scientific reports

Check for updates

OPEN

An open codebase for enhancing transparency in deep learning-based breast cancer diagnosis utilizing CBIS-DDSM data

Ling Liao[1,2]✉ & Eva M. Aagaard[3]

Accessible mammography datasets and innovative machine learning techniques are at the forefront of computer-aided breast cancer diagnosis. However, the opacity surrounding private datasets and the unclear methodology behind the selection of subset images from publicly available databases for model training and testing, coupled with the arbitrary incompleteness or inaccessibility of code, markedly intensifies the obstacles in replicating and validating the model's efficacy. These challenges, in turn, erect barriers for subsequent researchers striving to learn and advance this field. To address these limitations, we provide a pilot codebase covering the entire process from image preprocessing to model development and evaluation pipeline, utilizing the publicly available Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) mass subset, including both full images and regions of interests (ROIs). We have identified that increasing the input size could improve the detection accuracy of malignant cases within each set of models. Collectively, our efforts hold promise in accelerating global software development for breast cancer diagnosis by leveraging our codebase and structure, while also integrating other advancements in the field.

Breast cancer is the most common cancer among women in 157 out of 185 countries, leading to 670,000 deaths globally in 2022[1]. Despite technological advancements such as tomosynthesis, introduced to enhance breast cancer screening and early-stage diagnosis for more effective treatment, challenges persist in the frequent occurrence of false positive mammograms and variability among expert readers, resulting in increased patient anxiety, as well as financial and opportunity costs[2–6]. Given these obstacles, alongside efforts promoting screening access, significant endeavors have been made in developing software for computer-aided diagnosis (CAD) to interpret abnormal mammograms[7–14]. However, the effectiveness of various algorithms for CAD have been questioned due to difficulties in replication, a significant drop in performance, and constraints in training that have resulted from limited datasets[15–19].

Results

Overview of CBIS-DDSM mass dataset application for breast cancer diagnosis

In this publication, we conducted a case study utilizing the mass subset from CBIS-DDSM, a high-volume publicly available mammography dataset. This subset comprises 1696 abnormal ROIs and 1592 respective full images obtained from 892 patients[19,20]. Figure 1 provides an overview of applying CBIS-DDSM mass subset for breast cancer diagnosis.

Both benign and malignant full images and ROIs in the CBIS-DDSM mass subset are stored in Digital Imaging and Communications in Medicine (DICOM) format, with 16-bit depth for full images and 8-bit depth for ROIs. Upon conversion to grayscale portable network graphics (PNG) format without altering the bit quality, we aligned the ROIs with corresponding full images to identify abnormal regions. We then cropped the identified abnormal areas, centering them at the geometric center with a preliminary target size of 598×598 pixels for each mapped image, to capture most of the abnormal information. After processing the initial cropped images, as detailed in Fig. 2, we applied augmentation techniques such as flipping, rotation, zoom, shear, and shift to increase data variety and simulate real-world conditions. This resulted in generating five additional images for each cropped and processed output. After randomly selecting 80% of the processed 1696 images, each with

[1]Biomedical Deep Learning LLC, St. Louis, MO, USA. [2]Computational and Systems Biology, Washington University in St. Louis, St. Louis, MO, USA. [3]Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA. ✉email: 1995aileen@gmail.com

