CSE 6367: Final Project

Analysis of Breast Cancer Lesion Classification Models Trained using Transfer Leaning

Shiska Raut
Department of Computer Science
The University of Texas at Arlington

***Abstract*—Deep learning has revolutionized the technological industry, with applications extending to a wide range of domains; however, the application of deep learning in the medical domain still needs to be improved, given the black-box nature of deep learning models and the scarcity of extensive publicly accessible datasets for training and evaluation. While using pretrained weights for transfer learning has facilitated model training on smaller medical imaging datasets, whether these models are making decisions based on relevant features remains to be determined. This study applies two effective transfer learning strategies to breast mammogram data and evaluates the models' decisions based on a post hoc explanation method.**

***Index Terms*—transfer learning, explainable machine learning**

I. Introduction

The use of deep learning has surged across various sectors in recent years, thanks to improved access to data, cloud storage, computational power, and advanced hardware for parallel processing. While research on applying deep learning to diagnostic medicine has seen a recent uptick, several factors within healthcare systems hinder the practical implementation of these tools in the medical domain. [1]. One of these factors is the lack of large publicly available medical datasets. Nevertheless, there have been an astounding number of studies [2] that have explored the application of deep learning in the detection and diagnosis of diseases such as Alzheimer's and Cancer using medical data [3]. There are also notable efforts made by organizations such as the National Institute of Health(NIH) to make medical data available to the public for research purposes without compromising patient privacy [4].

In addition, 'transfer-learning' has allowed researchers to train and evaluate models on medical imaging data without the requirement of significantly large datasets. Regardless of a model's performance, understanding the rationale behind its decisions has become crucial, particularly in high-stakes situations. Surveys show that clinicians might be willing to trade-off between accuracy and explainability and are more likely to trust models that are paying attention to relevant parts of the image as opposed to highly accurate models with low model explainability [5].

A. Post Hoc Vs Ad Hoc Explanations

The major difference between post hoc and ad hoc explanations is that post hoc explanations attempt to explain the model's decision after the decision has been made. Some examples of post hoc analysis methods are Saliency maps, GRADCAM, SHAP, and LIME. On the contrary, ad hoc explanations are built into the model's architecture such that the model's decision is based on the explanation generated during the process. Examples of models implementing ad hoc explanations are prototype-based models that rely on training examples to provide case-based reasoning.

Although ad hoc explanations are superior to and more reliable than post hoc explanations, this study does not explore models with ad hoc explanations due to time constraints. Nevertheless, the initial part of building both most post hoc and ad hoc explanation models involves training a robust feature extractor; therefore, the findings of this study focus on providing an insight into the most common transfer learning strategies for medical imaging data and their effect on the resulting model's robustness in meaningful feature extraction.

II. Problem Statement and Related Work

Breast cancer is one of the biggest causes of cancer-related mortality in women worldwide and as the number of screening programs increases, so does the workload of radiologists and clinicians [6]. The integration of deep learning tools for detection and classification from mammogram screening images can be a great aid to radiologists and clinicians. However, as cancerous regions may only consist of a small part of the entire image, training directly on whole image data generally does not yield promising results [7]. Nevertheless, there have been a few studies [5], [6], [8] that implement different strategies to alleviate this issue by training models on both fine annotations as well as whole image annotated data.

The most common imaging modality for breast cancer screening is Mammography and many transfer learning

strategies have been implemented to alleviate this 'needle in the haystack' problem of detecting and classifying lesions from whole image mammograms [6]. Recent studies that have utilized the CBIS-DDSM dataset to train their models are [9] and [7]. The authors of [7] implemented a two-step transfer learning strategy to obtain a model that can classify whole image mammograms. They first used Convolutional Neural Network(CNN) models trained on the ImageNet data(VGG and ResNet) and a transfer learning algorithm to build a "patch classifier" that can recognize a patch from a whole mammogram as background, benign or malignant. Secondly, weights from the "patch classifier" were used to initialize a whole image classifier which was trained on whole image mammogram data. The authors in [9] base their work on [7] but implement a 2-view classifier - as opposed to a single-view classifier used in [7] that takes each view as a separate image - and concatenates the result to obtain the final verdict from the whole image classifier.

