On How to Unravel Bone Microscale Phenomena: A Mask-Guided Attention SR-microCT Image Classification Approach

Isabella Poles, Eleonora D'Arnese, Federica Buccino, Laura Vergani, and Marco D. Santambrogio
Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy
{isabella.poles, eleonora.darnese, federica.buccino, laura.vergani, marco.santambrogio}@polimi.it

***Abstract*— The global increase in elderly individuals has led to a rise in fragility fractures and chronic aging-related diseases, including osteoporosis. In this context, Deep Learning (DL) offers the potential to analyze bone images to aid researchers and clinicians in studying its health starting from the microscale. Previous studies demonstrate the effectiveness of DL in segmenting lacunae and classifying bone tissue microstates from Synchrotron-Radiation micro-Computed Tomography (SR-microCT) images. However, the generalizability of these models, the laborious work in labeling tiny structures in high-dimensional images, and the low inter-class variance in SR-microCT images remain a concern. To fill this void, this paper proposes a Mask-Guided Attention (MGA) approach that combines semi-supervised learning lacunae segmentation and attention methods for healthy and osteoporotic SR-microCT image classification. In particular, semi-supervised learning aims at reducing the number of labeled images required during segmentation. At the same time, the MGA approach exploits the pseudo-labels predicted to focus the network's attention on the informative lacunar structures. Our strategy allows achieving up to 5.64% and 12.17% accuracy improvements over *de-facto* lacunae image segmentation and image classification methods, as well as more interpretable results.**

***Clinical relevance*—The proposed MGA approach could enhance the understanding of bone microscale phenomena by exploiting SR-microCT images, supporting the study and the diagnosis of osteoporosis in individuals.**

I. Introduction

In the last decades, there has been a notable demographic increase of elderly individuals within the global population. This change has raised the urge to draw attention to the increasing occurrence of fragility fractures and chronic diseases as aging-related issues. In particular, osteoporosis stands out as a prominent contributor, leading to a substantial reduction in bone mineral density, especially in women [1]. Despite the numerous studies targeting the prevention of bone fracture, its comprehension is still limited to macro and mesoscale bone, where its occurrence is destructive [2], [3]. However, the role of microscale bone structures, such as the lacunar bone microporosities, is yet to be elucidated [4]. Therefore, the interest in bone microscale phenomena emerges in precisely defining the pathology-induced alterations at the scale where pharmacological treatments act to establish effective and targeted preventive strategies [5].

Preliminary attempts to shed some light on bone microscale mechanisms exploit high-resolution imaging techniques such as stereomicroscopy, scanning electron microscopy, computed micro-Tomography, and Synchrotron-Radiation micro-Computed Tomography (SR-microCT) [6]. Between them, SR-microCT electron beam images lacunae and microcracks at an unprecedented resolution of $\sim 1.6\mu$m thanks to its phase contrast and the *in-situ* simultaneous mechanical testing [7]. Despite SR-microCT potential, several image analysis criticalities emerge from thresholding approximations for lacunar detection to interpretability limitations that make microscale fine-grained differences between samples unrecognizable by human eyes.

In this complex scenario, Deep Learning (DL) shows great potential in the analysis of medical images as a tool for aiding clinicians in evaluating bone health at the microscale [8]. *Buccino et al.* have exploited a Convolutional Neural Network (CNN) to semantically segment lacunae and microcracks of healthy and osteoporotic patched SR-microCT human images [9]. However, the datasets employed for training, validation, and testing purposes consisted of distinct image slices obtained from the same set of patients, a choice that raised concerns about the model's generalizability to unseen patients' images. Furthermore, *Shen et al.* have successfully employed CNNs to classify mechanical states of cortical and trabecular bone tissue from SR-microCT images, but confined to healthy bovine bones [10]. Further problems arise when considering the high intra-class and small inter-class variances of microscale images that lead even DL methods to struggle in identifying visual differences [11]. Networks with attention layers can relieve this problem by suppressing the noisy areas from the final decision-making process using hard or soft masking mechanisms, making the network decision process more transparent and explainable [12], [13]. Nevertheless, while image segmentation and classification labeling need to be performed by a trained expert, generating manually fine boundaries of many small-sized bone structures across various high-dimensional images is more arduous because it increases the fatigue risk, compromising consistency and quality of segmentation. Finally, although DL has been used for SR-micro CT image analysis, no efforts have been made to interpret which bone structure imaged the model concentrates on.

