Impact of preprocessing strategies on segmentation of fibroglandular tissue on breast MRI using deep learning

Grzegorz Rydzyński(✉)[1], Ameya Markale[1,2], Shirin Heidarikahkesh[2], Chris Ehring[2], Lorenz Döppmann[2], Dominique Hadler[2], Lorenz A. Kapsner[2], Sabine Ohlmeyer[2], Michael Uder[2], Łukasz Jeleń[3], Tatyana Ivanovska[4], Sebastian Bickelhaupt[2], and Andrzej Liebert[1,2]

[1] Institute of Computer Science Polish Academy of Science Jana Kazimierza 5, 01-248 , Warszawa, Poland
[2] Institute of Radiology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) Maximiliansplatz 3, 91054 Erlangen, Germany
[3] Department of Computer Engineering, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław
[4] Department of Electrical Engineering, Media and Computer Science, Ostbayerische Technische Hochschule Amberg-Weiden, Kaiser-Wilhelm-Ring 23, 92224 Amberg, Germany

✉ grzegorz.rydzynski@ipipan.waw.pl

Abstract. Deep learning algorithms for fibro-glandular tissue (FGT) segmentation in breast MRI achieve good performance, however results can vary depending on the training data and preprocessing strategies. This study assesses the generalizability of deep learning architectures in FGT segmentation on a multi-centric, multi-vendor dataset with respect to breast denisty class and preprocessing steps, which included intensity clipping, contrast adjustment, histogram matching and breast masking. Our results show that architectures' responses varied across preprocessing steps. Among the tested models Swin UNETR achieved the highest weighted DICE scores among the evaluted networks - **0.80**, while nnUNet achieved the lowest Hausdorff distance - **37.51**. Among the tested configurations, breast masking and intensity clipping generally improved DICE scores of the models.

Keywords: FGT segmentation · Breast MRI · Deep learning · Generalization.

1 Introduction

The amount of fibroglandular tissue (FGT) is an important factor associated with breast cancer risk. During clinical evaluation, the amount of FGT is visually classified by radiologists into four categories under the Breast Imaging

Reporting and Data System (BI-RADS): A - almost entirely fat; B - scattered FGT, C - heterogeneous FGT, D - extreme FGT. However, such an evaluation has a limited inter-reader agreement. For this reason, quantitative methods for the evaluation of breast density based on automatic segmentation of FGT are used across different imaging modalities [21]. Advances in artificial intelligence have enabled highly accurate automated FGT segmentation in breast MRI, but despite strong state-of-the-art performance, important research gaps remain. Private single-center datasets are often used for network training and evaluation [22, 20], which doesn't provide insights into the networks' ability to generalize on data from other institutions [11]. Network results can also depend on the amount of FGT [1], which is not often reported. Additionally, image preprocessing techniques vary between implementations [13, 7, 3, 11], and it has been shown that normalization and data augmentation steps can influence network performance on downstream tasks [2].

To address the existing gap, our study evaluates the generalizability of deep neural networks on multi-centric data after training on publicly available datasets using different image preprocessing strategies. This study also stratifies each dataset by breast density classes for a more thorough evaluation of network performance.

2 Methods

For this study $n = 100$ cases with FGT and vessels segmentation's were selected from the Duke Breast dataset [18]. To assess the generalizability of FGT segmentation models on acquisitions from different MRI vendors, additional external datsets were selected. Two subsets of the ACRIN-6698 dataset [12] were selected. Each subset contained cases from different MRI vendors' scanners, including a GE Healthcare scanner and the Philips scanner. An additional, IRB approved, subset of a retrospective dataset (further denoted as "Internal"), acquired using a Siemens Healthineers scanner, was selected from among examinations used in previous scientific studies and performed at the Institute of Radiology, University Hospital Erlangen between 2015–2022.

For the above mentioned datasets, each examination was evaluated by a board-certified radiologist, with over 15 years of experience in breast MRI, to denote the amount of FGT (FGTa) observable on a precontrast T1-weighted (T1w) fat-saturated (FS) acquisition, using the standard BI-RADS classification scheme. All of the datasets used in this study were stratified according to FGTa class, and contained a relatively equal amount of cases from each breast density class. The Duke dataset subset was split into train, validation, and test splits with regard to FGTa class. For the external test set, FGT was segmented on the native T1-w FS acquisitions by medical students under the supervision of a board-certified radiologist. $n = 3$ cases containing mismatched lengths between images and masks were excluded from this study. The final DUKE subset was split into a 57/20/20 train/validation/test split after exclusion with an equal distribution of FGTa classes in the test set. The GE test set contained $n = 20$

cases ($n_{FGT_A}, n_{FGT_B}, n_{FGT_C}, n_{FGT_D} = 5$), Philips subset contained $n = 18$ cases ($n_{FGT_A}, n_{FGT_D} = 5, n_{FGT_B}, n_{FGT_C} = 4$) and the Internal test set contained $n = 34$ cases ($n_{FGT_A} = 8, n_{FGT_B} = 4, n_{FGT_C} = 6, n_{FGT_D} = 16$).

