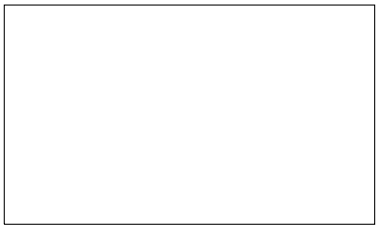
Graphical Abstract

Exploring the Influence of Attention for Whole Image Mammogram Classification

Marc Berghouse, George Bebis, Alireza Tavakkoli



Highlights

Exploring the Influence of Attention for Whole Image Mammogram Classification

Marc Berghouse, George Bebis, Alireza Tavakkoli

- Attention generally improves mammogram classification
- Newer attention methods don't consistently perform better
- Improvements due to attention depend on model architecture/complexity and characteristics of the image data
- Attention doesn't necessarily improve focus on the region of interest

Exploring the Influence of Attention for Whole Image Mammogram Classification

Marc Berghouse[a], George Bebis[b], Alireza Tavakkoli[b]

[a]*GPHS, University of Nevada, Reno, 89597, Nevada, USA*
[b]*Department of Computer Science and Engineering, University of Nevada, Reno, 89597, Nevada, USA*

Abstract

Attention is an important component of modern Convolutional Neural Networks (CNNs) that has been shown to improve baseline model performance for a wide variety of tasks. Attention has shown specific promise in the classification and segmentation of mammograms, but we have a limited understanding of why attention improves performance in these domains. In this paper, we present a robust comparison of different combinations of baseline models and attention methods at two resolutions for whole mammogram classification of masses and calcifications. We find that attention generally helps to improve baseline model performance. However, the extent of improvement is governed by a combination of model architecture and the statistical characteristics of the data. Specifically, we show that high amounts of pooling and model complexity may result in decreased performance for data with high variability. To better understand the effect of attention on mammogram classification, we employed LayerCAM, a hierarchical Class Activation Map (CAM) approach, to visualize where the network pays attention in the input image. This research provides statistical evidence that attention can improve the correlation between model performance and LayerCAM activation in the region of interest (ROI). However, these correlations are weak and variable, indicating that improvements in model performance due to attention are not necessarily caused by increased model activation near the ROI. Overall, our work provides novel insights to help guide future efforts in incorporating attention-based mechanisms for mammogram classification.

Email addresses: mberghouse@nevada.unr.edu (Marc Berghouse), bebis@unr.edu (George Bebis), tavakkol@unr.edu (Alireza Tavakkoli)

Keywords: Mammogram Classification, Attention, Deep Learning
PACS: 42.30.Tz, 87.57.rh
2010 MSC: 68T45, 62M45

1. Introduction

Attention has shown great promise in computer vision by guiding a model to focus on task-relevant local regions and channels [1]. In particular, it has been shown to increase performance in various classification [2], detection [3], and segmentation tasks [4]. Attention has also been shown to generally improve computer vision tasks in the field of medical imagery [2, 5, 6, 7, 8], where the region of interest often only comprises a small portion of the image.

Detecting and classifying abnormalities in mammograms is an active area of research that has potentially life-saving consequences [9]. It has been extensively studied, and a wide variety of deep learning model architectures have been proposed for whole image mammogram classification [10, 11, 12]. However, relatively less research has been performed on understanding how exactly attention impacts mammogram classification. Although it is generally accepted that attention may improve performance by helping the network to focus on relevant features [2, 13, 14, 15], there have not been any rigorous studies that investigate this in the context of mammogram classification. Many studies have shown that attention improves classification scores over the respective non-attention "baseline" models [16, 17, 18], but no comprehensive studies have been performed to compare combining different attention methods with various baseline models. Thus, it is unclear what attention models work best, and if attention generally leads to improvements in classification regardless of the baseline model. Furthermore, no studies have been performed on the impact of mammogram resolution or abnormality type on attention performance. Therefore, we do not currently have sufficient understanding of how attention generally impacts whole image mammogram classification.

To better understand the impact of attention on whole image mammogram classification, we have performed extensive experiments with three baseline models and three attention methods (i.e., 12 distinct models in total) on two different mammogram datasets (CBIS-DDSM and INbreast). Through this investigation, we provide a variety of theoretical and analytical conclusions that can help researchers design future attention-based models,

especially in the domain of mammogram classification. We find that attention generally improves the performance of the baseline models, but not always. Furthermore, the impacts of attention are dependent on dataset characteristics, and model complexity and structure. For datasets where the test set varies significantly from the training set (CBIS-DDSM), increased model complexity and the use of pooling lead to decreased performance, especially for calcifications. For datasets that have low variability between the train and test sets (INbreast), we find that pooling and complexity result in increased performance. Surprisingly, we also find that the most state of the art attention method doesn't consistently show the best performance. Finally, we only provide weak evidence to support the common assertion that attention improves model performance specifically by improving focus on the ROI. Thus, our work yields a wide array of results that improve our fundamental understanding of how attention impacts mammogram classification.

The preliminary results of this paper were presented at ISVC23 [19]. This paper significantly expands on the previously published results through the addition of a second mammogram dataset. The comparison of the results from each dataset adds significant complexity to the overall story of the impacts of attention on mammogram classficiation. Furthermore, we added other smaller parts such as an activation map analysis, supplementary figures, and expanded background and analyses that we weren't able to fit in the conference paper.

The rest of the paper is organized as follows: Section **2** reviews related work, providing a strong foundation for understanding our contributions to the literature. Section **3** describes our data and methods. In particular, we discuss data selection and preprocessing, selection of models, selection of attention models, and the training and testing workflow for each dataset. In Section **4**, we present our results and discussion. Specifically, we have considered the following issues: impact of attention on CNN performance, impact of model architecture on performance differences, impact of resolution on attention, impact of abnormality type on attention, and relationship between model activation and AU-ROC. Finally, Section **5** presents our conclusions and directions for future research.

2. Background

Attention is generally accepted to improve baseline performance through the adjustment of channel and/or spatial weights (Fig. 1). However, due

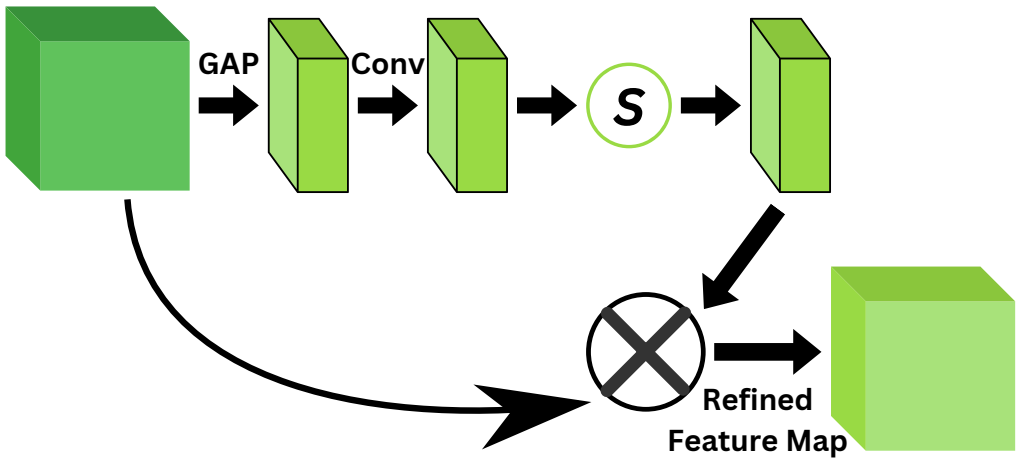


Figure 1: General structure of an attention module that returns an output of the same shape as the input. Most attention modules use global average pooling, convolution, and sigmoid attention. The final step is channel-wise or spatial-wise multiplication between the attended weights and the original feature map. Architectural details may be highly variable for different attention methods.

to a lack of rigorous comparisons between different attention methods with different baseline models, we luck some basic understanding of the sensitivity of attention to a variety of initial conditions. Mammogram classification studies often differ in the datasets used, image preprocessing methods, types of classification, model architecture, and training/testing methods [10, 11, 12, 16, 17, 18]. For example, many studies will classify patches of mammogram images instead of the whole image. Patch classifiers are effective because they don't require a decrease in the original image resolution, due to dividing the image into multiple sections. Attention has been shown to improve the performance of patch-based mammogram classification. Mao et al. performed a relatively comprehensive comparison of standard CNN baselines (Resnet50, Xception, and Densenet121), and found that the inclusion of the CBAM attention module improved the performance of each baseline [20]. However, patch classifiers alone are not helpful in diagnosing whole image mammograms, and often require a secondary model to process the outputs from the patch classifier [21]. This makes the training/testing workflow fairly complex. To reduce unnecessary complexity for the purposes of our comparative study, but still show clinically-relevant impacts, we chose to only train and test whole-image mammogram classifiers.

The literature on using attention-based methods to improve whole-image mammogram classification is relatively sparse. One of the earliest examples, introduced in [22], is a cross-view attention method (CVAM) to retain the latent information of all four views in a mammogram exam (Right and Left Cranial Caudal (CC) views, and Right and Left Mediolateral Oblique (MLO) views) for breast cancer classification. The authors tested a Resnet50 baseline model, Resnet50+CVAM, and Resnet50+CBAM (another attention method), and found that the Resnet50 baseline performed worse than the models with attention. In another whole-image classification study [16], it was shown that efficient channel attention (ECA) can be used to improve the performance of Resnet50, although the implementation of segmentation-based preprocessing methods and focal loss result in a more significant increase in performance. In addition to methods that fuse multiple views and simple applications of inserting attention into baseline models, some studies have used attention to develop multiple feature maps of the same image. For example, Xu et al. [17] present a multi-scale attention-guided network (MSANet) that uses multiple differently-sized feature maps and receptive fields to effectively focus on both the small and large features of a mammogram.

In contrast to whole image mammogram classification, more studies have demonstrated the benefits of using attention-based methods to improve performance in related tasks such as mass segmentation, mass localization, and breast tissue density classification. For example, an attention-guided dense-upsampling network (AUNet) for mass segmentation of whole image mammograms, which significantly outperformed a standard UNet, was proposed in [23]. A bilateral adaptive spatial and channel attention network (BASCNEt), which employs a similar cross-view attention method between the left and right images of the same view (CC or MLO), was proposed in [24] for breast density classification. By applying BASCNet to classical variations of Resnet, the authors found that attention improved performance in every case. ARFNet, which is a UNet-based encoder-decoder architecture that uses differently sized receptive fields and multi-scale attention, achieved state of the art performance on two separate mammogram datasets [25]. Furthermore, ARFNet showed significant performance increases when compared to a standard UNet architecture.

Two other primary ways the task of mammogram classification can vary are through image resolution and abnormality type. Mammogram classification performance has been shown to increase at higher image resolution [26].

Also, classification performance is generally better for masses than calcifications [27]. However, no previous studies have investigated how increases in resolution or changes in abnormality type may change the performance of attention-based models relative to their baseline model performance.

In addition to our investigation on the impacts of attention for multiple resolutions and abnormality types, our paper seeks to investigate exactly how attention impacts mammogram classification. It is widely accepted that attention improves classification performance by adjusting channel and spatial weights to help the model focus on the task-relevant region of interest [2, 6, 15, 16, 17, 28]. Theoretical results and activation heatmaps provide the vast majority of evidence for this claim [2, 6, 16, 29]. However, theoretical descriptions of attention don't allow us to fully understand or predict performance in various experimental settings. Because attention modules only comprise a small portion of the total CNN, a mathematical understanding of attention won't necessarily allow us to understand the overall function of attention within a model. Also, while activation heatmaps provide an effective visual representation of the model's focus in the feature space, they are susceptible to cherry picking, making it difficult to accept general conclusions from these results. Furthermore, recent research has shown that attention weights are generally uncorrelated with feature importance in the NLP domain [30]. Thus, there is no good evidence to indicate if attention generally improves scores by increasing focus on the ROI (regardless of dataset, classification objectives, and/or model structure). We seek to better address this issue, at least in the context of whole-image mammogram classification, by calculating the correlation between AU-ROC and the IOU between the mammogram mask and the activation heatmap.