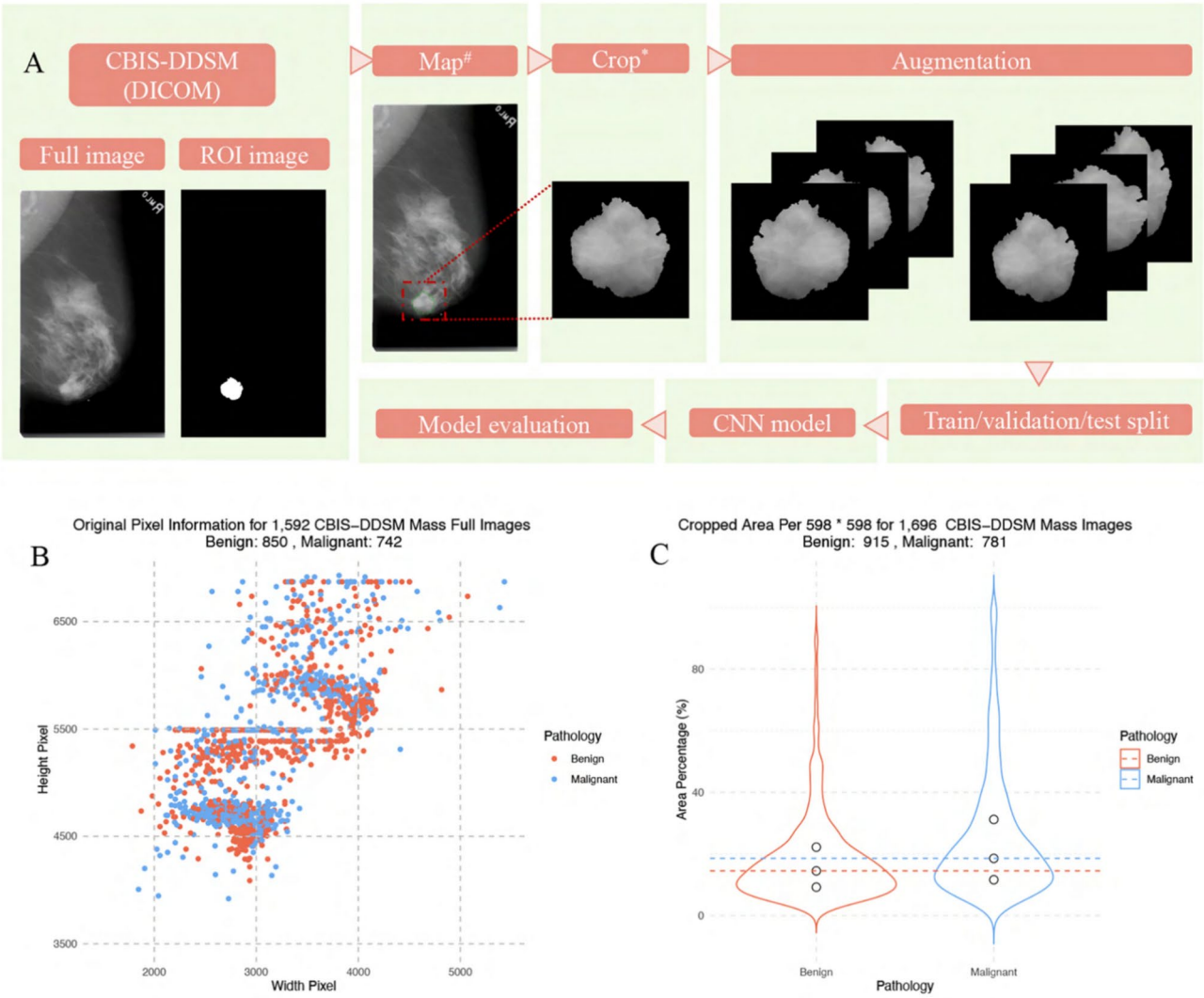

Fig. 1. Overview of CBIS-DDSM Mass Dataset Application for Breast Cancer Diagnosis. (**A**) A brief summary of our model training process. (**B**) Pixel information of 1,592 full CBIS-DDSM mass images. Width was depicted on the x-axis and height on the y-axis, with benign and malignant cases distinguished by colors. (**C**) Cropped abnormal area per 598 × 598 region from 1696 mapped CBIS-DDSM mass images. Some full images contain multiple abnormal regions, leading to a greater number of mapped images compared to the original full CBIS-DDSM mass images.

dimensions of 598 × 598 pixels, for training, the augmented versions of this training set were added as training input, while the validation and test sets were kept the same.

The variability in size among CBIS-DDSM mass full mammography images is illustrated in Fig. 1B, where dots on the plot represent various dimensions. In this study, we did not resize the original full images due to concerns about the varying pixel dimensions and the potential implications of standardizing all images to a fixed pixel dimension. While downsizing could alleviate some computing resource limitations, as performed in other studies, it may also result in the loss of important information. After cropping, only 3 out of 1696 cropped images were fully covered by the abnormal areas. Malignant cases showed a larger proportion per 598 × 598 pixel area compared to benign cases, as depicted in Fig. 1C and quantified across the 1st quartile, median, and 3rd quartile (9.2%, 14.5%, 22.2% for benign; 11.6%, 18.6%, 31.2% for malignant cases, respectively) ($P < 0.001$, Mann–Whitney U Test). These results underscore the importance of clearly specifying selected subsets of the CBIS-DDSM dataset when applying it for model training and testing. While often overlooked, it significantly contributes to our understanding and validation of model performance. Furthermore, these findings raised our concerns about the adequacy of collected abnormal information when researchers opt to crop only a small region, such as 224 × 224 or 299 × 299 pixels, for model training, a common practice observed in many publications. We subsequently compared the performance of two sets of models on different center-cropped image input sizes—224 × 224, 299 × 299, 448 × 448, and 598 × 598 pixels, with the results shown in Table 1 and Supplementary Fig. 1.

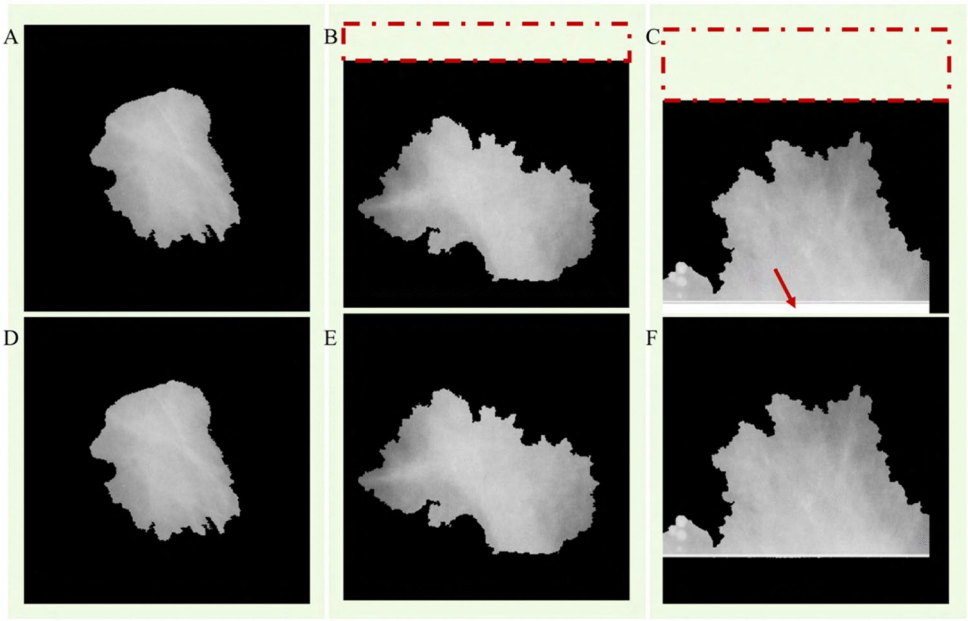

Fig. 2. Examples of Cropped Abnormal Areas at 598 × 598 pixels: Original vs. Processed Versions. Panels (**A**–**C**) showcase the original cropped images, while panels (**D**–**F**) depict the corresponding processed versions. The red dashed box indicates areas with missing pixels, and the red arrow highlights an example of unwanted white areas.

