Iranian Journal of Radiology
Volume 21 · Issue 3 · e147913

OPEN PEER REVIEW

# The Use of Radiomics Data Obtained from ADC Map of Lumbar MRI and Machine Learning in Diagnosis of Osteoporosis

Fatih Erdem, Emrah Akay, Gulen Demirpolat, Bahar Yanık Keyik, Erdogan Bulbul

| SUBMITTED | REVISED | ACCEPTED | DOI |
| --- | --- | --- | --- |
| 13 Apr 2024 | 29 Jun 2024 | 16 Jul 2024 | 10.5812/iranjradiol-147913 |

## About this report

This open peer review (OPR) report shows the public comments made on this article during peer review, by the reviewers, the associate editor and the editor-in-chief, for every revision round, together with the authors' replies. Comments appear only with the consent of the people who wrote them. Reviewers who chose to sign their report are named; the others are shown as "Referee". Comments are edited for language only, and confidential remarks to the editors are not included.

More about open peer review: brieflands.com/knowledgebase/opr

## Contents

1 **Original submission** Referee 1, Referee 2, Associate Editor, Editor-in-Chief ..... 2
2 **Revision 1** Referee 1, Editor-in-Chief ..... 6
**Authors' reply to reviewers** ..... 7
**Reviewer and editor comments** ..... 32
3 **Revision 2** Referee 1, Associate Editor, Editor-in-Chief ..... 34
**Authors' reply to reviewers** ..... 35
**Reviewer and editor comments** ..... 40
4 **Revision 3** Editor-in-Chief ..... 41
**Authors' reply to reviewers** ..... 42
**Reviewer and editor comments** ..... 44

## 1 Original submission

4 comments · 6 May 2024 – 17 Jul 2024

### Referee 1 Reviewer

MAJOR REVISION 6 May 2024

Thank you for submitting your manuscript. After carefully reviewing it, I have several concerns that warrant major revisions before it can be considered for publication. I would like to highlight the following points:

Abstract

1. Lack of Objectives Clarity: The objective could be more specific and measurable. It should clearly state the intended outcomes of the study, such as the performance metrics or evaluation criteria for the machine learning models.
2. Inadequate Methods Description: The methods section lacks important details. It doesn't mention the study design (cross-sectional or case-control), inclusion & exclusion criteria, data collection process, specific machine learning algorithms used, or the rationale behind choosing manual segmentation for creating the dataset.
3. Insufficient Results Summary: The results section does not provide specific numerical values or statistical measures to support the claims. Including relevant statistics, such as accuracy, sensitivity, specificity, or p-values, would strengthen the findings presented in the abstract.
4. Impact: The abstract does not clearly state the potential impact of the study's findings on the field of osteoporosis diagnosis. While the study identifies significant features and thresholds for diagnosing osteoporosis, it does not discuss the clinical significance of these findings or their potential implications for patient care.
5. There is a relevant and well-executed study titled 'Radiomics Based on Lumbar Spine Magnetic Resonance Imaging to Detect Osteoporosis' that can be utilized as a model for refining the writing structure.

Introduction

6. Inadequate Justification of Radiomics and ML: The introduction mentions the use of radiomics and machine learning (ML) without adequately justifying their selection as the approach for diagnosing osteoporosis. It would be beneficial to provide a stronger rationale for why radiomics and ML are promising methods that can complement or improve existing diagnostic techniques.
7. Absence of Specific Research Questions and Hypotheses: The introduction lacks clear research questions or hypotheses that the study aims to address. It is important to explicitly state the objectives of the study to provide readers with a clear understanding of the research goals and to guide the subsequent sections of the paper.

Methods, Results, and Tables

8. Lack of coherence and clarity: The methods section, although attempting to provide a thorough explanation, lacks coherence and clarity. The organization and structure of the section need improvement to enhance readability and understanding.
9. Excessive details in the results section: The textual description in the results section includes unnecessary details that could be streamlined. Focusing on the key steps and omitting non-essential specifics would make the section more concise and reader-friendly.
10. Tables copied from SPSS software: The results section contains tables that appear to have been directly copied from the SPSS software without proper formatting. This compromises the overall quality of the presentation. I recommend recreating the tables in a suitable format that aligns with the manuscript's style and ensures clarity for readers.

11. Missing figure: The absence of the figure related to the ROC analysis is a significant omission. Including this figure is crucial, as it provides valuable information and strengthens the presentation of the results.

Thank you for your attention to these matters. I look forward to reviewing the revised manuscript.

**Referee 2** Reviewer

MAJOR REVISION 7 Jul 2024

Dear Authors,

Thank you for submitting your valuable work.

Regarding the machine learning study presented, there are a few comments that are important in the context of such studies:

1. Your dataset is small and consists mainly of women, which affects the generalizability of this study.
2. The normal and pathological groups are disproportionate, which may influence the study's findings.
3. While using 10-fold cross-validation is appropriate for a small dataset, the selection of only five radiomic features increases the risk of overfitting. Additionally, the disproportionate subsets (normal and osteoporosis) might augment this effect, impacting machine learning performance. Despite cross-validation providing a more appropriate estimation in the context of a small sample size, these concerns should be addressed.
4. The choice of standardized parameters, such as bin width and filter kernel size, should be justified. Different settings could significantly influence the radiomic features.
5. Extracting 5580 features is comprehensive, but it increases the risk of overfitting. The manuscript should discuss strategies to mitigate this risk.
6. Only five features were selected using the FCBF method. While this reduces dimensionality, the manuscript should explain why five features were chosen and explore the implications of this choice.
7. The assessment of data distribution normality is noted, but the manuscript should discuss how deviations from normality, especially for skewed distributions, are handled.
8. The skewed gender distribution (94% female) limits the applicability of the results to the male population. This significant limitation should be acknowledged.
9. The statistically significant differences in body weight and BMI between osteoporotic and non-osteoporotic groups are noteworthy. The manuscript should discuss how these factors might influence radiomic features and model performance. Additionally, consider discussing whether using Z-scores could minimize this influence.
10. The neural network showed a low AUC (0.616), F1 score (0.108), and recall (0.095), indicating poor performance despite being the best model with all features. A detailed discussion on the possible reasons and implications of this finding is needed.
11. While the Naive Bayes model showed significant improvement after feature selection (AUC 0.913), the manuscript should discuss the potential reasons for this improvement and the clinical relevance of these findings.
12. The high number of significant features (58 with AUC > 0.70) raises concerns about potential overfitting. This should be addressed in the manuscript.

Overall, this study exhibits some concerns related to overfitting.

I hope these comments are helpful in improving your study.

## Associate Editor

MAJOR REVISION 16 Jul 2024

Dear authors,
Thank you for submitting your manuscript. After carefully reviewing it, we identified several concerns that warrant major revisions before it can be considered for publication. The more important concerns highlighted by the reviewers are as follows:
Reviewer 1:
Abstract
1. Lack of Objectives Clarity: The objective could be more specific and measurable. It should clearly state the intended outcomes of the study, such as the performance metrics or evaluation criteria for the machine learning models.
2. Inadequate Methods Description: The methods section lacks important details. It doesn't mention the study design (cross-sectional or case-control), inclusion & exclusion criteria, data collection process, specific machine learning algorithms used, or the rationale behind choosing manual segmentation for creating the dataset.
3. Insufficient Results Summary: The results section does not provide specific numerical values or statistical measures to support the claims. Including relevant statistics, such as accuracy, sensitivity, specificity, or p-values, would strengthen the findings presented in the abstract.
4. Impact: The abstract does not clearly state the potential impact of the study's findings on the field of osteoporosis diagnosis. While the study identifies significant features and thresholds for diagnosing osteoporosis, it does not discuss the clinical significance of these findings or their potential implications for patient care.
5. There is a relevant and well-executed study titled 'Radiomics Based on Lumbar Spine Magnetic Resonance Imaging to Detect Osteoporosis' that can be utilized as a model for refining the writing structure.
Introduction
6. Inadequate Justification of Radiomics and ML: The introduction mentions the use of radiomics and machine learning (ML) without adequately justifying their selection as the approach for diagnosing osteoporosis. It would be beneficial to provide a stronger rationale for why radiomics and ML are promising methods that can complement or improve existing diagnostic techniques.
7. Absence of Specific Research Questions and Hypotheses: The introduction lacks clear research questions or hypotheses that the study aims to address. It is important to explicitly state the objectives of the study to provide readers with a clear understanding of the research goals and to guide the subsequent sections of the paper.
Methods, Results, and Tables
8. Lack of coherence and clarity: The methods section, although attempting to provide a thorough explanation, lacks coherence and clarity. The organization and structure of the section need improvement to enhance readability and understanding.
9. Excessive details in the results section: The textual description in the results section includes unnecessary details that could be streamlined. Focusing on the key steps and omitting non-essential specifics would make the section more concise and reader-friendly.
10. Tables copied from SPSS software: The results section contains tables that appear to have been directly copied from the SPSS software without proper formatting. This compromises the overall quality of the presentation. I recommend recreating the tables in a suitable format that aligns with the manuscript's style and ensures clarity for readers.
11. Missing figure: The absence of the figure related to the ROC analysis is a significant omission. Including this figure is crucial, as it provides valuable information and strengthens the presentation of the results.