3. Data and Methods

3.1. Data Selection and Preprocessing

We used the CBIS-DDSM [31] and INbreast [32] datasets to analyze the impacts of attention on mammogram classification (benign or malignant). For CBIS-DDSM, in order to understand how attention impacts model performance for different abnormality types, we trained and tested masses and calcifications separately. CBIS-DDSM contains 1592 (1231 train, 361 test) film images (average resolution of 5220 $\times$ 3138 pixels) of masses and their respective segmented regions of interest and pathology, and 1513 (1227 train, 286 test) film images of calcifications and their respective segmented regions

of interest and pathology. CBIS-DDSM contains and official train-test split, so we used the train split for training (with a further 95-5 train-validation split), and the official test split for testing. INbreast only contains 410 digital images (average resolution of about 3700 $\times$ 2900 pixels), so we did not split the dataset by abnormality type. INbreast does not have binary benign-malignant classifications, but the dataset does contain BI-RADS classifications. To convert the BI-RADS classifications to a binary classification, we assume that all BI-RADS values less than or equal to 4 are benign, and all BI-RADS values above 4 are malignant. Because INbreast doesn't have an official train-test split, we used five-fold cross-validation (CV) to test the performance of each model. For each fold, there were 328 training images and 82 testing images.

Mammogram classification is highly dependent on image resolution where malignant masses or calcifications may be only a couple pixels wide [33]. To understand how the resolution of mammograms impacts attention methods, we trained and tested models at resolutions of 500x300 and 1000x600 (height x width) pixels for both datasets.

Images were preprocessed according to standard methods in classification of whole-image mammograms with deep learning [34]. The images were normalized, segmented, cropped, flipped and enhanced with CLAHE. Segmentation was performed through simple masking and morphological operations to separate the breast from the background. The foreground was cropped then flipped across the central vertical axis to ensure consistent orientation. After flipping, CLAHE was used to improve image contrast. The preprocessed images were then resized to the target size for the respective experiment (either 500x300 or 1000x600 pixels). For training, we used brightness, rotation, contrast, and flipping (across the central vertical and horizontal axes) augmentations.

3.2. Selection of Models

To understand how attention influences the classification of mammograms, we first selected three baseline models (Resnet50 [35], Densenet169 [36], Regnetx064 [37]). For each baseline model family, we trained and tested three different model sizes (eg. Resnet38, Resnet50, and Resnet101), and selected the particular model that performed the best. A graphical summary of each baseline model architeecture is provided in Figure 2a. We chose Resnet50 and Densenet169 due to their high popularity for mammogram classification [38]. Resnet50, introduced as part of the Residual Network (Resnet)

family, was groundbreaking for its use of deep residual learning with "skip connections" that enabled the training of much deeper networks than was previously feasible. This architecture significantly improved performance on image recognition tasks by alleviating the vanishing gradient problem commonly encountered in deep neural networks. Densenet169, a variant of the Dense Convolutional Network (Densenet) architecture, introduced a novel approach where each layer is connected to every other layer in a feed-forward fashion. This dense connectivity pattern ensured maximum information flow between layers in the network, making it more efficient and reducing the number of parameters compared to other architectures of the time. Regnetx064 is an architecture developed by Meta which represents a more state of the art version of Resnet50. To improve upon the designs of Resnet50, the authors of the Regnetx064 algorithm introduced a novel architecture that utilizes a systematic approach to network design, optimizing the trade-off between network depth, width, and resolution. This approach, based on the principle of quantized linear relationships between these variables, allows for more efficient and scalable networks compared to prior architectures, offering improved performance with lower computational costs. All of our baseline models (as well as the attention methods) come from the Huggingface PyTorch Image Models (TIMM) repository [39]. It should also be noted that for Resnet50 we used the specific model version from the TIMM repository with the name "tv_resnet50". For each of these baselines, we considered three different attention methods (see below), yielding 12 different models in total.

3.3. Selection of Attention Methods

The main criteria for our choice of attention methods were modularity and popularity. Based on initial experiments, we identified three top performing attention methods that could be easily integrated with our baseline models: Squeeze-and-Excitation (SE)[40], Efficient Channel Attention (ECA)[41], and the Convolutional Bottleneck Attention Module (CBAM) [42]. A graphical summary of each attention method architecture is shown in Figure 2b. SE and ECA utilize channel attention to appropriately weight task-relevant channels. The SE attention method, introduced as a novel architectural unit within neural networks, focuses on adaptively recalibrating channel-wise feature responses by explicitly modeling interdependencies between channels. This method was innovative for its time as it significantly enhanced the representational power of a network with minimal computational overhead, leading to improved performance in various visual recognition tasks. ECA was

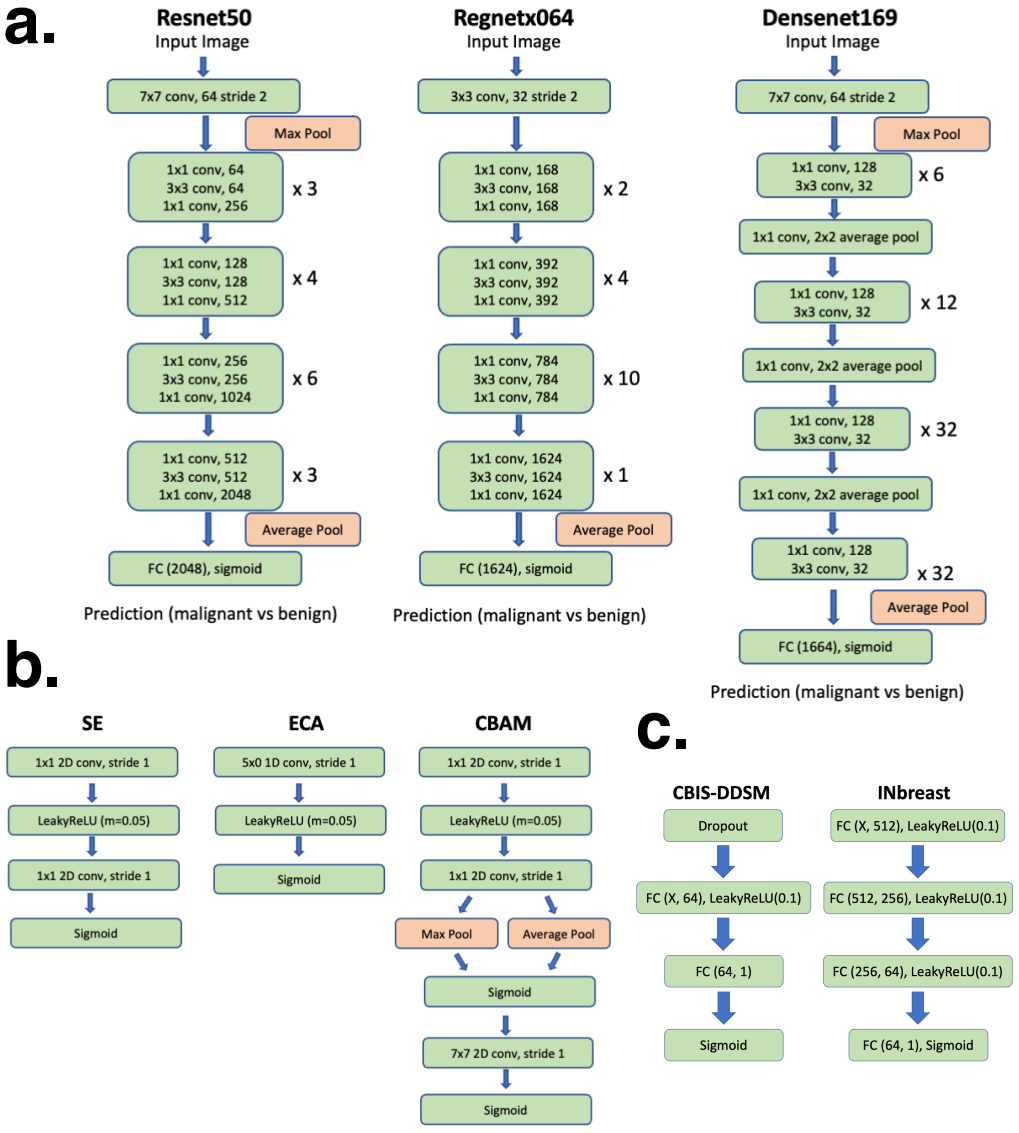


Figure 2: Baseline model and attention method architectures used in this study. (a) Baseline architectures. (b) Attention method architectures. (c) Classification head for CBIS-DDSM and INbreast.

developed to efficiently capture cross-channel interactions with a low computational cost. Unlike previous attention mechanisms, ECA dynamically determines the extent of the cross-channel interaction without dimensionality reduction (which occurs in both SE and ECA), offering an improved balance between performance and complexity. CBAM, distinguishes itself from SE and ECA through a sequential application of channel and spatial attention mechanisms. This structure was innovative for enhancing feature representation by focusing on informative features and suppressing less useful ones, significantly improving performance on various vision tasks with minimal computational overhead. For each baseline and attention combination, we performed ablation studies at low resolution to determine the best placement of the attention module within the baseline model. The attention modules were placed between dense blocks and transition blocks for Densenet169, and after the third convolutional layer in each block for Resnet50, and after the second convolutional layer in each block for Regnetx064. It should also be noted here that the standard version of Regnetx064 uses an SE module. So the model we refer to as Regnetx064+SE is the standard version, and the model we refer to as the baseline Regnetx064 model is the standard version with the SE modules removed.

Model complexity is an important characteristic of CNNs that can have a variety of impacts on classification performance. To understand the differences in model complexity for our various model combinations, we plotted the number of parameters for each model (Fig. 3). Between the attention methods, CBAM contains the largest number of parameters, followed by SE, then ECA. However, because the attention layers aren't repeated as much for Densenet169, there is only a small increase in the number of parameters (regardless of the attention method), meaning the complexity of all Densenet models are relatively equal. Consequently, Regnetx064+CBAM, Regnetx064+SE, Resnet50+CBAM and Resnet50+SE have the highest model complexities of the baseline+attention method combinations.

In addition to our primary comparison of modular attention methods in CNN baselines, we also present a comparison between these models and two models considered to be more state of the art (SOTA) - Twins-SVT [43] and DaViT [44]. We also ran preliminary tests on a number of other SOTA models, but were not able to train significantly complex models, or models that used specific input sizes, due to the requirements of our input size. Furthermore, many SOTA models did not show good initial performance, and as a result did not qualify for inclusion in our comparison study. Both

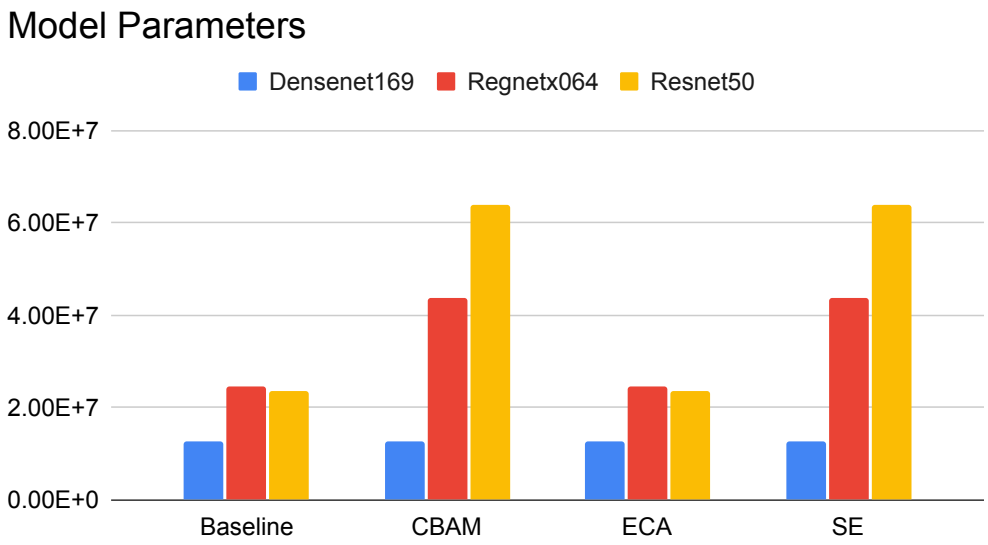


Figure 3: Number of parameters for each model variation used in this study.

