European Radiology (2019) 29:2207–2217
https://doi.org/10.1007/s00330-018-5846-8

COMPUTED TOMOGRAPHY

Vertebral body insufficiency fractures: detection of vertebrae at risk on standard CT images using texture analysis and machine learning

Urs J. Muehlematter[1] · Manoj Mannil[1] · Anton S. Becker[1] · Kerstin N. Vokinger[2,3] · Tim Finkenstaedt[1] · Georg Osterhoff[4] · Michael A. Fischer[5] · Roman Guggenberger[1]

Received: 12 August 2018 / Revised: 30 September 2018 / Accepted: 22 October 2018 / Published online: 5 December 2018

Abstract

Purpose To evaluate the diagnostic performance of bone texture analysis (TA) combined with machine learning (ML) algorithms in standard CT scans to identify patients with vertebrae at risk for insufficiency fractures.

Materials and methods Standard CT scans of 58 patients with insufficiency fractures of the spine, performed between 2006 and 2013, were analyzed retrospectively. Every included patient had at least two CT scans. Intact vertebrae in a first scan that either fractured ("unstable") or remained intact ("stable") in the consecutive scan were manually segmented on mid-sagittal reformations. TA features for all vertebrae were extracted using open-source software (MaZda). In a paired control study, all vertebrae of the study cohort "cases" and matched controls were classified using ROC analysis of Hounsfield unit (HU) measurements and supervised ML techniques. In a within-subject vertebra comparison, vertebrae of the cases were classified into "unstable" and "stable" using identical techniques.

Results One hundred twenty vertebrae were included. Classification of cases/controls using ROC analysis of HU measurements showed an AUC of 0.83 (95% confidence interval [CI], 0.77–0.88), and ML-based classification showed an AUC of 0.97 (CI, 0.97–0.98). Classification of unstable/stable vertebrae using ROC analysis showed an AUC of 0.52 (CI, 0.42–0.63), and ML-based classification showed an AUC of 0.64 (CI, 0.61–0.67).

Conclusion TA combined with ML allows to identifying patients who will suffer from vertebral insufficiency fractures in standard CT scans with high accuracy. However, identification of single vertebra at risk remains challenging.

Key Points

- *Bone texture analysis combined with machine learning allows to identify patients at risk for vertebral body insufficiency fractures on standard CT scans with high accuracy.*
- *Compared to mere Hounsfield unit measurements on CT scans, application of bone texture analysis combined with machine learning improve fracture risk prediction.*
- *This analysis has the potential to identify vertebrae at risk for insufficiency fracture and may thus increase diagnostic value of standard CT scans.*

Keywords Spine · Machine learning · Osteoporosis · Tomography, X-ray computed

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-018-5846-8) contains supplementary material, which is available to authorized users.

✉ Urs J. Muehlematter
urs.muehlematter@usz.ch

[1] Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Ramistrasse 100, 8091 Zurich, Switzerland

[2] University Hospital of Zurich, Zurich, Switzerland

[3] University of Zurich, Zurich, Switzerland

[4] Department of Trauma, University Hospital Zurich, Zurich, Switzerland

[5] Department of Radiology, University Hospital Balgrist, University of Zurich, Zurich, Switzerland

Abbreviations

ANN	Artificial neural networks
BMD	Bone mineral density
CCC	Concordance correlation coefficient
DXA	Dual-energy X-ray absorptiometry
FEA	Finite element analysis
GLCM	Gray-level co-occurrence matrix
GLRLM	Gray-level run-length matrix
HR-pQCT	High-resolution peripheral quantitative computed tomography
IH	Image histogram
ML	Machine learning
MLP	Multi-layer perceptron
RF	Random forest
ROI	Region of interest
SVM	Support vector machine
TA	Texture analysis
TLJ	Thoracic-lumbar junction

Introduction

Vertebral compression fractures can have pathologic, traumatic, or atraumatic causes. The latter may occur after low stress due to bone mineralization loss leading to reduced mechanical bone strength [1]. The most common underlying causes are osteopenia and osteoporosis [2], two systematic skeletal disorders affecting especially the elderly and chronically ill. Osteopenia and osteoporosis are characterized by the loss of bone tissue, skeletal fragility, and microarchitectural deterioration [3].

Osteoporotic vertebral compression fractures affect many patients worldwide, entailing significant morbidity and mortality: In 2000, 1.4 million vertebral fractures were estimated globally and approximately 214,000 occurred in the USA with patients at age 50 or older [4]. The lifetime risk of a vertebral fracture at age 50 is 15.6% for women and 5% for men [5].

In general, the diagnosis of vertebral compression fractures is based on X-ray examinations. However, a substantial amount of cases may be overlooked [6]. CT or MR imaging may be more appropriate to detect subtle cases.

In clinical practice, fracture risk is usually determined by dual-energy X-ray absorptiometry (DXA) [1]. The trabecular bone score based on gray-level textural metric can be extracted from DXA images and improves the fracture risk assessment [7]. Several studies demonstrate that automated bone mineral densitometry based on Hounsfield unit (HU) on clinical CT images is feasible [8], and that it correlates well with DXA measurements [9]. Bone microstructure can be assessed ex vivo with microcomputed tomography and in vivo with high-resolution peripheral quantitative computed tomography (HR-pQCT) at comparable resolution [10, 11]. Combination of HR-pQCT and finite element analysis (FEA) determines stresses in human bones, permits highly accurate estimation of individual fracture risk, and predicts fracture sites [12]. Recent studies implemented FEA in clinical CT scans [13]. However, these techniques require dedicated hardware and software which may hamper their application in clinical practice. Texture analysis (TA), on the other hand, is an objective and quantitative method to analyze the distribution and relationship of pixel or voxel gray levels in an image or volume [14], which can be applied retroactively to standard CT scans. Feasibility of TA on bone structure has been demonstrated for radiographs [7, 15–17] and CT scans [18, 19].

We hypothesize that TA and machine learning (ML) allow to predicting vertebral insufficiency fracture in standard CT. The goal of this retrospective case-control study is to evaluate the diagnostic performance of bone TA combined with ML algorithms in standard CT scans to identify patients with vertebrae at risk for insufficiency fractures.

Materials and methods

Study approaches

In a paired control study (approach A), vertebrae from patients developing vertebral fractures (cases) were compared to vertebrae of matched controls with normal bone density. In a within-subject study (approach B), it was investigated if it is possible to predict whether or not vertebrae will fracture. In both approaches, vertebrae were classified using ROC analysis of HU measurements and supervised ML techniques. Figure 1 depicts the two separate approaches, A and B.

