Deep learning-based classification of breast tumors using selected subregions of lesions in sonograms

Christian Schmidt
Westfälische Hochschule
University of Applied Sciences
Neidenburger Strasse 43
45897 Gelsenkirchen
Germany
christian.schmidt@w-hs.de

Heinrich Martin Overhoff
Westfälische Hochschule
University of Applied Sciences
Neidenburger Strasse 43
45897 Gelsenkirchen
Germany
heinrich-martin.overhoff@w-hs.de

ABSTRACT

Breast cancer, a prevalent disease among women, demands early detection for better clinical outcomes. While mammography is widely used for breast cancer screening, its limitation in e.g., dense breast tissue necessitates additional diagnostic tools. Ultrasound breast imaging provides valuable tumor information (features) which are used for standardized reporting, aiding in the screening process and precise biopsy targeting. Previous studies have demonstrated that the classification of regions of interest (ROIs), including only the lesion, outperforms whole image classification. Therefore, our objective is to identify essential lesion features within such ROIs, which are sufficient for accurate tumor classification, enhancing the robustness of diagnostic image acquisition. For our experiments, we employ convolutional neural networks (CNNs) to first segment suspicious lesions' ROIs. In a second step, we generate different ROI subregions: top/bottom half, horizontal subslices and ROIs with cropped-out center areas. Subsequently these ROI subregions are classified into benign vs. malignant lesions with a second CNN. Our results indicate that outermost ROI subslices perform better than inner ones, likely due to increased contour visibility. Removing the inner 66% of the ROI did not significantly impact classification outcomes ($p = 0.35$). Classifying half ROIs did not negatively impact accuracy compared to whole ROIs, with bottom ROI performing slightly better than top ROI, despite significantly lower image contrast in that region. Therefore, even visually less favorable images can be reliably analyzed when the lesion's contour is depicted. In conclusion, our study underscores the importance of understanding tumor features in ultrasound imaging, supporting enhanced diagnostic approaches to improve breast cancer detection and management.

Keywords
breast tumor, classification, CNN, ultrasound, tumor subregions

1 INTRODUCTION

According to the WHO, breast cancer stands as the most prevalent cancer among women worldwide [Who24], highlighting the need for effective screening and diagnostic methods. As early detection is crucial for achieving favorable patient outcomes, methods like mammography are widely used for breast cancer screening. The integration of artificial intelligence assistance for mammography screening has shown promising results [Lan23], demonstrating its potential to enhance detection rates. However, the complexity of breast tissue composition and the challenges posed by dense breast tissue in particular underscore the necessity for additional diagnostic tools and modalities. In comparison to alternative imaging modalities, ultrasound breast imaging is a non-ionizing, cost-effective, highly mobile, real-time imaging modality, making it widely accessible in most healthcare settings around the world. Ultrasound imaging excels in differentiating between different types of suspicious masses (e.g., cystic and solid lesions), and provides information about lesion shape, size, internal appearance, and other characteristics, which are essential for uniform reporting [Men13]. By enabling precise targeting of suspicious masses, ultrasound also aids in guiding biopsy procedures.

Our objective is to enhance the diagnostic accuracy of breast tumor ultrasound imaging by improving the understanding of lesion characteristics. This will facilitate the development of better-targeted, more effective diagnostic approaches, leading to improved cancer detection and patient care. In previous work [Sch23] convolutional neural network (CNN) based classification of suspicious masses into benign vs. malignant lesions was performed in whole breast sonograms vs. regions of interest (ROIs). In that work, the ROI is defined

