Neural Computing and Applications (2024) 36:14895–14908
https://doi.org/10.1007/s00521-024-09828-2

ORIGINAL ARTICLE

Multi-scale deep learning ensemble for segmentation of endometriotic lesions

Alessandro Sebastian Podda[1] · Riccardo Balia[1] · Silvio Barra[2] · Salvatore Carta[1] · Manuela Neri[3] · Stefano Guerriero[3] · Leonardo Piano[1]

Received: 4 April 2023 / Accepted: 12 April 2024 / Published online: 11 May 2024

Abstract

Ultrasound is a readily available, non-invasive and low-cost screening for the identification of endometriosis lesions, but its diagnostic specificity strongly depends on the experience of the operator. For this reason, computer-aided diagnosis tools based on Artificial Intelligence techniques can provide significant help to the clinical staff, both in terms of workload reduction and in increasing the overall accuracy of this type of examination and its outcome. However, although these techniques are spreading rapidly in a variety of domains, their application to endometriosis is still very limited. To fill this gap, we propose and evaluate a novel multi-scale ensemble approach for the automatic segmentation of endometriosis lesions from transvaginal ultrasounds. The peculiarity of the method lies in its high discrimination capability, obtained by combining, in a fusion fashion, multiple Convolutional Neural Networks trained on data at different granularity. The experimental validation carried out shows that: (i) the proposed method allows to significantly improve the performance of the individual neural networks, even in the presence of a limited training set; (ii) with a Dice coefficient of 82%, it represents a valid solution to increase the diagnostic efficacy of the ultrasound examination against such a pathology.

Keywords Endometriosis · Ultrasound images · Convolutional neural networks · Deep learning · Segmentation

✉ Alessandro Sebastian Podda
sebastianpodda@unica.it

Riccardo Balia
riccardo.balia@unica.it

Silvio Barra
silvio.barra@unina.it

Salvatore Carta
salvatore@unica.it

Manuela Neri
manuela.neri@unica.it

Stefano Guerriero
sguerriero@unica.it

Leonardo Piano
leonardo.piano@unica.it

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

[2] Department of Information Technology and Electrical Engineering, University of Naples Federico II, Via Claudio, 21, 80125 Naples, Italy

[3] Department of Obstetrics and Gynecology, University of Cagliari, Policlinico Universitario Duilio Casula, 09042 Monserrato, Italy

1 Introduction

Endometriosis is defined as a chronic benign inflammatory disease of the female genitalia and pelvic peritoneum, characterized by the abnormal presence of endometrial-like tissue outside the uterus. It affects approximately 10% of reproductive-age women worldwide [1]. Currently, there are no ways to prevent the onset of the disease, nor is there a definitive cure for endometriosis; existing treatments aim to control the symptoms.

The gold standard for the diagnosis of endometriosis is *laparoscopy*, a micro-invasive surgical method capable of accurately identifying the signs of the disease by direct visualization of the tissue, which allows obtaining histological diagnosis. However, the early diagnostic suspicion is based on a careful analysis of patients' symptoms, often associated with non-invasive examination techniques, including *Transvaginal Ultrasound* (TVUS) and *Magnetic Resonance Imaging* (MRI) [2]. While non-invasive techniques lack in accuracy compared to laparoscopy, in cases where the patient has developed cysts (i.e.,

endometriomas), adhesions or deep nodule forms, TVUS and MRI are effective in their detection [3].

In particular, whereas MRI may have some limitations related to costs and accessibility, medical ultrasonography is a risk-free and low-cost ultrasound-based procedure, whose main drawback is due to the high dependence between its diagnostic effectiveness and the expertise of the involved clinical staff. It is used: (i) as a supporting examination in suspected cases; (ii) to monitor the progress of the disease and the response to medical treatments; and, (iii) for adequate preparation for laparoscopy. Unfortunately, the use of such procedure by the hand of not sufficiently experienced professionals can result in 3–11 years of delay in the diagnosis of this condition [4].

To overcome these limitations, in this study, we propose a novel solution based on Artificial Intelligence techniques, to implement a *decision support system* (DSS) for the analysis of transvaginal ultrasound and to improve the diagnostic efficacy of this method.

Specifically, the proposed approach exploits the possibilities offered by Deep Learning and Computer Vision to define a *multi-scale ensemble* of *Convolutional Neural Networks* (CNNs), able to recognize and segment endometriosis lesions detected in ultrasound images. Its underlying idea lies in the findings of [5], according to which the optimal selection of the image resolution has the potential to enhance the performance of neural models in several radiology-based machine learning tasks. However, the choice of the best value is not trivial: higher resolutions often allow to improve the segmentation accuracy, particularly for lesion edges; conversely, an excessive level of detail may not necessarily help the networks' ability to discern between lesions and healthy tissue [5, 6]. Hence, through the proposed approach, we aimed to capture and combine the different peculiarities that can be extracted from multiple resolutions of the treated images, leveraging evidence that aggregating contextual information may improve segmentation accuracy [7]. The results obtained through the experimental validation confirm the validity of this intuition.

In light of the above, the main contributions of this work can be summarized as follows:

1. we propose a novel and pioneering Convolutional Neural Network ensemble-based approach for the automatic identification and segmentation of endometriotic lesions in transvaginal ultrasounds; notably, to the best of our knowledge, this is the first scientific work to tackle this task through Deep Learning and Computer Vision techniques;
2. we experimentally demonstrate that training and applying the models—in a fusion fashion—on multiple resolutions of the input images, through the proposed multi-scale approach, allows for significant improvement in the accuracy of the obtained segmentation; in particular, this behavior is observed for all the types of neural networks considered;
3. we test the proposed method on a dataset annotated by experienced medical staff, showing how not only the results achieved in the automatic segmentation task are satisfactory even in the presence of a limited training set, but that both in quantitative and qualitative terms they confirm the goodness of this approach for the prospective development of a Computer-Aided Diagnosis (CAD) system specifically dedicated to the endometriosis pathology.

The remainder of this manuscript is structured as follows. In Sect. 2, we explore the related work, with a specific focus on the existing solutions for the analysis of ultrasound images in the gynecological field. Then, in Sect. 3, we illustrate in detail the proposed multi-scale ensemble-based approach, while in Sect. 4, we describe the experimental setup adopted, with particular attention to the employed dataset and the augmentation and optimization techniques adopted. Results are shown and thoroughly discussed in Sect. 5. Finally, Sect. 6 concludes the work and outlines the most promising future research directions.

2 Related work

Ultrasound (US) serves as a non-invasive imaging technique for the examinations of the human body and internal structures. On the other hand, medical image segmentation aims to facilitate the differentiation and localization of anatomical changes in medical images, including ultrasound ones. Considering its impact on computer-aided systems, medical image segmentation stands out as a deeply explored problem in literature, where many methodologies tailored to diverse pathologies have been proposed. Among the most exploited techniques in the past decade have been thresholding [8], clustering [9], watershed [10], active contour models [11], and neural networks [12]. Neural networks represent a truly breakthrough in the field, and almost all of the recent literature is devoted to exploring and improving such technologies. Zhao et al. [13] take advantage of a *U-net*-like architecture for Nerve segmentation in ultrasound images, reaching a mean dice score of 65%. The authors in [14] explored segmentation of brachial plexus nerves from ultrasound images using different *Deep Convolutional Neural Networks* (Deep CNNs) combined with a preprocessing strategy to reduce speckle noise. Their best configuration consisted of a *M-Net* with a *Prewitt edge filter* that reached a Dice score of 88%. Xue et al. [15] developed a global guidance network (i.e., *GG-*

Net) for breast lesion segmentation. Podda, et al. [16] devised an end-to-end, fully automatic pipeline for the classification and segmentation of breast lesions introducing a cyclic mutual optimization strategy, which iteratively and reciprocally exploits the contribution of the classification step to improve the segmentation step, and vice versa. Similarly, Lei et al. [17] proposed a deep learning-based method for male pelvic multi-organ segmentation on transrectal ultrasound images. They developed an anchor-free Mask CNN-based architecture to segment prostate, bladder, rectum, and urethra simultaneously. Their method obtained a dice coefficient of 75% for the bladder, 93% for the prostate, 90% for the rectum, and 86% for the urethra. Furthermore, [18] introduces a novel *Multi Expert fusion (MXF)* framework to segment 3D transrectal ultrasound images of the prostate where three different CNNs are trained in parallel on a specific slice viewing and final segmentation volume is obtained through a specialized fusion network.

