mean ROC curve representing the average performance of Ensemble Stage I in predicting hip fracture risk, along with variability represented by standard deviation. The ROC curve illustrates the trade-off between sensitivity and specificity across different threshold values.

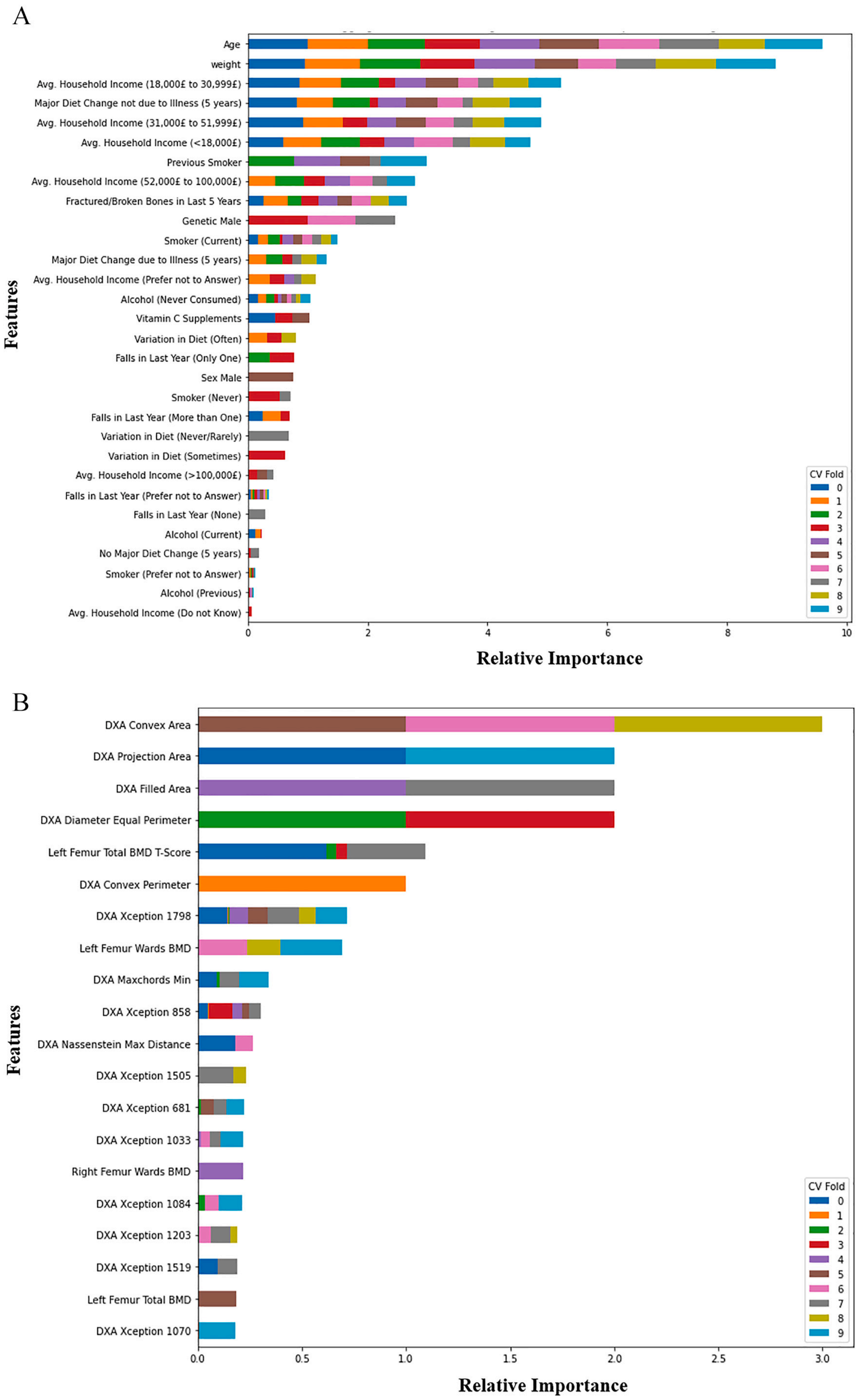
Fig. 3B. ROC curves for ensemble stage II

Fig. 3B. The plot shows the mean ROC curve for Ensemble Stage II, depicting the average performance across multiple iterations. The variability around the mean curve illustrates the uncertainty associated with the model's predictions.

