An Integrated Framework Using MedSAM-Driven Segmentation and Neural Network-Based Classifiers for Breast Cancer Detection on Mammograms

Zhuoxin Guo, Dongfang Cai, Harrison Dungey
Affiliation: Queen's University, Kingston, ON

Abstract

Breast cancer currently remains a leading cause of cancer-related death globally. Using mammography as a screening tool to detect breast tumours early can help reduce breast cancer mortality for patients. However, the traditional imaging method in assessing BI-RADS as a breast image findings system has limitations in diagnostic accuracy. This study combines the MedSAM model with different neural network-based classifiers to improve lesion segmentation and malignancy prediction in mammography using the CBIS-DDSM dataset. Using the MedSAM model, segmentation achieved mean Dice Similarity Coefficients (DSC) of 0.67 for calcifications and 0.72 for masses, with IoUs of 0.54 and 0.59. Significant correlations were observed between the longest diameter of lesions and segmentation accuracy. The classification was performed using segmented masks and whole images, utilizing EfficientNet and MedSAM in conjunction with perceptron architectures. EfficientNet achieved an accuracy of 0.56 when using segmented masks and improved to 0.69 when using the whole image as input. Combining EfficientNet with radiomics features outperformed all other methods, achieving an accuracy of 0.76 with balanced precision (0.75) and recall (0.78). MedSAM embeddings were further tested with perceptrons, where fine-tuning all weights yielded an accuracy of 0.64, and freezing the encoder while fine-tuning the decoder improved performance to 0.73. All models showed higher accuracy than BI-RADS. These results highlight the value of neural network-based classification models to improve breast cancer diagnosis on mammography and support future clinical use.

Background

Imaging modalities like mammography have decreased breast cancer mortality by 20% and remain the gold standard for cancer screening[1]. Nonetheless, the current method to obtain regions of interest (ROI) which is a specific area focused on detected breast masses is critical for accurate lesion analysis. Reliance on mammography for determining BI-RADS classification is hindered by the accuracy of mammographic interpretation; particularly dense tissue can absorb more radiation than normal fatty breast tissue, leading to decreased sensitivity and potential diagnostic inaccuracies[2]. These challenges may lead to a delay in diagnosing malignant lesions in mammograms which affects patient outcomes.

Recent studies have demonstrated the role of AI and deep learning in overcoming these limitations for the interpretation of mammograms as a valuable tool to assist radiologists. A new deep learning architecture with asymmetrical encoder and decoder blocks, AUNet, has successfully segmented mass regions in challenging complex breast tissue environments[3]. However, AUNet has not been benchmarked against expert radiologists, and its clinical reliability remains under evaluation. Feature extraction and concatenation obtained from pre-trained CNNs have also emerged as a strong classifier with high accuracy in distinguishing between benign and malignant lesions[4]. Their study focuses merely on the tumour region itself while missing the diagnostic value from surrounding tissues that may leak some potential pattern for malignancy prediction. Our work aims to improve breast cancer detection by leveraging MedSAM[5], a fine-tuned segmentation model, combined with advanced classification methods such as EfficientNet and perceptron (MLP). By incorporating features from both tumour regions and surrounding tissue, as well as testing different configurations of MedSAM, our approach provides a more comprehensive and accurate framework for malignancy prediction.

Hypothesis

We hypothesize that integrating the MedSAM model with a neutral network-based classifier will improve malignancy prediction in mammography images compared to traditional BI-RADS-based methods.

Methods

Dataset and Materials description

The Curated Breast Imaging Subset of the Digital Database of Screening Mammography (CBIS-DDSM) is an updated and standardized version of the original DDSM dataset[6]. This count (10,239 images from 1,566 patient cases) refers specifically to the breast imaging subset in the CBIS-DDSM dataset, excluding other data categories[7]. The improvements over the original DDSM include decompressed images in DICOM format, removal of questionable mass cases, and more accurate region of interest (ROI) annotations. The dataset also updated precision segmentation for mass margins by using a lesion segmentation algorithm, and standardized train/test splits for training models[6]. The metadata includes BI-RADS assessment, pathology outcomes, and mammographic views. The dataset is public and can be downloaded via The Cancer Imaging Archive.

All experiments were run using Python 3.9 and PyTorch 2.0 with GPU acceleration support for fast training. We used the MedSAM model, which is a variation of SAM for medical image segmentation, to extract embeddings and produce segmentation masks. We performed classification using MedSAM with a Multilayer Perceptron (MLP), and an EfficientNet-B0-based CNN, which were both trained on the CBIS-DDSM dataset.

