Deep Learning-Based Segmentation of the Lumbar Spine in DXA Images for Osteoporosis Assessment

Martino Ferrero[1,2], Mailen Gonzalez[1,2][0009−0000−2102−9222], José Manuel Fuertes García[3][0000−0001−6624−4102], and José María Massa[1][0000−0002−7456−9676]

[1] INTIA, Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Argentina
martinoferrero02@gmail.com
[2] Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina
[3] Departamento de Informática, Escuela Politécnica Superior, Universidad de Jaén, 23071 Jaén, Spain
jmf@ujaen.es

Abstract. Accurate segmentation of vertebral regions in dual-energy X-ray absorptiometry (DXA) scans is essential for reliable estimation of clinical indicators used in osteoporosis assessment, including bone mineral density. This work presents a preliminary U-Net-based study for the automatic segmentation of the lumbar vertebral region L1–L4 in DXA images. A U-Net architecture was trained using anonymised lumbar spine DXA images with expert-validated binary masks and was compared against the automatic segmentation produced by the densitometer's enCORE software. In a preliminary evaluation on eight reserved test cases, the proposed model achieved mean Dice, IoU, sensitivity, and accuracy values of 0.982, 0.965, 0.987, and 0.994, respectively, compared with 0.930, 0.869, 0.873, and 0.977 for the enCORE-derived masks. These results provide preliminary evidence that convolutional neural networks may improve agreement with expert lumbar spine annotations in DXA image analysis, particularly in cases where rule-based or software-generated regions of interest underestimate the vertebral area. However, the current study is limited by the small dataset size and requires validation on larger and more diverse clinical cohorts before drawing conclusions about generalisation or downstream diagnostic impact.

Keywords: DXA · Osteoporosis · Lumbar spine segmentation · U-Net · Deep learning · Medical image analysis

1 Introduction

Dual-energy X-ray absorptiometry (DXA) is widely used for assessing bone mineral density (BMD) and plays an important role in the diagnosis and monitoring of osteoporosis. Among the anatomical regions commonly evaluated, the lumbar

spine, particularly vertebrae L1 to L4, is clinically relevant due to its sensitivity to early degenerative bone changes and its role in standard osteoporosis assessment protocols [1]. Accurate delineation of the vertebral bodies is essential for obtaining reliable BMD measurements and for extracting additional structural indicators such as trabecular bone score (TBS), which provides indirect information about bone microarchitecture.

Small inaccuracies in the segmentation of vertebral regions can propagate through the DXA analysis pipeline and may affect BMD estimation. In clinical practice, these deviations are relevant because BMD values contribute to diagnostic categorisation, treatment decisions, fracture risk estimation, and longitudinal follow-up. Therefore, improving the robustness and reproducibility of vertebral region segmentation is an important step toward more reliable osteoporosis assessment.

Commercial DXA systems, including GE Healthcare devices using enCORE software, provide automatic or semi-automatic tools for defining vertebral regions of interest. These tools may perform adequately under standard imaging conditions, but their performance can be affected by anatomical variability, low contrast, spinal deformities, artefacts, or suboptimal positioning. Prior work has described technical and patient-related sources of error in DXA acquisition and interpretation, including issues related to region-of-interest placement and image artefacts [2]. This motivates the exploration of data-driven segmentation methods that can learn relevant anatomical patterns directly from examples.

Deep learning methods, particularly convolutional neural networks (CNNs), have achieved strong results in biomedical image segmentation. U-Net has become a reference architecture for medical segmentation tasks due to its encoder–decoder design and skip connections, which allow the model to combine contextual and spatial information [5]. In spinal imaging, U-Net-based and related architectures have been applied to vertebral segmentation in computed tomography, magnetic resonance imaging, and radiographic contexts [6–8]. Nevertheless, the use of expert-validated masks for lumbar spine segmentation in DXA remains comparatively less explored.