Preprocessing

In this work, several preprocessing pipelines were evaluated. The MRI volumes' axes were reordered, so that cases with different space encodings are uniformly oriented. Volume voxel spacing was adjusted through spline/bilinear interpolation to isotropic spacing of 1.0 mm. The baseline preprocessing pipeline only consisted of z-score normalization on the volume level. Additionally, the following preprocessing steps were evaluated: intensity clipping, contrast adjustment, histogram matching and breast masking.

Intensity clipping was performed at the 0.5th percentile and 99.5th percentile, to mitigate intensities from potential hyperintense artifacts. Contrast adjustment was based on the gamma correction transformation with $\gamma = 1.5$, and was performed to enhance the contrast between fat and FGT. Histogram matching was applied based on selected points of interest. Twenty points were selected between the 60th and 100th percentiles of the MRI volumes. The values were collected from volumes from the training and validation splits and averaged to produce a reference mapping. The volumes were then mapped to the reference values using piecewise linear interpolation. Histogram matching was applied in order to standardize the intensity profiles of all volumes. Breast masking consisted of applying a breast volume mask onto the MRI volumes. Breast masking was applied so that the networks focused only on the breast region, and to remove unnecessary information from the abdominal region of the scan. The breast volume masks were computed using a pretrained BreastDivider model [17].

In total, 10 configurations based on the mentioned preprocessing steps were created (see Tab.1). After preprocessing, the images were divided into 2D slices in the Z dimension. For model training, only slices with FGT masks that contained more than 100 foreground pixels were taken into account. This threshold was determined during the development of the image preprocessing pipeline by analyzing the histograms of volume masks. The threshold was applied to eliminate empty slices. After removing empty slices, the 2D slices were reconstructed back into 3D volumes and arranged into patches. After patch creation, data to be used in training and validation was passed to the model. In each training epoch, image augmentations were performed, including random flipping and random rotation, with a probability of 20% per call for each transform. This was performed in order to increase variability in data. Random operations, including data augmentation and sub-volume sampling, were initialized with a fixed seed to ensure reproducible and comparable experimental results between configurations.

Experiment Setup and Models

Various deep learning architectures achieving state-of-the-art performance in medical imaging segmentation tasks were used to evaluate the preprocessing

configurations, including the Attention U-Net, Swin UNETR, and a Mamba architecture - Mamba-HoME [14, 4, 15]. The Swin UNETR architecture as well as the Attention U-Net architecture contained 4–5 layers in the encoder and decoder stages. The nnU-Net framework [5] was used as a comparison for the evaluated models and configurations, as it is considered to be the gold standard for medical image segmentation.

All models, except nnU-Net were implemented using Python with the Pytorch, Pytorch Lightning and MONAI libraries. Evaluation metrics were calculated using numpy. The nnU-Net was trained using the same train/validation split as the other models, as opposed to the default five fold cross validation configuration. The batch size for model training was individually set for each network to match the maximum hardware capabilities. The training was performed on a single NVIDIA RTX 6000 Ada GPU with 48 GB of RAM.

With the exception of nnU-Net, all of the models' were allowed to train for a maximum of 100 epochs. A learning rate reduction scheduler and an early stopping mechanism were implemented. The learning rate decreased by a factor of 10^{-1} if the validation loss metric did not improve within 7 epochs, and the experiment stopped if no improvement was observed within 20 epochs. The Adam optimizer was used in the training pipeline. A DICE loss [10] implementation was used for model training. Two nnU-Net models were trained. The first model was trained using standard MRI volumes and corresponding FGT masks. The second model was trained using MRI volumes masked by the generated breast volume masks.

A grid search was performed on the model parameters and initial learning rate values. The models were trained using different initial learning rate values, ranging from $1e^{-3}$ to $1e^{-6}$. The model achieving the lowest validation loss value was further evaluated in the inference step.