Data analysis

Segmentation of calcifications and masses

The CC and MLO images in DICOM format were normalized and resized to 1024 × 1024 pixels to ensure compatibility with the MedSAM model. Binary masks were also resized to match the model's input resolution. The contours of the masks were identified to compute bounding boxes, which were expanded by 1 cm to simulate manually defined bounding boxes and ensure adequate coverage of the target region.

The pre-trained MedSAM model was then used for segmentation inference in three iterations. In each iteration, the model generated a refined mask based on the bounding box or prior segmentation results. Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) were calculated for each segmentation result to evaluate performance. Among the three iterations, the mask with the highest DSC was selected as the final segmentation output.

Classification of masses using MedSAM and Perceptron

To ensure balanced class distribution, the data was first split into training, validation, and test sets using stratified sampling. Specifically, 20% of the labelled data was reserved as the test set, while the remaining 80% was further divided into 80% for training and 20% for validation.

We explored three approaches for leveraging the MedSAM model to classify masses into benign and malignant categories. In the first approach, we used the pre-trained MedSAM model to generate feature embeddings directly from the input images. These embeddings were passed into a three-layer perceptron for classification. In the second approach, we fine-tuned only the decoder weights of the MedSAM model, freezing the encoder to preserve its pre-trained knowledge. This allowed the decoder to adapt to mammography-specific features while maintaining efficient training. The updated embeddings were subsequently fed into the same three-layer perceptron. In the third approach, we fine-tuned all weights of the MedSAM model, including both the encoder and decoder, enabling a full adaptation of MedSAM to the mammography dataset.

For all three methods, the embeddings were classified using a three-layer perceptron consisting of two hidden layers with 100 and 50 units, respectively, each followed by a ReLU activation. Dropout with a rate of 50% was applied after each hidden layer to prevent overfitting, and batch normalization was used to stabilize the training process. The final output layer mapped the features to two classes using a softmax activation function.

The models were trained using the cross-entropy loss function and optimized with the Adam optimizer, with a batch size of 64. At the end of each epoch, training and validation accuracies were recorded to monitor performance. The final evaluation of the test set involved calculating metrics such as accuracy, precision, recall, and F1-score to compare the effectiveness of the three approaches.

Radiomics Feature

We used the segmentation results from MedSAM to extract two regions around the detected mass: a 3 cm diameter region immediately surrounding the mass and a larger 6–9 cm diameter region representing the surrounding tissue. Gabor filters were applied to both regions to extract texture features, capturing patterns and structural variations. To ensure reliable comparisons, the feature vectors were normalized before computing the cosine similarity, reducing the impact of lesion size differences. Any portions of the extracted regions extending beyond the breast area were removed to ensure the analysis focused solely on relevant breast tissue.

Classification of masses using EfficientNet

For training of all the EfficientNet models, the CC and MLO images in DICOM format were first read and normalized, then converted to a PNG format, without information loss, for compatibility with the CNN input. To generate the segmentation-masked inputs, the previously obtained binary segmentation masks were applied to the image, retaining only the values where the binary mask has a positive value. The resulting masked image was cropped to a 224 x 224 square, with the masked section centered, maintaining a 20-pixel padding around the edges of the image. For both the masked and whole images, each was assigned benign or malignant based on the corresponding pathology metadata. These images were then split into training, test, and validation sets using the same method as the MedSAM perceptron. To improve the generalization of the model and account for the relatively small size of the dataset, data augmentation was applied, consisting of random horizontal rotations as well as random horizontal flips.

