Breast Tumor Segmentation in Dynamic Contrast-Enhanced MRI via Multi-Staged Training and Deep Ensembling of a Large Kernel MedNeXt

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Breast cancer is a major health concern [8], and Magnetic Resonance Imaging (MRI) plays an important role in its assessment, preoperative staging and treatment [14, 7]. T1-weighted dynamic contrast-enhanced (DCE) MRI can highlight tumor vascularity by using contrast agents, aiding in the localization of tumor lesions. Accurate segmentation of tumor boundaries on these images is clinically valuable, as it enables quantitative evaluation of tumor size, shape, and volume over time [16, 1]. High-quality tumor segmentation further facilitates advanced analyses such radiomics feature extraction [17] for other downstream tasks [2, 4], including pCR assessment [13] and the characterization of tumor types [5].

Manual biomedical image segmentation is highly time-consuming, and tends to suffer from inter- and intra-annotator variability sbject to the level of experience of a radiologist. Deep learning based automated segmentation methods have proven to be reliable in addressing the stated problems [9, 15, 12, 11], with some limitations in terms of data, and or architectural constraints. To achieve robust and generalizable performances, highly parameterized deep learning models need to be trained with sufficient enough labeled data. Readily available large-scale labeled medical imaging data is lacking, particularly for dynamic contrast-enhanced MRI segmentation.

We address the challenge of the segmentation of breast tumor lesions in multi-contrast MRI, using MedNeXt and a multi-staged training strategy of improving receptive field, and loss optimization based deep ensembling. Our segmentation method is motivated by the need to capture both the subtle enhancement patterns of tumors across multiple post-contrast time points and the broader breast tissue context, which larger receptive fields naturally accommodate. To effectively utilize large kernels without overfitting, we adopt a two-stage training strategy: we first train a MedNeXt model with the conventional $3 \times 3 \times 3$ kernel sizes, and then expand to $5 \times 5 \times 5$ kernel sizes via trilinear interpolation [15].

2 Methodology

2.1 Dataset and Preprocessing

We used a multicenter dataset of 1506 cases from over 20 institutions for training, and a held out 58 cases for testing [6]. Each case in the dataset includes a series of T1-weighted DCE-MRI volumes acquired at multiple time points: one pre-contrast and up to five post-contrast phases. For our experiments, we selected the pre-contrast image and the first two post-contrast images of each case.

The dataset comprises both unilateral and bilateral breast DCE-MRI scans. Data preprocessing, training and inference were done using the standard nnU-Net pipeline [9]. During training, we adopted nnU-Net's patch-based sampling strategy with a fixed input size of $128 \times 128 \times 128$ voxels.

To ensure sufficient exposure to tumor regions across both scan types, more than one-third of the patches in each batch were enforced to contain at least one foreground voxel. At inference time, segmentation was performed using overlapping sliding-window patches, ensuring full-volume coverage regardless of laterality.

2.2 MedNeXt Architecture

MedNeXt is a fully ConvNeXt [10] encoder-decoder U-shaped network for biomedical image segmentation [15]. Its inverted bottleneck design in the up and downsampling layers, and the compound scaling of depth, width and receptive field, makes it a highly capable segmentation method. It is further transformer-inspired in its scaling approach, and the use of large-kernel sizes to approximate attention. The added inductive bias provides the benefits of both convolutional- and transformer-based approaches, in capturing short and long range dependencies respectively.

$$M^{(5)} = \textit{UpKern}(M^{(3)}, \text{size} = 5) \tag{1}$$

The approximation of attention via larger kernel sizes of $5 \times 5 \times 5$ instead of the conventional $3 \times 3 \times 3$ sizes is achieved by first pretraining a conventional MedNeXt (M^3), and trilinearly interpolating its convolutional kernels to an initialized large-kernel MedNeXt (M^5), using an algorithm called *UpKern* [15] in Equation 1. The performance saturation usually observed with increasingly large kernel sizes is mitigated [3].

