Computer Methods and Programs in Biomedicine 260 (2025) 108540

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A multi-task framework for breast cancer segmentation and classification in ultrasound imaging

Carlos Aumente-Maestro, Jorge Díez, Beatriz Remeseiro *

Artificial Intelligence Center, Universidad de Oviedo, Gijón, 33204, Spain

ARTICLE INFO

Keywords:
Ultrasound medical imaging
Dataset curation
Breast cancer classification
Breast cancer segmentation
Multi-task learning

ABSTRACT

Background: Ultrasound (US) is a medical imaging modality that plays a crucial role in the early detection of breast cancer. The emergence of numerous deep learning systems has offered promising avenues for the segmentation and classification of breast cancer tumors in US images. However, challenges such as the absence of data standardization, the exclusion of non-tumor images during training, and the narrow view of single-task methodologies have hindered the practical applicability of these systems, often resulting in biased outcomes. This study aims to explore the potential of multi-task systems in enhancing the detection of breast cancer lesions.

Methods: To address these limitations, our research introduces an end-to-end multi-task framework designed to leverage the inherent correlations between breast cancer lesion classification and segmentation tasks. Additionally, a comprehensive analysis of a widely utilized public breast cancer ultrasound dataset named BUSI was carried out, identifying its irregularities and devising an algorithm tailored for detecting duplicated images in it.

Results: Experiments are conducted utilizing the curated dataset to minimize potential biases in outcomes. Our multi-task framework exhibits superior performance in breast cancer respecting single-task approaches, achieving improvements close to 15% in segmentation and classification. Moreover, a comparative analysis against the state-of-the-art reveals statistically significant enhancements across both tasks.

Conclusion: The experimental findings underscore the efficacy of multi-task techniques, showcasing better generalization capabilities when considering all image types: benign, malignant, and non-tumor images. Consequently, our methodology represents an advance towards more general architectures with real clinical applications in the breast cancer field.

1. Introduction

According to the World Health Organization (WHO), there were 2.3 million women diagnosed with breast cancer and 670,000 deaths worldwide in 2022 [1]. As with the majority of cancer types, early detection plays a crucial role in patient prognosis. In addition to physical exams, blood chemistry studies, or biopsies, screening tests are widely used for early detection. Screening tests allow the identification of breast cancer through imaging modalities at an early stage, when it can be treated and potentially cured [2]. Currently, some of the most common imaging modalities include mammograms, ultrasound (US) images, or digital breast tomosynthesis (DBT), and the choice depends on factors such as the patient's age, breast density, and specific clinical indications [3].

Ultrasound imaging can be especially helpful in women with dense breast tissue, where abnormal areas may be challenging to visualize with mammography [4]. Moreover, its non-ionizing nature makes it a safe screening option during some lifetime periods like pregnancy and youth [5]. However, despite these benefits, US images present drawbacks such as limited contrast, lower spatial resolution compared to alternative modalities, and the presence of speckle noise. These limitations make it difficult for radiologists to segment and classify images, especially to distinguish between benign and malignant tumors [6]. Consequently, computer-aided diagnosis (CAD) systems based on expert knowledge face considerable challenges due to biases present in image annotation.

Artificial intelligence models, particularly convolutional neural networks (CNNs), have been demonstrated to be promising candidates for CAD systems in breast lesions. Many CNN architectures have achieved remarkable results in recent years [7]. These approaches typically rely on previous annotated images to learn characteristic patterns and then

* Corresponding author.
E-mail address: bremeseiro@uniovi.es (B. Remeseiro).

https://doi.org/10.1016/j.cmpb.2024.108540
Received 24 July 2024; Received in revised form 8 November 2024; Accepted 28 November 2024
Available online 4 December 2024

generate segmentation or classification outputs [8]. Hence, having well-annotated datasets is essential for building robust systems capable of generalizing effectively. The vast majority of existing CAD systems rely on single-task methodologies, i.e., solving a single task at once, which typically overlook the intrinsic relationships between related tasks. While CNNs excel at extracting meaningful information from images and transforming it into predicted masks or tumor classes, those single-task methodologies fail to benefit from the correlation between multiple tasks [9].

In this research work, we propose a prediction-refining multi-task framework for breast cancer segmentation and classification. Unlike other methods that employ a single-task approach [10,11], our proposal effectively exploits the existing relationships between both tasks. This multi-task framework consists of different modules aimed at optimizing and refining the predictions while maintaining fairness throughout experimentation. The foundation of the framework is built upon a CNN; however, in contrast to other previous works where the architecture is fixed [12,13], ours offers flexibility in terms of the backbone architecture employed.

Moreover, several studies on automated medical image segmentation [14–16] exclude normal images, i.e., images that do not contain any tumor regions. While this simplifies the segmentation task, it requires manual human intervention to identify and discard non-tumor images, which is critical to avoid unnecessary surgeries in patients without tumors. To address this issue, our proposal allows for the inclusion of non-tumoral cases and incorporates a prediction-refining module to ensure coherence. As a result, it handles non-tumoral cases effectively, making it a fully automatic method applicable in practical settings.

On the other hand, the widespread use of public datasets has been crucial for advancing research across various fields, including computer vision and machine learning. However, the reliability and integrity of those datasets can be compromised by the presence of inconsistencies, potentially impacting the trustworthiness of research findings [17]. In this work, we present an algorithm developed to identify and address duplicated cases, ensuring the consistency of the dataset used for future analyses. Furthermore, we also provide a curated version of a public dataset to promote a more rigorous and fair comparison with future developments in the breast cancer field.

In summary, the contributions of this research work are as follows:

- We built a CAD system for breast cancer that exploits the existing correlations between segmentation and classification tasks, and proves its clinical application in practice.
- We performed an exhaustive ablation study to demonstrate how our multi-task framework outperforms single-task methods regardless of the backbone model chosen.
- We developed a duplicated image recognition algorithm to uncover inconsistencies in a widely used public breast cancer dataset. As a result, we also provide a curated version of this dataset to ensure fairness in the performance analysis of future research works.

The rest of the manuscript is organized as follows. Section 2 reviews the current state-of-the-art methods in breast cancer segmentation and classification from both perspectives: single-task and multi-task learning. The proposed method is described in Section 3. In Section 4 we present a detailed analysis of a public dataset and propose a curated version of it. Section 5 summarizes the implementation details followed in all the experiments along the manuscript. Results are presented and discussed in Section 6. Section 7 concludes the paper.

2. Related work

Over the past decade, the number of research works that employ CNNs to address biomedical segmentation tasks has increased exponentially [18]. This quick expansion was driven by one of the most widely recognized networks in the scientific community, the U-Net, originally introduced by [19]. The U-Net network is characterized by having an encoding branch that extracts the feature maps until reaching the bottleneck, and a decoding branch where the original spatial dimensions are reconstructed. Additionally, it contains skip connection layers between the encoding and decoding paths, facilitating the capture of both low-level and high-level features effectively. This base structure has been commonly used as a skeleton for many architectures in medical challenges [20–22]. Remarkable examples are: Attention U-Net [23], Residual U-Net [24], SegResNet [25], UNet^{++} [26], or SwinUNTER [27], each offering incremental enhancements over the original. It is also worth mentioning nnU-Net, a self-configuring method introduced by [28]. This U-Net-based approach autonomously adjusts preprocessing, training, post-processing, and structural components without manual intervention. nnU-Net outperformed existing methodologies across 23 publicly available datasets commonly used in competitions, demonstrating its high potential in this domain.

2.1. Single-task learning for breast cancer

Numerous studies have recently been conducted to investigate breast cancer segmentation and classification through US images. Almost all of them have adopted a single-task learning approach, which addresses either the classification or the segmentation problem. [14] introduced a two-stage multi-scale architecture prepared to handle speckle noise and shape variations in breast ultrasound cancer segmentation. [29] enhanced the U-Net architecture by incorporating bidirectional attention guidance, making use of saliency maps to capture global and local features from breast US images more effectively. Similarly, [15] presented two novel modules called global guidance block and boundary detection module for boosting the breast US lesion segmentation. All these research works and many others [10,16,30] have shown promising results in the breast ultrasound segmentation domain; however, not only did all the authors exclude normal (i.e., non-tumor) images from their studies but also performed a binary segmentation (without distinguishing between malignant and benign tumors). This omission poses a drawback for clinical applications, as the images are not pre-classified into normal, benign, and malignant categories prior to radiologist evaluation.