This work presents a preliminary feasibility study on automatic L1–L4 lumbar spine segmentation in DXA images using a standard but adapted U-Net architecture. Unlike approaches that rely exclusively on software-generated annotations, the proposed method is trained and evaluated using expert-segmented reference masks. The main contributions of this work are: (i) a processing pipeline for comparing commercial enCORE segmentations with expert masks; (ii) an implementation of a reduced-width U-Net tailored to the small DXA dataset available in this study; and (iii) a preliminary quantitative and qualitative evaluation showing improved overlap with expert annotations compared with enCORE-derived masks.

2 Related Work

DXA-based assessment depends on both image acquisition quality and correct definition of the anatomical region of interest. Technical artefacts, patient positioning, degenerative changes, and ROI placement errors can influence the measured values and their interpretation [2]. This concern is also reflected in studies focused on quality assurance and reporting errors in bone mineral densitometry. Martineau et al. emphasized that DXA-related errors may be technical, interpretative, or both, and may lead to inappropriate diagnosis and patient management [3]. Messina et al. reported that 451 out of 485 DXA examinations or reports presented at least one error, with most errors associated with data analysis rather than acquisition alone [4]. These limitations are especially relevant in lumbar spine analysis, where vertebral morphology, contrast, and superimposed structures may complicate robust segmentation. Therefore, robust and reproducible lumbar ROI segmentation is closely related to reducing analysis-dependent variability in DXA-based assessment.

Deep learning has been extensively studied for medical image segmentation. U-Net [5] remains a common baseline due to its strong performance in low-data biomedical scenarios. Several works have adapted U-Net-like models to spine analysis in CT and MRI, including two-stage Dense-U-Net approaches for vertebral localisation and segmentation [6], 3D U-Net-based analysis in lumbar MRI [7], and attention or multiscale variants for spine MRI segmentation [8].

In the context of DXA and bone densitometry, previous studies have explored deep learning for bone segmentation and osteoporosis-related assessment [9–12]. However, further work is needed to evaluate such methods against expert-defined masks and to analyse their behaviour when compared with commercial software outputs. The present study contributes to this direction by focusing on lumbar L1–L4 segmentation in DXA using expert-validated masks as reference.

3 Materials and Methods

3.1 Dataset

The dataset comprised 25 anonymised lumbar spine DXA scans acquired with a GE Lunar Prodigy Advance device. All images available for this study were acquired with the same device model; therefore, the dataset does not allow assessment of generalisation across manufacturers, scanner models, or clinical centres. Each case included an unprocessed DXA image (RAW) and a corresponding image containing visual overlays generated by the densitometer's enCORE software. The overlay image included the software-defined lumbar region of interest and horizontal grid lines separating vertebral levels L1 to L4. The images presented variable spatial dimensions across cases. The image width ranged from 268 to 300 pixels, while the image height ranged from 258 to 398 pixels. Considering the complete image size, the smallest observed resolution was 268×285 pixels and the largest was 300×398 pixels.



(a) RAW DXA image. (b) enCORE overlay.

Fig. 1: Example lumbar spine DXA case. The RAW image is used as model input, while the overlay image is processed to extract the commercial software segmentation for comparison.

Detailed clinical metadata such as age, body mass index, osteoporosis category, and presence of artefacts or deformities were not available in the anonymised dataset. Consequently, no stratified analysis according to clinical status could be performed.

Figure 1 shows an example RAW image and its corresponding enCORE overlay.

3.2 Reference Masks and enCORE-Derived Masks

Expert binary masks were generated by manually segmenting the lumbar L1–L4 region in the RAW images and were subsequently reviewed before being used as reference standard. These masks were used as the reference standard for training and evaluation. No interobserver variability analysis was performed in this preliminary study. In parallel, the enCORE overlay images were processed to derive binary masks corresponding to the commercial software segmentation. First, the lumbar contour was isolated and binarised. Then, the horizontal grid lines were detected, and the region between the first and last relevant lines was extracted to define the L1–L4 region. Because the enCORE masks were derived from visual overlays rather than exported binary masks, this extraction step may introduce minor contour inaccuracies. The comparison should therefore be interpreted as an evaluation against the available enCORE-derived segmentation representation.