Reviewer 2:
Regarding the machine learning study presented, there are a few comments that are important in the context of such studies:
1. Your dataset is small and consists mainly of women, which affects the generalizability of this study.
2. The normal and pathological groups are disproportionate, which may influence the study's findings.
3. While using 10-fold cross-validation is appropriate for a small dataset, the selection of only five radiomic features increases the risk of overfitting. Additionally, the disproportionate subsets (normal and osteoporosis) might augment this effect, impacting machine learning performance. Despite cross-validation providing a more appropriate estimation in the context of a small sample size, these concerns should be addressed.
4. The choice of standardized parameters, such as bin width and filter kernel size, should be justified. Different settings could significantly influence the radiomic features.
5. Extracting 5580 features is comprehensive, but it increases the risk of overfitting. The manuscript should discuss strategies to mitigate this risk.
6. Only five features were selected using the FCBF method. While this reduces dimensionality, the manuscript should explain why five features were chosen and explore the implications of this choice.
7. The assessment of data distribution normality is noted, but the manuscript should discuss how deviations from normality, especially for skewed distributions, are handled.
8. The skewed gender distribution (94% female) limits the applicability of the results to the male population. This significant limitation should be acknowledged.
9. The statistically significant differences in body weight and BMI between osteoporotic and non-osteoporotic groups are noteworthy. The manuscript should discuss how these factors might influence radiomic features and model performance. Additionally, consider discussing whether using Z-scores could minimize this influence.
10. The neural network showed a low AUC (0.616), F1 score (0.108), and recall (0.095), indicating poor performance despite being the best model with all features. A detailed discussion on the possible reasons and implications of this finding is needed.
11. While the Naive Bayes model showed significant improvement after feature selection (AUC 0.913), the manuscript should discuss the potential reasons for this improvement and the clinical relevance of these findings.
12. The high number of significant features (58 with AUC > 0.70) raises concerns about potential overfitting. This should be addressed in the manuscript.

Overall, this study exhibits some concerns related to overfitting.
I hope these comments are helpful in improving your study.

Regards.

**Editor-in-Chief**

MAJOR REVISION 17 Jul 2024

Dear authors,
Your article needs Major Revision.
Regards

2

# Revision 1

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

## Response to Reviewer 1:

Dear Reviewer 1,

Thank you for your prompt review and feedback. We appreciate your assessment that the manuscript, as is, is acceptable. Your input is valuable to us.

Best regards.

## Response to Reviewer 2:

Abstract:

Comment 1: Lack of Objectives Clarity: The objective could be more specific and measurable. It should clearly state the intended outcomes of the study, such as the performance metrics or evaluation criteria for the machine learning models.

Revision: Objectives: To evaluate the performance of machine learning (ML) models using lumbar MRI ADC map radiomics for osteoporosis detection, and to identify significant features and their diagnostic thresholds. Specific performance metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC) were assessed.

Comment 2: Inadequate Methods Description: The methods section lacks important details. It doesn't mention about the study design (cross-sectional or case-control), inclusion & exclusion criteria, data collection process, specific machine learning algorithms used, or the rationale behind choosing manual segmentation for creating the dataset.

Revision: A retrospective study was conducted on 140 cases, 21 of which had osteoporosis. Manual segmentation of lumbar vertebrae from ADC maps was used to create a dataset with 5580 radiomic features per case. Seven ML algorithms (k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, and Neural Network) were tested with a focus on the top five features selected by FCBF. Statistical tests and ROC curve analysis were conducted to determine feature significance and thresholds. The study design was cross-sectional, with inclusion criteria including concurrent lumbar MRI and DEXA within a year, and exclusion

criteria being infectious or neoplastic lumbar lesions, fractures, instrumentation, significant osteodegenerative changes, cases in which the first four lumbar vertebrae not included in the imaging field, and absence of diffusion weighted imaging.

Comment 3: Insufficient Results Summary: The results section does not provide specific numerical values or statistical measures to support the claims. Including relevant statistics, such as accuracy, sensitivity, specificity, or p-values, would strengthen the findings presented in the abstract.

Revision: Results: The Neural Network model demonstrated the highest performance with an AUC of 0.616 and a classification accuracy of 0.764 using all features. The Naive Bayes model, using the top five features selected by FCBF, showed the highest performance with an AUC of 0.913, accuracy of 0.907, sensitivity of 0.667, and specificity of 0.95. All ML models' performance were elevated by feature selection. Independent t-tests and Mann-Whitney U tests identified 521 and 670 significant features, respectively ($p<0.05$). ROC analysis revealed 58 features with AUC values above 0.70.

Comment 4: Impact: The abstract does not clearly state the potential impact of the study's findings on the field of osteoporosis diagnosis. While the study identifies significant features and thresholds for diagnosing osteoporosis, it does not discuss the clinical significance of these findings or their potential implications for patient care.

Revision: **Conclusions:** This study's findings suggest that ML models, particularly the Naive Bayes algorithm, can effectively use lumbar ADC map radiomics to diagnose osteoporosis. These findings could enhance early detection and treatment strategies, potentially improving patient outcomes and reducing the burden of osteoporotic fractures. This study also established threshold values for significant features.

Comment 5: There is a relevant and well-executed study titled 'Radiomics Based on Lumbar Spine Magnetic Resonance Imaging to Detect Osteoporosis' that can be utilized as a model for refining the writing structure.

Response to the comment: In accordance with Reviewer 2's suggestion, the relevant and well-structured study titled 'Radiomics Based on Lumbar Spine Magnetic Resonance Imaging to Detect Osteoporosis' was used as a model to enhance the writing structure.

Introduction

Comment 6: Inadequate Justification of Radiomics and ML: The introduction mentions the use of radiomics and machine learning (ML) without adequately justifying their selection as the approach for diagnosing osteoporosis. It would be beneficial to provide a stronger rationale for why radiomics and ML are promising methods that can complement or improve existing diagnostic techniques.

Revision: The second-to-last paragraph of the ‘Background’ section was revised as follows:

Radiomics and machine learning (ML) offer novel approaches for osteoporosis diagnosis by extracting high-dimensional data from medical images, which traditional methods might overlook. Radiomics can quantify tissue heterogeneity and capture subtle changes in bone microarchitecture, while ML algorithms can analyze these complex data sets to identify patterns indicative of osteoporosis. These techniques could complement existing diagnostic methods, such as DEXA, by providing more detailed and accurate assessments of bone health. Early and accurate detection of osteoporosis using these advanced techniques could potentially lead to better clinical outcomes by enabling timely intervention and personalized treatment plans, thus reducing the risk of fractures and associated complications.

Comment 7: Absence of Specific Research Questions and Hypotheses: The introduction lacks clear research questions or hypotheses that the study aims to address. It is important to explicitly state the objectives of the study to provide readers with a clear understanding of the research goals and to guide the subsequent sections of the paper.

Revision: The last parapgraph of the “Background” section was revised as follows:

The primary objective of this study is to evaluate the diagnostic performance of various machine learning models, specifically Naive Bayes and neural networks, in identifying osteoporosis using radiomic features extracted from lumbar MRI ADC maps. This will be measured using performance metrics such as Area Under the Curve (AUC), accuracy, sensitivity, specificity, F1 score, and recall. Additionally, the secondary objective is to identify the most significant radiomic features that contribute to accurate osteoporosis diagnosis. The hypotheses guiding this study are: 1- Machine learning models can significantly improve the diagnostic accuracy of osteoporosis compared to traditional methods; and 2- specific radiomic features from lumbar MRI ADC maps are significantly associated with osteoporosis and can reliably differentiate between osteoporotic and non-osteoporotic cases. These objectives and hypotheses are designed to

provide a comprehensive evaluation of the potential of ML in enhancing osteoporosis diagnosis, with a focus on both model performance and feature significance.

Methods, Results and Tables

Comment 8: Lack of coherence and clarity: The methods section, although attempting to provide a thorough explanation, lacks coherence and clarity. The organization and structure of the section need improvement to enhance readability and understanding.

Revision: The “Methods” section was revised as follows:

- **Study Design and Patient selection**

This cross-sectional study scanned the cases referred to the Radiology Department between 01/01/2015 and 31/12/2020 retrospectively. Lumbar MRIs were obtained from a Philips Ingenia 1.5T MRI system (Philips, Best, Netherlands) and bone mineral density (BMD) measurements were acquired using a DEXA device (Lunar Prodigy, model 8743; GE Lunar, Madison, WI, USA). Lumbar MRIs and DEXA examinations from the same patients, performed within one year of each other, were included unless they met the exclusion criteria.

Exclusion criteria were as follows:

1. Interval exceeding one year between lumbar MRI and DEXA examinations,
2. Presence of an infectious or a neoplastic lesion in any lumbar vertebral body
3. Fracture of a lumbar vertebral body, instrumentation, the osteodegenerative changes that could impair the measurements,
4. Cases in which the first four lumbar vertebrae not included in the imaging field,
5. Absence of diffusion weighted imaging.