https://www.doi.org/10.24132/CSRN.3401.26

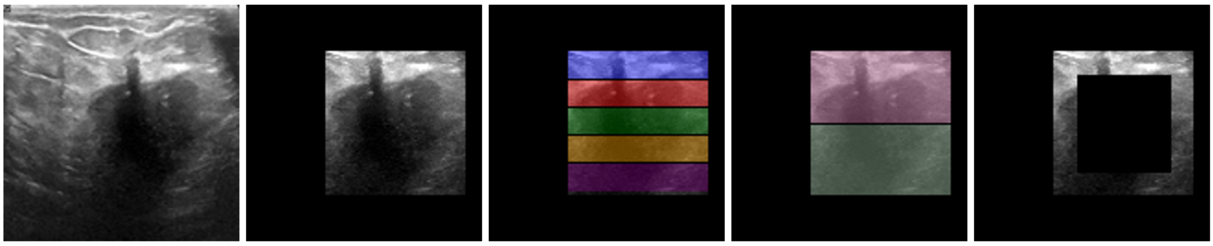

Figure 1: Overview of ROI subregions. From left to right: unprocessed ultrasound image (US), region of interest crop around the lesion mass ($\text{ROI}_{\text{whole}}$), subslices of ROI ($\text{ROI}_{1-5}$), ROI_{top} and $\text{ROI}_{\text{bottom}}$, ROI with cropped-out center area ($\text{ROI}_{\text{crop x\%}}$, here: $\text{ROI}_{\text{crop66}}$)

as a rectangular subregion around the suspicious lesion. Classification accuracy was considerably higher for ROI vs. whole sonogram (0.89 vs. 0.83). We hypothesize, that subregions of the lesion ROI yield similar classification accuracies to full ROIs. If so, a coarse subregion is sufficient for classification and the necessity of precise tumor segmentation and precise ROI positioning respectively diminishes. Our approach is to create different subregions of lesion ROIs (top/bottom half, horizontal subslices and ROIs with cropped-out center areas) and use a CNN to classify these subregions into benign and malignant lesions. By performing a comprehensive analysis of the classification results, we then identify which subregions contribute most to accurate lesion classification.

2 RELATED WORK

In the realm of breast tumor classification, researchers have explored various methodologies, each offering distinct advantages and insights into the diagnostic process. One prevalent approach relies heavily on deep learning techniques, particularly convolutional neural networks (CNNs), to analyze ultrasound tumor images. These CNN-based methods extract convolutional features through transfer learning, leveraging pre-trained architectures such as VGG16, YOLOv3, or GoogLeNet. Notable studies [Ald19, Chi19, Han17, Kal21] have demonstrated the efficacy of this approach in accurately classifying breast tumors based solely on learned features. However, deep learning methods come with certain trade-offs. While they offer automatic feature learning and high flexibility, requiring minimal manual intervention, they often demand large amounts of labeled data for training and significant computational resources. Additionally, deep learning models can be complex and challenging to interpret, potentially limiting their applicability in clinical settings where user scepticism towards AI is still widely present [Tam22, Che22] and interpretability of the employed algorithms is crucial for their trustworthiness.

In contrast, another avenue of investigation involves preselecting informative features from ultrasound images and subsequently employing classical machine learning algorithms for classification. Methods such as support vector machines and decision trees have been utilized in this context [Zha21, Cha14, Muh22]. These preselected features encompass a range of characteristics, including texture-based attributes derived from image data (e.g., gray-level co-occurrence matrix, contrast, homogeneity, energy) envelope and spectral-based properties extracted from raw radiofrequency data (e.g., K-distribution, Nakagami distribution). By focusing on specific features relevant to tumor characterization, these approaches offer a complementary perspective to the purely deep learning-based methodologies. Nevertheless, preselected feature classification methods also have their limitations. While they may offer greater interpretability and computational efficiency, they often rely on manual feature engineering, which can be labor-intensive and may not capture all relevant information in the data. Additionally, these methods may struggle to capture complex patterns and relationships in the data, potentially limiting their performance compared to deep learning approaches, especially in scenarios with large and heterogeneous datasets.

Recent investigations on a hybrid approach, such as [Dao20, Abh23, Saj23], suggest that integrating hand-crafted features with deep learning-based convolutional features can lead to further improvements in classification accuracy. This capitalizes on the strengths of both methodologies, potentially enhancing the robustness and reliability of breast tumor classification systems.