Study population

The study received institutional review board and local ethics committee approval. We identified 30,931 patients above 45 years undergoing a clinical CT scan that covered at least the thoracic or lumbar spine between January 2006 and December 2013 from the institutional PACS. Six hundred seventy patients were selected that received at least two CT scans within a year and a third scan at least 5 months after the second scan. The third scan was assessed for validation of the stable vertebra over a longer period. Eventually, 58 patients remained after a review for newly occurred osteoporotic vertebral compression fractures of the thoracic and/or lumbar spine using established criteria [20, 21] and excluding patients with traumatic fractures and metastasis of the spine (Fig. 2). These patients had two consecutive scans, showing intact vertebrae in the first scan that either fractured (“unstable” vertebra $n = 60$) or remained intact (“stable” vertebra $n = 60$).

We divided the spine into the following regions: thoracic spine (Th1-Th10), thoracic-lumbar junction (TLJ,

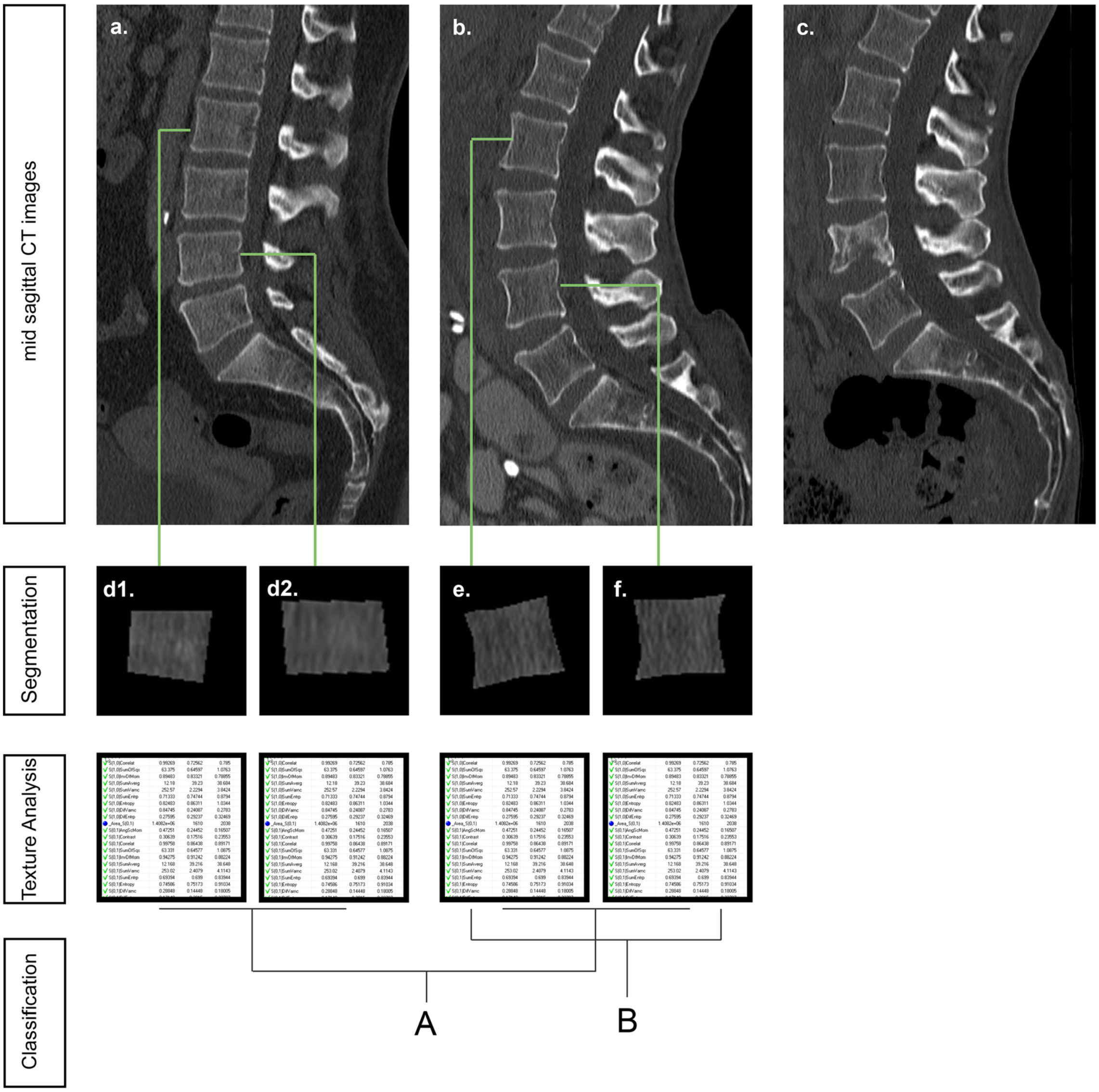

Fig. 1 Schematic shows our classification approaches to our study illustrated with an example. CT scan of a matched control (**a.**). Primary CT scan of a 59-year-old female subject (**b.**). Secondary CT scan of the same subject 4 months later with a newly occurred insufficiency fracture of vertebra L4 (**c.**). Cropped images of vertebra L2 (**d1.**) and L4 (**d2.**) of the matched control. Cropped image of vertebra L2 (**e.**) which remains intact in the secondary scan and of vertebra L4 (**f.**) which is broken in the secondary scan. Classification approach A is a paired control study comparing the study cohort (cases) with an external cohort presenting a normal BMD with DXA (controls). Approach B is a within-subject vertebra comparison between unstable vertebrae and intact reference vertebrae of the study cohort

Th11-L1) and lumbar spine (L2-L5). Stable and unstable vertebrae were selected from the same region. As a control set, images of 58 patients from a previous study with patients presenting a normal bone mineral density (BMD) with DXA were matched by age, sex, and region of the spine [18]. The mean age was 70 ± 9 years (range 48–90 years), including 26 women and 34 men in both the patient and the control groups. In the female subgroup, the mean age was 69 ± 10 years (range 48–89 years) and 71 ± 9 years (range 53–90 years) in the male subgroup. The age difference between these subgroups is not significant ($p = 0.519$). A flowchart of the patient selection is shown in Fig. 2.

CT data and post processing

Sagittal image stacks covering the spine were retrieved from the PACS and saved in uncompressed DICOM format. The images were acquired with different CT scanners: the SOMATOM-Force, Definition, Definition Flash, Definition AS and Sensation CT (all from Siemens). The collimation of the scanners ranged from 0.6 to 1.2 mm. The image section thickness of the sagittal images was 2.0 mm, and kilovoltage peak (kVp) ranged from 90 to 140. All images had been reconstructed using an edge-enhancing bone kernel and were rescaled to the coarsest in-plane resolution of 0.5 mm. No low-dose protocols were included.