Despite the fact, however, that endometriosis is a fairly common gynecological condition, to the best of our knowledge, this work is the first in the literature to specifically address the problem of automatically segmenting endometriosis lesions from transvaginal ultrasounds. For the above reason, the remainder of this section is focused on relevant approaches applied to correlated challenges.

In this context, a seminal work is represented by the one proposed by Singhal et al. [19]; here, the authors introduce a fully automated method to assess the endometrium thickness from 3D transvaginal ultrasound. Their method combines Deep Learning techniques with level set segmentation, embedding the output feature map of a Convolutional Neural Network in the segmentation energy function of a hybrid *variational curve propagation* model. Similarly, an automatic approach for the endometrium thickness measurement from 2D ultrasound has been proposed by Hu et al. [20]. Their pipeline involves an initial step of segmentation of the endometrium, employing a *VGG-based* U-Net, and a second step of endometrial thickness estimation through a medial axis transformation.

Within the same scope, Park et al. [21] developed a novel framework that provides robust endometrium segmentation against ambiguous boundaries and heterogeneous textures of TVUS images. The authors identified four key points, i.e., meaningful zones that are related to the characteristics of the endometrial morphology, to guide a discriminator network in distinguishing a predicted segmentation map from a ground-truth segmentation map. Such a key-point discriminator improved the baseline performance. On the other hand, Thampi et al. [22] focused on the automatic segmentation of endometrial cancer from ultrasounds images, through level set and *Otsu*'s thresholding methods. From a broader perspective, Usha et al. [23] investigate the automatic measurement of the ovarian size and its shape parameters assessment to help experts make a quick diagnosis. In contrast, Jin et al. [24] analyze the accuracy of segmentation algorithms based on multiple U-net models, applied to ultrasound images, in patients with ovarian cancer.

3 Materials and methods

Building on the results obtained from existing work, this study addresses the problem of automatically and robustly identifying and segmenting endometriosis lesions from transvaginal ultrasound images. Such an approach mainly exploits Deep Ensemble Learning and Computer Vision techniques, and explores the impact of the input TVUS image resolutions on the segmentation performance, to combine a set of individual models to improve the overall method effectiveness and accuracy.

3.1 CNN-based segmentation

The backbone of the proposed method is built upon *U-net* [25], a novel architecture of *Convolutional Neural Network* (CNN), originally proposed in 2015 and specifically designed to tackle image segmentation tasks in the biomedical field. Its success depends on the ability of such an architecture to return an output with: (i) a size similar to the input; and, (ii) a high spatial resolution, making it ideal for cases in which we aim to generate a mask that faithfully reproduces the shape of the target to be identified. To the best of our knowledge, this work is the first to propose the use of U-net for the automatic segmentation of endometriotic lesions.

The U-net architecture, depicted in Fig. 1, is composed of two parts. The first, namely the *contraction phase*, shares the typical encoder-based structure of many CNN classifiers: input is processed by the convolutional layers to extract the image features, while dimensions are down-sampled by the pooling layers. To recover the spatial information lost during the contraction phase, a second phase makes use of *skip* connections to concatenate feature maps with the same dimensions (shown as gray arrows in Fig. 1), acting as a decoder. These two phases give the layout of the architecture a typical *U-shape*, from which this network takes its name.

3.2 Multi-scale ensemble of U-nets

Analyzing the performance of U-net on a preliminary subset of TVUS images, we noticed a significant degree of variability in the segmentation performance obtained. This

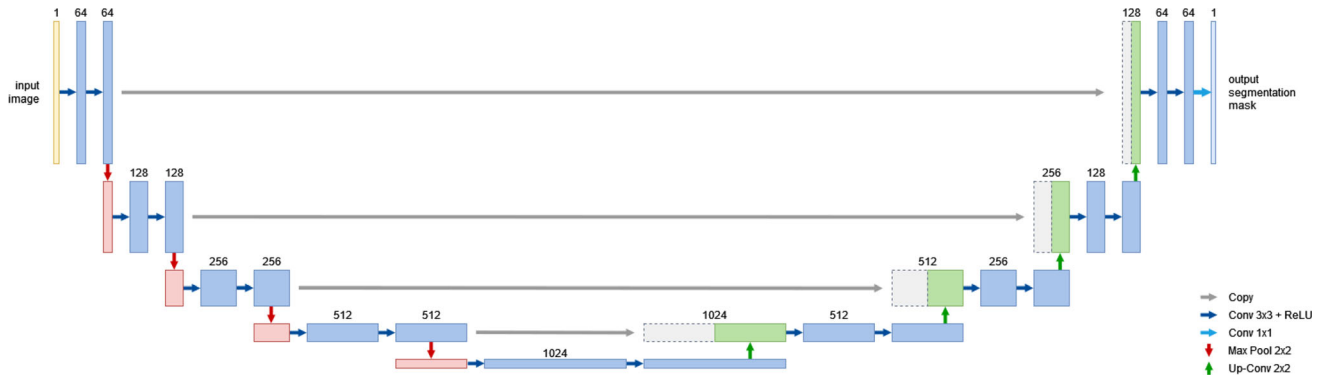

Fig. 1 Schematic diagram of the U-net architecture

behavior might depend on the considerable heterogeneity of endometriotic lesions, whose sizes, shapes, edge roughness, and surrounding tissues tend to differ markedly from one patient to another, similar to what has been observed in related medical tasks [26, 27]. However, since the goal of the proposed tool is to support the clinical staff in the identification and diagnosis of endometriosis, an essential requirement is to ensure the robustness and stability of the obtained predictions, thus minimizing the variance of the method.

To overcome this limitation, in this study, we propose a *multi-scale ensemble* of convolutional neural networks, where a single U-net architecture is trained multiple times with different input resolutions, in order to generate separate models capable of better capturing the characteristics of ultrasound images at different granularity. The proposed strategy is partially inspired by the *bagging predictors* [28], in which a set of identical models are trained in parallel on different random portions of the dataset and then aggregated through some voting process. In particular, bagging has been proven to be effective on *unstable* learning algorithms [28], i.e., those in which small variations in the training set may result in large variations in predictions. Neural networks are an example of unstable learning algorithms [29].

The basic idea is to obtain a set of complementary models: those trained on lower resolutions, thanks to a lower presence of details in the image, may be more suitable to identify the number and location of lesions; on the other hand, those trained on higher resolutions, having a deeper level of detail, are able to segment more accurately the existing lesions, although at the expense of potential exposure to a greater number of false positives. Through the use of appropriate fusion and ensemble techniques, the objective is then to balance the specificities of the different models obtained.