DaViT and Twins depend on specific types of attention that align with their transformer architectures, so we were unable to insert our modular attention methods into these models in a way that resulted in improved performance. Thus, we present the results for Twins and DaViT as a separate comparison to highlight the performance of our other models in a more global context, but are unable to discuss the impacts of attention in the SOTA models.

3.4. Training and Testing Process

All models were trained with fine-tuning [45]. Specifically, we used pre-trained weights for each model, modified the classification head for binary classification, then trained the model on the mammogram images at a relatively low learning rate (1e-5 to 1e-4). The training/testing processes were slightly different for CBIS-DDSM and INbreast.

For CBIS-DDSM, the last layer of each model was removed and replaced with a dropout layer, a fully connected (FC) layer with 1624, 1664, or 2048 input nodes (based on width of baseline network) and 64 output nodes, leaky ReLU activation with a slope of 0.1, another FC layer with 64 input nodes and 1 output node, and sigmoid activation in the final layer (Fig. 2c) to produce a binary prediction (malignant or benign). For INbreast, we used a

similar setup for the classification head, except we didn't use dropout, and the width and depth of the fully connected layers are slightly different (Fig. 2c). We used pre-trained weights from Imagenet [46] to initialize the models. For all models, even those with attention, we used pre-trained weights from the respective baseline model. Although some baseline+attention combinations had pre-trained weights in the TIMM repository, we used the baseline weights to keep the comparison fair. Thus, the added attention sections of all models always began with default PyTorch weight initializations. We experimented with freezing certain parts of the models for fine-tuning, but found that it either resulted in no significant difference, or a slight decrease, in model scores. Therefore, no layers were frozen in our final experiments.

Hyperparameter tuning was done in a two step process. In the first step, we used Optuna [47] to tune a large search space via a tree-structured Parzen estimator [48]. These results narrowed the search space, then final tuning was done manually for each model. After manually tuning, we trained and tested each model within an approximate range of best hyperparameters over 40 times (for INbreast, 20 five-fold CV runs). For CBIS-DDSM, we used the top 30 scores for each model to calculate the results. For INbreast, we used the top 10 values for each fold, then averaged across the folds to get the mean values for each run.

The low resolution (500x300) models were primarily trained on a machine with an RTX 3070 (8GB VRAM), while the high resolution (1000x600) models were all trained on a machine with an RTX 3090 (24GB VRAM). We used a batch size of 10-16 (depending on VRAM constraints), and the Adam optimizer. For CBIS-DDSM, we used cosine annealing with warm restarts for our scheduler. No scheduler was used for INbreast since cosine annealing was found not improve scores, and just made training take longer.

4. Results and Discussion

4.1. Impact of Attention on Baseline Model Performance

Our comprehensive experimental analysis reveals that attention methods generally bolster the performance of baseline models for both CBIS-DDSM and INbreast datasets. We show that, averaged over all model runs, ECA and SE provide more consistent improvements (in accuracy, AU-ROC and F1) than CBAM for both datasets (Fig. 4a and Fig. 5a).

Looking at the model-specific results for the CBIS-DDSM dataset, we observe distinct performance patterns among the baseline models (Fig. 4b).

CBAM notably elevates Densenet169's performance, yet it uniquely diminishes the effectiveness of Regnetx64. Across the board, Densenet169 and Regnetx64 typically surpass Resnet50 in terms of performance. An intriguing insight emerges when examining the impact of attention on these models: while Densenet169 and Resnet50 consistently benefit from all attention methods, Regnetx064 only shows performance gains with ECA and SE. This suggests a nuanced interaction between model architecture and attention mechanisms, particularly highlighting Regnetx064's limited compatibility with certain attention methods. The only models that showed significant ($p<.05$) increases in AU-ROC averaged over all model variations were Densenet169+CBAM and Densenet169+SE (Fig. 4b). However, it is crucial to note that these improvements, while statistically significant, are modest in magnitude. Thus, the choice of model architecture and attention module, although impactful, does not drastically alter outcomes in mammogram classification within the context of the CBIS-DDSM dataset.

The underpinning rationale for these varied performances can be attributed to model complexity. Notably, all models show high performance on the train set for CBIS-DDSM (Fig. A.13), indicating that, to some extent, errors on the test set are a result of overfitting on the train set. High-complexity models, characterized by an extensive parameter count, are prone to greater amounts of overfitting, a particularly acute issue given the fine-grained nature of mammogram classification and the limited size of the CBIS-DDSM dataset [49, 50, 51, 52]. Resnet50 and Regnetx64, with their larger parameter counts, especially when augmented with CBAM and SE, are susceptible to this pitfall. Thus, the poor performance of Resnet50 (alone, and in combination with attention) and Regnetx64+CBAM may be partially explained by the large number of parameters.

In addition to our analysis of the impacts of attention, it is important to note here that many of these models achieve relatively high performance on the CBIS-DDSM dataset for whole image mammogram classification [26]. Specifically, for the average resolution of our input images, all models with AU-ROC of .79 or higher can be considered state of the art (SOTA). Furthermore, although our Densenet and Resnet model combinations have been experimented with before, we have not found any published work using ECA with Regnetx064. Thus, the novelty of our work also lies in the presentation of this high-performing Regnetx064+ECA model.

The results from INbreast show a slightly different story (Fig. 5). Despite its smaller size compared to CBIS-DDSM, INbreast yields better results, po-

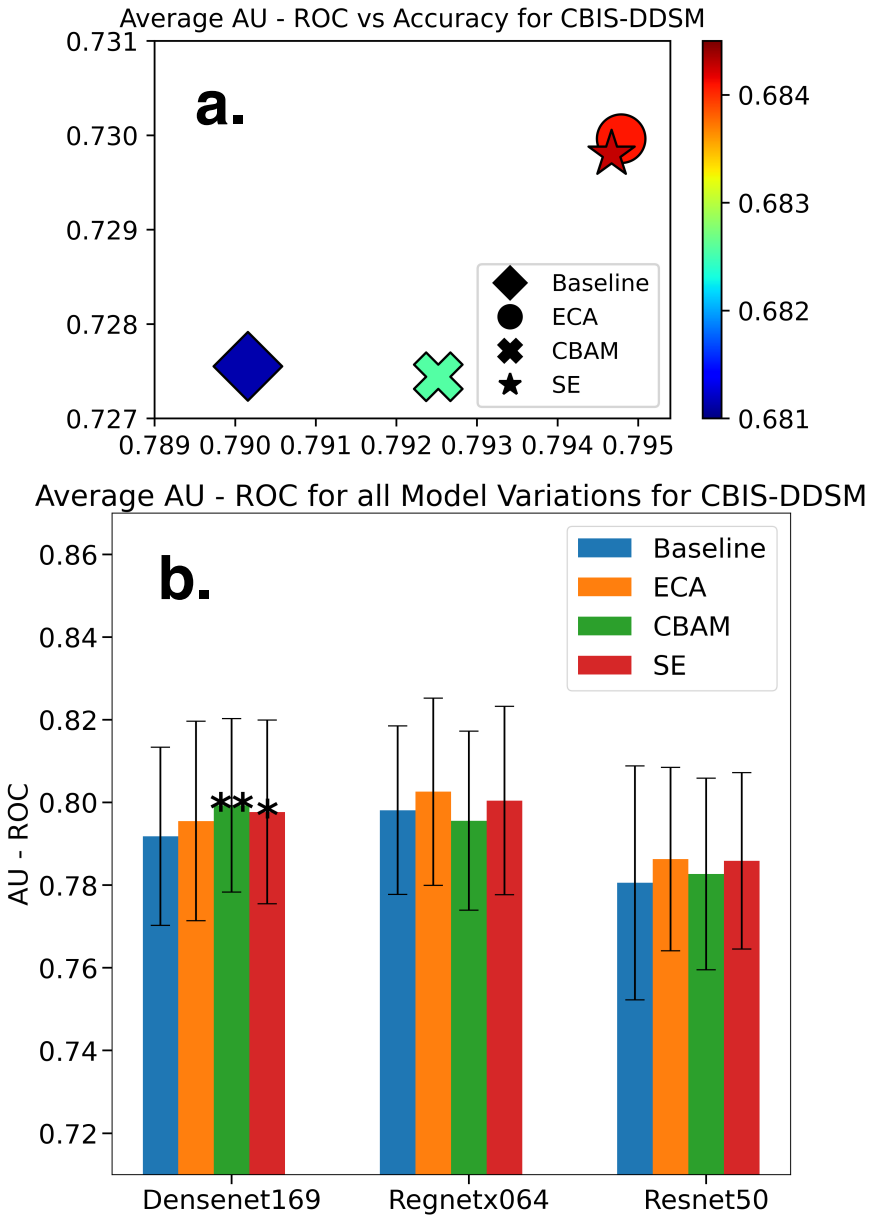


Figure 4: Average results for all CBIS-DDSM experiments. (a) Average AU-ROC vs accuracy with F1 colorbar. (b) Average AU-ROC for all model variations (network-specific breakdown of a). One star indicates a significant ($p<.05$) difference from the baseline. Two stars indicate a very significant ($p<.01$) difference from the baseline. These figures generally show that all attention methods improve baseline model performance, although ECA and SE offer more consistent gains than CBAM.

tentially reflecting clearer class distinctions or the higher quality and consistency of digital mammograms. Furthermore, the higher test scores generally indicate that overfitting is less significant for the INbreast dataset (in comparison with CBIS-DDSM). Between the different attention methods, SE shows superior performance (Fig. 5a). When compared to the results from CBIS-DDSM, ECA shows less improvement upon the baseline, but CBAM shows greater improvement upon the baseline. We still see the same general trend from the CBIS-DDSM results that Regnetx064 and Densenet169 have much better performance than Resnet50, but for the INbreast dataset, we see that Densenet169 offers less of an improvement when combined with attention than the other baseline models do (Fig. 5b). We theorize that these differences in results between datasets are primarily due to the differences in the characteristics of the data. Specifically, a greater distinction between classes, and/or higher quality image data allows more complex models to outperform less complex models.

Turning back to CBIS-DDSM, Figure 6 delineates the AU-ROC scores for each model variant across different resolutions and abnormality types. Attention yields the most substantial AU-ROC score increases for Regnetx064 and Resnet50 at 300x500 resolution for masses, and for Densenet169 at 600x1000 resolution for masses and 300x500 resolution for calcifications. The standout model performances include Densnet169 with any attention method for high-resolution masses, and Regnetx064 with ECA for high-resolution masses. Moreover, Figure 6 corroborates that ECA is the most effective attention method for CBIS-DDSM, followed by SE and then CBAM, based on the frequency of significant score improvements over the baseline.

In addition to a strict comparison of scores, we also found that models with CBAM were significantly harder to train than models with SE or ECA (for CBIS-DDSM). CBAM took a longer time to train than the other models, and the range of hyperparameters that produced good results was generally smaller (Fig. A.14). Thus, the relatively poor performance of CBAM on the CBIS-DDSM dataset may be a facet of its smaller window of acceptable hyperparameters rather than its architecture (high complexity and pooling). Furthermore, we found that the baseline models had the highest standard deviation of AU-ROC for different learning rates, indicating they have the smallest range of hyperparameters that reliably produce good results. Thus, attention results in less variation of mammogram classification scores due to changes in learning rate, which generally implies that models with attention will be less sensitive to hyperparameter choices.

