Machine Learning for Computational Science and Engineering (2025) 1:2
https://doi.org/10.1007/s44379-024-00004-w

RESEARCH

Interpretable machine learning classifiers for the reliable prediction of fall induced hip fracture risk

Rabina Awal[1,2] · Sarah C. Doll[1] · Mahmuda Naznin[3] · Tanvir R. Faisal[1]

Received: 21 August 2024 / Accepted: 19 September 2024

Abstract
Classical computing methods are costly and require advanced skills, limiting their clinical use. A data-driven framework offers an effective alternative for disease diagnosis and prediction. This study aims to apply and evaluate machine learning (ML) classifiers to predict hip fracture risk using a binary classification based on the fracture risk index (FRI) from Quantitative Computed Tomography-based Finite Element Analysis (QCT-based FEA). This study concentrated on comparing the performance of different ML models such as logistic regression, Support Vector Classifier (SVC), Decision Tree (DT), Categorical Boosting Model (CatBoost), Extreme Gradient Boost Model (XGBM), and Random Forest (RF) for the prediction of hip fracture probability. The models were trained with a dataset comprises clinical parameters, bone anatomy, and loading directions mimicking sideways fall postures. All the ML models were compared based on the performance metrics—precision, recall, F1 score, accuracy, and Area Under Receiver Operating Curve (AUROC). Both logistic regression and SVC exhibited the highest performance in assessing fracture risk with 82% accuracy and 87% AUROC. These models are also interpretable, as we used SHapley Additive exPlanations (SHAP) to identify the most important features and their impacts on the prediction process. Despite being trained on a limited dataset, this study demonstrates the viability of machine learning models in predicting hip fracture risk. To the best of our knowledge, these models are the first interpretable ML-based predictors of hip fracture risk.

Keywords Hip fracture · 3D proximal femur · QCT-based FEA · Machine learning · Logistic regression · SVC

1 Introduction

Hip fracture, also known as femoral fracture, is a major public health issue, especially among the elderly, causing long-term disability and significant mortality [1]. Annually, about 1.3 million hip fractures occur globally, leading to approximately 740,000 deaths [2]. This number is projected to exceed 6 million by 2050 [3]. In the USA, over 300,000 patients are hospitalized each year due to hip fractures, with 90% resulting from simple falls [4]. Accurate prediction of hip fracture risk is crucial for developing preventive measures and reducing the associated socio-economic burden. Current fracture risk assessment tools like FRAX [5], and Bone Mineral Density (BMD) measurements via Dual-Energy X-ray Absorptiometry (DXA) [6–9] have limitations. FRAX may be less accurate across different racial groups and can overestimate or underestimate BMD based on bone size [10].

Finite Element Analysis (FEA) is widely used in biomechanical modeling for cardiovascular systems, hemodynamics, implants, orthodontics, and surgical procedures [11–13]. Quantitative Computed Tomography-based FEA (QCT-based FEA) is a precise method for fracture assessment, considering bone density, femur morphology, and loading effects, [14–16] and offers patient-specific predictions of femur fracture strength [17, 18]. Despite its accuracy, QCT-based FEA is limited in clinical use due to high computational demands, expensive software, and the need for skilled personnel. We hypothesized that advancements in Artificial Intelligence (AI) and Machine Learning (ML)

✉ Tanvir R. Faisal
tanvir.faisal@louisiana.edu

[1] Department of Mechanical Engineering, University of Louisiana at Lafayette, Lafayette, LA 70503, USA

[2] Schlumberger Limited (SLB), 23400 Colonial Parkway, Katy, TX 77493, USA

[3] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh

Published online: 09 October 2024

could provide a more efficient and accurate data-driven approach to predict hip fracture risk. Thus, this study aims to evaluate various ML classifiers to predict fracture probability in geriatric patients.

The clinical application of machine learning in disease diagnosis [19–21] and injury prevention shows promising outcomes [22, 23]. ML techniques have been successfully used in orthopedic applications such as 3D bone reconstruction [24, 25], fracture detection [26–30], bone tumor diagnosis [31], and osteoarthritis grading [32, 33]. For example, Marcus et al. [34] used radiographs to predict hip fractures by analyzing patient and hospital variables, and Villamor et al. [35] explored ML models to predict fracture risk with a simplified 2D FE model. Ferizi et al. [36] compared fifteen ML classifiers using MRI data to predict osteoporotic bone fractures. However, most ML models focus on detecting and categorizing existing fractures. This study aims to evaluate the performance of various ML classifiers in predicting the probability of fall-induced hip fractures using the Fracture Risk Index (FRI) obtained from QCT-based FEA [37–39].

This research evaluated several classical ML classifiers [40–46] for predicting fracture risk, including Logistic Regression (Logit), Support Vector Classifier (SVC), and Decision Tree (DT), as well as ensemble methods like CatBoost, Extreme Gradient Boosting Model (XGBM), and Random Forest (RF). Additionally, the study examined the impact of various clinical and anatomical features and fall postures on prediction accuracy. The goals of this study were to 1) assess the effectiveness of ML classifiers in predicting hip fracture risk using a high-fidelity QCT-based FE dataset, and 2) identify features that influence or promote fracture risk from falls.

2 Materials and method

The framework for predicting hip or femoral fracture risk includes four processing steps: *data acquisition* (QCT image dataset) in Digital Imaging and Communications in Medicine (DICOM) format, *data preparation* (extracting features and target variable from QCT-based FEA), *feature engineering*, and *model training and performance evaluation* (Fig. 1).

2.1 Data acquisition

The dataset was built upon the QCT images of 97 anonymous patients, removing all the personal information, previously obtained from the Great-West Life PET/CT Center, Winnipeg, Canada in DICOM format [14, 39, 47]. DICOM images, 512 X 512 pixels, were obtained by SIEMENS S5VB40B CT scan machine (Siemens Medical Solution, Malvern, USA) with acquisition and reconstruction variables of 120 kVp and 244 mAs. The DICOM dataset was used for feature extraction as well as for conducting FEA to obtain datasets for training and target variables.