Alternatively, other previous works have focused on exploring breast cancer classification, typically including malignant, benign, and normal images. For instance, [31] developed a novel framework that combined metaheuristic optimization algorithms along with deep learning techniques for a more robust feature selection preceding US image classification. [32] proposed a method for addressing semantic similarity in the feature space to overcome the limitations associated with classification layer reliance, thereby enhancing generalizability across multiple datasets, including breast US. [33] used the gold-standard techniques coming from natural language processing to develop a highly robust yet efficient CNN-transformer hybrid model to classify the US images.

2.2. Multi-task learning for breast cancer

Single-task approaches often fail to fully exploit the inherent correlations presented between segmentation and classification tasks. Recognizing this limitation, some authors have explored the potential for handling both tasks jointly. [34] devised an architecture capable of classifying and segmenting volumetric breast ultrasound images. Through an iterative feature-refining strategy employed during training, they successfully mitigated noise in ambiguous boundaries and consequently boosted the model performance for both tasks with respect to single-task baseline models. Similarly, [35] also focused on fuzzy tumor boundaries and irregular shapes, but using 2D images. In this case, they thoroughly analyzed frequency characteristics extracted

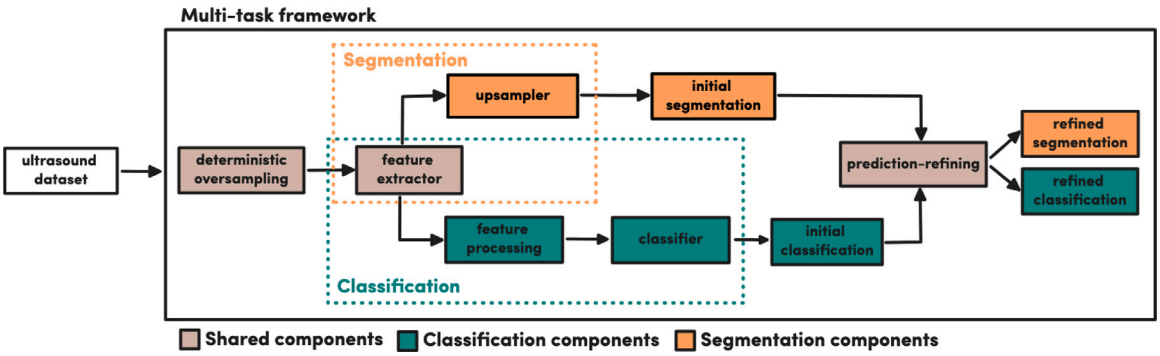

Fig. 1. Overview of the proposed prediction-refining multi-task learning framework. The deterministic oversampling module operates exclusively on the training set; validation and test sets maintain the original class distribution. Conversely, the prediction-refining module applies only to test sets during the inference phase.

from images using the discrete cosine transformation, incorporating this information into a CNN model. [12] proposed utilizing both encoder and decoder level feature maps not only for tumor segmentation but also for tumor classification. These are just a few examples; however, it is noteworthy that in all aforementioned studies, the exclusion of normal images persisted. By contrast, a few other researchers have emphasized the significance of including non-tumor images for the development of clinically applicable CAD systems. For example, [36] developed a multi-branch U-Net architecture capable of performing breast cancer segmentation and classification using malignant, benign, and normal images. The authors introduced an additional auto-encoder branch for image reconstruction, which worked as a regularization mechanism.

Although several authors have conducted studies about breast lesions, the domain is still insufficiently explored. The exclusion of non-tumor images and the lack of reliability and integrity within datasets make most of the above single-task and multi-task learning methods unreliable in clinical practice.

3. Methodology

We propose a fully automated end-to-end multi-task framework for breast ultrasound images. The framework aims to jointly address classification and segmentation tasks by taking breast US images as inputs and producing two outputs: a binary segmentation mask and a multi-class tumor label. Fig. 1 depicts the proposed method, and their components described below.

3.1. Shared components

The shared components encompass all common pieces required to solve both tasks. These are the *deterministic oversampling*, the *feature extractor*, and the *prediction-refining* modules.

1. **Deterministic oversampling**: The inherent randomness in oversampling techniques used to address class imbalance may significantly influence model performance. This phenomenon is particularly noticeable when working with small medical image datasets (especially with ultrasound images), where the low quality of certain images can hinder model learning. Additionally, factors such as the accuracy in the annotation of ground truth masks or anomalies within the dataset can either favor or hinder model performance for specific instances. The proposed module offers a suitable solution to this issue and it has been designed to ensure fairness within the experimentation.
   The strategy followed by the *deterministic oversampling* module involves replicating all the images within a class, rather than individual instances randomly. The process of balancing the dataset begins by calculating the proportion of each class (P_C) in relation to the total number of examples, which provides a measure of class distribution. Next, a replication factor (RF) is computed for each class. This factor determines how many times a class needs to be replicated to achieve balance, ideally bringing its adjusted proportion closer to 1. Specifically, the replication factor for a class c is calculated as $RF_c = \lceil \frac{1}{P_c} \rceil$. Finally, all examples of each class are replicated according to their corresponding replication factor, ensuring a more balanced number of examples per class. For example, in a training set where the distribution is as follows: benign 49.3%, malignant 36.4%, and normal 14.2%, the scaling factor for each class would be 2, 3, and 7, respectively; resulting in the new class distribution: benign 32.2%, malignant 35.6%, and normal 32.2%. This approach ensures that, although the classes are not completely balanced, the model is not affected by randomness-induced factors.
2. **Feature extractor**: It refers to the component responsible for extracting meaningful information, also called features, from input images. The numeric features are extracted by the encoding branch of the backbone network selected. This branch typically aggregates a sequence of convolutional, max-pooling, and activation layers to extract relevant information, which is subsequently utilized for both segmentation and classification tasks. In our method, the *feature extractor* is flexible and its role can be performed by the backbone of a CNN architecture.
3. **Prediction-refining**: The objective of this refinement module is to bring consistency to the segmentation and class predicted by the multi-task approach, as well as to enhance its performance. The *prediction-refining* module follows two strategies:

   - Segmentation refinement: When the predicted segmentation contains tumor pixels, but the predicted class is normal, there is a lack of consistency from the medical point of view. To address this issue, the *prediction-refining* module corrects the predicted segmentation by marking those pixels that had been labeled as tumor pixels as non-tumor pixels.
   - Classification refinement: When the predicted segmentation does not contain any pixels labeled as a tumor, but the predicted class is malignant or benign, again there is an incoherence. In these cases, the *prediction-refining* module corrects the predicted class and it is assigned as normal.

   Regardless of whether segmentation and classification are treated as isolated tasks or using a multi-task methodology, such contradictory results may occur, making this module indispensable.

3.2. Segmentation components

The segmentation elements are responsible for generating a binary segmentation mask where each pixel is labeled as 0 (non-tumor) or 1 (tumor).

1. **Upsampler**: The segmentation components comprise an *upsampler* module, responsible for reconstructing the features extracted by the *feature extractor*. This module is essentially a decoding branch, often utilizing transposed convolutions, that restores the spatial dimensions of the input images. As it happens in the *feature extractor* where the backbone is flexible, our *upsampler* module is interchangeable and it also allows the use of deep supervision layers [37] to enhance the segmentation maps output.

The segmentation branch provides an initial binary segmentation that feeds the *prediction-refining* module above described to output a refinement segmentation.

3.3. Classification components

The classification components categorize the input US images into three possible classes (benign, malignant, or normal) through two modules: the *feature processing* and the *classifier*.

1. **Feature processing**: The level just before the bottleneck captures features with higher spatial resolution and more local context, while the bottleneck contains low-level features and a more abstract representation. Previous authors demonstrated that combining these features enables the network to capture both local details and high-level contextual information [34]. Consequently, the *feature processing* module is responsible for merging the multi-scale features. Initially, it upsamples the bottleneck to recover the spatial resolution of the previous level. Then, it concatenates these feature maps with those coming from the encoding and decoding branches, resulting in a multi-scale feature map.
2. **Classifier**: The concatenated multi-scale features pass through a convolutional layer with 512 3×3 kernels, followed by normalization and ReLU activation layers. A 2D adaptive average pooling is applied over the feature volume and then flattened to feed into a densely connected layer with 256 neurons. Finally, an output layer with three neurons and a softmax activation function is added, allowing to associate probabilities to each predicted class.