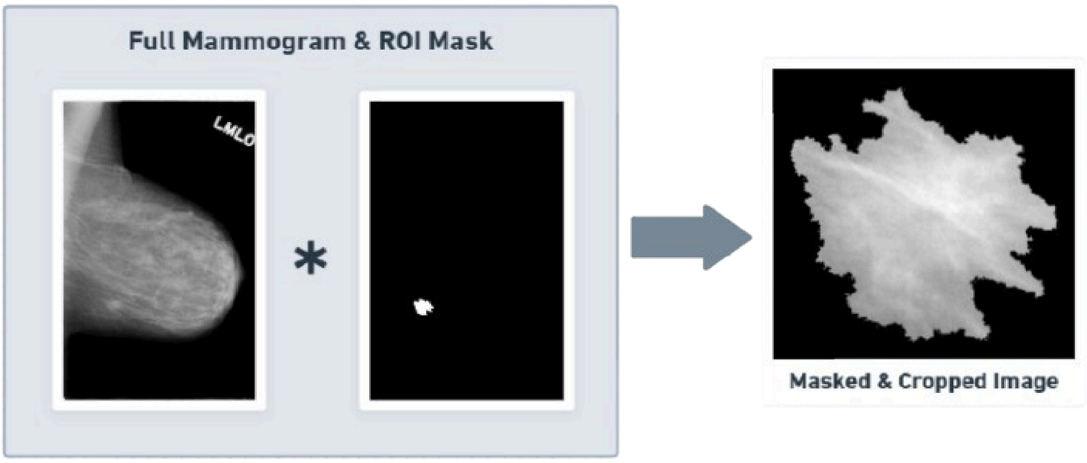


Figure 1. Diagram of the masking process, consisting of multiplying the full mammogram with the ROI mask to obtain a 224x224 cropped and padded image, masked by the segmentation.

A custom architecture for transfer learning with EfficientNet was used. It consisted of the EfficientNet-B0 model, pre-trained on the ImageNet dataset[8], with the final output neurons removed. This was followed by a global average pooling layer, two fully connected layers with ReLU activations, and ending in a single output node with sigmoid activation. This transfer learning approach allowed for the utilization of EfficientNet's pretrained capacity for the classification task. The model was trained with a batch size of 32 over 200 epochs. The model performance was again evaluated using accuracy, precision, recall, F1 scores, and a confusion matrix.

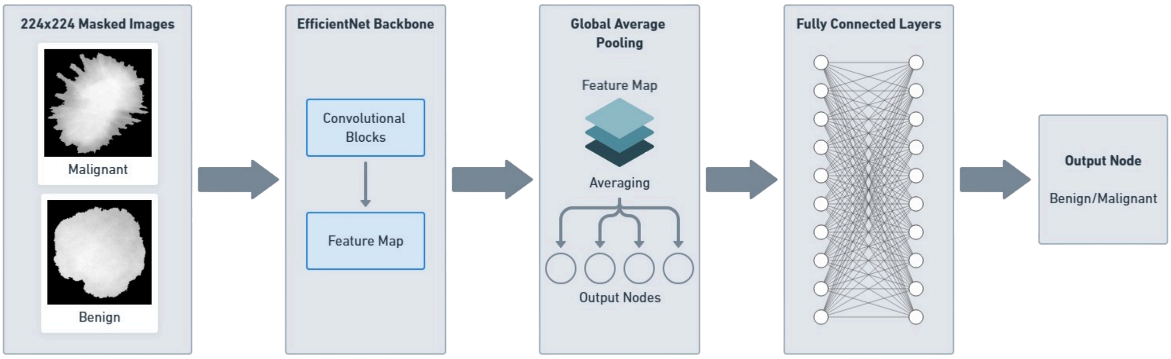
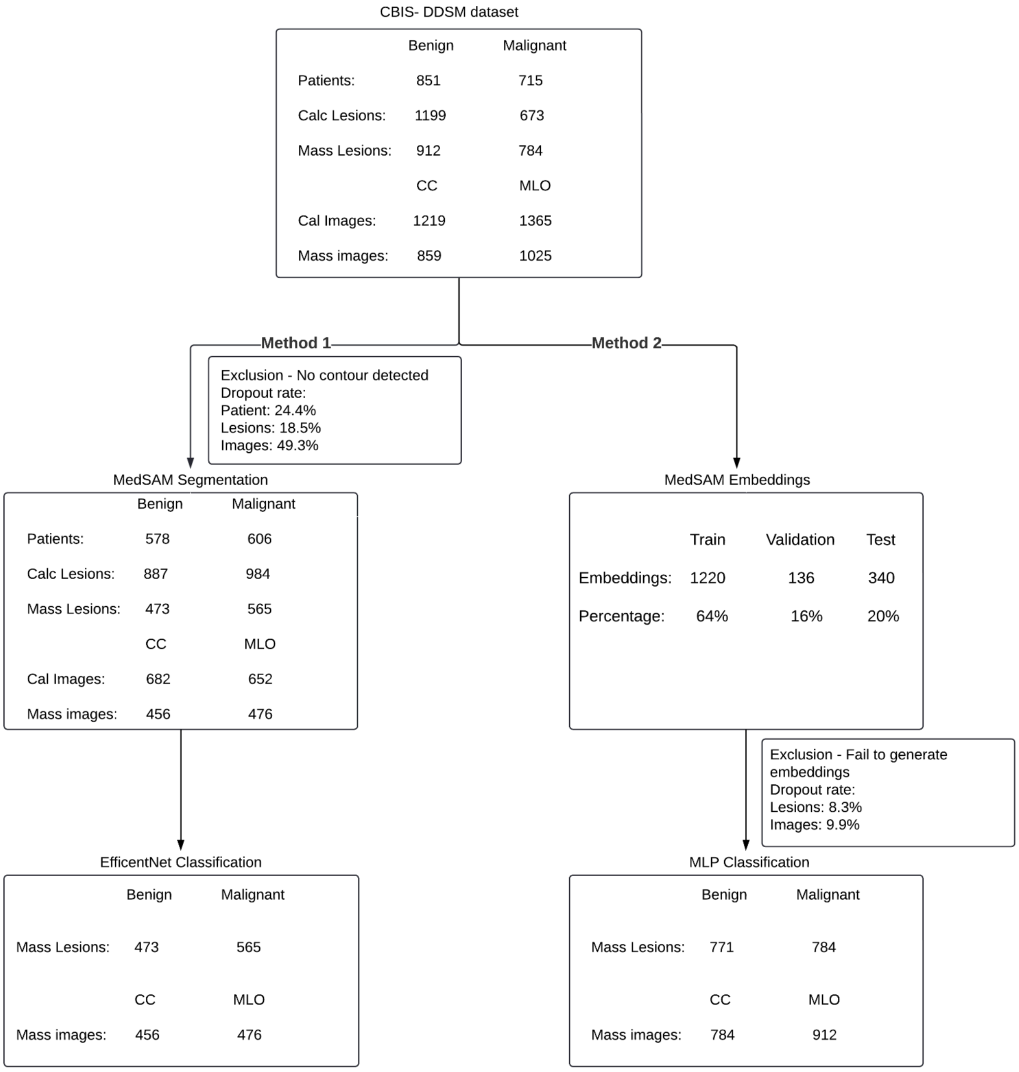