3.3 Fusion strategy

To achieve the aforementioned purposes, we trained different U-net models with 64×64, 128×128, 224×224 and 256×256 resolution images, respectively; then, we compared the segmentation performance of the models obtained for each of these resolutions.

In order to make the method more robust, reducing the variance of the results obtained, the proposed fusion strategy leverages on an ensemble strategy in which, basically, the single-scale model that shows the highest accuracy in the validation set is denoted as a *strong learner* (S_L), whereas the remaining configurations are marked as *weak learners* (W_L). The motivation behind this choice lies in the fact that, from an aprioristic point of view, the strong learner shows superior performances in comparison with the other (weaker) models; consequently, unifying all the predictions through a *flat* voting policy may degrade such a result. On the contrary, enhancing the outputs of the weak models in a fusion fashion allow to correct the possible imperfections of the strong learner prediction and thus improve the general accuracy of the method. Note that this hypothesis has been experimentally verified through some preliminary tests. These experiments showed that the proposed fusion method performed systematically better, for all the backbone architectures considered, than a *peer learners* approach (i.e., an approach where all the single-scale models have equal voting dignity), in both a *soft-voting* and a *hard voting* context—note that such techniques have, besides, led to very similar results. We also tested the possibility of selecting the strong learner according to different criteria than electing the model with the best performance in validation (e.g., the single-scale model with the higher input resolution, or a randomly determined model, or a pair of two S_Ls instead of one only). In all these attempts, the proposed fusion method proved to be significantly more accurate (with an improvement $> 2\%$ in every considered scenario), and therefore, we decided to continue on this path.

Figure 2 graphically illustrates the proposed pipeline. The adopted procedure is as follows: the weak learners' candidate segmentations are combined into a single better segmentation mask through *hard voting* (i.e., a majority voting that assigns equal weights to each candidate and designates each pixel with the label that the most segmentations agree on). Since the segmentation masks returned by U-net are *probability maps*, while the final mask we want to obtain is binary, a threshold t is applied (we set $t = 0.5$). Hence, pixels with probability values above this threshold are considered in *foreground* (i.e., part of the endometriosis lesion), while the others are labeled as *background*.

As we employ three W_L models, each pixel receives three votes. Hence, we consider such pixel as belonging to a lesion if it receives at least two votes (i.e., the majority). This pixel-wise ensemble decision can be formalized as follows:

$$\hat{y} = \max_{j=1}^{C} \sum_{t=1}^{T} v_{t,j} \tag{1}$$

where C represents the number of classes (*background* and *foreground*, in our case), T is the number of ensemble models, and the summation $\sum_{t=1}^{T} v_{t,j}$ indicates the sum of the votes assigned from the models $\{W_{L0}, ..W_{Lt}\}$ to each class j.

The ensemble of weak learners, that for sake of clarity we hereafter denote as W_{Ens}, is then combined—in our pipeline—with the prediction of the strong learner by following a more sophisticated strategy.

We thus generate two new images, indicated with I and U. The first one is obtained as the *intersection* between the segmentation masks $M_a = mask_{W_{Ens}}$ and $M_b = mask_{W_{Ens}}$ produced, respectively, by the weak learners' ensemble and the strong learner (i.e., $I = M_a \cap M_b$), that considers only the region of interest common to both masks. The above solution is employed to reduce the probability of considering regions that do not really belong to a lesion.

In an analogous way, the second image U is determined as the *union* between the mask generated by the weak learners' ensemble and the mask produced by the strong learner, respectively (i.e., $U = M_a \cup M_b$). It contains all the regions identified by the two models, including those not in common. The newly generated images I and U are then finally merged through *morphological reconstruction*. This process can be conceptually summarized as an iterative combination of the mathematical procedures of *dilation* and *erosion* to refine an image, called *marker*, until its contours fit under a second image, called *mask* [30]. In our pipeline, the marker is the image resulting from the intersection (I), while the mask image is the result of the union (U). With this procedure, our method aims to better identify the area of interest that most likely belongs to the lesion and, subsequently, to reconstruct the lesion shape of the lesion which can be lost after the image intersection.

To sum up, the proposed ensemble consists of two phases: (i) a first step where masks generated by a set of weak learners are fused through a hard-voting policy; and (ii) a second stage to combine, by means of the morphological reconstruction, the image obtained from the previous step and the mask generated by the strong learner, to capture the peculiarities of each.

4 Experimental setup

The proposed solution was developed in *Python 3.8* language, equipped with the `OpenCV 4.5.1`, `scikit-image 0.18.1`, `scikit-learn 0.24.2`, `Numpy 1.19.5`, `Keras 2.6.0` and `Tensorflow 2.6.0` libraries.

We ran the experiments on a desktop computer with 4.10 GHz CPU, 32GB RAM, and a NVIDIA GeForce GTX 1060 Max-Q graphic card with 6GB dedicated DDR5 RAM and 1280 CUDA Cores.

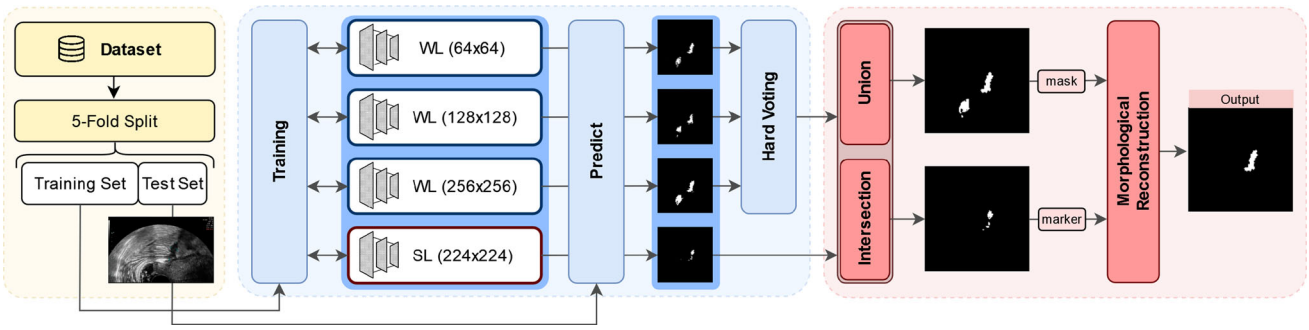

Fig. 2 Schema of the proposed multi-scale strategy, based on four parallel CNN models. The best performing network (at the validation stage) is denoted as *strong learner* (SL), while the rest serve as *weak learners* (WLs) whose predicted masks are fused by a hard voting. The result is then combined with the SL's output through morphological reconstruction to produce a robust final prediction

4.1 Dataset collection and annotation

The dataset used has been collected and anonymized by the personnel of the "*Duilio Casula*" Hospital of Cagliari, Italy, and kindly granted for the purposes of this study. It includes 75 transvaginal ultrasound images acquired from patients with endometriosis, including, in particular, bowel deep endometriotic lesions. Notably, all the samples have been manually annotated by expert medical operators, i.e., each image is accompanied by a pixel-wise annotation which represents its ground-truth segmentation mask.

For the annotation phase, we provided *Hasty*,[1] a free in-browser tool, to the medical operators. It allows to easily annotate data to train models for artificial intelligence tasks. In particular, it shows to be effective for the visual marking of elements of interest in image data, as for the endometriosis lesions of our case. Figure 3 illustrates the main screen of a Hasty project. The left panel features both automatic and manual tools; the first ones exploit AI-driven algorithms to facilitate the marking process. For the purpose of this study, the annotation has been done with manual tools only, such as the brush and the polygon area selector.