The classification branch provides an initial multi-class prediction, which also feeds the *prediction-refining* module to output a refinement multi-class prediction.

3.4. Loss function

As part of our proposed methodology to simultaneously segment and classify ultrasound images, we implemented a custom loss function.

The Dice similarity coefficient [38] used in this work can be defined as follows for two segmentation masks S and $\hat{S}$:

$$\text{Dice}\,(S, \hat{S}) = \frac{2\,S \cdot \hat{S} + \epsilon}{S^2 + \hat{S}^2 + \epsilon} \tag{1}$$

where S represents the segmentation map from the proposed multi-task framework, $\hat{S}$ the ground truth segmentation, and ϵ a constant to avoid zero divisions set to 1.

Since the *upsampler* component could benefit from deep supervision layers, our method may output more than one segmentation map. To manage the contribution of the ith segmentation map to the overall loss function, we introduced a weighting parameter w_i. The weight assigned to each of these segmentation maps is inversely proportional to their depth in the network. This weighting strategy ensures that the loss function gives more importance to the segmentation maps from the top layers than those coming from hidden layers. Eq. (2) shows the segmentation loss function used in this work.

$$\mathcal{L}_{seg} = \sum_{i}^{n} w_i(1 - \text{Dice}) \tag{2}$$

where Dice is defined in Eq. (1), n is the number of deep supervision layers, and w_i is defined as i^{-1}.

Conversely, Eq. (3) shows the focal loss function [39] employed for classification.

$$\mathcal{L}_{cls} = -(1 - p_i)^{\gamma} \log(p_i) \tag{3}$$

where p_i is the predicted probability of class i and γ is the focusing parameter set to 2 according to [39].

In our framework, we combined classification and segmentation losses into a single multi-task loss function as follows:

$$\mathcal{L}_{multi-task} = \lambda \mathcal{L}_{seg} + (1 - \lambda)\mathcal{L}_{cls} \tag{4}$$

where λ is a modularization hyperparameter that balances the weight given to each task.

4. Dataset

This section introduces a widely used public dataset for breast cancer segmentation and classification and how it was prepared for this research work.

4.1. BUSI

Breast UltraSound Imaging (BUSI) dataset was originally published by Al-Dhabyani et al. [40] in 2020 to make progress in the breast cancer domain. BUSI is comprised of 780 ultrasound images categorized into three classes: benign (56.0%), malignant (26.9%), and normal (17.1%). Both benign and malignant images were annotated by expert radiologists to provide a pixel-map mask (ground truth segmentation). The images have a wide range of sizes, varying from 190 to 1048 pixels. It is worth mentioning that the images of BUSI dataset do not contain both malignant and benign tumor pixels simultaneously within the same image. However, some cases contain more than one tumor region belonging to the same class.

This dataset has been widely used within the scientific community in recent years [41]. Nevertheless, previous authors have detected some anomalies in the annotations. For example, [17] demonstrated that some cases appear to be incoherent. After carefully exploring the dataset we realized that many cases seem to be repeated and misclassified. To further investigate the nature of BUSI and avoid bias in the model evaluation, a *duplicate image recognition* algorithm is proposed in this work.

4.2. Duplicate image recognition algorithm

To quantify the level of similarity between the images of the dataset, we took advantage of the Structural Similarity Index Measure (SSIM) [42]. It was originally designed as an alternative to MSE or MAE to assess the quality of an image, since they were not always well-aligned with human perception. SSIM, by contrast, takes into account important aspects such as image structure and texture, and it is commonly used to determine the perceptual similarity between two images. Specifically, SSIM evaluates three main components to calculate the similarity: the images' luminance, contrast, and structure. The values of these three components are then combined to produce an SSIM value in the range [0, 1]. On this scale, 1 suggests perfect similarity, whereas 0 represents complete dissimilarity.

Our *duplicate image recognition* algorithm relies on SSIM to analyze and uncover duplicated images within the BUSI dataset, and it allowed us to find all the anomalies presented in it. See more details in the following algorithm.

Step 1. The SSIM is calculated for every pair of images in the given dataset, excluding comparisons with themselves.
Step 2. Choose an image, which will be called a reference image (RI), and sort its comparisons in descending order by SSIM.

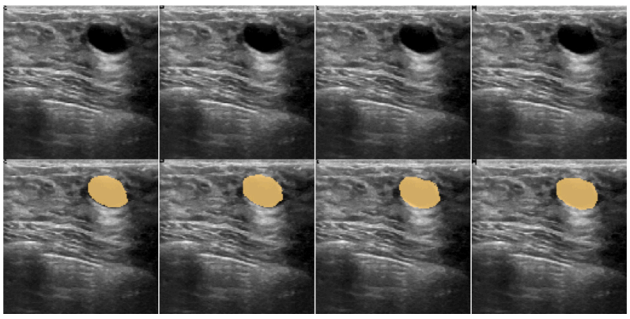

Fig. 2. Example of four repeated images (quadruplet). From left to right: ID 139 benign, ID 157 benign, ID 65 benign, ID 99 benign.

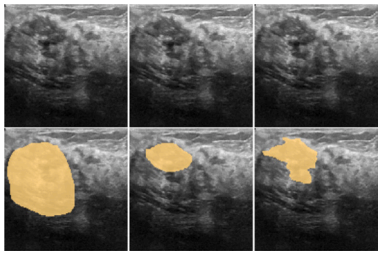

Fig. 3. Example of three repeated images (triplet). From left to right: ID 131 benign, ID 42 benign, ID 51 malignant.

Step 3. Select the top-ranked image based on the SSIM and manually check whether it is apparently identical to the RI.

Step 4a. When the selected image and the RI are different, the process concludes for the RI chosen, determining that the RI has no more duplicates within the dataset.

Step 4b. When the selected image and the RI are identical, a duplicate of the RI has been found. Remove the duplicated image found from the ranking and go to Step 3.

Step 5. If there are images left to analyze, go to step 2.

4.3. Curated BUSI

This section presents the results, both qualitative and quantitative, obtained after applying the *duplicate image recognition* algorithm to the BUSI dataset. We also suggested a post-processed dataset called Curated BUSI, which is a standardized version of the original one.

Once the algorithm carefully described in Section 4.2 was applied for all the images in BUSI, a total of three different cases were uncovered.

(a) Quadruplets: A total of 5 quadruplets of repeated images were found in the original BUSI dataset. Fig. 2 illustrates four images extracted from the BUSI dataset and their corresponding segmentation masks, all of them slightly different but belonging to the same tumor class.

(b) Triplets: A total of 22 triplets of repeated images were found in the original BUSI dataset. Fig. 3 shows three ultrasound images, representing two benign classes and one malignant class, each of them provided with a notably different segmentation mask by the radiologists.

(c) Duplets: A total of 122 duplets of repeated images were found in the original BUSI dataset. Fig. 4 shows again two ultrasound images where both the mask and the class are different for the same image.

Similar discrepancies are observed in other numerous images detected by the algorithm, however, it is essential to determine whether

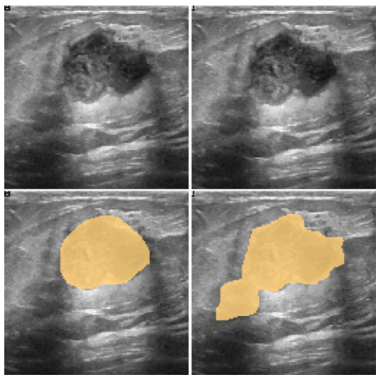

Fig. 4. Example of two repeated images (duplet). From left to right: ID 406 benign, ID 94 malignant.

Table 1
Average and standard deviation from duplicated and non-duplicated SSIM distributions.