3.3 Preprocessing and Data Augmentation

The RAW images and their corresponding masks were first cropped to ensure that the lumbar region of interest was contained within the selected image area. Since the selected region was smaller than 256 pixels in both spatial dimensions

in all cases, zero-padding was then applied to obtain a common input size of 256 $\times$ 256 pixels while preserving the relevant pixels corresponding to the L1–L4 region. Pixel intensities were normalised to the range $[0, 1]$ before training.

A conservative online data augmentation strategy was applied to reduce overfitting in this small-data setting. The final augmentation pipeline included random brightness variation with a multiplicative factor sampled from the interval $[0.8, 1.2]$, and random horizontal flipping with probability $p = 0.5$. More aggressive geometric augmentation strategies, such as rotations, zoom transformations, elastic deformations, or noise injection, were not included in this experiment.

3.4 Segmentation Model

The segmentation model follows the U-Net encoder–decoder architecture, which is a widely used baseline for biomedical image segmentation. The encoder progressively reduces spatial resolution while increasing the number of feature channels, whereas the decoder restores spatial resolution and combines high-level contextual information with fine-grained spatial features from the encoder. The final layer produces a binary mask representing the segmented L1–L4 region.

The architecture used in the reported experiments followed a four-level U-Net design with an additional bottleneck stage. Each encoder and decoder level contained two main 3×3 convolutional layers followed by batch normalisation and ReLU activation, resulting in 18 main convolutional layers: eight in the encoder, eight in the decoder, and two in the bottleneck. A final 1×1 convolution was used to map the decoder output to the binary segmentation output. The number of feature channels started at 32 in the first encoder block and was doubled after each downsampling step, reaching 512 channels at the bottleneck. This reduced-width configuration was selected instead of the classical 64–1024 U-Net configuration to decrease the number of trainable parameters and reduce overfitting risk in the small-data setting.

All 3×3 convolutions used stride 1 and padding 1, preserving the spatial size within each convolutional block. Therefore, changes in feature-map resolution were produced only by the downsampling and upsampling operations. Downsampling was performed using 2×2 max pooling. In the decoder, upsampling was implemented using bicubic interpolation followed by a 1×1 convolution, instead of transposed convolutions, to avoid checkerboard artefacts and keep the upsampling operation parameter-efficient. Dropout with a rate of 0.2 was included as a regularisation mechanism.

Figure 2 summarizes the U-Net architecture used in the reported experiments.

3.5 Training Strategy

Of the 25 available cases, 17 were used for training and validation using four-fold cross-validation, while the remaining 8 cases were reserved for final testing. Due to the absence of complete clinical metadata and the small sample size,