Fig. 3C. ROC curves for the staged model

Fig. 3C depicts Receiver Operating Characteristic (ROC) curves for the staged hip fracture risk prediction model. The mean ROC curve represents the average performance across multiple iterations, with the shaded area indicating the variability or uncertainty associated with the model's predictions.

of 0.62 (95 % CI of 0.55–0.69), an accuracy of 0.76, sensitivity of 0.42 and specificity of 0.83, our model again exhibited superior performance (p < 0.001). The DeLong tests yielded p-values, indicating the significance of the differences in AUC between different model pairs. For instance, the p-value for comparing Staged vs. FRAX with BMD (left femur) was 0.004, and for comparing Staged vs. FRAX without BMD, it was <0.0001). Despite FRAX being a widely used tool for fracture risk assessment, our staged approach demonstrated enhanced accuracy and



(caption on next page)

Fig. 4A. Feature importance – Ensemble stage I
Fig. 4A illustrates the fold-aggregated ensemble-averaged absolute feature importance for stage I ensemble models. This assessment determines the importance of each feature based on its contribution to the predictive performance of the ensemble model.
Fig. 4B. Feature importance – Ensemble stage II
Fig. 4B showcases the top 20-fold aggregated ensemble averaged absolute feature importance for Stage II ensemble modeling. The importance of each feature is determined based on its contribution to the predictive performance of the model.