Model Performance	ROC AUC [95% CI]	Accuracy [95% CI]	Precision [95% CI]	Recall [95% CI]	F1 Score [95% CI]
ResNet-50 448 × 448	0.7421 [0.6651, 0.8147]	0.6824 [0.6118, 0.7529]	0.6800 [0.5774, 0.7848]	0.6296 [0.5172, 0.7350]	0.6538 [0.5556, 0.7356]
ResNet-50–224 × 224	0.7269 [0.6500, 0.7965]	0.6765 [0.6059, 0.7471]	0.6562 [0.5428, 0.7733]	0.5600 [0.4516, 0.6703]	0.6043 [0.5074, 0.6950]
Xception—599 × 599	0.6820 [0.6006, 0.7642]	0.6824 [0.6118, 0.7529]	0.6197 [0.5074, 0.7286]	0.6197 [0.5113, 0.7314]	0.6197 [0.5203, 0.7043]
Xception-299 × 299	0.7024 [0.6176, 0.7764]	0.6059 [0.5294, 0.6765]	0.6232 [0.5000, 0.7273]	0.5119 [0.4069, 0.6154]	0.5621 [0.4595, 0.6460]
Confusion Matrix	**Benign-Benign**	**Benign-Malignant**	**Malignant-Benign**	**Malignent-Malignent**	–
ResNet-50 448 × 448	65(73.03%)	24(26.97%)	30(37.04%)	51(62.96%)	–
ResNet-50–224 × 224	73(76.84%)	22(23.16%)	33(44%)	42(56%)	–
Xception—599 × 599	72(72.73%)	27(27.27%)	27(38.03%)	44(61.97%)	–
Xception-299 × 299	60(69.77%)	26(30.23%)	41(48.81%)	43(51.19%)	–

Table 1. Evaluation matrix of model's performance on processed CBIS-DDSM mass image data. CI: confidential interval. True Negative refers to Benign-Benign, False Positive refers to Benign-Malignant, False Negative refers to Malignant-Benign, and True Positive refers to Malignant-Malignant. Percentage (%) was calculated based on the sum of the first two or last two numbers per row.

Cropped abnormal region procession

As noted, the cropped images shown in Fig. 1 depicted the final version after initial cropping, and many original outputs were smaller than 598 × 598 pixels due to the geometric centering of the abnormal region within corresponding mapped images. Figure 2 presents a detailed illustration of both the initial and processed versions of the cropped images, with images D-F representing the processed versions of images A-C. The original cropped images were categorized based on size into three groups: those equal to the desired 598 × 598 pixels size, those smaller than the desired size, and 18 images with unwanted wide background, an example highlighted by the red arrow in Fig. 2C. While these white edges were identified as ROIs by CBIS-DDSM contributors, we believe we are the first to describe that they do not represent the abnormal regions we aimed to collect. To confirm this, we traced further back to the original full images and verified that these areas were pure backgrounds. For images resembling category B, we added blank background to achieve the desired size while keeping the abnormal region centered. In the case of category C images, we removed the white background by setting their pixel value to 0, and subsequently applied the same process as for category B images.

Model development

Two sets of models were constructed with different transfer learning techniques. As illustrated in Fig. 3, a couple of convolutional layers, Batch Normalization, and LeakyReLU activation were initially applied to extract

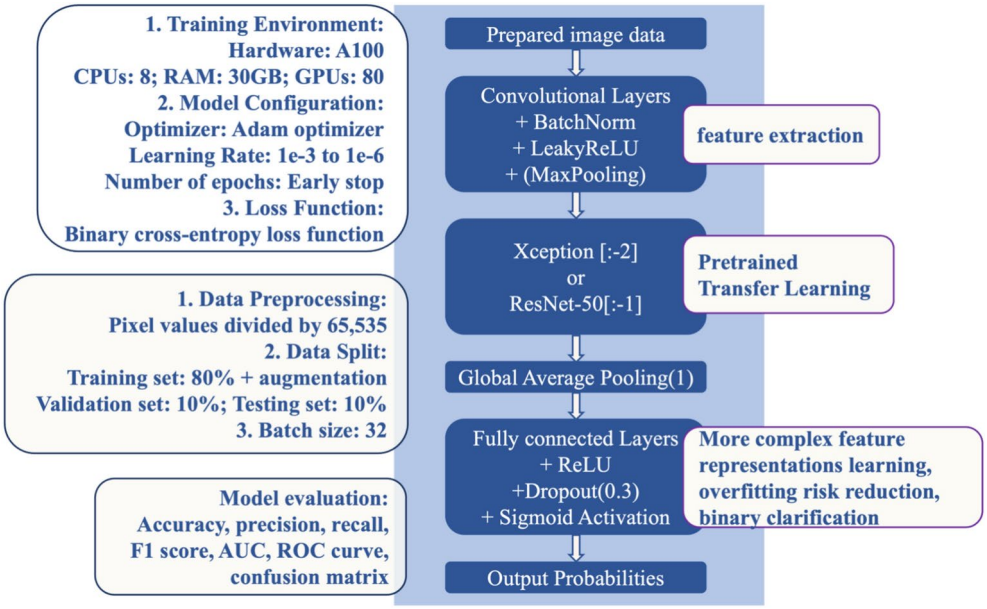

Fig. 3. Key aspects in model development, training, and evaluation, as well as the workflow of model structure.