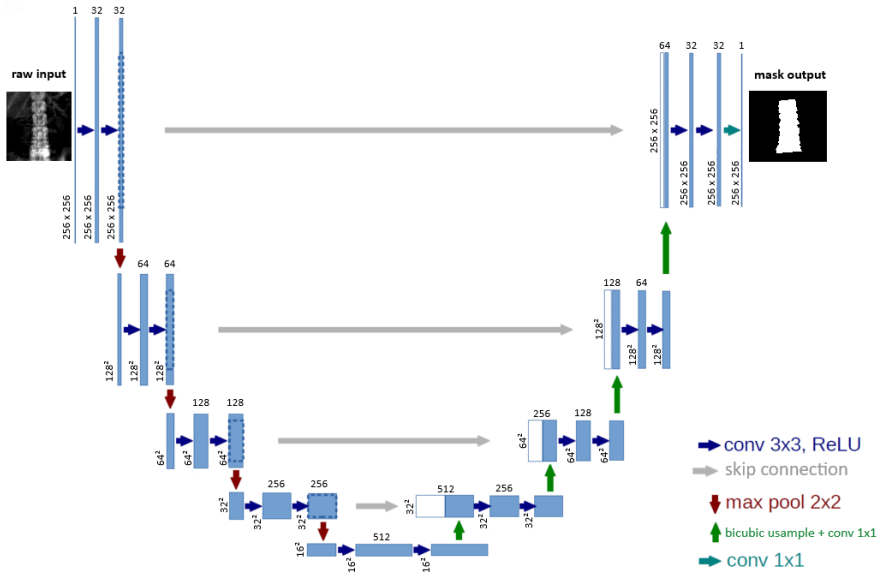


Fig. 2: U-Net architecture used for lumbar spine segmentation.

the split was not clinically stratified. The model was trained using binary cross-entropy with logits as the loss function and the Adam optimiser. The learning rate followed a One-Cycle policy. Weight decay, dropout, and data augmentation were used to reduce overfitting.

Cross-validation was used to evaluate the training configuration and assess its behaviour across different training/validation splits. After this stage, the final model evaluated on the independent test set was trained once using all 17 cases from the training/validation subset. The 8 reserved cases were kept separate throughout the process and were used only for the final evaluation. Given the limited number of cases, the final estimates are expected to be sensitive to the specific test-set composition.

The One-Cycle learning-rate schedule started at 10^{-4}, increased up to 10^{-2}, and then decreased to 10^{-5} during training. The model was trained for 750 epochs with a batch size of 2, dropout rate of 0.2, and weight decay of 10^{-4}. Since the final model was trained on all 17 training/validation cases, no validation-based early stopping was applied during this final training stage; regularisation relied on weight decay, dropout, augmentation, and the training configuration selected during cross-validation. All reported experiments were run using a fixed random seed of 42. The implementation was based on PyTorch 2.6.

A post-processing step was applied to remove small spurious segmented regions. Connected regions occupying less than 12% of the image area were discarded. This threshold was defined empirically from area proportions observed outside the spinal column in the development data, with the aim of removing

Table 1: Training configuration used for the reported experiments.

Parameter	Value
Input size	256×256
Model	U-Net
Loss function	Binary cross-entropy with logits
Optimiser	Adam
Learning-rate schedule	One-Cycle policy
Learning-rate range	$10^{-4} \rightarrow 10^{-2} \rightarrow 10^{-5}$
Batch size	2
Epochs	750
Dropout	0.2
Weight decay	10^{-4}
Random seed	42
Framework	PyTorch 2.6
Cross-validation	4 folds on 17 training/validation cases
Final training	Single model trained on all 17 training/validation cases
Final test set	8 reserved cases

small non-column components while preserving the L1–L4 region. The empirical nature of this threshold is a limitation, and future work should optimise or validate post-processing parameters using larger validation cohorts.

3.6 Evaluation Metrics

Segmentation performance was evaluated using Dice coefficient (DSC), intersection over union (IoU), pixel-wise accuracy, sensitivity, and specificity. These metrics were computed by comparing each predicted binary mask with the corresponding expert reference mask. Dice and IoU measure spatial overlap, sensitivity quantifies the proportion of vertebral pixels correctly detected, and specificity quantifies the proportion of background pixels correctly classified. All metrics were computed independently for each test image and then averaged across the reserved test set, following a macro-averaging strategy. Mean and standard deviation values were computed across the eight reserved test cases.

4 Results

The proposed U-Net model was evaluated on the eight reserved test cases and compared with the enCORE-derived masks. Table 2 reports the mean and standard deviation values across the eight reserved test cases.

Within this limited test set, the U-Net model achieved higher mean Dice coefficient, IoU, accuracy, and sensitivity than the enCORE-derived masks. The standard deviation values provide a descriptive estimate of variability across the eight test cases. The improvement in sensitivity indicates that the model detected a larger proportion of the expert-defined vertebral region, reducing the

Table 2: Mean and standard deviation of segmentation performance between enCORE-derived masks and the proposed U-Net model on the reserved test set.