Moreover, the discussion between whole image classification and subregion-based analysis has garnered some attention within the research community. Further studies [Cha14, Sch23] have underscored the superiority of the analysis of selected tumor subregions over features derived from entire images. This finding aligns with the notion that localized analysis can provide more precise insights into tumor characteristics and facilitate more accurate classification outcomes.

3 MATERIALS AND METHODS

3.1 Dataset

In this study we employed the BUSI [Ald20] (breast ultrasound images) dataset for our analysis. This dataset comprises of 780 breast sonograms (437 benign masses, 210 malignant masses, 133 normal) of women aged between 25-75 years. For each sonogram in the dataset, ground truth lesion segmentations are available, enabling precise localization of the lesions. To ensure comparability with other state-of-the-art classification research, we focused exclusively on benign and malignant masses, excluding normal images, with no masses, from our consideration. Before training our models, all images were uniformly resized to dimensions of 128 × 128 pixels, and the grayscale values were normalized to span the range [0, 1]. This preprocessing step standardizes the image dimensions and intensity values, facilitating consistent model training and evaluation across the dataset. The root mean square (RMS) contrast, which is given by

$$\text{RMS Contrast} = \sqrt{\frac{1}{N}\sum_{x,y}(I(x,y)-\bar{I})^2}, \qquad (1)$$

where N is the total number of pixels in the image, $I(x,y)$ is the intensity value of a pixel at position (x,y), and $\bar{I}$ denotes the mean intensity value of all image pixels, serves as an indicator of image quality. Notably, the top half of the ROIs (ROI_{top}) exhibit an RMS contrast of 0.22 ± 0.12, while the bottom half ($\text{ROI}_{\text{bottom}}$) show an RMS contrast of 0.14 ± 0.07, with a statistically significant difference ($p < 0.0001$) based on the unpaired t-test.

3.2 ROI subregions

Because ultrasound B-mode imaging analyzes reflected waves, which disperse while passing through different tissues, image contrast decreases in deeper layers of tissue, as shown above. These B-mode images are also prone to artifacts, especially along the direction of the wavefront. This geometric condition is reflected in the implementation and choice of the image versions. We created the following four versions of ROI subregions (Fig. 1) to evaluate which lesion areas are most important for classification:

- $\text{ROI}_{\text{whole}}$: For the entire ROI ($\text{ROI}_{\text{whole}}$), a contour segmentation model [Sch23] is applied on the original unprocessed BUSI ultrasound dataset. A tight rectangular area is cropped around the resulting segmentation to obtain the desired ROI. This image version ensures that the entire lesion area, along with some surrounding tissue, is included for comprehensive analysis, capturing the full extent of tumor features.
- ROI_{1-5}: To evaluate the effects of artifacts and better understand the distribution of pertinent information among $\text{ROI}_{\text{whole}}$ images, we divide it into five, evenly spaced, horizontal subslices ROI_{1-5} (indexing applies top down). This segmentation strategy allows us to explore variations in lesion characteristics across different depths within the tissue, providing insights into the spatial distribution of features relevant to classification.
- ROI_{top} and $\text{ROI}_{\text{bottom}}$: To evaluate the effects of potentially worse image quality (low contrast) in the lower lesion area, we horizontally divide $\text{ROI}_{\text{whole}}$ into two halves ROI_{top} and $\text{ROI}_{\text{bottom}}$. By separately analyzing the upper and lower halves of the lesion, we can assess whether image quality variations (contrast) impact classification performance differently across different regions of the ROI.
- $\text{ROI}_{\text{crop x\%}}$: To gain better insights into the importance of the outer, contour regions vs. the inside of the lesions, we crop out a rectangle with dimensions of $x\%$ of the original $\text{ROI}_{\text{whole}}$ dimensions, to create $\text{ROI}_{\text{crop x\%}}$. This approach allows us to systematically evaluate the significance of the lesion's outer boundary and surrounding tissue in classification, providing valuable information on the spatial localization of discriminative features.