features[21–23]. In between, a MaxPooling layer was employed to reduce dimension of the inputs, which were either 448 × 448 and 598 × 598 pixels[24]. We then employed ImageNet pretrained transfer learning with ResNet-50 or Xception to utilize pretrained weight and the residual architecture or depth wise separable convolutions[25,26]. To adapt to our specific task, we made further modifications, including removing the last one layer of ResNet-50 or last two layers of Xception and incorporating a one-channel global average pooling layer to align output with our subsequent linear layer and prediction process[27]. Moreover, we implemented dropout regularization to further prevent overfitting by randomly deactivating 30% of neurons during training and forcing the model to learn more robust and generalized representations[28]. For prediction, we applied sigmoid activation to produce binary outcomes and utilized the binary cross-entropy loss function to calculate loss[29]. Model evaluation encompasses metrics including accuracy, precision, recall, F1 score, ROC AUC, ROC curve, and the confusion matrix for assessment of effectiveness and performance.

For environment and configuration, the training, validation and testing process were executed with an A100 GPU alongside 80GiB GPUs, 8 CPUs, and 30GiB of RAM. The model was optimized with Adam optimizer and trained with a learning rate ranging from 1e−3 to 1e−6. Model checkpoints were saved, and early stopping was implemented if the validation loss increased for three consecutive epochs after reaching its lowest value. The epoch with the lowest loss before early stopping was selected as the final checkpoint for testing. A detailed description is provided in the Methods section, with an example shown in Fig. 4.

As outlined below, the processed 1696 598 × 598 pixel images (benign: 915; malignant: 781) were randomly split for training (80%), validation (10%), and testing (10%). The augmented versions (6780 images) of the randomly selected training set were included as training input, while the validation and test sets remain the same. All pixel values of the images were normalized to between 0 and 1. Depending on the input requirements across models, the images were either center-cropped to dimensions of 224 × 224, 299 × 299, or 448 × 448, or they remained unmodified in terms of size.

Performance evaluation

The set of metrics for evaluating the model performance at 224 × 224, 299 × 299, 448 × 448 and 598 × 598 pixels were depicted in Table 1. ResNet-50 448 × 448 employed the ResNet-50 architecture with transfer learning and an input image size of 448 × 448 pixels. It achieved the highest performance for detecting malignant cases, with a ROC AUC of 0.7421 and a malignant detection rate of 62.96%. ResNet-50–224 × 224 also utilized ResNet-50 with transfer learning but used a smaller input size of 224 × 224 pixels, resulting in a lower malignant detection rate of 56%, despite similar overall performance metrics. Xception—599 × 599 applied the Xception architecture with transfer learning and an input size of 599 × 599 pixels, demonstrating performance with a malignant detection rate of 61.97%. Finally, Xception-299 × 299, which employed the same Xception architecture but with a smaller input size of 299 × 299 pixels, yielded the lowest malignant detection rate of 51.19%. Overall, increasing the input image size generally improved the detection of malignant cases, with ResNet-50 based models outperforming Xception models in this regard, particularly at larger input sizes. The AUC curves for these models were shown in Supplementary Fig. 1. These results are generally consistent with those reported in some recent studies that examined the robustness of deep networks for mammography across public datasets (~ 70% ROC AUC on CBIS-DDSM dataset)[15,16].

The observed performance differences might be attributed to two key factors. First, larger input image sizes enable a more detailed representation of mammography images, including critical features such as the shape of