However, neither of the studies conducted a comprehensive post hoc analysis of the initial patch classifier model to assess whether the model relied on relevant features for decision-making. This study uses two different transfer learning approaches to train a patch classifier. It evaluates the performance of the resulting models not only using classification metrics but also using a post hoc explanation method. The contributions of this study are:

1) Implementation of two different transfer learning pipelines to train a lesion classification model with pre-trained weights from VGG and ResNet.
2) Visualization of parts of the image considered important by the model during classification using the gradient-weighted class activation mapping (Grad-CAM) technique.
3) Comparison of activation maps generated by Grad-CAM for accurately classified data points to the regions highlighted by radiologists in the ground truth annotations.

III. CBIS-DDSM Dataset

The CBIS-DDSM dataset [10] is the largest publicly available mammogram dataset. A compressed version of this dataset downloaded from 'kaggle' was used in this study due to the limitation of computation and storage resources. The dataset is publicly available at https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset. The data consists of 4 classes: benign calcification, benign mass, malignant calcification, and malignant mass. The class 'benign without callback' was not included as a target class for this study as the class lacked confirmation with biopsy. The dataset contains two types of images: whole-image mammograms and ROI masks corresponding to each whole-image mammogram.

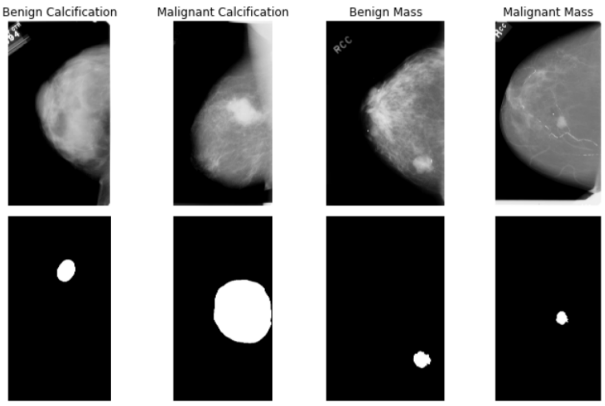

Fig. 1: Classes in CBIS-DDSM Dataset

IV. Methodology

A. Preprocessing and Data Preparation

The metadata, train, and test CSV files containing pathology information were used to organize the dataset into train and test folders. The original train-test split provided with the meatdata was used for this study in order to make a fair comparison with other studies. Training a lesion classifier required extracting smaller regions from the whole image containing lesions. This was done using with ground truth radiologist annotation provided with the data and implementation of the following pipeline: 1) a mammogram image was read, 2) the corresponding ground truth image(binary mask) was pulled up, 3) non-max suppression was applied on the binary mask to obtain the approximate position of the lesion in the whole image, 4) a bounding box was drawn around the lesion to obtain a cropped image of the lesion and the corresponding mask.

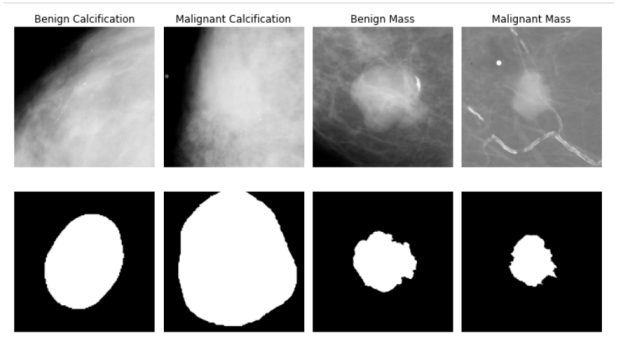

Fig. 2: Cropped Images for Training Patch Classifier

To build a whole image mammogram classifier by upscaling a patch classifier, it is necessary to train the patch classifier that can not only differentiate between various types of lesions but also differentiate a lesion patch from a background patch. To accomplish this, a distinct class for background patches was established. Background patches were generated by extracting 550x550-sized patches from background regions of randomly selected images belonging to the 'benign without callback' class.