Model evaluation

During inference, quantitative metrics were calculated to evaluate model performance. Firstly DICE score, Hausdorff distance (HD) and average symmetric surface distance (ASSD) were calculated on the volume level for each test dataset case. Inspired by Nowakowska et al. [13], generated breast volume masks were used to measure the relative amount of FGT to the breast volume ($FGT_{\%}$). The calculated weights were used for a weighted DICE score calculation:

$$FGT_{\%,i} = \frac{FGT_{vol,i}}{breast_{vol,i}} \cdot 100\% = w_i, \qquad DSC_w = \frac{\sum_i w_i \cdot DSC_i}{\sum_i w_i} \tag{1}$$

Where $breast_{vol,i}$ denotes the number of voxels in the breast volume mask for case i, $FGT_{vol,i}$ denotes the number of voxels in the FGT ground truth (GT) mask for case i and DSC_i represents the DICE score for case i. The weighted DICE score was used to reduce the penalization of cases with a lower relative amount of FGT. Models' performance was additionally evaluated using Bland-Altman plots, which represented if the models under- or overpredicted when

generating the FGT mask. A repeated measures ANOVA with Friedman test followed by a Tukey's test for pairwise comparisons were performed to determine statistically significant differences in the evaluated metrics between the selected preprocessing strategies and architecture combination's . Both tests use $\alpha = 0.05$.

3 Results

Model inference was performed on the complete test set with n=92 cases, with a relatively similar FGTa class distribution. Comparison between the different preprocessing-architecture combinations shows that weighted DICE score improves, when breast masking is used in the preprocessing pipeline (see Tab.1). nnUNet architecture achieved comparable results to configurations without breast masking and slightly lower as compared to configurations that utilized breast masking. Greater variations in model performance per configuration can be observed for HD and ASSD. Configurations that utilized breast masking achieved lower median ASSD for all models. The nnUNet models achieved a lower ASSD as compared to most model and preprocessing configurations altough the lowest value was achieved for the Swin UNETR architecture. The nnUNet models also achieved a significantly lower median HD as compared to most configurations. It can be observed that the effect of different preprocessing steps on median HD is not uniform across the evaluated architectures.

A qualitative analysis was conducted by comparing the segmentation results of the evaluated models. The comparison was performed on slices from random volumes from the DUKE dataset on FGTa classes A–D (see Fig. 1). The models achieved good performance when segmenting FGTa class B–D cases. The results are comparable across models, with small amounts of false-positive (red pixels) and false-negative (green pixels) regions. The models correctly outlined tissue boundaries for FGTa class A (see Fig. 1, FGTa A, right breast). The nnUNet 3d model incorrectly segmented fragments of the chest wall (see Fig. 1, FGTa B, left breast).

Table 1. Weighted DICE score (DSC) (↑), median, IQR Hausdorff distance (HD) (↓) and median, IQR ASSD (↓) for all model configurations for the complete test dataset (n=92 cases). B - Baseline; IC - Intensity clipping; CA - Contrast Adjustment; HIST - histogram matching; BM - breast masking.

Configurations	Weighted DSC					Median, IQR HD					Median, IQR ASSD				
	Att. U-Net	Swin UNETR	Mamba HoME	nnUNet	nnUNet + BM	Att. U-Net	Swin UNETR	Mamba HoME	nnUNet	nnUNet + BM	Att. U-Net	Swin UNETR	Mamba HoME	nnUNet	nnUNet + BM
B	0.69	0.72	0.69	0.72	0.77	128.04 (109.28, 150.91)	132.57 (106.00, 189.89)	163.54 (138.55, 198.43)	111.62 (87.48, 131.43)	111.56 (58.78, 289.12)	11.19 (4.08, 20.21)	4.56 (2.31, 11.74)	9.43 (3.91, 21.69)	4.94 (2.53, 9.58)	3.55 (1.83, 16.65)
B + IC	0.66	0.68	0.71			136.02 (115.76, 157.89)	143.56 (124.39, 172.23)	142.60 (119.89, 167.13)			16.20 (6.29, 30.98)	6.59 (2.99, 16.53)	8.39 (3.18, 19.20)		
B + IC + CA	0.67	0.70	0.67			126.71 (88.97, 143.26)	128.22 (110.07, 172.43)	127.53 (111.95, 159.99)			11.36 (3.85, 25.85)	5.59 (2.44, 14.21)	10.52 (3.44, 23.02)		
B + IC + HIST	0.64	0.69	0.71			146.77 (122.98, 173.93)	159.28 (135.47, 180.60)	138.35 (115.78, 170.42)			16.50 (7.20, 36.43)	8.88 (4.10, 20.85)	8.39 (3.47, 15.90)		
B + IC + HIST + CA	0.73	0.67	0.64			123.37 (90.49, 144.46)	146.11 (120.49, 187.93)	142.35 (116.54, 175.83)			5.82 (2.54, 12.13)	7.97 (3.80, 24.70)	12.59 (7.06, 24.77)		
B + BM	0.78	0.80	0.76			109.19 (67.63, 247.94)	239.17 (166.02, 266.14)	246.51 (113.39, 270.30)			4.26 (1.64, 6.35)	2.74 (1.39, 5.31)	4.52 (2.11, 8.08)		
B + IC + BM	0.78	0.80	0.79			117.83 (73.23, 256.87)	231.00 (110.96, 259.74)	245.86 (201.02, 271.83)			3.82 (1.66, 7.14)	2.76 (1.33, 5.74)	3.36 (1.53, 6.63)		
B + IC + CA + BM	0.79	0.77	0.76			211.24 (81.75, 262.81)	118.02 (83.50, 235.07)	150.93 (85.35, 252.86)			3.48 (1.94, 7.74)	3.97 (1.90, 7.82)	4.51 (2.03, 8.73)		
B + IC + HIST + BM	0.77	0.80	0.77			106.48 (72.79, 266.91)	86.83 (64.16, 235.31)	92.65 (65.59, 194.28)			3.87 (2.17, 8.94)	2.37 (1.40, 6.52)	3.68 (1.96, 8.91)		
B + IC + HIST + CA + BM	0.74	0.79	0.76			215.96 (93.76, 268.31)	98.06 (67.99, 160.09)	99.90 (70.61, 250.95)			5.93 (2.76, 14.79)	2.79 (1.47, 5.34)	4.04 (2.37, 8.19)		

