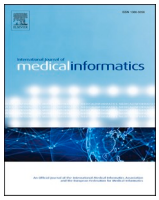
International Journal of Medical Informatics 189 (2024) 105522

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Segmentation-based BI-RADS ensemble classification of breast tumours in ultrasound images

Maciej Bobowicz [a,1,*], Mikołaj Badocha [a,1], Katarzyna Gwozdziewicz [a], Marlena Rygusik [a], Paulina Kalinowska [b], Edyta Szurowska [a], Tomasz Dziubich [c]

[a] 2nd Department of Radiology, Medical University of Gdansk, 17 Smoluchowskiego Str., Gdansk 80-214, Poland
[b] Department of Thoracic Radiology, Karolinska University Hospital, Anna Steckséns g 41, Solna 17176, Sweden
[c] Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, 11/12 G. Narutowicza Str., Gdańsk 80-233, Poland

ARTICLE INFO

Keywords:
Breast cancer
Ultrasound
Segmentation
Image classification
BI-RADS
Ensemble approach

ABSTRACT

Background: The development of computer-aided diagnosis systems in breast cancer imaging is exponential. Since 2016, 81 papers have described the automated segmentation of breast lesions in ultrasound images using artificial intelligence. However, only two papers have dealt with complex BI-RADS classifications.
Purpose: This study addresses the automatic classification of breast lesions into binary classes (benign vs. malignant) and multiple BI-RADS classes based on a single ultrasonographic image. Achieving this task should reduce the subjectivity of an individual operator's assessment.
Materials and Methods: Automatic image segmentation methods (PraNet, CaraNet and FCBFormer) adapted to the specific segmentation task were investigated using the U-Net model as a reference. A new classification method was developed using an ensemble of selected segmentation approaches. All experiments were performed on publicly available BUS B, OASBUD, BUSI and private datasets.
Results: FCBFormer achieved the best outcomes for the segmentation task with intersection over union metric values of 0.81, 0.80 and 0.73 and Dice values of 0.89, 0.87 and 0.82, respectively, for the BUS B, BUSI and OASBUD datasets. Through a series of experiments, we determined that adding an extra 30-pixel margin to the segmentation mask counteracts the potential errors introduced by the segmentation algorithm. An assembly of the full image classifier, bounding box classifier and masked image classifier was the most accurate for binary classification and had the best accuracy (ACC; 0.908), F1 (0.846) and area under the receiver operating characteristics curve (AUROC; 0.871) in the BUS B and ACC (0.982), F1 (0.984) and AUROC (0.998) in the UCC BUS datasets, outperforming each classifier used separately. It was also the most effective for BI-RADS classification, with ACC of 0.953, F1 of 0.920 and AUROC of 0.986 in UCC BUS. Hard voting was the most effective method for dichotomous classification. For the multi-class BI-RADS classification, the soft voting approach was employed.
Conclusions: The proposed new classification approach with an ensemble of segmentation and classification approaches proved more accurate than most published results for binary and multi-class BI-RADS classifications.

1. Introduction

Breast cancer is the most common female cancer, accounting for around 2.2 million new cases and more than 0.6 million cancer-related deaths around the globe in 2020 [1]. A breast ultrasound scan (USS) detects and characterises breast lesions and assesses locoregional lymph nodes. USS assessment requires extensive training and certification because it is highly influenced by the ultrasonographer's experience and the equipment used. Lesion characterisation is standardised by the Breast Imaging Reporting & Data System (BI-RADS) atlas developed by the American College of Radiology (ACR) [2]. The physician assigns one of nine possible classes based on the specific breast lesion's features, including the shape, orientation, margin, echo pattern, posterior features, presence of calcifications and associated features. Based on the BI-

* Corresponding author.
E-mail addresses: maciej.bobowicz@gumed.edu.pl (M. Bobowicz), mikolaj.badocha@gumed.edu.pl (M. Badocha), katarzyna.gwozdziewicz@gumed.edu.pl (K. Gwozdziewicz), marlena.rygusik@gumed.edu.pl (M. Rygusik), paulina.kalinowska@regionstockholm.se (P. Kalinowska), edyta.szurowska@gumed.edu.pl (E. Szurowska), tomasz.dziubich@eti.pg.edu.pl (T. Dziubich).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijmedinf.2024.105522
Received 13 March 2024; Received in revised form 19 May 2024; Accepted 5 June 2024
Available online 6 June 2024

RADS class and associated malignancy probability, oncologists decide whether an invasive breast biopsy is required. Pathological assessment of biopsy specimens is the only examination that provides indisputable information on tumour characteristics. Proper USS assessment and the assignment of BI-RADS classes are essential in guiding clinical decisions.

Numerous approaches have been proposed for detecting and classifying breast lesions using computer assistance. These can be classified into classic computer vision–based and machine learning (ML)–based algorithms. The former class can be categorised into region-based methods [3,4], edge-based methods [5,6], thresholding [7] and energy function–based methods [8]. Their main limitations are low solution generality and relatively long computation times.

The second group includes unsupervised methods that can effectively handle data that are not clearly separated [9] and supervised methods, including neural networks (ANN), which rely heavily on the quality and diversity of labelled training data. It remains crucial to address bias, overfitting and generalisation challenges to ensure the models' reliable and ethical deployment [10].