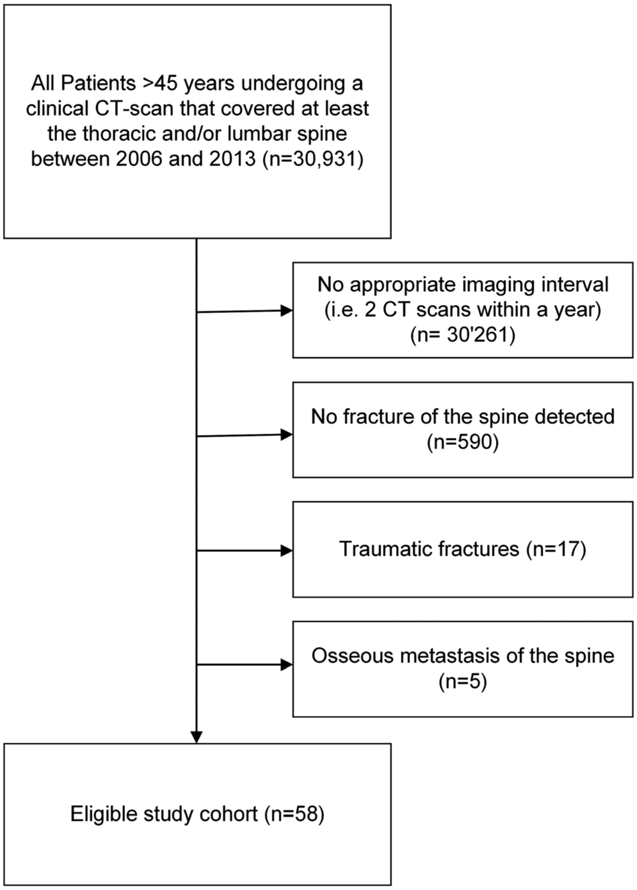
All Patients >45 years undergoing a clinical CT-scan that covered at least the thoracic and/or lumbar spine between 2006 and 2013 (n=30,931)
No appropriate imaging interval (i.e. 2 CT scans within a year) (n= 30'261)
No fracture of the spine detected (n=590)
Traumatic fractures (n=17)
Osseous metastasis of the spine (n=5)
Eligible study cohort (n=58)

Fig. 2 Flowchart of the patient selection process

Image classification

One radiologist (UJM, with 2 years of experience in skeletal radiology) manually segmented trabecular bone of stable and unstable vertebrae on single, mid-sagittal images of the primary scans, using freehand-drawn regions of interest (ROIs). TA was performed using open-source software (MaZda, version 4.6) [22] with the identical technique as with the control subjects [18]. The TA software calculated 305 features from 6 different statistical image descriptors (Supplements Table 1) for each ROI. Since ROI characteristics (e.g., location, size, shape) can influence texture features, only reproducible features (i.e., features with excellent intra- and inter-reader agreement (intraclass correlation coefficient (ICC) ≥ 0.81) that were defined in a previous study [18] were included. Mean CT HU values were obtained from the identical ROIs that were used for TA feature extraction.

Human readout

Two radiologists (RG, 15 years and ASB, 3 years of experience in skeletal radiology) visually rated vertebral trabecular bone texture with respect to number, length, and thickness of bone trabeculae, using a 5-point Likert-like scale (1, age-appropriate; 2, rather age-appropriate; 3, unsure; 4, rather age-inappropriate; 5, age-inappropriate bone texture) in approaches A and B, based on reported correlations of structural trabecular bone appearance and fracture risk [23–25]. In these vertebra-based analyses, the cropped vertebral images showing only the spongiosa were presented with identical windowing (HU width 1720, length 535) in random order as mid-sagittal reformations on a standard reporting workstation in two separate readout sessions for approaches A and B. Readers were allowed to change the window settings and take ROI measurements. Figure 1d–f shows examples of the cropped images.

Statistical analysis

The statistical analysis was performed in R version 3.4.2 (R Foundation for Statistical Computing). The mean and standard deviation of the mean, median, and interquartile range and range were used for descriptive statistics of continuous variables, where appropriate. The chi-square test was used to compare ordinal and nominal protocol parameters. The Mann-Whitney *U* test was used to investigate the influence of protocol settings on TA features. Tenfold cross validation with stratified sampling was used with 1/3 of the data as test set and 2/3 as training set. The features and folds were consistent across all classifiers. Data standardization using data scaling and data centering and removal of redundant features (Pearson correlation coefficient $R \geq 0.80$) were used as pre-processing. Selected features were compared by using the Wilcoxon test and co-correlation was assessed with Pearson correlation. For classification, the following ML classifiers from the caret package version 6.0-77 [26] were used: multi-layer perceptron (MLP) with 3 hidden layers, artificial neural networks (ANN) with a single hidden layer, random forest (RF), support vector machine (SVM) with linear kernel and naïve Bayesian classifier. Feature importance was calculated for ANN using the method of Gevrey et al [27], for RF using permutation, and for SVM ROC analysis. ROC analysis was performed for ML classifiers and the human readers for approaches A and B. Diagnostic accuracy was expressed as the AUC. The nonparametric test by DeLong et al was used to compare AUCs [28]. $P < 0.05$ was considered indicative for significant differences, with Bonferroni correction for multiple comparisons where appropriate. All tests were two-tailed. Inter-reader agreement was assessed using the concordance correlation coefficient (CCC) [29].

Results

Study population

Sixty stable and 60 unstable vertebrae of 58 patients were included. Twenty-two vertebrae of the thoracic spine, 24 of the TLJ, and 14 of the lumbar spine region were affected. Reasons for referral for the primary CT scan were oncologic diseases ($n = 23$), status after vascular repair ($n = 22$), acute gastrointestinal problems ($n = 7$), non-oncologic diseases of the lung ($n = 5$), and suspected fracture ($n = 3$). The mean time difference between the primary and secondary scan was 237 ± 117.5 days, between latter and third 548 ± 386 days.

CT data and post processing

We found no significant difference in contrast agent administration among case and control cohort ($p = 0.432$). However, there was a significant difference in tube voltage among case and control cohort with higher tube voltage in the case cohort ($p = 0.019$). Figure 3 depicts the used kVp in the study and control cohort. Furthermore, we found no significant influence of the tube voltage on the selected features within the study cohort (U ranged from 1968 to 3419, p ranged from 0.184 to 1) or the control cohort (U ranged from 5269 to 6667, p ranged from 0.120 to 1).