Figure 2. Architecture of the proposed custom CNN architecture consisting of segmented and masked input images, EfficientNet-B0 pre-trained on ImageNet with output nodes removed, a global average pooling layer, two fully connected layers, and an output node, arranged in sequence.



CBIS- DDSM dataset

	Benign	Malignant
Patients:	851	715
Calc Lesions:	1199	673
Mass Lesions:	912	784
	CC	MLO
Cal Images:	1219	1365
Mass images:	859	1025

Exclusion - No contour detected
Dropout rate:
Patient: 24.4%
Lesions: 18.5%
Images: 49.3%

MedSAM Segmentation

	Benign	Malignant
Patients:	578	606
Calc Lesions:	887	984
Mass Lesions:	473	565
	CC	MLO
Cal Images:	682	652
Mass images:	456	476

MedSAM Embeddings

	Train	Validation	Test
Embeddings:	1220	136	340
Percentage:	64%	16%	20%

Exclusion - Fail to generate embeddings
Dropout rate:
Lesions: 8.3%
Images: 9.9%

EfficentNet Classification

	Benign	Malignant
Mass Lesions:	473	565
	CC	MLO
Mass images:	456	476

MLP Classification

	Benign	Malignant
Mass Lesions:	771	784
	CC	MLO
Mass images:	784	912

Figure 3. Consort diagram illustrating the flow of participants through the two classification models

Results

MedSAM segmentation

The average DSC and IoU values across iterations are shown in Figure 4. The first inference achieved the highest performance, with DSC at 0.77 and IoU at 0.65. A positive correlation was observed between DSC and ground-truth lesion diameter (r = 0.125, p <0.0001), as shown in Figure 5. An example of MedSAM segmentation results is demonstrated in Figure 6.