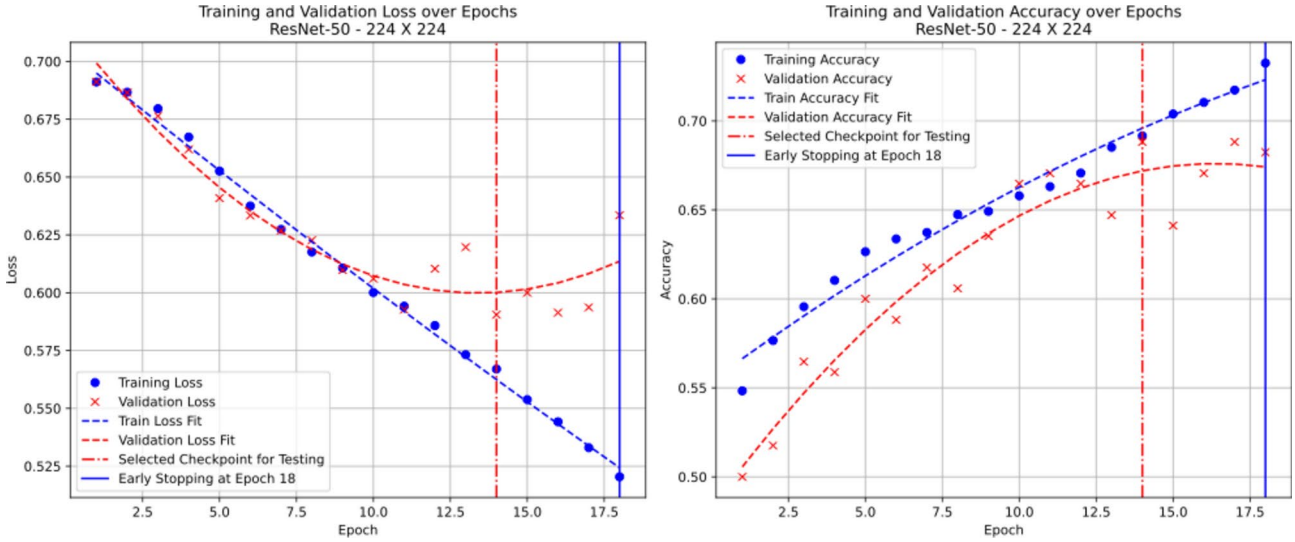

Fig. 4. Training and Validation Loss and Accuracy over Epochs for Processed Input Images with 224 × 224 Pixels on Our Designed Model.

the ROI, whether it is oval or irregular. This enhanced detail can improve the performance of models, particularly CNNs which are designed to detect fine-grained patterns and edges. Additionally, as noted before, malignant cases exhibited a larger proportion of the image area compared to benign cases ($P < 0.001$, Mann–Whitney U Test). Conversely, smaller image sizes might result in some loss of important details, leading to diminished performance in detecting malignant cases. Second, the ResNet-50 architecture, with its residual learning framework, effectively trains deep networks and captures complex patterns in data, potentially leading to better performance on our dataset. In contrast, while Xception's depthwise separable convolutions offer efficient feature extraction, they may not surpass ResNet-50 in this context.

Discussion

Computer models designed to evaluate breast cancer screening have evolved significantly over the decades. Carter et al.[30] introduced a model simulating the progression of breast cancer from the first malignant cell, incorporating thresholds for growth and spread to assess screening effectiveness[30]. Kowal et al.[31] developed a CAD system using biopsy images, while Huang et al.[32] applied a fruit fly optimization algorithm enhanced support vector machine for high-level feature analysis[31,32]. Zhang et al.[33] introduced a bionic fixation system for MRI-guided biopsies, and Ahmed et al.[34] proposed a reinforcement federated learning strategy for processing medical data[33,34]. These innovations underscore the diverse approaches driving forward breast cancer diagnostic research and serve as foundation upon which our paper is built.

The CBIS-DDSM dataset, central to our study, has been instrumental in numerous research publications utilizing innovative methods. However, these studies often exhibit inconsistent and controversial performance due to variations in subset selection, image processing techniques, model development details and evaluation methods. Our study aims to empower transparency in deep learning-based breast cancer diagnosis by providing an open codebase which covers the entire process from image preprocessing to model construction and evaluation pipeline utilizing the publicly available CBIS-DDSM mass subset and ROIs. Additionally, we have made our entire codebase publicly available, spanning from data preprocessing to model evaluation. As the old saying goes, *the devil is in the details*. A complete and annotated codebase is essential for guaranteeing the reproducibility, transparency, effectiveness, and overall reliability of our model, thereby further advancing research communication and improvement in this field.

We meticulously attended to details throughout this study. We documented the selection of subsets from the CBIS-DDSM dataset, maintained consistent bit depth during image format conversion for both full images and ROIs, ensured size congruence for mapping, confirmed sufficient abnormal regions in output images after cropping, and appended blank ground and removed white edges before augmentation as necessary. Furthermore, we intentionally built two sets of ImageNet pre-trained neural network based CNNs, fine-tuned various hyperparameters, detailed test checkpoint selection and early stopping criteria, and applied ROC AUC, accuracy, precision, recall, F1 score, and confusion matrix to thoroughly evaluate the model's performance on 4 sets of input sizes.

Based on the findings, we raised concerns about how mammography images were generally processed, especially in resizing and cropping, with the visualized full image size and the processed area per 598 × 598 pixels in Fig. 1B and C, and in the performance results of the models detailed in Table 1. For model training, validation, and testing, collecting enough abnormal area is critical to ensure robustness and accuracy in the detection and prediction of pathologies, particularly in tasks such as breast cancer diagnosis. Furthermore, we concerned about the use of denoising and filtering technologies, such as Contrast Limited Adaptive Histogram Equalization, to enhance image quality, as observed in several publications, without adequate validation by

trained mammographers. For computational researchers without professional guidance, relying solely on self-implemented mammography image denoising feels akin to navigating blindly and hoping for positive outcomes.

Additionally, while the CBIS-DDSM dataset has been widely utilized for model development since its publication in 2017, we recognized the inherent limitations in the accuracy of the automatically generated ROIs we employed to crop abnormal regions. This limitation was exemplified in Fig. 2C and supported by data contributors, acknowledging the challenges posed by the exorbitantly large workload associated with hand-drawn annotations. In the future, we aim to engage more expert radiologists in the model development process, and we encourage all in this field to promote transparency in the decision-making process for data utilization, particularly publicly available data, among outstanding laboratories and researchers worldwide.

Methods

Dataset description

Dataset utilized in this study is the mass subset of CBIS-DDSM data, which is publicly accessible and does not raise additional privacy concerns from the authors' standpoint. CBIS-DDSM represents an enhanced and standardized subset extracted from the Digital Database for Screening Mammography. Curated by a trained mammographer, this subset features converted and decompressed mammography images in DICOM format. Additional updates have been made to the segmentation of ROIs, bounding boxes, and pathologies. Our downloaded data from CBIS-DDSM includes DICOM images (Mass-Training Full Mammogram Images, Mass-Training ROI and Cropped Images, Mass-Test Full Mammogram Images, and Mass-Test ROI and Cropped Images) and description CSV files (Mass-Training-Description and Mass-Test-Description).