	Duplicated images	Non-duplicated images
SSIM	0.614 ± 0.172	0.103 ± 0.040

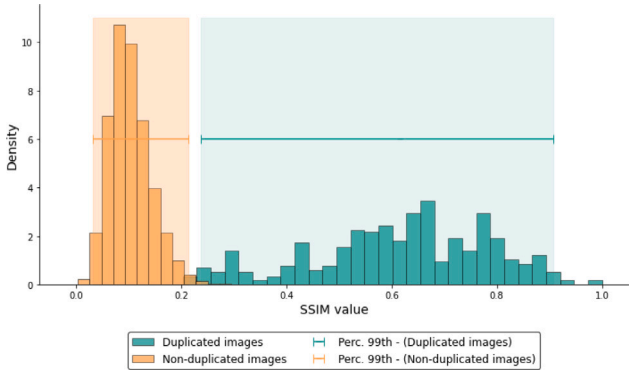

Fig. 5. Distribution of SSIM (Structural Similarity Index) values for duplicated (green) and non-duplicated (orange) images. Histograms represent the normalized densities of SSIM values for both groups. The shaded regions indicate the range of SSIM values encompassing the 99th percentile intervals.

the SSIM values derived from repeated images are statistically significantly higher than those from non-repeated images (see Table 1). Initially, the Lilliefors test [43] was applied to both SSIM distributions, the one coming from duplicated and non-duplicated images. This statistical test assumes as a null hypothesis, H_0, that data come from a normally distributed population. The test resulted negative in both cases, rejecting the null hypothesis and concluding that none of the samples followed a normal distribution. Consequently, we perform a non-parametric test to analyze the discrepancies, specifically the Mann–Whitney test [44]. This statistical test is built under the null hypothesis, H_0, that the distributions of both populations are identical. The p-value obtained from the test was statistically significant enough (<0.0001) to reject H_0. Therefore, this statistical test not only confirms the visual inspection findings but also provides robust quantitative evidence supporting the presence of duplicated cases within the BUSI dataset.

To complement the previous results, Fig. 5 presents the distribution of SSIM values for duplicated and non-duplicated images, alongside a shadowed range representing the 99th percentile. The plot visually supports the statistically significant differences found by the Mann–Whitney test and suggests a threshold SSIM value ≈ 0.23 to uncover potential duplicates.

Table 2
Grade of overlap in the ground truth annotations provided in the BUSI dataset for duplicated images. IoU: Intersection-over-Union metric.

Overlapping grade	Count	Distribution
IoU = 1	88	40.4%
IoU < 0.9	92	42.2%
IoU < 0.7	17	7.8%
IoU < 0.5	7	3.2%
IoU < 0.3	9	4.1%
IoU = 0	5	2.3%

Table 3
Class distributions for BUSI and Curated BUSI datasets.

Dataset	Normal	Benign	Malignant	Total
BUSI	133 (17.1%)	437 (56.0%)	210 (26.9%)	780
Curated BUSI	64 (14.2%)	222 (49.3%)	164 (36.4%)	450

Apart from studying the similarity between the US images, we also analyzed the consensus between the masks provided within BUSI for those duplicated images. Since triplets and quadruplets are respectively a set of three and four duplicated images, all the possible combinations in a set were analyzed pairwise. Table 2 shows the grade of overlap between duplicated images based on the Intersection-over-Union metric. As can be seen, there are quite a few ground truth masks coming from similar US images whose overlapping is very poor.

To conclude this section, we would like to mention the potential harm of using low-quality datasets on scientific achievements. When identical images appear in both the training and test sets, model evaluation can be misleading: if the images share the same ground truth, the model's performance may be overestimated; however, if they have different ground truths, the performance may be underestimated. The authors of the BUSI dataset mentioned in the original paper [40] that duplicated images required to be removed as well as incorrect annotation; however, it seems the task was not entirely completed. Given the disparity in the annotations, in our view, all those duplicated images found within the dataset must be removed to ensure consistency in the results provided by CAD systems. The proposed curated version of BUSI has been thoroughly analyzed to preserve the integrity and reliability of the results within the scientific community. All the information relating to this post-processed dataset along with the duplicated image identifiers (IDs) are available in our repository.[1,2] Table 3 provides a comparison between BUSI and the proposed Curated BUSI datasets in terms of class distribution and total images. Compared to the original BUSI, the majority and minority classes still consist of benign and normal images, respectively, but the total number of images has decreased by 330, representing a 42% reduction.

5. Experimental setup

This section outlines the setup considered for the experiments carried out and provides a summary of the metrics employed for evaluating the models.

5.1. Dataset details

After the findings in the original BUSI dataset presented in Section 4.3 about repeated images, in some cases with different classes and masks (see Figs. 2, 3, and 4), all the experiments conducted in this research work were performed using Curated BUSI dataset. Inspired by previous works [29] using BUSI and to be fair in the evaluation of the model performance, we adopted a 4-fold cross-validation strategy. For future comparisons, the IDs used for testing in every fold are publicly available in our repository.

Before feeding the models with the ultrasound images, we applied data augmentation on-the-fly to increase the variability during the training phase. Every epoch the input images were flipped along both axes with probability 0.5 and rotated between [0, 360) degrees. As a result, the number of train images per epoch was the same, but they were slightly different. Finally, given the size variability of the images, all the images were resized to 128 × 128 pixels.

5.2. Performance measures

In order to measure the performance derived from the experiments conducted in this research, a set of metrics widely used in both segmentation and classification tasks was selected.

Accuracy (ACC) and F1-score (F1) are commonplace metrics in classification. Both are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{5}$$

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

F1 score is particularly useful in medical diagnosis, where both false positives and false negatives have significant consequences, especially in scenarios in which there is an unbalanced class distribution, such as in the Curated BUSI dataset. Therefore, the combination of these two metrics allows us to assess whether the system presents bias for some specific class or it generalizes correctly.

Concerning the segmentation task, the Dice similarity coefficient (DSC) is one of the most popular overlap-based performance measures in the medical imaging context. It is defined as follows:

$$\mathrm{DSC} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

Apart from the DSC, other segmentation metrics like Hausdorff distance (HD), Jaccard index (JAC), recall (REC), and precision (PRE) were used in this manuscript to report the results [45].

5.3. Implementation details

We implemented our proposal using Python 3.9 and PyTorch 2.0. All the experiments were run on an NVIDIA GeForce RTX 3080 10 GB GPU. The framework was trained during 200 epochs using the Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 2. Regarding the λ hyperparameter from Eq. (4), it was determined that setting it to 0.85 resulted in optimal performance across both tasks.

6. Results and discussion

This section presents the experiments carried out in this research work.

6.1. Breast cancer segmentation on curated BUSI

Since previous works did not undertake the cleaning procedures outlined in Section 4, we conducted a comparative study on the Curated BUSI dataset. We took some of the state-of-the-art architectures listed in Section 2 and customized them to address a binary segmentation task. That is, pixels within the segmentations are categorized as either tumor (regardless of malignancy) or non-tumor.

Table 4 shows the results obtained by the selected architectures. Alongside diverse metrics and the model utilized in each experiment, the table also includes a "Subset" column. This column specifies the dataset upon which the model was trained and tested. The idea behind

[1] https://github.com/caumente/multi_task_breast_cancer.
[2] Available after paper acceptance.

Table 4
Results obtained in breast cancer binary segmentation with each architecture trained on the Curated BUSI dataset using a 4-fold cross-validation.