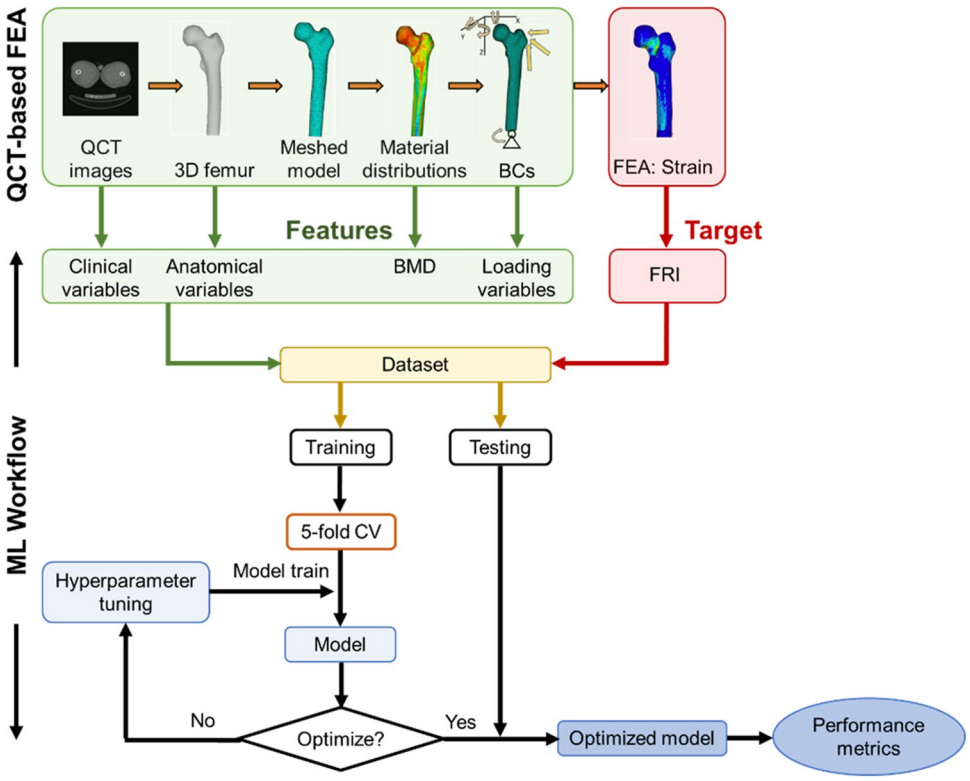

Fig. 1 An overview of ML based framework considering clinical variables, femur anatomy, and loading directions as the input features, and FRI as the target variable

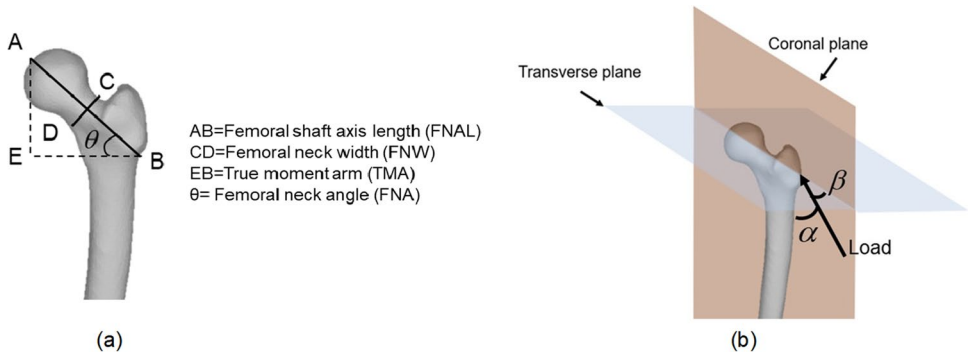

Fig. 2 **a** A proximal femur anatomy and its different geometric parameters that are considered in this study. **b** Loading direction α on coronal plane, varying from 0° to 30°, and β on transverse plane, ranging from -15° to + 15° to mimic different sideways fall postures

2.2 Data preparation

2.2.1 Feature extraction

The features extracted from the DICOM dataset and QCT-based FEA (Fig. 1) included clinical parameters, bone anatomy, and loading orientations. Key variables affecting the FRI [37, 39] were prioritized. Clinical and demographic data, along with BMD distributions in the 3D femur [38, 39], were derived from the DICOM data. Anatomical features, such as *Femoral Neck Axis Length (FNAL)*–AB, *Femoral Neck Width (FNW)*–CD, *True Moment Arm (TMA)*–EB, *Femoral Neck Angle (FNA)* –θ, and the horizontal component of FNAL on the transverse plane were obtained from 2D projections of the 3D femur model (Fig. 2a). Different sideways fall postures were simulated by adjusting the loading angle (α) from 0° to 30° in 15° intervals on the coronal plane relative to the shaft axis, and the angle (β) from -15° to + 15° in 15° intervals on the transverse plane relative to the neck axis (Fig. 2b). These angles were chosen based on typical fall orientations and previous experimental studies where femoral fractures were most frequent [48].

2.2.2 Target variables via QCT-based FEA

The target variable, FRI, was obtained from QCT-based FEA as shown in Fig. 1. Prior research by Schileo et al. [49] as well other studies [50, 51] demonstrated that fracture risk estimation can be more accurately predicted in FEA, considering principal strain. Therefore, for obtaining ground truth, 1st and 3rd principal strain-based FRI was adopted in the QCT-based FEA as follows.

$$FRI = \frac{\varepsilon^{T}_{\max}}{0.0073} \tag{1}$$

$$FRI = \frac{|\varepsilon^{c}_{\max}|}{0.0104} \tag{2}$$

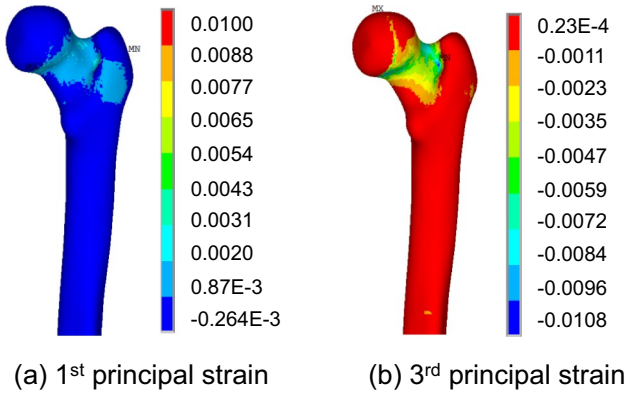

Fig. 3 Typical strain distributions in sideways fall, with $\alpha = \beta = 0^{o}$, obtained via FEA. **a** 1st principal (Tensile) strain and (**b**) 3rd principal (Compressive) strain