For each trained architecture, a preprocessing configuration was chosen that achieved the lowest validation loss during training. Their performance was then evaluated for each FGTa class and compared to nnUNet configurations (see Tab.2) statistically over the whole dataset. Segmentation performance improved with the increase in FGTa. Class-wise mean DICE scores were similar across all the tested datasets, except for classes A–C in the Internal dataset. Variations in DICE scores across classes reflected the properties of the DICE metric, which is correlated with the amount of GT pixels in the mask [16]. Mean HD was high for all FGTa classes for the trained models across test datasets. The nnUNet 3d configuration consistently achieved lower HD across datasets. The nnUNet 3d + BM achieved the highest HD scores for the Philips and GE datasets, but performed best for the DUKE dataset. Mean ASSD was comparable for models for FGTa classes within each test dataset. ASSD values for the nnUNet 3d + BM model were the highest for the Philips and GE datasets. The Swin UNETR achieved comparable ASSD to nnUNet 3d.

Significant differences were observed for the selected preprocessing-architecture combinations (see Tab.2) for all evaluation metrics (all p-values < 0.001). The nnUnet models achieved significantly lower DS on the test set when compared to other evaluated model-preprocessing configurations in Tab.2 (all p-values <0.001

Table 2. Mean DSC (↑), Hausdorff distance (HD) (↓) and ASSD (↓) comparison across different breast density classes and datasets for the nnUNet and for trained models using the best-performing preprocessing configuration. Bold values represent the best score for each FGTa class.