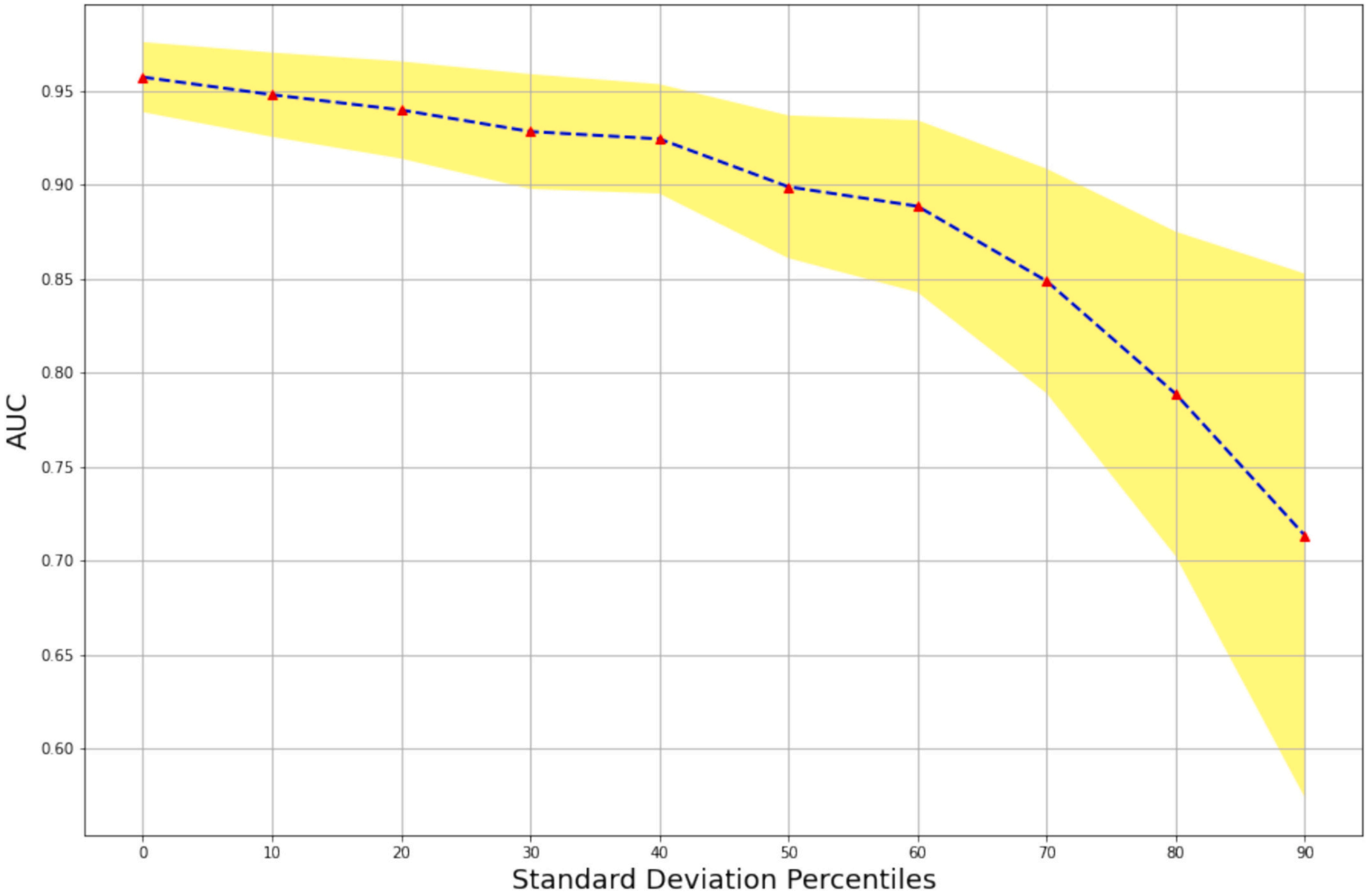


Fig. 5. Ensemble 2's AUC with 95 % DeLong Confidence Intervals(CI) across standard deviation percentiles or more extreme values. This analysis offers insights into the model's performance variability across varying levels of uncertainty.

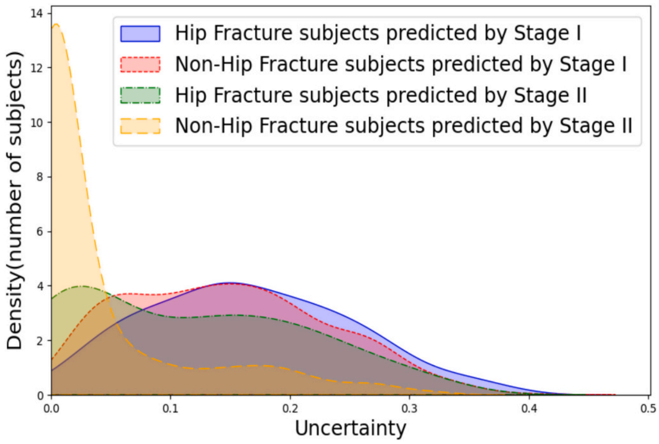


Fig. 6. The distribution of uncertainty observed in both Stage I, where only clinical factors are utilized, and Stage II, where clinical factors and DXA imaging features are incorporated.

reliability in predicting hip fractures, underscoring the effectiveness of incorporating imaging features alongside clinical variables in a stepwise, staged manner.

4.4. *Cost and radiation reduction*

Our staged approach for hip fracture risk prediction not only improves diagnostic accuracy but also addresses cost and radiation concerns. It identified that 54.49 % of the patients did not require DXA scanning, while 45.52 % did. For subjects with BMD (left femur) T-score ≤ -1.0 and > -1.0, 55 % and 52 %, respectively, had low uncertainty in Stage I, indicating DXA scans were unnecessary for determining high fracture risk. This efficiency in distinguishing high-risk patients early in the diagnostic process optimizes resource allocation, leading to significant cost savings.