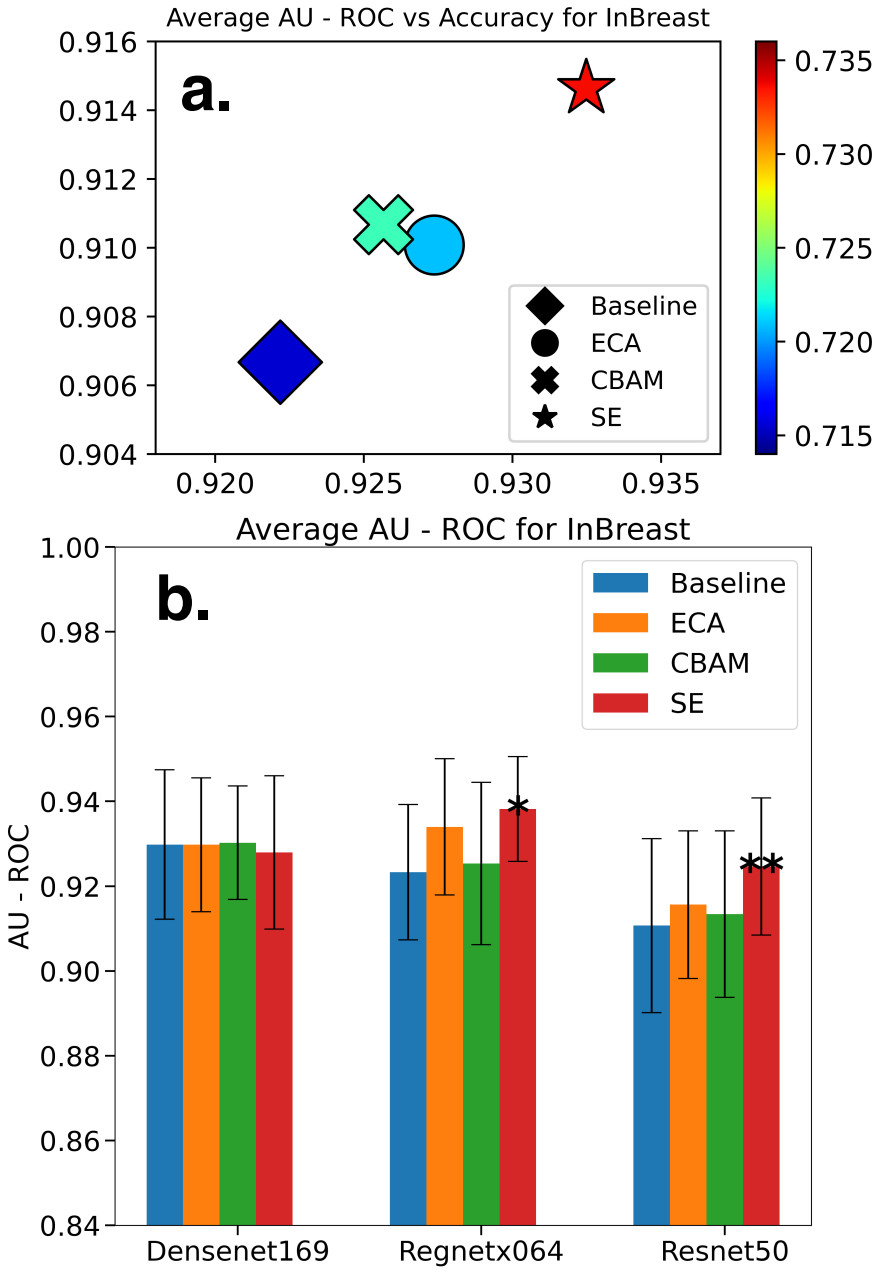


Figure 5: Average results for all INbreast experiments. (a) Average AU-ROC vs accuracy with F1 colorbar. (b) Average AU-ROC for all model variations (network-specific breakdown of a). These figures confirm attention generally improves baseline model performance for mammogram classification across multiple datasets. The results are similar to those from CBIS-DDSM, but ECA shows worse relative performance for INbreast.

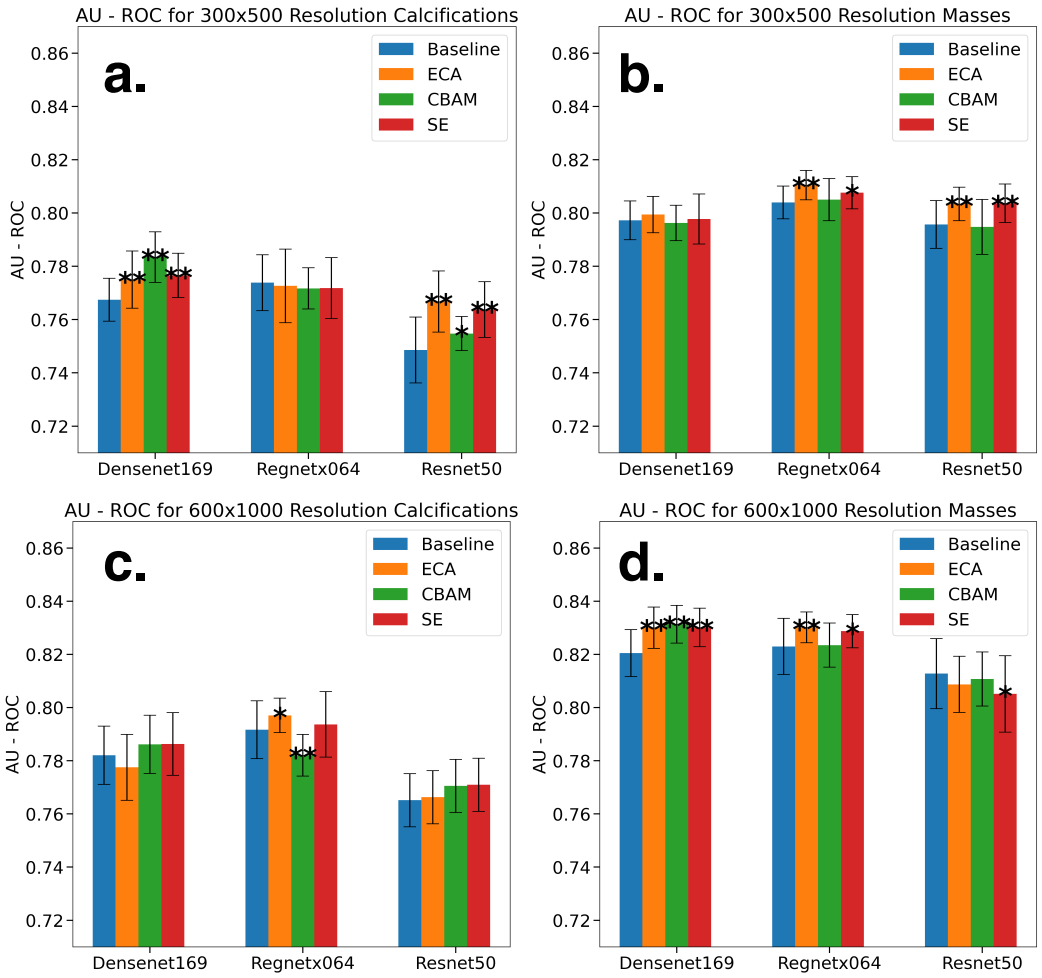


Figure 6: Average AU-ROC for each model variation under each training/testing scenario for mammogram classification of the CBIS-DDSM dataset. (a) Low resolution calcifications. (b) Low resolution masses. (c) High resolution calcifications. (d) High resolution masses. This figure generally shows improved performance due to attention, especially for ECA and SE.

4.2. Impact of Model Architecture on Performance Differences

To further explain these differences in performance, we took a closer look at the model architecture of each baseline method (Fig. 2). Densenet169 has a slower/more gradual increase in the number of feature maps (channels) than the other baseline networks do (Fig. 2a). Furthermore, all Densenet169 feature maps for each layer within a block are connected, meaning that no relevant information is lost within a block [38]. Densenet169 also ends up with less channels than Resnet50 (1664 vs 2048). Regnetx064 has less changes in the number of channels than Resnet50, and also ends up with less channels (1624 vs 2048). Given the smaller growth rates of channel size and the similar final channel size for Densenet169 and Regnetx064, the relatively poor performance of Resnet50 on the CBIS-DDSM dataset might be due to its relatively wide architecture and/or bigger changes in number of channels between each layer. While broader models generally correlate with enhanced performance [53, 54], the specific demands of mammogram classification, a task characterized by its fine-grained nature and smaller data volumes, suggest that a propensity for overfitting in wider models could detrimentally impact their effectiveness in this domain.

For INbreast, Resnet50 also performs worse than other baseline models. Since overfitting is generally less of an issue with INbreast, these results indicate that the baseline Resnet50 model is fundamentally less capable of classifying mammograms as good as Densenet169 or Regnetx064. One possible explanation for this is the extra max pooling layer in Resnet50 (compared to Regnetx064). Pooling may cause the model to focus more on only the most salient features, which causes a relative loss of fine-grained information. Also, the difference in the width of layers and repetition of blocks may lead to more refined feature representation or better gradient flow. Although a mathematical analysis of each model architecture is beyond the scope of this paper, it may prove useful to understanding exactly why Resnet50 seems to have a less effective latent space structure than the other baseline models.

The architectures of the ECA and SE attention methods are relatively similar as they both use convolution to adjust the weights of the channel dimension (Fig. 2b). The difference is that SE has two 2D convolutional layers, and ECA only has one 1D convolutional layer. The first part of CBAM (channel attention) is nearly identical to the SE module. However, CBAM employs max and average pooling before sigmoid activation, which may restrict the predictive power of small features. After channel attention, the new feature map is passed to a spatial 7x7 2D convolutional filter followed

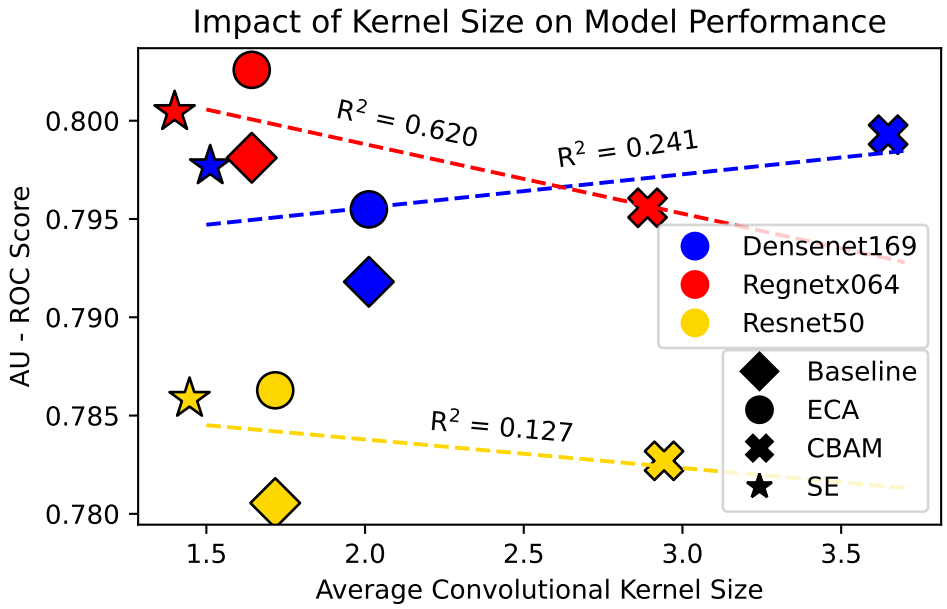


Figure 7: Impact of convolutional kernel size on model performance. The Regnetx models show the strongest correlation between kernel size and AU-ROC score.

by another sigmoid activation. This spatial convolution with a high kernel size further reinforces prediction of only the biggest spatial features, reducing the input of small or less relevant features that may still be important for mammogram classification. We further investigate this idea by plotting model performance against the average 2D convolutional kernel size (Fig. 7). We find that there is generally a small negative correlation between kernel size and AU-ROC score, although the results are highly dependent on the baseline model. There is a strong negative correlation for Regnetx064, a weak negative correlation for Resnet50, and a weak positive correlation for Densenet169. Thus, although kernel size may have a significant impact on model performance, it also clearly fails to explain the variance in scores between all models. Besides the size of the kernel, the type of convolution also plays a role in overall performance. Spatial convolution has been shown to result in overfitting [55], which might explain the relatively poor performance of CBAM on the CBIS-DDSM dataset. For the INbreast dataset, CBAM performance was much more consistent. Thus, unlike Resnet50, these results suggest that CBAM is not inherently incapable of matching the other

attention methods in their ability to improve the classification of mammograms, but that CBAM also has a high chance of overfitting depending on the quality of the data.