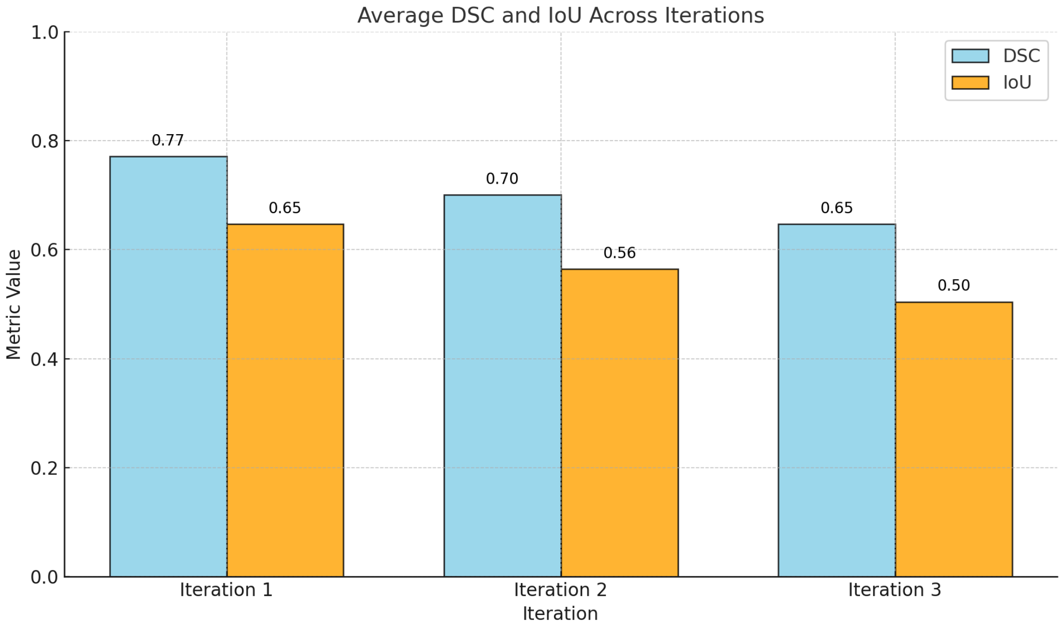


Figure 4. Average DSC and IoU values across iterations for the MedSAM segmentation model. The first iteration achieved the highest performance (DSC = 0.77, IoU = 0.65), while subsequent iterations showed a decline in both metrics.

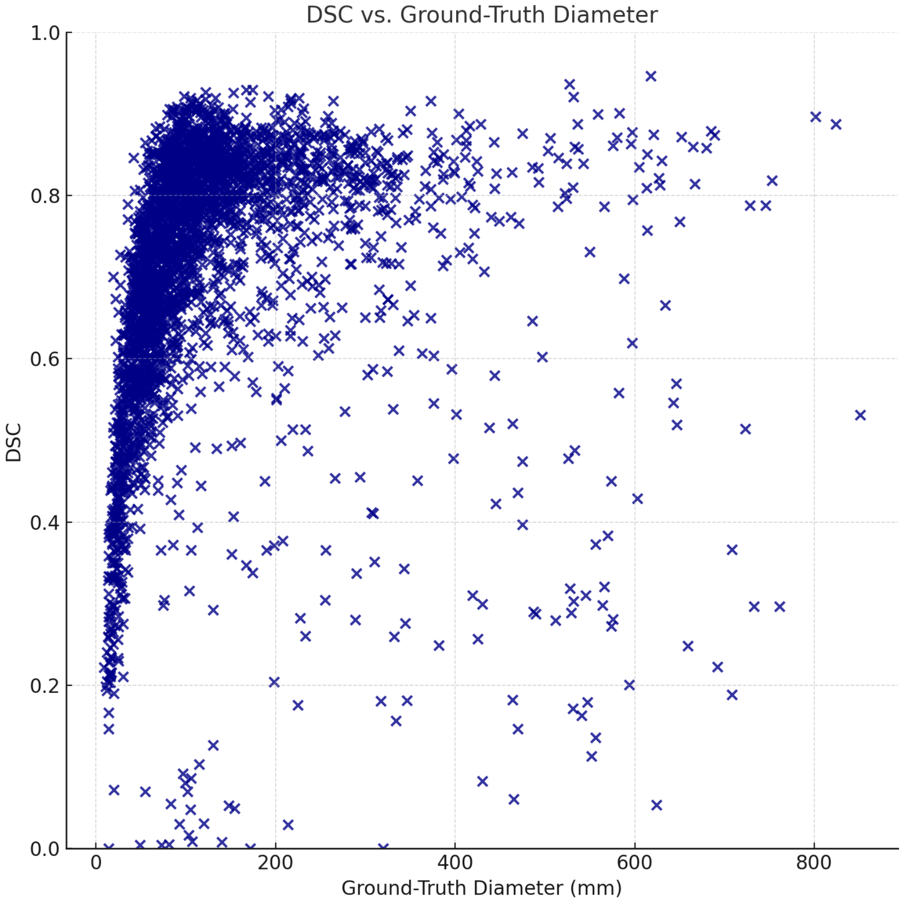


Figure 5. Scatter plot of DSC versus ground-truth lesion diameter. A positive correlation is observed (r = 0.125, p <0.0001), indicating better segmentation performance for larger lesions.

