MeisenMeister: A Simple Two Stage Pipeline for Breast Cancer Classification on MRI

Benjamin Hamm[1,2], Yannick Kirchhoff[1,3,4], Maximilian Rokuss[1,3], Klaus Maier-Hein[1,2,3,4,5]

[1]German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany
[2]Medical Faculty, Heidelberg University, Germany
[3]Faculty of Mathematics and Computer Science, Heidelberg University
[4]HIDSS4Health, Karlsruhe/Heidelberg, Germany
[5]Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
{benjamin.hamm, yannick.kirchhoff, maximilian.rokuss}@dkfz-heidelberg.de

INTRODUCTION

The ODELIA Breast MRI Challenge 2025 addresses a critical issue in breast cancer screening: improving early detection through more efficient and accurate interpretation of breast MRI scans. Even though methods for general-purpose whole-body lesion segmentation [1] as well as multi–time-point analysis [2] exist, breast cancer detection remains highly challenging, largely due to the limited availability of high-quality segmentation labels. Therefore, developing robust classification-based approaches is crucial for the future of early breast cancer detection, particularly in applications such as large-scale screening. In this write-up, we provide a comprehensive overview of our approach to the challenge. We begin by detailing the underlying concept and foundational assumptions that guided our work. We then describe the iterative development process, highlighting the key stages of experimentation, evaluation, and refinement that shaped the evolution of our solution. Finally, we present the reasoning and evidence that informed the design choices behind our final submission, with a focus on performance, robustness, and clinical relevance. We release our full implementation publicly at https://github.com/MIC-DKFZ/MeisenMeister

MATERIALS AND METHODS

Divide And Conquer Pipeline

We initiated our approach by designing a customized Divide-and-Conquer pipeline to address the computational challenges posed by high-resolution breast MRI volumes. Given the extremely high spatial resolution and large dimensionality of the input scans, processing an entire volume in a single forward pass is impractical due to memory limitations. Although the challenge organizers provided a unilateral cropping script to isolate a single breast per scan, our analysis showed that the resulting subvolumes remained too large for efficient model inference. To mitigate this, we developed a dedicated dataset and segmentation model, presented as *Divide and Conquer: A Large-Scale Dataset and Model for Left–Right Breast MRI Segmentation* [3]. To improve performance and generalizability, we employed an active learning strategy that leveraged
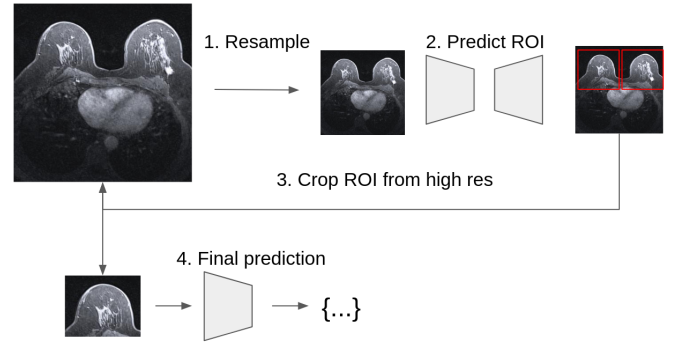

Fig. 1. The pipeline follows a divide-and-conquer strategy to efficiently and accurately analyze high-resolution breast MRI volumes. (1) The original high-resolution 3D MRI is first resampled to a lower resolution to reduce computational complexity. (2) A segmentation model is applied to the resampled image to delineate the breast regions. Bounding boxes are then derived from the segmentation masks to identify the spatial extent of each breast. (3) Using these predicted bounding boxes, the corresponding high-resolution regions of interest (ROIs) are cropped directly from the original image, preserving anatomical detail. (4) Each high-resolution ROI is processed independently by a final prediction model, enabling focused, high-accuracy analysis at the level of individual breasts.

predictions from non–contrast-enhanced T1-weighted images as pseudo-ground truth for co-registered sequences, building on previously demonstrated reliability of cross-sequence annotation transfer [4]. This model enables accurate localization of individual breasts and the extraction of tight bounding boxes, thereby enabling efficient, ROI-focused processing in subsequent stages. A hierarchical overview of this pipeline is illustrated in Figure 1.

Public Data

To enhance the challenge data, we sought to incorporate publicly available datasets. A primary challenge was identifying datasets with benign-labeled cases and harmonizing imaging modalities across varying acquisition protocols. We ultimately integrated two external datasets into our pipeline: the Advanced-MRI-Breast-Lesions (AMBL) dataset [5], which includes both malignant and benign cases, and the Duke-Breast-Cancer-MRI (DUKE) dataset [6], which contains only

malignant cases. While DUKE does not include benign annotations, we utilized it in a separate experimental setting described later in the paper.

