Article

Improving Benign and Malignant Classification in Mammography with ROI-Aware Deep Learning

Kenji Yoshitsugu [*] , Kazumasa Kishimoto , Tadamasa Takemura

Article

Improving Benign and Malignant Classification in Mammography with ROI-Aware Deep Learning

Kenji Yoshitsugu [1,*], Kazumasa Kishimoto [2] and Tadamasa Takemura [1]

[1] University of Hyogo, Graduate School of Information Science

[2] Kyoto University Hospital, Division of Medical Information Technology and Administration Planning

* Correspondence: smokyjp@gmail.com; Tel.: +81-090-3262-7027

Abstract

Deep learning has achieved widespread adoption for medical image diagnosis, with extensive research dedicated to mammographic image analysis for breast cancer screening. This study investigates the hypothesis that incorporating region-of-interest (ROI) mask information for individual mammographic images during deep learning can improve the accuracy of benign/malignant diagnoses. We used Swin Transformer and ConvNeXtV2 deep learning models to evaluate their performance on the public VinDr and CDD-CESM datasets. Our approach involved stratifying mammographic images based on the presence or absence of ROI masks, performing independent training and prediction for each subgroup, and subsequently merging the results. Baseline prediction metrics (sensitivity, specificity, F-score, and accuracy) without ROI-based separation were the following: VinDr/Swin Transformer (0.00, 1.00, 0.00, 0.85), VinDr/ConvNeXtV2 (0.00, 1.00, 0.00, 0.85), CDD-CESM/Swin Transformer (0.29, 0.68, 0.41, 0.48), and CDD-CESM/ConvNeXtV2 (0.65, 0.65, 0.65, 0.65). Subsequent analysis with ROI-based separation demonstrated marked improvements in these metrics: VinDr/Swin Transformer (0.93, 0.87, 0.90, 0.87), VinDr/ConvNeXtV2 (0.90, 0.86, 0.88, 0.87), CDD-CESM/Swin Transformer (0.65, 0.65, 0.65, 0.65), and CDD-CESM/ConvNeXtV2 (0.74, 0.61, 0.67, 0.68). These findings provide compelling evidence validating our hypothesis and affirming the utility of considering ROI mask information for enhanced diagnostic accuracy in mammography.

Keywords: deep learning; mammography; region of interest

1. Introduction

For breast cancer, which remains a prevalent malignancy among women worldwide, early detection and accurate diagnosis are crucially important for improving survival. Mammography is the widely adopted standard for breast cancer screening, but its interpretation demands extensive expertise. Challenges persist related to diagnostic discrepancies and missed diagnoses among radiologists.

Beyond mammography, breast cancer diagnosis incorporates various methods, including visual inspection, palpation, and ultrasound examination. When these examinations reveal abnormalities, clinicians often perform highly invasive procedures such as cytological and histological examinations for definitive diagnosis. If deep learning-based image analysis of minimally invasive mammographic images can achieve high diagnostic accuracy, then it could reduce the need for highly invasive procedures. This approach would simultaneously alleviate burdens on radiologists and breast surgeons responsible for interpreting these images.

Recent rapid advancements in artificial intelligence (AI) technology, particularly deep learning, have accelerated the development of automated analysis and diagnostic support systems for mammographic images remarkably. For various image recognition tasks, deep learning algorithms, especially convolutional neural networks (CNNs), now demonstrate performance comparable to or

exceeding human capabilities. For medical image diagnosis, these technologies often achieve superior accuracy and efficiency compared to conventional methodologies.