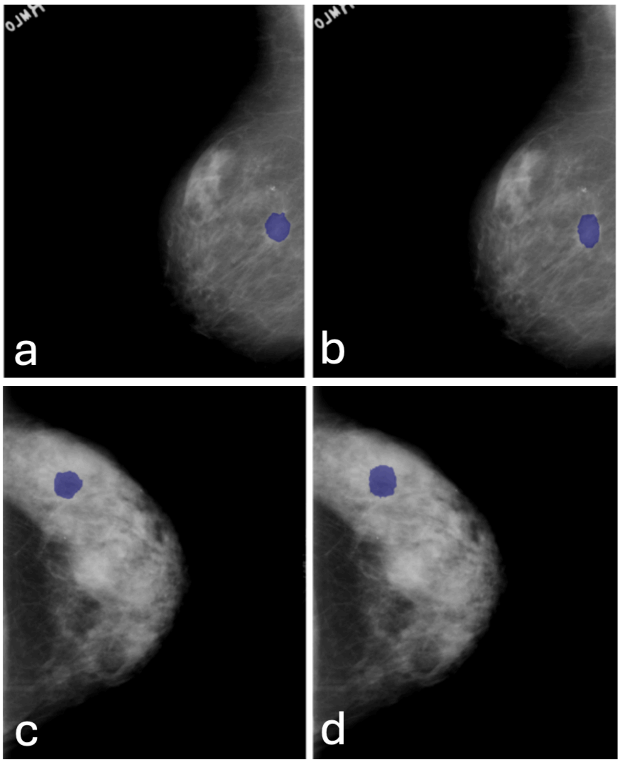


Figure 6. The ground truth and MedSAM segmentations overlaid on mammography images. The MedSAM segmentation of calcifications (blue area, b) demonstrates high overlap with the ground truth (blue area, a) on MLO images, achieving an MSC of 0.91 and an IoU of 0.83. The blue areas in c and d represent the ground truth of a breast mass and the corresponding MedSAM segmentation on CC images, respectively, with an MSC of 0.80 and an IoU of 0.67.

EfficientNet and MedSAM perceptron classification

The performance of EfficientNet using whole images, segmentation-masked images, and whole images combined with the extracted radiomics data has been evaluated on the subset of images containing masses. Using only the masked images with a DSC greater than 0.6, the model obtained an accuracy of 56%. However, the accuracy using whole images, and whole images combined with the radiomics data was 69% and 76% respectively. On this best-performing EfficientNet and radiomics model, we also observed that the performance is balanced between precision and recall, with slightly lower scores for the benign class.

Similar evaluations were conducted on the MedSAM perceptron models using only the whole images. The results indicate that the model without any fine-tuning achieved an accuracy of 63%, the model with all the weights retrained obtained an accuracy of 64%, and the model with only the decoder retrained achieved an accuracy of 73%.

The comparison of model performance, including precision, recall, and F1 scores, as well as BI-RADS performance, is shown in Table 1. The Confusion matrix for the EfficientNet and radiomics model, which obtained the highest accuracy, is displayed in Figure 7.

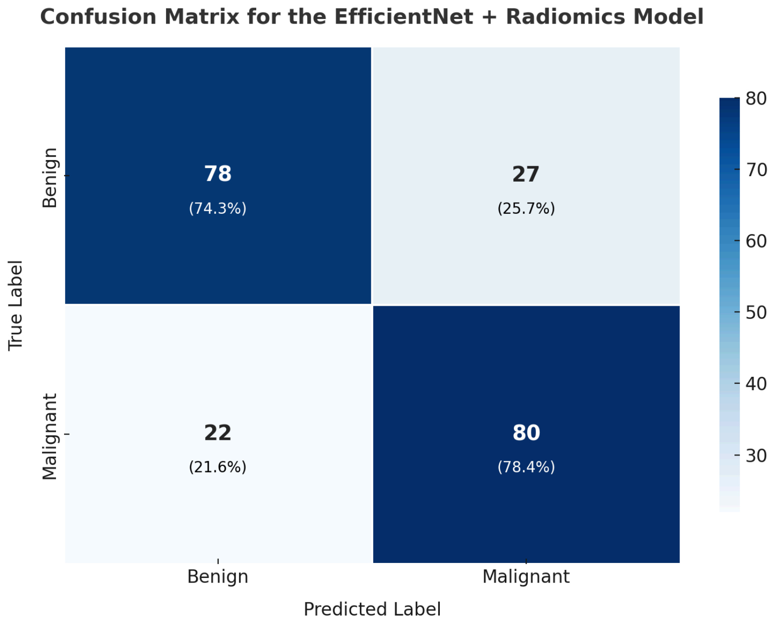
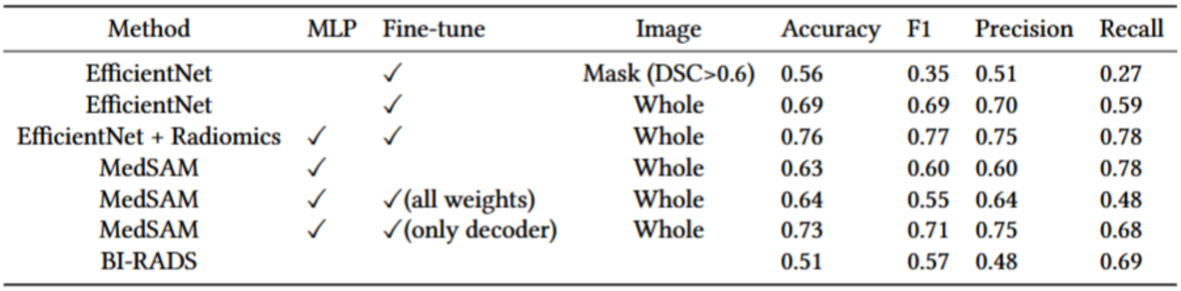