Although limited in size, the endometriosis dataset collected and annotated using the aforementioned procedure is, to our knowledge, the first in the literature based on ultrasound images. However, a similar dataset collected by Leibetseder et al. [31] is available, which provides segmentation and bounding box masks, but whose images are extracted from video sequences during laparoscopic procedures.

4.2 Data augmentation

To overcome the size limitations of the dataset, we adopted a preliminary step of data augmentation. Instead of the canonical augmentation methods offered by the `Keras` library, we opted for those provided by the `Albumentation` one [32], which implements a larger variety of transformations. More precisely, we applied some augmentation techniques to the sample in the training set only, to increase the variability of the data and improve the generalization capabilities of the model. For a fair presentation of the results, these augmentation techniques were applied identically to all models used, including the state-of-the-art competitor methods considered in the remainder of this work. Conversely, no transformation aimed at improving the quality of the ultrasound images is applied at inference time.

Hence, for the training set only, we generated no. 5 new transformed images for every original one: each new sample is the result of several transformations applied in combination, according to a given probability. Specifically, our augmentation pipeline consists of the following transformations: (i) horizontal flip, (ii) ISO and multiplicative noise, (iii) random zoom, (iv) transposition, (v) grid distortion, and (vi) contrast limited adaptive histogram equalization (CLAHE).

4.3 Metrics

To evaluate the results obtained by the proposed method, we employ two metrics commonly used in the context of segmentation tasks in several application domains. The first is the *Dice* coefficient (Dice), defined as:

$$Dice = \frac{2 \cdot \|A \cap B\|}{\|A\| + \|B\|} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

while the second is the *Jaccard similarity* coefficient (Jac), sometimes referred also as Intersection over Union (IoU) and expressed as:

$$Jac = \frac{Intersection\ Area}{Union\ Area} = \frac{\|A \cap B\|}{\|A \cup B\|} = \frac{TP}{TP + FP + FN} \tag{3}$$

Moreover, to determine how many lesions are actually correctly identified, we adopted a detection system based on the *overlap criterion*, according to which a lesion is correctly identified only if the obtained mask is not empty and its Jaccard similarity coefficient is greater than 0.5. Thus, based on the generated labels, we calculate the *detection accuracy* (dACC), with the prefix d denoting the dependence on the detection system chosen. Its equation is then defined as it follows:

$$dACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

4.4 Training hyperparameters

The neural networks involved in this work were all trained by adopting the following parameters, selected within the validation set: the batch size is set to 16, the optimizer is *Adam* [33] and the maximum number of training epochs is fixed to 100. However, an *early stopping* strategy with 30 epochs patience is employed (i.e., the training is interrupted if the validation Dice coefficient has not improved in the last 30 epochs). We also tested several loss functions commonly adopted in segmentation tasks, like the region-based *Dice Loss* [34] and the *Tversky loss* [35], although the best results were obtained by using the distribution-based *Binary Cross-Entropy loss*, defined as:

[1] https://app.hasty.ai/.

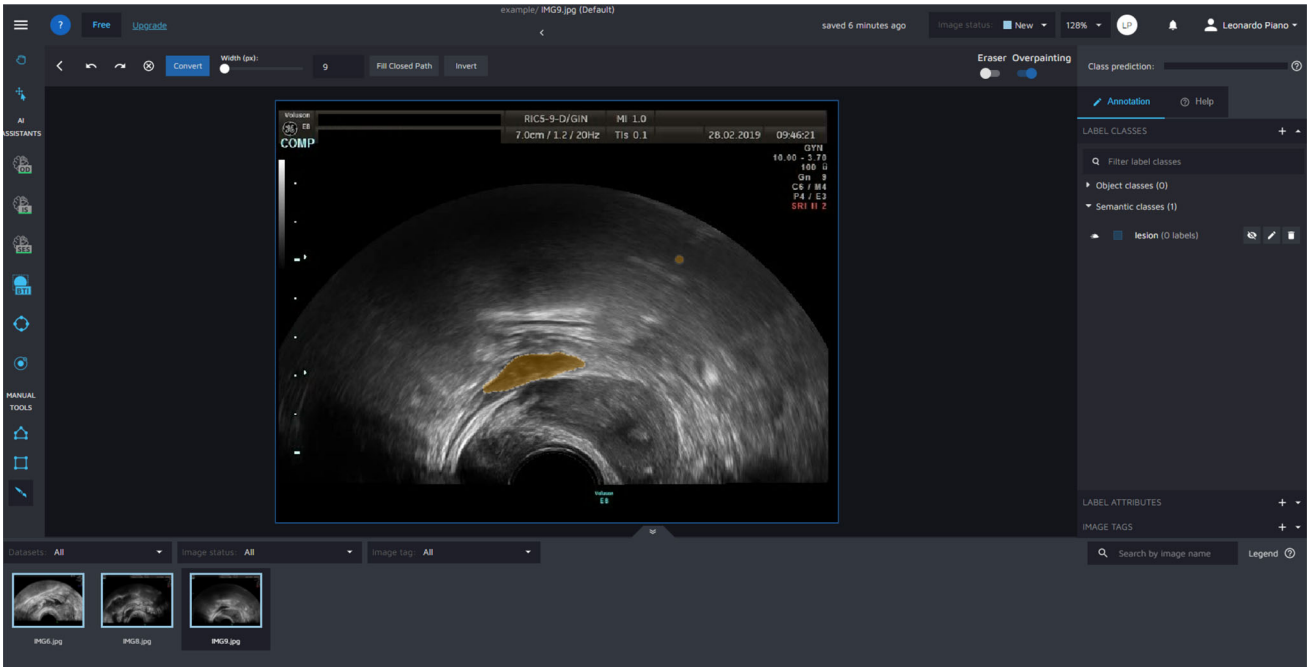

Fig. 3 Lesion marked with the *Hasty* annotation tool

$$\mathrm{BCE} = -\frac{1}{c}\sum_{i=1}^{c} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

where y_i denotes the pixel label (i.e., class 1 for foreground/lesion and class 0 for background/normal tissue), $\hat{y}_i$ indicates the predicted probability of the pixel belonging to class 1, and $(1\text{-}\hat{y}_i)$ is the probability of the same pixel belonging to class 0. Finally, in order to preserve the heterogeneity of information arising from the multi-scale approach, we kept the same parameters for the models trained on all the input resolutions.

5 Results and discussion

We evaluated our method over three different U-net architectures. Two of them employ a state-of-the-art network backbone for the contraction phase, i.e., *DenseNet121* and *VGG19* [36, 37]. The third one consists instead of an architecture built from scratch as part of this study. Table 1 reports a brief summary of the employed networks and training hyperparameters. As described in detail later in this section, where we analyze the performance of the proposed method, all the aforementioned architectures showed significant improvement employing the proposed ensemble approach.

5.1 Quantitative comparison

To evaluate our method, a *fivefold* cross-validation was adopted to provide statistical significance. Specifically, 80% of the samples were used for the training and validation steps (split into 70 and 10%, respectively), with the remaining 20% used as the test set. The final results are obtained by averaging all folds.

In Table 2, we report the results returned by our pipeline, implemented with the DenseNet, VGG, and custom backbone network, respectively, by first considering the four single scales and, subsequently, the proposed ensemble performance (which exploits all the aforementioned resolutions in combination). First, it can be observed that each of the evaluated architectures has different behavior depending on the input resolution used. In particular, the DenseNet121-based U-net performs significantly better with the 224×224 scale, achieving a 67.8% in terms of Jaccard similarity and a 79.3% of Dice coefficient. On the other hand, the VGG19-based network reaches its best result with the 128×128 scale (71.6% of Dice), while our custom implementation tends to show a fairly stable performance across the different resolutions. Overall, however, the DenseNet121 architecture performs on average better than the others when considering all the scales.