Data processing and visualization

The downloaded DICOM images were converted to PNG format, with full images being converted to 16-bit depth and ROIs and cropped images to 8-bit depth, using Pydicom and the Python Imaging Library (PIL). Following conversion, cropped images were manually removed from consideration. The remaining images, including both full images and ROIs, were named with corresponding patient IDs, indicating the left or right breast, and the image view (CC or MLO). This information was extracted from the description CSV files.

The named ROIs were mapped to the corresponding full images. This process started with verifying whether the dimensions of the ROI matched to the corresponding full image or not. If a match was found, the circled region, indicative of the white area within the ROI, was utilized to discern the abnormal region within the full image. This involved masking out the abnormal region and setting pixel values outside this area to 0. In cases where discrepancies were identified, the ROI was resized to match the dimensions of the full image before repeating the mapping process.

Subsequently, the execution of cropping a maximum of 598×598 pixel area from each mapped image involved: (1) determining the geometric center, denoted as $\mathrm{Center}_X, \mathrm{Center}_Y)$, of the abnormal region (the non-zero pixel region); (2) calculating the coordinates of the cropping area based on this geometric center and the maximum output size (598), ensuring that the cropping area remained within the image boundaries; and (3) executing the cropping process. The functions to calculate the center and coordinates were as follows:

$$(\mathrm{Center}_X, \mathrm{Center}_Y) = \left(\frac{\sum_{(x_i,y_i)\varepsilon S} x_i}{|S|}, \frac{\sum_{(x_i,y_i)\varepsilon S} y_i}{|S|} \right) \tag{1}$$

where $|S|$ represented the number of non-zero pixels, $\sum_{(x_i,y_i)\epsilon S} x_i$ denoted the sum of the *x*-coordinates of all non-zero pixels, and $\sum_{(x_i,y_i)\epsilon S} y_i$ denoted the sum of the *y*-coordinates of all non-zero pixels.

The functions to determine the coordinates of the cropping area were:

$$\mathrm{Top}_{\mathrm{left}}\,(l,t) = \left(\max\left(0, \mathrm{Center}_X - \frac{598}{2}\right), \max\left(0, \mathrm{Center}_Y - \frac{598}{2}\right) \right) \tag{2}$$

$$\mathrm{Bottom}_{\mathrm{right}}\,(b,r) = \left(\min\left(\mathrm{Center}_X + \frac{598}{2}, \mathrm{Width}\right), \min\left(\mathrm{Center}_Y + \frac{598}{2}, \mathrm{Height}\right) \right) \tag{3}$$

where width and height represent the mapped image size; top left and right bottom denoted the coordinates where to start and end the cropped images. All cropped images at 598×598 pixel were then manually reviewed, revealing 18 with unwanted white edges. The identified white backgrounds were removed by setting pixels in the corresponding white areas to 0.

All processed images were examined using the image.size function and classified into two categories: those equal to the desired pixels and those smaller than the desired size. For images smaller than the desired pixels, blank backgrounds were appended while keeping the center of the abnormal region unchanged. The full image size and the percentage of abnormal area per 598×598 pixel area, presented in Fig. 1B and C, were calculated using the image.size function from the PIL package and the formula *np.count_nonzero/np.array(input image).size * 100*. Subsequently, ggplot function from the tidyverse package was employed to visualize the full image size and the percentage of abnormal area per cropped image. Mann–Whitney U Test P-value was calculated with function *wilcox.test*. The next step involved augmentation, wherein flipping, rotation, zoom, shear, and shift were applied to generate five additional images for each extracted region using the ImageDataGenerator from the keras.preprocessing.image package (benign: 4575; malignant: 3905).These transformations were randomly

generated within specified ranges to increase data variety and better simulate real-world conditions. Specifically, images were rotated within a range of ± 20 degrees, shifted by up to ± 10% of the total pixel values, sheared by ± 20 degrees, zoomed in or out between 80 and 120% of their original size, and randomly flipped horizontally.

Model development

A new CSV file named all.csv was created to compile the paths of all 10,176 (1696 + 8480) processed images along with their respective pathology information extracted from the description.csv files. The pathology information was then encoded into labels, with the BENIGN and BENIGN_WITHOUT_CALLBACK categories mapped to 0 and the MALIGNANT category mapped to 1. This conversion was done using a dictionary and the LabelEncoder from the scikit-learn library. Subsequently, a custom dataset class, CustomDataset, was developed to handle image data loading and preprocessing tasks. This class extracted the file paths and labels, the pathology information, from the DataFrame obtained from all.csv. Using the Image.open() function from the PIL library, it loaded each image. The 1696 file paths from unaugmented images, extracted from all.csv, were then randomly split into 80% for training, 10% for validation, and 10% for testing[35,36]. Augmented versions of selected images in the training set were included as training input, while the validation and test sets remained unchanged.

Custom transformations include converting the image to a tensor, normalizing its values to [0, 1] by dividing by 65,535.0, and cropping images while preserving the center of the image remained unchanged to generate input images of 224 × 224, 299 × 299, and 448 × 448 pixels with function transforms.CenterCrop() when necessary. DataLoader was then applied to facilitate batch-wise data loading, shuffling the training set to introduce randomness during training, while keeping the validation and test sets unshuffled.