Figure 7. Confusion matrix for the images analyzed using the custom EfficientNet + Radiomics architecture. Each cell displays the number of classifications and percentages for the benign and malignant classes separately. Rows represent true labels and columns represent predicted labels.



Method	MLP	Fine-tune	Image	Accuracy	F1	Precision	Recall
EfficientNet		✓	Mask (DSC>0.6)	0.56	0.35	0.51	0.27
EfficientNet		✓	Whole	0.69	0.69	0.70	0.59
EfficientNet + Radiomics	✓	✓	Whole	0.76	0.77	0.75	0.78
MedSAM	✓		Whole	0.63	0.60	0.60	0.78
MedSAM	✓	✓(all weights)	Whole	0.64	0.55	0.64	0.48
MedSAM	✓	✓(only decoder)	Whole	0.73	0.71	0.75	0.68
BI-RADS				0.51	0.57	0.48	0.69

Table 1. Comparison of BI-RADS and various models in predicting malignancy on mammography.

* BI-RADS > 3 is interpreted as predicting "Malignant" while BI-RADS < 4 is interpreted as "Benign" in the current study.

Discussion

In this study, we evaluated the performance of the MedSAM segmentation model and various CNN-based classification approaches in predicting malignancy in mammography images. Our results demonstrated that MedSAM achieved accurate segmentation, particularly for larger lesions, while both EfficientNet and MedSAM outperformed traditional BI-RADS assessment, achieving an accuracy of up to 76%. These findings highlight the potential of integrating CNN-based methods to improve breast cancer diagnosis.

MedSAM Segmentation Performance

The MedSAM model showed strong performance in segmenting calcifications and breast masses, with higher accuracy for larger lesions, as smaller ones like calcifications are inherently harder to detect. Lesion size significantly influenced segmentation performance. Accurate segmentation provides precise lesion boundaries, enhancing classification accuracy and aiding radiologists in identifying subtle abnormalities. Despite overall strong results, improving performance on smaller lesions remains a challenge. Fine-tuning mammography-specific datasets may help address this.

Our findings also showed that a single inference achieves the best segmentation, with DSC and IoU declining in subsequent iterations. This suggests the initial segmentation effectively captures lesion boundaries, making further refinements unnecessary and potentially less accurate. Optimizing single-pass inference may thus be more efficient.

Classification Performance

We investigated the impact of combining segmentation, global feature extraction, and radiomic features for predicting malignancy in mammography images. Models using global features consistently outperformed those relying solely on segmentation masks, highlighting the importance of broader contextual information. Even mask-based models outperformed BI-RADS, demonstrating the robustness of our approach in extracting high-dimensional features for classification.

Adding radiomic features with Gabor filters further improved performance by capturing texture patterns in surrounding tissue. Malignant tumours often exhibit invasive textures, while benign ones retain normal tissue appearance. Comparing these patterns allowed the model to detect subtle differences, particularly in dense breasts where lesions are less visible. This novel approach effectively accounts for breast density variations and tumour invasion, significantly enhancing prediction accuracy. To address lesion size effects, we applied feature vector normalization during cosine similarity calculations, ensuring consistent comparisons across cases.

For MedSAM, fine-tuning only the decoder outperformed full-weight fine-tuning, likely because freezing the encoder preserves pre-trained global features while adapting the decoder for domain-specific refinements. EfficientNet outperformed MedSAM in classification, probably because it is optimized for classification tasks, while MedSAM, derived from SAM, focuses on segmentation. MedSAM embeddings may not directly translate to high classification performance.