By leveraging existing imaging data, such as DXA scans, our approach reduces the need for additional imaging tests, minimizing radiation exposure and associated risks. This targeted use of clinical and imaging data not only enhances diagnostic accuracy but also lowers the likelihood of missed diagnoses or unnecessary treatments, contributing to overall healthcare cost reduction.

4.5. Interpretability of our model

Our model's clinical relevance is highlighted by its identification of significant predictors of hip fracture risk as age, weight, dietary changes, and DXA parameters consistent with established literature on fracture risk factors. These findings have the potential to guide clinical decision-making by enabling the early identification of individuals at high risk of fractures, thus facilitating the implementation of tailored interventions to effectively reduce fracture risk. Additionally, our staged modeling approach offers the potential for further refinement and expansion by incorporating additional features, such as genetic data, to enhance its predictive capabilities.

4.6. Future feasibility

Extracting the features used in our model from DXA data presents challenges, especially as raw data may not be accessible in standard clinical settings. To address this, collaboration with DXA manufacturers to develop tools for direct feature extraction is essential. In the meantime, independent software solutions for processing DXA outputs could serve as an interim measure. Ensuring practical implementation will require efforts to standardize procedures and integrate manufacturer support.

4.7. Limitations

First, this study involved a relatively small number of subjects, which is an inherent limitation in the study design. Increasing the sample size would improve the statistical power and generalizability of our findings. Additionally, the performance and feasibility of the data-driven system might be influenced by the quality of the data. For instance, inconsistencies or inaccuracies in clinical data and DXA images could impact the model's predictive accuracy. Second, there were missing features in the UKBiobank repository. Not all potential risk factors for hip fractures were captured or included in the analysis. This limitation might have resulted in an incomplete representation of each patient's health profile. Incorporating additional relevant features such as genetic data, comprehensive environmental factors, and more detailed medical history could further refine the model's predictive capabilities. Third, the ethnicity of all patients in our sample is British, emphasizing the homogeneity of ethnic background within our study population. This lack of ethnic diversity may limit the generalizability of the model to more diverse populations. Lastly, and perhaps most importantly, this study did not include external validation using datasets from other populations or healthcare settings.

5. Conclusion

We developed a staged approach combining clinical data and DXA hip images for hip fracture risk prediction. By considering various factors like age, weight, and bone health alongside images with machine learning and uncertainty quantification, our staged model offers a cost-effective holistic view of patients' health. Through rigorous evaluation, we found that our staged approach could identify individuals at risk with a high accuracy while reducing the need for DXA scans by 54.49 %. This staged approach shows great promise to guide interventions to prevent hip fracture with reduced cost and radiation.

CRediT authorship contribution statement

Anjum Shaik: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Conceptualization. **Kristoffer Larsen:** Writing – review & editing, Software, Methodology, Formal analysis. **Nancy E. Lane:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Chen Zhao:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kuan-Jui Su:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Joyce H. Keyak:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Qing Tian:** Writing – review & editing, Project administration, Data curation, Conceptualization. **Qiuying Sha:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Hui Shen:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition, Conceptualization. **Hong-Wen Deng:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Weihua Zhou:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

All authors declare that there are no conflicts of interest.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research has been conducted using the UK Biobank Resource under application number [61915]. It was supported in part by grants from the National Institutes of Health, USA (P20GM109036, R01AR069055, U19AG055373, R01AG061917, and R15HL172198) and NASA Johnson Space Center, USA contracts NNJ12HC91P and NNJ15HP23P.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bonr.2024.101805.

References

Awal R, Faisal T. QCT-based 3D finite element modeling to assess patient-specific hip fracture risk and risk factors. J. Mech. Behav. Biomed. Mater. 2024 Feb 1;150: 106299.

Cha Y, Kim J-T, Park C-H, Kim J-W, Lee SY, Yoo J-I. Artificial intelligence and machine learning on diagnosis and classification of hip fracture: systematic review. J. Orthop. Surg. Res. 2022 Dec 1;17(1):520. (PMCID: PMC9714164).

Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions [Internet]. arXiv; 2017 [cited 2024 May 20]. Available from: http://arxiv.org/abs/1610.02357.

Cosman, F., de Beur, S.J., LeBoff, M.S., Lewiecki, E.M., Tanner, B., Randall, S., Lindsay, R., 2014. Clinician's guide to prevention and treatment of osteoporosis. Osteoporos. Int. 25 (10), 2359–2381 (PMCID: PMC4176573).

Dhanwal, D.K., Dennison, E.M., Harvey, N.C., Cooper, C., 2011. Epidemiology of hip fracture: worldwide geographic variation. Indian J. Orthop. 45 (1), 15–22. Jan. (PMCID: PMC3004072).

Emmerson BR, Varacallo M, Inman D. Hip Fracture Overview. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 [cited 2024 May 20]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK557514/ PMID: 32491446.

Faulkner, K.G., 2005. The tale of the T-score: review and perspective. Osteoporos. Int. 16 (4), 347–352. Apr. (PMID: 15565352).

Gullberg, B., Johnell, O., Kanis, J.A., 1997. World-wide projections for hip fracture. Osteoporos. Int. 7 (5), 407–413. Sep 1.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for Cancer classification using support vector machines. Mach. Learn. 1 (46), 389–422. Jan.

Hong, N., Park, H., Kim, C.O., Kim, H.C., Choi, J.-Y., Kim, H., Rhee, Y., 2021. Bone Radiomics score derived from DXA hip images enhances hip fracture prediction in older women. J. Bone Miner. Res. 36 (9), 1708–1716. Sep. (PMID: 34029404).

Kroell, N., 2021. Imea: a Python package for extracting 2D and 3D shape measurements from images. Journal of Open Source Software. 6 (60), 3091. Apr 6.

Lex, J.R., Di Michele, J., Koucheki, R., Pincus, D., Whyne, C., Ravi, B., 2023. Artificial intelligence for hip fracture detection and outcome prediction. JAMA Netw. Open 6 (3), e233391. Mar 17. (PMCID: PMC10024206).

Murphy, E.A., Ehrhardt, B., Gregson, C.L., von Arx, O.A., Hartley, A., Whitehouse, M.R., Thomas, M.S., Stenhouse, G., Chesser, T.J.S., Budd, C.J., Gill, H.S., 2022 Feb 8.

Machine learning outperforms clinical experts in classification of hip fractures. Sci Rep. Nature Publishing Group 12 (1), 2058.

Nethander, M., Coward, E., Reimann, E., Grahnemo, L., Gabrielsen, M.E., Wibom, C., Team, Estonian Biobank Research, Mägi, R., Funck-Brentano, T., Hoff, M., Langhammer, A., Pettersson-Kymmer, U., Hveem, K., Ohlsson, C., 2022. Assessment of the genetic and clinical determinants of hip fracture risk: genome-wide association and Mendelian randomization study. Cell Rep Med. 3 (10), 100776. Oct 18. (PMCID: PMC9589021).

Resource 502 [Internet]. [cited 2024 May 20]. Available from: https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=502.

Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [Internet]. arXiv.org. 2014 [cited 2024 May 20]. Available from: https://arxiv.org/abs/1409.1556v6.

van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.-C., Pieper, S., Aerts, H.J.W.L., 2017. Computational Radiomics system to decode the radiographic phenotype. Cancer Res. 77 (21), e104–e107. Oct 31.

Zhao C, Keyak JH, Cao X, Sha Q, Wu L, Luo Z, Zhao L, Tian Q, Qiu C, Su R, Shen H, Deng H-W, Zhou W. Multi-view information fusion using multi-view variational autoencoders to predict proximal femoral strength [Internet]. arXiv; 2023 [cited 2023 Oct 31]. Available from: http://arxiv.org/abs/2210.00674.