The study continued with 140 participants, 21 of whom had osteoporosis, after exclusion (Figure 1).

The cases’ height and weight were recorded. Anterior-posterior lumbar vertebrae BMD were routinely measured. The DEXA scan used L1-4 as the basis for T-scores. The cases were classified as “osteoporosis” if the T-score was ≤-2.5, “non-osteoporosis” if the T-score was >2.5.

- **Image acquisition**

The lumbar MRI diffusion-weighted imaging sequence was acquired axially on a Philips Ingenia 1.5T MRI device (Philips, Best, Netherlands). The sequence utilized a spin-echo technique with fat suppression (SPIR) and parallel imaging method (SENSE). The pulse repetition time was 7505 milliseconds, the echo time was 60 milliseconds, and the echo train length was 43. The slice thickness was 3 mm, with a 3 mm gap between slices. The field of view (FOV) was 225x225 $mm^2$, and the b-values were 0 and 650 $s/mm^2$.

The PACS software available in our hospital was used as the archiving system.

Lumbar MRI diffusion-weighted images (b values of 0 and 650 s/mm2) of the cases that met the criteria were transferred to the 3DSlicer 4.8.1 (www.slicer.org) program in DICOM (Digital Imaging and Communications in Medicine) format. First, monoexponential ADC maps were created with the module named DWModeling.

- **Segmentation**

Segmentation was performed manually by an experienced board-certified radiologist. To ensure its consistency, the intraobserver reliability was evaluated using intraclass correlation coefficient (ICC) that measured the absolute agreement between the features obtained from 2 separate segmentation sets performed in one month interval. Features that achieved good ($0.75 \leq ICC < 0.9$) to excellent ($ICC \geq 0.9$) intraobserver reliability were considered stable.

The contours were meticulously drawn to closely follow the cortical bone while remaining within the boundaries of the vertebral body. Each vertebral level was segmented separately and named as L1, L2, L3, and L4. Additionally, inspired by the DEXA assessment, which provides average bone mineral density values for L1, L2, L3, and L4 levels, the segmentations were combined into two separate sets: L1-4 and L2-4. Consequently, a total of six segmentations—L1, L2, L3, L4, L1-4, and L2-4—were obtained for each case (Figure 2).

- **Preprocessing and radiomic feature extraction**

Following the segmentation process, features were extracted from segmented volumes obtained using the PyRadiomics-based Radiomics module. Prior to feature extraction, a consistent configuration was applied across all cases. Specifically, under the module's "resampling and filtering" section, the bin width was fixed at 25. Additionally, the Laplacian of Gaussian filter kernel size was set to 3 mm, and wavelet-based features were enabled to generate transformed filtered images. This standardization aimed to achieve uniformity among cases,

increase the volume of data, and capture high dimensionality in the radiomics dataset.

Subsequently, for each of the L1, L2, L3, L4, L1-4, and L2-4 segmentations, the following features were extracted:

- First-order features (18 features) from the original images
- Second-order features based on GLCM (24 features)
- Second-order features based on GLDM (14 features)
- High-order features based on GLRLM (16 features)
- High-order features based on GLSZM (16 features)
- High-order features based on NGTDM (10 features)

Additionally, features were extracted from Laplacian of Gaussian-filtered images with a kernel size of 3 mm:

- First-order features (18 features)
- Second-order features based on GLCM (24 features)
- Second-order features based on GLDM (9 features)
- High-order features based on GLRLM (16 features)
- High-order features based on GLSZM (16 features)
- High-order features based on NGTDM (5 features)

Furthermore, a total of 744 features were obtained from wavelet-transformed images for each case, resulting in a grand total of 5580 features across all cases.

- **Feature selection, model validation, classification and performance evaluation**

For the classification using ML algorithms, the Orange data mining program version 3.30 (available at https://orange.biolab.si/) was employed.

The dataset consisted of 5580 features and 140 cases, which were subjected to classification using seven distinct ML algorithms: k-Nearest Neighbors (kNN), Decision Tree, Random Forest, Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Neural Network. The classification was performed before and after the feature selection process using these seven ML algorithms to see the positive effect of feature selection on the performance metrics. As a feature selection method, the algorithm-based FCBF (Fast Correlation-Based Filter) was utilized to select five features. Stratified 10-fold cross-validation was employed as the internal validation technique.

The performance evaluation of the ML models included metrics such as AUC (Area Under the Curve), Classification Accuracy (CA), F1 score, Precision (Positive Predictive Value), Recall (Sensitivity), and Specificity.

- **Conventional statistical analysis**

IBM SPSS Statistics for Windows, version 22 (IBM Corp., Armonk, NY, USA) were used for conventional statistical analysis.

After the ML model performance evaluation, independent samples t-test and Mann-Whitney U test were performed to determine whether each feature showed a statistically significant difference in osteoporosis status. Normal distribution of data was assessed by Kolmogorov-Smirnov test and histogram graphs. Means and SDs for normally distributed data, medians and interquartile ranges for non-normally distributed data were expressed as mean (SD), median (interquartile range), respectively. The categorical data were expressed as counts and percentages. Receiver operating characteristic (ROC) curve analysis was performed for the features found to be significant in these tests, and threshold values were determined. $P < 0.05$ was considered significant in all statistical results.

- **Ethical considerations**

This study was approved by the Clinical Research Ethics Committee of Balikesir University Faculty of Medicine at 26/05/2021 with the decision number of 2021/129.

The study doesn't require informed consent because of its retrospective nature.

Comment 9: Excessive details in the results section: The textual description in the results section includes unnecessary details that could be streamlined. Focusing on the key steps and omitting non-essential specifics would make the section more concise and reader-friendly.

Revision: The "Results" section was revised as follows:

A total of 140 cases (132 females, 8 males; mean age 65.32 ± 8.50 years) were included. For women, the mean age was 65.20 ± 8.19 years, and for men, it was 67.22 ± 13.28 years. Female subjects had a median height of 1.53 ± 0.08 m, mean body weight of 73.21 ± 13.81 kg, and mean BMI of 31.43 ± 5.53 kg/m². Male subjects had a mean height of 1.67 ± 0.08 m, mean body weight of 73.25 ± 15.64 kg, and mean BMI of 26 ± 3.59 kg/m².

Age, body weight, and BMI fit a normal distribution for all cases and by gender, except height, which was normally distributed only for males. An

independent samples t-test showed no significant age difference between osteoporotic and non-osteoporotic groups (p=0.889), but body weight and BMI were significantly different (p=0.003 and p=0.001, respectively). Osteoporotic group had lower mean body weight (64.90 ± 10.13 kg) and BMI (27.40 ± 4.38 kg/m²) compared to the non-osteoporotic group (74.68 ± 13.94 kg and 31.77 ± 5.52 kg/m²) (Table 1). No significant difference in height was found (p=0.412).

The median interval between DEXA and lumbar MRI was 1 month (max 11.87 months) (Figure 3).

Neural network was the top-performing ML model using 5580 radiomic features for osteoporosis detection (AUC 0.616, CA 0.764, F1 score 0.108, precision 0.125, recall 0.095, specificity 0.882) (Figure 4). Feature selection with FCBF improved performance across seven ML models (Table 2), with Naive Bayes achieving the highest performance (AUC 0.913, CA 0.907, F1 score 0.683, precision 0.7, recall 0.667, specificity 0.95) (Table 3, 4 and Figure 5 A, B).

All features had ICC values >0.75, indicating stability. Independent samples t-test and Mann-Whitney U tests identified 521 and 670 significant features (p<0.05) for osteoporosis status, respectively. ROC analysis showed 58 features with AUC >0.70, predominantly from wavelet-transformed images (42 features), mainly from the LHL subvolume (30 features) and the L1 vertebra level (38 features). The Youden index determined threshold values for these features, with corresponding sensitivity and specificity provided (Table 5). Among these 58 features, there were 30 features belonging to the L1 vertebra level obtained from the images created by wavelet transformation. Of these 30 features, 9 were first order, 8 were GLCM-based, 3 were GLDM-based, 5 were GLRLM-based, 4 were GLSZM-based, and 1 was NGTDM-based.

Comment 10: Tables copied from SPSS software: The results section contains tables that appear to have been directly copied from the SPSS software without proper formatting. This compromises the overall quality of the presentation. I recommend recreating the tables in a suitable format that aligns with the manuscript's style and ensures clarity for readers.

Response: Dear Reviewer,

Thank you for your feedback regarding the tables in the results section. I appreciate your attention to detail. However, I'd like to clarify that the tables were not directly copied from SPSS; rather, they were manually created in

Microsoft Word. While I aimed to maintain clarity and accuracy, I understand the importance of proper formatting.

Given the complexity of the results, I believe that the meticulous presentation of data—reflecting the nuances captured during analysis—is best achieved through the current table format. Nevertheless, I will maintain the current format of the tables to accurately convey the complexity of the results while ensuring clarity.

Comment 11: Missing figure: The absence of the figure related to the ROC analysis is a significant omission. Including this figure is crucial, as it provides valuable information and strengthens the presentation of the results.

