Multi-View Attention Based Methods for Breast Cancer Detection

Yaswanth Badugu
50418798
ybadugu@buffalo.edu

Mahesh Bhosale
50418912
mbhosale@buffalo.edu

Yu Lin Chen
50168689
yulinchi@buffalo.edu

Chia Chen Chen
50385918
cchen248@buffalo.edu

GITHUB(please check readme for execution instructions)
https://github.com/bhosalems/Breast_cancer_detection

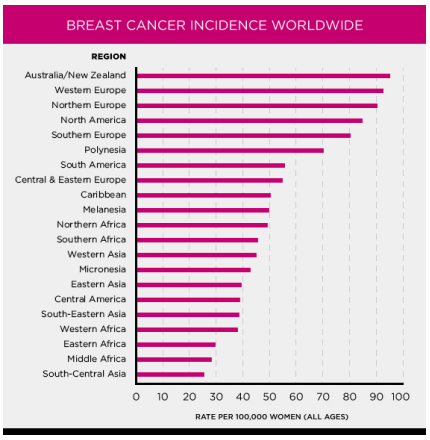
Abstract

Cancer detection is an incredibly interesting problem because of its academic difficulty and importance. The problem is quite simple if it is solved by traditional methods of classification, but it's useful-ness in the real world is questionable as it misses to capture the information from the different views for the single case. Multi stream model with each stream operating on single view of the model is therefore desired for the muti-task learning. We add attention methods in multi-stream view wise model to further promote the collaborative learning.

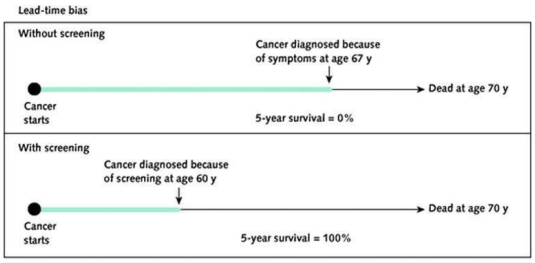
1 Introduction

Breast cancer is the second most common cancer occurring among women in the United States and internationally.[3] There were over 2 million new cases recorded in 2018. The most common and least invasive method of breast cancer diagnosis is mammography, which, according to the American Cancer Society reduces the rate of cancer death through early detection by 20-40%. Therefore, accurate and early breast cancer detection is of vital importance. According to different protocols of each country, multiple Radiologists take a look at the mammograms; often, a single mammogram doesn't reveal much information, and differences in multiple consecutive mammograms are found useful by Radiologists to make an accurate diagnosis. Computer-Aided systems (CAD), which include Computer-Aided Diagnosis(CADx) and Computer-Aided Detection(CADe) and assistance has been introduced decades ago and has been approved for medical use, but has failed to improve the performance of readers in real-world settings. Radiologists mainly look for four abnormalities in breast tissues- microcalcifications, masses, architectural distortions, and asymmetries. Many attempts have been made to only look for Regions of Interest (ROI) in an image to find out these abnormalities, rather than looking at a complete mammogram in an attempt to increase the accuracy. Oftentimes, these approaches apply techniques such as machine learning, deep learning, visual composition, boundary simplicity, Gabor filters, segmentation, sparse coding, and saliency-based. Following the recent success of deep learning methods such as convolutional neural networks in computer vision-related applications at a large scale, many attempts have been made to detect abnormalities in breast tissues from complete mammograms.

Cancer Detection is an interesting and challenging problem because of its complexity and the real-world consequences of automated decision-making on the prognosis. There have been multiple(and sometimes successful yet complex) attempts at using Deep Learning algorithms and integrating them with radiology imaging to create vastly useful outputs in Mammography analysis.



