# Replication and Reproducibility Analysis of Breast Tumor Classification and Segmentation Benchmarks on the BreastDM Dataset

Anonymous

**Abstract.** Reproducing deep learning benchmarks on medical imaging datasets is challenging due to differences in data organization, preprocessing, model adaptation, and evaluation protocols. This study systematically reproduces segmentation and classification benchmarks on the BreastDM DCE-MRI dataset, covering 2D/3D segmentation models, 2D/3D classification models, and SENet–ViT fusion strategies. The reproduced results largely recover the main reported trends—2D segmentation outperforms 3D segmentation, benign/malignant classification is feasible, subtraction sequences do not consistently improve performance, and LG-CAFN remains a strong fusion baseline—although exact model rankings and threshold-dependent metrics differ. We document several reproducibility challenges encountered during this process. As an exploratory extension, we combine 3D CNN ensemble and ViT branch output probabilities without training a new backbone, observing AUC values of 0.9187 in Exp-1 and 0.8929 in Exp-2, which suggests potential complementarity between 3D sequence/depth features and Transformer-based global representations.

## 1 Introduction

Breast cancer is one of the most prevalent malignant tumors among women worldwide, and early detection plays a critical role in improving clinical outcomes and reducing mortality [16, 5]. Among various imaging modalities, dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) provides rich temporal enhancement information that can reveal tumor morphology, location, and contrast-enhancement patterns, with advantages over mammography and ultrasound in certain clinical settings [10]. However, manual interpretation of DCE-MRI is time-consuming and highly dependent on radiologist experience [11]. Developing reliable automated methods for breast tumor segmentation and benign/malignant classification therefore remains an important goal.

Deep learning methods have achieved strong performance in breast tumor segmentation and classification by automatically learning higher-level feature representations from imaging data [11]. However, the reliable evaluation of such methods depends on publicly available, well-annotated datasets. For breast DCE-MRI, public datasets that simultaneously support both segmentation and classification remain relatively limited [17]. The BreastDM dataset partially addresses this gap, providing 232 breast DCE-MRI cases with pre-contrast, post-contrast,

and subtraction sequences [17]. The original BreastDM study evaluated multiple 2D, 3D, and fusion-based methods, and proposed a local–global cross-attention fusion network (LG-CAFN) to improve classification performance [17].

The reproducibility of deep learning benchmarks remains a significant challenge [1, 2]. Even when the dataset and model architecture are identical, differences in training stochasticity, preprocessing, implementation details, checkpoint selection, and evaluation protocols can lead to notable performance variations [1]. This issue is particularly important in medical imaging, where benchmark conclusions may influence subsequent algorithm development. Although BreastDM provides an important public benchmark, systematic independent reproduction and reproducibility analysis of its results remain limited.

This paper presents a systematic reproduction and reproducibility analysis of the BreastDM benchmark, including 2D and 3D segmentation models, 2D and 3D classification models, and the original SENet–ViT fusion strategies. Beyond reproduction, we explore a lightweight probability-level fusion that combines 3D CNN ensemble outputs with ViT branch outputs to examine whether 3D sequence/depth information and Transformer-based global information provide complementary classification cues. The main contributions are threefold:

1. A systematic reproduction of the BreastDM segmentation, classification, and fusion benchmarks, verifying key reported trends.
2. Documentation of practical reproducibility challenges related to implementation details, data preprocessing, model adaptation, and evaluation protocols.
3. An exploratory 3D–Transformer probability-level fusion that, without training a new backbone, suggests complementarity between 3D CNN and ViT representations as reflected in improved AUC.

## 2 Methods

### 2.1 Dataset

BreastDM is a DCE-MRI dataset containing 232 cases (147 malignant, 85 benign), each with pre-contrast, post-contrast, and subtraction sequences [17]. Two classification settings were evaluated: Exp-1 used one pre-contrast and eight post-contrast sequences (9-channel input), while Exp-2 additionally included eight subtraction sequences (17-channel input). For segmentation, tumor annotations were generated with reference to the pre-contrast, second post-contrast, and corresponding subtraction images.