The same Table 2, last row, also shows the results obtained by applying the proposed multi-scale ensemble approach to each of the three examined architectures. We can state that although the training set contained an

Table 1 Summary of the architectures and parameters adopted

Backbone	Year	Encoding layers	# of Trainable params (M)	Training hyperparameters*
DenseNet121	2017	117x Conv, 3x Transition	16	`batch_size` = 16; `loss` = *binary cross-entropy*; `optimizer` = *Adam*; `max_epochs` = 100
VGG19	2014	16x Conv, 3x Dense, 5x MaxPool	33	
Custom	–	7x Conv, 3x MaxPool	4	

∗The reported values are the same for all the architectures

Table 2 Comparison between single scales and ensemble

InputSize	DenseNet121 U-net			VGG19 U-net			Custom U-net		
	Jac	Dice	dAcc	Jac	Dice	dAcc	Jac	Dice	dAcc
64×64	0.592	0.716	0.773	0.568	0.699	0.706	0.582	0.709	0.679
128×128	0.617	0.731	0.800	0.594	0.716	0.720	0.580	0.703	0.706
224×224	0.678	0.793	0.867	0.503	0.622	0.600	0.585	0.709	0.693
256×256	0.647	0.759	0.853	0.502	0.627	0.559	0.588	0.705	0.706
Average	0.633	0.750	0.823	0.542	0.666	0.646	0.584	0.706	0.696
Ensemble	**0.712**	**0.818**	**0.906**	**0.638**	**0.758**	**0.800**	**0.650**	**0.767**	**0.773**

Bold values represent the best results for each corresponding metric

exiguous number of images our method achieved excellent generalization capabilities. Ultrasound images are more difficult to interpret respect to other radiological images, they are disturbed by the typical speckle noise and the shadow produced by the probe could be mistaken for a lesion, for which normally many more images would be needed to avoid such mistakes. It is possible to observe that all of them achieved a remarkable increase in performance, not only with respect to the average value observed with the corresponding single-scale approach, but also in relation to their best resolution configuration. DenseNet121 has improved by $\approx 3\%$ its best Jaccard and Dice results obtained with the 224×224 scale (also showing a $+7/8\%$ compared to its average behavior). It also outperformed a detection accuracy of 90%, which implies that more than 9 out of 10 predictions generated with such an ensemble have a Jaccard similarity coefficient > 0.5. Similarly, VGG19 and the custom architecture show a significant boost, both with improvements of over 6% in terms of the Jaccard and Dice metrics.

5.2 Comparison against state-of-the-art models

To evaluate and compare the performance of our method with other state-of-the-art methods that exploit variable object scales for semantic segmentation, we proceeded to train DeepLabV3+ [38] and MSUnet [39]. DeepLabV3+ improves the popular segmentation network DeepLabV3 introducing an encoder-decoder structure where the encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, while the decoder module refines the segmentation results along object boundaries. MSUnet replaces the original U-net convolution blocks with *multi-scale blocks* that are composed of multiple convolution sequences with different receptive fields. This enables the network to extract more semantic features from the images and generate more detailed feature maps. The experiments were conducted by applying the same experimental setting used with the networks described in our manuscript; the dataset images were instead resized to 256px per side. The results obtained reported in Table 3 show that these networks are still not superior to our ensemble performance, and while MSUnet manages to have generalization capabilities similar to the Densenet121-U-net used in this work, DeepLabV3+

Table 3 Performance comparison of our best configuration against literature models that exploit multi-scale features based on single-network architectures

Method	Jac	Dice	dAcc
MSUnet[a] [39]	0.685	0.796	0.866
DeepLabv3+[b] [38]	0.557	0.683	0.680
Proposed multi-scale ensemble	**0.712**	**0.818**	**0.906**

Bold values represent the best results for each corresponding metric

[a]https://github.com/CN-zdy/MSU_Net

[b]https://keras.io/examples/vision/deeplabv3_plus/

behaves more like VGG19-U-net, sometimes failing to converge to an optimal solution. Our multi-scale ensemble yields competitive results compared with the state-of-the-art.

5.3 Qualitative evaluation

In order to better explain the contribution of our method, Fig. 4 shows a graphical comparison of the performance of a single-scale configuration and its corresponding multi-scale ensemble pipeline, both based on the VGG19 backbone. The comparison highlights some relevant instances where the prediction of the single-scale model exhibit imperfections and inconsistencies that the ensemble was able to compensate effectively, e.g., by clearing background regions improperly classified as lesions (Fig. 4a) or by better defining the shape and edges of an existing lesion that is not well segmented by the model that employs the single resolution (Fig. 4b–d).

In this regard, Fig. 5 better outlines how the choice of the input resolution may impact the behavior of the neural networks, by means of a visual representation of the neuronal activation generated through the popular *Gradient-weighted Class Activation Mapping* (GradCam) [40] technique. Basically, GradCam employs the gradients of any target concept (a *lesion* area, in our case) flowing into the final convolutional layer to produce a coarse localization map that highlights the important regions in the image used for predicting the concept.

For the considered samples, we notice that the resolutions 128×128 and 224×224 present areas characterized by poor or wrong neuronal activations, resulting in incorrect or incomplete predictions. Vice versa, a better localization of the lesion is observed by adopting the lowest resolution (64×64), where the heatmap correctly emphasizes the area corresponding to the endometriosis lesion in all four ultrasound images.

The aforementioned evidences allows us to consolidate some significant assumptions underlying the proposed method: (i) choosing the best resolution represents a non-trivial task and depends on both the profile of the input and the considered model; (ii) higher resolutions are not