Comparing the methods using the area under the receiver operating characteristics (AUROC) curve metric, the ML approaches achieve a value of approximately 0.97, texture feature–based methods achieve 0.89 and morphological features achieve 0.93 [11]. Some authors have reviewed various 2D and 3D semantic segmentation strategies using deep convolutional neural networks (CNNs), demonstrating outstanding outcomes for biomedical image analysis [12–14]. In [15], the authors provided a compilation of the performance of 11 models, with the average dice similarity coefficient (DSC) ranging from 0.61 to 0.89, depending on the dataset.

In this study, we compare automatic methods for breast lesion segmentation in ultrasound images, PraNet [16], CaraNet [17] and FCBFormer [18], with the U-Net model [19] used as a reference. All models were adapted to the specific segmentation task. A new automatic classification method for breast lesions based on a single ultrasonographic breast image is developed and described based on an ensemble of selected segmentation approaches. This method is tested on various datasets for binary (benign vs. malignant) and multi-class BI-RADS classifications. Achieving this task should reduce the subjectivity of individual operator assessments, consequently minimising inter-reader variability, and benefit both patients and doctors.

2. Literature review

A literature review was conducted in June 2023 using the IEEE Xplore database with the keywords 'segmentation', 'breast' and 'cancer'. All publications from 2016 were included, resulting in a total of 633 articles, of which 81 were USS-related.

Most studies (62/81) used ML, while 35 used CNNs. CNNs are shown to be effective in processing image input through convolutional layers. Different CNN architectures were used, such as the U-Net model [20–22], which is popular due to its extendable architecture that allows problem adaptation. In [20], the authors extended the U-Net network with the fusion and multi-scale dilated convolution modules to segment irregular and large breast lesions. They proposed class imbalance minimisation with a focal-DSC loss function. Other modifications of CNN architecture include recursive methods that allow variable-length inputs, such as the Faster R-CNN [23] and the attention module [24].

Studying lesions involves two main tasks: segmentation and classification. Fifty-three papers addressed both tasks, while 26 focused solely on segmentation. Although segmentation provides valuable information about lesion shape, only a few studies tackled BI-RADS multi-class assignments due to their challenging nature and lack of appropriate datasets.

Only two studies dealt with BI-RADS classification. In [25], the authors used a private dataset consisting of 641 images (413 benign cases and 228 malignant). They trained a binary classifier on the CNN-based architecture, which consisted of four convolutional layers and two fully connected layers with a single neuron in the last layer, returning the answer's certainty (benign or malignant). Then, fixed threshold values were assigned to the six BI-RADS classes using doctors' knowledge and a trial-and-error method. The final classification was performed by comparing the activation/certainty of the last neuron from the binary classification with the predetermined threshold values for each class and determining the interval into which the value fell. This two-stage framework attained an accuracy of 0.998 for Category 3 BI-RADS, with the lowest accuracy of 0.734 for Category 4B.

In contrast, the authors of [26] attempted to train a classifier mimicking the human decision-making process. They employed four dCNNs, utilising a sliding window approach. The initial model aimed to classify metadata from any tissue type, while the second focused on distinguishing normal healthy tissue from regions containing irregularities or lesions. The third model aimed to differentiate between cysts and soft tissue lesions, and the fourth was tasked with discerning lesions warranting follow-up from those necessitating biopsy. Unfortunately, the authors did not provide details about the proposed architecture. The training utilised 1019 images from 582 patients, with testing conducted on 144 images. The classification accuracy for differentiating BI-RADS 2 from BI-RADS 3–5 lesions was 87.1 %, and it was 93.1 % for BI-RADS 2–3 versus BI-RADS 4–5.

Both recent cases had the disadvantage of analysing entire images rather than focusing on the areas of the lesions.

3. Materials and methods

3.1. Proposed solution

Our method relies on deep neural networks to identify the important features of objects in images. Instead of looking at entire images indiscriminately, this approach focuses on specific regions (tumours) deemed informative by the neural network. It further refines these regions to make them more distinct and useful for distinguishing between different objects. This process is likely to involve some form of image processing or feature extraction to enhance the discriminative power of these regions.

To achieve this, the prediction is made as a result of the decisions made by three CNN classifiers, each operating on a different input image (see Fig. 1). The first classifier, the full image classifier (FIC), takes an entire original image as its input. Next, the bounding box classifier (BBC) uses a cropped area containing a lesion as an input. The size of this area is based on segmentation (the segmenter is described later). It corresponds to a bounding box surrounding the lesion with a certain fixed margin around it. The last classifier, the masked image classifier (MIC), takes a modified semantic mask generated by a segmenter as input. The mask is negatively thresholded, removing all values below certain thresholds. It leaves the centre of the lesion unmasked while not changing the remaining part. This approach was chosen because the irregular shape of lesion edges is a clinical discriminatory feature. This study explored ensemble methods to enhance classification performance, including hard voting for malignancy classification and averaging predictions for BI-RADS classification. Hard voting ensured robust decisions in binary malignancy classification, while averaging predictions (soft voting) with the highest probability emerged as the best method for multi-class BI-RADS classification, capturing the collective knowledge of the ensemble and ensuring reliable classification outcomes.

3.2. Segmentation models

Four segmentation approaches were utilised, including the baseline U-Net [19]. The remaining three models, which were first investigated for segmenting gastrointestinal polyps, incorporated attention modules.