### 2.2 Replicated Models and Fusion Strategies

For segmentation, we reproduced seven 2D models (FCN [12], DeepLabV3 [3], Unet [14], Unet-VGG16 [15], and UNeXt) and three 3D models (3D-Unet [4], 3D-VNet [13], and 3D-DenseSeg). For 2D classification, we reproduced CNN-based baselines (VGG [15], ResNet [7], DenseNet [9], SENet [8]) and Transformer-based baselines (ViT-6, ViT-7 [6]). For 3D classification, we reproduced 3DResNet18/50/101, 3DResNeXt101, 3DShuffleNet/V2, and 3DMobileNet/V2. We also

reproduced three original fusion strategies: feature-level fusion, decision-level fusion, and LG-CAFN, all based on SENet and ViT branches.

### 2.3 3D Ensemble and 3D–Transformer Fusion Extension

After reproducing the 3D classification models, we explored a lightweight probability-level fusion extension. First, we selected 3D CNN models for ensembling based exclusively on validation-set AUC, without using any test-set information. In Exp-1, the three selected models were 3DResNet50 (validation AUC = 0.8566), 3DMobileNet (0.8707), and 3DShuffleNetV2 (0.8150). In Exp-2, the selected models were 3DMobileNetV2 (0.8898), 3DMobileNet (0.8320), and 3DShuffleNetV2 (0.9028). Two ensembling strategies were evaluated: simple averaging and validation-AUC-weighted averaging. In the weighted scheme, each model's weight was computed as $w_i = \mathrm{AUC}_i^{\mathrm{val}} / \sum_j \mathrm{AUC}_j^{\mathrm{val}}$, using only validation-set performance.

The resulting 3D ensemble probability was then fused with the ViT-7 branch probability to examine whether 3D information and Transformer-based global information provide complementary cues. For the average fusion variant, the 3D ensemble and ViT-7 probabilities were equally weighted ($w_{\mathrm{3D}} = w_{\mathrm{ViT}} = 0.5$); for the weighted variant, each branch's weight was proportional to its validation-set AUC.

This extension did not train a new backbone or introduce an additional trainable checkpoint; it used the output probabilities of already reproduced and individually evaluated models for post-hoc analysis. Therefore, the extension should be interpreted as an exploratory analysis of cross-branch complementarity rather than a new trainable architecture.

### 2.4 Evaluation Metrics and Implementation

Segmentation was evaluated using DSC, mIoU, and PPV; classification was evaluated using sensitivity, specificity, accuracy, precision, and AUC, all following the definitions in [17]. This study placed particular emphasis on AUC as a threshold-independent metric. For threshold-dependent metrics, the decision threshold was selected on the validation set using Youden's index and fixed for the held-out test set. All experiments were implemented in PyTorch. For CNN-based models, ImageNet-pretrained first convolutional layers were adapted to the corresponding input-channel number. For ViT-based models, patch and positional embeddings were adapted according to the input size and channel number. All final reported metrics were recomputed from saved checkpoints.

## 3 Results

### 3.1 Segmentation Results

Table 1 compares the original and reproduced segmentation results. The reproduced 2D results were slightly lower than the original on several metrics, but

**Table 1.** Original and reproduced test-set segmentation results on BreastDM.

| Model | DSC (%) | | mIoU (%) | | PPV (%) | |
|---|---|---|---|---|---|---|
| | Orig. | Repr. | Orig. | Repr. | Orig. | Repr. |
| FCN-50 | 71.10 | 68.36 | 74.50 | 78.23 | 77.00 | 69.13 |
| FCN-101 | 71.40 | 67.83 | 79.00 | 77.88 | 76.20 | 71.92 |
| DeepLabV3-50 | 72.90 | 68.91 | 79.30 | 78.14 | 80.30 | 65.28 |
| DeepLabV3-101 | 73.40 | 68.24 | 78.70 | 78.07 | **86.50** | 71.08 |
| Unet | **73.70** | 72.86 | **80.30** | **82.22** | 83.60 | **75.83** |
| Unet-VGG16 | 70.40 | **73.69** | 77.30 | 82.12 | 79.60 | 74.10 |
| UNeXt | 70.10 | 70.77 | 78.50 | 79.93 | 86.30 | 71.45 |
| 3D-Unet | 66.50 | 68.07 | 75.10 | 79.15 | 83.30 | 66.76 |
| 3D-VNet | 65.60 | 64.90 | 74.20 | 77.91 | 77.10 | 64.55 |
| 3D-DenseSeg | 67.00 | 62.61 | 74.90 | 76.77 | 84.10 | 68.96 |