Comparison with Previous Studies

Compared to Duggetto et al. (71% accuracy on CBIS-DDSM) [9], our approach achieved higher accuracy (73–76%). However, it did not match the highest reported performances, such as Al-Masni et al. (97% with YOLO5) and Nguyen et al. (100% with CNN on the larger DDSM dataset) [10-11]. Jafari et al. also reported 92–96% accuracy using advanced ensemble models like ResNet50, ConvNeXt, and EfficientNet [12]. This performance gap can be attributed to several factors. First, our use of the smaller CBIS-DDSM dataset limits generalizability compared to larger datasets like DDSM and RSNA. Second, previous studies used more complex architectures, benefiting from transfer learning on extensive datasets. Finally, differences in data augmentation and training strategies, such as extensive augmentations, enhanced model robustness in prior works.

Limitations and Future Directions

The main limitation of this study is the small CBIS-DDSM dataset, which may hinder generalization. Future work should use larger, high-quality datasets and adopt K-fold cross-validation for more robust evaluation. Advanced data augmentation techniques, such as simulating breast density variations, and domain-specific pretraining could address data constraints. While radiomic features and embeddings were integrated, the feature fusion approach could benefit from refinement. Further validation is needed to align saliency maps with real lesion locations for clinical interpretability. Subgroup analyses of challenging BI-RADS categories (e.g., 4a, 4b) and dense breast cases could offer deeper insights into performance across clinical scenarios [13].

Conclusion

In this study, we demonstrated that integrating MedSAM segmentation with CNN-based classification and radiomic features enhances malignancy prediction in mammography images. Compared to traditional BI-RADS assessment, our approach achieved better performance, offering a valuable tool to support more accurate clinical decision-making in breast cancer diagnosis.

Team Reflection

Our team consists of two master's students in bioinformatics and a fourth-year undergraduate in computer science.

References

1. Farkas, A. H., & Nattinger, A. B. (2023). Breast Cancer Screening and Prevention. *Annals of Internal Medicine*, *176*(11), ITC161–ITC176. https://doi.org/10.7326/AITC202311210
2. *Management of Women With Dense Breasts Diagnosed by Mammography*. (n.d.). Retrieved November 18, 2024, from https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2015/03/management-of-women-with-dense-breasts-diagnosed-by-mammography
3. Sun, H., Li, C., Liu, B., Zheng, H., Feng, D. D., & Wang, S. (2019). *AUNet: Attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms* (No. arXiv:1810.10151). arXiv. https://doi.org/10.48550/arXiv.1810.10151
4. 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. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005
5. Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). *Segment Anything in Medical Images* (No. arXiv:2304.12306). arXiv. https://doi.org/10.48550/arXiv.2304.12306
6. Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, *4*(1), 170177. https://doi.org/10.1038/sdata.2017.177
7. CBIS-DDSM. (n.d.). *The Cancer Imaging Archive (TCIA)*. Retrieved November 21, 2024, from https://www.cancerimagingarchive.net/collection/cbis-ddsm/

8. *ImageNet*. (n.d.). Retrieved November 22, 2024, from https://www.image-net.org/

9. Duggento, A., Guerrisi, M., Toschi, N., Scimeca, M., Urbano, N., Bonanno, E., Aiello, M., Cavaliere, C., Cascella, G. L., Cascella, D., & Conte, G. (2019). A random initialization deep neural network for discriminating malignant breast cancer lesions. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2019*, 912–915. https://doi.org/10.1109/EMBC.2019.8856740

10. Al-Masni, M. A., Al-Antari, M. A., Park, J. M., Gi, G., Kim, T. Y., Rivera, P., Valarezo, E., Choi, M. T., Han, S. M., & Kim, T. S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods and programs in biomedicine*, *157*, 85–94. https://doi.org/10.1016/j.cmpb.2018.01.017

11. Nguyen, N. V., Huynh, H. T., & Le, P. L. (2023, November). Deep Learning Techniques for Segmenting Breast Lesion Regions and Classifying Mammography Images. In *International Conference on Future Data and Security Engineering* (pp. 471-483). Singapore: Springer Nature Singapore.

12. Jafari Z, Karami E. Breast Cancer Detection in Mammography Images: A CNN-Based Approach with Feature Selection. *Information*. 2023; 14(7):410. https://doi.org/10.3390/info14070410

13. Ghaemian, N., Haji Ghazi Tehrani, N., & Nabahati, M. (2021). Accuracy of mammography and ultrasonography and their BI-RADS in detection of breast malignancy. *Caspian journal of internal medicine*, *12*(4), 573–579. https://doi.org/10.22088/cjim.12.4.573