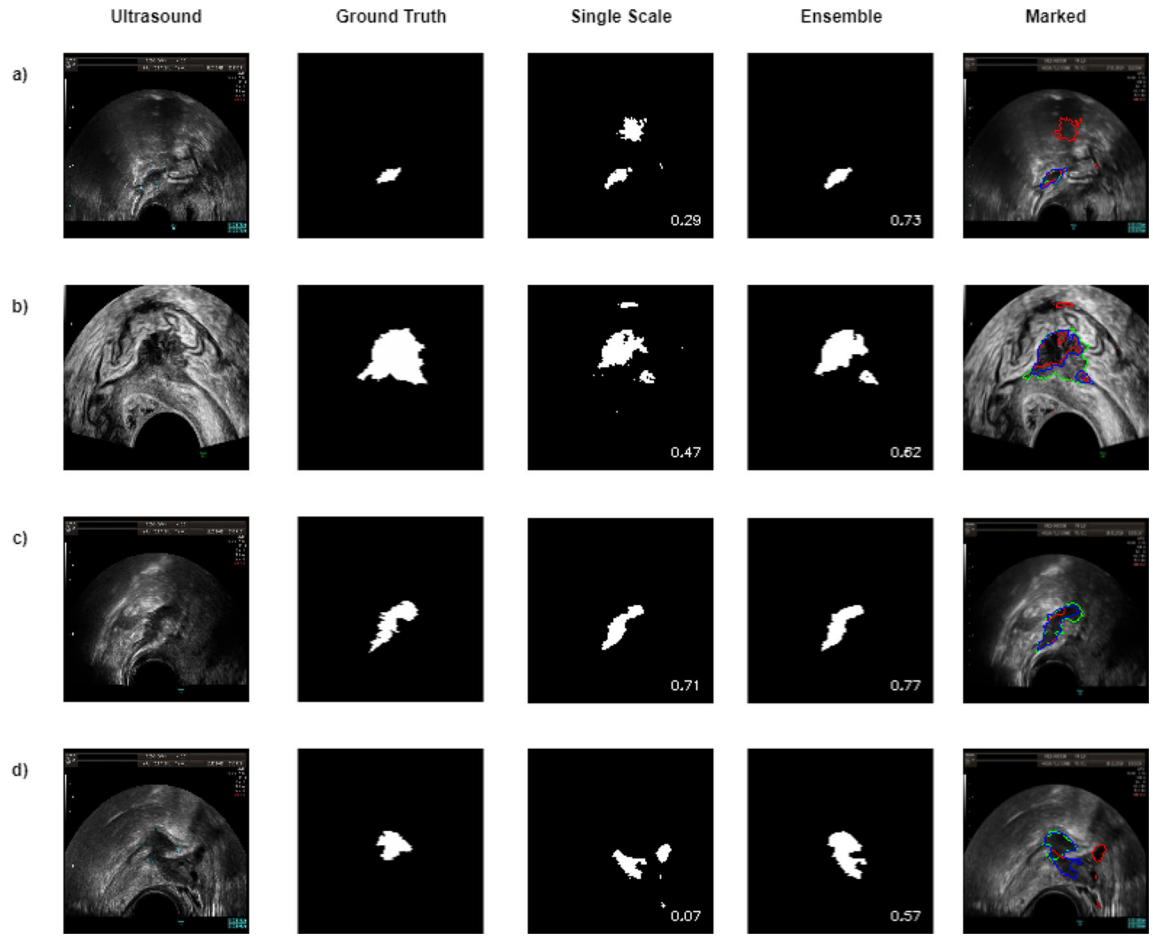

Fig. 4 Graphic comparison between the performance of the VGG19-based U-net in single-scale mode and with the proposed multi-scale ensemble (Jaccard's similarity reported in the lower right corner)

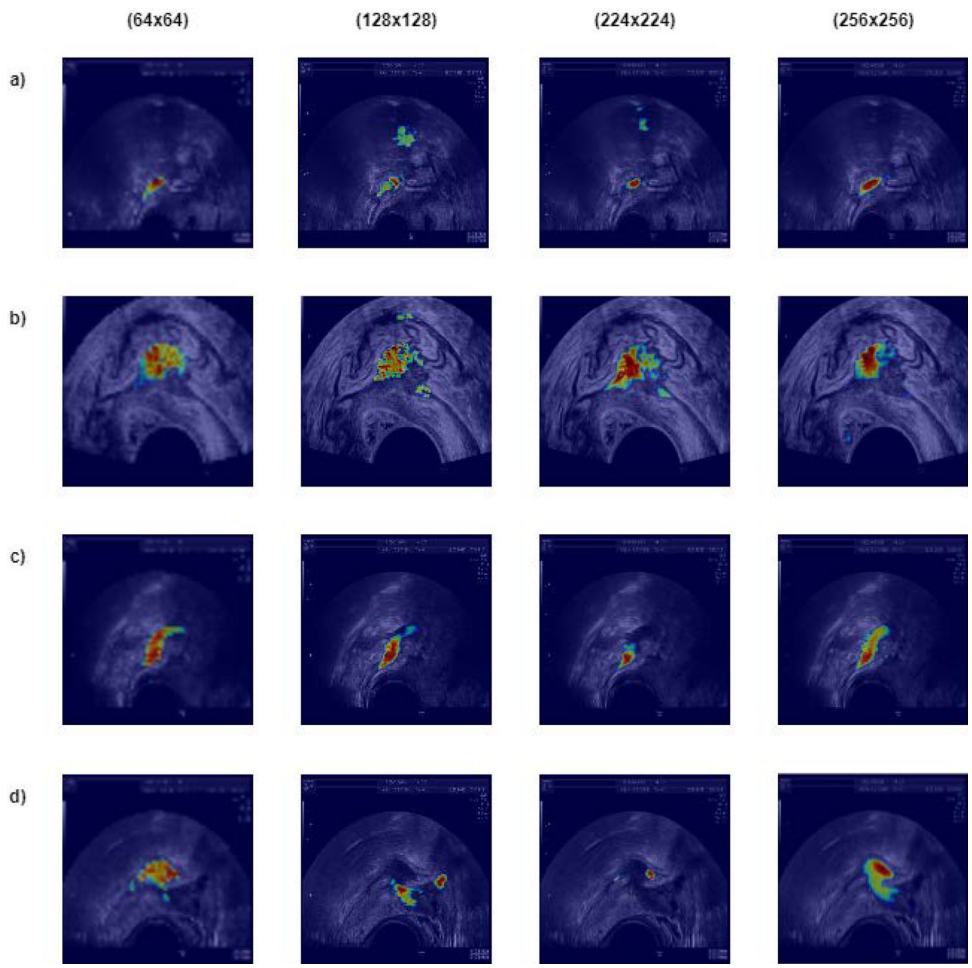

Fig. 5 Gradcam heatmaps of lesions. It highlights the regions identified by the single-scale models at different resolutions, to show how the choice of the *ideal* input size is not trivial and might significantly affect the final output

necessarily more effective in achieving the final result, as the higher level of detail may make the differentiation between lesion and healthy tissue less precise; (iii) the proposed multi-scale ensemble is quantitatively and qualitatively effective in capturing and combining the specificities of the different resolutions, allowing for more stable, accurate and robust results even in the presence of limited and highly heterogeneous training sets.

5.4 Limitations

While exhibiting encouraging results, such an experimental validation presents some limitations. Specifically, the dataset utilized is not only limited in size but also exclusively comprises ultrasound images featuring the presence of lesions (i.e., true positives). This is because such datasets come from a clinical study where selected patients had already been diagnosed with endometriosis or showed a strong suspicion of symptomatology of such a pathology. Thus, our study aimed to investigate a useful solution for the detection and segmentation of lesions in already established or strongly suspected cases of endometriosis. However, we investigated whether our method would perform with non-pathological lesion-free images. Since we could not obtain images of patients without endometriosis, we conducted an exploratory analysis of the ability of our models to avert false positives in healthy tissue images. For this purpose, we therefore attempted to emulate images of healthy tissue by synthetically generating them using photo editing software and AI tools. We clarify that to do so we have reserved a few images for each fold to be used as a test set (for a total number of 10 items), modifying them appropriately before running the prediction engine. This ensured that the model had never observed in previous training samples the portions of healthy tissue used to generate the synthetic test images. Figure 6 shows a visual outcome of such an experiment. In the figure, the first column shows the original image, the second column presents the synthetically generated image (after removing the endometriosis lesion), while the third column contains the prediction generated by our model. From such an exploratory experiment, we observe that our proposed approach, despite being trained only on images having lesions, seems able to generalize to images where the lesion is not present, predicting an empty mask 7 times out of 10. In