Table 1 Feature and target variable for predicting fracture risk based on strain-based FRI

Feature			Target
Clinical variables	Anatomical variables	Loading variables	
Age	FNA	α	FRI
Weight	FNW	β	
Sex	TMA		
BMD	FNAL		

where $\varepsilon^{T}_{\max}$ and $\varepsilon^{c}_{\max}$ are the maximum principal strain in tension (1st principal strain) and compression (3rd principal strain), respectively. Figure 3 shows typical 1st and 3rd principal strain distributions in proximal femurs. A more detailed description of the QCT-based FEA can be found in our prior works [37–39].

An FRI value above 1 signifies a higher risk of hip fracture. Therefore, a binary classification was used: "fracture probability" is coded as 1, and "no fracture probability" as 0. Table 1 summarizes all categorical features and target variables in this study.

2.3 Feature engineering

To enhance the predictive models' effectiveness, feature engineering was performed during the data preprocessing stage. This involved tasks such as cleaning, correlation analysis, data splitting, and feature scaling. These steps aimed to streamline the data complexity and enhance the accuracy of the models.

2.3.1 Data cleaning

The target variable FRI displayed a positively skewed distribution and dispersion. In this study, FRI values exceeding 1.5 times the interquartile range were identified as outliers and removed from the dataset. This exclusion involved two femurs from two patients and resulted in 192 out of 194 femurs for analysis.

2.3.2 Correlation analysis

The association between the features was analyzed using the Pearson correlation coefficient to evaluate if the features have any form of association with each other to affect the target variable. The correlation of the features has been shown by a heatmap (Fig. 4), which demonstrates no correlation among the features, and therefore, no features were excluded.

2.3.3 Data splitting

The dataset was split into 80–20 ratio for training and testing of the ML models. To optimize and increase the robustness of the model by reducing data bias, the training data was further split into 5 folds in the same 80–20 ratio. Splitting was done based on the patient's unique identifier to prevent

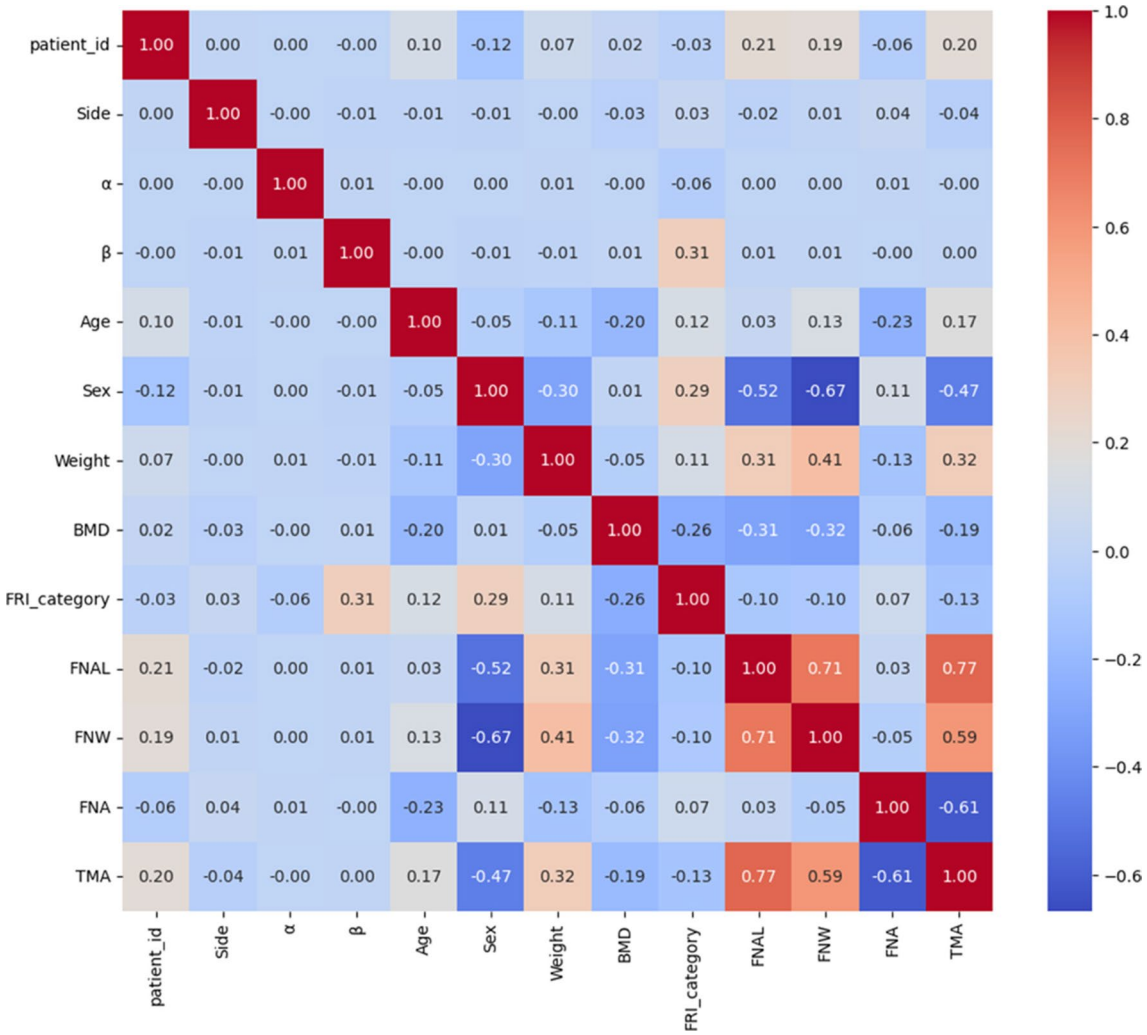

Fig. 4 A heatmap showing the correlation among the features. Any datapoint on heatmap represents the correlation between corresponding row and column index. A higher correlation between the features indicates a similar effect on FRI

data leakage. The models' performance was tested with a completely unseen dataset.