Metric	enCORE software	U-Net model
Dice coefficient	0.930 ± 0.020	0.982 ± 0.006
IoU	0.869 ± 0.035	0.965 ± 0.012
Accuracy	0.977 ± 0.006	0.994 ± 0.002
Sensitivity	0.873 ± 0.050	0.987 ± 0.008
Specificity	0.999 ± 0.001	0.995 ± 0.003

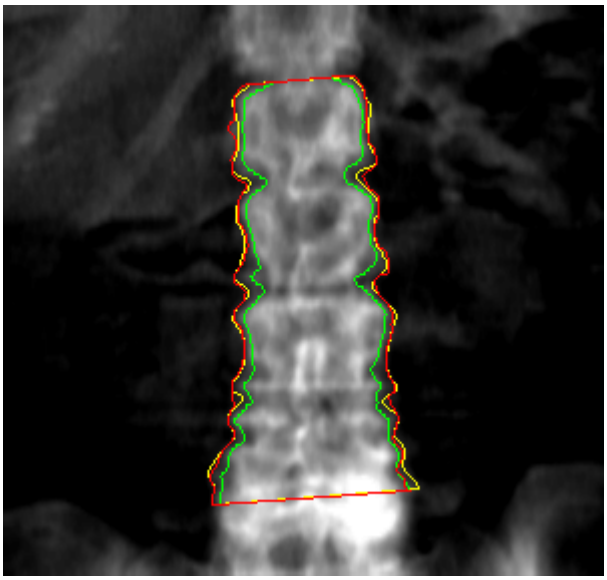


Fig. 3: Representative qualitative comparison on a randomly selected test image. Ground truth is shown in yellow, the U-Net prediction in red, and the enCORE-derived segmentation in green.

under-segmentation observed in the commercial software output. The enCORE-derived masks achieved slightly higher specificity, which is consistent with a more conservative segmentation that classifies most background pixels correctly but may omit relevant vertebral areas.

Figure 3 shows one randomly selected qualitative example to illustrate the typical under-segmentation pattern observed in the enCORE-derived mask, comparing the contours produced by the enCORE-derived mask, the U-Net prediction, and the expert reference mask.

5 Discussion

The preliminary results suggest that, in the analysed test cases, the U-Net-based approach produced lumbar spine segmentations with higher agreement with expert annotations than the enCORE-derived masks. The main difference was observed in sensitivity, suggesting that the proposed model reduced under-segmentation of the vertebral region. This is relevant because omitted vertebral pixels may affect downstream quantitative analyses based on the segmented region.

Nevertheless, these findings must be interpreted cautiously. The dataset is small, and the reserved test set contains only eight cases. Therefore, the reported results should be considered preliminary evidence of feasibility rather than definitive proof of clinical generalisation. Data augmentation and cross-validation can help reduce overfitting and provide a more robust estimate within the available data, but they do not replace external validation on larger and more diverse cohorts. Although standard deviations are reported, the absence of confidence intervals, statistical significance tests, and detailed per-case analysis limits the strength of the quantitative conclusions that can be drawn from these results.

Another important consideration is that this work evaluates segmentation accuracy against expert masks, not the final clinical impact on BMD, TBS, diagnostic classification, or fracture-risk prediction. Consequently, claims should be limited to segmentation performance. Future studies should quantify how changes in segmentation affect downstream DXA-derived clinical indicators.

6 Limitations and Future Work

This study has several limitations. First, the dataset size is small, with only 25 cases and 8 reserved test cases, which restricts the ability to assess generalisation and makes the reported mean and standard deviation values sensitive to individual cases. Second, the evaluation was performed on images acquired with a single device model, so external validation across manufacturers, scanner models, acquisition protocols, and clinical centres is required. Third, detailed clinical metadata were not available, preventing stratified analysis by osteoporosis status, age, body mass index, artefacts, or anatomical variability. Finally, the impact of the improved segmentation on downstream BMD estimation, TBS, diagnostic classification, or fracture-risk assessment was not directly evaluated.