relative trends remained largely consistent: Unet-VGG16 achieved the highest reproduced DSC (73.69%) and Unet the highest mIoU (82.22%). The reproduced 3D models performed worse than 2D models, consistent with the original observation. Figure 1 shows qualitative results on representative test samples; most 2D models localized the main tumor regions well, while 3D models showed more noticeable boundary deviations.

### 3.2 Classification Results

Table 2 summarizes the original and reproduced AUC for all classification models. Among reproduced 2D models, ViT-6 achieved the highest AUC in Exp-1 (0.8602) and ResNet101 in Exp-2 (0.8640). For 3D models, 3DResNet50 led in Exp-1 (0.8582) and 3DMobileNetV2 in Exp-2 (0.8634). Consistent with the original observation, adding subtraction sequences in Exp-2 did not uniformly improve performance.

### 3.3 Fusion and Extension Results

Table 3 shows the reproduced AUC for the three original fusion strategies. LG-CAFN remained the most competitive, achieving the highest reproduced AUC of 0.8847 in Exp-1. In Exp-2, its advantage over simpler baselines was less pronounced. Table 4 shows the 3D–Transformer fusion extension results. Without training a new backbone, the 3D Ensemble + ViT fusion achieved AUC of 0.9187 in Exp-1 and 0.8929 in Exp-2, exceeding both the 3D ensemble alone and the reproduced original fusion baselines. However, the higher AUC did not translate into uniformly higher threshold-dependent metrics; the extension achieved high specificity and precision but lower sensitivity, reflecting a more conservative operating point compared with LG-CAFN. This pattern indicates that the extension improves overall ranking ability rather than providing uniformly better classification at any single decision threshold.