Many studies have explored deep learning applications for mammographic image diagnosis. For instance, Zhang et al. [1] performed two-stage classification (normal/abnormal and benign/malignant) of two-view mammograms (CC and MLO) on the public DDSM dataset using a multi-scale attention DenseNet. Lång et al. [2] evaluated the potential of AI for identifying normal mammograms by classifying cancer likelihood scores with a deep learning model on a private dataset, comparing the obtained results to radiologists' interpretations. Another study by Lång et al. [3] indicated that deep learning models trained on a private dataset can reduce interval cancer rates without supplementary screening. Zhu et al. [4] predicted future breast cancer development in negative subjects during an eight-year period using a deep learning model with a private dataset. Kerschke et al. [5] compared human versus deep learning AI accuracy for benign–malignant screening using a private dataset, highlighting the need for prospective studies. Nica et al. [6] reported high-accuracy benign–malignant classification of cranio-caudal view mammography images using an AlexNet deep learning model and a private dataset. Rehman et al. [7] achieved high-accuracy architectural distortion detection using image processing and proprietary depth-wise 2D V-net 64 convolutional neural networks on the PINUM, CBIS-DDSM, and DDSM datasets. Yirgin et al. [8] used a public deep learning diagnostic system on a private dataset, concluding that combined assessment by the deep learning model and radiologists yielded the best performance. Tzortzis et al. [9] demonstrated superior performance for efficiently detecting abnormalities on the public INBreast dataset using their tensor-based deep learning model, particularly showing robustness with limited data and reduced computational requirements. Pawar et al. [10] and Hsu et al. [11] both reported high-accuracy Breast Imaging Reporting and Data System (BIRADS) category classification, respectively, using proprietary multi-channel DenseNet architecture and a fully convolutional dense connection network on private datasets. Elhakim et al. [12] investigated the feasibility of replacing the first reader with AI in double-reading mammography using a commercial AI system with a private dataset, emphasizing the importance of an appropriate AI threshold. Jaamour et al. [13] improved the segmentation accuracy for mass and calcification images from the public CBIS-DDSM dataset by application of transfer learning. Kebede et al. [14] developed a model combining EfficientNet-based classifiers with a YOLOv5 object detection model and an anomaly detection model for mass screening on the public VinDr and Mini-DDSM datasets. Ellis et al. [15], using the UK-national OPTIMAM dataset, developed a deep learning AI model for predicting future cancer risk in patients with negative mammograms. Elhakim et al. [16] further investigated replacement of one or both readers with AI in double-reading mammography, emphasizing clinical implications for accuracy and workload. Sait et al. [17] reported high segmentation accuracy and generalizability in multi-class breast cancer image classification using an EfficientNet B7 model within a LightGBM model on the CBIS-DDSM and CMMD datasets. Chakravarthy et al. [18] reported high classification accuracy for normal, benign, and malignant cases using an ensemble method with a modified Gompertz function on the BCDR, MIAS, INbreast, and CBIS-DDSM datasets. Liu et al. [19] achieved high classification accuracy on four binary tasks using a CNN and a private mammography image dataset, suggesting a potential to reduce unnecessary breast biopsies. Finally, Park et al. [20] reported improved diagnostic accuracy, especially in challenging ACR BIRADS categories 3 and 4 with breast density exceeding 50%, by learning both benign–malignant classification and lesion boundaries using a ViT-B DINO-v2 model on the public CBIS-DDSM dataset. AlMansour et al. [21] reported high-accuracy BIRADS classification using MammoViT, a novel hybrid deep learning framework, on a private dataset.

Despite these advancements, several points of difficulty hinder the reproducibility of claims in deep learning applications for mammographic image diagnosis. Studies using private, non-public datasets or proprietary deep learning models with undisclosed details make verification challenging. Methods incorporating subject information alongside mammographic images as training data also face reproducibility issues caused by limited commonalities across different datasets. Similarly,

studies combining mammographic images with other modality images require specific data combinations, thereby complicating claim reproduction.

Given these considerations, we prioritized reproducible research by particularly addressing studies using publicly available datasets and open-source deep learning models. Furthermore, we emphasized the generalizability of claims across multiple public datasets and various deep learning models.

Therefore, this study tested the hypothesis that prediction accuracy improves when images are divided into those with and without annotated mask information for regions of interest, with subsequent separate training and prediction for each of the four mammographic views (RCC, LCC, RMLO, LMLO), before merging the results. This approach is compared to cases for which image data are not separated based on the availability of mask information for regions of interest. Using two public datasets and two deep learning models, we validated this hypothesis, particularly addressing the presence or absence of annotated mask information as a novel feature.