To overcome these limitations, we propose a Mask-Guided Attention (MGA) approach to localize the discriminative lacunar image regions for fine-grained healthy and osteoporotic SR-microCT image classification. Our approach, summarized in Figure 1, consists of a semi-supervised learning

Project funded by Polisocial Award 2022 - Politecnico di Milano

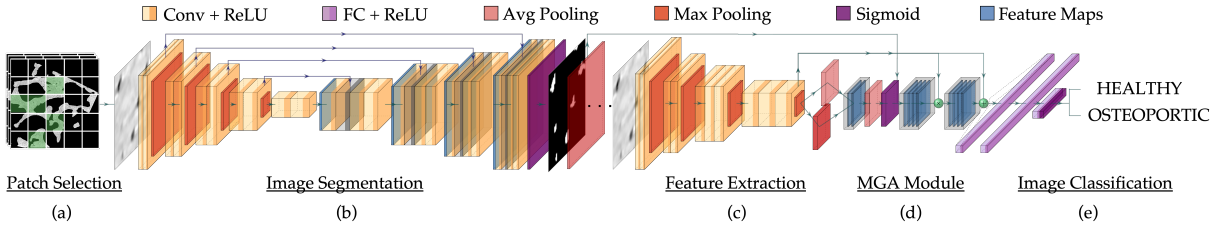

Fig. 1: High-level description of the proposed methodology from patch-selection, to image segmentation and classification.

approach to let a U-Net semantically segment the lacunae, an attention module that exploits the predicted masks to guide the vgg16 to classify healthy from osteoporotic images even with fined-grained differences, and the visualization of the class-defining features using GradCAM [14]. The main contributions of our work are the following: **(i)** a semi-supervised learning approach to perform semantic segmentation of lacunae in SR-microCT images to reduce the number of labeled images needed; **(ii)** a vgg16 integrated with an MGA module to classify healthy and osteoporotic bone SR-microCT images; **(iii)** a strategy to improve the classifier awareness in detecting lacunar structure when distinguishing subtle inter-class SR-microCT image differences.

II. Proposed Methodology

In this Section, we present an in-depth description of the proposed solution from the dataset, its pre-processing, the image segmentation, and the image classification method.

A. Dataset Description and Selection Criteria

This study exploits SR-microCT images from trabecular samples from three healthy and three osteoporotic human femoral heads. These samples were imaged using a synchrotron and captured at different strain states for each anatomical direction: Z1 in the frontal plane, Z2 in the coronal plane, and Z3 in the posterior plane. The testing phase consists of preloading the sample from the zero-load position using three compression step cycles of a motor plunger. The displacement during these cycles was constant, while SR-microCT images were acquired simultaneously. A comprehensive set of 3D images comprising 2048 slices per volume was obtained for each scan during loading with a field of view measuring $3.28 \times 3.28 \times 2$ mm. Each side of the resulting images had an average dimension of 3300 pixels and pixel size of 1.6μm. The resulting dataset consisted of single-channel image volumes containing 32-bit per pixel intensity information. A selection criterion was applied to reduce the dataset's high dimensionality, including 440 and 4 images per patient for the classification and lacunae image segmentation tasks, respectively. Nonconsecutive image slices were selected from each image volume, ensuring correspondence between patients. Furthermore, image volumes exhibiting video intensity artifacts were excluded to mitigate noise disturbances, while the pixels belonging to lacunar structures visible from the images selected for the segmentation task were manually labeled by two operators. Finally, to avoid having the same patient's images appear in both the training and validation/testing datasets, we trained all the models by dividing the image patches patient-wise in an 80/10/10 pattern given batches of 30 randomly shuffled image patches and related labels.

B. Data Pre-processing

We design a pre-processing phase with an image enhancement stage to increase the bone structure image contrast and a patch extraction stage to reduce the memory footprint. The image enhancement process starts with identifying two distinct image sets characterized by different gray level ranges: [0, 1] and [-1, 0], the latter arising from image acquisition errors. Therefore, we implemented an enhancement pipeline consisting of three main steps: pre-enhancement, image segmentation, and morphological post-processing, where the first and last steps are common to both image sets. First, we employ *min-max* normalization to ensure optimal consistency across various acquisition methods and textures. Moreover, we effectively enhance the output contrast by saturating the bottom and top 1% of pixel values. The second step focuses on segmenting bone structures, where we utilize for the [0, 1] range the Otsu method, while the K-means pixel clustering method for the second image set. Finally, in the last step, we smooth the segmentation contours and prevent voids by employing binary morphological operations with square- and disk-shaped structuring elements for opening and closing operations, respectively. Once the binary segmentation is obtained, we apply the mask to the previously adjusted images, setting background pixels to 0-values while retaining the adjusted image content within the bone region.