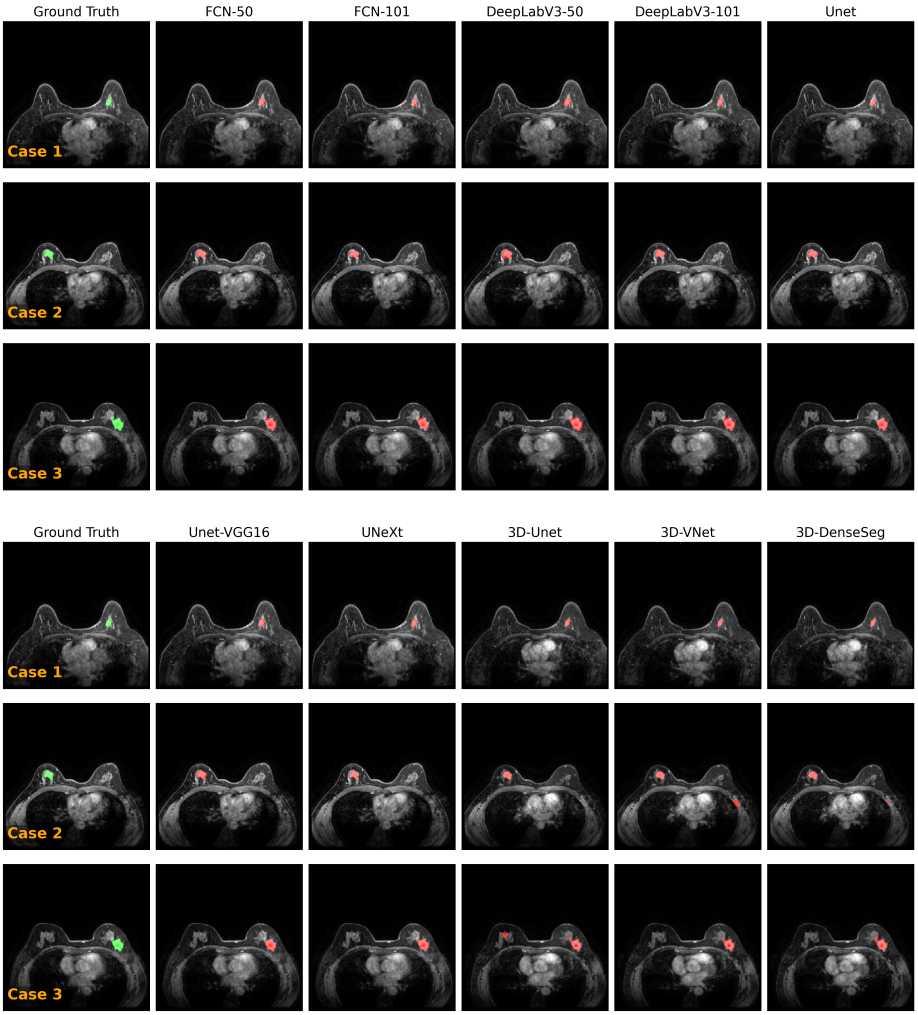


**Fig. 1.** Qualitative segmentation results on selected test cases. Upper: 2D models; lower: additional 2D and 3D models. Green: ground truth; red: predictions.

**Table 2.** Original and reproduced classification AUC on BreastDM.

| | | Exp-1 AUC | | Exp-2 AUC | |
|---|---|---|---|---|---|
| Type | Model | Orig. | Repr. | Orig. | Repr. |
| 2D | VGG16 | 0.8453 | 0.7423 | 0.8636 | 0.7773 |
| | VGG19 | 0.8655 | 0.8103 | 0.8533 | 0.7563 |
| | ResNet50 | 0.8786 | 0.8560 | 0.8576 | 0.8360 |
| | ResNet101 | 0.8681 | 0.7621 | 0.8456 | **0.8640** |
| | DenseNet121 | 0.8321 | 0.8372 | 0.8512 | 0.8087 |
| | DenseNet169 | 0.8690 | 0.7947 | 0.8780 | 0.8193 |
| | SENet50 | 0.8757 | 0.8177 | 0.8555 | 0.8316 |
| | SENet101 | 0.8632 | 0.7415 | 0.8529 | 0.8508 |
| | ViT-6 | 0.8755 | **0.8602** | 0.8555 | 0.7896 |
| | ViT-7 | **0.8811** | 0.8000 | **0.8887** | 0.7919 |
| 3D | 3DResNet18 | 0.8612 | 0.8113 | 0.8661 | 0.8366 |
| | 3DResNet50 | 0.8650 | **0.8582** | 0.8400 | 0.7994 |
| | 3DResNet101 | 0.8329 | 0.7983 | 0.8276 | 0.7648 |
| | 3DResNeXt101 | **0.9029** | 0.8219 | **0.8927** | 0.8443 |
| | 3DShuffleNet | 0.8877 | 0.7976 | 0.8599 | 0.8315 |
| | 3DShuffleNetV2 | 0.8834 | 0.8346 | 0.8722 | 0.8554 |
| | 3DMobileNet | 0.8832 | 0.8507 | 0.8726 | 0.8585 |
| | 3DMobileNetV2 | 0.8801 | 0.8388 | 0.8923 | **0.8634** |

**Table 3.** Original and reproduced fusion AUC (SENet + ViT branches).

| | Exp-1 | | Exp-2 | |
|---|---|---|---|---|
| Strategy | Orig. | Repr. | Orig. | Repr. |
| Feature-level | 0.8811 | 0.8313 | 0.8657 | **0.8314** |
| Decision-level | 0.8908 | 0.8272 | 0.8495 | 0.8242 |
| LG-CAFN | **0.9154** | **0.8847** | **0.8826** | **0.8314** |

**Table 4.** 3D ensemble and 3D–Transformer probability-level fusion results. Models selected by validation-set AUC only (see Section 2.3).