Response: Dear Reviewer,

Thank you for your thorough review and valuable feedback. I appreciate your attention to detail.

Regarding Comment 11, I apologize for any confusion. The figures related to the ROC analysis are indeed included in the manuscript. They can be found among the existing figures. If there is any specific aspect of the ROC analysis you would like me to highlight further, please let me know, and I will ensure its clarity.

Once again, thank you for your thoughtful review, and I will make any necessary adjustments based on your input.

**Response to Reviewer 3:**

Dear Reviewer,

Thank you for your insightful assessment of our manuscript. We appreciate your recognition of the study's innovative approach in leveraging radiomics data from lumbar MRI ADC maps and machine learning algorithms for osteoporosis diagnosis. Your acknowledgment of its potential impact is encouraging.

We take your feedback seriously, and we are committed to addressing the areas that need improvement. Enhancing clarity, robustness, and applicability will be our priority as we revise the manuscript.

Once again, thank you for your thoughtful review, and we look forward to refining our work based on your valuable insights.

Best regards.

Background:

Comment 1: While the background mentions the use of ADC values and radiomic features, can you provide a more detailed explanation of other modalities that are used in this area? Is there any other way to diagnose osteoporosis? What other imaging modalities are used for this issue, and what are the advantages and disadvantages of each?

Revision: The revisions were made and added to “Background” as the fourth, the fifth, and the sixth paragraph:

In addition to DEXA, several other imaging modalities are used for the diagnosis and evaluation of osteoporosis. Quantitative computed tomography (QCT) provides 3D imaging and volumetric measurements of BMD, offering a more accurate assessment of trabecular and cortical bone separately. QCT is less affected by degenerative changes and can measure bone density at various skeletal sites, making it a valuable tool for osteoporosis diagnosis. However, it involves higher radiation doses compared to DEXA and is more expensive, which limits its widespread use (9).

Quantitative ultrasound (QUS) is another modality used to assess bone density. It is portable, non-invasive, and does not involve radiation, making it suitable for use in various settings. QUS is particularly useful for evaluating bone density at peripheral sites, such as the heel. However, QUS is less accurate and reliable compared to DEXA and QCT, and it is primarily used for screening rather than definitive diagnosis (10).

High-resolution peripheral quantitative computed tomography (HR-pQCT) provides detailed images of bone microarchitecture at peripheral sites, such as the distal radius and tibia. It can improve fracture prediction beyond what DEXA offers and may be useful for monitoring anti-osteoporotic therapy and assessing bone disorders related to chronic kidney disease. However, HR-pQCT's broader use is limited by factors such as high cost, limited machine availability, the need for better reproducibility, and a lack of reference normative data. Despite these challenges, HR-pQCT offers more comprehensive bone quality assessment compared to DEXA (11).

Comment 2: The background could be enhanced by a brief discussion on the clinical implications of early and accurate osteoporosis detection using advanced techniques.

Revision: The second-to-last paragraph of the ‘Background’ section was revised as follows:

Radiomics and machine learning (ML) offer novel approaches for osteoporosis diagnosis by extracting high-dimensional data from medical images, which traditional methods might overlook. Radiomics can quantify tissue heterogeneity and capture subtle changes in bone microarchitecture, while ML algorithms can analyze these complex data sets to identify patterns indicative of osteoporosis. These techniques could complement existing diagnostic methods, such as DEXA, by providing more detailed and accurate assessments of bone health. Early and accurate detection of osteoporosis using these advanced techniques could potentially lead to better clinical outcomes by enabling timely intervention and personalized treatment plans, thus reducing the risk of fractures and associated complications.

Comment 3: Clearly state the hypothesis and objectives of the study. While the primary objective is mentioned, the secondary objectives need more explicit articulation.

Revision: The last paragraph of the “Background” was revised as follows:

The primary objective of this study is to evaluate the diagnostic performance of various machine learning models, specifically Naive Bayes and neural networks, in identifying osteoporosis using radiomic features extracted from lumbar MRI ADC maps. This will be measured using performance metrics such as Area Under the Curve (AUC), accuracy, sensitivity, specificity, F1 score, and recall. Additionally, the secondary objective is to identify the most significant radiomic features that contribute to accurate osteoporosis diagnosis. The hypotheses guiding this study are: 1- Machine learning models can significantly improve the diagnostic accuracy of osteoporosis compared to traditional methods; and 2- specific radiomic features from lumbar MRI ADC maps are significantly associated with osteoporosis and can reliably differentiate between osteoporotic and non-osteoporotic cases. These objectives and hypotheses are designed to provide a comprehensive evaluation of the potential of ML in enhancing osteoporosis diagnosis, with a focus on both model performance and feature significance.

Methods:

Comment 4: The process of manual segmentation needs additional information on the radiologist's qualifications, and the consistency of the segmentation process (e.g., inter-rater reliability, intra-rater reliability) has to be mentioned.

Revision: Segmentation was performed manually by an experienced board-certified radiologist. To ensure its consistency, the intraobserver reliability was evaluated using intraclass correlation coefficient (ICC) that measured the absolute agreement between the features obtained from 2 separate segmentation sets performed in one month interval. Features that achieved good ($0.75 \leq ICC < 0.9$) to excellent ($ICC \geq 0.9$) intraobserver reliability were considered stable.

Comment 5: Why were certain filter sizes or bin widths selected? How do these choices impact the radiomic features?

Response: The bin width was set to the default value of 25 in the radiomics module of the 3D Slicer application. This choice was made to ensure standardization across all cases, facilitating consistent and comparable feature extraction.

The choice of a 3 mm Laplacian of Gaussian (LoG) filter size in radiomic analysis is guided by its ability to capture relevant spatial frequencies within the medical images. This filter size is effective in detecting edges and fine structures at a scale that is representative of the anatomical and pathological features relevant to the study. In the context of medical imaging, selecting an appropriate filter size is crucial to balance between sensitivity to fine details and noise suppression.

Using a 3 mm LoG filter in the radiomics plugin of 3D Slicer, for example, is a standardized approach to ensure consistency across analyses. This size is particularly chosen to optimize the detection of structural variations within the given image resolution, which is vital for accurate feature extraction and subsequent analysis.

Comment 6: The method mentions creating ADC maps and segmenting images but lacks detail on how data was managed post-extraction. Briefly describe any normalization techniques, data splitting (e.g., training and testing sets), and feature reduction methods before applying ML algorithms.

Response:

Normalization techniques were not applied in this study. Data splitting was not performed due to the limited number of cases available. Instead, 10-fold cross-validation was used for model validation to ensure the robustness of the machine learning algorithms. As already mentioned in the text, the feature reduction method used was the Fast Correlation-Based Filter (FCBF), which effectively

identified and selected the most relevant features from the radiomic dataset. This method was chosen to ensure that the feature set remained manageable and informative despite the constraints in data size.

Comment 7: Incorporate information on how the ML models were validated. Describe cross-validation methods, if used, and how performance metrics were calculated to ensure robust model evaluation.

Revision: In this study, the machine learning (ML) models were validated using 10-fold cross-validation to ensure robust evaluation. This method involves dividing the dataset into ten equal parts, where each part is used as a validation set while the remaining nine parts are used for training. This process is repeated ten times, with each part being used as the validation set once. The performance metrics were calculated by averaging the results from each fold, which provides a more reliable estimate of the model's performance compared to a single train-test split.

The performance metrics, including AUC (Area Under the Curve), Classification Accuracy (CA), F1 score, Precision (Positive Predictive Value), Recall (Sensitivity), and Specificity, were calculated for each fold. These metrics provide a comprehensive evaluation of the model’s diagnostic capabilities, measuring its ability to correctly identify positive and negative cases of osteoporosis. The use of cross-validation ensures that the models are not overfitted to a particular subset of the data and can generalize well to unseen data.

Comment 8: The authors have to provide more details on the demographic characteristics of the study population, such as age distribution, gender ratio, and relevant clinical history, to give a better context for the sample.

Response: The demographic characteristics of the study population have been thoroughly detailed in the results section of the manuscript. Specifically, the first paragraph of the results section provides the following information:

"A total of 140 cases (132 females, 8 males; mean age 65.32 ± 8.50 years) were included. For women, the mean age was 65.20 ± 8.19 years, and for men, it was 67.22 ± 13.28 years. Female subjects had a median height of 1.53 ± 0.08 m, mean body weight of 73.21 ± 13.81 kg, and mean BMI of 31.43 ± 5.53 kg/m². Male subjects had a mean height of 1.67 ± 0.08 m, mean body weight of 73.25 ± 15.64 kg, and mean BMI of 26 ± 3.59 kg/m²."

This section provides a comprehensive overview of the demographic details, ensuring a clear context for the sample used in the study.

Comment 9: Specify the reasons for choosing the specific MRI sequence parameters and the PyRadiomics settings.