Image classification

After removal of 88/305 (28.9%) features with poor reproducibility and 188/217 (86.6%) redundant features, 29 features were considered for classification in both approaches, A and B. In approach A, 16/29 (55.2%) texture features showed a significant difference between cases and controls after Bonferroni correction. There were no significant differences between selected texture features in approach B after correction. Table 1 depicts means and differences of all selected features.

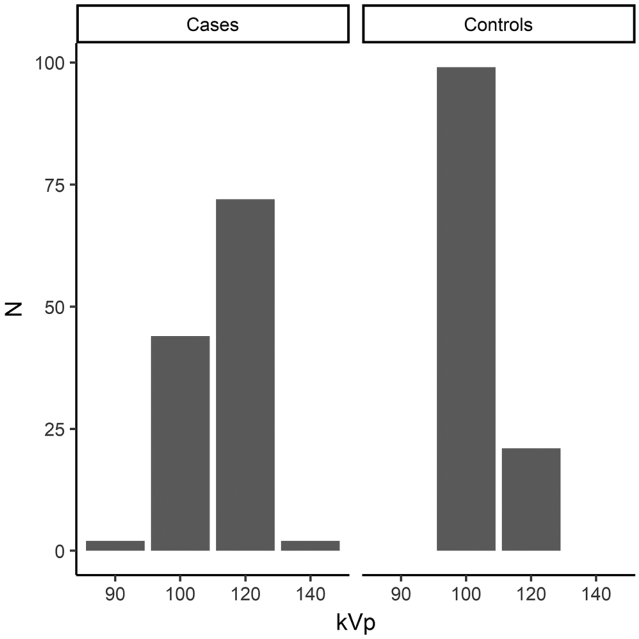

Fig. 3 Barplots show the distribution of applied kilovoltage peak (kVp) in the case and control cohort

In approach A, ROC analysis using mean HU values for classification yielded AUC of 0.83 (95% confidence interval [CI], 0.77–0.88). All ML classifier yielded higher accuracy (AUC ranging from 0.88 to 0.96, Table 2), and the highest test accuracy was achieved using SVM with AUC of 0.97 (CI, 0.96–0.98). There were no significant differences between AUC of SVM and RF ($p = 0.07$). AUC of SVM was significantly higher compared to MLP, ANN, and naïve Bayesian classifier ($p = 0.029$, $p < 0.001$, and $p < 0.001$, respectively). AUC using SVM and texture features was significantly higher than ROC analysis using mean HU values ($p < 0.001$). ROC curves of ROC analysis and ML classification are shown in Fig. 4. Figure 5 demonstrates estimated feature importance for the MLP, RF, and SVM classifiers. Of the combined top 10 important features of all ML classifiers, 3/17 (17.6%) belong to the image histogram (IH)-, 2/17 (11.8%) to the gray-level run-length matrix (GLRLM)-, 4/17 (23.5%) to gray-level co-occurrence matrix (GLCM)-, 3/17 (17.6%) to autoregressive model (AR)-, and 5/17 (29.5%) to wavelet transformation-derived features, respectively. Combined top 10 important features showed low co-correlations as demonstrated in Fig. 6.

In approach B, ROC analysis using mean HU values for classification yielded AUC of 0.52 (CI, 0.42–0.63). All ML showed low accuracy (AUC ranging from 0.50 to 0.64), and the highest test accuracy was achieved using SVM with AUC of 0.64 (CI, 0.61–0.67). SVM classified stable and unstable vertebra significantly better than mean HU values ($p = 0.029$).

Human readout

Diagnostic accuracy in approach A was low for both readers with AUC of 0.57 (CI, 0.50–0.63) for reader 1 and 0.48 (CI, 0.41–0.55) for reader 2. However, reader 1 classified cases versus control images significantly better than reader 2 ($p = 0.03$), but only slightly better than chance. Both readers showed random within-subject classification of stable versus unstable vertebrae (reader 1 AUC 0.53, CI 0.44–0.62 and reader 2 AUC 0.49, CI 0.39–0.59), with no significant difference between the readers ($p = 0.569$). Inter-reader agreement was low for both approaches A (CCC 0.12, CI 0.04–0.21) and B (CCC 0.15, CI 0.05–0.25). Table 2 summarizes test accuracy (AUC) of automated and human reader image classification.