3.3 Classification experiments

To obtain the desired lesion ROI, we applied a *segmentation* model detailed in [Sch23]. It consists of seven convolutional layers (number of filters: 32, 64, 128, 256, 128, 64, 32; kernel size 3×3), which each perform feature extraction by applying learnable filters to the input images, followed by a max-pooling and a dropout layer to enhance model generalization. This segmentation process delineates the lesion region of interest from the surrounding tissue, enabling focused analysis and classification of suspicious masses. The network architecture (Table 1) we used for our *classification* experiments is also a CNN with sequential layers. The CNN architecture incorporates fully connected dense layers, which are widely employed in classification tasks. These dense layers process the flattened output from the convolutional layers to generate meaningful class predictions. The final output layer consists of two neurons for binary classification, providing predicted probabilities for each class. These probabilities serve as classifier parameters for our receiver operating

Layer (Type)	Output Shape	Channels
Conv	(64, 64, 4)	4
Conv_1	(32, 32, 8)	8
Conv_2	(16, 16, 16)	16
Conv_3	(8, 8, 32)	32
Flatten	(None, 2048)	-
Conv_4	(None, 32)	32
Dense	(None, 2)	2

Table 1: Summary of the network structure for our classification models.

characteristic (ROC) curve evaluation in the results section, enabling comprehensive performance assessment. During training, the classification model was initialized randomly and trained from scratch for 100 epochs using the binary cross-entropy loss function, the Adam optimizer and a learning rate of $\eta = 1 \cdot 10^{-4}$. The dataset was split into training, test, and validation data at a $60:30:10$ ratio for all experiments. Additionally, data augmentation in the form of horizontal flip and scaling was applied before training to improve model generalization. To account for potential variability in the dataset and improve the reliability of performance estimates, we employed a 10-fold cross-validation for each classification experiment. This technique involves partitioning the dataset into ten subsets, training the model on nine subsets, and evaluating its performance on the remaining subset. By repeating this process with different subsets for evaluation, we obtain more stable and representative performance estimates for our classification models. The area under the curve (AUC) metric was employed to compare classification accuracy across all models, providing a quantitative measure of model performance and efficacy in distinguishing between benign and malignant lesions.

We trained three distinct models for our evaluation: The first model was exclusively trained on $\mathrm{ROI_{whole}}$ data (whole net, WN). The WN is used to evaluate the subslice experiment ($\mathrm{ROI_{1-5}}$) as well as the crop experiment ($\mathrm{ROI_{crop\,x\%}}$). In addition to the WN, we trained two additional models: the "top net" (TN) and the "bottom net" (BN). These models were specifically trained on $\mathrm{ROI_{top}}$ and $\mathrm{ROI_{bottom}}$, respectively. The TN and BN models were designed to explore the effects of contrast variations across different regions of the lesion and to investigate whether robust classification models can be effectively trained using only half of the ROI information. By training separate models on the upper and lower halves of the lesion ROI, we aimed to discern any disparities in classification performance and ascertain the significance of ROI composition in model training and evaluation. In the following results section, the significance threshold was set at .05.

4 RESULTS

We found that our baseline model ($\mathrm{AUC}(\mathrm{ROI_{whole}}) = 0.922 \pm 0.032$) performed in line with recent, comparable studies [Ghe22, Byr21], which also used the BUSI dataset. These studies applied vastly more complex vision transformers as well as transfer learning-based, large convolutional models to address this task and to achieve similar classification accuracies to our baseline. This suggests that while more complex models may offer marginal improvements, our approach maintains competitiveness within the current state-of-the-art. Results of the subslice experiment (Fig. 2) showed that subslices $\mathrm{ROI_1}$ and $\mathrm{ROI_5}$ performed better than inner ones (Fig. 3). We observed that the model classified $\mathrm{ROI_5}$ significantly more accurately than $\mathrm{ROI_3}$ ($\mathrm{AUC}(\mathrm{ROI_5}) = 0.824 \pm 0.071$ vs. $\mathrm{AUC}(\mathrm{ROI_3}) = 0.760 \pm 0.052, p = 0.036$), while $\mathrm{ROI_1}$ vs. $\mathrm{ROI_3}$ did not reach statistical significance ($p = 0.074$).