In the patch extraction phase, we extract fixed-size patches by sliding a size $k \times k$ patch with a stride of s over images (Figure 1(a)). This process makes a total number of $[1+W-ks] \times [1+H-ks]$ patches where W and H are image width and height, respectively. We choose $k = s = 55$ in our experiments, considering the available GPU memory. Finally, we exploited the bone structure image masks to retain the sole patches that contained at least the 99% of bone tissue and guaranteeing at least one patch per image volume slice.

C. Semi-Supervised Image Lacunae Segmentation

The encoder-decoder 2D U-Net with skip-connections inspired the proposed neural network for binary segmenting lacunar structures [15]. Figure 1(b) shows the U-Net CNN with its 13 layers. The encoder consists of two repeated 3×3 convolutions, each followed by a LeakyReLU activation function and a 2×2 max pooling operation with $s = 2$ for downsampling. The decoder consists of a feature map upsampling followed by a 2×2 convolution that halves the

number of feature map channels, a concatenation with the cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a LeakyReLU. A final 1×1 convolution maps each 64-component feature vector to the desired number of classes. We let the U-Net learn in a semi-supervised mode. In particular, we trained and validated the first U-Net model (U-Net-1) on the subset of 8 fined labeled images, corresponding to ~ 584 patches. After convergence, we exploit the knowledge gained to generate pseudo-labels to use as ground truth in the second image subset ($\sim 1.2k$ patches). Finally, we trained and validated the same U-Net model (U-Net-2) on a set of 8 manually labeled, and 8 pseudo-labeled images. We trained both models from scratch using a combination of the Dice Similarity Coefficient (DSC) and focal losses on foreground pixels by masking the loss to address the label imbalance problem between the foreground lacunar structures and the background. The network takes batches of 30 randomly shuffled image patches and related labels while Adam optimizes the training with a learning rate exponentially decaying from $1e-3$ for 300 epochs.

D. Attention-based Image Classification

The proposed classification architecture can be divided into three parts. A CNN to extract the image features Figure 1(c), which are then input together with their U-Net predicted segmentations (Section II-C) in the Mask-Guided Attention (MGA) module (Figure 1(d)) to aid the binary image classification task forcing the feature extractor to recognize the informative and discriminative lacunar regions in images (Figure 1(e)). In particular, the first part employs a vgg16 pre-trained on ImageNet with 16 layers made by 3×3 filters with $s = 1$ and LeakyReLU activation function, followed by a max pool layer of a 2×2 filter with $s = 2$, and 3 fully-connected layers. We take only the feature maps from the last max pool layer to output the $f_{img}^{H \times W \times C}$ image features, with C the features map channel number.

The second step involves the MGA module, which starts by performing the f_{img} average and max pooling, to output the average Am_{avg} and maximum Am_{max} spatial attention maps, respectively. Then, their concatenation is convolved and normalized through a sigmoid to output the final normalized Am one-channel attention map. Lastly, the current U-Net predicted image mask drives the learning of the attention map. In particular, we average pooled the mask to the f_{img} size and normalized it to be more noise tolerant (A_{MGA}). To learn the MGA minimizes the Mean Squared Error (MSE) loss $L_{MSE} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left\| A_{MGA}^{i,j} - A_m \right\|$.

As regards the classification step, it takes advantage of both the original and MGA features, weighting them to obtain the final output features $f_{out} = 0.5 * f_{img} + 0.5 * (A_{MGA} * f_{img})$, to be sent to the fully final connected layer for classification. The classification task minimizes the combination of the MSE and Cross-Entropy loss functions $L_{fin} = \lambda \cdot L_{MSE} + L_{CE}$, where $\lambda = 0.1$ using stochastic gradient descent, and $1e-2$ learning rate for 300 epochs. Finally, to show whether the model uses the lacunar structures features or others for classification, we use GradCAMs.

TABLE I: Comparison of lacunae and background segmentation and classification performances. All results are in percentage [%] and the best underlined.