Model Training

We utilized the nnU-Net framework [7] to create a classification training framework. Given the heterogeneity in available input modalities—particularly the variable number of post-contrast images and the optional presence of T2-weighted sequences—we conducted a series of ablation studies to determine the optimal input configuration. Specifically, we investigated (i) how many post-contrast phases were necessary for robust performance, and (ii) whether the inclusion of T2-weighted images contributed meaningful gains.

After finalizing the input channel configuration, we proceeded to ablate various backbone architectures for the classification model. Specifically, we compared three architectures: ResNet-18 [8], ResEncL (a scaled-up variant of the nnU-Net encoder) and ResEncL with Squeeze-and-Excitation (SE) blocks [9]. A key consideration during this process was architectural compatibility with supervised segmentation-based pretraining.

We first conducted supervised pretraining on the MAMMA-MIA [10] dataset using the standard nnU-Net framework. For this, we trained a single model (no ensembling) on the full dataset for 2000 epochs. While this prolonged training clearly led to overfitting, it served as a valuable initialization point for our downstream task. Following pretraining, we explored several fine-tuning strategies on the target classification task. Specifically, we compared: (i) linear probing, where only the classification head is trained; (ii) full fine-tuning of all model weights; and (iii) fine-tuning with learning rate warm-up, including two variants—one with a warm-up from 1e-4 to 1e-2, and another from 1e-5 to 1e-3.

After finalizing our fine-tuning strategy, we performed an ablation of data augmentation (DA) techniques. Effective augmentation was essential due to the limited number of training samples and the high susceptibility of models to overfitting in this setting. However, applying overly aggressive DA—particularly intensity-based transformations—can obscure diagnostically relevant features, effectively reducing the task to class distribution matching rather than promoting the learning of meaningful representations. To balance augmentation strength and fidelity, we systematically ablated individual DA components using only fold 0 (with CAM held out as the test center), as this fold showed good signal and resource constraints prevented full 5-fold ablation. The full list of compared DA techniques can be seen in TableII. These include contrast adjustment, gamma correction, Gaussian blur, Gaussian noise, multiplicative brightness, simulated low resolution, full 3D spatial transformations, in-plane spatial transformations, scaling, and elastic deformations. Each technique was tested in isolation to assess its effect

on generalization and inform the final augmentation strategy used in our pipeline. In addition, we found that cropping away non-informative background regions by setting them to zero using the segmentation masks generated by our Divide-and-Conquer model led to significant performance gains.

Finally, we explored the potential of incorporating the DUKE dataset into our training pipeline, given its availability. However, DUKE contains only malignant cases, making it incompatible with the original multi-class classification task. To address this, we reformulated the problem as a binary classification task with two labels: healthy and lesion-present (i.e., benign + malignant). At inference time, we mapped the binary predictions back to the original three-class challenge setting. Specifically, if the model predicted lesion-present, we assigned its probability to the malignant class and attributed the remaining probability to benign. Conversely, if the model predicted healthy, we assigned that probability to the healthy class and the complement to benign. This approach inherently prevents the model from explicitly predicting the benign class. While this might appear limiting, we thought it could reduce confusion between healthy and benign, as well as between malignant and benign, ultimately improving overall class separation. Given that benign is a minority class in the dataset, we hoped this strategy could still yield a strong overall performance under macro-averaged metrics, despite the lack of a dedicated prediction path for benign cases.

RESULTS AND DISCUSSION

The results of our incremental pipeline improvements are summarized in TableI. We report the mean AUROC, averaged across five cross-validation folds, where each fold corresponds to holding out one acquisition center as the test set. Including the AMBL dataset provides a modest gain, while switching to the ResEncL backbone further improves performance. Fine-tuning strategies with learning rate warm-up offer clear advantages over linear probing, with warm-up schedules producing the largest single improvement—highlighting the impact of pretraining. Data augmentation and background masking contribute additional gains. The final configuration—combining ResEncL, warm-up fine-tuning, and data augmentation—achieves a strong baseline, which is further enhanced by the background masking step.

The results of our data augmentation ablation study are summarized in TableII. Some techniques—particularly gamma transformation, elastic deformation, and simulated low resolution—led to performance degradation. These findings underscore the importance of carefully selecting augmentations in medical imaging tasks, where overly strong or unrealistic distortions can obscure clinically relevant features. The baseline model without augmentation achieves an AUROC of 0.826, which is exceeded by several well-calibrated augmentations, most notably Gaussian noise (0.851) and scaling (0.843).