As a result, the "patch" dataset used to train the patch classifier consisted of 5 classes: background, benign calcification, benign mass, malignant calcification, and malignant mass. The total number of images is listed in the table below:

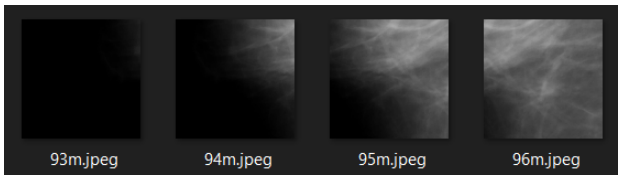

Fig. 3: Cropped Images for Background Class

B. Image Normalization

Due to the high dissimilarity between medical imaging data and source data for the pre-trained weights(ImageNet), normalizing the data based on ImageNet statistics - which would generally be a rule of thumb - could hurt the training process. In addition, the average intensity values between different classes in medical image data can significantly differ. Images with mass lesions tend to have brighter regions in the middle, whereas images containing a background tend to have darker regions and low-intensity values. Data was normalized separately for each class to account for the variation in brightness and contrast without hurting the prominent characteristics of each class. Table 1 contains the total number of data points in each class.

Total Number of Images		
Class	Train	Test
Background	644	310
Malignant Mass	599	144
Benign Mass	554	183
Malignant Calcification	375	110
Benign Calcification	358	110

TABLE I: Total number of Training and Test Images

C. Model Selection

As model simplicity is a key factor affecting explainability, most studies implementing ad hoc explanations for medical imaging data had utilized shallower architectures from VGG and ResNet [11]. Although post hoc studies such as [9] had utilized more complex architectures such as 'EfficientNetB0' for improved accuracy, most ad hoc implementations have avoided using complex models for simplicity and explainability. Although this study uses a post hoc explanation method, pre-trained backbone models for experimentation were picked to be suitable for both post hoc and ad hoc explainable models. The following models were selected in the initial part of the experiment: VGG16, VGG19, ResNet34, and ResNet50. Of the four, VGG19 and ResNet50 performed better in accuracy than VG16 and ResNet34. Due to limited time and computational resources, only one model from each family(VGG19 and ResNet50) was chosen for further hyperparameter tuning and evaluation.

D. Transfer Learning

Pytorch was used as the primary module for implementing network architecture. Pytorch Lightning was used to implement the transfer learning pipelines as it offers improved scalability across multiple pre-trained backbones. Two transfer learning strategies were used for training and fine-tuning the models, along with six sets of hyperparameters for each model and training strategy combination.

1) Configuring Optimizers, Loss and Learning Rates: Adam was used as the optimizer, with Cross-Entropy as the loss for each model. In CNNs that have already been trained, layers close to the input layer learn primitive features such as corners, edges, and shapes, whereas the layers close to the output layer learn features more specific to the data domain. Using the same learning rate for all layers could destroy features in the initial layers and be detrimental to the training process. Therefore, a differential learning rate was used depending on the depth of the latter. Values are discussed in detail in the 'Results' section.

2) Layer Training Strategies: For each pre-trained model, the fully connected block was removed and replaced by a fully connected layer to reflect the number of classes in the training dataset. Two different transfer learning strategies were used to train each model. A train-validation split of 0.85:0.15 was used for each model for fair comparison and consistency.

Strategy 1: Gradual Unfreezing: This strategy involves gradually unfreezing and training the network, starting from the fully connected output layer and moving inside the convolutional layers. The initial half of the convolutional layer was kept frozen throughout the transfer learning process, and the final half was gradually unfrozen and trained.

Strategy 2: Simultaneous Training: This strategy involves unfreezing all the layers(except the initial half, which was kept frozen) and training all layers simultaneously with a differential learning rate. Similar to strategy 1, the network's initial half was frozen throughout the process, and layers in the final half (including the fully connected layer) were trained. For both strategies, layers close to the output had a higher learning rate, and layers close to the input had a lower learning rate.

Early Stopping The number of original data points in the CBIS-DDSM dataset is extremely low in comparison to large image datasets like ImageNet and Stanford Cars. The small number of data points makes the models highly susceptible to overfitting. Therefore, an early stopping callback was added to the training module with the patience of 7 validation epochs. 'Validation Loss' was

used as the primary metric for early stopping with a minimum improvement constraint of 0.01 after seven epochs. During each part of the transfer-learning routine, checkpoints were saved for the best model(with the lowest validation loss) and used in the next part. This strategy prevents the use of an over-fitted model during subsequent phases of transfer learning and speeds up the training process, as the model does not have to be trained for the maximum number of epochs.