Response: The selection of MRI sequence parameters and PyRadiomics settings was guided by a combination of established protocols, previous research, and practical considerations to optimize image quality and feature extraction. The specific reasons are as follows:

MRI Sequence Parameters:

Apparent Diffusion Coefficient (ADC) Maps:

The use of ADC maps is based on their established utility in assessing the microstructural changes in bone marrow, which are indicative of osteoporosis. ADC values decrease in osteoporotic patients due to the narrowing of the extracellular space from increased adipose tissue, as suggested by previous studies.

Sequence Settings:

The MRI sequence parameters, such as echo time (TE), repetition time (TR), and b-values, were chosen to balance image quality and acquisition time while maximizing sensitivity to diffusion changes in the bone marrow. These settings are consistent with those used in similar studies to ensure comparability and reproducibility of results.

PyRadiomics Settings:

Bin Width:

The bin width for radiomic feature extraction was set to 25, the default value in the PyRadiomics module of 3D Slicer. This standardization ensures consistency across all cases and facilitates the comparison of radiomic features by maintaining uniform quantization of the image intensities.

Laplacian of Gaussian (LoG) Filter Size:

A filter size of 3 mm was selected for the LoG filter based on its effectiveness in enhancing edge detection and texture characterization at the spatial resolution of our MRI images. This size is commonly used in radiomic studies to capture fine details while minimizing noise.

Feature Reduction Method:

Fast Correlation Based Filter (FCBF) was employed as the feature reduction method due to its ability to identify the most relevant features while reducing dimensionality, thereby improving the performance of machine learning models.

These choices reflect a careful consideration of both literature precedents and the technical capabilities of the imaging and analysis tools, aiming to produce robust and reliable radiomic features for the diagnosis of osteoporosis.

Comment 10: Describe the steps taken to ensure the robustness of the stratified 10-fold cross-validation process.

Revision: The last two paragraphs was revised as follows:

In this study, the machine learning (ML) models were validated using 10-fold cross-validation to ensure robust evaluation. This method involves dividing the dataset into ten equal parts, where each part is used as a validation set while the remaining nine parts are used for training. This process is repeated ten times, with each part being used as the validation set once. The performance metrics were calculated by averaging the results from each fold, which provides a more reliable estimate of the model's performance compared to a single train-test split.

The performance metrics, including AUC (Area Under the Curve), Classification Accuracy (CA), F1 score, Precision (Positive Predictive Value), Recall (Sensitivity), and Specificity, were calculated for each fold. These metrics provide a comprehensive evaluation of the model’s diagnostic capabilities, measuring its ability to correctly identify positive and negative cases of osteoporosis. The use of cross-validation ensures that the models are not overfitted to a particular subset of the data and can generalize well to unseen data.

Comment 11: Elaborate on the criteria for determining the threshold values for significant radiomic features.

Response: The criteria for determining the threshold values for significant radiomic features were based on the Youden index. This index is a well-established method used to identify the optimal threshold that maximizes the sum of sensitivity and specificity. By applying the Youden index, we ensured that the chosen cut-off points provided the most effective discrimination between osteoporotic and non-osteoporotic cases. This method allows for a balanced evaluation of the diagnostic performance, improving the robustness and reliability of the model's predictions. This approach and the resulting threshold values are detailed in the corresponding table within the study.

Method section revision: “Threshold values for significant radiomic features were determined using the Youden index. The Youden index was employed to identify

the optimal cut-off points that maximize the sum of sensitivity and specificity, thereby ensuring a balanced and effective discrimination between osteoporotic and non-osteoporotic cases.”

Results:

Comment 12: The quality of the figures included in the manuscript is too low. Clear and high-resolution images are essential.

Response: The figures included in the manuscript were uploaded at a resolution of 300 dpi to ensure clarity and high quality. This resolution is typically sufficient for publication standards, providing clear and detailed images for the readers' better understanding.

Comment 13: In Table 3, the authors mention that the Naive Bayes and neural network methods are the two top networks used, but in Figure 5, the opposite can be seen. Please explain this.

Response: In Table 3, the performance metrics of different machine learning (ML) algorithms are summarized. It shows that the Naive Bayes algorithm has the highest Area Under the Curve (AUC) of 0.913 and Classification Accuracy (CA) of 0.907, indicating it as the top-performing model. However, the neural network algorithm, despite being the second in terms of AUC (0.752) and CA (0.864), performs poorly in terms of F1 score (0.296) and Recall (0.19), which are critical for evaluating the model’s performance in identifying true positive cases.

The confusion arises in Figure 5 because it presents Receiver Operating Characteristic (ROC) curves for the models. ROC curves and AUC values help in understanding the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). The figure shows the ROC curves of various models, and it can visually appear that certain models perform better in specific regions of the curve. However, the comprehensive performance of models is better summarized by metrics like AUC, CA, F1, and Recall provided in Table 3.

Thus, while Figure 5 visually compares the sensitivity and specificity of the models, Table 3 provides a more detailed performance comparison using multiple metrics. This explains why Naive Bayes and Neural Network might appear differently in the figure compared to the detailed analysis in the table. The figure helps to understand the model performance visually, but Table 3 offers a more accurate and detailed numerical performance assessment.

Moreover, it is important to note that the Naive Bayes model's ROC curve in Figure 5 clearly demonstrates superior performance compared to the other models. The visual distinction is apparent, and the confusion may arise from the

relative performance of the second-best model, the Neural Network. This model's performance, while still notable, does not surpass the Naive Bayes model in comprehensive evaluation metrics, leading to potential misinterpretation when only considering visual ROC curves.

Comment 14: Table 5 contains a lot of information about which no conclusions have been made and cannot be checked or followed. It is better to display this large amount of information in another format.

Response: Thank you for your feedback. However, we believe that the detailed information presented in Table 5 is crucial for the comprehensiveness of our study. The table provides an in-depth analysis that supports our findings and allows for thorough verification and follow-up. Displaying this information in another format may result in loss of critical details that are essential for understanding the nuances of our research. We appreciate your understanding and are open to any suggestions for improving the clarity or presentation of the data within the constraints of our current format.

Discussion:

Comment 15: Address the limitations more thoroughly. For example, discuss the potential impact of the low number of male cases on the generalizability of the findings and how future studies might address this imbalance.

Revision: The suggested revision was made on second last paragraph as follows:

Apart from the retrospective nature of our study, there were some limitations. One significant limitation is the low number of male cases, with only 8 male participants out of a total of 140 cases. This gender imbalance reflects the higher prevalence of osteoporosis in women but poses a challenge for the generalizability of our findings. The predominance of female cases may limit the applicability of our results to the male population, which could exhibit different patterns and responses to osteoporosis. Future studies should aim to include a larger and more balanced sample size with more male participants to ensure that the findings are generalizable to both genders. This approach would enhance the robustness and applicability of the results across different demographics. Instead of the hold-out cross-validation, the stratified 10-fold cross-validation technique was preferred as the number of cases was limited, and the use of hold-out cross-validation would further limit the data size of the dataset. The lack of external validation with an independent dataset is another important limitation. Although the determination of a maximum one-year interval between two examinations as the inclusion criterion is seen as a relatively long interval, it is understood that DEXA and MRI examinations of the same case were frequently taken at close intervals in the statistical analysis made for the month difference between the two

examinations (the median value of the month difference is 1 month, the interquartile range is 3.44 months).

**Response to Reviewer 4:**

Comment 1: Your dataset is small and consists mainly of women, which affects the generalizability of this study.

Response: We acknowledge the concerns regarding the small sample size and the predominance of female participants in our study. These factors indeed pose limitations on the generalizability of our findings. Osteoporosis is more prevalent among women, which explains the higher number of female cases in our dataset. However, we recognize the importance of including a more balanced gender distribution to enhance the applicability of the results across both sexes.

Future research should aim to include a larger and more diverse population. This approach will help achieve a more representative sample, ensuring that the findings are robust and generalizable. Additionally, collaborative studies with other institutions could provide a larger pool of participants, thereby addressing the gender imbalance and sample size constraints.

Comment 2: The normal and pathological groups are disproportionate, which may influence the study's findings.

Response: We acknowledge that the disproportionate representation of normal and pathological groups may influence our study's findings. The imbalance in group sizes could potentially affect the statistical power and the robustness of the results. However, we have applied advanced machine learning techniques and cross-validation methods to mitigate these effects and ensure the reliability of our outcomes.

The distribution of cases in our dataset is a result of our inclusion and exclusion criteria applied to the records available in our archives within the selected timeframe. Despite our efforts, the resultant dataset reflected the natural occurrence of cases, leaving limited scope for modification.

In future studies, collaborative efforts with other research institutions could provide access to larger and more balanced datasets, thereby improving the robustness of the study outcomes and enhancing their generalizability.

Comment 3: While using 10-fold cross-validation is appropriate for a small dataset, the selection of only five radiomic features increases the risk of overfitting. Additionally, the disproportionate subsets (normal and osteoporosis) might augment this effect, impacting machine learning performance. Despite

cross-validation providing a more appropriate estimation in the context of a small sample size, these concerns should be addressed.

Response: We appreciate your feedback regarding the potential risk of overfitting when selecting only five radiomic features, especially given the disproportionate subsets of normal and osteoporosis cases. While 10-fold cross-validation is a robust method for small datasets, we acknowledge that feature selection and dataset imbalance can influence machine learning performance.