Table 1 Differences of means of selected texture features for approaches A and B

		Approach A			Approach B		
TA feature	TA feature class	Cases	Controls	*p* value	Unstable	Stable	*p* value
Mean	IH	1132.178 (± 50.367)	1184.546 (± 31.301)	< 0.001	1132.178 (± 49.286)	1132.758 (± 54.64)	1
MinNorm	IH	959 (± 76.354)	1020 (± 90.439)	< 0.05	955.5 (± 74.871)	961.5 (± 76.354)	1
Variance	IH	3202.415 (± 1595.107)	2734.304 (± 2298.383)	1	3253.623 (± 1758.937)	3201.59 (± 1396.207)	1
Skewness	IH	0.201 (± 0.228)	0.206 (± 0.375)	1	0.201 (± 0.292)	0.201 (± 0.194)	1
Kurtosis	IH	0.235 (± 0.365)	0.102 (± 0.294)	1	0.289 (± 0.4)	0.19 (± 0.33)	1
S01SumVarnc	GLCM	408.016 (± 21.081)	415.64 (± 16.181)	< 0.001	406.256 (± 22.22)	408.921 (± 20.962)	1
S01SumEntrp	GLCM	1.897 (± 0.015)	1.893 (± 0.018)	1	1.897 (± 0.014)	1.897 (± 0.015)	1
S50InvDfMom	GLCM	0.091 (± 0.011)	0.099 (± 0.015)	< 0.05	0.093 (± 0.01)	0.09 (± 0.012)	1
S50SumAverg	GLCM	64.588 (± 0.818)	64.703 (± 1.302)	1	64.49 (± 0.65)	64.626 (± 0.923)	1
S50DifVarnc	GLCM	67.861 (± 9.588)	64.117 (± 8.976)	< 0.05	65.627 (± 11)	69.195 (± 8.901)	1
S05InvDfMom	GLCM	0.1 (± 0.013)	0.105 (± 0.016)	< 0.001	0.098 (± 0.011)	0.1 (± 0.013)	1
S05SumEntrp	GLCM	1.816 (± 0.023)	1.831 (± 0.029)	0.8	1.813 (± 0.024)	1.818 (± 0.019)	1
S55AngScMom	GLCM	0.001 (± 0)	0.001 (± 0)	< 0.001	0.001 (± 0)	0.001 (± 0)	1
S55InvDfMom	GLCM	0.086 (± 0.01)	0.087 (± 0.007)	1	0.086 (± 0.011)	0.086 (± 0.009)	1
S55SumEntrp	GLCM	1.776 (± 0.02)	1.786 (± 0.025)	0.5	1.776 (± 0.021)	1.776 (± 0.018)	1
S55DifVarnc	GLCM	73.113 (± 8.342)	71.741 (± 15.371)	1	71.724 (± 11.908)	74.966 (± 6.153)	1
HorzlFraction	GLRLM	0.947 (± 0.015)	0.941 (± 0.014)	< 0.001	0.947 (± 0.014)	0.948 (± 0.016)	1
dgr45ShrtREmp	GLRLM	0.965 (± 0.012)	0.96 (± 0.008)	< 0.05	0.964 (± 0.011)	0.966 (± 0.013)	1
dgr135GLevNonU	GLRLM	51.023 (± 20.339)	37.604 (± 22.356)	< 0.001	52.48 (± 19.787)	50.515 (± 21.215)	1
GrSkewness	AG	0.847 (± 0.212)	0.865 (± 0.23)	1	0.802 (± 0.229)	0.869 (± 0.179)	1
Teta1	AR	0.543 (± 0.142)	0.61 (± 0.163)	0.9	0.524 (± 0.121)	0.557 (± 0.164)	1
Teta3	AR	0.823 (± 0.133)	0.884 (± 0.061)	< 0.001	0.834 (± 0.12)	0.812 (± 0.147)	1
Teta4	AR	0.047 (± 0.044)	0.038 (± 0.052)	1	0.047 (± 0.041)	0.048 (± 0.05)	1
WavEnLHs_1	WT	75.723 (± 49.188)	55.361 (± 20.707)	< 0.001	77.161 (± 49.99)	75.723 (± 48.876)	1
WavEnHHs_1	WT	19.15 (± 17.078)	13.278 (± 10.385)	< 0.05	19.15 (± 16.68)	18.384 (± 17.619)	1
WavEnHLs_2	WT	397.498 (± 97.826)	309.857 (± 78.727)	< 0.001	400.063 (± 105.281)	396.666 (± 93.921)	1
WavEnHHs_3	WT	92.76 (± 25.297)	76.854 (± 48.12)	1	93.771 (± 26.788)	90.588 (± 27.559)	1
WavEnLLs_4	WT	20,355.787 (± 2000.417)	18,944.148 (± 2238.257)	< 0.001	20,000.831 (± 1626.904)	20,652.393 (± 2024.275)	1
WavEnHHs_4	WT	77.993 (± 49.635)	93.434 (± 75.329)	< 0.05	77.861 (± 48.673)	78.136 (± 50.548)	1

Values expressed as median (± median absolute deviation). All comparisons performed with controlling for multiple comparisons

TA texture analysis, *IH* image histogram, *GLCM* gray-level co-occurrence matrix, *GLRLM* gray-level run-length matrix, *AG* absolute gradient, *AR* autoregressive model, *WT* wavelet transformation

Table 2 Classification and human readout results

	AUC approach A (CI)	AUC approach B (CI)
ROC analysis		
HU measurements	0.83 (0.77–0.88)	0.52 (0.42–0.63)
ML classification with texture features		
RF	0.96 (0.96–0.97)	0.57 (0.55–0.59)
MLP	0.95 (0.94–0.97)	0.51 (0.46–0.57)
ANN	0.93 (0.93–0.94)	0.55 (0.53–0.57)
SVM	0.97 (0.97–0.98)	0.64 (0.61–0.67)
Naïve Bayesian	0.88 (0.86–0.90)	0.50 (0.46–0.54)
Human readout on images		
Reader 1	0.57 (0.50–0.63)	0.53 (0.44–0.62)
Reader 2	0.48 (0.41–0.55)	0.49 (0.39–0.59)

CI 95% confidence interval, *ML* machine learning, *RF* random forest, *MLP* multi-layer perceptron, *ANN* artificial neural network, *SVM* support vector machine, *Naïve Bayesian* naïve Bayesian classifier

Discussion

In this study, we investigated the use of TA combined with ML in clinical CT scans for the differentiation of patients as well as vertebrae at risk for fracture. We found that an SVM classifier utilizing 29 texture features yielded a high AUC of 0.97 for identifying patients at risk for insufficiency fractures. Differentiation using mean CT density measurements alone yielded an AUC of only 0.83 ($p < 0.001$). Given the same data, an experienced radiologist reached a performance only slightly better than chance (AUC = 0.57). However, neither TA/ML nor the radiologists could reliably distinguish between the single vertebral bones at risk for fracture and the neighboring segments in the same individuals.