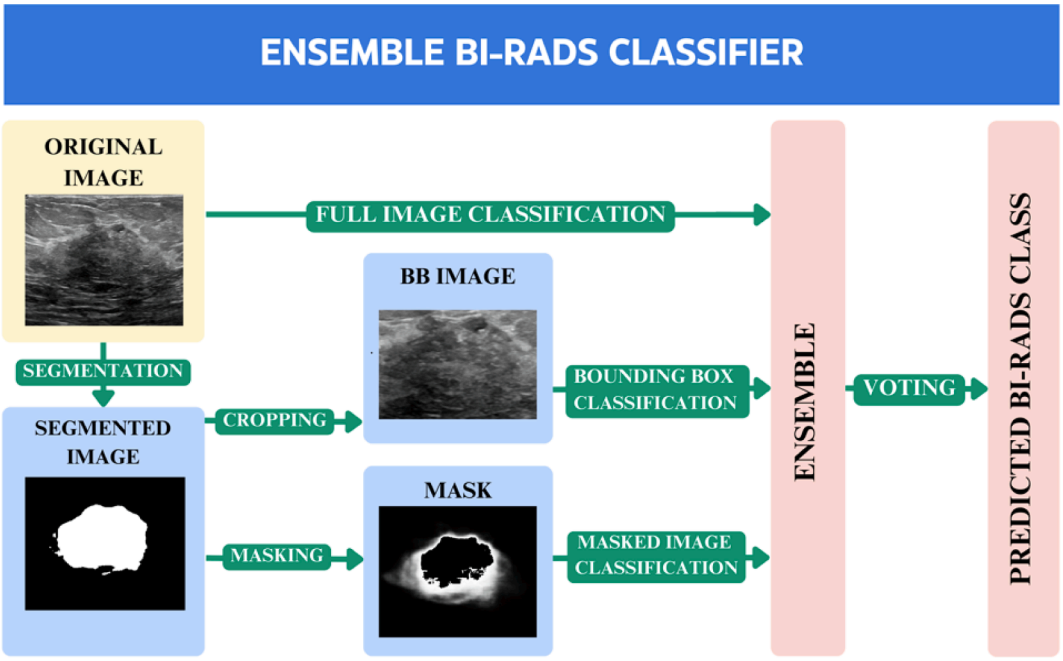


Fig. 1. Flowchart for our proposed method: The process begins with an original ultrasound image. Using a segmentation model, we derive a segmented image. This image undergoes two transformations: 1) cropping with a 30-px margin to create a bounding box image that primarily contains the tumour and 2) applying a negative mask where all values above 0.95 are set to 0, highlighting the lesion's border. These processed and original images are inputs for the three convolutional neural network classifiers. The outputs are then combined through ensemble methods for final decision-making.

3.2.1. PraNet

Parallel reverse attention network (PraNet) [16] is an architecture developed to segment polyps in colonoscopy images. These images share characteristics with ultrasonographic images of breast lesions, such as varying sizes, textures and indistinct boundaries with surrounding tissues. The first step in the model involves aggregating image features at high-level layers through a parallel partial decoder. This generates a global map that is subsequently processed by successive components. Through reverse attention, masks of increasingly high resolution are generated, thereby improving segmentation accuracy with each step.

3.2.2. CaraNet

The context axial reverse attention network (CaraNet) model [17] is similar to PraNet, with the addition of contextual feature pyramid (CFP) blocks before the reverse attention modules. These blocks provide separate dilation rate values for each channel. In this study, the reverse attention modules were replaced with axial reverse attention modules. The input to these modules comes from the CFP blocks and consists of feature maps with general information about mask localisation. They mimic self-attention mechanisms, which aim to assign appropriate weights to each value. However, this can be computationally demanding, especially when dealing with large two-dimensional maps that require weight assignment. The authors replaced this with two one-dimensional vectors along both axes. This reduces the number of weights for a 256 × 256-pixel input to just 512.

3.2.3. Fcbformer

The FCN-Transformer [18], or FCBFormer, architecture was designed to analyse colon polyp images and combines fully convolutional networks (FCNs) and transformers. The FCN component is excellent at extracting spatial features and capturing detailed information. The transformer component then processes these spatial features, which uses its self-attention mechanism to consider the relationships between different image regions, incorporating the global context. This fusion of local and global understanding allows the FCBFormer to segment the USS accurately. It can be trained end to end, improving its efficiency and effectiveness.

3.3. Classifiers

3.3.1. Masked image classifier

The masked image classifier (MIC) considers clinically significant features, such as lesions' shape, border, symmetry, surrounding area continuity and clarity as well as the presence of 'spiculations'. It concentrates only on the surroundings of the main part of the breast lesion and not on its centre. It uses the information from the whole 'areola' of the lesion with all clinically significant features without distinguishing or separate analysis of each feature, as proposed by some authors [27]. We prepared the input by generating a semantic mask and prioritising the lesion's shape and boundaries. We modified the mask by first normalising the pixel values in the mask to a range of 0 to 1. We then set the pixel values in the centre of the lesion to zero by changing all values higher than 0.95 to zero. This step ensured that the central region would be masked out while maintaining the normalised pixel values within the desired range.

3.3.2. Bounding box classifier

The BBC extracts information from a lesion's surroundings using the bounding box obtained from the segmentation process as input. Through a series of experiments, we determined that adding an extra 30-pixel margin to the segmentation would counteract the potential errors introduced by the segmentation algorithm.

3.4. Ensemble methods

The following ensemble approaches have been explored to enhance classification performance:

- Majority voting (hard majority) [28]: This is an algorithm in which each classifier makes a prediction, and the ensemble's prediction is a majority vote of each individual classifier.