To address these concerns, we have taken the following steps:

1. Advanced Feature Selection: We used the Fast Correlation-Based Filter (FCBF) method to select the most relevant radiomic features, minimizing redundancy and enhancing the robustness of the model.
2. Cross-Validation Approach: Our use of 10-fold cross-validation helps mitigate overfitting by ensuring that each data point is used for both training and validation, providing a more reliable estimate of model performance.
3. Performance Metrics: We reported multiple performance metrics (AUC, accuracy, sensitivity, specificity, F1 score, and recall) to provide a comprehensive evaluation of our models and to ensure that the selected features contribute to robust predictive performance.
4. Acknowledgment of Limitations: We explicitly discuss the limitations of our study, including the small dataset size and the disproportionate class distribution, and suggest that future studies with larger and more balanced datasets could further validate our findings.

We understand the importance of addressing these concerns and are committed to refining our methodologies in future research to enhance the generalizability and reliability of our results.

Comment 4: The choice of standardized parameters, such as bin width and filter kernel size, should be justified. Different settings could significantly influence the radiomic features.

Response: To justify the choice of standardized parameters such as bin width and filter kernel size, we aimed to ensure consistency and reproducibility across all cases in our study. Standardized parameters help in minimizing variability that could arise from different settings, which is crucial for reliable and comparative radiomic feature extraction.

Bin Width:

We used a bin width of 25 in the PyRadiomics module of the 3D Slicer application. This value is a default setting in the software and was chosen to

maintain standardization across all cases. The use of a consistent bin width ensures that the histogram-based features are comparable, reducing the impact of differing image acquisition parameters or patient anatomy on the extracted features.

Filter Kernel Size:

The choice of a 3 mm Laplacian of Gaussian filter size was based on literature recommendations and the need to capture relevant texture details without over-smoothing the image. A 3 mm filter size effectively balances the detection of fine details and noise reduction, which is essential for accurate radiomic feature extraction. Studies have shown that this kernel size is suitable for capturing relevant textural patterns in medical images while maintaining computational efficiency and accuracy (Lambin et al., 2012).

By standardizing these parameters, we aimed to reduce the variability in radiomic feature extraction and ensure that the features were robust and comparable across different cases. This standardization is particularly important in a study with a relatively small dataset, as it helps to maintain the integrity and reliability of the extracted features and subsequent machine learning analysis.

Comment 5: Extracting 5580 features is comprehensive, but it increases the risk of overfitting. The manuscript should discuss strategies to mitigate this risk.

Response: Extracting 5580 features indeed poses a risk of overfitting, especially with a relatively small dataset. To address this concern, we implemented several strategies to mitigate overfitting and ensure the robustness of our machine learning models.

Feature Selection and Reduction

We employed the Fast Correlation Based Filter (FCBF) method to reduce the number of features before applying machine learning algorithms. FCBF is an effective feature selection technique that identifies and retains the most relevant features while discarding those that are redundant or irrelevant. This approach helps in reducing the dimensionality of the data and minimizing the risk of overfitting by ensuring that only the most informative features are used for model training.

Cross-Validation

We used 10-fold cross-validation to evaluate the performance of our models. Cross-validation is a robust method for assessing the generalizability of a model, particularly with small datasets. It involves partitioning the data into ten subsets

and iteratively training and testing the model on different subsets. This method helps in providing a more reliable estimate of model performance and reduces the risk of overfitting by ensuring that the model is tested on multiple independent subsets of data.

Balanced Dataset

While we acknowledge the challenge of having a disproportionate number of nonosteoporotic and osteoporotic cases, we ensured that our feature selection and model validation methods are robust to such imbalances. The cross-validation method helps in balancing the training and testing subsets, providing a fair assessment of the model's performance across different data distributions.

By implementing these strategies, we aimed to mitigate the risk of overfitting and enhance the reliability of our findings. Future studies with larger and more balanced datasets can further validate the robustness of these models.

Comment 6: Only five features were selected using the FCBF method. While this reduces dimensionality, the manuscript should explain why five features were chosen and explore the implications of this choice.

Response: Thank you for your insightful comment. We selected five features using the Fast Correlation-Based Filter (FCBF) method to balance dimensionality reduction and retaining significant predictive power. The choice of five features was guided by preliminary experiments, which indicated that this number provided an optimal trade-off between model complexity and performance. Selecting too many features could increase the risk of overfitting, particularly with our relatively small dataset. Conversely, selecting too few features might overlook important information. We acknowledge the need for further studies to validate this approach with larger datasets and explore the impact of selecting different numbers of features on model performance.

Revision:

Methods:

"In this study, we used the Fast Correlation-Based Filter (FCBF) method to select five features from the extracted radiomic data. The choice of five features was based on preliminary experiments, aiming to balance the trade-off between reducing dimensionality and retaining significant predictive power. This approach helps mitigate the risk of overfitting by focusing on the most relevant features while avoiding the inclusion of redundant or less informative ones. We recognize that the optimal number of features may vary with different datasets and encourage further research to explore this aspect."

Discussion:

"Selecting five features using the FCBF method helped in reducing the dimensionality of the dataset while retaining the most relevant predictive information. This decision was based on preliminary analyses, which suggested that this number provided a good balance between model complexity and performance. While this approach reduces the risk of overfitting, it is essential to validate these findings with larger datasets. Future studies should explore the impact of selecting different numbers of features on the performance and generalizability of the models."

Comment 7: The assessment of data distribution normality is noted, but the manuscript should discuss how deviations from normality, especially for skewed distributions, are handled.

Response: Thank you for highlighting the need to address deviations from normality in the dataset. In our analysis, we assessed the normality of the data distribution for each feature. When deviations from normality were identified, particularly for skewed distributions, we employed appropriate statistical methods to manage these deviations. Specifically, non-parametric tests were used for statistical analyses that do not assume normal distribution, ensuring robust results.

Comment 8: The skewed gender distribution (94% female) limits the applicability of the results to the male population. This significant limitation should be acknowledged.

Response: Thank you for pointing out the skewed gender distribution in our study. We acknowledge that the high proportion of female participants (94%) limits the generalizability of our findings to the male population. This significant limitation is inherent due to the available sample from our institution's archives, which predominantly included female cases. Future studies should aim to recruit a more balanced gender distribution to enhance the applicability of the results across different populations.

Revision: Apart from the retrospective nature of our study, there were some limitations. One significant limitation is the low number of male cases, with only 8 male participants out of a total of 140 cases. This gender imbalance reflects the higher prevalence of osteoporosis in women but poses a challenge for the generalizability of our findings. The predominance of female cases may limit the applicability of our results to the male population, which could exhibit different patterns and responses to osteoporosis. Future studies should aim to include a larger and more balanced sample size with more male participants to ensure that the findings are generalizable to both genders. This approach would enhance the robustness and applicability of the results across different demographics.

Comment 9: The statistically significant differences in body weight and BMI between osteoporotic and non-osteoporotic groups are noteworthy. The manuscript should discuss how these factors might influence radiomic features and model performance. Additionally, consider discussing whether using Z-scores could minimize this influence.

Response: Thank you for your insightful comment regarding the differences in body weight and BMI between the osteoporotic and non-osteoporotic groups. These factors could indeed influence radiomic features and model performance. We will include a discussion on how body weight and BMI might affect our findings and consider the use of Z-scores as a potential method to minimize these influences.

Revision: "The statistically significant differences in body weight and BMI between the osteoporotic and non-osteoporotic groups are noteworthy. These factors could potentially influence the radiomic features extracted from MRI images, as body composition might affect tissue properties and, consequently, the radiomic signatures. Differences in body weight and BMI might also impact the performance of machine learning models by introducing bias related to these physical attributes. To mitigate these potential influences, future studies could consider using Z-scores for body weight and BMI, which would standardize these variables relative to the population mean and standard deviation. This approach could help minimize the impact of body composition differences on radiomic feature extraction and model performance. Additionally, including body weight and BMI as covariates in the analysis could provide more robust results, ensuring that the observed differences are not solely attributable to these factors."

Comment 10: The neural network showed a low AUC (0.616), F1 score (0.108), and recall (0.095), indicating poor performance despite being the best model with all features. A detailed discussion on the possible reasons and implications of this finding is needed.

Revision: The neural network model demonstrated low performance metrics, including an AUC of 0.616, an F1 score of 0.108, and a recall of 0.095, despite being identified as the best model when utilizing all features. Several factors may explain this outcome. Firstly, the significant class imbalance in the dataset likely biased the model toward the majority class, impairing its ability to accurately predict the minority class. Secondly, the small dataset size limited the model's capacity to generalize and effectively learn from the data. Neural networks typically require large volumes of data to achieve optimal performance, and the lack of sufficient data may have hindered the model's efficacy. Additionally, the complexity of neural networks means they are highly sensitive to hyperparameter settings, and inadequate tuning could result in suboptimal model configurations. Moreover, the model might have been prone to overfitting due to the high

dimensionality of the feature set and the relatively small sample size. Overfitting occurs when the model learns noise and details from the training data to an extent that negatively impacts its performance on new, unseen data. This can be particularly problematic in complex models like neural networks, which have a high capacity to fit the training data too closely. These findings highlight the importance of addressing class imbalance, ensuring sufficient data volume, performing rigorous hyperparameter optimization, and implementing techniques to prevent overfitting in future studies to enhance the predictive performance of neural networks in similar applications.