2.3 Training Strategy

All networks were trained with deep supervision, stochastic gradient descent (SGD) optimization, and a cosine annealing learning rate schedule initialized at 1×10^{-4}, using an A100 NVIDIA GPU. For the base $M^3{}_{Base}$, training was conducted using a five-fold cross-validation split and optimized using Dice cross-entropy loss. Each fold was trained independently for 250 epochs. Following the completion of cross-validation, we identified the fold that achieved the highest mean Dice coefficient on its respective validation set, whose weights were used in the second stage of training.

In the subsequent stage, we employed the *UpKern* strategy 1 to resample the learned $M^3{}_{Base}$ convolutional kernel weights into $M^5{}_{Base}$ via trilinear interpolation. This approach enabled a smooth transition to a large-kernel configuration, thus expanding the effective receptive field of the network without introducing instability often associated with training large kernels from random initialization [3]. The newly initialized $M^5{}_{Base}$ was then fine-tuned for making use of the entire training set, with all other architectural and training settings held constant.

Additionally, we generated a second M^5 by applying the *UpKern* algorithm and fine-tuning the pretrained weights, forming $M^5{}_{Focal}$. In this stage, the network was optimized using a composite loss that combines Dice–cross-entropy and focal loss to better penalize small lesion segmentation errors and address class imbalance arising from large foreground–background differences:

$$\mathcal{L}_{\text{total}} = 0.25\,\mathcal{L}_{\text{Dice-CE}} + 0.75\,\mathcal{L}_{\text{Focal}} \tag{2}$$

where $\mathcal{L}_{\text{Dice-CE}}$ denotes the combined Dice and cross-entropy loss used in the first training stage, and $\mathcal{L}_{\text{Focal}}$ is the focal loss component that increases the weighting of hard-to-segment regions. Code is publicly available along with the implemented composite loss functions[1].

3 Results

Segmentation performance was evaluated on the held-out testing set. The results are summarized in Table 1. M^3 achieved a Dice score of 0.64 and a normalized Hausdorff Distance (NormHD) of 0.3. Upon applying the *UpKern* strategy and fine-tuning the best performing single model, $M^5{}_{Base}$ improved the Dice score to 0.66 and the NormHD to 0.29 (see Figure 1. Further ensembling with $M^5{}_{Focal}$ led to minimal increase in Dice to 0.67, and NormHD to 0.24. The reported baseline is a 5-fold nnU-Net ensemble by [6], also trained on the 1506 training dataset and evaluated on the test set.