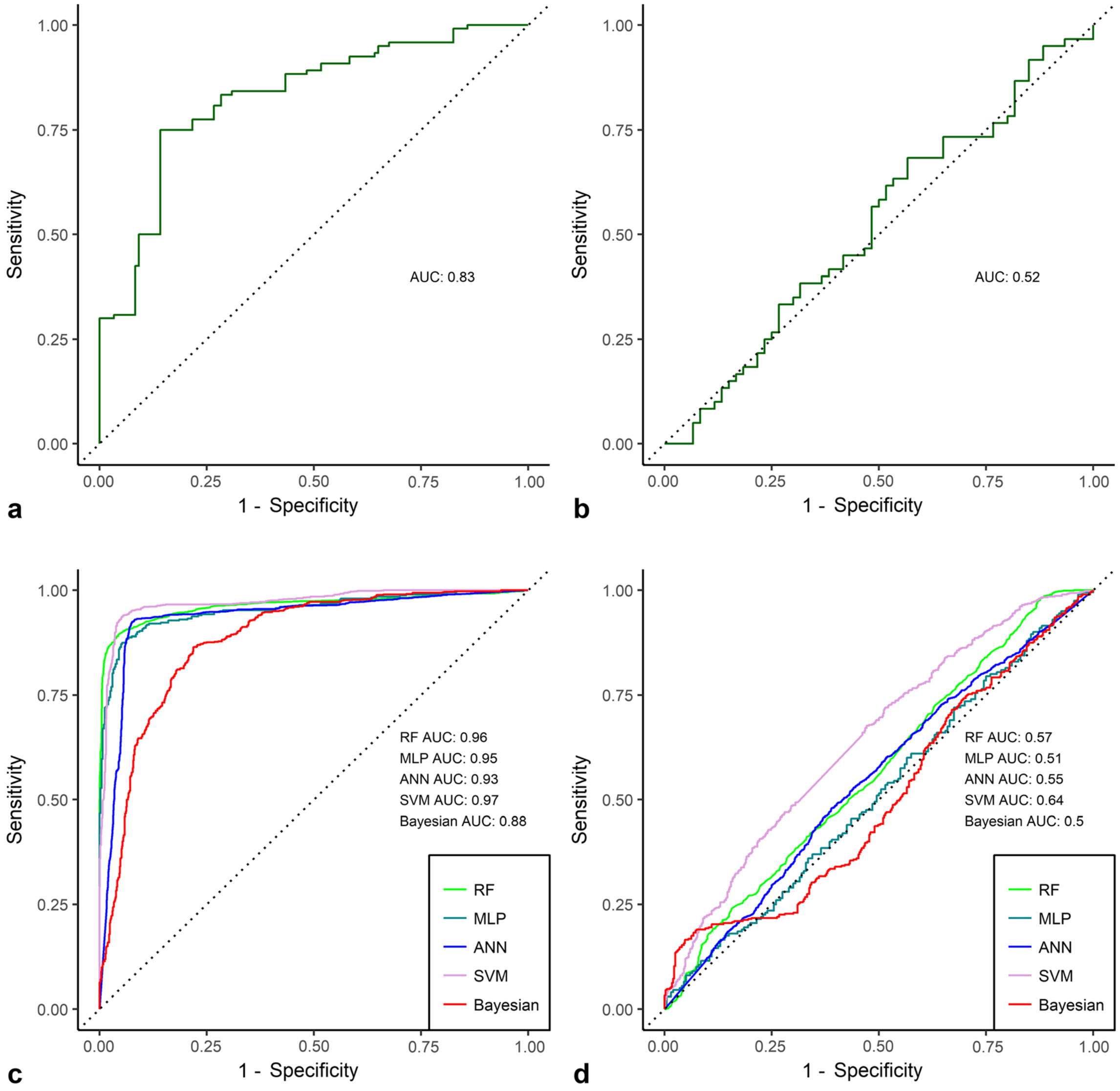

Fig. 4 Upper row shows receiver operating characteristic performance for classification in case/control (**a**) and stable/unstable (**b**) vertebrae using Hounsfield units measurements solely. Lower row shows receiver operating characteristic performance for classification in case/control (**c**) and stable/unstable (**d**) vertebrae using 29 texture analysis features and 5 machine learning classifiers

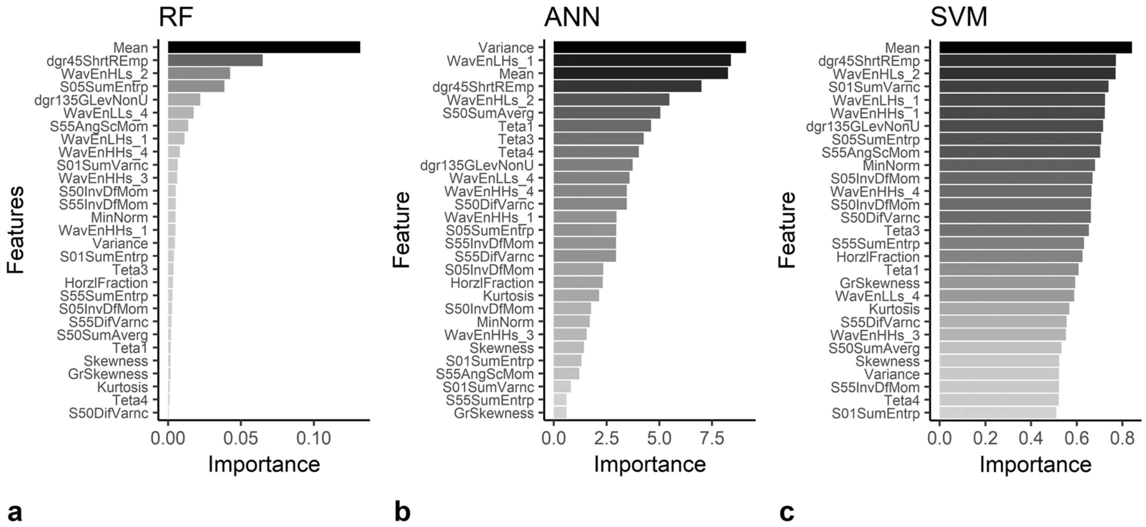

Fig. 5 Barplots show estimated texture feature importance for classification in case/control vertebrae for random forest classifier (**a**), artificial neural network classifier (**b**), and support vector machine classifier (**c**)

Increased computational power and the successful development of new algorithms in the last decades have led to promising approaches in various fields of radiology [14, 30] and TA has been shown to be efficient in the differentiation of osteoporotic and healthy subjects in HR-pQCT data [31].

In our study, the combination of ML and TA revealed several important TA features when discriminating between healthy and diseased bones. The mean signal intensity crystallized is one of the most important factors in our analyses. This is also consistent with the literature, reporting that bone mineralization is an important factor for bone strength [1, 32]. However, as already previously suggested [15, 33, 34], the increased performance of models including TA features, as well as the missing correlation with the mean signal, supports