Comment 11: While the Naive Bayes model showed significant improvement after feature selection (AUC 0.913), the manuscript should discuss the potential reasons for this improvement and the clinical relevance of these findings.

Revision: The Naive Bayes model showed significant improvement after feature selection, achieving an AUC of 0.913. This improvement can be attributed to several factors. Feature selection likely removed irrelevant or redundant features, reducing the dimensionality of the data and allowing the model to focus on the most informative variables. This process enhances the model's ability to generalize from the training data to new, unseen data by minimizing the risk of overfitting. In the clinical context, this improvement is particularly relevant. A higher AUC indicates a better ability of the model to distinguish between osteoporotic and non-osteoporotic cases, which is crucial for early diagnosis and intervention. By identifying the most significant features, clinicians can better understand the underlying factors contributing to osteoporosis, leading to more targeted and effective treatment strategies. Additionally, the use of a more accurate model in clinical practice can improve patient outcomes by enabling more precise and reliable diagnostic processes. Thus, the feature selection process not only enhances the performance of the Naive Bayes model but also has meaningful implications for improving the clinical management of osteoporosis. Future studies should continue to explore and validate these selected features to ensure their robustness and relevance in diverse patient populations.

Comment 12: The high number of significant features (58 with AUC > 0.70) raises concerns about potential overfitting. This should be addressed in the manuscript.

Response: We understand the concern raised regarding the potential for overfitting with a high number of significant features. However, it is important to clarify that the identification of significant features with AUC > 0.70 was based on univariate analysis (Mann Whitney U test or independent samples T-test) and subsequent ROC analysis, not on the application of machine learning models at this stage.

Our intention was to highlight the individual features that demonstrated strong discriminatory power. Overfitting concerns typically arise in the context of model training, particularly when a model is overly complex relative to the amount of training data. However, in this scenario, the analysis was performed to identify promising features for further study, not to train a predictive model.

Therefore, while overfitting is a critical issue to consider in the development of machine learning models, it is not directly relevant to the univariate analysis of features presented in this section. This stage serves as a preliminary step to identify features that could potentially improve model performance when appropriately validated and used in a well-regularized machine learning framework.

## 2 Revision 1

2 comments · 13 Aug 2024 – 1 Sep 2024

**Referee 1** Reviewer

MINOR REVISION 13 Aug 2024

Dear Authors,
Thank you for addressing the initial comments and revising your manuscript. The clarity and completeness have notably improved, and your efforts are appreciated. However, a few areas still require further refinement to enhance the overall quality of the manuscript. Please consider the following suggestions:

1. Cohesive Connections and Structure: The "Methods" section has been presented more clearly and comprehensively. Nonetheless, some disconnection between sentences remains. To improve cohesion, I recommend reorganizing this section as follows:
“Methods: This retrospective study employed a cross-sectional design, with a total of 140 cases, including 21 with osteoporosis. The study's inclusion criteria consisted of concurrent lumbar MRI and DEXA within a year, while exclusion criteria included infectious or neoplastic lumbar lesions, fractures, instrumentation, significant osteodegenerative changes, cases where the first four lumbar vertebrae were not included in the imaging field, and absence of diffusion-weighted imaging. Manual segmentation of lumbar vertebrae from ADC maps was performed to create a comprehensive dataset, comprising 5,580 radiomic features per case. Subsequently, the top five features selected by FCBF were used to test the performance of seven machine learning algorithms (k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, and Neural Network). Statistical tests and ROC curve analysis were conducted to determine the significance and thresholds of these features.”

2. Demographical Data in the Abstract Results: The abstract results section has been clarified and now includes effect sizes of diagnostic values. For further improvement, please consider starting with a presentation of key demographic and clinical data of the sample. This will provide readers with better context and understanding of the findings and their representativeness.

3. Introduction Structure and Focus: The introduction now contains extensive information about osteoporosis, imaging modalities, and the potential of radiomics and machine learning. However, it could benefit from a more focused structure. I recommend revising the introduction according to the following order to improve clarity:
a) Introduction to Osteoporosis and its Clinical Importance: Briefly describe osteoporosis, its public health impact, and the importance of early diagnosis.
b) Problem Statement: Highlight the limitations of current diagnostic methods (e.g., DEXA) and the need for more accurate and comprehensive tools.
c) Introduction of Radiomics and ML: Present radiomics and machine learning as promising technologies that address the limitations of traditional methods by providing more detailed and accurate assessments of bone health.
d) Relevance and Significance: Emphasize the novelty and potential impact of your study in improving osteoporosis diagnosis and patient outcomes.
e) Study Objectives and Hypotheses: Clearly articulate the primary and secondary objectives of the study and the hypotheses being tested.

These revisions will further enhance the readability and focus of your manuscript. Thank you for your attention to these details, and I look forward to seeing the revised version.

**Editor-in-Chief** MINOR REVISION 1 Sep 2024

Dear authors,
Your article needs Minor revision.
Regards

3

# Revision 2

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

**Response to Reviewer**

Dear Reviewer,

We appreciate your constructive feedback and the opportunity to further improve our manuscript. Below, we have addressed each of your comments point by point:

**1. Cohesive Connections and Structure in the Methods Section**

**Comment:** The "Methods" section has been presented more clearly and comprehensively. Nonetheless, some disconnection between sentences remains. To improve cohesion, I recommend reorganizing this section as follows: “Methods: This retrospective study employed a cross-sectional design, with a total of 140 cases, including 21 with osteoporosis. The study's inclusion criteria consisted of concurrent lumbar MRI and DEXA within a year, while exclusion criteria included infectious or neoplastic lumbar lesions, fractures, instrumentation, significant osteodegenerative changes, cases where the first four lumbar vertebrae were not included in the imaging field, and absence of diffusion-weighted imaging. Manual segmentation of lumbar vertebrae from ADC maps was performed to create a comprehensive dataset, comprising 5,580 radiomic features per case. Subsequently, the top five features selected by FCBF were used to test the performance of seven machine learning algorithms (k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, and Neural Network). Statistical tests and ROC curve analysis were conducted to determine the significance and thresholds of these features.”

**Response:** We have reorganized the "Methods" section according to your suggestion to improve the flow and coherence of the content. The revised section now reads as follows:

**Revised Text:** "Methods: This retrospective study employed a cross-sectional design, with a total of 140 cases, including 21 with osteoporosis. The study's inclusion criteria consisted of concurrent lumbar MRI and DEXA within a year, while exclusion criteria included infectious or neoplastic lumbar lesions, fractures, instrumentation, significant osteodegenerative changes, cases where the first four lumbar vertebrae were not included in the imaging field, and absence of diffusion-weighted imaging. Manual segmentation of lumbar vertebrae from ADC maps was performed to create a comprehensive dataset, comprising 5,580 radiomic features per case. Subsequently, the top five features selected by FCBF were used to test the performance of seven machine learning algorithms (k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, and Neural Network). Statistical tests and ROC

curve analysis were conducted to determine the significance and thresholds of these features."

### 2. Demographical Data in the Abstract Results

**Comment:** The abstract results section has been clarified and now includes effect sizes of diagnostic values. For further improvement, please consider starting with a presentation of key demographic and clinical data of the sample. This will provide readers with better context and understanding of the findings and their representativeness.

**Response:** We have updated the abstract results section to include key demographic and clinical data at the beginning. This provides better context for the diagnostic values reported.

**Revised Text: " Results:** The study included 140 cases, with 132 females (94.3%) and 8 males (5.7%), and a mean age of 65.32 ± 8.50 years. The mean BMI was 31.43 ± 5.53 kg/m² for females and 26 ± 3.59 kg/m² for males. In terms of demographic differences, no significant age difference was found between the osteoporotic and non-osteoporotic groups (p=0.889). However, the osteoporotic group had significantly lower mean body weight (64.90 ± 10.13 kg vs. 74.68 ± 13.94 kg, p=0.003) and BMI (27.40 ± 4.38 kg/m² vs. 31.77 ± 5.52 kg/m², p=0.001) compared to the non-osteoporotic group. The median interval between DEXA and lumbar MRI was 1 month (range 0.1–11.87 months). The Neural Network model demonstrated the highest performance with an AUC of 0.616 and a classification accuracy of 0.764 using all features. The Naive Bayes model, using the top five features selected by FCBF, showed the highest performance with an AUC of 0.913, accuracy of 0.907, sensitivity of 0.667, and specificity of 0.95. All ML models' performance were elevated by feature selection. Independent t-tests and Mann-Whitney U tests identified 521 and 670 significant features, respectively ($p<0.05$). ROC analysis revealed 58 features with AUC values above 0.70."