- Soft voting [29]: This algorithm works by initially having each classifier assign a probability to each class. The ensemble's prediction is the class with the highest total probability.
- Max: Similar to soft voting, each classifier assigns a probability to each class. The ensemble's prediction is a class with the singular highest probability across all classifiers.
- Dense: Here, a small neural network is added, connecting all three classifiers. The network has one linear layer with eight neurons. The classifiers are first trained separately and then connected together and frozen to train ensemble classifiers.

Majority voting was the most effective method for dichotomous classification. With only two available classes (benign and malignant) and three classifiers, it ensured that one label always received two or three votes. However, majority voting with a secondary conflict resolution method could not yield satisfactory results in the multi-class BI-RADS classification. Instead, soft voting was utilised.

3.5. Materials

3.5.1. Dataset description

The essential characteristics of the publicly available (BUSI [30], BUS B [31] and OASBUD [32]) and private (UCC BUS) datasets used in the study are presented in Table 1. More information can be sought in the original publications [30–32]. The UCC BUS dataset was curated by four researchers. All images were obtained with a single ultrasonography scanner by a single doctor with more than 10 years' experience in breast ultrasound. Every image contained only a single breast lesion. Two doctors, one with five years and the other with more than 10 years of experience in breast ultrasonography, scored the BI-RADS classifications. When there was no consensus, a third opinion was sought from a breast radiologist with more than 15 years of experience. Images with divergent scoring were independently assessed, and the final score was based on a consensus discussion. All lesions were biopsied and confirmed using pathology assessment. Ground truth labels were provided at the image level. The pre-processing of DICOM files involved anonymisation with the CTP anonymiser. Then, the actual image was cut out using an automatic Python script to remove all associated information. This image, with pseudo-ID, was then associated with a file containing the BI-RADS category and pathology results. The major characteristics of the dataset are included in Table 1.

Table 1
Breast ultrasound scan database characteristics.

Dataset with reference	BUS B [31]	OASBUD [32]	BUSI [30]	UCC BUS (private)
Year of data collection	2012	2017	2018	2022
Source	Diagnostic Centre of Parc Tauli, Sabadell, Spain	Oncology Institute, Warsaw, Poland	Baheya Hospital, Cairo, Egypt	University Clinical Centre, Gdansk, Poland
Ultrasound device	Siemens ACUSON Sequoia C512	Ultrasonix SonixTouch research device	LOGIQ E9	LOGIQ E9
Resolution (pixels)	Variable; average 760–570	685 × 868	500 × 500	782 × 782
File format	PNG	RF ultrasound echoes, mat lab format	PNG	PNG
Age range	26–78	N/A	25–75	16–88
Average age	N/A	N/A	N/A	46
Age variability (SD)	N/A	N/A	N/A	14.7
Ethnic group	N/A	European	N/A	European
No. of women	163	78	600	610
No. of images	163	200	780	6774*
No. of malignant images	53	104	210	3867*
No. of benign images	110	96	437	2907*
No. of cases without lesions	0	0	133	0
Benign vs. malignant classification	YES	YES	YES	YES
BI-RADS classification	NO	YES	NO	YES
Segmentation mask	YES	YES	YES	NO

Note: *In the UCC dataset, multiple cross-sectional images were obtained from different parts of the tumour in different axes of the lesion.

All datasets were employed separately for binary classification to ensure the reliability and robustness of the results. Only images with lesions were analysed. Only the UCC and OASBUD datasets were used for multi-class BI-RADS classification, as the other datasets lacked this variable. By leveraging these diverse datasets with varying imaging conditions, lesion types and patient populations, the study could comprehensively evaluate the proposed methodologies, enhancing the credibility and generalisability of the findings. All data came from retrospective studies.

3.5.2. Evaluation metrics

Following [33], we used the Dice similarity coefficient (DSC; Eq. (1) and the intersection over union (IoU; Eq. (2), or the Jaccard index, as the most suitable metrics for the segmentation problem.

$$DSC = F1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

where TP is true positive, FP is false positive, and FN is false negative.

DSC better handles unbalanced classes. When one class exhibits a substantially smaller pixel count, the Jaccard metric may yield a near-zero value, failing to reflect the extent of class overlap. We opted for a threshold of 0.7 since it consistently delivered the best results in our experiments. Similarly, for classification problem accuracy, recall precision and F1 score are defined in Eqs. (3), 4, 5 and 1.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

AUROC and average precision (AP), which summarises a precision-recall curve, were other metrics used. In the multi-class BI-RADS classification, accuracy was counted by summing the statistics over all labels, and in all other classifications, it was counted by calculating the statistics for each label and averaging them. Dataset splitting was performed before the pre-processing steps. All models were validated with internal random five-fold cross-validation. The next step in the proposed solution is external validation.

3.5.3. System and classification specifications

The classification specifications were as follows:

- Model: EfficientNet v2 L (118.5 M parameters, 56.08 GFLOPS)
- Loss function: cross-entropy
- Optimiser: Adam

- Batch size: 36
- Learning rate: 0.0003
- Input image size: 224 × 224 px
- Augmentations: horizontal flip, vertical flip, 180° rotation
- Normalisation: mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]

Transfer learning was utilised with default weights from the TorchVision library. A weighted loss function was tested because of an imbalance in the number of benign and malignant lesions. Experiments showed that this had no effect on the results. Hyperparameters were chosen by a grid search. The code used for the experiments can be found at https://github.com/Dumbldore/BI-RADS_classification. CNN training/validation and testing were done using a computer system with the following characteristics:

- Intel(R) Core (TM) i7-12700 K CPU @ 3.60 GHz
- 32 GB RAM (2 × 16 GB) DDR4 @ 3600 MHz RAM memory
- RTX 3090 24 GB GPU

3.5.4. Ethics statement

The authors declare that the presented work was carried out in accordance with the Code of Ethics of the World Medical Association for experiments involving humans. The necessary consent of the Bioethics Committee for Scientific Research at the Medical University of Gdansk was obtained. All images were fully anonymised to ensure the patients' privacy.