For input images with 598 × 598 or 299 × 299 pixels, the model architecture began with a sequence of convolutional layers incorporating Batch Normalization and LeakyReLU activation functions. The initial configuration of the model involved a single input channel, increasing the channel count from 1 to 3. Following the first two convolutional layers, max pooling was applied with a kernel size of 2 when input size equals to 598 × 598. The ultimate layer employed a 1 × 1 kernel to generate 3 output channels, utilizing LeakyReLU activation. The model then integrated an ImageNet pretrained Xception feature extractor with the exclusion of its final two layers or ResNet-50 feature extractor with exclusion of its final layer. This was succeeded by global average pooling and fully connected layers, nn. Linear (2048, 512), employing ReLU activation and dropout. Finally, the model produced its output through sigmoid activation. For input images with 448 × 448 or 224 × 224 pixels, the only two differences are (1) the pretrained model is ResNet-50, and (2) max pooling was applied with a kernel size of 2 when input size equals to 448 × 448.

For model training, early stopping was implemented when the validation loss continued to increase for three consecutive epochs after reaching its lowest value during training. Each epoch involved optimizing the model's parameters using the Adam optimizer with a binary cross-entropy loss function. Data was transferred to the GPU for faster computation. Throughout training, the model's predictions were compared to the ground truth labels to compute loss and update parameters via backpropagation. Accuracy was determined by comparing predicted labels to actual ones. Similarly, during validation, the model assessed performance on the validation dataset without parameter updates. Training and validation losses and accuracies were printed for each epoch to observe the model's performance, like convergence, possibly overfitting, and others. Model checkpoints were saved as *model_epoch_{epoch + 1}.pth* for future model restoration. The model's performance was evaluated on the test set using the checkpoint with the lowest validation loss. An example of how the best check point was selected is outlined in Fig. 4. As specified, the checkpoint selected for testing was at epoch 14, while early stopping occurred at epoch 18.

Performance evaluation

The model's effectiveness was evaluated using various metrics, including accuracy, precision, recall, F1 score, and a confusion matrix, as outlined in Table 2.

In Table 1, accuracy represents the proportion of correct outcomes out of the total, while precision measures the proportion of true positive results among all positive outcomes. Recall indicates the ability to correctly identify all relevant instances within a class. The F1 Score balances precision and recall, providing a single metric to evaluate model performance. The Confusion Matrix offers a detailed breakdown of a model's outcomes, showing the counts of true positives, true negatives, false positives, and false negatives. Additionally, we visualized the ROC curve and calculated the AUROC to assess the model's performance comprehensively. The ROC curve plots the true positive rate against the false positive rate at various classification thresholds, while AUROC summarizes the discriminative performance across all thresholds, with values ranging from 0 to 1. Higher values indicate superior discriminative performance.

Metric	Formula
Accuracy	(Number of correct predictions)/(total number of predictions)
Precision	(True positives)/(true positives + false positives)
Recall (sensitivity)	(True positives)/(true positives + false negatives)
F1 score	(2 * precision * recall)/(precision + recall)
Confusion matrix	Tabular representation of various predictions#

Table 2. Model evaluation methods[37]. # Confusion matrix directly presents the count of true positive, true negative, false positive and false negative instances.

Data availability

The official CBIS-DDSM dataset can be downloaded from: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629#2251662935562334b1e043a3a0512554ef512cad and https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629#22516629accaef0469834754b89af9e007760b10

Code availability

All codes applied in this research is available at https://github.com/lingliao/Transparency-in-CABCDTD/tree/main

Received: 11 April 2024; Accepted: 4 November 2024
Published online: 09 November 2024