Model	DICE							
	Class A	Class B	Class C	Class D	Class A	Class B	Class C	Class D
	Duke				GE			
Att. U-Net B + IC + CA + BM	**0.60 ± 0.07**	0.76 ± 0.01	**0.81 ± 0.05**	**0.87 ± 0.02**	**0.59 ± 0.10**	**0.70 ± 0.07**	0.82 ± 0.06	**0.85 ± 0.04**
Swin UNETR B + IC + BM	0.57 ± 0.10	0.76 ± 0.02	**0.81 ± 0.05**	**0.87 ± 0.04**	0.57 ± 0.12	**0.70 ± 0.06**	**0.83 ± 0.06**	**0.85 ± 0.05**
Mamba-HoME B + IC + BM	0.58 ± 0.10	**0.77 ± 0.02**	**0.81 ± 0.05**	**0.87 ± 0.04**	0.58 ± 0.11	**0.70 ± 0.05**	**0.83 ± 0.05**	**0.85 ± 0.05**
nnU-Net 3d	0.52 ± 0.09	0.70 ± 0.03	0.79 ± 0.05	0.85 ± 0.04	0.52 ± 0.13	0.67 ± 0.08	0.76 ± 0.03	0.75 ± 0.15
nnU-Net 3d + BM	0.52 ± 0.06	0.71 ± 0.02	0.77 ± 0.07	0.86 ± 0.04	0.44 ± 0.13	0.62 ± 0.08	0.65 ± 0.15	0.80 ± 0.07
	Philips				Internal			
Att. U-Net B + IC + CA + BM	0.55 ± 0.07	0.74 ± 0.03	0.86 ± 0.02	0.85 ± 0.05	0.39 ± 0.13	0.59 ± 0.07	0.73 ± 0.05	0.84 ± 0.05
Swin UNETR B + IC + BM	0.57 ± 0.07	0.74 ± 0.02	0.86 ± 0.02	**0.86 ± 0.04**	0.37 ± 0.15	0.61 ± 0.08	0.73 ± 0.08	**0.85 ± 0.06**
Mamba-HoME B + IC + BM	**0.59 ± 0.04**	**0.75 ± 0.02**	**0.87 ± 0.00**	0.85 ± 0.05	0.35 ± 0.13	0.61 ± 0.08	0.72 ± 0.08	0.84 ± 0.06
nnU-Net 3d	0.53 ± 0.10	0.62 ± 0.13	0.69 ± 0.24	0.73 ± 0.11	**0.42 ± 0.17**	0.61 ± 0.08	**0.75 ± 0.03**	0.74 ± 0.13
nnU-Net 3d + BM	0.55 ± 0.09	0.71 ± 0.05	0.82 ± 0.04	0.83 ± 0.03	**0.42 ± 0.13**	**0.62 ± 0.10**	0.73 ± 0.08	0.83 ± 0.06
	HD							
	Class A	Class B	Class C	Class D	Class A	Class B	Class C	Class D
	Duke				GE			
Att. U-Net B + IC + CA + BM	139.82 ± 78.25	81.26 ± 71.72	96.73 ± 78.83	189.17 ± 84.16	174.67 ± 102.53	162.79 ± 85.22	211.93 ± 67.63	231.48 ± 94.48
Swin UNETR B + IC + BM	187.52 ± 71.44	95.73 ± 59.78	112.41 ± 80.60	200.16 ± 81.19	233.04 ± 81.05	206.86 ± 82.44	252.99 ± 19.21	262.88 ± 31.83
Mamba-HoME B + IC + BM	226.27 ± 19.56	229.52 ± 30.79	187.94 ± 65.02	161.21 ± 103.73	281.67 ± 41.03	233.89 ± 66.95	230.51 ± 78.69	270.16 ± 22.25
nnU-Net 3d	102.75 ± 47.04	88.83 ± 36.58	79.55 ± 40.31	71.62 ± 40.01	**112.15 ± 18.16**	**111.01 ± 17.13**	**111.03 ± 21.23**	**99.82 ± 49.35**
nnU-Net 3d + BM	**41.69 ± 8.87**	**38.85 ± 7.75**	**35.95 ± 5.23**	**33.54 ± 10.05**	323.94 ± 28.73	312.57 ± 40.97	296.90 ± 22.97	293.08 ± 17.82
	Philips				Internal			
Att. U-Net B + IC + CA + BM	**119.99 ± 63.68**	**72.89 ± 14.09**	173.09 ± 76.21	207.74 ± 87.87	182.04 ± 102.70	266.64 ± 17.54	179.72 ± 91.81	267.65 ± 53.61
Swin UNETR B + IC + BM	184.19 ± 87.28	182.75 ± 73.93	238.60 ± 15.48	177.40 ± 86.68	222.31 ± 94.31	203.79 ± 90.96	177.40 ± 87.20	217.29 ± 77.33
Mamba-HoME B + IC + BM	202.01 ± 71.68	171.15 ± 73.18	208.20 ± 70.74	225.96 ± 80.03	198.96 ± 102.43	223.48 ± 69.18	241.63 ± 73.82	239.29 ± 77.82
nnU-Net 3d	153.27 ± 24.67	114.54 ± 4.86	**124.09 ± 12.99**	**139.46 ± 27.17**	131.76 ± 62.79	122.04 ± 20.39	109.32 ± 29.52	114.17 ± 34.44
nnU-Net 3d + BM	296.90 ± 16.74	271.94 ± 20.64	255.08 ± 27.59	279.94 ± 31.32	**117.02 ± 67.59**	**105.60 ± 8.44**	**80.18 ± 22.50**	**95.58 ± 45.64**
	ASSD							
	Class A	Class B	Class C	Class D	Class A	Class B	Class C	Class D
	Duke				GE			
Att. U-Net B + IC + CA + BM	**1.88 ± 0.66**	0.90 ± 0.25	0.89 ± 0.30	1.40 ± 0.45	10.64 ± 7.76	6.03 ± 2.11	9.51 ± 11.41	**7.26 ± 4.28**
Swin UNETR B + IC + BM	2.15 ± 0.81	**0.89 ± 0.23**	**0.85 ± 0.23**	1.25 ± 0.45	**9.46 ± 4.44**	**5.65 ± 1.42**	**6.22 ± 5.34**	9.41 ± 6.30
Mamba-HoME B + IC + BM	2.41 ± 1.16	1.10 ± 0.24	1.08 ± 0.35	1.24 ± 0.23	11.82 ± 6.69	6.99 ± 2.19	10.75 ± 12.23	9.24 ± 5.38
nnU-Net 3d	2.93 ± 1.27	1.50 ± 0.40	1.18 ± 0.47	1.33 ± 0.30	16.17 ± 9.82	5.84 ± 2.26	10.46 ± 5.01	6.38 ± 2.40
nnU-Net 3d + BM	1.84 ± 0.28	1.07 ± 0.40	0.94 ± 0.35	**1.01 ± 0.21**	67.52 ± 41.08	49.40 ± 30.03	81.77 ± 76.98	54.27 ± 49.03
	Philips				Internal			
Att. U-Net B + IC + CA + BM	4.95 ± 3.65	1.57 ± 0.76	2.46 ± 1.94	2.05 ± 0.86	16.32 ± 19.70	7.49 ± 4.96	11.07 ± 9.91	9.04 ± 8.55
Swin UNETR B + IC + BM	**4.39 ± 3.58**	**1.45 ± 0.43**	4.32 ± 6.13	**1.48 ± 0.90**	16.18 ± 20.14	5.68 ± 3.06	3.90 ± 1.53	**4.30 ± 3.50**
Mamba-HoME B + IC + BM	5.73 ± 5.42	1.49 ± 0.43	**1.28 ± 0.81**	1.90 ± 1.58	16.32 ± 19.10	6.52 ± 3.67	5.54 ± 2.12	5.10 ± 4.81
nnU-Net 3d	5.78 ± 2.21	5.32 ± 2.70	10.46 ± 10.45	8.69 ± 6.11	13.01 ± 15.89	7.52 ± 4.91	3.97 ± 0.80	9.29 ± 4.59
nnU-Net 3d + BM	29.09 ± 22.02	16.31 ± 25.33	6.65 ± 6.68	5.80 ± 5.93	**12.68 ± 16.80**	**3.81 ± 1.39**	**3.72 ± 1.71**	4.80 ± 3.80