4. Results and discussion

4.1. Image segmentation results

Table 2. shows the performance metrics for the different segmenters. DSC and IoU metrics are presented for three datasets used (BUS B, BUSI and OASBUD). Most results come from the original ESTAN model publication [15], with additional results from UNet, PraNet, CaraNet and FCBFormer calculated during the project. A comparison with the SK-UNet model [34] was added for the two datasets. The results are based on five-fold cross-validation, ensuring reproducibility.

In each of the three described datasets, the FCBFormer model outperformed the other methods. In BUS B, the model achieved an IoU value of 0.02 higher than PraNet and CaraNet and 0.07 higher than ESTAN. This was similar for the DSC metric (0.02 and 0.07, respectively).

A graphical comparison of the segmentation results between the four models, with the ground truth superimposed over the masks (UNet as the baseline), is shown in Fig. 2.

Table 2
Segmentation results for different architectures. Modified from B. Shareef et al. (2022) [15], with additional results for UNet, PraNet, CaraNet and FCBFormer architectures.

Dataset	Architecture	IoU	DSC
BUS Dataset B	AlexNet	0.47	0.61
	SegNet	0.60	0.71
	UNet	0.65	0.75
	CE-Net	0.61	0.71
	MultiResUNet	0.66	0.75
	RDAU-NET	0.67	0.77
	SCAN	0.65	0.74
	DenseU-Net	0.60	0.69
	STAN	0.70	0.78
	ESTAN	0.74	0.82
	PraNet	0.79	0.87
	CaraNet	0.79	0.86
	FCBFormer	**0.81**	**0.89**
BUSI	AlexNet	0.55	0.68
	SegNet	0.62	0.72
	UNet	0.63	0.73
	CE-Net	0.64	0.73
	MultiResUNet	0.67	0.75
	RDAU-NET	0.68	0.73
	SCAN	0.63	0.72
	DenseU-Net	0.64	0.72
	STAN	0.66	0.75
	ESTAN	0.70	0.78
	SK-Unet	ND	0.70
	PraNet	0.74	0.82
	CaraNet	0.73	0.81
	FCBFormer	**0.80**	**0.87**
OASBUD	SK-Unet	ND	0.72
	PraNet	0.64	0.74
	CaraNet	0.65	0.76
	FCBFormer	**0.73**	**0.82**

4.2. Image classification results

4.2.1. Ablation study

An ablation study was performed to test the different ensemble methods on cancer versus no cancer and BI-RADS classes. The results are summarised in Table 3.

4.2.2. Cancer vs. No cancer classification

The binary classification with FIC yielded the best results among single methods (except for the BUS dataset), yet it was only slightly better than the BBC (see Table 4), with an ensemble of classifiers being superior to any single classifier used. The aspect of margin selection in the BBC was also examined (F1 score averaged between datasets: 0 px: 0.8353, 30 px: 0.884, and 60 px: 0.8385), with the conclusion that strict cropping of the segmented area leads to the loss of certain discriminative features located at the lesion's boundaries. In contrast, the broader margin provided unnecessary input information, thereby reducing effectiveness. The achieved results were still lower than those obtained for the entire image. Both individual methods significantly outperformed the segmentation-based model.

The application of an ensemble significantly improved the classification quality. Using the majority voting approach resulted in a 1.2 % increase in classification accuracy averaged between datasets. As the problem involved binary classification and the number of models was odd, there were no ties in the voting.

4.2.3. BI-RADS classification

During the BI-RADS classification, the classifiers predicted classes labelled from 2 to 5 on the UCC dataset and from 3 to 5 on the OASBUD dataset. Classes 0 and 1 were not included because of insufficient data (see Table 5).

The soft voting approach was employed because the majority voting method was unfeasible. The outcomes of all networks were aggregated on a per-class basis, and the final predicted class was based on the highest value in the list of aggregated scores.

A noteworthy increase in the weighted F1 score was observed, amounting to 3.3 % for the OASBUD and 1.6 % for the UCC datasets, compared to the best results achieved by individual models. This improvement in classification quality is comparable to the findings presented in Section 4.2.1 and may indicate the proposed method's generalisability.

The ACC reached values similar to FIC for both datasets. Aggregating both metrics indicated better model performance in underrepresented classes.

5. Conclusion

The proposed new classification approach, based on an ensemble of selected segmentation and classification approaches, proved more