### 3. Introduction Structure and Focus

**Reviewer Comment:** The introduction now contains extensive information about osteoporosis, imaging modalities, and the potential of radiomics and machine learning. However, it could benefit from a more focused structure. I recommend revising the introduction according to the following order to improve clarity: a) Introduction to Osteoporosis and its Clinical Importance: Briefly describe osteoporosis, its public health impact, and the importance of early diagnosis. b) Problem Statement: Highlight the limitations of current diagnostic methods (e.g., DEXA) and the need for more accurate and comprehensive tools. c) Introduction of Radiomics and ML: Present radiomics and machine learning as promising technologies that address the limitations of traditional methods by providing more

detailed and accurate assessments of bone health. d) Relevance and Significance: Emphasize the novelty and potential impact of your study in improving osteoporosis diagnosis and patient outcomes. e) Study Objectives and Hypotheses: Clearly articulate the primary and secondary objectives of the study and the hypotheses being tested.

**Response: We have revised the introduction section by deleting, summarizing and merging the parts containing extensive information, to follow the suggested structure, which improves the focus and clarity of the manuscript.**

**Revised Text: Background:**

Osteoporosis is a systemic skeletal disease characterized by decreased bone density and impairment in bone tissue microarchitecture, causing more fragile bones and a greater likelihood of fractures (1). Osteoporosis is a common public health problem worldwide, and it depends on many factors, such as ethnicity, gender, age, nutritional status, drug use, and diseases affecting bone density (1,2). Osteoporosis and osteoporotic fractures have undesirable consequences on the individual and society in terms of financial, psychosocial, physical, and quality of life (1,2). That is why early diagnosis of osteoporosis, especially before the development of fracture, is essential.

World Health Organization (WHO) adopted the bone mineral density (BMD) test with dual-energy X-ray absorptiometry (DEXA) as a reference standard in 1994 (3). The T score, which reflects how much standard deviation (SD) above or below the mean BMD of a young, healthy white woman, is used by WHO to identify osteoporosis. When the T score is -2.5 SD or less, osteoporosis is diagnosed (2). While DEXA is economical and non-invasive, it has low sensitivity to predict fractures and low-dose ionizing radiation (4,5). Furthermore, it has some limitations, such as the possibility that reference data for BMD provided by manufacturers may not be appropriate for local populations, being affected by body size, osteodegenerative changes, vascular calcifications, prior injected contrast media, fractures and deformities, 2-dimensional measurement of BMD, and operator dependency (4,6). Other imaging modalities such as Quantitative Computed Tomography (QCT), Quantitative Ultrasound (QUS), and High-Resolution Peripheral Quantitative Computed Tomography (HR-pQCT) emerged as an probable alternative to DEXA. QCT provides 3D imaging and accurate bone density measurements but is limited by higher radiation doses and costs. QUS is portable and radiation-free, suitable for screening but less accurate than DEXA and QCT. HR-pQCT offers detailed bone microarchitecture assessment, improving fracture prediction and monitoring therapy, but its use is restricted by high costs and limited availability (7–9).

Many studies in the literature have evaluated and showed the use of various magnetic resonance imaging (MRI) techniques in the diagnosis of osteoporosis. In osteoporotic bone, the proportion of adipose tissue increases as bone production declines, which is thought to be caused by the differentiation of stromal cells into adipocytes rather than osteoblasts (10,11). The amount of fat, protein, water, and bone cells determines how the bone marrow appears on the MRI sequence. (5,12–20). Among these studies, the apparent diffusion coefficient (ADC) values decreased in osteoporotic patients, which was hypothesized to be related to a decrease in diffusion due to the narrowing of the extracellular space as a result of increasing adipose tissue (5,12,15,16,19,20). ADC denotes a biological tissue's particular diffusion capacity, microscopic structure, and organization (21).

Radiomics and machine learning (ML) offer novel approaches for osteoporosis diagnosis by extracting high-dimensional data from medical images, which traditional methods might overlook. Radiomics can quantify tissue heterogeneity and capture subtle changes in bone microarchitecture, while ML algorithms can analyze these complex data sets to identify patterns indicative of osteoporosis (14,17,22,23). These techniques could complement existing diagnostic methods, such as DEXA, by providing more detailed and accurate assessments of bone health (22). Early and accurate detection of osteoporosis using these advanced techniques could potentially lead to better clinical outcomes by enabling timely intervention and personalized treatment plans, thus reducing the risk of fractures and associated complications.

The primary objective of this study is to evaluate the diagnostic performance of various machine learning models, specifically Naive Bayes and neural networks, in identifying osteoporosis using radiomic features extracted from lumbar MRI ADC maps. This will be measured using performance metrics such as Area Under the Curve (AUC), accuracy, sensitivity, specificity, F1 score, and recall. Additionally, the secondary objective is to identify the most significant radiomic features that contribute to accurate osteoporosis diagnosis. The hypotheses guiding this study are: 1- Machine learning models can significantly improve the diagnostic accuracy of osteoporosis compared to traditional methods; and 2- specific radiomic features from lumbar MRI ADC maps are significantly associated with osteoporosis and can reliably differentiate between osteoporotic and non-osteoporotic cases. These objectives and hypotheses are designed to provide a comprehensive evaluation of the potential of ML in enhancing osteoporosis diagnosis, with a focus on both model performance and feature significance.

We have implemented all the suggested revisions, which we believe have significantly enhanced the manuscript. Thank you for your valuable feedback and guidance.

Kind regards.

Authors

## 3 Revision 2

3 comments · 8 Sep 2024 – 8 Oct 2024

### Referee 1 Reviewer

ACCEPT 8 Sep 2024

The authors have successfully applied the recommended revisions to the manuscript, particularly addressing the clarity and cohesion in the "Methods" section, as well as the presentation of demographic data in the abstract results. Additionally, the introduction now follows a more focused and logical structure, improving the overall flow of the paper. These improvements significantly enhance the manuscript's quality, and I now believe it is suitable for publication in IJ Radiology.

### Associate Editor

ACCEPT 23 Sep 2024

The authors have successfully applied the recommended revisions to the manuscript, particularly addressing the clarity and cohesion in the "Methods" section, as well as the presentation of demographic data in the abstract results. Additionally, the introduction now follows a more focused and logical structure, improving the overall flow of the paper. These improvements significantly enhance the manuscript's quality, and I now believe it is suitable for publication in IJ Radiology.
Regards.

### Editor-in-Chief

MINOR REVISION 8 Oct 2024

Dear authors,
Your article needs Minor Revision.
Regards

4

# Revision 3

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

**Dear Associate Editor,**

Thank you for your valuable feedback on our manuscript. We have carefully addressed each of your comments and implemented the following revisions:

1. The term "radiomic" has been updated to "radiomics" throughout the manuscript.
2. The graphical abstract has been added as an appendix at the end of the paper.
3. The phrase in lines 82-83 has been corrected to: “in terms of financial, psychosocial, physical, and quality of life aspects (1,2).”
4. The title "Methods" has been changed to "Patients and Methods."
5. An "Objective" section has been added before the "Patients and Methods" section.
6. Sequential numbering has been added to all titles and subtitles.
7. The figure 1 legend has been updated to: "Patients’ Enrollment Diagram."
8. The phrase in line 167 has been corrected to: “T-score was >-2.5.”
9. The "x" character in line 176 has been replaced with the multiplication symbol (×) (225×225).
10. The word "stability" in line 306 has been changed to "reliability."
11. The sentence in lines 402-408 has been clarified and shortened for better readability.
12. Journal names in the references section have been abbreviated according to ISO standards.
13. Table 1 has been revised to include only the mean±SD of each variable in each group, along with the p-value; all other data have been removed.
14. The legend for Table 1 has been updated to: “Comparison of Age, Body Weight and BMI Between Osteoporotic and Non-Osteoporotic Subjects.”
15. The full terms of all abbreviations used in the tables have been added as footnotes.
16. Spaces have been inserted in Tables 2 and 5 where necessary (e.g., Sum Squares, first order, Large Area High Gray Level Emphasis, etc.).
17. The term "tree" has been replaced with "decision tree" in the legend for Table 3 and Figure 5.
18. The term "radiomics" has been correctly used in Table 5.
19. Full terms for abbreviations (e.g., SVM, KNN) have been added to Figure 5.
20. The X-axis in Figure 3 now includes "months" as the time unit.
21. The term "radiomics" has been corrected in Figure 4 and throughout the manuscript.

22. Full terms for all abbreviations (e.g., ADC, DEXA, FCBF, BMI, AUC) have been provided at their first usage, both in the abstract and the main body of the manuscript.

23. The phrase "lumbar MRI ADC map radiomics" has been revised for clarity to: “radiomics features of lumbar MRI ADC maps.”

24. The sentence in line 12 has been corrected to: “Objectives: In this study, we are going to evaluate the performance of….”

We have attached the revised manuscript for your review. We truly appreciate your constructive feedback and hope these revisions meet the journal's requirements.

Thank you once again for your time and consideration.

Sincerely,
Authors

## 4 Revision 3

1 comment · 21 Oct 2024

### Editor-in-Chief

ACCEPT 21 Oct 2024

Dear authors,
Your article is acceptable for publication in IJR.
Regards