2.3.4 Feature scaling

Feature scaling is another vital step towards standardizing the features to increase the predictability of a model. In this step, features were scaled to a 0 to 1 range using a MinMax scaler based on the minimum and maximum values of each feature, and it was applied separately to the training and testing datasets.

2.4 Predictive models

After an exhaustive trial with several ML models, we considered the following ML classifiers based on our problem domain, modeling framework, and our ultimate objectives.

2.4.1 Logistic regression

Logistic regression belongs to the group of statistical models called generalized linear models [52]. It is a classical approach commonly applied in binary and linear classification problems [53]. This process involves two main steps: initially, it computes a linear combination of the independent/predictor variables while incorporating a bias term; subsequently, it applies a sigmoid function to estimate the probability of belonging to a particular class (Eq. 3).

$$P\left(class \mid x_1, \cdots, x_N\right) = \frac{1}{1 + e^{-\left(w_o + \sum_{k=1}^{N} w_k . x_k\right)}} \tag{3}$$

where $w_k (k \in [0, N])$ are the model parameters, N is the number of predictor variables, and x_k are these variables for a given patient.

2.4.2 Support vector classifier

A support vector classifier is a supervised machine-learning algorithm that handles both linear and nonlinear problems [54]. Originally designed for binary classification, SVCs can also perform regression tasks. In classification, a SVC identifies the optimal hyperplane in the transformed space to separate classes effectively. It is memory efficient, using only a subset of training data in its decision function, and creates a generalized model with minimal error by drawing a margin along the regression line [55].

2.4.3 Decision tree

Decision tree is a simple yet powerful tool for multi-variable analysis. As a hierarchical supervised learning model, DT predicts the target variable by asking a series of questions about predictor variables, partitioning data into subsets through internal decision nodes and terminal leaves [53].

2.4.4 XGBM

XGBM is a popular boosting algorithm that combines predictions from multiple decision trees, built sequentially to correct errors from previous trees. It uses parallel processing and regularization to reduce overfitting and enhance performance, making XGBM more predictive and faster than other gradient boosting methods [56].

2.4.5 CatBoost

CatBoost is a high-performance classifier that requires minimal data preprocessing compared to other ML algorithms [57]. It combines random permutations, ordered boosting, and gradient-based optimization to enhance gradient boosting accuracy and efficiency, making it effective for large, complex datasets [57]. CatBoost is particularly well-suited for handling categorical features and managing overfitting.

2.4.6 Random forest

Random forest represents a significant adaptation of bagging, constructing a substantial ensemble of uncorrelated trees and subsequently averaging them [58]. While RF often exhibits performance akin to boosting on numerous tasks, they are generally simpler to train and fine-tune. The RF classifier utilizes randomly chosen features or combinations of features at each node to construct a tree.

2.5 Model training and performance metrics for model evaluation

2.5.1 Model training and testing

To train the ML models, k-fold cross-validation (here, k = 5) was used and each model was iterated multiple times by tuning hyperparameters by a random search method until an optimized model was built. Both training and testing models were iterated with the same hyperparameters until each fold contributes as a testing dataset.

2.5.2 Performance metrics

To evaluate the performance of the ML models, performance metrics such as *precision*, *recall* (*sensitivity*), *accuracy*, *F1 score*, and *Area Under Receiver Operating Characteristics (AUROC)* curve [59, 60] were considered. In this study, the possibility of fracture is positive and

classified as 1, and no fracture is negative and denoted as 0. Considering QCT-based FE outcomes as the ground truth, in this work,

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$F1\ score = \frac{2 \times \text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{7}$$

where *True Positive* (*TP*) represents correctly predicted cases of fracture risk, and *True Negative* (*TN*) represents correctly predicted cases of no fracture risk by the ML classifiers. *False Positive* (*FP*) indicates when the ML model falsely predicts fracture, while there is no fracture risk in ground truth, and *False Negative* (*FN*) denotes the ML model that is falsely predicting no fracture when ground truth indicates a fracture.

2.5.3 SHAP values

To enhance interpretability, we used SHapley Additive exPlanations (SHAP) to identify key features and their impacts on model predictions. SHAP values provide a universal method for interpreting ML models by showing the average change in model output and the impact of each variable, both positive and negative. They offer individual values for each feature in each instance, resulting in a plot that highlights important features and their effects on the model's predictions.

2.6 Testbed description

The testing environment for our experiments involved a cluster computer with an Intel ®Xeon ® E5-2600 v4 CPU running at 2.00 GHz, 1 Tesla GPU of RAM of 16 GB and 256 GB RAM. We chose to use Python 3 [61] because of its ease of use and the availability of relevant libraries. For the implementation of the model, several Python libraries were incorporated in this work. We used TensorFlow [62] and SciPy [63] libraries for scientific and technical computing. We also used the libraries associated with Google Colab, and Visual Studio Code that are widely used for Python development.

3 Results

To evaluate the performance of the adopted ML classifiers for predicting hip fracture risk based on FRI, we compared precision, recall, F1 score, and accuracy across models. Both logistic regression and SVC exhibit the maximum precision (0.79), followed by CatBoost (0.78), RF (0.71), XGBM (0.69), and DT (0.62), (Fig. 5). High precision reduces false positives, ensuring healthy femurs are rarely mislabeled as at risk.

Logistic regression and SVC also had the highest recall (0.81), followed by CatBoost (0.77), RF (0.72), XGBM (0.71), and DT (0.64). High recall indicates a model's effectiveness in identifying true fracture risks, reducing the likelihood of missing actual hip fractures.

The F1 score evaluates a model's predictive capability, offering insights into its performance across different classes. Logistic regression, SVC, and CatBoost all exhibit the highest F1 score of ~ 0.79. Both XGBM and RF demonstrate nearly the same value of 0.7 for F1 score, whereas DT performs poorly with 0.57 only. The F1 score employs a harmonic means to provide a balanced assessment of a model's predictive performance by considering both precision and recall.