E. Model Evaluation

The models were evaluated in two ways: First, metrics such as accuracy, f1-score, and one-vs-rest average AUC (area under the curve) were used to evaluate the model's performance for the classification task. Second, the Gradient-weighted Class Activation Mapping(GradCAM) algorithm was used to obtain the class activation map with respect to the predicted class for each correctly classified sample, which was then used to calculate an average localization score for each model. This task was accomplished by taking the Intersection over the Union(IoU) of the activation map (after using thresholding to highlight highly activated regions) with the ground truth annotation, which is demonstrated in Figure 4.

Gradient-weighted Class Activation Mapping(GradCAM)

GradCAM is a post hoc model analysis algorithm that produces class activation maps for the predicted(or specified) class at a specific layer. Class activation maps for a specific layer help visualize the spatial regions from the feature map with high activation. When computed for the final convolution layer, regions of the image that contributed the most to the model's decision can be visualized [12].

V. Results

The sections below compare the performance of each model in terms of classification and localization scores. Results from this study were also compared with the performance of patch classifiers from the study [9]. However, since 110 images from the training set were used as test data for the calcification classes to compensate for the low number of test data points in the calcification class, a fair comparison could not be made.

A. Classification Evaluation

Accuracy

For each model and transfer learning strategy combination (four in total), Table 2 lists the results for the best-performing model in each category. Each of the models listed in the table had an initial learning rate of 0.0001 for the fully connected layer and 0.00003 for the final convolutional layer(except for ResNet50 with training strategy 2, which had a learning rate of 0.0003 for the final convolutional layer).

TABLE II: Model Accuracy on Test Data

Base Model	Training Strategy	Batch Size	Accuracy	Epochs[1]
VGG19	Strategy 1	16	0.66	47
ResNet50	Strategy 1	8	0.67	30
VGG19	Strategy 2	16	0.64	40
ResNet50	Strategy 2	16	**0.70**	22
Petrini et al.		40	**0.76**	

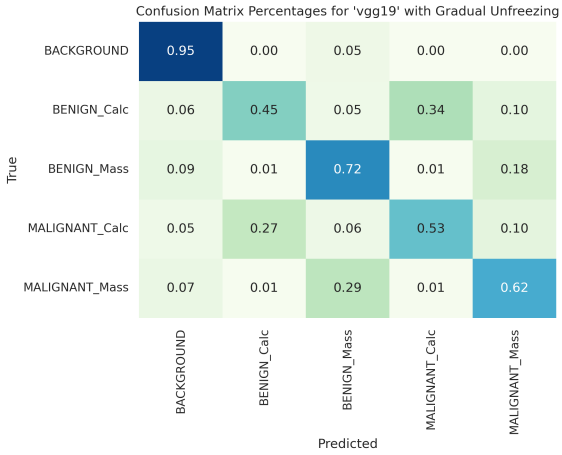

Fig. 4: Confusion Matrix VGG19 Gradual Unfreezing(Strategy 1)

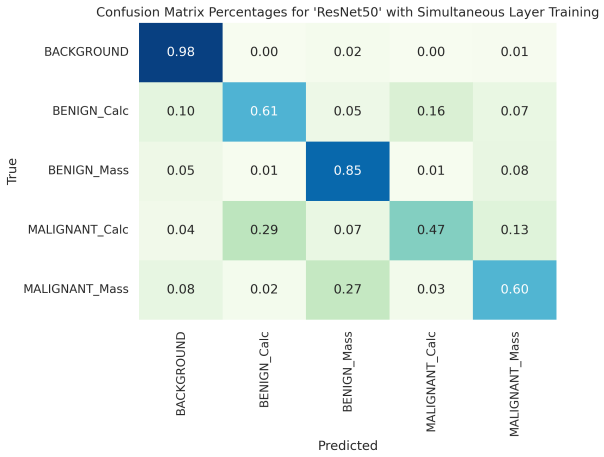

Fig. 5: Confusion Matrix Resnet50 Simultaneous Layer Training(Strategy 2)

The patch classification model with the ResNet50 backbone trained with simultaneous layer training(strategy 2) performed the best in terms of accuracy. For models with the VGG19 backbone, training with gradual unfreezing(strategy 1) resulted in a slightly better performance (0.66 vs. 64) than training with strategy 2.
newline

[1]Training epoch at minimum validation loss

Area Under the Curve(AUC)
Since the problem entails a multi-class classification, a 'one vs. rest,' strategy was used to calculate the AUC score for each class and averaged to obtain the average AUC score for the model, as shown in Table 3.