SE seems to strike a good balance between high level feature representation and overfitting, such that it is a clearly effective attention method for both the CBIS-DDSM and INbreast datasets. The greater model complexity (compared to ECA) means a slightly greater potential to overfit, which is why ECA slightly outclasses SE for the CBIS-DDSM dataset. However, for INbreast, where overfitting is less detrimental to performance, SE clearly outperforms ECA. The greater model complexity, and the greater dimension of initial convolution (2D vs 1D), mean that SE is able to retain more relevant spatial information than ECA is.

Generally, our results seem to suggest that more complex networks that use more pooling are likely to result in overfitting for mammogram classification. However, when there is less inter-image variability, or a greater distinction between classes (such as the case with INbreast), overfitting has less of a negative impact on classification performance. Although Densenet169 uses more pooling than any other architecture, it generally shows better performance than Resnet50 across both datasets because of the inter-connectedness of all feature maps, which ensures that small-but-important features are not lost. Thus, we posit that pooling is only problematic if it reduces the weights of small features that are highly relevant to the classification task. If the underlying network architecture promotes strong connectivity between layers, then pooling may act to improve model performance, or at least doesn't represent an apparent hindrance to model performance.

4.3. Impact of Attention at Different Resolutions

Next, we investigated the impacts of attention on classification for different mammogram image resolutions. For CBIS-DDSM, our results indicate that attention has a greater impact at low resolution (Fig. 8a), although this is largely due to the much more significant increase in AU-ROC due to attention for Resnet50 at low resolution (Figs. 8b and 8c). Regnetx064, however, shows a slightly greater increase in AU-ROC when combined with ECA and SE at high resolution than it does at lower resolution.

As previously discussed, the general explanation for the relatively poor performance of Resnet50 across both datasets is that it has both high model complexity and an inefficient latent space utilization. For CBIS-DDSM, at low resolution, all attention methods are able to more effectively guide

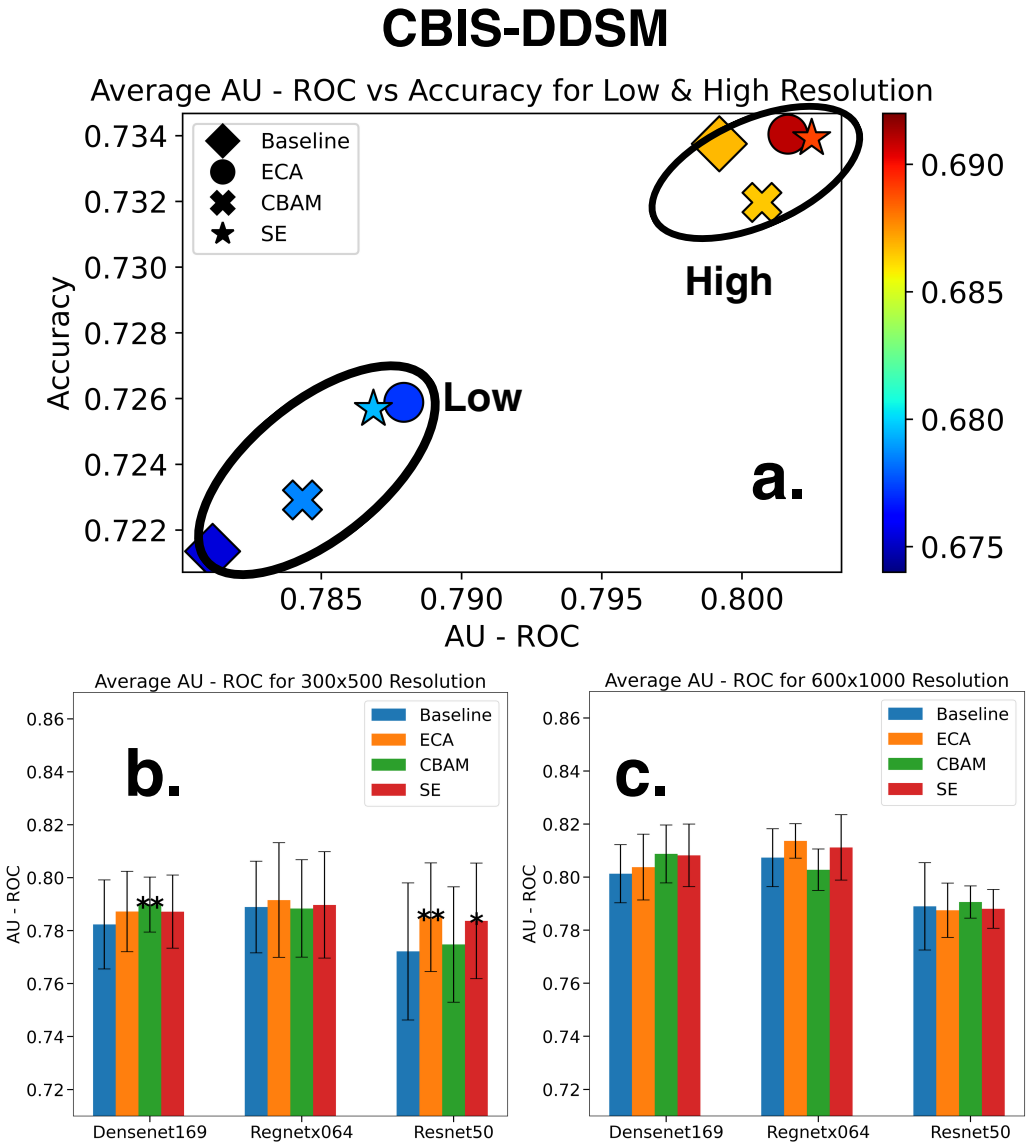


Figure 8: Average scores of each model for each resolution for the CBIS-DDSM dataset. (a) Average AU-ROC vs. accuracy with F1 colorbar. (b) Average AU-ROC of each model variation for low resolution. (c) Average AU-ROC of each model for high resolution. These figures generally show higher scores at high resolution, and a higher impact of attention at low resolution.

the latent space structure (focus on task-relevant features). At high resolution, the latent space structure for all models is improved by the addition of data/more learnable features. This results in significant increases from low to high resolution for the Resnet50 baseline and Resnet50+CBAM. However, we don't see any real improvement between resolutions for Resnet50+ECA and Resnet50+SE. Although the resolution increase, and likely the attention methods, improve the focus on task-relevant features, overfitting is so strong that the a small improvement in latent space structure does not lead to increased scores. Essentially, the scores for Resnet50 at high resolution have reached a limit - simple additions to the architecture, without changing the underlying architecture, are unlikely to offer improvements. However, like CBAM, it is also plausible that the "best" set of hyperparameters were not well determined for Resnet50 at high resolution, especially since the Resnet50 baseline showed the highest standard deviation of score based on choice of learning rate (Fig. A.14). Although significant effort was made to ensure the best hyperparameters for each model were used, given the large number of models tested, it is conceivable that some, especially those at high resolution, were not trained in a way to produce the best possible scores.

For the INbreast dataset, attention offers more of an improvement to the baseline score at high resolution (Fig. 9). Also in contrast to the CBIS-DDSM results, all attention methods offer strong improvement to Resnet50, especially at high resolution. Because the INbreast dataset is more generous for models that tend to overfit, there is less of a limit on how much an improvement in latent space structure can actually increase scores. Another primary difference between the results from the two datasets is that for INbreast, none of the attention methods offer improvements to the Densenet169 baseline. The likeliest explanation for this is that Densenet169 doesn't have enough learnable parameters. Densenet169 has the lowest model complexity, and the addition of attention methods cause the smallest increase in model complexity (compared to the other baseline models). Thus, because the scores for INbreast are already so high, the addition of attention models to Densenet169 is not able to shift the feature space enough to capture the complex relationships required to produce better scores.

Besides looking at how attention improves scores at different resolutions, we can also observe some general trends in the impact of resolution on mammogram classification. Clearly, classification scores are generally better at a higher resolution. Furthermore, there is much less variability for all model scores at high resolution. This has important implications for mammogram

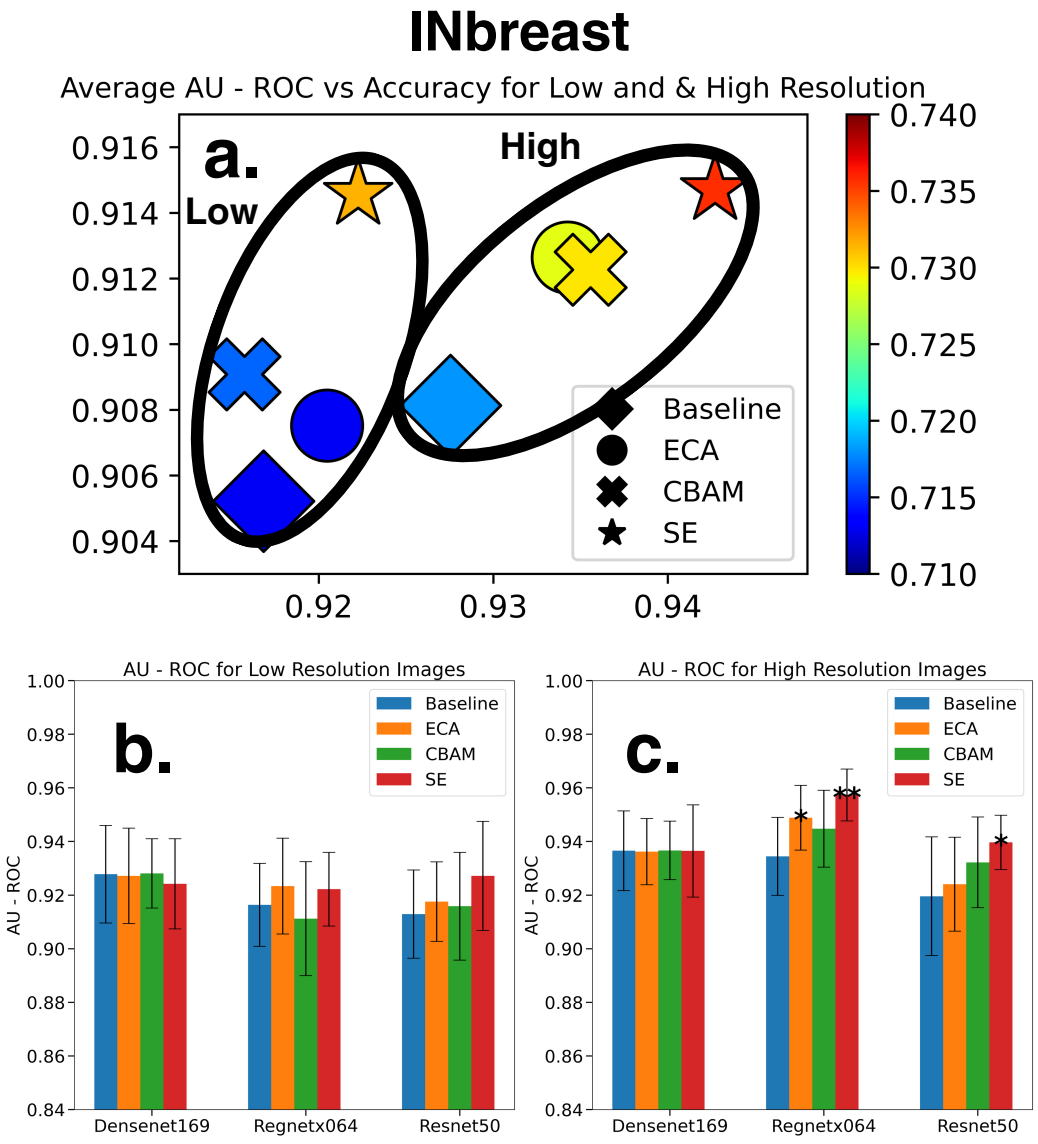


Figure 9: Average scores of each model for each resolution for the INbreast dataset. (a) Average AU-ROC vs. accuracy with F1 colorbar. (b) Average AU-ROC of each model variation for low resolution. (c) Average AU-ROC of each model for high resolution. These figures generally show higher scores at high resolution, and a higher impact of attention at low resolution.

classification, since images are often heavily downsized to fit in various models [12, 13, 14, 16].