The accuracy of an ML model measures the rate of correct predictions, considering both true positives (possibility of hip fracture) and true negatives (possibility of no hip fracture) equally. Unlike precision and recall, which focus solely on predicting hip fracture risk, accuracy values of both the correct predictions of risk and no risk. Logistic Regression and SVC demonstrated the highest accuracy at 0.82 each (Fig. 5). The accuracy for ensemble models was 0.76 for CatBoost, 0.71 for XGBM, and 0.74 for RF. The DT model performed poorly, with an accuracy of 0.57.

To compare the performance of ML classifiers, AUROC curves were plotted for all models (Fig. 6a). These curves show the tradeoff between the true positive rate and the

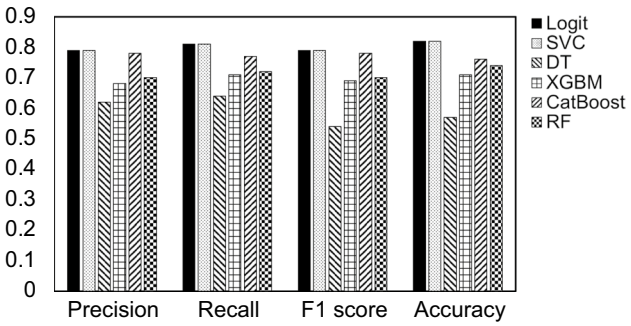

Fig. 5 The performance evaluation of the adopted ML classifiers in the following metrics: precision, recall (sensitivity), F1 score, and accuracy

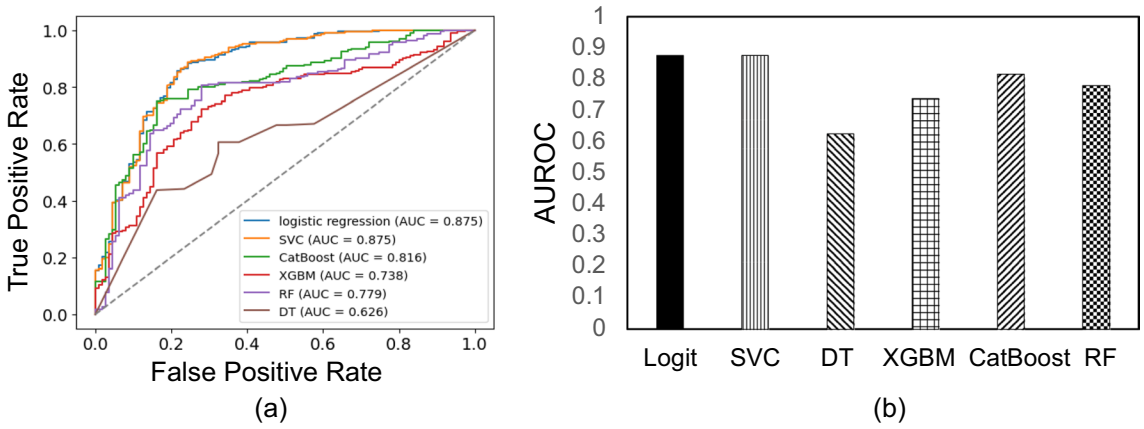

Fig. 6 Comparison of AUROC of all the ML models adopted in this study. **a** AUROC curves and (**b**) maximum AUROC values of the adopted ML models

false positive rate, helping to evaluate model performance. Logistic regression and SVC had the highest AUC values at 0.88, indicating they outperform the ensemble models—CatBoost (0.81), XGBM (0.73), and RF (0.78) (Fig. 6b). The DT performed poorly again with an AUC of 0.61 (Fig. 6b). The higher AUROC values for logistic regression and SVC highlight their effectiveness in predicting fracture risk.

While performance metrics give an overview of the models' effectiveness, understanding how various features impact predictions is crucial. SHAP plots (Fig. 7) highlight important features and their influence on model predictability. The x-axis shows SHAP (impact) values, and the y-axis lists features in descending order of impact. The distribution of red and blue dots indicates the directionality of features' impact. SHAP values thus help determine the significance of each feature in the prediction results.

The SHAP plots for logistic regression and SVC indicate that weight, angle β (loading direction) on the transverse plane, and age are the most significant features for these models. Conversely, the SHAP plots for CatBoost and XGBM highlight BMD as the key feature, followed by others. For RF, the most important features are FNA, weight, and FNW, followed by age and β. The decision tree SHAP plot suggests that β, FNA and sex are the most critical features. SHAP values help us interpret which features have the most significant impact on model output.

4 Discussion

This research evaluates hip fracture risk from sideways falls in the elderly using an ML pipeline integrated with QCT-based FEA. We present a novel approach combining clinical, anatomical, loading, and 3D biomechanical data through various ML classification algorithms to predict hip fractures. QCT-based FEA [39] was used to construct the dataset and establish ground truth, as it provides reliable data with over 95% correlation to experimental results for femur strength [17, 64–66]. Comparing precision, recall, F1 score, and accuracy across six ML models, logistic regression and SVC proved more accurate in predicting fracture risk than CatBoost, XGBM, RF, and DT. Logistic regression and SVC achieved 82% accuracy, effectively predicting both fracture and non-fracture cases.

AUROC is more pertinent over accuracy in binary classification, especially misclassification costs are unclear, or classes are imbalanced. The AUROC value of 0.87 for logistic regression and SVC indicates an 87% likelihood of correctly classifying a random patient's fracture risk (positive or negative cases with no fracture risk). Since the ML models are suitable for screening purposes, it ensures accurate detection of patients with fracture risk.

The high performance of logistic regression and SVC is attributed to the regularization parameter "C", which focuses on critical features and avoids overfitting by minimizing weights for less important features. Lower accuracy in other models may result from the continuous variables in the limited dataset. DT models can lose information when continuous data is discretized during splitting [67]. Ensemble models like CatBoost, XGBM, and RF, which are based on DT, also face the similar issue. Among them, CatBoost performs the best, likely due to its effective handling of categorical variables such as sex and femur side.