We found that cropping out the inner 66% of the ROI (Fig. 4) did not lead to a significant deterioration in AUC ($\mathrm{AUC}(\mathrm{ROI_{whole}}) = 0.922 \pm 0.032$ vs. $\mathrm{AUC}(\mathrm{ROI_{crop66}}) = 0.897 \pm 0.058$, $p = 0.35$). AUC began to significantly increase with $\mathrm{ROI_{whole}}$ vs. $\mathrm{ROI_{crop75}}$ ($p = 0.016$).

Additionally, $\mathrm{BN\_ROI_{whole}}$ performed significantly better than $\mathrm{BN\_ROI_{bottom}}$ ($p = 0.045$), despite the upper half not being known during the training process (Fig. 5). The BN demonstrated superior generalization to $\mathrm{ROI_{whole}}$ compared to the TN ($\mathrm{AUC}(\mathrm{BN\_ROI_{whole}}) = 0.922 \pm 0.032$ vs. $\mathrm{AUC}(\mathrm{TN\_ROI_{whole}}) = 0.876 \pm 0.051$, $p = 0.029$).

5 DISCUSSION

We hypothesized that ROI subsections would achieve classification results comparable to using the entire ROI. Our hypothesis was supported by the results, as $\mathrm{ROI_{top}}$ and $\mathrm{ROI_{bottom}}$ indeed demonstrated classification performance similar to that of $\mathrm{ROI_{whole}}$. However, for smaller subslices, this assumption was proven false. We observed a significant deterioration in classification accuracy in every subslice $\mathrm{ROI_{1-5}}$ compared to $\mathrm{ROI_{top}}$, $\mathrm{ROI_{bottom}}$ and $\mathrm{ROI_{whole}}$. The observed decline in performance is likely attributed to an excessive reduction in contextual information. In addition, we observed that outer subslices tend to perform better, likely due to the preservation of more visible contour information. We further conducted verification tests by removing rectangular areas inside and outside the lesion, confirming that the inner lesion area is not critical for classification with our CNN model.

Additionally, we investigated whether analyzing only the upper or lower half of the lesion ROI would yield

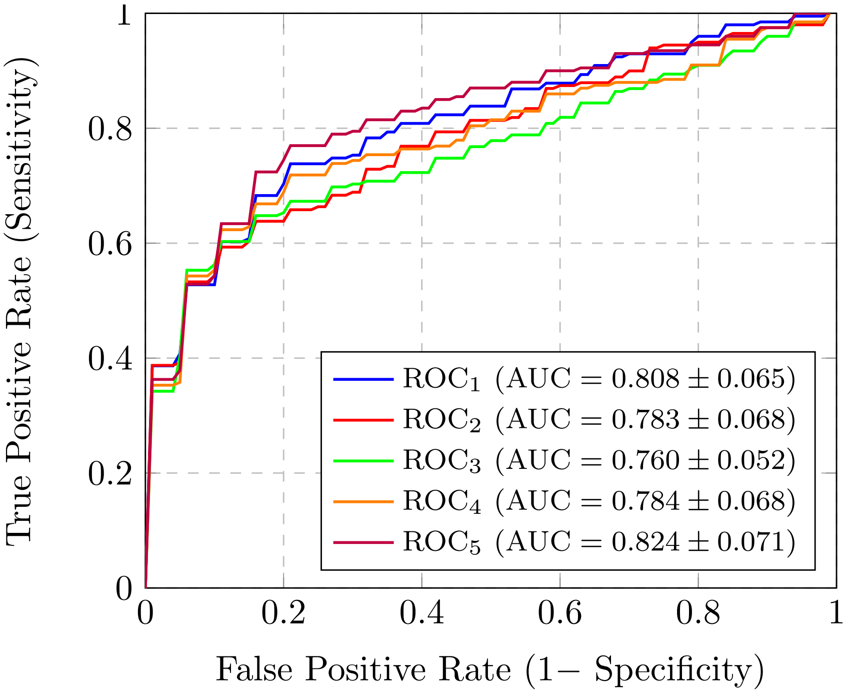

Figure 2: Results of subslice classification experiment.