TABLE III: Class and Average AUC scores

AUC Score[2]						
Base Model	**Avg**	**B**	**BC**	**BM**	**MC**	**MM**
VGG19(S1)	0.78	0.94	0.69	0.79	0.72	**0.76**
ResNet50(S1)	0.78	**0.95**	0.7	0.86	0.64	0.75
VGG19(S2)	0.77	0.93	0.64	0.77	**0.74**	0.75
ResNet50(S2)	**0.81**	**0.95**	**0.76**	**0.87**	0.71	**0.76**
Petrini et al.	0.80					

The model with the ResNet50 backbone trained with strategy 1 had the highest average as well as per class 'one-vs-rest' AUC score for each class, except for the 'Malignant Calcification(MC)' class, for which model with VGG19 backbone trained with strategy 2 had the highest score.

Lesion Localization(LL) Score
The average lesion localization score was calculated by taking the average of the ioU scores for all accurately classified data points for each model. The results are demonstrated in Table 4.

TABLE IV: Model Accuracy on Test Data

Base Model	**LL Score[3]**	**Acc**	**Acc x LL Score**
VGG19(S1)	0.12	0.66	0.08
ResNet50(S1)	0.128	0.67	0.085
VGG19(S2)	0.1	0.64	0.064
ResNet50(S2)	**0.2**	**0.70**	**0.14**

The model with the ResNet50 backbone trained using strategy 2 was found to have the highest localization score as well as accuracy for patch classification, followed by ResNet50 trained using strategy 1, which had the second highest scores.

VI. Discussion

Accuracy and AUC Score
For both transfer learning strategies, models with the ResNet50 backbone performed better than models with the VGG19 backbone. These results suggest that ResNet50 might be better suited as a base model for transfer learning on breast mammogram data. In addition, models with the ResNet50 backbone also converged within a few epochs compared to models with the VGG19 backbone, which took significantly longer. However, ResNet-based models were more susceptible to over-fitting during training, especially when a higher batch size was used (32 and 64), whereas VGG19-based models were found to be less prone to over-fitting. Since the size of the test dataset was fairly small, the difference in performance between the best(0.70) and the worst(0.64) is not significant. Therefore, further research and additional regularization strategies could lead to better performance for both VGG and ResNet-based models.

The confusion matrices in Figures 4 and 5 show that both VGG and Resnet-based models performed fairly well in classifying background patches but had a poor performance in classifying patches with calcification. The imbalance in training data may have led to this, as the calcification classes had the least number of training samples. Another likely explanation is that calcifications are small and difficult to spot(even in patches) compared to more prominent masses. This is demonstrated in Figure 6.

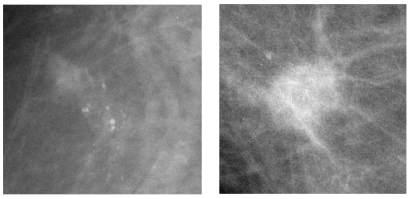

Fig. 6: A calcification(first image) vs a mass(second image)

Lesion Localization(LL) Score
Results show that models with high accuracy also had a higher localization score. However, even for the best-performing model(0.70 accuracy), the localization score was extremely low(0.2). These results suggest that, although the model correctly classified most of the samples for some classes, regions that correspond to the most important part of the image did not result in a high activation in the last layer of the network. Therefore, the model's decision may have been based on confounding information such as the background. Analysis of the attribution maps confirmed this, as a majority of the samples had high activation for background regions(as shown in figure 7).