For interpretability, we used SHAP plots to identify key features and their impact on fracture risk predictions (Fig. 7). SHAP values reveal the contribution of individual features to predictions. CatBoost analysis shows significant learning from variables such as β, the loading angle on the coronal plane. Logistic regression and SVC highlight that obesity, age, and increased rotation angle during a fall significantly affect hip fracture risk. SHAP values provide insights into feature contributions for individual predictions,

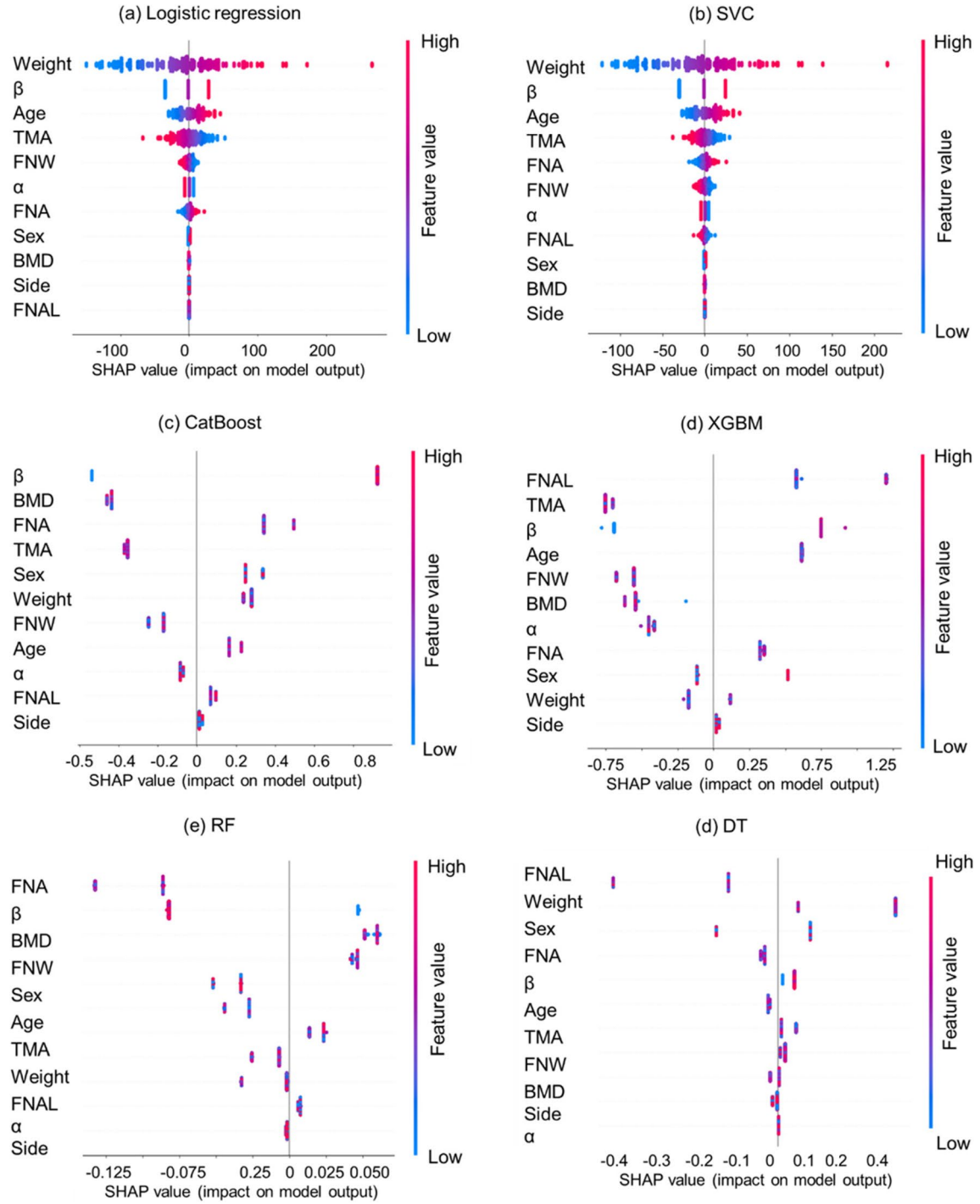

Fig. 7 The waterfall plot of the SHAP values to interpret the most contributed features towards the prediction of hip fracture risk. Red and blue color mean higher and lower value of a feature, respectively

offering more detail than techniques that only show aggregated results.

Furthermore, the feature importance shown in the SHAP plots of logistic regression and SVC align with statistical analysis that was conducted with the ground truth (QCT-based FEA) data to identify the effect of variables on fracture assessment [37]. However, BMD, though important, appears less significant in these models. This disparity can be minimized by increasing the dataset variance in BMD due to considering different feature importance SHAP plots for other ML models show different feature impacts, leading to varied performance metrics.

Although the proposed data-driven frameworks show encouraging outcomes, there are some limitations in this study. The ML-driven framework was built upon the data obtained from one geographical region, Canada, limiting its versatility. The models need to be trained with data from different ethnicities and geographical locations to make the ML models more robust. The current dataset may not truly reflect the elderly population as it includes patients well below 65 years of age. A dataset of older patients aged 65 and above may train the ML model more appropriately for the population group. However, the inclusion of younger adults increases the variability, which might be beneficial for developing a generalized ML model. Most importantly, a dataset with clinically diagnosed osteoporosis or information of prior fractures will increase the efficiency of model training. The major limitation of this work is the smaller dataset, which restricts achieving higher accuracy.