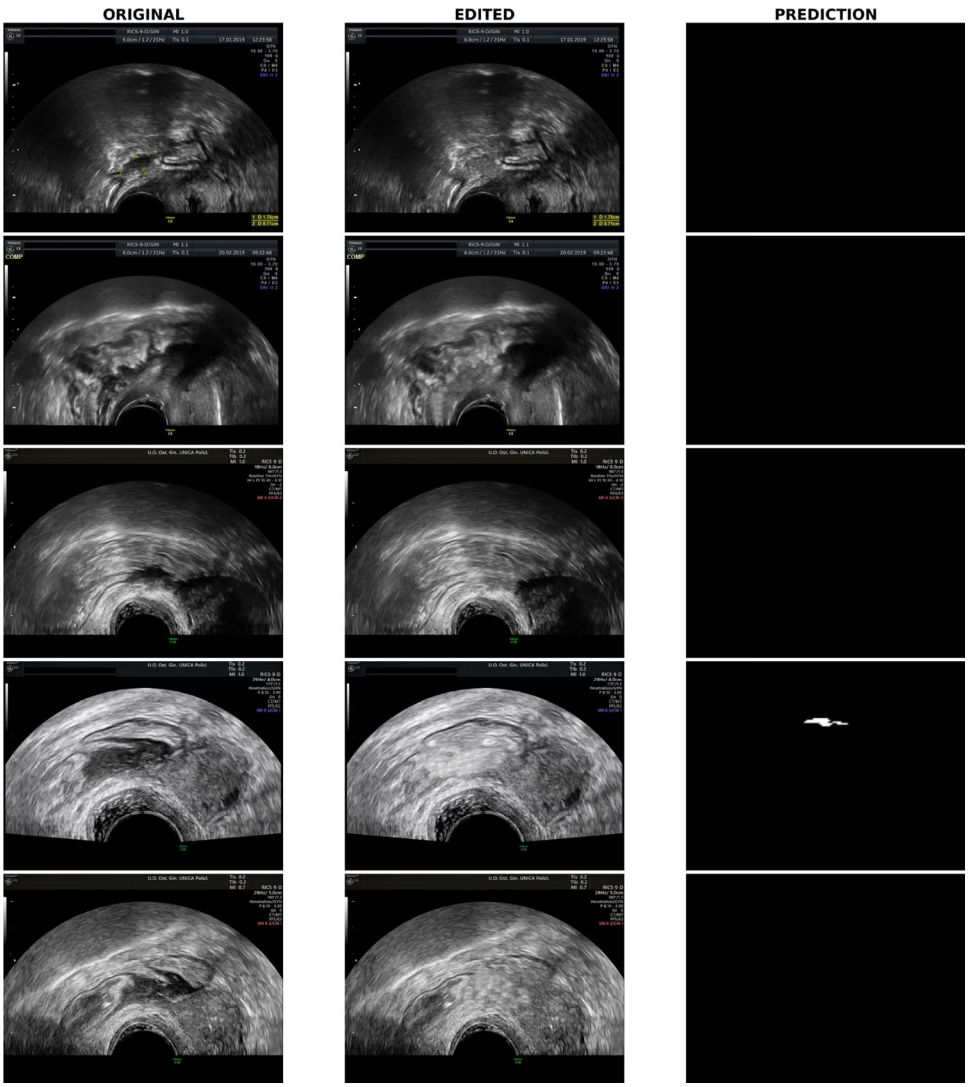

Fig. 6 Qualitative results of the prediction test performed on synthetically generated lesion-free images

addition, even for the false positives, the mask generated is very circumscribed. This lets us argue that our proposed model is also promising in diagnostic settings in which transvaginal ultrasound is performed solely for preventive purposes, i.e., there is a possibility of obtaining true negatives (lesion-free images). Hence, in the event that such images will be available for training purposes in the future, we expect further improvements in such results.

5.5 Solution impact

To speculate on the possible impact of the proposed solution in a real-world scenario, we first remark on how the training and experience of human operators may condition the accuracy of ultrasound diagnosis of endometriosis [41]. Clinical studies show that only in the hands of experienced operators (with more than 10 years of experience in ultrasound gynecology), the accuracy in the diagnosis of some endometriotic lesions is high [42, 43]. For example, an average sensitivity of 78.5% emerged from meta-analyses about the diagnosis of deep endometriosis with transvaginal ultrasound performed by experienced operators [43]. This value, although not directly comparable, can still be related to the metric `dAcc` (detection accuracy) reported in previous Sect. 5.1, which shows a value of 90.6% for the best configuration of our approach. Such evidence consolidates the motivation of our work and the goodness of the results obtained and further highlights how aid methods based on machine/deep learning techniques would be useful tools especially for less experienced sonographers to identify and delineate endometriosis lesions.

6 Conclusions

Transvaginal ultrasound is a safe and low-cost method for the diagnosis of gynecological diseases and reproductive health, whose accuracy, however, is highly dependent on

clinician operator experience. It has been estimated that the learning curve to reach an acceptable ability to recognize deep endometriosis lesions requires at least 100–150 cases, with great individual variability [44]. Since, for the above reasons, tools capable of improving the diagnostic ability of ultrasound are of great interest in research in the field of endometriosis, in this work, we proposed the first automatic segmentation method—based on a multi-scale ensemble pipeline of convolutional neural networks—to support the clinical staff in establishing the diagnosis of such a disease from ultrasound images.

The experiments carried out confirmed the robustness and reliability of the method, showing, in the best configuration, an accuracy of 71% in terms of Jaccard's similarity and a 82% in terms of Dice coefficient. In addition, the deployment of the multi-scale approach proved decisive in boosting the performance of single resolution models, with improvements averaging more than 5% for all the considered architectures.

However, despite promising results, this work stands as pioneering in the field of Computer-Aided Diagnosis applied to endometriosis pathology and therefore still has limitations: in particular, the employed dataset, besides being constrained in size, included only ultrasound images characterized by the presence of lesions (i.e., *true positives*), so the performance of the system in classifying between healthy and diseased tissue was not evaluated. Second, the study focused on the analysis of two-dimensional ultrasound scans, but the increasing prevalence of 3D ultrasound equipment requires generalizing the method to this type of imaging. Nevertheless, we believe that the proposed method and the presented results provide a solid basis for future research in this field.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

Data availability The datasets analyzed during the current study are not publicly available, as they were kindly granted by the *Cagliari Hospital Authority* (AOU Cagliari) for the purposes of this work. However, for reasons of verification and reproducibility of the results, or for further scientific studies, interested parties may expressly request them by sending an email to *sebastianpodda [at] unica.it* and *manuela.neri [at] unica.it*.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Informed consent The dataset used as part of this study is completely anonymous and contains no personal data or attributes with which participants can be traced.

References

1. Zondervan KT, Becker CM, Missmer SA (2020) Endometriosis. N Engl J Med 382(13):1244–1256. https://doi.org/10.1056/NEJMra1810764
2. Hsu AL, Khachikyan I, Stratton P (2010) Invasive and non-invasive methods for the diagnosis of endometriosis. Clin Obstet Gynecol 53(2):413–419. https://doi.org/10.1097/GRF.0b013e3181db7ce8
3. Guerriero S, Saba L, Pascual MA, Ajossa S, Rodriguez I, Mais V, Alcazar JL (2018) Transvaginal ultrasound vs magnetic resonance imaging for diagnosing deep infiltrating endometriosis: systematic review and meta-analysis. Ultrasound Obstetr Gynecol 51(5):586–595. https://doi.org/10.1002/uog.18961
4. Husby GK, Haugen RS, Moen MH (2003) Diagnostic delay in women with pain and endometriosis. Acta Obstet Gynecol Scand 82(7):649–653. https://doi.org/10.1034/j.1600-0412.2003.00168.x
5. Sabottke CF, Spieler BM (2020) The effect of image resolution on deep learning in radiography. Radiol: Artif Intell 2(1):190015. https://doi.org/10.1148/ryai.2019190015
6. Haque MIU, Dubey AK, Hinkle JD (2021) The effect of image resolution on automated classification of chest x-rays. J Med Imaging 10(4):044503–044503. https://doi.org/10.1101/2021.07.30.21261225
7. Vo DM, Lee S-W (2018) Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions. Multimedia Tools Appl 77:18689–18707. https://doi.org/10.1007/s11042-018-5653-x
8. Zhu H, Zhuang Z, Zhou J, Zhang F, Wang X, Wu Y (2017) Segmentation of liver cyst in ultrasound image based on adaptive threshold algorithm and particle swarm optimization. Multimedia Tools Appl 76:8951–8968. https://doi.org/10.1007/s11042-016-3486-z
9. Huang Q, Huang Y, Luo Y, Yuan F, Li X (2020) Segmentation of breast ultrasound image with semantic classification of superpixels. Med Image Anal 61:101657. https://doi.org/10.1016/j.media.2020.101657
10. Gómez W, Leija L, Alvarenga AV, Infantosi AFC, Pereira WCA (2010) Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation. Med Phys 37(1):82–95. https://doi.org/10.1118/1.3265959
11. Cvancarova M, Albregtsen F, Brabrand K, Samset E (2005) Segmentation of ultrasound images of liver tumors applying snake algorithms and GVF. In: International congress series, vol