[1] `https://github.com/toufiqmusah/caladan-mama-mia`

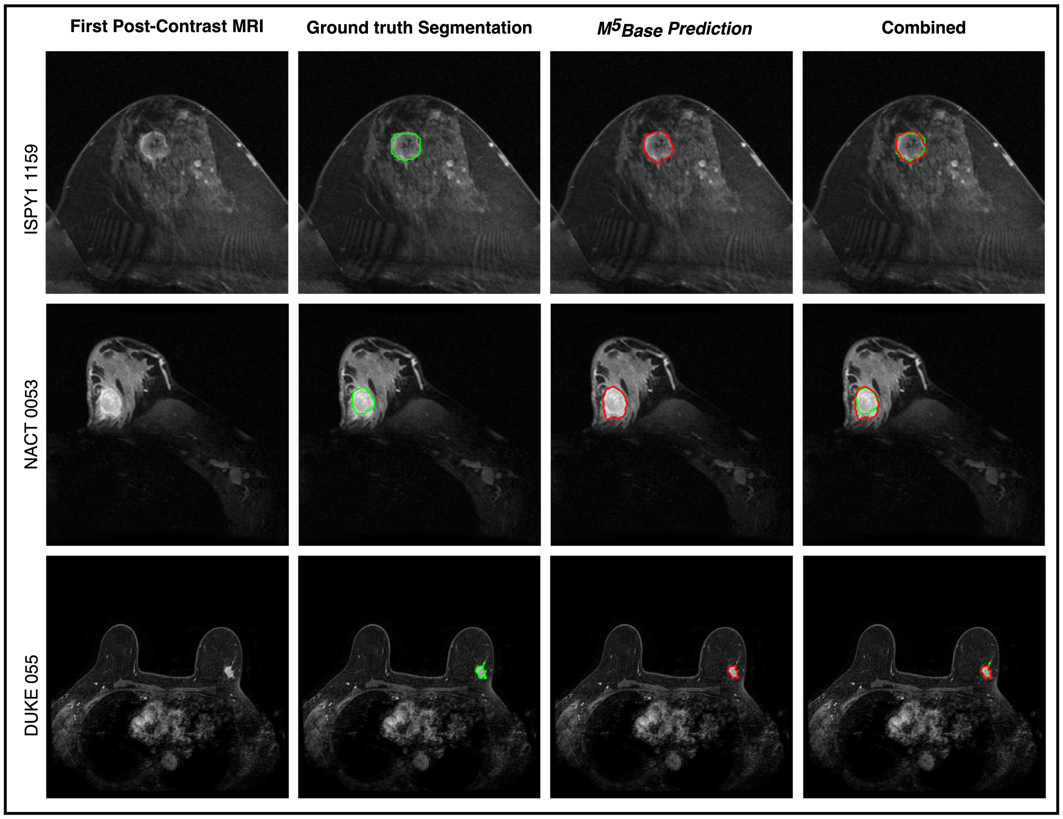

Figure 1: Qualitative segmentation performance of ${M^5}_{Base}$ on validation samples from various centers in in red, and ground truth in green.

Table 1: Performance comparison between methods. Metrics are reported as Dice Score and 95th-Percentile Normalized Hausdorff Distance. Total number of combined parameters are denoted in in millions.

Architecture	**Dice Score (↑)**	**NormHD (↓)**	**Parameters (↓) (M)**
nnU-Net (5-Folds) [6]	0.65	0.30	700
M^3 (5-Folds)	0.64	0.30	154
${M^5}_{Base}$	0.66	0.29	**32.1**
${M^5}_{Base}$ + ${M^5}_{Focal}$	**0.67**	**0.24**	64.2

The 5-fold M^3 ensemble's performance improved after applying the *UpKern* strategy to expand the receptive field. Fine-tuning the single best-performing fold into ${M^5}_{Base}$ led to a Dice score of 0.66 and a reduction in NormHD to 0.29. Further ensembling ${M^5}_{Base}$ with ${M^5}_{Focal}$, which was trained using Dice-cross-entropy and focal loss, resulted in a Dice score of 0.67 and a NormHD of 0.24. These results validate the hypothesis that larger receptive fields improve segmentation performance by capturing broader anatomical context. Notably, both ${M^5}_{Base}$ and ${M^5}_{Base}$ + ${M^5}_{Focal}$ outperformed the nnU-Net baseline [6] and achieved a lower NormHD, despite having substantially fewer parameters per model (32.1M vs. 140M per model instance) and using only two models in the final ensemble compared to the five-model nnU-Net baseline. Architectural efficiency and targeted loss function strategies can deliver improved performance while reducing computational requirements. Our study is limited in scope to the evaluation of the proposed *UpKern* strategy within the MedNeXt architecture; we did not test its generalizability across other network families. While ensemble performance was reported, we did not isolate and report the standalone performance of individual models within the ensemble, which could provide further insights into complementarity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction summarizes the contributions of a two stage training strategt and ensembling of large-kernel MedNeXt.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of not reporting on individual ensemble component gains.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper demonstrates increased segmentation performance with UpKern and loss-optimized ensembles.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All information needed for reproducibility is provided in the methodology.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data is openly available [6]. And the code is added as a footnote in section 2.3

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The training and test splits are specified in the methodology.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Not enough experimental repetitions or folds were conducted to produce statistically significant results or perform formal statistical tests. Reported performance metrics therefore reflect central tendencies from the available runs rather than statistical estimates.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute used is specified in the methodology.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The code of ethics was reviewed, and this research conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The positive impact of reliable biomedical image segmentation is discussed in the introduction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No new model or dataset with potential for misuse was released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Data and code sources are well cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced are the loss composite loss functions, available in the provided code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing was done for this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects research was conducted

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: No large language model was used in the design, analysis, or generation of results; only standard writing assistance was used.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

References

[1] Khadijeh Askaripour and Arkadiusz Zak. Breast mri segmentation by deep learning: Key gaps and challenges. *IEEE Access*, 11:117935–117946, 2023.