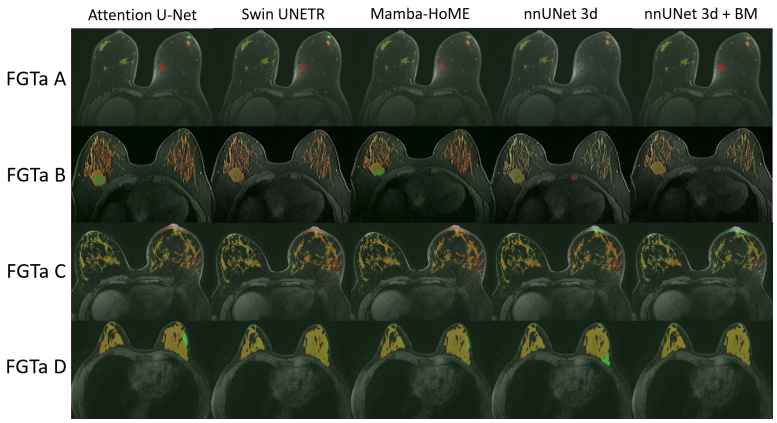

Fig. 1. Comparison between model outputs for FGTa A–D cases. Green pixels - ground truth (GT); Red pixels - model predictions; Yellow pixels - overlap.

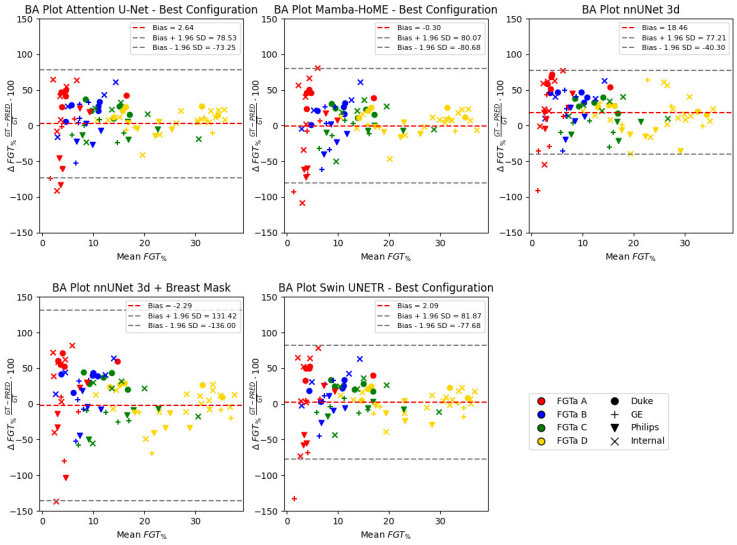

Fig. 2. Bland-Altman plots for the models in Tab.2. GT on the plot refers to the ground truth voxel amount relative to the breast volume. PRED on the plot refers to model predictions' voxel amount relative to the breast volume.

for comparisons between nnUnet and nnUnet+BM vs other configurations). No statistically significant differences in DS where observed between the selected Attention U-Net, Swin UNETR and Mamba configurations (lowest pairwise test p-value of 0.287 between Swin UNETR and Attention U-Net). Pairwise tests showed that Swin UNETR achieved significantly better ASSD than other mod-

els from Tab.2 ($p<0.001$), while both nnUNet models performed significantly better with regard to HD ($p<0.001$).