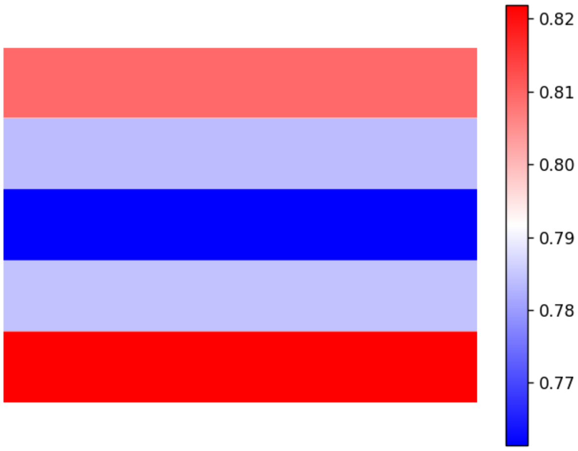

Figure 3: Color-coded results (AUC) of subslice classification experiment, mapped onto the ROI$_{\text{whole}}$ image version.

significantly different classification accuracies. Results showed, that even visually less favorable images can be reliably analyzed if the upper or lower edge of the lesion is depicted. Surprisingly, our experimental findings demonstrated that the variation in image contrast, particularly the substantially lower contrast observed in ROI$_{\text{bottom}}$, did not adversely impact the classification accuracy of our proposed method. Therefore, successful classification is possible even for images or image subregions of low quality, highlighting the robustness of our approach across varying imaging conditions.

To further strengthen the validity and generalizability of our findings, future research should include experiments conducted on other datasets beyond the BUSI dataset. While our results demonstrate promising classification performance using ROI subsections, it is essential to validate these findings on diverse datasets with varying characteristics such as imaging protocols, patient demographics, and lesion types. This broader exploration will help assess the external validity of our approach and its applicability across different clinical settings.

Additionally, conducting further experiments to compare our results with "heatmaps" generated by classification networks could provide valuable insights into the discriminative features utilized by our CNN model. Heatmaps visualize regions of interest within the images that contribute most to the classification decision, offering a deeper understanding of the underlying mechanisms driving our model's performance. By comparing the performance of our ROI subsections with the spatial distribution of discriminative features identified by heatmaps, we can gain further insights into the robustness and interpretability of our classification approach. These experiments would contribute to the ongoing efforts to enhance the transparency and interpretability of deep learning models in medical image analysis.

6 REFERENCES

[Abh23] Abhisheka B, Biswas S, Purkayastha B, Das S. (2023). Integrating Deep and Handcrafted Features for Enhanced Decision-Making Assistance in BreastCancer Diagnosis on Ultrasound Images. 10.21203/rs.3.rs-3276190/v1.

[Ald19] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Deep Learning Approaches for Data Augmentation and Classification of Breast Masses using Ultrasound Images. International Journal of Advanced Computer Science and Applications 10 (2019).

[Ald20] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data in Brief. 2020;28:104863.

[Byr21] Byra M. Breast mass classification with transfer learning based on scaling of deep representations. Biomedical Signal Processing and Control. 2021;69:102828.

[Cha14] Chaudhury B et al., Using features from tumor subregions of breast DCE-MRI for estrogen receptor status prediction, 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 2014, pp. 2624-2629, doi: 10.1109/SMC.2014.6974323.

[Che22] Chen M, Zhang B et al., Acceptance of clinical artificial intelligence among physicians and medical students: A systematic review with cross-sectional survey, Frontiers in Medicine 2022 Volume 9, doi: 10.3389/fmed.2022.990604

[Chi19] Chiao JY, Chen KY, Liao K, Hsieh I, Zhang G, Huang TC. Detection and classification the breast tumors using mask R-CNN on sonograms. Medicine 98 (2019), e15200.