Subset	Architecture	HD	DSC	JACC	REC	PRE
Benign & Malignant (386 images)	U-Net [19]	29.916 ± 2.615	0.672 ± 0.014	0.552 ± 0.021	0.698 ± 0.022	0.739 ± 0.019
	Attention U-Net [23]	22.398 ± 2.670	0.748 ± 0.012	0.648 ± 0.017	0.794 ± 0.023	0.783 ± 0.018
	UNet^{++} [26]	14.812 ± 2.888	**0.762 ± 0.012**	**0.669 ± 0.019**	**0.798 ± 0.017**	0.817 ± 0.016
	SwinUNETR [27]	18.211 ± 2.857	0.733 ± 0.022	0.634 ± 0.026	0.779 ± 0.019	0.772 ± 0.016
	Residual U-Net [24]	24.071 ± 1.432	0.625 ± 0.019	0.515 ± 0.023	0.707 ± 0.023	0.756 ± 0.027
	SegResNet [25]	16.809 ± 2.351	0.739 ± 0.011	0.645 ± 0.014	0.772 ± 0.020	0.817 ± 0.017
	nnU-Net [28]	**13.903 ± 2.631**	0.751 ± 0.029	0.659 ± 0.033	0.796 ± 0.028	**0.820 ± 0.020**
	BTS U-Net [46]	15.604 ± 2.203	0.752 ± 0.018	0.654 ± 0.021	0.788 ± 0.022	0.810 ± 0.018
Normal & Benign & Malignant (450 images)	U-Net	30.383 ± 1.782	0.575 ± 0.018	0.472 ± 0.022	0.694 ± 0.026	0.732 ± 0.005
	Attention U-Net	20.446 ± 2.009	0.642 ± 0.015	0.556 ± 0.019	0.791 ± 0.016	0.790 ± 0.013
	UNet^{++}	12.937 ± 3.219	**0.680 ± 0.038**	**0.602 ± 0.040**	0.805 ± 0.012	**0.832 ± 0.014**
	SwinUNETR	18.498 ± 3.208	0.625 ± 0.019	0.539 ± 0.024	0.771 ± 0.026	0.769 ± 0.028
	Residual U-Net	16.409 ± 2.196	0.601 ± 0.031	0.511 ± 0.031	0.686 ± 0.042	0.791 ± 0.028
	SegResNet	16.073 ± 3.306	0.655 ± 0.023	0.574 ± 0.026	0.796 ± 0.011	0.811 ± 0.009
	nnU-Net	**12.673 ± 2.545**	0.651 ± 0.021	0.573 ± 0.024	0.790 ± 0.025	0.827 ± 0.025
	BTS U-Net	15.409 ± 2.183	0.652 ± 0.015	0.570 ± 0.021	**0.809 ± 0.023**	0.818 ± 0.028

this differentiation is that many authors proposed segmentation models excluding normal images, arguing that images and segmentation without any tumor pixels are meaningless. Nonetheless, we believe that incorporating normal images in experimentation is essential to have a fully automatic CAD system in clinical practice.

Our evaluation of breast cancer binary segmentation on the Curated BUSI dataset highlights a superior performance in terms of overlapping for UNet^{++}. Specifically, it attained the highest average DSC scores, 0.762 and 0.680 for the Benign & Malignant and Normal & Benign & Malignant subsets, respectively. These results align with previous works, where authors demonstrated the generalization power of UNet^{++} in segmenting medical images from different domains [26,47]. Additionally, it obtained competitive results in the remaining metrics.

It is important to highlight how the performance obtained when using Benign & Malignant & Normal instead of only Benign & Malignant images decreased considerably for all the methods tested. However, despite the inclusion of normal images increasing the training time and model complexity, end-to-end CAD systems must consider these cases to be fully automatic.

In view of the results, UNet^{++} was selected as one of the backbone networks to further evaluate our method. Additionally, we also chose nnU-Net given its popularity in recent years within the research community and its noticeable results in various medical imaging challenges [28].

6.2. Results of the proposed multi-task method and ablation study

In order to demonstrate the importance of each module of our proposed framework, we performed a comprehensive ablation study taking UNet^{++} and nnU-Net as backbone networks.

Tables 5 and 6 show the results obtained by each network configuration, including those trained following a single-task (ST) approach for segmentation (Seg) and classification (Cls), and employing the multi-task (MT) approach. Additionally, we explore variations such as integrating the *prediction-refining* module (PR) and the *deterministic oversampling* module (DO) into the single-task experiments and multi-task approach. Results are carefully presented for both segmentation and classification tasks, as well as across tumor classes.

6.2.1. Segmentation analysis results

Comparing the figures from multi-task against single-task approaches, the average DSC values were increased by 13.6% for the UNet^{++} network, from 0.661 to 0.751. Similarly, the average DSC gain was about 15.5% for the nnU-Net, from 0.653 to 0.754. Analyzing the segmentation results from any of Tables 5 and 6, it can be seen that the improvement has substantially stemmed from the accurate segmentation of normal images, in part attributed to the *prediction-refining module*. In most of the previous works utilizing the BUSI dataset, normal images were not considered. As demonstrated in Table 4, they tend to significantly worsen segmentation results. However, our proposal successfully addressed this issue by incorporating non-tumor images while resulting in superior performance. As a result, our CAD system becomes fully automatic, requiring no manual intervention.

Furthermore, a qualitative analysis was conducted on the experimental results. Fig. 6 depicts a collection of cases, showing both successful (green boxes) and wrong (red boxes) outcomes. The first three rows exemplify successful cases in which segmentation predicted outperformed single-task (Seg). On the other hand, the red box reveals that the multi-task framework encountered challenges in assessing some cases.

6.2.2. Classification analysis results

A similar scenario is observed for the classification task. When examining the single-task models (Cls), it is evident that their main challenge lies in classifying normal images. By contrast, the multi-task framework enhanced the F1 metric for this kind of images. Moreover, not only did the F1 metric improve notably with respect to the single-task method for all the classes, but it also improved the accuracy considerably. Therefore, in addition to outperforming on average the single-task segmentation approach, the multi-task approach can reliably classify the tumor. Specifically, improvements in accuracy for UNet^{++} were 14.9%, from 0.698 to 0.802, while for the nnU-Net they were 14.7%, from 0.680 to 0.780.

A confusion matrix is depicted in Table 7 to go further in the interpretation results. It presents a classification performance comparison between the single-task method UNet^{++} (Cls) and our proposed approach, UNet^{++} (MT + PR + DO). Based on the figures, the UNet^{++} (Cls) configuration achieved an accuracy of 73.9%, 63.4%, and 71.9% along the diagonal for the benign, malignant, and normal classes, respectively. By contrast, our proposed method demonstrated substantial improvements, achieving an accuracy of 83.3%, 74.4%, and 79.7% for the same classes, respectively.

The findings presented in this section demonstrate how the prediction-refining multi-task framework effectively leverages existing synergies between classification and segmentation tasks. In particular, when the system is used in its entirety, considering the DO and PR modules, the multi-task approach boosts average performance in both tasks. The PR and DO modules made that both ST methods improved their performance. However, solving the tasks disjointly is computationally more expensive and still does not reach the performance of the multi-task approach. It is worth mentioning that the ST approach generated a ratio of 15.3% incoherent predictions, i.e., cases classified as normal by the classifier even if they contain tumor pixels or cases that do not contain tumor pixels but are classified as non-normal. By contrast, that ratio decreased to 10.9% for the MT model, showing how optimizing both tasks at once favors learning.

Table 5
Ablation study of the proposed prediction-refining multi-task learning framework for UNet^{++} trained on Curated BUSI. ST: Single-task. MT: Multi-task. DO: Deterministic oversampling module. PR: Prediction-refining module.

UNet^{++} architecture	Segmentation				Classification				
	DSC Benign	DSC Malignant	DSC Normal	DSC Average	F1 Benign	F1 Malignant	F1 Normal	F1 Weighted	ACC
ST (Seg)	0.799 ± 0.036	**0.731 ± 0.040**	0.000 ± 0.000	0.661 ± 0.017	–	–	–	–	
ST (Seg) + DO	0.797 ± 0.056	0.730 ± 0.052	0.234 ± 0.469	0.692 ± 0.044	–	–	–	–	
ST (Cls)	–	–	–	–	0.707 ± 0.057	0.673 ± 0.069	0.694 ± 0.048	0.693 ± 0.038	0.698 ± 0.038
ST (Cls) + DO	–	–	–	–	0.745 ± 0.018	0.717 ± 0.077	0.643 ± 0.088	0.720 ± 0.042	0.725 ± 0.036
ST (Seg + Cls) + PR	0.744 ± 0.051	0.726 ± 0.041	0.719 ± 0.081	0.734 ± 0.031	0.707 ± 0.057	0.673 ± 0.069	0.694 ± 0.048	0.693 ± 0.038	0.698 ± 0.038
ST (Seg + Cls) + PR + DO	0.770 ± 0.046	0.730 ± 0.052	0.656 ± 0.231	0.739 ± 0.023	0.751 ± 0.025	0.717 ± 0.077	0.679 ± 0.116	0.729 ± 0.049	0.733 ± 0.043
MT	**0.806 ± 0.055**	0.713 ± 0.066	0.000 ± 0.000	0.658 ± 0.025	0.788 ± 0.018	0.687 ± 0.070	0.643 ± 0.162	0.730 ± 0.036	0.742 ± 0.028
MT + PR	0.787 ± 0.057	0.707 ± 0.062	0.594 ± 0.277	0.731 ± 0.025	0.788 ± 0.018	0.687 ± 0.070	0.643 ± 0.162	0.730 ± 0.036	0.742 ± 0.028
MT + DO	0.773 ± 0.050	0.717 ± 0.058	0.703 ± 0.156	0.742 ± 0.018	**0.830 ± 0.015**	0.789 ± 0.023	0.721 ± 0.094	0.799 ± 0.018	0.802 ± 0.027
MT + PR + DO (Our proposal)	0.773 ± 0.050	0.711 ± 0.057	**0.781 ± 0.157**	**0.751 ± 0.018**	0.826 ± 0.029	**0.791 ± 0.025**	**0.741 ± 0.060**	**0.801 ± 0.026**	**0.802 ± 0.018**