	Image Segmentation		
	Model	Lacunae DSC	Background DSC
(a)	vgg16-1	57.48	99.39
	U-Net-1	58.52	99.12
	vgg16-2	61.92	99.42
	ours	**63.12**	**99.37**

	Image Classification				
	Model	Accuracy	Precision	Recall	F1-score
(b)	ResNet50	50.72	50.66	50.51	50.58
	AlexNet	50.91	56.42	50.62	53.36
	vgg16	51.21	51.72	51.51	51.61
	ours	**62.89**	**65.53**	**61.11**	**63.24**

III. Experimental Results

The PyTorch framework (version 1.13.0) was employed for conducting all experiments. The data pre-processing operations were executed on an AMD Ryzen 7 5800X 8-Core Processor, while training and inference tasks were performed on an NVIDIA RTX A5000 with 24 GB RAM.

A. Semi-Supervised Image Lacunae Segmentation

We start evaluating the proposed U-Net-1 model DSC accuracy in predicting the pseudo-labels. We proceed with the U-Net-2 performance evaluation and conclude by comparing the performances achieved retraining on our image set the state-of-the-art vgg16-based segmentation model without (vvg16-1) [9] and with the semi-supervised strategy (vvg16-2). Table I (a) shows that the U-Net-1 model achieves 58.52% and 99.12% lacunae and background average DSC, respectively. Given the higher U-Net-1 accuracy on the lacunar structures than the vgg16-1 model, we employed the first CNN architecture to learn in a semi-supervised approach. The rationale behind the choice is supported by the 4.60% and 5.64% lacunae performance improvements of the subsequent U-Net-2 training from the U-Net-1 and vgg16-1. Indeed, it achieves a 63.12% average in segmenting the lacunae. At the same time, qualitative evaluation confirms the model's capability to distinguish the lacunae from the background. An example of our method outputs is shown in Figure 2, where it is possible to notice the resemblance between our results and their corresponding labels. Furthermore, our method outperforms vgg16-2-based segmentation model up to 1.20% in DSC improvement for lacunar structures segmentation with comparable background accuracy. Finally, it is worth noting that we computed the DSC by exploiting the manually labeled image patches and not the pseudo ones during all the experiments, thus demonstrating the semi-supervised method's potentiality in accurately segmenting bone lacunae even with a restricted amount of training labeled images.

B. Attention-based Image Classification

We evaluate the robustness of our classification method by comparing accuracy, precision, recall, and F1-score metrics

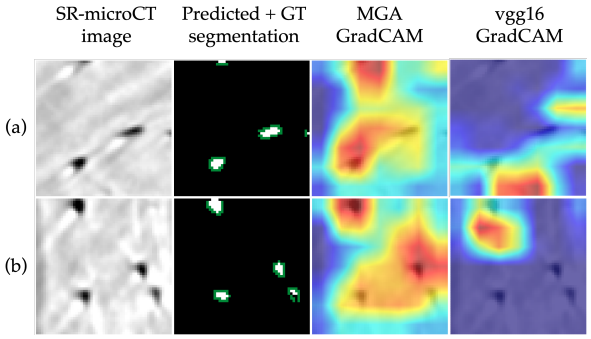

Fig. 2: Segmentation with its ■ Ground Truth (GT) and MGA and vgg16 classification with their GradCAM results for a healthy (a) and osteoporotic (b) SR-microCT image patch.

achieved using majority voting per patient with the ResNet50 employed by *Shen et al.* [10] and the AlexNet and vgg16 networks selected from State Of the Art. We trained all the baselines from the ImageNet pre-trained weights. Table I (b) displays how both shallower and deeper standard networks struggle to distinguish the subtle differences between healthy and osteoporotic images. This aspect is caused by the absence of architecture adaptations able to find the image class-defining feature with a low inter-class image variance. Although the baseline performances slightly overtake random guessing, we exploited the best-performing vgg16 baseline as the feature-extraction backbone, given its highest accuracy among both classes. Thanks to this choice, our attention method allows us to achieve up to 11.68% average accuracy improvement over the not-guided vgg16 version, while up to 12.17% of accuracy improvement over the ResNet50 exploited by *Shen et al.*. Furthermore, the trustiness percentage of the model, when it says that a sample is osteoporotic, is the highest among all the models given the 65.53% average precision obtained. Result confirmed by the 12.66% F1-score improvement given as the harmonic mean between the precision and recall metrics. Finally, GradCAMs support the numerical results achieved. Figure 2 clearly shows that the MGA model interprets the lacunae and their neighbor more precisely than the not-guided vgg16 baseline. This latter result confirms that the MGA strategy really employed the lacunar structure's features as the most important ones to distinguish between healthy and osteoporotic conditions.