4.4. Impact of Attention for Different Abnormality Types

For the CBIS-DDSM dataset, we also investigated the impact of attention on classification of different abnormality types. In general, we found that SE and CBAM result in greater improvements for calcifications than for masses (Fig. 10a). However CBAM also results in a large decrease in AU-ROC for calcifications when combined with Regnetx64. Thus, the performance of CBAM is more reliant on baseline model architecture for calcifications than for masses. ECA doesn't show any significant difference between scores for calcifications and masses. For Densenet169 and Resnet50, we found that attention results in greater score increases for the classification of calcifications than of masses (Figs. 10b and 10c). However, for Regnetx64, the opposite is true. One possible explanation for this trend is that Regnetx64 uses much less pooling than Densenet169 and Resnet50, meaning its architecture may be more favorable for small features such as calcifications. Thus, due to the constraints of resolution, attention is unable to offer as much of an impact in directing an already-good baseline model to more task-relevant features.

For calcification classification, the model scores may be significantly limited by the resolution of the image. Even at a resolution of 1000x600, some calcifications may only comprise very few pixels. Consequently, Resnet50 and Densenet169, which both use a large amount of pooling, do not see much of a performance gain between low and high resolution (Fig 8). Regnetx64, which uses less pooling, sees much more significant increases in scores at high resolution. These results suggest that even at high resolution, too much pooling can result in a loss of relevant information for calcifications. Densenet169, which uses the largest amount of pooling, shows a relatively small increase for calcifications going from low to high resolution, but the largest increase in scores for masses going from low to high resolution (Fig. 6). The large amount of pooling may allow the model to focus on more relevant features for masses, but with calcifications, the pooling causes too much of a loss of information, even with the interconnectedness of the dense layers. This loss of information likely doesn't significantly impact Densenet169 at low resolution because the small features that would be lost to pooling have already been lost during downsizing.

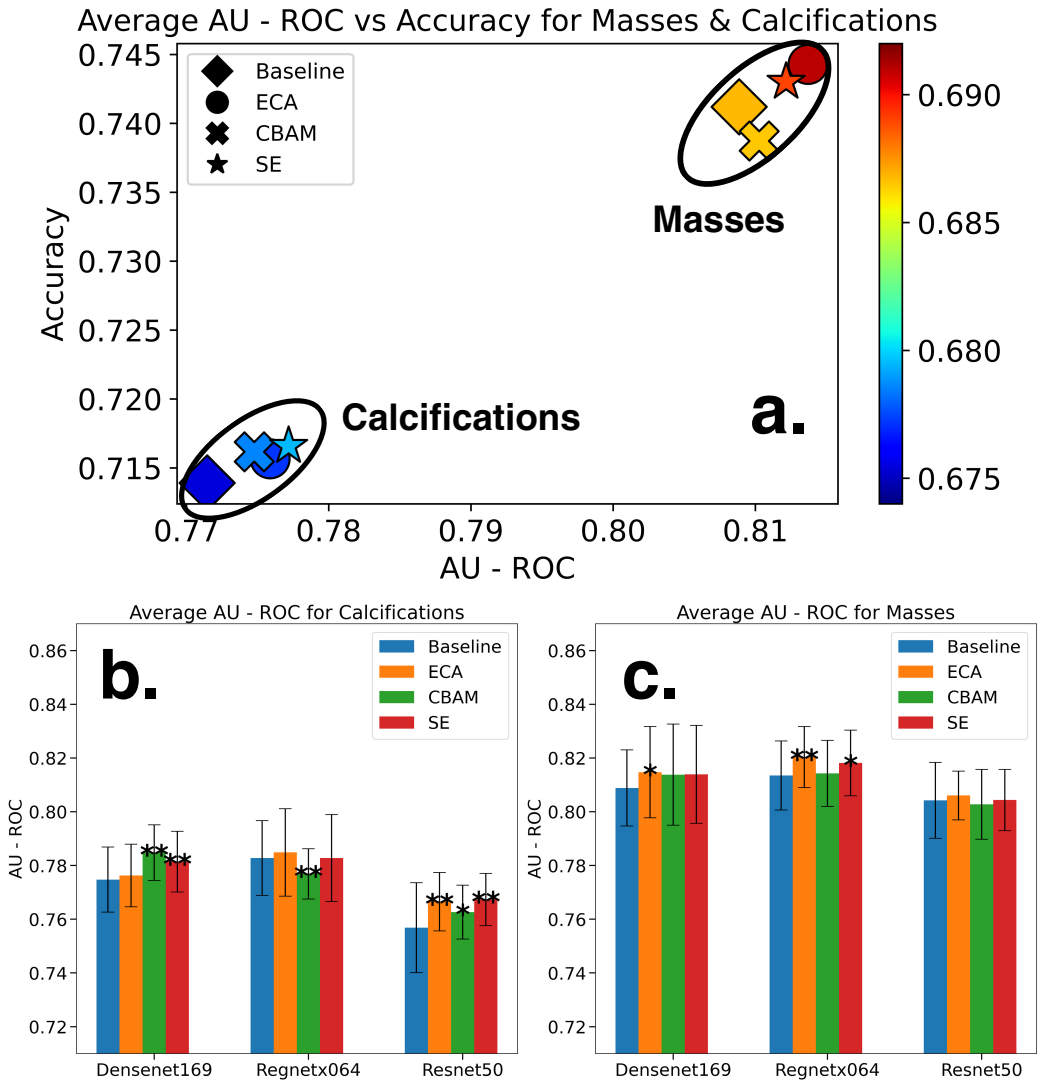


Figure 10: Average scores of each model for masses and calcifications for the CBIS-DDSM dataset. (a) Average AU-ROC vs. accuracy with F1 colorbar. (b) Average AU-ROC of each model variation for calcifications. (c) Average AU-ROC of each model for masses. These figures generally show higher scores for mass classification, and a higher impact of attention for calcifications.

4.5. Exploring the Link Between Model Activation and Attention

To investigate how well attention methods guide the baseline models to focus on the most promising areas in the input image, we analyzed the relation between class activation maps (CAMs) and AU-ROC. CAMs are frequently used in computer vision to visualize the model focus through heatmap overlays [56, 57, 58]. Clasically, CAMs are produced from the gradients of the final convolutional layer of a CNN. However, we used LayerCAM [59], which is a flexible CAM algorithm that uses information from multiple layers to generate accurate heatmaps. Because of the high variability in model architectures tested in this study, we determined that LayerCAM was an effective method to allow for consistent comparisons due to its versatility through use of multiple layers. Furthermore, LayerCAM has shown improved performance over other CAM algorithms, especially in the context of fine-grained objects. We then binarized the CAMs for each test image and calculated the Jaccard index, or IOU, between the binarized heatmap and its respective mask. We used a simple ascending search to determine the threshold for heatmap binarization that results in the highest IOU score.

LayerCAM activation heatmap analysis for the CBIS-DDSM dataset (Fig. 11) reveals a few interesting trends, although not necessarily at a statistical level. The left series of images are for a malignant calcification, with the only interesting result being the false negative (FN) prediction by Regnetx064+CBAM. In this case, we can see much less model activation for any part of the heatmap, but especially less in the region of interest (white mask). The right figures show a small malignant mass. In this case, all baseline models show FN predictions, along with Regnetx064+CBAM. All other baseline+attention combinations result in true positive predictions, indicating that attention may help specifically with prediction of small masses. However, since testing this theory at a statistically relevant scale was beyond the scope of this paper, more investigation is needed to conclude if the size of masses impacts the performance of attention.

To understand the model activations at a statistical important level, we calculated the mean IOU (over all test images) between the mammogram mask and the activation heatmap, and plotted these against AU-ROC for CBIS-DDSM (Fig. 12). Our results indicate that all attention methods show more correlation between IOU and AU-ROC than the baseline models (Fig. 12a), but that this relationship is much more significant for ECA (Fig. 12b). This implies that attention does to some extent cause a given model to focus on more relevant features in mammograms. However, given the low

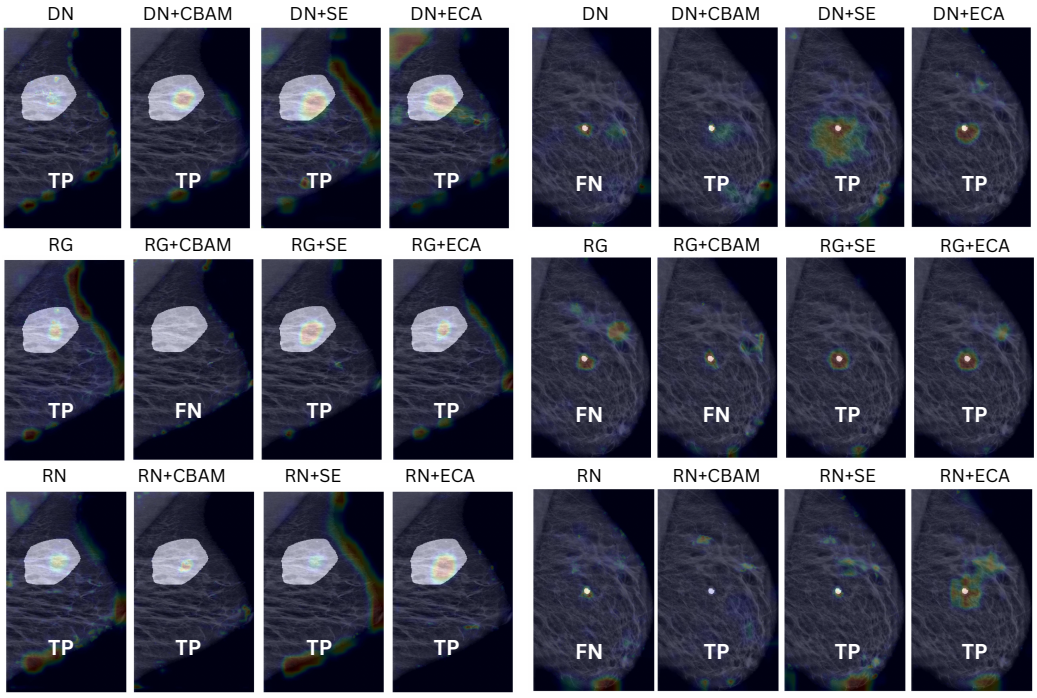


Figure 11: LayerCAM activation heat maps for CBIS-DDSM. "DN" corresponds to Densenet169, "RG" corresponds to Regnetx064, and "RN" corresponds to Resnet50. The figures on the left side show a mammogram with a malignant calcification. For these figures, the only false negative (FN) came from Regnetx064+CBAM. The figures on the right show a mammogram with a malignant mass. For these figures, the attention methods generally improve classification, but there is little to no apparent difference in model activation.