2. Materials and Methods

2.1. Materials

This study used two publicly available mammography datasets with region of interest (ROI) annotations: VinDr [22] and CDD-CESM [23]. Both datasets include ROI mask information, but not all mammographic images within them have corresponding mask images available.

During the training and prediction phases, our study exclusively considered the presence or absence of corresponding ROI images for individual mammographic images. The ROI images themselves were not used as input data.

The VinDr dataset provides BI-RADS information, but it lacks explicit benign–malignant classifications. Consequently, images categorized as BI-RADS 2 and 3 were classified as benign lesions, whereas those categorized as BI-RADS 4 and 5 were classified as malignant.

The CDD-CESM dataset includes predefined normal, benign, and malignant classifications. For this analysis, we used benign and malignant data exclusively.

Because the CDD-CESM dataset does not provide a predefined train–test split, we divided the data into training and testing sets with a 10:1 ratio.

Compositions of the respective datasets are presented in Tables 1–4.

Table 1. Distribution of Cases in Training and Testing Subsets of the VinDr Dataset.

	training	testing	total
malignant	790	198	988
benign	4486	1120	5606
total	5276	1318	6594

Table 2. Presence of ROIs in the VinDr Dataset.

	with ROIs	without ROIs	total
malignant	952	36	988
benign	816	4790	5606
total	1768	4826	6594

Table 3. Distribution of Cases in Training and Testing Subsets of the CDD-CESM Dataset.

	training	testing	total
malignant	300	31	331
benign	300	31	331
total	600	62	662

Table 4. Presence of ROIs in the CDD-CESM Dataset.

	with ROIs	without ROIs	total
malignant	326	5	331
benign	322	9	331
total	648	14	662

2.2. Methods

2.2.1. Image Preprocessing

For preprocessing, window processing was applied during conversion of DICOM images to JPEG format. This preprocessing was followed by contrast adjustment using Contrast-Limited Adaptive Histogram Equalization (CLAHE).

2.2.2. Image Classification Models

Swin Transformer [24] and ConvNeXtV2 [25] were selected as image classification models because of their superior performance among the various models evaluated.

2.2.3. Validation Procedure

The validation procedures were identical for both image classification models, involving the following steps.

1. Mammographic images were segregated based on the presence or absence of ROI mask images.
2. Images were divided into four standard views: right craniocaudal (RCC), left craniocaudal (LCC), right mediolateral oblique (RMLO), and left mediolateral oblique (LMLO).
3. Training was performed on mammographic images without ROI mask images, with separate training for each view.
4. Prediction was performed on mammographic images without ROI mask images, with separate prediction for each view.
5. Training was then performed on mammographic images with ROI mask images, again with separate training for each view.
6. Prediction was performed on mammographic images with ROI mask images, with separate prediction for each view.
7. Finally, the prediction results were merged.

2.2.4. Comparative Validation Procedure

The comparative validation procedure differed from primary validation only in that the presence or absence of ROI mask images was not considered during processing. The steps used for this procedure were the following.

1. Mammographic images were divided into the four standard views: RCC, LCC, RMLO, and LMLO.
2. Training was performed on mammographic images without ROI mask images, separately for each view.
3. Prediction was performed on mammographic images without ROI mask images, separately for each view.
4. Training was performed on mammographic images with ROI mask images, separately for each view.
5. Prediction was performed on mammographic images with ROI mask images, separately for each view.
6. The prediction results were merged.

2.2.5. Training Hyperparameters and Computational Environment

Training hyperparameters were determined through five-fold cross-validation. They remained consistent for both image classification models: a learning rate of 0.0001, 100 epochs, and image size of 384 × 384 pixels. All other hyperparameters were maintained at their respective default values for each model.

Validation was conducted on a system running Windows 11 Pro, equipped with a 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz processor, 128 GB of memory, and an NVIDIA RTX 3090 GPU.