5 Conclusions

We have demonstrated the use of ML classifiers integrated with QCT-based FEA for predicting hip fracture risk in the elderly. Classical QCT-based FEA involves expensive and time-consuming 3D femur reconstruction, limiting its clinical use. Data driven modeling can be an alternative to overcome the limitations, while achieving acceptable predictive accuracy in a timely manner. In this work, we have investigated the application of different ML models and evaluated their performance in detecting fracture risk utilizing image based scientific computation. The outcomes of this work substantiate the viability of AI/ML in predicting fracture risk and will act as a fundamental work for future research directions. While ML models have proven effective in various biomedical applications, their application to fracture risk assessment shows promise. The superior performance of logistic regression, SVC, and CatBoost with limited datasets indicates their potential for accurate fracture risk prediction. Additionally, the modeling framework interpreted/identified the factors influencing hip or femoral fracture. This will ensure proper prevention and control as well as treatment plan to mitigate or reduce the occurrence of hip fracture. These research efforts have the potential to significantly advance the field of AI in healthcare, particularly in the areas of AI/ML, and biomedical engineering, leading to more accurate diagnoses, personalized treatment plans, and improved patient outcomes.

Author contribution Tanvir Faisal (T.F.) conceived the idea, and T.F., and Rabina Awal (R.A.) designed the study. Sarah Doll (S.D.) conducted image processing and 3D model generation, R.A. developed the computational model and conducted simulation. T.F., M.N., and R.A. conducted the data analysis and interpretation of data. All authors discussed the results and contributed to the initial drafting of this manuscript. T.F. and M.N. reviewed and approved the final manuscript.

Funding No funding was received for conducting this study.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Cooper C, Cole Z, Holroyd C, Earl S, Harvey NC, Dennison EM, et al. Secular trends in the incidence of hip and other osteoporotic fractures. Osteoporos Int. 2011;22:1277–88.
2. Johnell O, Kanis J. An estimate of the worldwide prevalence, mortality and disability associated with hip fracture. Osteoporos Int. 2004;15:897–902.
3. Kannus P, Parkkari J, Sievänen H, Heinonen A, Vuori I, Järvinen M. Epidemiology of hip fractures. Bone. 1996;18(1):S57–63.
4. Aschkenasy MT, Rothenhaus TC. Trauma and falls in the elderly. Emerg Med Clin. 2006;24(2):413–32.
5. Kanis J, Johnell O, Odén A, Johansson H, McCloskey E. FRAX™ and the assessment of fracture probability in men and women from the UK. Osteoporos Int. 2008;19(4):385–97.
6. Adams JE. Advances in bone imaging for osteoporosis. Nat Rev Endocrinol. 2013;9(1):28.
7. Albertsson D, Mellström D, Petersson C, Thulesius H, Eggertsen R. Hip and fragility fracture prediction by 4-item clinical risk score and mobile heel BMD: a women cohort study. BMC Musculoskelet Disord. 2010;11(1):55.
8. Luo Y, Ferdous Z, Leslie W. A preliminary dual-energy X-ray absorptiometry-based finite element model for assessing osteoporotic hip fracture risk. Proc Inst Mech Eng [H]. 2011;225(12):1188–95.
9. Aldieri A, Terzini M, Audenino AL, Bignardi C, Morbiducci U. Combining shape and intensity dxa-based statistical approaches for osteoporotic HIP fracture risk assessment. Comput Biol Med. 2020;127:104093.
10. Ferizi U, Besser H, Hysi P, Jacobs J, Rajapakse CS, Chen C, et al. Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from MRI data. J Magn Reson Imaging. 2019;49(4):1029–38.
11. Singh S, Mogra S, Shetty VS, Shetty S, Philip P. Three-dimensional finite element analysis of strength, stability, and stress distribution in orthodontic anchorage: a conical, self-drilling miniscrew implant system. Am J Orthod Dentofac Orthop. 2012;141(3):327–36.
12. Post A, Kendall M, Koncan D, Cournoyer J, Hoshizaki TB, Gilchrist MD, et al. Characterization of persistent concussive syndrome using injury reconstruction and finite element modelling. J Mech Behav Biomed Mater. 2015;41:325–35.
13. McCulloch A, Guccione J, Waldman L, Rogers J. Large-scale finite element analysis of the beating heart. High-Perform Comput Biomed Res. 2020;27–49. https://doi.org/10.1201/9781003068136-3
14. Faisal TR, Luo Y. Study of the variations of fall induced hip fracture risk between right and left femurs using CT-based FEA. Biomed Eng Online. 2017;16(1):116. https://doi.org/10.1186/s12938-017-0407-y.