[2] Nathaniel M Braman, Maryam Etesami, Prateek Prasanna, Christina Dubchuk, Hannah Gilmore, Pallavi Tiwari, Donna Plecha, and Anant Madabhushi. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri. *Breast Cancer Research*, 19:1–14, 2017.

[3] Honghao Chen, Xiangxiang Chu, Yongjian Ren, Xin Zhao, and Kaiqi Huang. Pelk: Parameter-efficient large kernel convnets with peripheral convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5557–5567, 2024.

[4] Kawtar Debbi, Paul Habert, Anaïs Grob, Anderson Loundou, Pascale Siles, Axel Bartoli, and Alexis Jacquier. Radiomics model to classify mammary masses using breast dce-mri compared to the bi-rads classification performance. *Insights into Imaging*, 14(1):64, 2023.

[5] Khuram Faraz, Grégoire Dauce, Amine Bouhamama, Benjamin Leporq, Hajime Sasaki, Yoshitaka Bito, Olivier Beuf, and Frank Pilleul. Characterization of breast tumors from mr images using radiomics and machine learning approaches. *Journal of Personalized Medicine*, 13(7):1062, 2023.

[6] Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier Del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. *Scientific data*, 12(1):453, 2025.

[7] Briete Goorts, Kelly MA Dreuning, Janneke B Houwers, Loes FS Kooreman, Evert-Jan G Boerma, Ritse M Mann, Marc BI Lobbes, and Marjolein L Smidt. Mri-based response patterns during neoadjuvant chemotherapy can predict pathological (complete) response in patients with breast cancer. *Breast Cancer Research*, 20(1):34, 2018.

[8] Gabriel N Hortobagyi, Jaime de la Garza Salazar, Kathleen Pritchard, Dino Amadori, Renate Haidinger, Clifford A Hudis, Hussein Khaled, Mei-Ching Liu, Miguel Martin, Moise Namer, et al. The global breast cancer burden: variations in epidemiology and survival. *Clinical breast cancer*, 6(5):391–401, 2005.

[9] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[11] Toufiq Musah. Large kernel mednext for breast tumor segmentation and self-normalizing network for pcr classification in magnetic resonance images. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pages 72–80. Springer, 2025.

[12] Toufiq Musah, Chinasa Kalaiwo, Maimoona Akram, Ubaida Napari Abdulai, Maruf Adewole, Farouk Dako, Adaobi Chiazor Emegoakor, Udunna C Anazodo, Prince Ebenezer Adjei, and Confidence Raymond. Towards trustworthy breast tumor segmentation in ultrasound using monte carlo dropout and deep ensembles for epistemic uncertainty estimation. *arXiv preprint arXiv:2508.17768*, 2025.

[13] JPM O'Donnell, SA Gasior, MG Davey, E O'Malley, AJ Lowery, J McGarry, AM O'Connell, MJ Kerin, and P McCarthy. The accuracy of breast mri radiomic methodologies in predicting pathological complete response to neoadjuvant chemotherapy: A systematic review and network meta-analysis. *European journal of radiology*, 157:110561, 2022.

[14] Selvi Radhakrishna, S Agarwal, Purvish M Parikh, K Kaur, Shikha Panwar, Shelly Sharma, Ashish Dey, KK Saxena, Madhavi Chandra, and Seema Sud. Role of magnetic resonance imaging in breast cancer management. *South Asian journal of cancer*, 7(2):69, 2018.

[15] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.

[16] Guoliang Shao, Linyin Fan, Juan Zhang, Gang Dai, and Tieming Xie. Association of dw/dce-mri features with prognostic factors in breast cancer. *The International journal of biological markers*, 32(1):118–125, 2017.

[17] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.