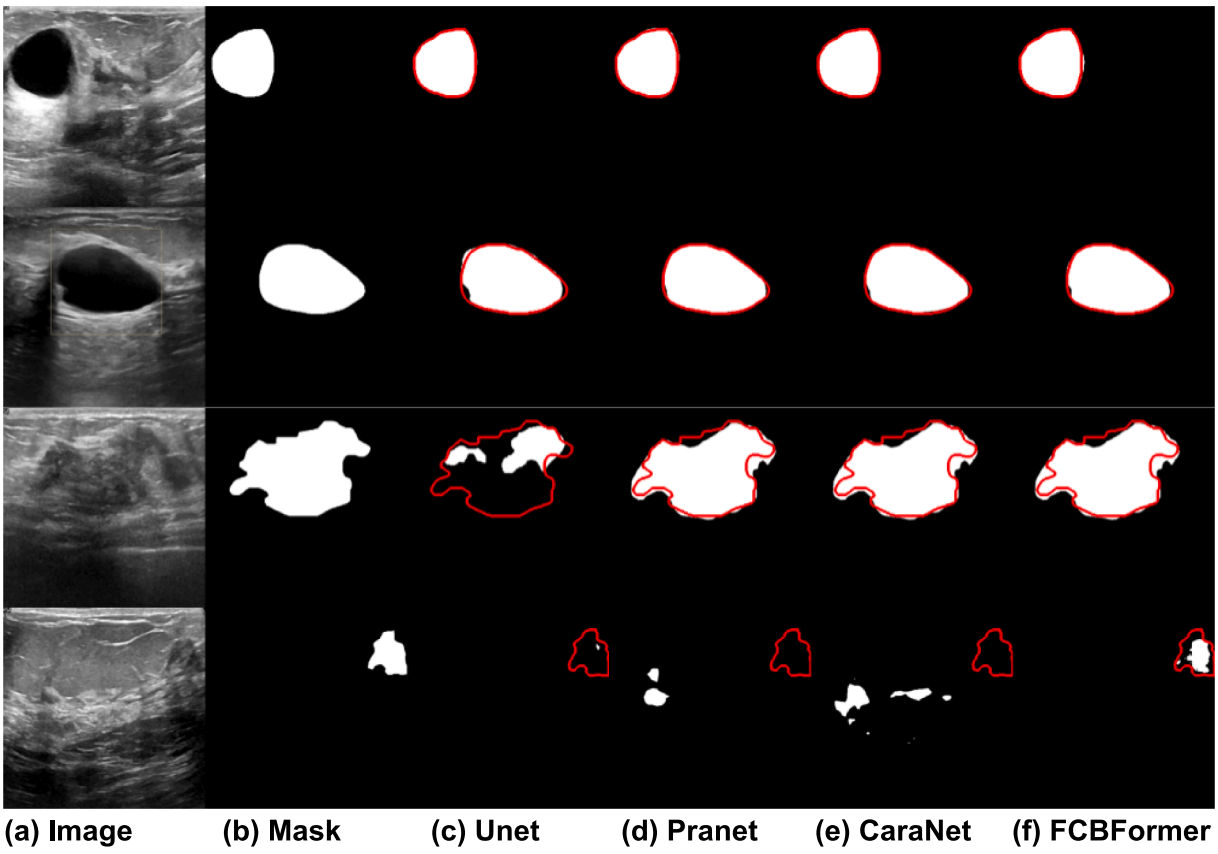


Fig. 2. Comparison of the exemplary segmentation outcomes among the different models.

Table 3
Ablation study conducted on the combined datasets.

Used classifiers	Ensemble method	ACC	F1
Cancer vs. no cancer classification			
FIC	–	0.888	0.871
BBC	–	0.870	0.848
MIC	–	0.827	0.794
FIC + MIC	Max	0.898	0.880
FIC + BBC	Max	0.907	0.890
FIC + BBC + MIC	Max	0.907	0.890
FIC + MIC	Soft majority	0.8695	0.877
FIC + BBC	Soft majority	0.904	0.887
FIC + BBC + MIC	Soft majority	0.910	0.895
FIC + BBC + MIC	Dense	0.901	0.887
FIC + BBC + MIC	Majority voting	0.919	0.907
BI-RADS classification			
FIC	–	0.504	0.431
BBC	–	0.516	0.380
MIC	–	0.446	0.356
FIC + BBC + MIC	Max	0.537	0.393
FIC + BBC + MIC	Soft majority	0.553	0.455
FIC + BBC + MIC	Dense	0.528	0.441
FIC + BBC + MIC	Majority voting + max	0.520	0.370
FIC + BBC + MIC	Majority voting + soft majority	0.537	0.435

Note: a) BUSI, BUS B, OASBUD and a subset of UCC for cancer vs. no cancer classification; b) OASBUD and a subset of UCC for BI-RADS classification. The results indicate that hard voting is most effective for binary classification, while soft voting excels in multi-class scenarios.

accurate than most published results for binary and multi-class BI-RADS classifications. The MIC was designed to improve the evaluation of breast lesions by enhancing the focus on their shape and boundaries, prioritising these critical features by modifying the semantic mask to exclude the tumour's centre and normalising the pixel values. However, considering its sensitivity to segmenter errors, it should be utilised as part of a more comprehensive approach, as proposed in the paper. Borders of the breast lesions add crucial information, although too-large margins introduce unnecessary noise and irrelevant information to the input of the BBC. By focusing on the informative region within the bounding box, the BBC can effectively capture and analyse the key features associated with the lesion. This classifier plays a crucial role in our overall methodology, providing valuable insights into breast lesion classifications.

Further research that eliminates our study limitations, involving larger multi-institutional datasets containing BIRADS 0 and 1 scans with external validation and assessments of added benefit, is needed to evaluate the proposed methodology before it can be used in clinical practice. The proposed methodology is currently being utilised as a method for active learning during the annotation of ground truth datasets to expand the UCC dataset.

CRediT authorship contribution statement

Maciej Bobowicz: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mikołaj Badocha:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Katarzyna Gwozdziewicz:** Writing – review & editing, Formal analysis, Data curation. **Marlena Rygusik:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Paulina Kalinowska:** Writing – review & editing, Formal analysis, Data curation. **Edyta Szurowska:** Writing – review & editing, Supervision, Methodology. **Tomasz Dziubich:** Writing – review & editing, Writing –

Table 4
Cancer vs. no cancer classification results of five-fold cross-validation.