IV. Conclusions and Future Work

Two problems hinder the performance of CNN-based methods during SR-microCT image classification: few labeled images due to the laborious manual work while segmenting tiny structures on high-dimensional images and the low inter-class image differences. To handle these issues, we proposed a semi-supervised segmentation approach to delineate lacunae reducing the number of labeled images, and a method to help the classifier focus on the most informative imaged lacunar structures exploiting the mask generated as an attention map. Experimental results demonstrate that this approach achieves up to 5.64% and 12.17% accuracy improvements over state-of-the-art lacunae segmentation and image classification methods, with more interpretable results.

Future work - We will integrate our strategies into a multiple-head standalone network for simultaneous image segmentation and classification to explore if the two tasks can benefit from one another while learning.

Acknowledgment

The authors acknowledge Elettra Sincrotrone Trieste for providing access to its synchrotron radiation facilities.

Compliance with Ethical Standards

Femoral heads are collected with prior authorization from the Ethics Committee (approval date: 13/05/2020, ClinicalTrials.gov ID: NCT04787679) of San Raffaele Hospital (Milan, Italy) and signed approval consent of the patients.

References

[1] A. Oden, E. V. McCloskey, J. A. Kanis, N. C. Harvey, and H. Johansson, "Burden of high fracture probability worldwide: secular increases 2010–2040," *Osteoporosis International*, vol. 26, pp. 2243–2248, 2015.

[2] N. Pradhan, V. S. Dhaka, and H. Chaudhary, "Classification of human bones using deep convolutional neural network," in *IOP conference series: materials science and engineering*, vol. 594, no. 1. IOP Publishing, 2019, p. 012024.

[3] F. Rehman, S. I. Ali Shah, M. N. Riaz, S. O. Gilani *et al.*, "A region-based deep level set formulation for vertebral bone segmentation of osteoporotic fractures," *Journal of digital imaging*, vol. 33, no. 1, pp. 191–203, 2020.

[4] R. Nalla, J. Kruzic, J. Kinney, M. Balooch, J. Ager Iii, and R. Ritchie, "Role of microstructure in the aging-related deterioration of the toughness of human cortical bone," *Materials Science and Engineering: C*, vol. 26, no. 8, pp. 1251–1260, 2006.

[5] H. Portier, C. Jaffré, C. Kewish, C. Chappard, and S. Pallu, "New insights in osteocyte imaging by synchrotron radiation," *Journal of Spectral Imaging*, vol. 9, 2020.

[6] F. Buccino, C. Colombo, and L. M. Vergani, "A review on multiscale bone damage: From the clinical to the research perspective," *Materials*, vol. 14, no. 5, p. 1240, 2021.

[7] M. Mastrogiacomo, G. Campi, R. Cancedda, and A. Cedola, "Synchrotron radiation techniques boost the research in bone tissue engineering," *Acta Biomaterialia*, vol. 89, pp. 33–46, 2019.

[8] I. Poles, E. D'Arnese, F. Buccino, L. Vergani, and M. D. Santambrogio, "Towards an informed cnn for bone sr-microct image classification with an unsupervised patched-based image clustering," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023.

[9] F. Buccino, I. Aiazzi, A. Casto, B. Liu, M. C. Sbarra, G. Ziarelli, G. Banfi, and L. M. Vergani, "The synergy of synchrotron imaging and convolutional neural networks towards the detection of human micro-scale bone architecture and damage," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 137, p. 105576, 2023.

[10] S. C.-y. Shen, M. P. Fernández, G. Tozzi, and M. J. Buehler, "Deep learning approach to assess damage mechanics of bone tissue," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 123, p. 104761, 2021.

[11] X. Yu, Y. Zhao, Y. Gao, X. Yuan, and S. Xiong, "Benchmark platform for ultra-fine-grained visual categorization beyond human performance," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 10 265–10 275.

[12] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1306–1315, 2019.

[13] J. Wang, X. Yu, and Y. Gao, "Mask guided attention for fine-grained patchy image classification," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1044–1048.

[14] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?" *arXiv preprint arXiv:1611.07450*, 2016.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015.