The configurations from Tab.2 were further evaluated using Bland-Altman (BA) plots (see Fig.2). The input data consisted of the percentage of the GT mask and the predicted mask voxels relative to the breast volume mask voxels. It can be observed that for all the models with the exception of nnUNet 3d the bias of the models was close to 0. It can also be observed that nnUNet 3d had the lowest standard deviation. nnUNet 3d that utilized breast masking had the highest standard deviation and outlier values up to -500%. All models with the exception of nnUNet 3d had low value outliers. The SWIN, Attention U-Net and Mamba networks segmented a similar amount of voxels to the GT.

4 Discussion

In this study, several image preprocessing configurations were tested and their impacts on FGT segmentation were evaluated. Additionally, the ability to generalize on multi-vendor scanner data in FGT segmentation was tested on $n = 92$ T1-w FS breast MRI acquisitions.

Overall, breast masking improved model performance for all trained models. These findings suggest that breast masking is a beneficial preprocessing step for FGT segmentation. A combination of breast masking, intensity clipping and contrast adjustment yielded good DICE scores for the test set and achieved comparable results to nnUNet across all FGTa classes. HD increased for most preprocessing configurations except the nnUNet for the DUKE test set. High HD could have been caused by model predictions on MRI volume fragments where no ground truth was present, as HD is highly sensitive to outliers [16]. ASSD decreased for most preprocessing configurations for the evaluated architectures and was comparable to nnUNet for most of the test datasets for all FGTa classes. Importantly in our results no universally optimal preprocessing strategy could be identified among the tested models. As such different combinations of preprocessing steps should be considered as a potential hyper-parameter during the training of new architectures for FGT segmentation. Our evaluations achieved a weighted DICE scores of 0.80 for the test set for certain configurations, which are comparable to other works: Nowakowska et al. [13] achieved a 0.864 weighted DICE score, Geissler et al. [3] achieved a mean DICE score of 0.78 and Lew et al. [7] achieved a DICE score of 0.86.

Previous studies in this research area often used single-center datasets for model training and evaluation [22, 9, 20, 7, 1]. Additionally, some studies do not include breast density class in their assessment [22, 9, 19] even tough it was previously shown that the results differ depending on the breast density class [1]. Because of these limitations in previous studies, in this study our network was trained using data from a single-site dataset and then tested on a multi-vendor, multi-site dataset in which we could also evaluate the dependency from FGTa. Thanks to this our study could show that when tested on external data sets an over all drop in performance can be observed for breast with FGTa classes of

A through C. Interestingly the performance of the models didn't deteriorate for acquisitions of class D.

Our study is not without limitations. Firstly, the size of the training set and test sets should be considered as limited. This is an important issue as breast MRI is highly heterogeneous in regards to the image appearance of acquisitions performed on scanners of different vendors. It should also be noted that the models achieved an overall high HD for the test datasets. However such high HD scores are hard to interpret when outliers are present. In the future, Hausdorff Distance 95th Percentile (HD95) should be evaluated alongside HD, as it is more robust to outliers. Our study also didn't evaluate the performance of segmentation in cases with hyper-intense artifacts which can be often caused by coil-flare-ups and erroneous fat-saturation in T1w FS acquisitions [6, 8]. In future works, the robustness of different preprocessing strategies, under such conditions should be evaluated.

5 Conclusions

This study investigated how different preprocessing strategies affect the performance and ability to generalize of different models for FGT segmentation. Addition of Breast masking and intensity clipping to standard z-score normalization generally lead to an improvement of performance. Addition of further preprocessing steps should be considered for some models.