Dataset	Architecture	ACC	F1	Recall	Precision	AUROC	AP
BUS Dataset B	FIC	0.895	0.834	**0.771**	0.919	**0.910**	**0.901**
	BBC	0.907	0.836	0.720	**1.0**	0.892	0.886
	MIC	0.877	0.801	0.761	0.850	0.870	0.852
	Ensemble	**0.908**	**0.846**	0.762	0.953	0.871	0.806
BUSI	FIC	0.935	**0.897**	**0.877**	0.919	**0.953**	0.927
	BBC	0.925	0.884	0.875	0.897	0.951	**0.937**
	MIC	0.874	0.800	0.782	0.822	0.899	0.832
	Ensemble	**0.937**	0.886	0.851	**0.927**	0.933	0.933
OASBUD	FIC	0.864	0.854	0.831	0.885	**0.906**	**0.922**
	BBC	0.859	0.850	0.803	0.910	0.0894	0.897
	MIC	0.780	0.786	0.804	0.775	0.804	0.781
	Ensemble	**0.899**	**0.882**	**0.851**	**0.933**	0.895	0.866
UCC	FIC	0.971	0.974	0.964	0.985	0.994	0.995
	BBC	0.961	0.966	0.965	0.967	0.985	0.987
	MIC	0.922	0.935	0.974	0.899	0.972	0.973
	Ensemble	**0.982**	**0.984**	**0.988**	**0.980**	**0.998**	**0.999**

Table 5
BI-RADS classification results of five-fold cross-validation.

Dataset	Architecture	ACC	F1	Recall	Precision	AUROC	AP
OASBUD	FIC	0.585	0.600	0.585	**0.686**	0.796	0.661
	BBC	0.637	0.619	0.637	0.637	0.812	0.694
	MIC	0.522	0.497	0.522	0.527	0.717	0.500
	Ensemble	**0.669**	**0.652**	**0.669**	0.655	**0.920**	**0.748**
UCC	FIC	0.909	0.904	0.909	0.903	0.985	0.948
	BBC	0.889	0.829	0.798	0.879	0.978	0.898
	MIC	0.849	0.785	0.761	0.820	0.957	0.837
	Ensemble	**0.953**	**0.920**	**0.925**	**0.916**	**0.986**	0.923

original draft, Supervision, Software, Methodology, Investigation, Formal analysis, 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.

Acknowledgements

No acknowledgements.

Funding

This research received no specific grant funding from agencies in the public, commercial or not-for-profit sectors.

References

[1] Globocan 2020: New global cancer data, https://www.uicc.org/ news/globocan-2020-new-global-cancer-data, accessed: March 16, 2023.
[2] L. Levy, M. Suissa, J.F. Chiche, G. Teman, B. Martin, BIRADS ultrasonography, Eur. J. Radiol. 61 (2) (2007) 202–211, https://doi.org/10.1016/j.ejrad.2006.08.035.
[3] J. Shan, H. Cheng, Y. Wang, Completely automated segmentation approach for breast ultrasound images using multiple-domain features, Ultrasound Med. Biol. 38 (2) (2012) 262–275, https://doi.org/10.1016/j.ultrasmedbio.2011.10.022. PMID: 22230134.
[4] X. Jiang, Y. Guo, H. Chen, Y. Zhang, Y. Lu, An adaptive region growing based on neutrosophicset in ultrasound domain for image segmentation, IEEE Access 7 (2019) 60584–60593, https://doi.org/10.1109/ACCESS. 2019.2911560.
[5] M.I. Daoud, A.A. Atallah, F. Awwad, M. Al-Najar, Accurate and fully automatic segmentation of breast ultrasound images by combining image boundary and region information, IEEE, 2016, pp. 718–721, https://doi.org/10.1109/ISBI.2016.7493367.
[6] S.R. Rao, S.E. Shelton, P.A. Dayton, The "fingerprint" of cancer extends beyond solid tumor boundaries: assessment with a novel ultra sound imaging approach, IEEE Trans. Biomed. Eng. 63 (5) (2016) 1082–1086, https://doi.org/10.1109/TBME.2015.2479590.
[7] X. Li, C. Yang, S. Wu, Automatic segmentation algorithm of breast ultrasound image based on improved level set algorithm, IEEE, 2016, pp. 319–322, https://doi.org/10.1109/SIPROCESS.2016.7888276.
[8] B. Li, S.T. Acton, Active contour external force using vector field con volution for image segmentation, IEEE Trans. Image Process. 16 (8) (2007) 2096–2106, https://doi.org/10.1109/TIP.2007.899601.
[9] E. Samundeeswari, P. Saranya, R. Manavalan. Segmentation of breast ultrasound image using regularized k-means (rekm) clustering, in: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 1379– 1383. doi: 10.1109/ WiSPNET.2016.7566362.
[10] T. Prabhakar, S. Poonguzhali. Automatic detection and classification of benign and malignant lesions in breast ultrasound images using texture morphological and fractal features, in: 2017 10th Biomedical Engineering International Conference (BMEiCON), 2017, pp. 1–5. doi: 10.1109/ BMEiCON.2017.8229114.
[11] I. Bakkouri, K. Afdel, Breast tumor classification based on deep convolutional neural networks, in: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2017, pp. 1– 6. doi:10.1109/ATSIP. 2017.8075562.
[12] W. Gómez-Flores, W. Coelho de Albuquerque Pereira, A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound, Comput. Biol. Med. 126 (2020) 104036, https://doi.org/10.1016/j.compbiomed.2020.104036.
[13] I. Enitan, U. Chaumrattanakul, S. Makhanov, Methods for the segmentation and classification of breast ultrasound images: a review, J. Ultrasound 24 (2021), https://doi.org/10.1007/s40477-020-00557-5.
[14] M.A. Al-antari, M.A. Al-masni, M.-T. Choi, S.-M. Han, T.-S. Kim, A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inf. 117 (2018) 44–54, https://doi.org/10.1016/j.ijmedinf.2018.06.003.
[15] B. Shareef, A. Vakanski, M. Xian, P.E. Freer, Estan: Enhanced small tumor-aware network for breast ultrasound image segmentation, Healthcare (Basel) 10 (11) (2022) 2262, https://doi.org/10.3390/healthcare10112262.