3. Results

Our mammographic image classification results are presented for two scenarios: with and without the inclusion of ROI mask images. For the VinDr dataset, the classification results obtained using Swin Transformer and ConvNeXtV2 are shown respectively in Tables 5 and 6. Similarly for the CDD-CESM dataset, Tables 7 and 8 present the classification results obtained respectively using Swin Transformer and ConvNeXtV2.

Table 5. SwinTransformer / Vindr.

	without consideration of ROI mask image presence	with consideration of ROI mask image presence
Sensitivity	0.00	0.93
Specificity	1.00	0.87
F-score	0.00	0.90
Accuracy	0.85	0.87

Table 6. ConvNeXt2 / Vindr.

	without consideration of ROI mask image presence	with consideration of ROI mask image presence
Sensitivity	0.00	0.90
Specificity	1.00	0.86
F-score	0.00	0.88
Accuracy	0.85	0.87

Table 7. SwinTransformer / CDD-DESM.

	without consideration of ROI mask image presence	with consideration of ROI mask image presence
Sensitivity	0.29	0.65
Specificity	0.68	0.65
F-score	0.41	0.65
Accuracy	0.48	0.65

Table 8. ConvNeXt2 / CDD-DESM.

	without consideration of ROI mask image presence	with consideration of ROI mask image presence
Sensitivity	0.65	0.74
Specificity	0.65	0.61
F-score	0.65	0.67
Accuracy	0.65	0.68

4. Discussion

Some data lead to diagnosis as malignant, but lack a visible region of interest (ROI). This discrepancy is likely attributable to factors such as dense breast tissue, which can obscure ROIs by causing the entire image to appear uniformly opaque. In such cases, a malignant diagnosis reached despite the absence of a clear ROI on the image likely reflects corroborating results from other diagnostic modalities such as biopsies. Also observed are instances of data diagnosed as normal but exhibiting an ROI. The presence of an ROI in a "normal" diagnosis looks strange and suggests a potential misrepresentation or artifact in the diagnostic labeling process. Such anomalous data points, whether they involve a malignant diagnosis without a discernible ROI or a normal diagnosis with an ROI, introduce noise into deep learning models. This noise can strongly hinder the model's ability to learn accurate patterns. The noise consequently diminishes its predictive performance. Preprocessing the dataset to identify and remove or re-evaluate these inconsistent data points before training might enhance the learning and prediction accuracy of deep learning algorithms for medical image analysis.

This study used data with pre-existing ROI mask images. However, mammographic images requiring benign–malignant classification do not always have corresponding mask images available. Therefore, future research should specifically examine generation of mask images for mammographic data lacking existing masks, employing techniques such as semantic segmentation or object detection, and subsequently validating these approaches.

The deep learning models used for this study, Swin Transformer and ConvNeXtV2, demonstrated superior accuracy in both training and prediction compared to other deep learning models. We hypothesize that this improved performance derives from differences in their respective layer architectures. However, detailed analysis of this phenomenon is left as a subject for future investigation.

Whereas this study specifically addressed benign–malignant classification, mammographic data are typically categorized initially into normal versus abnormal findings, with abnormal cases subsequently classified as either benign or malignant. An important area for future investigation is assessment of whether our methodology can also classify normal versus abnormal cases effectively. If successful, this capability would enable diagnostic prediction for a broader range of arbitrary mammographic data.

5. Conclusions

Our findings indicate that pre-segmentation of mammographic data based on the presence or absence of ROI mask images, followed by separate training and prediction processes for each segment and subsequent merging of the results, can enhance classification accuracy for the benign–malignant classification of mammographic images using deep learning.