[Dao20] Daoud, M.I.; Abdel-Rahman, S.; Bdair, T.M.; Al-Najar, M.S.; Al-Hawari, F.H.; Alazrai, R. Breast Tumor Classification in Ultrasound Images Using Combined Deep and Handcrafted Features. Sensors 2020, 20, 6838.

[Ghe22] Gheflati B, Rivaz H. Vision Transformer for Classification of Breast Ultrasound Images. 2022.

[Han17] Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine and Biology 62 (2017).

[Kal21] Kalafi E, Jodeiri A, Setarehdan K, Ng W, Rahmat K, Mohd Taib NA et al. Classification of Breast Cancer Lesions in Ultrasound Images by Using Attention Layer and Loss Ensemble in Deep Convolutional Neural Networks. Diagnostics 11 (2021), p. 1859.

[Lan23] LÃ¥ng K, Josefsson V, Larsson AM, Larsson S. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, singleblinded, screening accuracy study. The Lancet Oncology. 2023;24(8):936-44.

[Men13] Mendelson E, BÃ¶hm-VÃ©lez M. ACR BI-RADS Ultrasound. Reston, VA, 2013.

[Muh22] Muhtadi S. Breast Tumor Classification Using Intratumoral Quantitative Ultrasound Descriptors. Comput Math Methods Med. 2022 Mar 7;2022:1633858. doi: 10.1155/2022/1633858. PMID: 35295204; PMCID: PMC8920646.

[Saj23] Sajid U et al., Breast cancer classification using deep learned features boosted with handcrafted features. Biomedical Signal Processing and Control 2023 Volume 86, Part C, doi: 10.1016/j.bspc.2023.105353 Breast cancer classification using deep learned features boosted with handcrafted features

[Sch23] Schmidt C, Overhoff HM. Applicability of BI-RADS Criteria for Deep Learning-based Classification of Suspicious Masses in Sonograms. Bildverarbeitung für die Medizin 2023. Ed. by Deserno TM, Handels H, Maier A. Wiesbaden: Springer Fachmedien Wiesbaden, 2023:108-13.

[Tam22] Tamori H, Yamashina H, Mukai M, Morii Y, Suzuki T, Ogasawara K. Acceptance of the Use of Artificial Intelligence in Medicine Among Japan's Doctors and the Public: A Questionnaire Survey. JMIR Hum Factors. 2022 Mar 16;9(1):e24680. doi: 10.2196/24680.

[Who24] Fact sheet: Breast cancer, World Health Organization. https://www.who.int/newsroom/fact-sheets/detail/breast-cancer/. Accessed: 2023-07-31.

[Zha21] Zhang B, Song L, Yin J. Texture Analysis of DCE-MRI Intratumoral Subregions to Identify Benign and Malignant Breast Tumors. Front Oncol. 2021 Jul 8;11:688182. doi: 10.3389/fonc.2021.688182.