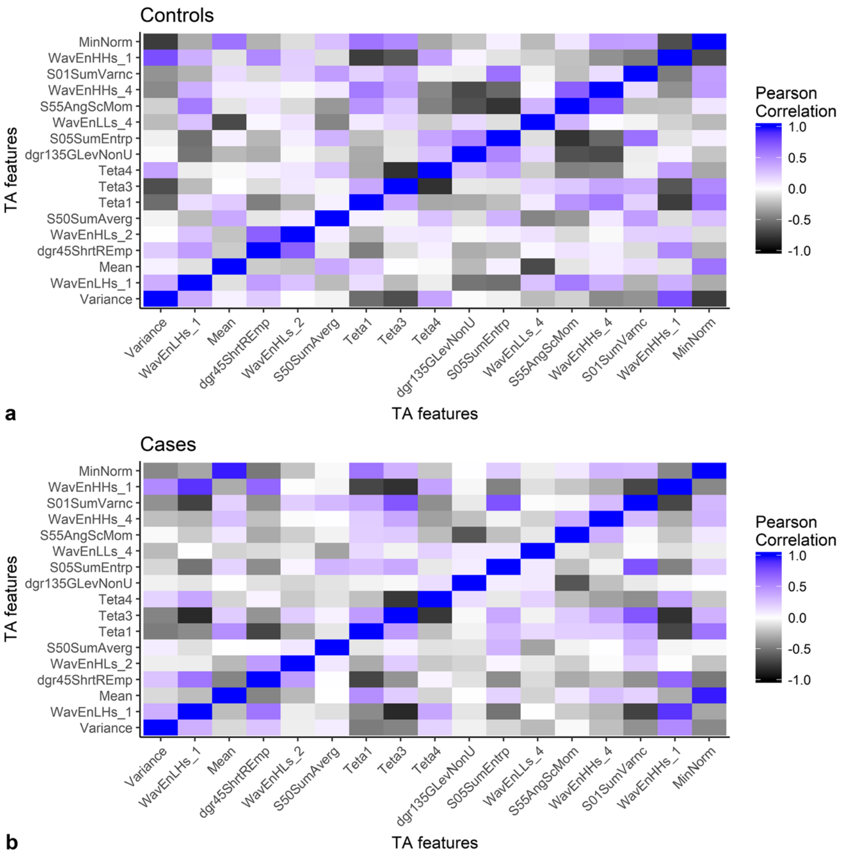

Fig. 6 Heatmap shows the combined top 10 important features of all classifier for controls (**a**) and cases demonstrating low co-correlation between features (**b**)

the notion that these features are independent surrogate markers for trabecular microarchitecture.

Particularly, complex TA features that showed a high importance in our classification (see Fig. 5) could be linked to known pathomorphological changes in the osteoporotic bone: Some GLRLM features (e.g., dgr45ShrtREmp, Table 1) that are defined over information of consecutive pixels of the same value in a given direction are negatively correlated with trabecular bone volume measured by histomorphometric methods, which is increased in osteoporotic individuals [34, 35]. Moreover, some wavelet features (e.g., WavEnHLs_2, Table 1) showed an important role in our classification. Because of their filter function, they can be considered as a breakdown of an image into a set of spatially oriented frequency channels useful to detect horizontal and vertical lines in images as well as crossings and corners. Wavelet transformation-derived texture features were previously reported to enhance the robustness and accuracy in trabecular bone classification on radiographs [17]. Moreover, they may aid detection of trabecular bone lesions in CT [36]. The GLCM feature entropy (S05SumEntrp, Table 1) that is defined over the distribution of co-occurring pixel values at a given offset was also of great importance in our classification. Entropy has been suggested as impaired skeletal integrity [19]. Further TA features, such as skewness (a measure of the asymmetry of the image histogram) or kurtosis (a measure of the peakedness of the image histogram), did not show a high importance in our classification.

In our study cohort, no reliable discrimination between a vertebra that will fracture within the following 4–12 months and a vertebra that will remain intact in the same patient was possible applying TA and ML. Several factors could have led to these non-significant results: For example, the differences might be too small for the sample size of this study. Alternatively, external factors, such as static or kinetic effects that have a determining influence on site of fracture [37] may not be contained in the data. Furthermore, standard feature extraction used in this paper could have limited discrimination of small differences. A recent study using deep features extracted from convolution neural networks, for example, revealed improved results in osteoporosis classification on X-ray images [38].

Our results from human reader analyses suggest that healthy and osteoporotic trabecular bone shows little visual differences leading to a low accuracy regarding the identification of osteoporotic bone as well as limited inter-reader agreement. Similar results were reported for radiographs [39]. However, in a real clinical setting, the accuracy of bone quality estimation is usually higher since visual signs of osteoporosis, such as vertebral cortical bone thinning and fractures of adjacent vertebrae in a larger spine region, can be considered.

In this retrospective study, we consider the male predominance in the case group (26 women and 34 men) as a selection bias since women have a higher prevalence and incidence of osteoporotic vertebral fractures. Male predominance could partially be explained by the inclusion criteria of 2 CT scans within a year that lead to a high number of patients with vascular repair (22/58).

There are several other limitations to our study. First, this was a retrospective case-control study, using a relatively small data set of 240 vertebrae of patients from a single academic institution. Second, the included CT images in our studies were acquired using various CT imaging protocols and different CT scanners, which may introduce biases to the texture features. Third, our analysis did not include other factors like cortical thickness or vertebra size, which are reported to be important in assessing bone quality. Fourth, TA features extraction in our study were limited to the features available in the MaZda software, and more recent methods for feature extraction in bone TA were not included [40]. Lastly, prebuilt machine learning models in our study were not thoroughly and individually “fine-tuned” with, e.g., a grid search, since this was not necessary for our proof-of-principle study.

In conclusion, TA combined with ML allows to identifying patients who will suffer from future vertebral insufficiency fractures in standard CT scans with a high accuracy. However, identification of single vertebra at risk remains challenging.

Funding The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Roman Guggenberger.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap The control cohort used in the present study (58 patients) is part of a study population of a previously published study establishing normative values for CT-based texture analysis of vertebral bodies. Mannil M, Eberhard M, Becker AS, et al (2017) Normative values for CT-based texture analysis of vertebral bodies in dual X-ray absorptiometry-confirmed, normally mineralized subjects. *Skeletal Radiology* 46:1541–1551. https://doi.org/10.1007/s00256-017-2728-0).