TABLE I
ABLATION RESULTS ACROSS KEY COMPONENTS OF OUR PIPELINE.
REPORTED AUROC VALUES ARE AVERAGED ACROSS 5
CROSS-VALIDATION FOLDS, EACH CORRESPONDING TO ONE HELD-OUT
ACQUISITION CENTER. BOLD ENTRIES INDICATE THE SETTING WE
COMMIT TO USE FURTHER.

Setting	Scheme	Mean AUROC
Data	Only Odelia Data	0.623
	Add AMBL Data	**0.649**
Networks	ResNet18	0.608
	ResEncL SE	0.631
	ResEncL	**0.649**
Finetuning (Mama Mia)	Linear Probing	0.651
	Full Finetuning	0.725
	Warmup 1e-4 1e-2	0.729
	Warmup 1e-5 1e-3	**0.738**
Data Augmentation	Data Augmentation	**0.749**
Preprocessing	Mask Background	**0.765**
	ResEncL + Warmup + DA	0.736

TABLE II
ABLATION STUDY OF INDIVIDUAL DATA AUGMENTATION TECHNIQUES.
EACH AUGMENTATION WAS APPLIED IN ISOLATION USING THE RESENCL
BACKBONE WITH WARM-UP FINE-TUNING. REPORTED AUROC VALUES
CORRESPOND TO FOLD 0, WHERE CAM WAS USED AS THE HELD-OUT
TEST CENTER. GREEN CELLS INDICATE IMPROVEMENT OVER THE
NO-AUGMENTATION BASELINE (0.826), WHILE RED CELLS INDICATE A
PERFORMANCE DROP.

Augmentation Technique	AUROC
Contrast Transform	0.837
Gamma Transform	0.823
Gaussian Blur Transform	0.819
Gaussian Noise Transform	0.851
Multiplicative Brightness Transform	0.827
Simulate Low Resolution Transform	0.812
Spatial Transform	0.836
Spatial Transform Inplane	0.813
Scaling	0.843
Elastic Deform	0.797
No Augmentation	0.826

Batch size results are summarized in TableIII. Reducing the batch size increases gradient noise during training, which can serve as a beneficial regularizer. The results show a clear trend: smaller batch sizes consistently improve performance. A batch size of 1 yielded the highest mean AUROC of 0.765, averaged across five cross-validation folds. While training with very small batch sizes can introduce instability or slow convergence, in our setting it proved to be an effective and simple regularization strategy that improved generalization.

Given the variability in available MRI sequences across datasets, it was important to identify a modality combination that works for all centers. Including T2-weighted images alongside pre- and post-contrast sequences resulted in degraded performance (0.584), suggesting that the T2 modality either introduced noise or lacked sufficient cross-dataset consistency. Using only pre-contrast and two post-contrast phases

TABLE III
EFFECT OF BATCH SIZE ON CLASSIFICATION PERFORMANCE USING THE
RESENCL BACKBONE WITH WARM-UP FINE-TUNING. REPORTED AUROC
VALUES ARE AVERAGED ACROSS 5 CROSS-VALIDATION FOLDS, EACH
CORRESPONDING TO ONE HELD-OUT ACQUISITION CENTER. SMALLER
BATCH SIZES LED TO IMPROVED PERFORMANCE, WITH A BATCH SIZE OF 1
ACHIEVING THE HIGHEST MEAN AUROC.

Batch Size	Mean AUROC
4	0.745
2	0.740
1	**0,765**

(middle and last) significantly improved performance (0.636). The best results were achieved using the first and second post-contrast phases together with the pre-contrast scan (0.649), indicating that early dynamic enhancement patterns are particularly informative for classification. Results are shown in TableIV,

TABLE IV
EFFECT OF INPUT CHANNEL CONFIGURATION ON CLASSIFICATION
PERFORMANCE. REPORTED AUROC VALUES ARE AVERAGED ACROSS 5
CROSS-VALIDATION FOLDS, EACH CORRESPONDING TO ONE HELD-OUT
ACQUISITION CENTER. THE COMBINATION OF PRE-CONTRAST AND EARLY
POST-CONTRAST PHASES YIELDS THE HIGHEST PERFORMANCE, WHILE
ADDING T2 LEADS TO DEGRADATION.