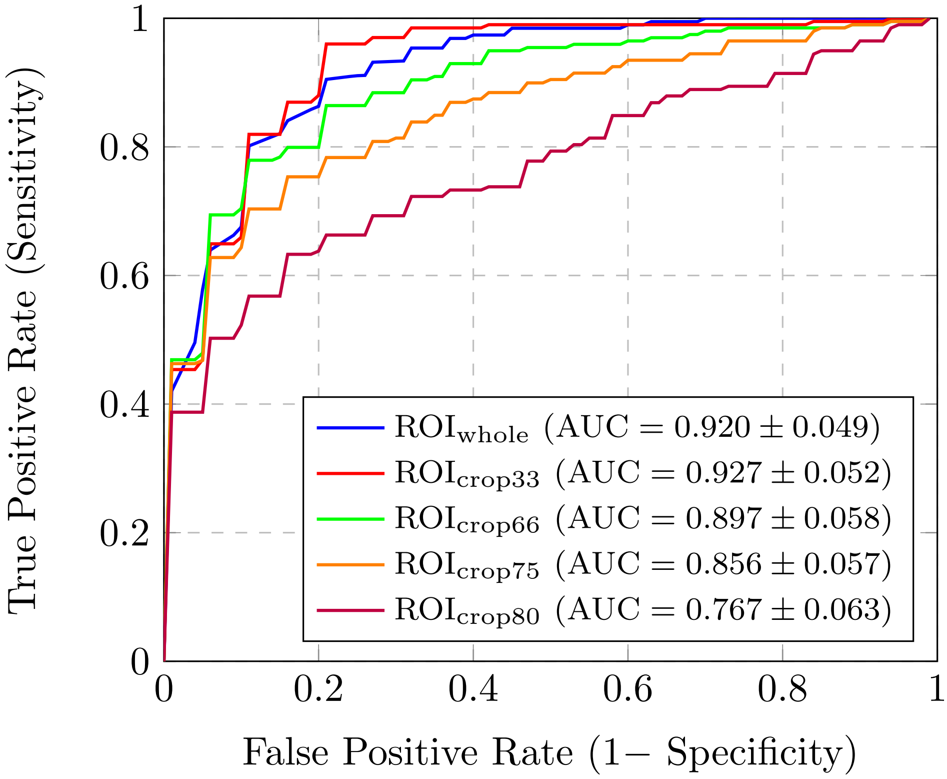

Figure 4: Results of crop classification experiment.

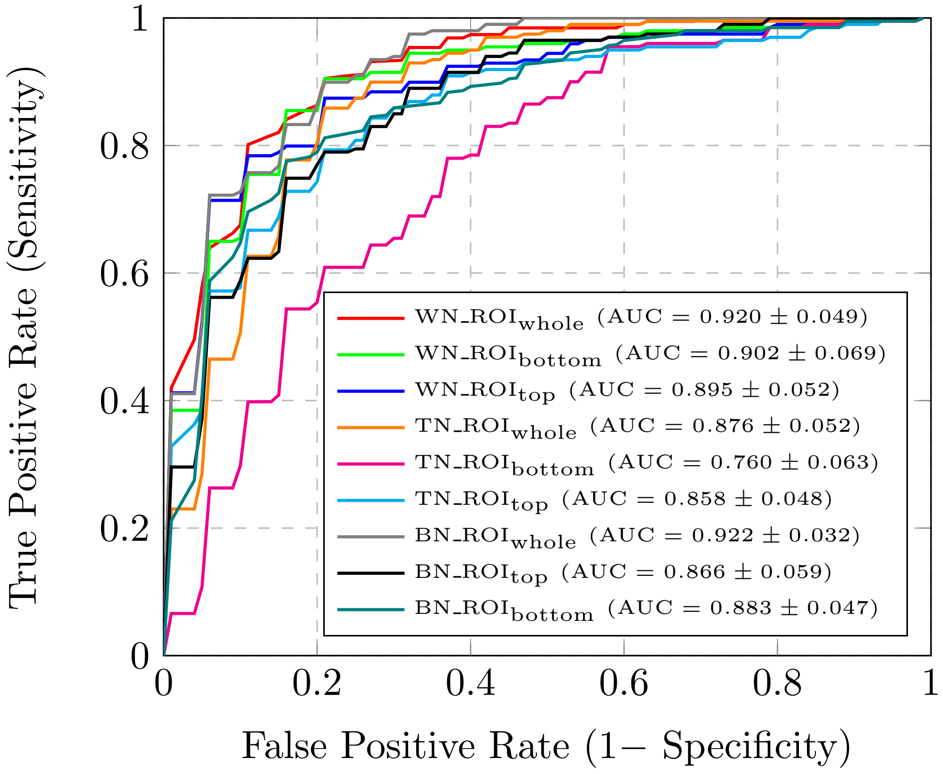

Figure 5: Results of top/bottom halves experiment.