[16] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao. Pranet: Parallel reverse attention network for polyp segmentation (2020). doi: 10.48550/ARXIV.2006.11392.
[17] A. Lou, S. Guan, H. Ko, M. Loew, CaraNet: CaraNet: Context axial reverse attention network for segmentation of small medical objects, J. Med. Imaging 10 (1) (2023) 014005, https://doi.org/10.1117/1.jmi.10.1.014005.
[18] E. Sanderson, B.-J. Matuszewski. FCN-Transformer feature fusion for polyp segmentation, in: Yang, G., Aviles-Rivero, A., Roberts, M., Schönlieb, CB. (eds) Medical Image Understanding and Analysis. MIUA 2022. Lecture Notes in Computer Science, vol 13413. Springer, Cham. doi: 10.1007/978-3-031-12053-4_65.
[19] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. doi: 10.1007/978-3-319-24574-4_28.
[20] J. Li, L. Cheng, T. Xia, H. Ni, J. Li, Multi-scale fusion u-net for the segmentation of breast lesions, IEEE Access 9 (2021) 137125–137139, https://doi.org/10.1109/ACCESS.2021.3117578.
[21] R. Almajalid, J. Shan, Y. Du, M. Zhang, Development of a deep-learning-based method for breast ultrasound image segmentation, IEEE, 2018, pp. 1103–1108, https://doi.org/10.1109/ICMLA.2018.00179.
[22] S. Hussain, X. Xi, I. Ullah, Y. Wu, C. Ren, Z. Lianzheng, C. Tian, Y. Yin, Contextual level-set method for breast tumor segmentation, IEEE Access 8 (2020) 189343–189353, https://doi.org/10.1109/ACCESS. 2020.3029684.
[23] S.Y. Shin, S. Lee, I.D. Yun, S.M. Kim, K.M. Lee, Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images, IEEE Trans. Med. Imaging 38 (3) (2019) 762–774, https://doi.org/10.1109/TMI. 2018.2872031.
[24] M. Xu, K. Huang, X. Qi, Multi-task learning with context-oriented self-attention for breast ultrasound image classification and segmentation, IEEE, 2022, pp. 1–5, https://doi.org/10.1109/ISBI52829.2022.9761685.
[25] B. Zeimarani, M.G.F. Costa, N.Z. Nurani, S.R. Bianco, W.C. De Albuquerque Pereira, C.F.F.C. Filho, Breast lesion classification in ultrasound images using deep convolutional neural network, IEEE Access 8 (2020) 133349–133359, https://doi.org/10.1109/ACCESS.2020.3010863.
[26] A. Ciritsis, C. Rossi, M. Eberhard, M. Marcon, A.S. Becker, A. Boss, Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making, Eur. Radiol. 29 (10) (2019) 5458–5468, https://doi.org/10.1007/s00330-019-06118-7.
[27] C. Thomas, M. Byra, R. Marti, M.H. Yap, R. Zwiggelaar, BUS-Set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets, Med. Phys. 50 (5) (2023) 3223–3243, https://doi.org/10.1002/mp.16287.
[28] O. O. Awe, G. O. Opateye, C. A. G. Johnson, O. T. Tayo, R. Dias. Weighted hard and soft voting ensemble machine learning classifiers: Application to anaemia diagnosis, in: Awe, O. O., Vance, E. A. (eds) Sustainable statistical and data science methods and practices. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health. Springer, Cham. (2023). doi: 10.1007/978-3-031-41352-0_18.
[29] S. Karlos, G. Kostopoulos, S. Kotsiantis, A soft-voting ensemble-based co-training scheme using static selection for binary classification problems, Algorithms 13 (1) (2020) 26, https://doi.org/10.3390/a13010026.
[30] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultra sound images, Data Brief 28 (2019) 104863, https://doi.org/10.1016/j.dib.2019.104863.
[31] M.H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. Davison, R. Marti, Automated breast ultrasound lesions detection using convolutional neural networks, IEEE J. Biomed. Health Inform. 22 (4) (2018) 1218–1226, https://doi.org/10.1109/JBHI.2017.2731873.
[32] H. Piotrzkowska-Wroblewska, K. Dobruch-Sobczak, M. Byra, A. Nowicki, Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions, Med. Phys. 44 (2017), https://doi.org/10.1002/mp.12538.
[33] A.A. Taha, A. Hanbury, Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool, BMC Med. Imaging 15 (1) (2015) 29, https://doi.org/10.1186/s12880-015-0068-x.
[34] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda Fournier, L. Olson, M. O'Boyle, C. Comstock, M. Andre, Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network, Biomed. Signal Process. Control 62 (2020) 1–10, https://doi.org/10.1016/j.bspc.2020.102027.