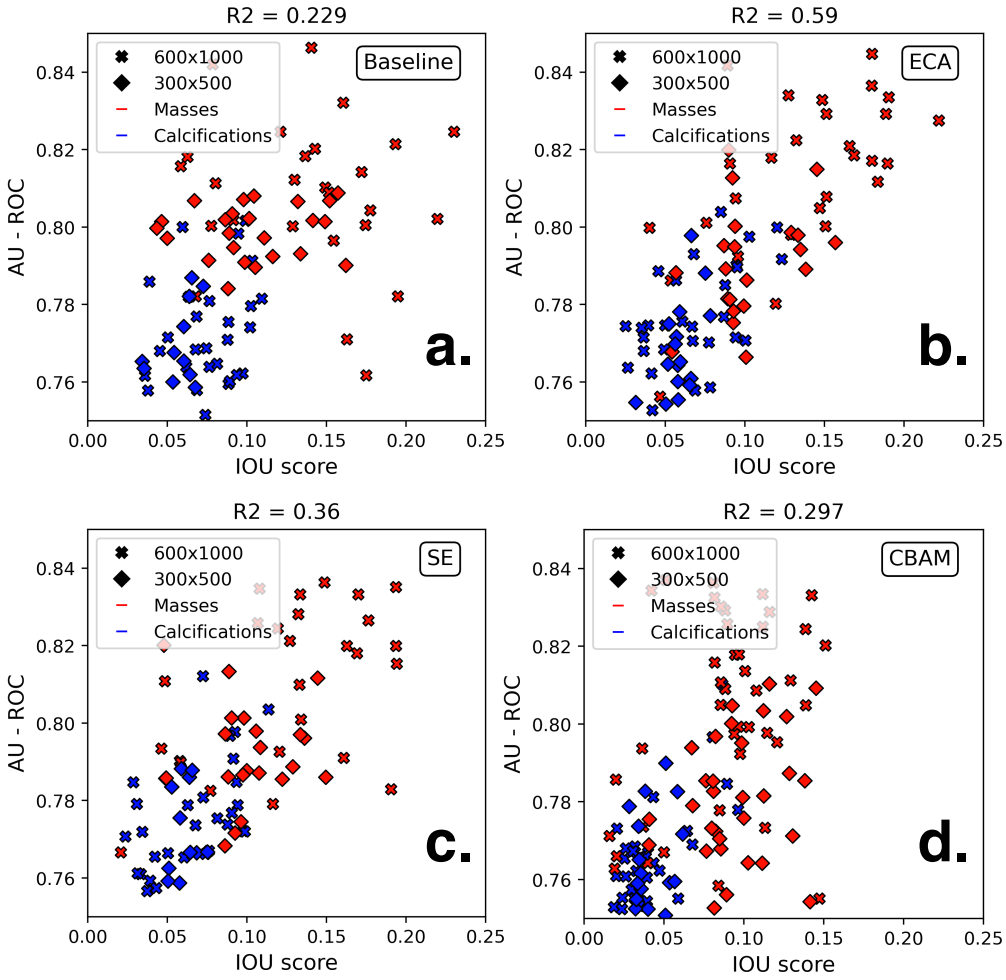


Figure 12: Relationship between IOU and AU-ROC. Correlation coefficient (R2) is given at the top of each graph. (a) All baseline models. (b) All models with ECA. (c) All models with SE. (d) All models with CBAM. This figure generally shows that attention methods provide a stronger correlation between IOU and AU-ROC, which indicates that attention helps the models focus on task-relevent features.

	CBIS-DDSM		INbreast	
	Low Res	High Res	Low Res	High Res
Twins-SVT-small	0.795	0.805		0.945
DaViT-small	0.795	0.779		
ECAregnetx064	0.792	0.814	0.925	0.933
SEregnetx064	0.790	0.811	0.922	0.932
ECAdensenet169	0.787	0.804	0.921	0.945
CBAMdensenet169	0.790	0.809	0.926	0.941
SEdensenet169	0.787	0.808	0.921	0.954
SEresnet50	0.784	0.788	0.925	0.925

Table 1: Caption

values of the statistical correlations between AU-ROC and IOU, this does not provide strong evidence that attention primarily increases model performance through better focusing the model on the ROI. Furthermore, we did not find any significant relationships between IOU and AU-ROC for the INbreast dataset, even though attention showed a clear impact on performance. Especially with complex images like mammograms, more research is needed to determine if improving focus on the ROI, all else considered equal, results in a increase in classification performance.

Generally, these results also show that mass classification in high resolution mammograms provides the greatest IOU scores. Also, at high resolution, the correlation between AU-ROC and IOU slightly increases for all attention methods. Thus, at higher resolutions, attention should theoretically provide a greater increase in scores. Although we observed the opposite trend for CBIS-DDSM (Fig. 8), we did observe this trend of increasing impact of attention at high resolution for INbreast (Fig. 9). Thus, the nature of the relationship between IOU and AU-ROC may depend on the specific dataset.

4.6. Comparison with SOTA Transformer Models

5. Conclusions

Our comprehensive study has elucidated the multifaceted impact of attention mechanisms on mammogram classification for multiple resolutions, abnormality types and datasets. The primary finding of this work is that attention generally improves baseline model performance across numerous

training/testing conditions. Notably, SE and ECA emerge as superior attention methods, significantly outperforming CBAM. Furthermore, although ECA outperforms SE by a small margin for CBIS-DDSM, SE greatly outperforms SE for INbreast. Thus, although ECA is considered to be the most "state-of-the-art" of the attention models we tested (given its high performance on ImageNet), we find that these relative rankings don't necessarily transfer to the task of mammogram classification.

Between the baseline models, we found that Regnetx064 had the best performance, then Densenet169, then Resnet50. All baseline models showed variable increases due to attention depending on the dataset's statistical characteristics. Densenet169 shows the highest increases due to attention for CBIS-DDSM, but the lowest increases for INbreast because it didn't increase model complexity enough improve performance. On the other side, Regnetx064 and Resnet50, which are more complex, but are also more likely to overfit, show a greater increase in scores due to attention for the INbreast dataset. We also theorize that differences in model architecture, besides just model complexity, can provide reasonable explanations for the differences we observed in model performance. Specifically, we found that architectures with significant amounts of pooling are more likely to overfit than those with less pooling. This impact was more noticeable for calcifications than for masses, which implies that the small features found in mammograms are often lost during pooling.

In addition, we show evidence of a correlation between IOU and AUROC scores. This correlation improves due to attention, with ECA showing the most significant correlation among the three attention methods examined. Thus, our analysis provides evidence that attention improves model performance specifically by focusing on task-relevant regions. However, we also show that this relationship is weak, and was completely absent for the INbreast dataset. Hence, it is unclear if this commonly accepted theory of attention holds for mammogram classification. To investigate further, future studies could try to develop a loss function that uses segmentation loss to improve predictions for mammogram classification, since this would provide direct evidence for the ability of attention to improve scores by focusing on task-relevant regions.

Besides our investigation of the impacts of attention on mammogram classification, our study provides a number of advances related to general mammogram classification. Our result that complex models tend to overfit on datasets where the train and test sets have significant variability should

extend to models without attention as well. Also, we provide evidence to bolster the theory that whole image mammogram classifcation is highly dependent on resolution and abnormality type. Finally, we test the performance of a number novel baseline+attention method combinations. The best performing novel model, Regnetx064+ECA, achieves state of the art scores on the CBIS-DDSM dataset for whole-image classification [26].

Our study presents the first large-scale investigation of attention for whole-image mammogram classification. We hope that our results provide a road map to help guide model choice depending on the statistical characteristics of the data. The present study does contain several limitations. As already stated, given the large number of models, and finite time allowed for training, it's possible that we did not obtain the best possible results for each model. Also, we only used a small selection of attention methods and baseline models. A larger selection of both would produce results for a wider range of model complexities, allowing us to understand the issues discussed in this paper at a more statistically significant level. Furthermore, the incorporation of transformer modules into the baseline models, and a more rigorous theoretical examination of the differences in model architecture, could reveal more insightful trends. We did some preliminary experiments with transformer-based models, but we found that it was beyond the scope of the paper to rigorously analyze transformers in comparison to our baseline+attention model combinations. However, because transformers have been show to be competitive with CNNs in medical image diagnosis [60], and our quick experiments showed competitive performance, future studies should explore a more rigorous comparison of CNN-based attention and transformer-based attention.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used GPT-4 in order to summarize previous literature and improve manuscript clarity and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Guo, M.H., et al.: Attention Mechanisms in Computer Vision: A Survey. Comp. Visual Media **8**(5), 331–368 (2022)

[2] Datta, S.K., Shaikh, M.A., Srihari, S.N., and Gao, M.: Soft Attention Improves Skin Cancer Classification Performance. In: Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. Lecture Notes in Computer Science, Springer Cham., **12929** (2021)

[3] Zhang, K., Wang, W., Lv, Z., Fan, Y., and Song, Y.: Computer vision detection of foreign objects in coal processing using attention CNN. Engineering Applications of Artificial Intelligence **102**, 104242 (2021)

[4] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W.: CCNet: Criss-Cross Attention for Semantic Segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 603-612, Seoul, Korea (South) (2019)

[5] Anaya-Isaza, A., Mera-Jiménez, L., Zequera-Diaz, M.: An overview of deep learning in medical imaging. Informatics in Medicine Unlocked **26**, 100723 (2021)

[6] Gonçalves, T., Rio-Torto, I., Teixeira, L.F., Cardoso, J.S.: A survey on attention mechanisms for medical applications: are we moving towards better algorithms?. IEEE Access, (2022)

[7] Mao, N., Zhang, H., Dai, Y., Li, Q., Lin, F., Gao, J., Zheng, T., Zhao, F., Xie, H., Xu, C. and Ma, H. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. British Journal of Cancer, **128**(5), 793-804 (2023)

[8] Roy, A.G., Navab, N. and Wachinger, C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE transactions on medical imaging, **38**(2), 540-549 (2018)

[9] Hassan, N.M., Hamad, S. Mahar, K.: Mammogram breast cancer CAD systems for mass detection and classification: a review. Multimed Tools Appl **81**, 20043–20075 (2022)

[10] Altan, G.: Deep Learning-based Mammogram Classification for Breast Cancer. International Journal of Intelligent Systems and Applications in Engineering, **8**(4), 171–176 (2020)

[11] Shen, L., Margolies, L.R., Rothstein, J.H., et al.: Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep, **9**, 12495 (2019)

[12] Adedigba, A.P., Adeshina, S.A., Aibinu, A.M.: Performance evaluation of deep learning models on mammogram classification using small dataset. Bioengineering, **9**(4), 161 (2022)

[13] Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Attention Deeplabv3+: Multi-level Context Attention Mechanism for Skin Lesion Segmentation. In: Bartoli, A., Fusiello, A. (eds) Computer Vision – ECCV 2020 Workshops. ECCV 2020. Lecture Notes in Computer Science, Springer, Cham., **12535** (2020). https://doi.org/10.1007/978-3-030-66415-2_16

[14] Sinha, A. and Dolz, J.: Multi-scale self-guided attention for medical image segmentation. IEEE journal of biomedical and health informatics, **25**(1), pp.121-130 (2020)

[15] Tang, H., Yuan, C., Li, Z. and Tang, J.: Learning attention-guided pyramidal features for few-shot fine-grained recognition. Pattern Recognition, **130**, 108792 (2022)

[16] Lou, Q., Li, Y., Qian, Y., Lu, F., Ma, J.: Mammogram classification based on a novel convolutional neural network with efficient channel attention. Computers in Biology and Medicine, **150**, 106082 (2022)

[17] Xu, C., Lou, M., Qi, Y., Wang, Y., Pi, J., Ma, Y.: Multi-Scale Attention-Guided Network for mammograms classification, Biomedical Signal Processing and Control, **68**, 102730 (2021)

[18] Niu, J., Li, H., Zhang, C. and Li, D.: Multi-scale attention-based convolutional neural network for classification of breast masses in mammograms. Medical physics, **48**(7), pp.3878-3892 (2021)

[19] Berghouse, M., Bebis, G. and Tavakkoli, A.: Investigating the Impact of Attention on Mammogram Classification. In International Symposium on Visual Computing. Cham: Springer Nature Switzerland, 30-43 (2023)