As the difference between all models in both accuracy and localization score were not significant, it is difficult to conclude whether the training strategy affected the feature extractor's robustness. However, since all models performed poorly in terms of localization, larger datasets, and novel transfer learning techniques may be needed to develop models that are capable of extracting meaningful features from medical imaging data. While it may be true that features learned by a patch classifier can be scaled to develop a full image mammogram classification model, using a patch classifier that makes decisions based on non-

[2]B = Background; BC = Benign Calcification; BM = Benign Mass; MC = Malignant Calcification; MM = Malignant Mass; S1 = Strategy 1; S2 = Strategy 2

[3]The range for Lesion Localization(LL) Score is 0 - 1 with 0 corresponding to no localization for any samples and 1 corresponding to perfect localization for all samples.

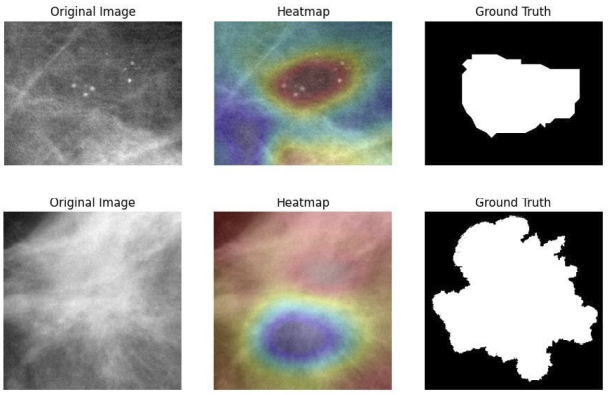

Fig. 7: Classification based on region of interest(first row) vs background information(second row)

important parts of the image may not result in a robust and reliable whole image classification model.

VII. Conclusion

Although there are numerous studies that propose high performing models for medical image classification, extending the application of these models into a clinical setting is challenging due to their black-box nature [13]. Trusting a model in high stakes scenarios requires analyzing the model's decision making process and evaluating whether the decisions are based on relevant features. This study shows that even models with a fair performance on medical imaging data trained using the most effective transfer learning techniques may be relying on non-important parts of the image most of the time while making a decision. Evaluating such models on publicly available test datasets may not be good enough as small datasets are rarely representative of the true population and prone to inherent bias.

VIII. Challenges

The use of pre-trained weights for training a model on a different classification problem can be a time and resource-saving approach. However, this is effective only when there is a substantial similarity between the source dataset (like ImageNet) and the target dataset. In medical imaging, datasets significantly differ from natural image datasets such as ImageNet, making transfer learning with medical datasets challenging, particularly when the dataset is small.

This study dedicated considerable time to identifying the appropriate normalization technique for smooth training and selecting optimal hyperparameters for good performance. Dealing with a small dataset, a major challenge involved combating overfitting during the training and fine-tuning of the deep learning model. Initially, data augmentation was employed to increase the number of training images, but this approach yielded diminishing results due to a lack of domain expertise. Further investigation revealed that most studies using transfer learning with the CBIS-DDSM had been conducted on the original dataset without augmentation.

IX. Future Directions

Training a robust and dependable patch classifier is imperative to transition from a patch classifier model to a whole mammogram image classifier. Exploring advanced transfer learning techniques may result in a model with enhanced accuracy and improved localization scores. A thorough examination of a diverse set of hyperparameters, coupled with implementing more robust regularization techniques, could improve overall model performance. Additionally, this research suggests potential issues with using post-hoc explanation methods like grad-CAM. Since the explanation is created after the model's decision and relies on correct classifications, it only considers explanations for accurately classified samples. Furthermore, while class activation maps identify image regions that strongly influenced the model's decision, they don't reveal how this information was utilized. This highlights the need for more reliable explanation methods when evaluating a model's decision.

References

[1] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *NPJ digital medicine*, vol. 5, no. 1, pp. 48–48, 2022.

[2] M. A. Gulum, C. M. Trombley, and M. Kantardzic, "A review of explainable deep learning cancer detection models in medical imaging," *Applied sciences*, vol. 11, no. 10, p. 4573, 2021.

[3] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," *Multimodal technologies and interaction*, vol. 2, no. 3, p. 47, 2018.

[4] M. S. Iqbal, W. Ahmad, R. Alizadehsani, S. Hussain, and R. Rehman, "Breast cancer dataset, classification and detection using deep learning," *Healthcare (Basel)*, vol. 10, no. 12, p. 2395, 2022.