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* **74**(3), 229–263. https://doi.org/10.3322/caac.21834 (2024).
2. Ho, T. H. et al. Cumulative probability of false-positive results after 10 years of screening with digital breast tomosynthesis vs digital mammography. *JAMA Netw Open.* **5**(3), e222440 (2022).
3. Redondo, A. et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *Br. J. Radiol.* **85**(1019), 1465–70. https://doi.org/10.1259/bjr/21256379 (2012).
4. Loving, V. A., Aminololama-Shakeri, S. & Leung, J. W. T. Anxiety and Its association with screening mammography. *J. Breast Imaging* **3**(3), 266–272. https://doi.org/10.1093/jbi/wbab024 (2021).
5. Stout, N. K. et al. Benefits, harms, and costs for breast cancer screening after US implementation of digital mammography. *JNCI: J Natl Cancer Inst* https://doi.org/10.1093/jnci/dju092 (2014).
6. Keen, J. D. Opportunity cost of annual screening mammography. *Cancer.* **124**(6), 1297–1298. https://doi.org/10.1002/cncr.31197 (2018).
7. Lotter, W. et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat. Med.* **27**, 244–249. https://doi.org/10.1038/s41591-020-01174-9 (2021).
8. Shen, L. et al. Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* **9**, 12495. https://doi.org/10.1038/s41598-019-48995-4 (2019).
9. Baccouche, A., Garcia-Zapirain, B. & Elmaghraby, A. S. An integrated framework for breast mass classification and diagnosis using stacked ensemble of residual neural networks. *Sci Rep* **12**, 12259. https://doi.org/10.1038/s41598-022-15632-6 (2022).
10. Busaleh, M., Hussain, M., Aboalsamh, H. A. & Amin, F. E. Breast mass classification using diverse contextual information and convolutional neural network. *Biosensors (Basel).* **11**(11), 419 (2021).
11. Nasser, M. & Yusof, U. K. Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. *Diagnostics (Basel).* **13**(1), 161 (2023).
12. Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M. & Atashi, A. Prediction of breast cancer using machine learning approaches. *J. Biomed. Phys. Eng.* **12**(3), 297–308 (2022).
13. Botlagunta, M. et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Sci Rep.* **13**, 485. https://doi.org/10.1038/s41598-023-27548-w (2023).
14. Omondiagbe, D. A., Veeramani, S. & Sidhu, A. S. Machine learning classification techniques for breast cancer diagnosis. *IOP Conf. Ser. Mater. Sci. Eng.* **495**, 012033. https://doi.org/10.1088/1757-899X/495/1/012033 (2019).
15. Velarde, O. M., Lin, C., Eskreis-Winkler, S. & Parra, L. C. Robustness of Deep Networks for Mammography: Replication Across Public Datasets. *J Imaging Inform Med.* **10**, 536. https://doi.org/10.1007/s10278-023-00943-5 (2024).
16. Logan, J., Kennedy, P. J. & Catchpoole, D. A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Sci. Data* **10**, 595. https://doi.org/10.1038/s41597-023-02430-6 (2023).
17. Wang, X. et al. Inconsistent performance of deep learning models on mammogram classification. *J. Am. Coll. Radiol.* **17**(6), 796–803. https://doi.org/10.1016/j.jacr.2020.01.006 (2020).
18. Hsu, W. et al. External validation of an ensemble model for automated mammography interpretation by artificial intelligence. *JAMA Netw. Open* **5**(11), e2242343. https://doi.org/10.1001/jamanetworkopen.2022.42343 (2022).
19. Lee, R. S. et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* https://doi.org/10.1038/sdata.2017.177 (2017).
20. Sawyer-Lee, R., Gimenez, F., Hoogi, A. & Rubin, D. Curated breast imaging subset of digital database for screening mammography (CBIS-DDSM). *Cancer Imaging Arch.* https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY (2016).
21. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324. https://doi.org/10.1109/5.726791 (1998).
22. Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning (ICML), 448–456 (2015). Retrieved from http://proceedings.mlr.press/v37/ioffe15.html.
23. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. *Proc. Int. Conf. Mach. Learn. (ICML)* **30**(1), 3 (2013).
24. Boureau, Y. L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In: proceedings of the 27th international conference on machine learning (ICML), 111–118.
25. Chollet, F. Xception: Deep learning with depthwise separable convolutions. Computer Vision and Pattern Recognition (CVPR), (2017). Available at: https://arxiv.org/abs/1610.02357
26. 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 (CVPR) (pp. 770–778) (2016). https://doi.org/10.1109/CVPR.2016.90
27. Lin, M., Chen, Q., & Yan, S. Network in network. arXiv preprint arXiv:1312.4400 (2013).
28. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).
29. Svetoslav, M., & Kyurkchiev, N. Sigmoid functions: Some approximation and modelling aspects.
30. Carter, K. J., Castro, F., Kessler, E. & Erickson, B. A computer model for the study of breast cancer. *Comput. Biol. Med.* **33**(4), 345–360. https://doi.org/10.1016/s0010-4825(03)00003-9 (2003).
31. Huang, H. et al. A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features. *BMC Bioinform.* **20**(Suppl 8), 290. https://doi.org/10.1186/s12859-019-2771-z (2019).
32. Kowal, M., Filipczuk, P., Obuchowicz, A., Korbicz, J. & Monczak, R. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Comput. Biol. Med.* **43**(10), 1563–1572. https://doi.org/10.1016/j.compbiomed.2013.08.003 (2013).
33. Zhang, T. & Liu, Yh. Optimal design of bionic flexible fixation system for MRI-guided breast biopsy. *J. Bionic. Eng.* **16**, 1116–1126. https://doi.org/10.1007/s42235-019-0123-3 (2019).

34. Ahmed, S., Groenli, T. M., Lakhan, A., Chen, Y. & Liang, G. A reinforcement federated learning based strategy for urinary disease dataset processing. *Comput. Biol. Med.* **163**, 107210. https://doi.org/10.1016/j.compbiomed.2023.107210 (2023).
35. Prinzi, F. et al. A yolo-based model for breast cancer detection in mammograms. *Cogn. Comput.* **16**, 107–120. https://doi.org/10.1007/s12559-023-10189-6 (2024).
36. Aly, G. H., Marey, M., El-Sayed, S. A. & Tolba, M. F. YOLO based breast masses detection and classification in full-field digital mammograms. *Comput. Methods Progr. Biomed.* **200**, 105823. https://doi.org/10.1016/j.cmpb.2020.105823 (2021).
37. Kai Ming Ting Confusion matrix. In *Encyclopedia of machine learning* (eds Sammut, Claude & Webb, Geoffrey I.) 209–209 (Springer US, Boston, MA, 2010). https://doi.org/10.1007/978-0-387-30164-8_157.

Acknowledgements

L.L. acknowledges the strong support from McDonnell International Scholars Academy, Washington University in St. Louis, MO, USA.

Author contributions

L.L. developed the codebase, conducted the experiments and wrote the manuscript. E.A. and L.L. conceived the project and reviewed the manuscript.

Declarations

Competing interests

L.L. is the founder of Biomedical Deep Learning LLC. E.A. declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-78648-0.

Correspondence and requests for materials should be addressed to L.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.