Acknowledgments. This study was funded by the Polish National Agency for Academic Exchange within the Polish Returns Programme.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dalmış, M.U., Litjens, G., Holland, K., Setio, A., Mann, R., Karssemeijer, N., Gubern-Mérida, A.: Using deep learning to segment breast and fibroglandular tissue in mri volumes. Medical physics **44**(2), 533–546 (2017)
2. Dohmen, M., Klemens, M.A., Baltruschat, I.M., Truong, T., Lenga, M.: Similarity and quality metrics for mr image-to-image translation. Scientific Reports **15**(1), 3853 (2025)
3. Geissler, K., Meine, H., Laue, H., Grimm, R., von Busch, H., Diekmann, S., Szolar, D., Ohlmeyer, S., Szurowska, E., Fischer, U., et al.: Multi-institutional classification of fibroglandular tissue and background parenchymal enhancement in breast mri using deep learning. Journal of Medical Imaging **13**(3), 034501–034501 (2026)
4. Hatamizadeh, et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

6. Kapsner, L.A., Ohlmeyer, S., Folle, L., Laun, F.B., Nagel, A.M., Liebert, A., Schreiter, H., Beckmann, M.W., Uder, M., Wenkel, E., et al.: Automated artifact detection in abbreviated dynamic contrast-enhanced (dce) mri-derived maximum intensity projections (mips) of the breast. European Radiology **32**(9), 5997–6007 (2022)
7. Lew, C.O., Harouni, M., Kirksey, E.R., Kang, E.J., Dong, H., Gu, H., Grimm, L.J., Walsh, R., Lowell, D.A., Mazurowski, M.A.: A publicly available deep learning model and dataset for segmentation of breast, fibroglandular tissue, and vessels in breast mri. Scientific reports **14**(1), 5383 (2024)
8. Liebert, A., Das, B.K., Kapsner, L.A., Eberle, J., Skwierawska, D., Folle, L., Schreiter, H., Laun, F.B., Ohlmeyer, S., Uder, M., et al.: Smart forecasting of artifacts in contrast-enhanced breast mri before contrast agent administration. European Radiology **34**(7), 4752–4763 (2024)
9. Ma, X., Wang, J., Zheng, X., Liu, Z., Long, W., Zhang, Y., Wei, J., Lu, Y.: Automated fibroglandular tissue segmentation in breast mri using generative adversarial networks. Physics in Medicine & Biology **65**(10), 105006 (2020)
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
11. Müller-Franzes, G., Müller-Franzes, F., Huck, L., Raaff, V., Kemmer, E., Khader, F., Arasteh, S.T., Lemainque, T., Kather, J.N., Nebelung, S., et al.: Fibroglandular tissue segmentation in breast mri using vision transformers: a multi-institutional evaluation. Scientific Reports **13**(1), 14207 (2023)
12. Newitt, et al.: Acrin 6698/i-spy2 breast dwi [data set](2021). DOI: https://doi.org/10.7937/tcia. kk02-6d95 (2021)
13. Nowakowska, S., Borkowski, K., Ruppert, C.M., Landsmann, A., Marcon, M., Berger, N., Boss, A., Ciritsis, A., Rossi, C.: Generalizable attention u-net for segmentation of fibroglandular tissue and background parenchymal enhancement in breast dce-mri. Insights into Imaging **14**(1), 185 (2023)
14. Oktay, et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
15. Płotka, S., Mert, G., Chrabaszcz, M., Szczurek, E., Sitek, A.: Mamba goes home: Hierarchical soft mixture-of-experts for 3d medical image segmentation (2026), https://arxiv.org/abs/2507.06363
16. Reinke, A., Tizabi, M.D., Sudre, C.H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., et al.: Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642 (2021)
17. Rokuss, M., Hamm, B., Kirchhoff, Y., Maier-Hein, K.: Divide and conquer: A large-scale dataset and model for left–right breast mri segmentation. arXiv preprint arXiv:2507.13830 (2025)
18. Saha, et al.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. The Cancer Imaging Archive **10** (2021)
19. Samperna, R., Moriakov, N., Karssemeijer, N., Teuwen, J., Mann, R.M.: Exploiting the dixon method for a robust breast and fibro-glandular tissue segmentation in breast mri. Diagnostics **12**(7), 1690 (2022)
20. Wei, D., Jahani, N., Cohen, E., Weinstein, S., Hsieh, M.K., Pantalone, L., Kontos, D.: Fully automatic quantification of fibroglandular tissue and background parenchymal enhancement with accurate implementation for axial and sagittal breast mri protocols. Medical physics **48**(1), 238–252 (2021)

21. Wengert, G., et al.: Inter-and intra-observer agreement of bi-rads-based subjective visual estimation of amount of fibroglandular breast tissue with magnetic resonance imaging: comparison to automated quantitative assessment. European radiology **26**(11), 3917–3922 (2016)
22. Zhang, Y., Chan, S., Chen, J.H., Chang, K.T., Lin, C.Y., Pan, H.B., Lin, W.C., Kwong, T., Parajuli, R., Mehta, R.S., et al.: Development of u-net breast density segmentation method for fat-sat mr images using transfer learning based on non-fat-sat model. Journal of digital imaging **34**(4), 877–887 (2021)