[5] P. Xue, J. Wang, D. Qin, H. Yan, Y. Qu, S. Seery, Y. Jiang, and Y. Qiao, "Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis," *NPJ digital medicine*, vol. 5, no. 1, pp. 19–19, 2022.

[6] L. Balkenende, J. Teuwen, and R. M. Mann, "Application of deep learning in breast cancer imaging," *Seminars in nuclear medicine*, vol. 52, no. 5, pp. 584–596, 2022.

[7] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, pp. 12495–12, 2019.

[8] M. A. Al-masni, M. A. Al-antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system," *Computer methods and programs in biomedicine*, vol. 157, pp. 85–94, 2018.

[9] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, "Breast cancer diagnosis in two-view mammography using end-to-end trained efficientnet-based convolutional network," *IEEE access*, vol. 10, pp. 77723–77731, 2022.

[10] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, *Current Status of the Digital Database for Screening Mammography*, pp. 457–460. Dordrecht: Springer Netherlands, 1998.

[11] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin, "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," *Nature machine intelligence*, vol. 3, no. 12, pp. 1061–1070, 2021.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

X. Supplemental Figures

Fig. 8

vgg19 Classification Report with Gradual Unfreezing

	precision	recall	f1-score	support
BACKGROUND	0.77	0.95	0.85	123
BENIGN_Calc	0.62	0.45	0.52	110
BENIGN_Mass	0.63	0.72	0.67	135
MALIGNANT_Calc	0.59	0.53	0.56	109
MALIGNANT_Mass	0.66	0.62	0.64	142
accuracy			0.66	619
macro avg	0.65	0.66	0.65	619
weighted avg	0.66	0.66	0.65	619

Fig. 9

ResNet50 Classification Report with Gradual Unfreezing

	precision	recall	f1-score	support
BACKGROUND	0.83	0.96	0.89	123
BENIGN_Calc	0.54	0.51	0.53	110
BENIGN_Mass	0.60	0.90	0.72	135
MALIGNANT_Calc	0.55	0.36	0.43	109
MALIGNANT_Mass	0.76	0.55	0.64	142
accuracy			0.67	619
macro avg	0.66	0.65	0.64	619
weighted avg	0.66	0.67	0.65	619

Fig. 10

vgg19 Classification Report with Simultaneous Layer Training

	precision	recall	f1-score	support
BACKGROUND	0.75	0.96	0.84	123
BENIGN_Calc	0.64	0.33	0.43	110
BENIGN_Mass	0.65	0.64	0.65	135
MALIGNANT_Calc	0.56	0.60	0.58	109
MALIGNANT_Mass	0.59	0.64	0.61	142
accuracy			0.64	619
macro avg	0.64	0.63	0.62	619
weighted avg	0.64	0.64	0.63	619

Fig. 11

ResNet50 Classification Report with Simultaneous Layer Training

	precision	recall	f1-score	support
BACKGROUND	0.78	0.98	0.87	123
BENIGN_Calc	0.65	0.61	0.63	110
BENIGN_Mass	0.68	0.85	0.76	135
MALIGNANT_Calc	0.69	0.47	0.56	109
MALIGNANT_Mass	0.71	0.60	0.65	142
accuracy			0.71	619
macro avg	0.70	0.70	0.69	619
weighted avg	0.70	0.71	0.70	619

Fig. 12

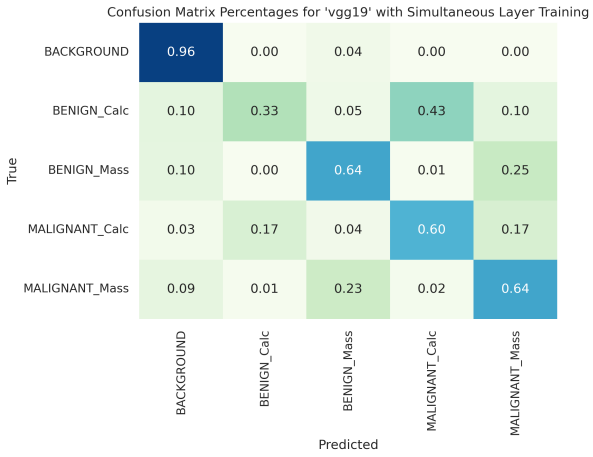

Fig. 13