Table 6
Ablation study of the proposed prediction-refining multi-task learning framework for nnU-Net trained on Curated BUSI. ST: Single-task. MT: Multi-task. DO: Deterministic oversampling module. PR: Prediction-refining module.

nnU-Net architecture	Segmentation				Classification				
	DSC Benign	DSC Malignant	DSC Normal	DSC Average	F1 Benign	F1 Malignant	F1 Normal	F1 Weighted	ACC
ST (Seg)	0.793 ± 0.024	0.720 ± 0.084	0.000 ± 0.000	0.653 ± 0.019	–	–	–	–	
ST (Seg) + DO	0.773 ± 0.047	0.698 ± 0.070	0.594 ± 0.194	0.720 ± 0.031	–	–	–	–	
ST (Cls)	–	–	–	–	0.740 ± 0.025	0.665 ± 0.052	0.276 ± 0.271	0.647 ± 0.060	0.680 ± 0.043
ST (Cls) + DO	–	–	–	–	0.742 ± 0.027	0.721 ± 0.035	0.618 ± 0.112	0.717 ± 0.033	0.720 ± 0.028
ST (Seg + Cls) + PR	0.783 ± 0.027	0.715 ± 0.082	0.219 ± 0.282	0.678 ± 0.054	0.740 ± 0.025	0.665 ± 0.052	0.276 ± 0.271	0.647 ± 0.060	0.680 ± 0.043
ST (Seg + Cls) + PR + DO	0.747 ± 0.039	0.694 ± 0.067	0.766 ± 0.164	0.730 ± 0.034	0.758 ± 0.026	0.725 ± 0.031	**0.756 ± 0.086**	0.746 ± 0.021	0.747 ± 0.021
MT	**0.806 ± 0.041**	**0.722 ± 0.056**	0.141 ± 0.281	0.681 ± 0.034	0.770 ± 0.033	0.747 ± 0.029	0.655 ± 0.025	0.745 ± 0.029	0.747 ± 0.028
MT + PR	0.783 ± 0.038	0.721 ± 0.058	0.656 ± 0.108	0.742 ± 0.017	0.773 ± 0.034	0.747 ± 0.029	0.674 ± 0.055	0.750 ± 0.032	0.751 ± 0.031
MT + DO	0.762 ± 0.054	0.696 ± 0.080	0.672 ± 0.180	0.725 ± 0.031	0.795 ± 0.045	**0.752 ± 0.022**	0.696 ± 0.095	0.765 ± 0.032	0.769 ± 0.034
MT + PR + DO (Our proposal)	0.779 ± 0.046	0.717 ± 0.049	**0.766 ± 0.103**	**0.754 ± 0.035**	**0.806 ± 0.005**	0.751 ± 0.040	0.741 ± 0.109	**0.777 ± 0.019**	**0.780 ± 0.017**

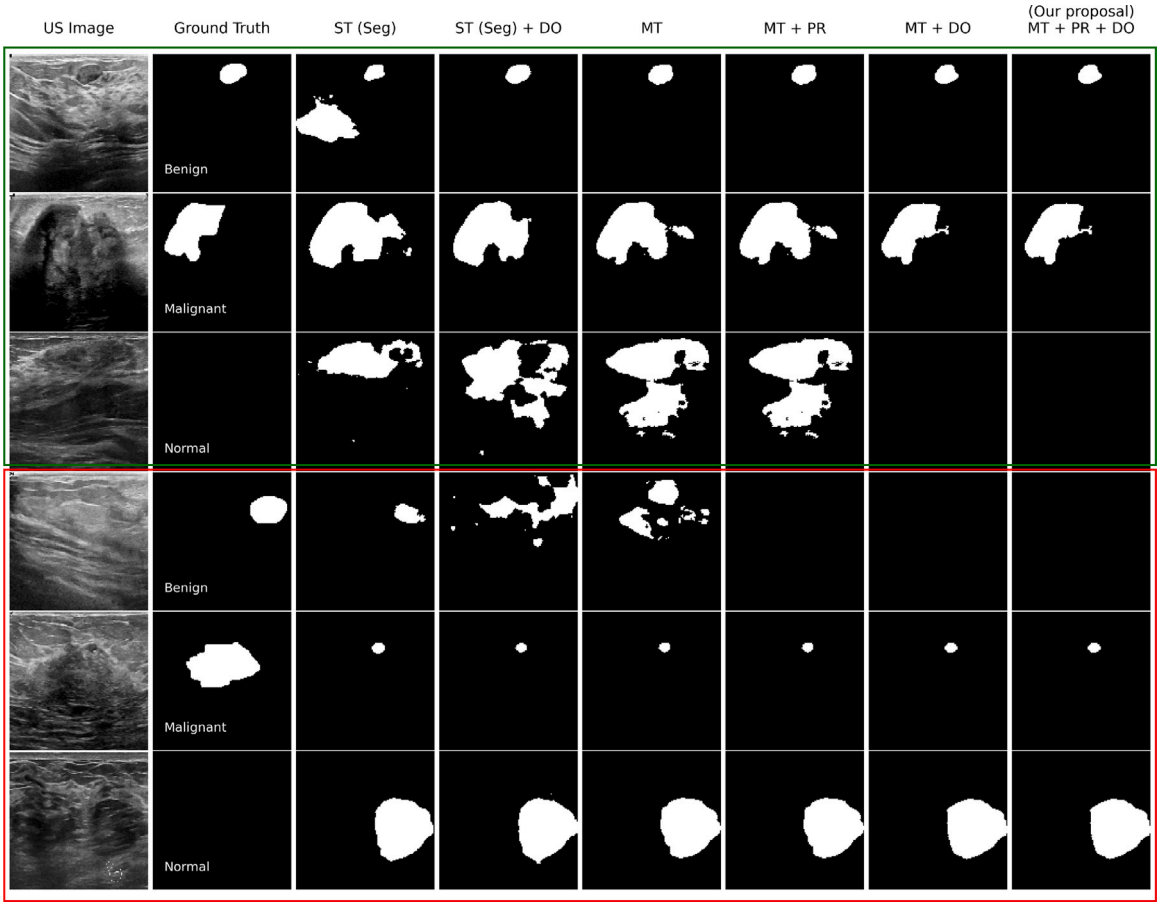

Fig. 6. Qualitative results depicting successful (green box) and unsuccessful (red box) segmentation cases. The first two columns display the ultrasound image and its corresponding ground truth. Subsequent columns depict the predicted segmentation by each UNet^{++} configuration. The cases shown, from top to bottom, are: Benign 337, Malignant 176, Normal 36, Benign 61, Malignant 141, and Normal 113. ST: Single-task. MT: Multi-task. DO: Deterministic oversampling module. PR: Prediction-refining module.

6.3. Comparison with state-of-the-art multi-task methods

In order to fully evaluate the performance of our method, we conducted a comparison with, to the best of our knowledge, the sole end-to-end multi-task approach identified in the literature wherein the authors employed the BUSI dataset and incorporated normal images [36]. However, as the authors did not provide the IDs of the images employed for testing their method, and considering that our

Table 7
Confusion matrix illustrating classifications for the benign, malignant, and normal classes. The top table depicts results obtained with the UNet^{++} single-task configuration (Cls). The bottom table shows outcomes from the multi-task framework (our proposal).

		Predicted class		
		Malignant	Benign	Normal
Single-task	Malignant	63.4%	35.4%	1.2%
	Benign	16.7%	73.9%	9.5%
	Normal	1.6%	26.6%	71.9%
Multi-task	Malignant	76.2%	20.1%	3.7%
	Benign	9.9%	83.8%	6.3%
	Normal	7.8%	14.1%	78.1%

Table 8
Segmentation performance comparison of different multi-task methods tested on Curated BUSI dataset.