| Exp. | Method | Sens. | Spec. | Acc. | Prec. | AUC |
|---|---|---|---|---|---|---|
| Exp-1 | 3D Ens. (avg) | 82.83 | 78.95 | 81.84 | 91.97 | 0.8747 |
| | 3D Ens. (wt.) | 83.13 | 78.95 | 82.06 | 92.00 | 0.8751 |
| | 3D Ens.+ViT (avg) | 69.88 | 92.11 | 75.56 | 96.27 | 0.9184 |
| | 3D Ens.+ViT (wt.) | 67.77 | 92.11 | 73.99 | 96.15 | **0.9187** |
| Exp-2 | 3D Ens. (avg) | 53.63 | 89.47 | 63.77 | 92.81 | 0.8689 |
| | 3D Ens. (wt.) | 59.86 | 89.47 | 68.24 | 93.51 | 0.8687 |
| | 3D Ens.+ViT (avg) | 83.74 | 81.58 | 83.13 | 92.02 | **0.8929** |
| | 3D Ens.+ViT (wt.) | 76.47 | 85.09 | 78.91 | 92.86 | 0.8896 |

### 3.4 Classification Visualization

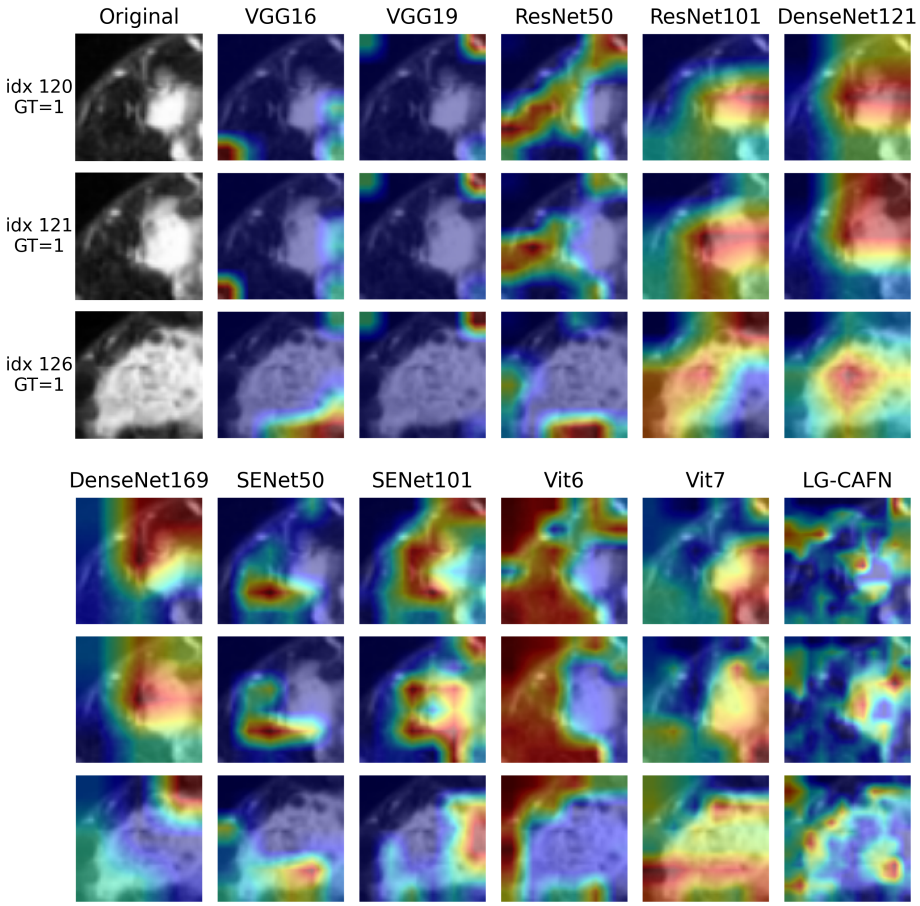


**Fig. 2.** Classification visualization on correctly classified malignant test cases. CNN models: Grad-CAM; ViT models: attention rollout; LG-CAFN: Grad-CAM on fused features.

Figure 2 shows classification heatmaps on selected malignant test cases. CNN-based models produced localized Grad-CAM responses, while ViT-based models showed more distributed attention patterns. LG-CAFN exhibited a multi-focal response, consistent with the interpretation that it combines local CNN and global Transformer features.