References

1. Zhang, C.; Zhao, J.; Niu, J.; Li, D. New convolutional neural network model for screening and diagnosis of mammograms. *PLoS One* **2020**, *15*(8), e0237674. doi:10.1371/journal.pone.0237674
2. Lång, K.; Dustler, M.; Dahlblom, V.; Åkesson, A.; Andersson, I.; Zackrisson, S. Identifying normal mammograms in a large screening population using artificial intelligence. *European Radiology* **2021**, *31*(3), 1687–1692. doi:10.1007/s00330-020-07165-1
3. Lång, K.; Hofvind, S.; Rodríguez-Ruiz, A.; Andersson, I. Can artificial intelligence reduce the interval cancer rate in mammography screening?. *European Radiology* **2021**, *31*(8), 5940–5947. doi:10.1007/s00330-021-07686-3
4. Zhu, X.; Wolfgruber, T.K.; Leong, L.; Jensen, M.; Scott, C.; Winham, S.; Sadowski, P.; Vachon, C.; Kerlikowske, K.; Shepherd, J.A. Deep Learning Predicts Interval and Screening-detected Cancer from Screening Mammograms: A Case–Case–Control Study in 6369 Women. *Radiology* **2021**, *301*(3), 50–558. doi:10.1148/radiol.2021203758

5. Kerschke, L.; Weigel, S.; Rodriguez-Ruiz, A.; Karssemeijer, N.; Heindel, W. Using deep learning to assist readers during the arbitration process: a lesion-based retrospective evaluation of breast cancer screening performance. *European Radiology* **2022,** *32*(2), 842–852. doi:10.1007/s00330-021-08217-w

6. Nica, R.E.; Șerbănescu, M.S.; Florescu, L.M.; Camen, G.C.; Streba, C.T.; Gheonea, I.A. Deep Learning: a Promising Method for Histological Class Prediction of Breast Tumors in Mammography. *Journal of Digital Imaging* **2021**, *34*(5). 1190–1198. doi:10.1007/s10278-021-00508-4

7. Rehman, K.U.; Li, J.; Pei, Y.; Yasin, A.; Ali, S.; Saeed, Y. Architectural Distortion-Based Digital Mammograms Classification Using Depth Wise Convolutional Neural Network. *Biology* **2021**, *11*(1), 15. doi:10.3390/biology11010015

8. Kizildag Yirgin, I.; Koyluoglu, Y.O.; Seker, M.E.; Ozkan Gurdal, S.; Ozaydin, A.N.; Ozcinar, B.; Cabioğlu, N.; Ozmen, V.; Aribal, E. Diagnostic Performance of AI for Cancers Registered in A Mammography Screening Program: A Retrospective Analysis. *Technology in Cancer Research & Treatment* **2022**, *21*, 15330338221075172. doi:10.1177/15330338221075172

9. Tzortzis, I.N.; Davradou, A.; Rallis, I.; Kaselimi, M.; Makantasis, K.; Doulamis, A.; Doulamis, N. Tensor-Based Learning for Detecting Abnormalities on Digital Mammograms. *Diagnostics (Basel, Switzerland)* **2022**, *12*(10), 2389. 1 Oct., doi:10.3390/diagnostics12102389

10. Pawar, S.D.; Sharma, K.K.; Sapate, S.G.; Yadav, G.Y.; Alroobaea, R.; Alzahrani, S.M.; Hedabou, M. Multichannel DenseNet Architecture for Classification of Mammographic Breast Density for Breast Cancer Detection. *Frontiers in Public Health* **2022**, *10*, 885212. doi:10.3389/fpubh.2022.885212

11. Hsu, S.Y.; Wang, C.Y.; Kao, Y.K.; Liu, K.Y.; Lin, M.C.; Yeh, L.R.; Wang, Y.M.; Chen, C.I.; Kao, F.C. Using Deep Neural Network Approach for Multiple-Class Assessment of Digital Mammography. *Healthcare (Basel, Switzerland)* **2022**, *10*(12), 2382. doi:10.3390/healthcare10122382

12. Elhakim, M.T.; Stougaard, S.W.; Graumann, O.; Nielsen, M.; Lång, K.; Gerke, O.; Larsen, L.B.; Rasmussen, B.S.B. Breast cancer detection accuracy of AI in an entire screening population: a retrospective, multicentre study. *Cancer imaging: The official publication of the International Cancer Imaging Society* **2023**, *23*(1), 127. 20 Dec., doi:10.1186/s40644-023-00643-x