Metric	[36]	MT UNet^{++}	MT nnU-Net
DSC Benign	0.778 ± 0.059	0.773 ± 0.050	**0.779 ± 0.046**
DSC Malignant	0.716 ± 0.070	0.711 ± 0.057	**0.717 ± 0.049**
DSC Normal	0.000 ± 0.000	**0.781 ± 0.157**	0.766 ± 0.103
DSC Average	0.645 ± 0.026	0.751 ± 0.018	**0.754 ± 0.035**

Table 9
Classification performance comparison of different multi-task methods tested on Curated BUSI dataset.

Metric	[36]	MT UNet^{++}	MT nnU-Net
F1 Benign	0.783 ± 0.033	**0.826 ± 0.029**	0.806 ± 0.005
F1 Malignant	0.727 ± 0.041	**0.791 ± 0.025**	0.751 ± 0.040
F1 Normal	0.681 ± 0.015	**0.741 ± 0.060**	0.741 ± 0.109
F1 Weighted	0.748 ± 0.027	**0.801 ± 0.026**	0.777 ± 0.019
ACC	0.749 ± 0.026	**0.802 ± 0.018**	0.780 ± 0.017

study uses the cleaned version of the BUSI, we implemented the system described in their work and tested it on Curated BUSI. Tables 8 and 9 show the performance obtained for the segmentation and classification tasks, respectively, by our method using UNet^{++} and nnU-Net as backbone network as well as the proposal from [36].

The results demonstrate two key findings. First, our multi-task method obtained improvements statistically significant with respect to [36] in both tasks. In terms of segmentation performance, both MT nnU-Net and MT UNet^{++} exhibited competitive results compared to Adityan et al. However, MT nnU-Net consistently outperformed MT UNet^{++}, achieving higher DSC scores across all segmentation metrics (DSC Benign, DSC Malignant, DSC Normal, DSC Average). Notably, our analysis revealed a significant discrepancy in the segmentation performance of Adityan et al. particularly in assessing normal images, where the method yielded poor results. This underscores its limitation in accurately segmenting certain images, highlighting the necessity for more sophisticated approaches like ours. On the other hand, for classification tasks, MT UNet^{++} demonstrated superior performance over both Adityan et al. and MT nnU-Net across most categories (F1 Benign, F1 Malignant, F1 Weighted, and ACC). While MT nnU-Net showed comparable performance to MT UNet^{++} in certain classification metrics, such as F1 Normal, MT UNet^{++} consistently achieved higher scores overall.

6.4. Evaluation on other breast ultrasound dataset

In order to further evaluate our method, we chose an additional dataset called BUS-UCLM, which includes ground truth segmentations and multiclass labels. The dataset BUS-UCLM contains 683 breast ultrasound images categorized as follows: 90 malignant (13.2%), 174 benign (25.5%), and 419 normal (61.3%) images. The experiments carried out include single-task approaches, ST (Seg) for segmentation and ST (Cls) for classification, along with the multi-task approach and our proposed method.

For experimentation with BUS-UCLM, given that it contains approximately 50% more images than Curated BUSI, the number of epochs was increased by 50 up to a total of 250 epochs. Additionally, a hyperparameter search was conducted for λ, which balances the given importance to segmentation and classification within the loss function. In this case, the optimal λ value on the validation set was 0.4, compared to 0.85 obtained for Curated BUSI. This suggests that conferring more weight to the classification task helped optimize the system's performance. It is worth noting that values between 0.1 and 0.95 yielded satisfactory results regardless of the dataset. For values outside this range, however, the system became unstable due to excessive prioritization of one of the tasks within the loss function. Results are displayed in Tables 10 and 11.

As with the BUSI dataset, the MT approach outperformed most metrics achieved by ST methods on the BUS-UCLM dataset; however, our proposed method achieved superior results compared to them. Significantly, there was a significant improvement in the segmentation of benign and malignant tumors, increasing the DICE metric by up to 20 points. Moreover, our method reduced the standard deviations of numerous metrics, suggesting an increment in stability and robustness.

We carefully analyzed the confusion matrices to identify the sources of model errors. The analysis determined that our method classifies malignant, benign, and normal classes more accurately than the ST approach, showing improvements of 6.7, 8.7, and 3.1 percentage points, respectively. Beyond these gains, there was a reduction in both false positives and false negatives.

It is important to highlight how contrasting the class distributions of the evaluated datasets are. Curated BUSI contains 14.2% normal, 49.3% benign, and 36.4% malignant images, while BUS-UCLM includes 61.3% normal, 25.5% benign, and 13.2% malignant images. These differences represent a major shift in class proportions. Normal class is the minority in Curated BUSI, in contrast to their majority in BUS-UCLM. Moreover, the number of images in Curated BUSI is nearly 50% lower than in BUS-UCLM. Despite this disparity in class distributions, our method demonstrated a consistent generalization capability, maintaining high and stable performance across both datasets. This robustness is a key attribute for practical clinical applications, where CAD systems are expected to generalize effectively across various scenarios without exhibiting biases.

7. Conclusions

This research work presents an end-to-end prediction-refining multi-task framework for breast cancer segmentation and classification on ultrasound images. Our method aims to provide a robust CAD system with practical applicability in clinical settings. It has been demonstrated that leveraging the inherent correlation between both tasks enhances performance compared to traditional single-task approaches, which often have a more limited scope.

One of the remarkable strengths of this framework lies in its flexibility and modular structure, making it adaptable to diverse problems. One of the modules has been incorporated to ensure fair experimentation, addressing a gap observed in previous studies that overlooked the specific characteristics of the BUSI dataset. Other module is prepared to exploit the synergies between the tasks. In addition to the framework, the manuscript presents a comprehensive analysis of the BUSI dataset, revealing its irregularities and anomalous cases, making it an unreliable dataset for fair research. In response, we provide an alternative post-processed version of it named Curated BUSI, meticulously prepared to meet the needs of future research works within the community.

The main limitation found in this work includes the lack of consensus in the segmentation masks provided by the radiologists, resulting in a reduction in the number of cases within Curated BUSI. Additionally, our approach has the limitation that the *prediction-refining* module

Table 10
Segmentation performance comparison of the proposed prediction-refining multi-task learning framework for nnU-Net and UNet^{++} trained on BUS-UCLM. ST: Single-task. MT: Multi-task.

Metric	nnU-Net			UNet^{++}		
	ST (Seg)	MT	Our proposal	ST (Seg)	MT	Our proposal
DSC Benign	0.591 ± 0.174	0.706 ± 0.046	**0.716 ± 0.060**	0.584 ± 0.084	0.618 ± 0.027	**0.707 ± 0.020**
DSC Malignant	0.558 ± 0.211	0.718 ± 0.066	**0.750 ± 0.050**	0.525 ± 0.212	0.686 ± 0.043	**0.778 ± 0.027**
DSC Normal	0.654 ± 0.446	0.838 ± 0.046	**0.914 ± 0.041**	0.826 ± 0.084	0.890 ± 0.100	**0.924 ± 0.028**
DSC Average	0.601 ± 0.277	0.754 ± 0.053	**0.793 ± 0.050**	0.645 ± 0.127	0.732 ± 0.057	**0.803 ± 0.025**

Table 11
Classification performance comparison of the proposed prediction-refining multi-task learning framework for nnU-Net and UNet^{++} trained on BUS-UCLM. ST: Single-task. MT: Multi-task.

Metric	nnU-Net			UNet^{++}		
	ST (Cls)	MT	Our proposal	ST (Cls)	MT	Our proposal
F1 Benign	0.694 ± 0.028	0.692 ± 0.027	**0.756 ± 0.036**	0.607 ± 0.030	0.701 ± 0.034	**0.757 ± 0.038**
F1 Malignant	0.597 ± 0.020	0.613 ± 0.063	**0.718 ± 0.053**	0.508 ± 0.158	0.650 ± 0.040	**0.765 ± 0.088**
F1 Normal	0.884 ± 0.012	0.891 ± 0.015	**0.911 ± 0.018**	0.837 ± 0.060	0.896 ± 0.016	**0.919 ± 0.010**
F1 Weighted	0.798 ± 0.011	0.804 ± 0.005	**0.846 ± 0.014**	0.735 ± 0.049	0.814 ± 0.007	**0.858 ± 0.025**
ACC	0.801 ± 0.012	0.804 ± 0.005	**0.848 ± 0.013**	0.735 ± 0.056	0.814 ± 0.009	**0.858 ± 0.025**

corrects inconsistencies without flagging uncertain cases, which could reduce transparency in some cases. Implementing a warning system for these cases could improve reliability for clinicians.