Future work will focus on increasing the dataset size, including additional devices and clinical centres, evaluating advanced U-Net variants, and extending the method to individual vertebral segmentation. Additional experiments should also assess the relationship between segmentation differences and downstream clinical indicators such as BMD, T-score, and TBS. The automatic segmentation of the hip and femoral neck regions is another relevant direction for future research.

7 Conclusion

This paper presented a preliminary deep learning-based approach for automatic segmentation of the L1–L4 lumbar spine region in DXA images. Using a U-Net architecture trained with expert-validated masks, the proposed method achieved higher Dice coefficient, IoU, sensitivity, and accuracy than enCORE-derived masks in the available test set. These results suggest that CNN-based segmentation may help address under-segmentation observed in the available enCORE-derived masks, but only as preliminary feasibility evidence. However, due to the limited dataset size, larger external validation and downstream clinical evaluation are required before making claims about robustness, generalisation, downstream diagnostic impact, or clinical integration.

Acknowledgements This work was supported by the National Scientific and Technical Research Council (CONICET). The authors thank the Instituto de Diagnóstico e Investigaciones Metabólicas (IDIM) for providing the DXA images used in this study, and the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA) for institutional support.

Disclosure of Interests The authors have no competing interests to declare.

References

1. Slart, R., et al.: Updated practice guideline for dual-energy X-ray absorptiometry (DXA). European Journal of Nuclear Medicine and Molecular Imaging 52(2), 539–563 (2025)
2. Qutbi, M., et al.: Technical and patient-related sources of error and artefacts in bone mineral densitometry using dual-energy X-ray absorptiometry: A pictorial review. Indian Journal of Radiology and Imaging 30(3), 362–371 (2020)
3. Martineau, P., Morgan, S.L., Leslie, W.D.: Bone Mineral Densitometry Reporting: Pearls and Pitfalls. Canadian Association of Radiologists Journal 72(3), 490–504 (2021). https://doi.org/10.1177/0846537120919627
4. Messina, C., Bandirali, M., Sconfienza, L.M., D'Alonzo, N.K., Di Leo, G., Papini, G.D.E., Ulivieri, F.M., Sardanelli, F.: Prevalence and type of errors in dual-energy x-ray absorptiometry. European Radiology 25, 1504–1511 (2015). https://doi.org/10.1007/s00330-014-3509-y
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer, Cham (2015)
6. Cheng, P., Yang, Y., Yu, H., He, Y.: Automatic vertebrae localization and segmentation in CT with a two-stage Dense-U-Net. Scientific Reports 11, 22156 (2021)
7. Domi, A., et al.: Automated MR lumbar vertebrae analysis: 3D U-Net and finite element analysis for precise segmentation and robust bone strength assessment. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–4. IEEE (2024)
8. Wang, B., et al.: MLKCA-Unet: Multiscale large-kernel convolution and attention in U-Net for spine MRI segmentation. Optik 272, 170277 (2023)

9. Zhou, L., Nguyen, T., Choi, S., Yoon, J.: U-Net-based deep learning hybrid model: Research and evaluation for precise prediction of spinal bone density on abdominal radiographs. Bioengineering 12(4), 385 (2025)
10. Hussain, D., Naqvi, R.A., Loh, W.K., Lee, J.: Deep learning in DXA image segmentation. Computers, Materials & Continua 66, 2587–2598 (2021)
11. Fathima, S.M.N., Tamilselvi, R., Beham, M.P., Nagaraj, A.: A deep learning approach on segmentation of bone for BMD measurement from DEXA scan images. In: 2020 Sixth International Conference on Bio Signals, Images, and Instrumentation (ICBSII), pp. 1–5. IEEE (2020)
12. Fathima, S.M.N., Tamilselvi, R., Beham, M.P., Sabarinathan, D.: Diagnosis of osteoporosis using modified U-Net architecture with attention unit in DEXA and X-ray images. Journal of X-Ray Science and Technology 28(5), 953–973 (2020)