Methodology
- retrospective
- case-control study
- performed at one institution

References

1. Sambrook P, Cooper C (2006) Osteoporosis. Lancet 367:2010–2018. https://doi.org/10.1016/S0140-6736(06)68891-0
2. Kim DH, Vaccaro AR (2006) Osteoporotic compression fractures of the spine; current options and considerations for treatment. Spine J 6:479–487. https://doi.org/10.1016/j.spinee.2006.04.013
3. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy (2001) Osteoporosis prevention, diagnosis, and therapy. JAMA 285:785–795. https://doi.org/10.1001/jama.285.6.785
4. Johnell O, Kanis JA (2006) An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 17:1726–1733. https://doi.org/10.1007/s00198-006-0172-4
5. Johnell O, Kanis J (2005) Epidemiology of osteoporotic fractures. Osteoporos Int 16:S3–S7. https://doi.org/10.1007/s00198-004-1702-6
6. Delmas PD, van de Langerijt L, Watts NB et al (2005) Underdiagnosis of vertebral fractures is a worldwide problem: the IMPACT study. J Bone Miner Res 20:557–563. https://doi.org/10.1359/JBMR.041214
7. Silva BC, Leslie WD, Resch H et al (2014) Trabecular bone score: a noninvasive analytical method based upon the DXA image. J Bone Miner Res 29:518–530. https://doi.org/10.1002/jbmr.2176
8. Burns JE, Yao J, Summers RM (2017) Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology 284:788–797. https://doi.org/10.1148/radiol.2017162100
9. Schreiber JJ, Anderson PA, Rosas HG, Buchholz AL, Au AG (2011) Hounsfield units for assessing bone mineral density and strength: a tool for osteoporosis management. J Bone Joint Surg 93:1057–1063. https://doi.org/10.2106/JBJS.J.00160
10. Krug R, Burghardt AJ, Majumdar S, Link TM (2010) High-resolution imaging techniques for the assessment of osteoporosis. Radiol Clin North Am 48:601–621. https://doi.org/10.1016/j.rcl.2010.02.015
11. Damilakis J, Maris TG, Karantanas AH (2007) An update on the assessment of osteoporosis using radiologic techniques. Eur Radiol 17:1591–1602. https://doi.org/10.1007/s00330-006-0511-z
12. Imai K, Ohnishi I, Bessho M, Nakamura K (2006) Nonlinear finite element model predicts vertebral bone strength and fracture site. Spine (Phila Pa 1976) 31:1789–1794
13. Schwaiger BJ, Kopperdahl DL, Nardo L et al (2017) Vertebral and femoral bone mineral density and bone strength in prostate cancer patients assessed in phantomless PET/CT examinations. Bone 101:62–69. https://doi.org/10.1016/j.bone.2017.04.008
14. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ (2017) CT texture analysis: definitions, applications, biologic correlates, and challenges. Radiographics 37:1483–1503. https://doi.org/10.1148/rg.2017170056
15. Rachidi M, Marchadier A, Gadois C, Lespessailles E, Chappard C, Benhamou CL (2008) Laws' masks descriptors applied to bone texture analysis: an innovative and discriminant tool in osteoporosis. Skeletal Radiol 37:541–548. https://doi.org/10.1007/s00256-008-0463-2
16. Thevenot J, Hirvasniemi J, Pulkkinen P et al (2014) Assessment of risk of femoral neck fracture with radiographic texture parameters: a retrospective study. Radiology 272:184–191
17. Zou Z, Yang J, Megalooikonomou V, Jennane R, Cheng E, Ling H (2016) Trabecular bone texture classification using wavelet leaders. Proc. SPIE 9788, Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging, 97880E. https://doi.org/10.1117/12.2216452
18. Mannil M, Eberhard M, Becker AS et al (2017) Normative values for CT-based texture analysis of vertebral bodies in dual X-ray absorptiometry-confirmed, normally mineralized subjects. Skeletal Radiol 46:1541–1551. https://doi.org/10.1007/s00256-017-2728-0
19. Tabari A, Torriani M, Miller KK, Klibanski A, Kalra MK, Bredella MA (2016) Anorexia nervosa: analysis of trabecular texture with CT. Radiology 283:178–185
20. Torres C, Hammond I (2016) Computed tomography and magnetic resonance imaging in the differentiation of osteoporotic fractures from neoplastic metastatic fractures. J Clin Densitom 19:63–69. https://doi.org/10.1016/j.jocd.2015.08.008
21. Genant HK, Wu CY, van Kuijk C, Nevitt MC (2009) Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res 8:1137–1148. https://doi.org/10.1002/jbmr.5650080915
22. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A (2009) MaZda—a software package for image texture analysis. Comput Methods Programs Biomed 94:66–76. https://doi.org/10.1016/j.cmpb.2008.08.005
23. Andresen R, Radmer S, Banzer D (1998) Bone mineral density and spongiosa architecture in correlation to vertebral body insufficiency fractures. Acta Radiol 39:538–542
24. Ito M, Ikeda K, Nishiguchi M et al (2005) Multi-detector row CT imaging of vertebral microstructure for evaluation of fracture risk. J Bone Miner Res 20:1828–1836. https://doi.org/10.1359/JBMR.050610
25. Issever AS, Link TM, Kentenich M et al (2010) Assessment of trabecular bone structure using MDCT: comparison of 64- and 320-slice CT using HR-pQCT as the reference standard. Eur Radiol 20:458–468. https://doi.org/10.1007/s00330-009-1571-7
26. Kuhn M (2008) Building predictive models in *R* using the caret package. J Stat Softw 28. https://doi.org/10.18637/jss.v028.i05
27. Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Modell 160:249–264
28. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
29. Lin LI-K (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268. https://doi.org/10.2307/2532051
30. Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine learning for medical imaging. Radiographics 37:505–515. https://doi.org/10.1148/rg.2017160130
31. Valentinitsch A, Patsch J, Mueller D et al (2010) Texture analysis in quantitative osteoporosis assessment. In: Biomedical imaging: from nano to macro, 2010 IEEE International Symposium on. IEEE, pp 1361–1364
32. Boivin GY, Chavassieux PM, Santora AC, Yates J, Meunier PJ (2000) Alendronate increases bone strength by increasing the mean degree of mineralization of bone tissue in osteoporotic women. Bone 27:687–694
33. Guggenbuhl P, Bodic F, Hamel L, Baslé MF, Chappard D (2006) Texture analysis of X-ray radiographs of iliac bone is correlated with bone micro-CT. Osteoporos Int 17:447–454. https://doi.org/10.1007/s00198-005-0007-8
34. Chappard D, Guggenbuhl P, Legrand E, Baslé MF, Audran M (2005) Texture analysis of X-ray radiographs is correlated with bone histomorphometry. J Bone Miner Metab 23:24–29. https://doi.org/10.1007/s00774-004-0536-9
35. Kimmel DB, Recker RR, Gallagher JC, Vaswani AS, Aloia JF (1990) A comparison of iliac bone histomorphometric data in

post-menopausal osteoporotic and normal subjects. Bone Miner 11: 217–235

36. Reddy TK, Kumaravel N (2010) Wavelet based texture analysis and classification of bone lesions from dental CT. Int J Med Eng Inf 2:319–327

37. Rohlmann A, Zander T, Bergmann G (2006) Spinal loads after osteoporotic vertebral fractures treated by vertebroplasty or kyphoplasty. Eur Spine J 15:1255–1264. https://doi.org/10.1007/s00586-005-0018-3

38. Paul R, Alahamri S, Malla S, Quadri GJ (2017) Make your bone great again: a study on osteoporosis classification. Available via http://arxiv.org/abs/1707.05385. Accessed 02 Jan 2018

39. Wagner S, Stäbler A, Sittek H et al (2005) Diagnosis of osteoporosis: visual assessment on conventional versus digital radiographs. Osteoporos Int 16:1815–1822. https://doi.org/10.1007/s00198-005-1937-x

40. Ngo VQ, Dinh TN (2016) Bone texture characterization based on Contourlet and Gabor tranforms. Int J Comput Theory Eng 8:177–181. https://doi.org/10.7763/IJCTE.2016.V8.1040