To handle the limitations included in our work, future research includes considering the potential effects of previous works evaluated on the BUSI dataset instead of its curated version, given that the results may be biased to some extent. Furthermore, the proposed methodology could be extended by making the *prediction-refining* module based on probability class distributions instead or even as a learnable parameter of the architecture.

CRediT authorship contribution statement

Carlos Aumente-Maestro: Writing – original draft, Visualization, Validation, Software, Methodology, Data curation. **Jorge Díez:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Beatriz Remeseiro:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Breast Cancer Statistics, World Health Organization, 2024, https://www.who.int/news-room/fact-sheets/detail/breast-cancer, (Accessed 14 March 2024).

[2] T. Crook, R. Leonard, K. Mokbel, A. Thompson, M. Michell, R. Page, A. Vaid, R. Mehrotra, A. Ranade, S. Limaye, et al., Accurate screening for early-stage breast cancer by detection and profiling of circulating tumor cells, Cancers 14 (14) (2022) 3341.

[3] G.J. Wengert, T.H. Helbich, P. Kapetas, P.A. Baltzer, K. Pinker, Density and tailored breast cancer screening: practice and prediction–an overview, Acta Radiol. Open 7 (9) (2018) 2058460118791212.

[4] Breast Cancer Information, American Cancer Society, 2024, https://www.cancer.org, (Accessed 14 March 2024).

[5] Breast Cancer, Cancer Research UK, 2024, https://www.cancerresearchuk.org, (Accessed 14 March 2024).

[6] T. Ying, C. Ya-Ling, Y. Yu, H. Rui-Qing, Breast ultrasound image despeckling using multi-filtering DFrFT and adaptive fast BM3D, Comput. Methods Programs Biomed. 246 (2024) 108042.

[7] A.E. Ilesanmi, U. Chaumrattanakul, S.S. Makhanov, Methods for the segmentation and classification of breast ultrasound images: a review, J. Ultras. (2021) 1–16.

[8] M. Carrilero-Mardones, M. Parras-Jurado, A. Nogales, J. Pérez-Martín, F.J. Díez, Deep learning for describing breast ultrasound images with BI-RADS terms, J. Imaging Inform. Med. (2024) 1–15.

[9] Q. Huang, L. Ye, Multi-task/single-task joint learning of ultrasound BI-RADS features, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 69 (2) (2021) 691–701.

[10] S. Sun, C. Fu, S. Xu, Y. Wen, T. Ma, GLFNet: Global-local fusion network for the segmentation in ultrasound images, Comput. Biol. Med. (2024) 108103.

[11] P. Celard, E.L. Iglesias, J.M. Sorribes-Fdez, R. Romero, A.S. Vieira, L. Borrajo, A survey on deep learning applied to medical images: from simple artificial neural networks to generative models, Neural Comput. Appl. 35 (3) (2023) 2291–2323.

[12] A.K. Mishra, P. Roy, S. Bandyopadhyay, S.K. Das, A multi-task learning based approach for efficient breast cancer detection and classification, Expert Syst. 39 (9) (2022) e13047.

[13] J. Torres-Soto, E.A. Ashley, Multi-task deep learning for cardiac rhythm detection in wearable devices, NPJ Digit. Med. 3 (1) (2020) 116.

[14] W. Qi, H. Wu, S. Chan, MDF-net: A multi-scale dynamic fusion network for breast tumor segmentation of ultrasound images, IEEE Trans. Image Process. (2023).

[15] C. Xue, L. Zhu, H. Fu, X. Hu, X. Li, H. Zhang, P.-A. Heng, Global guidance network for breast lesion segmentation in ultrasound images, Med. Image Anal. 70 (2021) 101989.

[16] G. Chen, L. Zhou, J. Zhang, X. Yin, L. Cui, Y. Dai, Esknet: An enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation, Expert Syst. Appl. 246 (2024) 123265.

[17] A.S. Podda, R. Balia, S. Barra, S. Carta, G. Fenu, L. Piano, Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images, J. Comput. Sci. 63 (2022) 101816.

[18] D. Sarvamangala, R.V. Kulkarni, Convolutional neural networks in medical image understanding: a survey, Evol. Intell. 15 (1) (2022) 1–22.

[19] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[20] J. Wang, P. Lv, H. Wang, C. Shi, SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in computed tomography, Comput. Methods Programs Biomed. 208 (2021) 106268.

[21] W. Wang, D. Qin, S. Wang, Y. Fang, Y. Zheng, A multi-channel unet framework based on SNMF-DCNN for robust heart-lung-sound separation, Comput. Biol. Med. (2023) 107282.

[22] F. Bougourzi, C. Distante, F. Dornaika, A. Taleb-Ahmed, PDAtt-Unet: Pyramid dual-decoder attention unet for Covid-19 infection segmentation from CT-scans, Med. Image Anal. 86 (2023) 102797.

[23] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[25] A. Myronenko, 3D MRI brain tumor segmentation using autoencoder regularization, in: 4th International Workshop Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Held in Conjunction with MICCAI 2018, Revised Selected Papers, Part II 4, Springer, 2019, pp. 311–320.

[26] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested U-net architecture for medical image segmentation, in: 4th International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Held in Conjunction with MICCAI 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[27] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 272–284.
[28] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[29] G. Chen, Y. Dai, J. Zhang, C-Net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation, Comput. Methods Programs Biomed. 225 (2022) 107086.
[30] F. Tang, L. Wang, C. Ning, M. Xian, J. Ding, CMU-Net: A strong ConvMixer-based medical ultrasound image segmentation network, in: IEEE 20th International Symposium on Biomedical Imaging, IEEE, 2023, pp. 1–5.
[31] A.A. Alhussan, M.M. Eid, S. Towfek, D.S. Khafaga, Breast cancer classification depends on the dynamic dipper throated optimization algorithm, Biomimetics 8 (2) (2023) 163.
[32] F. Chen, J. Wang, H. Liu, W. Kong, Z. Zhao, L. Ma, H. Liao, D. Zhang, Frequency constraint-based adversarial attack on deep neural networks for medical image classification, Comput. Biol. Med. 164 (2023) 107248.
[33] O.N. Manzari, H. Ahmadabadi, H. Kashiani, S.B. Shokouhi, A. Ayatollahi, MedViT: a robust vision transformer for generalized medical image classification, Comput. Biol. Med. 157 (2023) 106791.
[34] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, D. Shen, Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images, Med. Image Anal. 70 (2021) 101918.
[35] C. Zhu, X. Chai, Z. Wang, Y. Xiao, R. Zhang, Z. Yang, J. Feng, DBL-Net: A dual-branch learning network with information from spatial and frequency domains for tumor segmentation and classification in breast ultrasound image, Biomed. Signal Process. Control 93 (2024) 106221.
[36] M.L. Adityan, H. Sharma, A. Paul, Segmentation and classification-based diagnosis of tumors from breast ultrasound images using multibranch unet, in: IEEE International Conference on Image Processing, IEEE, 2023, pp. 2505–2509.
[37] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Artificial Intelligence and Statistics, PMLR, 2015, pp. 562–570.
[38] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Fourth International Conference on 3D Vision, IEEE, 2016, pp. 565–571.
[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[40] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020) 104863.
[41] X. Wen, X. Guo, S. Wang, Z. Lu, Y. Zhang, Breast cancer diagnosis: A systematic review, Biocybern. Biomed. Eng. 44 (1) (2024) 119–148.
[42] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[43] H.W. Lilliefors, On the Kolmogorov–Smirnov test for normality with mean and variance unknown, J. Am. Stat. Assoc. 62 (318) (1967) 399–402.
[44] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. (1947) 50–60.
[45] A.A. Taha, A. Hanbury, Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool, BMC Med. Imaging 15 (2015) 1–28.
[46] C. Aumente-Maestro, D. Rodríguez González, D. Martinez, B. Remeseiro, BTS U-Net: A Data-Driven Approach to Brain Tumor Segmentation Through Deep Learning, Available at SSRN 4327638.
[47] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 1055–1059.