[20] Mao, N., Zhang, H., Dai, Y., Li, Q., Lin, F., Gao, J., Zheng, T., Zhao, F., Xie, H., Xu, C. and Ma, H. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. British Journal of Cancer, **128**(5), 793-804 (2023)

[21] Shen, L.: End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv preprint. arXiv:1711.05775, (2017)

[22] Zhao, X., Yu, L. and Wang, X. Cross-view attention network for breast cancer screening from multi-view mammograms. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 1050-1054 (2020)

[23] Sun, H., Li, C., Liu, B., Liu, Z., Wang, M., Zheng, H., Feng, D.D. and Wang, S. AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. Physics in Medicine Biology, **65**(5), 055005 (2020)

[24] Zhao, W., Wang, R., Qi, Y., Lou, M., Wang, Y., Yang, Y., Deng, X. and Ma, Y. BASCNet: Bilateral adaptive spatial and channel attention network for breast density classification in the mammogram. Biomedical Signal Processing and Control, **70**, 103073 (2021)

[25] Xu, C., Qi, Y., Wang, Y., Lou, M., Pi, J. and Ma, Y. ARF-Net: An Adaptive Receptive Field Network for breast mass segmentation in whole mammograms and ultrasound images. Biomedical Signal Processing and Control, **71**, 103178 (2022)

[26] Wei, T., Aviles-Rivero, A.I., Wang, S., Huang, Y., Gilbert, F.J., Schönlieb, C.B. and Chen, C.W.: Beyond fine-tuning: Classifying high resolution mammograms using function-preserving transformations. Medical Image Analysis, **82**, 102618 (2022)

[27] Baccouche, A., Garcia-Zapirain, B., Olea, C.C. and Elmaghraby, A.S.: Breast Lesions Detection and Classification via YOLO-Based Fusion Models. Computers, Materials & Continua, **69**(1) (2021)

[28] Zagoruyko, S. and Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint. arXiv:1612.03928, (2016)

[29] Vuckovic, J., Baratin, A. and Combes, R.T.D.: A mathematical theory of attention. arXiv preprint. arXiv:2007.02876, (2020)

[30] Wiegreffe, S. and Pinter, Y.: Attention is not not explanation. arXiv preprint. arXiv:1908.04626, (2019)

[31] Lee, R., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data **4**, 170177 (2017)

[32] Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J. and Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic radiology, **19**(2), pp.236-248 (2012)

[33] Saunders, R.S. Jr, Baker J.A., Delong D.M., Johnson J.P., and Samei E.: Does image quality matter? Impact of resolution and noise on mammographic task performance. Med Phys. **34**(10), 3971-81 (2007)

[34] Abdel-Nasser, M., Melendez, J., Moreno, A., and Puig, D.: The Impact of Pixel Resolution, Integration Scale, Preprocessing, and Feature Normalization on Texture Analysis for Mass Classification in Mammograms. International Journal of Optics **2016**, 1370259 (2016)

[35] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016)

[36] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.: Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708 (2017)

[37] Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., and Dollár, P.: Designing Network Design Spaces. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10425-10433, Seattle, WA, USA (2020)

[38] Chen, Y., Zhang, Q., Wu, Y., Liu, B., Wang, M., and Lin, Y.: Fine-Tuning ResNet for Breast Cancer Classification from Mammography. In: Wu, C., Chyu, MC., Lloret, J., Li, X. (eds) Proceedings of the 2nd International Conference on Healthcare Science and Engineering . ICHSE 2018. Lecture Notes in Electrical Engineering, **536**. Springer, Singapore (2019)

[39] Wightman, R. https://github.com/huggingface/pytorch-image-models.

[40] Hu, J., Shen, L., and Sun, G.: Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132-7141, Salt Lake City, UT, USA (2018)

[41] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q.: ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531-11539, Seattle, WA, USA (2020)

[42] Woo, S., Park, J., Lee, J.Y., and Kweon, I.S. CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, **11211**, Springer, Cham. (2018). https://doi.org/10.1007/978-3-030-01234-2_1

[43] Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H. and Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in neural information processing systems, 34, 9355-9366 (2021)

[44] Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J. and Yuan, L.: Davit: Dual attention vision transformers. In European conference on computer vision, 74-92 (2022)

[45] Falconi, L.G., Perez, M., Aguilar, W.G. and Conci, A.: Transfer learning and fine tuning in breast mammogram abnormalities classification on

CBIS-DDSM database. Adv. Sci. Technol. Eng. Syst. J, 5(2), 154-165 (2020)

[46] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248-255 (2009)

[47] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2623-2631 (2019)

[48] Bergstra, J., Bardenet, R., Bengio, Y. and Kégl, B.: Algorithms for hyper-parameter optimization. Advances in neural information processing systems, **24** (2011)

[49] Kebria, P.M., et al. Deep Imitation Learning: The Impact of Depth on Policy Performance. In: Cheng, L., Leung, A., Ozawa, S. (eds) Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science(), **11301**. Springer, Cham. (2018). https://doi.org/10.1007/978-3-030-04167-0_16

[50] Rice, L., Wong, E., and Kolter, Z.: Overfitting in adversarially robust deep learning. In International Conference on Machine Learning, 8093-8104, PMLR (2020)

[51] Sun, S., Chen, W., Wang, L., Liu, X. and Liu, T.Y.: On the depth of deep neural networks: A theoretical view. In Proceedings of the AAAI Conference on Artificial Intelligence **30**(1) (2016)

[52] Li, Z., Gu, T., Li, B., Xu, W., He, X., and Hui, X.: ConvNeXt-Based Fine-Grained Image Classification and Bilinear Attention Mechanism Model. Applied Sciences, **12**(18), 9016 (2022)

[53] Doimo, D., Glielmo, A., Goldt, S., and Laio, A.: Redundant representations help generalization in wide neural networks. Advances in Neural Information Processing Systems, **35**, 19659-19672 (2020)

[54] Casper, S., Boix, X., D'Amario, V., Guo, L., Schrimpf, M., Vinken, K., and Kreiman, G.: Frivolous units: Wider networks are not really that

wide. In Proceedings of the AAAI Conference on Artificial Intelligence **35**(8), 6921-6929 (2021)

[55] Yu, S., Jia, S., and Xu, C.: Convolutional neural networks for hyperspectral image classification. Neurocomputing, **219**, 88-98 (2017)

[56] Dong, N., Yan, S., Tang, H., Tang, J. and Zhang, L.: Multi-view information integration and propagation for occluded person re-identification. Information Fusion, **104**, 102201 (2024)

[57] Dong, N., Zhang, L., Yan, S., Tang, H. and Tang, J.: Erasing, transforming, and noising defense network for occluded person re-identification. In IEEE Transactions on Circuits and Systems for Video Technology (2023)

[58] Zha, Z., Tang, H., Sun, Y. and Tang, J.: Boosting few-shot fine-grained recognition with background suppression and foreground alignment. In IEEE Transactions on Circuits and Systems for Video Technology (2023)

[59] Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., and Wei, Y.: LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. In: IEEE Transactions on Image Processing, **30**, 5875-5888 (2021)

[60] Matsoukas, C., Haslum, J.F., Söderberg, M. and Smith, K.: Is it time to replace cnns with transformers for medical images? arXiv:2108.09038 (2021)

Appendix A. Supplementary Figures

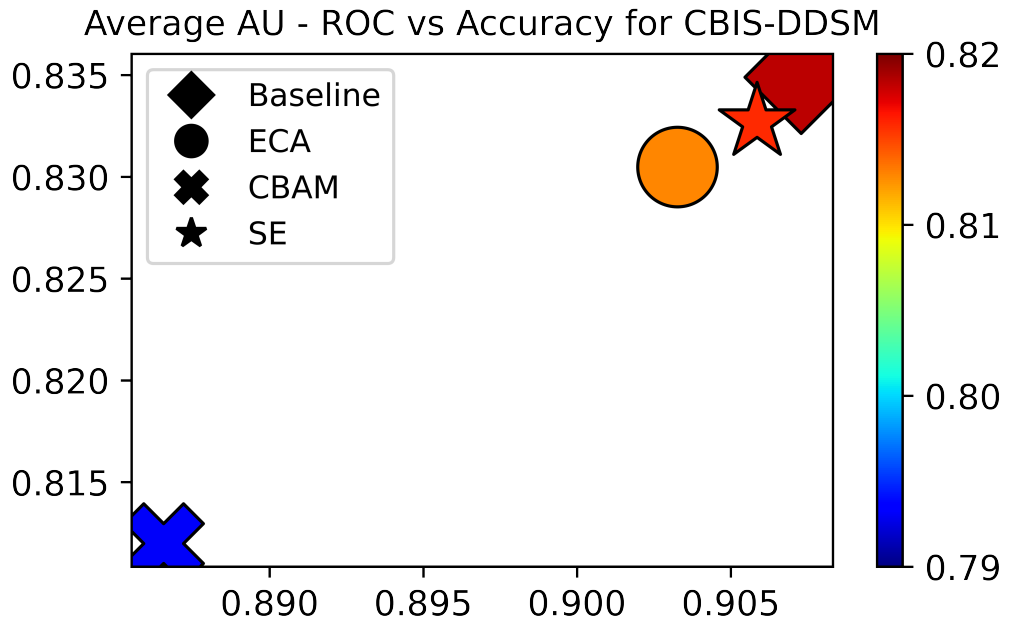


Figure A.13: Average AU-ROC vs accuracy with F1 colorbar for the training set for CBIS-DDSM. This figure generally shows much higher scores on the training set, indicating a high amount of overfitting for all models.

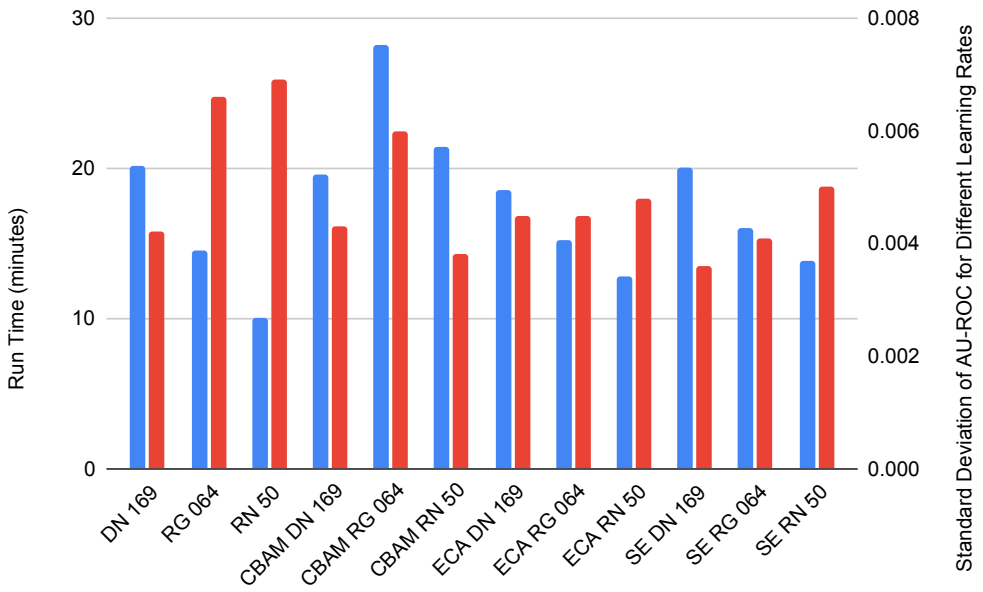


Figure A.14: Average run time (blue) and standard deviation of AU-ROC for different learning rates (red) on the CBIS-DDSM dataset. Standard deviations were calculated from 5 learning rate bins spanning the range of learning rates used for each model ($min - LR : 3 \times 10^{-5}, 3 \times 10^{-5} : 5 \times 10^{-5}, 5 \times 10^{-5} : 7 \times 10^{-5}, 7 \times 10^{-5} : 9 \times 10^{-5}, 9 \times 10^{-5} : max - LR$). Model names are abbreviated to fit in the figure. "DN" stands for Densenet, "RG" stands for Regnetx, and "RN" stands for Resnet.