1281. Elsevier, pp 218–223. https://doi.org/10.1016/j.ics.2005.03.190
12. Ma J, Wu F, Jiang T, Zhao Q, Kong D (2017) Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int J Comput Assist Radiol Surg 12:1895–1910. https://doi.org/10.1007/s11548-017-1649-7
13. Zhao H, Sun N (2017) Improved u-net model for nerve segmentation. In: International conference on image and graphics, pp 496–504. https://doi.org/10.1007/978-3-319-71589-6_43
14. Abraham N, Illanko K, Khan NM, Androutsos D (2019) Deep learning for semantic segmentation of brachial plexus nervesin ultrasound images using u-net and m-net. 2019 3rd international conference on imaging, signal processing and communication (ICISPC), pp 85–89. https://doi.org/10.1109/ICISPC.2019.8935668
15. Xue C, Zhu L, Fu H, Hu X, Li X, Zhang H, Heng P-A (2021) Global guidance network for breast lesion segmentation in ultrasound images. Med Image Anal 70:101989. https://doi.org/10.1016/j.media.2021.101989
16. Podda AS, Balia R, Barra S, Carta S, Fenu G, Piano L (2022) Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images. J Comput Sci 63:101816. https://doi.org/10.1016/j.jocs.2022.101816
17. Lei Y, Wang T, Roper J, Jani AB, Patel SA, Curran WJ, Patel P, Liu T, Yang X (2021) Male pelvic multi-organ segmentation on transrectal ultrasound using anchor-free mask CNN. Med Phys 48(6):3055–3064. https://doi.org/10.1002/mp.14895
18. Beitone C, Troccaz J (2022) Multi-expert fusion: An ensemble learning framework to segment 3d trus prostate images. Med Phys 49(8):5138–5148. https://doi.org/10.1002/mp.15679
19. Singhal N, Mukherjee S, Perrey C (2017) Automated assessment of endometrium from transvaginal ultrasound using deep learned snake. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), pp 283–286. https://doi.org/10.1109/ISBI.2017.7950520
20. Hu S-Y, Xu H, Li Q, Telfer BA, Brattain LJ, Samir AE (2019) Deep learning-based automatic endometrium segmentation and thickness measurement for 2d transvaginal ultrasound. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 993–997. https://doi.org/10.1109/EMBC.2019.8856367
21. Park H, Lee HJ, Kim HG, Ro YM, Shin D, Lee SR, Kim SH, Kong M (2019) Endometrium segmentation on transvaginal ultrasound image using key-point discriminator. Med Phys 46(9):3974–3984. https://doi.org/10.1002/mp.13677
22. Thampi LL, Malarkhodi S (2013) An automatic segmentation of endometrial cancer on ultrasound images. In: 2013 international conference on communication and signal processing, pp 139–143. https://doi.org/10.1109/iccsp.2013.6577032
23. Usha BS, Sandya S (2013) Measurement of ovarian size and shape parameters. In: 2013 annual IEEE india conference (INDICON), pp 1–6. https://doi.org/10.1109/INDCON.2013.6726079
24. Jin J, Zhu H, Zhang J, Ai Y, Zhang J, Teng Y, Xie C, Jin X (2021) Multiple u-net-based automatic segmentations and radiomics feature stability on ultrasound images for patients with ovarian cancer. Front Oncol. https://doi.org/10.3389/fonc.2020.614201
25. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
26. Chen G, Yin J, Dai Y, Zhang J, Yin X, Cui L (2022) A novel convolutional neural network for kidney ultrasound images segmentation. Comput Methods Programs Biomed 218:106712. https://doi.org/10.1016/j.cmpb.2022.106712
27. Wang K, Zhang X, Zhang X, Lu Y, Huang S, Yang D (2022) Eanet: iterative edge attention network for medical image segmentation. Pattern Recogn 127:108636. https://doi.org/10.1016/j.patcog.2022.108636
28. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1007/BF00058655
29. Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198. https://doi.org/10.5555/3013545.3013549
30. Gonzalez RC, Woods RE, Eddins SL (2010) Morphological reconstruction. Digital image processing using MATLAB, MathWorks
31. Leibetseder A, Kletz S, Schoeffmann K, Keckstein S, Keckstein J (2020) GLENDA: gynecologic laparoscopy endometriosis dataset. In: MultiMedia modeling—26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II. Lecture notes in computer science, vol 11962. Springer, pp 439–450. https://doi.org/10.1007/978-3-030-37734-2_36
32. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA (2020) Albumentations: fast and flexible image augmentations. Information 11(2):125. https://doi.org/10.3390/info11020125
33. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
34. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, pp 240–248. https://doi.org/10.1007/978-3-319-67558-9_28
35. Salehi SSM, Erdogmus D, Gholipour A (2017) Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International workshop on machine learning in medical imaging. Springer, pp 379–387. https://doi.org/10.1007/978-3-319-67389-9_44
36. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708. https://doi.org/10.1109/CVPR.2017.243
37. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
38. Chen L, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision—ECCV 2018—15th European conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII. Lecture notes in computer science, vol 11211. Springer, pp 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
39. Su R, Zhang D, Liu J, Cheng C (2021) Msu-net: Multi-scale u-net for 2d medical image segmentation. Front Genet. https://doi.org/10.3389/fgene.2021.639930

40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626. https://doi.org/10.1109/ICCV.2017.74
41. Exacoustos C, Manganaro L, Zupi E (2014) Imaging for the evaluation of endometriosis and adenomyosis. Best Pract Res Clin Obstetr Gynaecol 28(5):655–681
42. Guerriero S, Ajossa S, Minguez J, Jurado M, Mais V, Melis G, Alcazar J (2015) Accuracy of transvaginal ultrasound for diagnosis of deep endometriosis in uterosacral ligaments, rectovaginal septum, vagina and bladder: systematic review and meta-analysis. Ultrasound Obstetr Gynecol 46(5):534–545
43. Bazot M, Thomassin I, Hourani R, Cortez A, Darai E (2004) Diagnostic accuracy of transvaginal sonography for deep pelvic endometriosis. Ultrasound Obstetr Gynecol: Off J Int Soc Ultrasound Obstetr Gynecol 24(2):180–185
44. Indrielle-Kelly T, Fischerova D, Hanuš P, Frühauf F, Fanta M, Dundr P, Lavu D, Cibula D, Burgetova A (2020) Early learning curve in the assessment of deep pelvic endometriosis for ultrasound and magnetic resonance imaging. BioMed Res Int. https://doi.org/10.1155/2020/8757281