15. Bettamer A. Prediction of proximal femur fracture: finite element modeling based on mechanical damage and experimental validation. 2012.
16. Liu Y, Zhang A, Wang C, Yin W, Wu N, Chen H, et al. Biomechanical comparison between metal block and cement-screw techniques for the treatment of tibial bone defects in total knee arthroplasty based on finite element analysis. Comput Biol Med. 2020;125:104006.
17. Black DM, Bouxsein ML, Marshall LM, Cummings SR, Lang TF, Cauley JA, et al. Proximal femoral structure and the prediction of hip fracture in men: a large prospective study using QCT. J Bone Mineral Res. 2008;23(8):1326–33.
18. Dragomir-Daescu D, Buijs JOD, McEligot S, Dai Y, Entwistle RC, Salas C, et al. Robust QCT/FEA models of proximal femur stiffness and fracture load during a sideways fall on the hip. Ann Biomed Eng. 2011;39:742–55.
19. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944.
20. Sajjad M, Khan S, Muhammad K, Wu W, Ullah A, Baik SW. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. J Comput Sci. 2019;30:174–82.
21. Fathima AJ, Fasla MN. A comprehensive review on heart disease prognostication using different artificial intelligence algorithms. Comput Methods Biomech Biomed Eng. 2024;27(11):1357–74.
22. Iliou T, Anagnostopoulos C-N, Anastassopoulos G. Osteoporosis detection using machine learning techniques and feature selection. Int J Artif Intell Tools. 2014;23(05):1450014.
23. Kong SH, Ahn D, Kim B, Srinivasan K, Ram S, Kim H, et al. A novel fracture prediction model using machine learning in a community-based cohort. JBMR Plus. 2020;4(3):e10337.
24 Zhou Y, Klintström E, Klintström B, Ferguson SJ, Helgason B, Persson C. A convolutional neural network-based method for the generation of super-resolution 3D models from clinical CT images. Comput Methods Prog Biomed. 2024;245:108009.
25 Sultana J, Naznin M, Faisal TR. SSDL-an automated semi-supervised deep learning approach for patient-specific 3D reconstruction of proximal femur from QCT images. Med Biol Eng Comput. 2024. https://doi.org/10.1007/s11517-023-03013-8.
26. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol. 2019;48(2):239–44.
27. Kim D, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73(5):439–45.
28. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci. 2018;115(45):11591–6.
29. Liu Q, Cui X, Chou Y-C, Abbod MF, Lin J, Shieh J-S. Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders. Biomed Signal Process Control. 2015;21:146–56.
30. Memiş A, Varlı S, Bilgili F. Image based quantification of the proximal femur shape deformities in 3D by using the contralateral healthy shape structure: a preliminary study. Biomed Signal Process Control. 2022;71:103079.
31. Do BH, Langlotz C, Beaulieu CF. Bone tumor diagnosis using a naïve Bayesian model of demographic and radiographic features. J Digit Imaging. 2017;30(5):640–7.
32. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One. 2017;12(6):e0178992.
33 Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep. 2018;8(1):1727.
34. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019;2(1):1–10.
35. Villamor E, Monserrat C, Del Río L, Romero-Martín J, Rupérez MJ. Prediction of osteoporotic hip fracture in postmenopausal women through patient-specific FE analyses and machine learning. Comput Methods Programs Biomed. 2020;193:105484.
36. Ferizi U, Besser H, Hysi P, Jacobs J, Rajapakse CS, Chen C, et al. Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from MRI data. J Magn Reson Imaging. 2019;49(4):1029–38.
37. Awal R, Ben Hmida J, Luo Y, Faisal T. Study of the significance of parameters and their interaction on assessing femoral fracture risk by quantitative statistical analysis. Med Biol Eng Compu. 2022;60(3):843–54. https://doi.org/10.1007/s11517-022-02516-0.
38. Awal R, Faisal TR. Multiple regression analysis of hip fracture risk assessment via finite element analysis. J Eng Sci Med Diagn Ther. 2021;4(1):011006.
39. 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;150:106299. https://doi.org/10.1016/j.jmbbm.2023.106299.
40. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003;2(3 Suppl):S75-83.
41. Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. Adv Exp Med Biol. 2011;696:191–9. https://doi.org/10.1007/978-1-4419-7046-6_19.
42. Cao Y, Geddes TA, Yang JYH, Yang P. Ensemble deep learning in bioinformatics. Nat Mach Intell. 2020;2(9):500–8.
43. Bartoszewicz JM, Seidel A, Rentzsch R, Renard BY. DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. Bioinformatics (Oxford, England). 2020;36(1):81–9. https://doi.org/10.1093/bioinformatics/btz541.
44. Cao Y, Liu L, Chen X, Man Z, Lin Q, Zeng X, et al. Segmentation of lung cancer-caused metastatic lesions in bone scan images using self-defined model with deep supervision. Biomed Signal Process Control. 2023;79:104068.
45. Hu Y, Zhao L, Li Z, Dong X, Xu T, Zhao Y. Classifying the multi-omics data of gastric cancer using a deep feature selection method. Expert Syst Appl. 2022;200:116813.
46. Lorente D, Martínez-Martínez F, Rupérez MJ, Lago M, Martínez-Sober M, Escandell-Montero P, et al. A framework for modelling the biomechanical behaviour of the human liver during breathing in real time using machine learning. Expert Syst Appl. 2017;71:342–57.
47. Faisal TR, Luo Y. Study of stress variations in single-stance and sideways fall using image-based finite element analysis. Bio-Med Mater Eng. 2016;27(1):1–14.
48 Ford CM, Keaveny TM, Hayes WC. The effect of impact direction on the structural capacity of the proximal femur during falls. J Bone Miner Res. 1996;11:377–283.
49. Schileo E, Taddei F, Cristofolini L, Viceconti M. Subject-specific finite element models implementing a maximum principal strain criterion are able to estimate failure risk and fracture location on human femurs tested in vitro. J Biomech. 2008;41(2):356–67.
50. Altai Z, Qasim M, Li X, Viceconti M. The effect of boundary and loading conditions on patient classification using finite element predicted risk of fracture. Clin Biomech. 2019;68:137–43.
51. Marco M, Giner E, Caeiro-Rey JR, Miguélez MH, Larraínzar-Garijo R. Numerical modelling of hip fracture patterns in human femur. Comput Methods Programs Biomed. 2019;173:67–75.
52. Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Stat Methodol. 1958;20(2):215–32.

53. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Springer; 2009.
54. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Adv Neural Inf Process Syst. 1996;28(7):779–84.
55. Tsochantaridis I, Hofmann T, Joachims T, Altun Y. Support vector machine learning for interdependent and structured output spaces. Proceedings of the Twenty-First International Conference on Machine Learning 2004. p. 104.
56. Chang Y-C, Chang K-H, Wu G-J. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Appl Soft Comput. 2018;73:914–20.
57. Al DE. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. Int J Comput Inf Eng. 2019;13(1):6–10.
58. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40.
59. Erickson BJ, Kitamura F. Magician's corner: 9 Performance metrics for machine learning models. Radiol Soc North Am. 2021;3(3):e200126.
60. Murphy E, Ehrhardt B, Gregson CL, von Arx O, Hartley A, Whitehouse M, et al. Machine learning outperforms clinical experts in classification of hip fractures. Sci Rep. 2022;12(1):2058.
61. Van Rossum G, Drake FL. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform; 2009.
62. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.
63 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72.
64. Engelke K, van Rietbergen B. Zysset PJCrib, metabolism m. FEA Meas Bone Strength: Rev. 2016;14:26–37.
65. Dragomir-Daescu D, Buijs JOD, McEligot S, Dai Y, Entwistle RC, Salas C, Melton LJ, Bennet KE, Khosla S, Amin S. Robust QCT/FEA models of proximal femur stiffness and fracture load during a sideways fall on the hip. Ann Biomed Eng. 2011;39(2):742–55.
66. Cody DDGG, Hou FJ, Spencer HJ, Goldstein SA. Fyhrie DP Femoral strength is better predicted by finite element models than QCT and DXA. J Biomechanics. 1990;32:1013–20.
67. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002;35(5–6):352–9.