13. Jaamour, A.; Myles, C.; Patel, A.; Chen, S.J.; McMillan, L.; Harris-Birtill, D. A divide and conquer approach to maximise deep learning mammography classification accuracies. *PLoS One* **2023**, *18*(5),5 e0280841. doi:10.1371/journal.pone.0280841

14. Kebede, S.R.; Waldamichael, F.G.; Debelee, T.G.; Aleme, M.; Bedane, W.; Mezgebu, B.; Merga, Z.C. Dual view deep learning for enhanced breast cancer screening using mammography. Scientific Reports **2024**, *14*(1), 3839. doi:10.1038/s41598-023-50797-8

15. Ellis, S.; Gomes, S.; Trumble, M.; Halling-Brown, M.D.; Young, K.C.; Chaudhry, N.S.; Harris, P.; Warren, L.M. Deep Learning for Breast Cancer Risk Prediction: Application to a Large Representative UK Screening Cohort. *Radiology. Artificial Intelligence* **2024**, *6*(4), e230431. doi:10.1148/ryai.230431

16. Elhakim, M.T.; Stougaard, S.W.; Graumann, O.; Nielsen, M.; Gerke, O.; Larsen, L.B.; Rasmussen, B.S.B. AI-integrated Screening to Replace Double Reading of Mammograms: A Population-wide Accuracy and Feasibility Study. *Radiology. Artificial Intelligence* **2024**, *6*(6), e230529. doi:10.1148/ryai.230529

17. Sait, A.R.W.; Nagaraj, R. An Enhanced LightGBM-Based Breast Cancer Detection Technique Using Mammography Images. *Diagnostics (Basel, Switzerland)* **2024**, *14*(2), 227. doi:10.3390/diagnostics14020227

18. Altameem, A.; Mahanty, C.; Poonia, R.C.; Saudagar, A.K.J.; Kumar, R. Breast Cancer Detection in Mammography Images Using Deep Convolutional Neural Networks and Fuzzy Ensemble Modeling Techniques. *Diagnostics (Basel, Switzerland)* **2022**, *12*(8), 1812. doi:10.3390/diagnostics12081812

19. Liu, C.; Sun, M.; Arefan, D.; Zuley, M.; Sumkin, J.; Wu, S. Deep learning of mammogram images to reduce unnecessary breast biopsies: a preliminary study. *Breast Cancer Research: BCR* **2024**, *26*(1), 82. doi:10.1186/s13058-024-01830-9

20. Park, J.H.; Lim, J.H.; Kim, S.; Heo, J. A Multi-label Artificial Intelligence Approach for Improving Breast Cancer Detection with Mammographic Image Analysis. In vivo *(Athens, Greece)* **2024**, *38*(6), 2864–2872. doi:10.21873/invivo.13767

21. Al Mansour, A.G.M.; Alshomrani, F.; Alfahaid, A.; Almutairi, A.T.M. MammoViT: A Custom Vision Transformer Architecture for Accurate BIRADS Classification in Mammogram Analysis. *Diagnostics (Basel, Switzerland)* **2025**, *15*(3), 285. doi:10.3390/diagnostics15030285

22. Nguyen, H.T.; Nguyen, H.Q.; Pham, H.H.; Lam, K.; Le, L.T.; Dao, M.; Vu, V. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data* **2023**, *10*(1), 277. doi:10.1038/s41597-023-02100-7

23. Khaled, R.; Helal, M.; Alfarghaly, O.; Mokhtar, O.; Elkorany, A.; El Kassas, H.; Fahmy, A. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. *Scientific Data* **2022**, *9*(1), 122. doi:10.1038/s41597-022-01238-0

24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows" Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012-10022

25. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z; Kweon, I.S.; Xie, S. "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 16133–16142, doi: 10.1109/CVPR52729.2023.01548.