# 4 Discussion

## 4.1 Comparison with Original Study and Reproducibility Challenges

This study reproduced several core trends from the original BreastDM paper: 2D segmentation models outperformed 3D models; benign/malignant classification was feasible; subtraction sequences did not consistently improve classification; and LG-CAFN was a strong fusion baseline. Exact metric values differed, likely due to preprocessing, implementation choices, and checkpoint selection.

The main reproducibility challenges were: (1) inconsistent data-loading assumptions across released scripts (e.g., RGB loading vs. multi-channel .npy inputs); (2) misalignment between paper-reported metrics and training-log metrics, requiring recomputation from TP/FP/FN/TN; (3) the need for a unified checkpoint, threshold, and test-set evaluation protocol; (4) adaptation of pretrained models for multi-channel inputs; and (5) the absence of a single end-to-end pipeline for some experiments, requiring partial reconstruction.

## 4.2 2D, 3D, and Fusion Behavior

The different behaviors of 2D, 3D, and fusion models suggest that DCE-MRI information usage depends on both the task and the architecture. In classification, 3D CNNs exploited sequence/depth information and showed competitive AUC. In segmentation, 3D models did not outperform 2D models, likely due to fewer 3D samples and sparser mask supervision. The 3D–Transformer fusion extension improved AUC in both settings by combining 3D enhancement patterns with ViT global representations at the probability level, without additional training. However, this advantage was mainly reflected in threshold-independent discrimination; sensitivity, specificity, and accuracy still depend on the chosen operating point. This study has several limitations: all experiments used only the BreastDM dataset without external validation, some experiments were partially reconstructed from released components, and the 3D–Transformer fusion is a post-hoc strategy rather than an end-to-end trainable architecture.

# 5 Conclusion

This paper presented a systematic reproduction and reproducibility analysis of the BreastDM benchmarks, covering 2D/3D segmentation, 2D/3D classification, and fusion strategies. The main reported trends were largely reproduced, and we documented practical reproducibility challenges. As an exploratory extension, a lightweight 3D–Transformer probability-level fusion achieved competitive AUC (0.9187 in Exp-1, 0.8929 in Exp-2) without training a new backbone, suggesting that 3D CNN ensemble outputs and ViT branch outputs may provide complementary information for DCE-MRI classification.

## References


1. Bosma, J.S., Peeters, D., Alves, N., Saha, A., Saghir, Z., Jacobs, C., Huisman, H.: Reproducibility of training deep learning models for medical image analysis. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 227, pp. 1269–1287. PMLR (2024)
2. Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepah, N.M., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Serdyuk, D., Arbel, T., Pal, C., Varoquaux, G., Vincent, P.: Accounting for variance in machine learning benchmarks. In: Proceedings of Machine Learning and Systems. vol. 3, pp. 747–769 (2021)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. pp. 424–432. Springer (2016)
5. DeSantis, C.E., Ma, J., Gaudet, M.M., Newman, L.A., Miller, K.D., Goding Sauer, A., Jemal, A., Siegel, R.L.: Breast cancer statistics, 2019. CA: A Cancer Journal for Clinicians **69**(6), 438–451 (2019). https://doi.org/10.3322/caac.21583
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
10. Karahaliou, A., Vassiou, K., Arikidis, N.S., Skiadopoulos, S., Kanavou, T., Costaridou, L.: Assessing heterogeneity of lesion enhancement kinetics in dynamic contrast-enhanced MRI for breast cancer diagnosis. The British Journal of Radiology **83**(988), 296–309 (2010). https://doi.org/10.1259/bjr/50743919
11. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571 (2016)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer (2015)

15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
16. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71**(3), 209–249 (2021). https://doi.org/10.3322/caac.21660
17. Zhao, X., Liao, Y., Xie, J., He, X., Zhang, S., Wang, G., Fang, J., Lu, H., Yu, J.: BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. Computers in Biology and Medicine **164**, 107255 (2023). https://doi.org/10.1016/j.compbiomed.2023.107255