Input Configuration	AUROC
Pre + Post middle + Post last + T2	0.584
Pre + Post middle + Post last	0.636
Pre + Post 1 + Post 2	**0.649**

To compare performance under different task formulations and preprocessing variants, we evaluated both the original three-class classification task ("Regular Task") and a simplified binary reformulation ("Binary Task") as described earlier. The results are summarized in TableV. Each row incrementally adds components to the baseline configuration (ResEncL with warm-up fine-tuning), allowing us to isolate their effects. For the regular task, the best performance (0.776) was achieved with the inclusion of data augmentation, background masking and applying isotropic spacing. In contrast to our initial hypothesis, the binary task formulation did not consistently benefit from the incremental addition of components. Its best performance (0.757) was achieved when isotropic spacing was included, yet this still lagged behind the corresponding result in the regular task (0.776). These findings suggest that the trade-off between the number of benign cases and the confusion among the remaining classes did not yield the expected benefit.

CONCLUSION

Overall, we found it challenging to extract a stable and informative training signal from the available data, even after incorporating additional public datasets. Models exhibited a strong tendency to overfit, and training outcomes were highly

TABLE V

Setting	Scheme	Mean AUROC
Regular Task	ResEncL + Warmup + Data Augmentation + Mask Background	0.765
	+ Isotropic Spacing	0.776
Binary Task	ResEncL + Warmup + Data Augmentation + Mask Background	0.722
	+ Isotropic Spacing	0.757

sensitive to initialization and data splits, with substantial performance variance across runs using identical settings. This instability suggests a high level of noise in the learning process, making reproducibility difficult.

There remain several promising directions we have yet to explore. These include multi-resolution feature aggregation (e.g., in the spirit of HRNet [11]), additional regularization strategies to further combat overfitting, improved model selection techniques during training, and more extensive pretraining approaches. Additionally, gaining access to more data through secure federated learning [12], [13] or by leveraging efficient annotation tools such as nnInteractive [14] could be highly beneficial, as the challenge remains primarily a data limitation rather than a lack of strong architectural solutions.

REFERENCES

[1] M. Rokuss, Y. Kirchhoff, S. Akbal, B. Kovacs, S. Roy, C. Ulrich, T. Wald, L. T. Rotkopf, H.-P. Schlemmer, and K. Maier-Hein, "Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3d whole-body imaging," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 872–30 885.

[2] M. R. Rokuss, Y. Kirchhoff, S. Roy, B. Kovacs, C. Ulrich, T. Wald, M. Zenk, S. Denner, F. Isensee, P. Vollmuth *et al.*, "Longitudinal segmentation of ms lesions via temporal difference weighting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 64–74.

[3] M. Rokuss, B. Hamm, Y. Kirchhoff, and K. Maier-Hein, "Divide and conquer: A large-scale dataset and model for left-right breast mri segmentation," 2025. [Online]. Available: https://arxiv.org/abs/2507.13830

[4] T. Wald, B. Hamm, J. C. Holzschuh, R. El Shafie, A. Kudak, B. Kovacs, I. Pflüger, B. von Nettelbladt, C. Ulrich, M. A. Baumgartner *et al.*, "Enhancing deep learning methods for brain metastasis detection through cross-technique annotations on space mri," *European Radiology Experimental*, vol. 9, no. 1, pp. 1–14, 2025.

[5] D. Daniels, D. Last, K. Cohen, Y. Mardor, and M. Sklair-Levy, "Standard and delayed contrast-enhanced mri of malignant and benign breast lesions with histological and clinical supporting data (advanced-mri-breast-lesions) (version 2)," https://doi.org/10.7937/C7X1-YN57, 2024, the Cancer Imaging Archive [dataset].

[6] A. Saha, M. R. Harowicz, L. J. Grimm, J. Weng, E. H. Cain, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski, "Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]," *The Cancer Imaging Archive*, vol. 10, 2021.

[7] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[10] L. Garrucho, K. Kushibar, C.-A. Reidel, S. Joshi, R. Osuala, A. Tsirikoglou, M. Bobowicz, J. Del Riego, A. Catanese *et al.*, "A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations," *Scientific data*, vol. 12, no. 1, p. 453, 2025.

[11] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[12] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[13] B. Hamm, Y. Kirchhoff, M. Rokuss, P. Schader, P. Neher, S. Parampottupadam, R. Floca, and K. Maier-Hein, "Efficient privacy-preserving medical cross-silo federated learning," *Authorea Preprints*, 2025.

[14] F. Isensee, M. Rokuss, L. Krämer, S. Dinkelacker, A. Ravindran, F. Stritzke, B. Hamm, T. Wald, M. Langenberg, C. Ulrich *et al.*, "nninteractive: Redefining 3d promptable segmentation," *arXiv preprint arXiv:2503.08373*, 2025.