(a) Breast Cancer Incidence per 100,000 women, Worldwide[1]



(b) Survival Rates of women w.r.t detection time[2]

2 Datasets

We are dealing with 2 datasets in this project primarily, INBreast[4] and DDSM[5],[7]. Both of these datasets have similar mechanisms defining theie structures. They have two kinds of views, CC(Craniocaudal) and MLO(Mediolateral Oblique) which are top to bottom view and side view respectively. Therefore, the two views combined, give a full picture of a breast's mammogram. One way that both of these datasets come to evaluate the criticality(or the malignancy) of the calcification in breast tissue is through codifying a numerical terminology with the recommendation of American College of Radiology(ACR) termed as Breast Imaging Reporting and Data System(BI-RADS) scale. Based on the level of suspicion, the lesions are placed in one of six BI-RADS categories:

Category	Description
0	Needs additional imaging evaluation and/or prior mammograms for comparison
1	Negative
2	Benign finding(s)
3	Probably benign finding(s). Short-interval follow-up is suggested.
4	Suspicious anomaly. Biopsy should be considered.
5	Highly suggestive of malignancy. Appropriate action should be taken.
6	Biopsy proven malignancy

2.1 INBreast Dataset

This dataset is farmed through Breast Centre in CHSJ, Porto, Portugal, under the guidance of Hospital's Ethics Committee and the National Committee of Data Protection. Every Image is of 3328 x 4084 or 2569 x 3328-pixel resolution depending on the compression plate used during the acquisition of the mammograms(this is also reliant on the breast size of the patient). All of these images are stored in '.DICOM' (Digital Imaging and Communications in Medicine) format. This Dataset represents 115 cases of diagnosis, from which 90 have both the image views i.e., MLO and CC of each breast, and the remaining 25 instances comprise women who had a mastectomy. Two views of the singular breast were included[4].

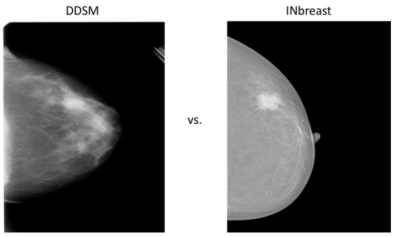


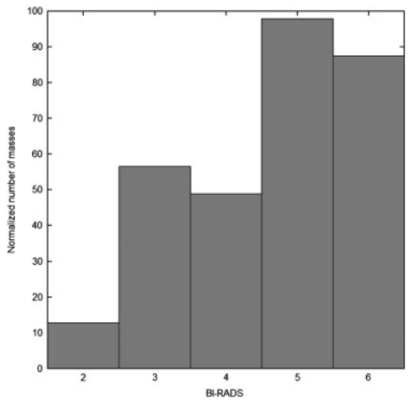
Figure 2: difference between the images present in DDSM and INBreast

2.1.1 DDSM Dataset

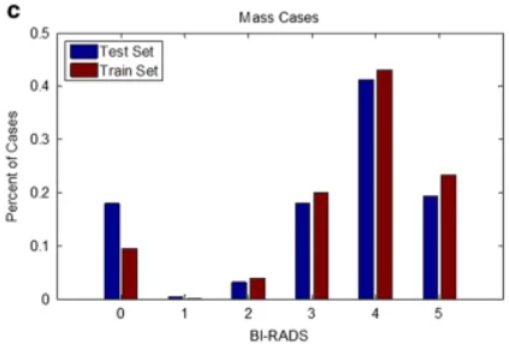
CBIS-DDSM(Curated Breast Imaging Subset of DDSM, Digital Database for Screening Mammography) is a database of 2,620 scanned films of mammographs[5]. It contains mammographs

whose abnormalities fall under three categories: Normal, Benign and Malignant with a verified pathological information. The images are in DICOM format just like the aforementioned dataset and the image data is structured in such a way that each participant might have multiple patient IDs, which help in intuitively understanding a timeline of a malignancy's progression in a patient. This makes it appear as though there are 6,671 participants when we verify the .DICOM metadata, but in reality, there are only 1,566 actual participants in the cohort.

The main difference between DDSM and INBreast is that the former is a screen-film mammography dataset while the latter is a digital mammography dataset. In addition, the difference in the optic density due to different screening metrics will result in requiring different methods to process and impart ROIs in the mammographs[4].



(a) mass distribution w.r.t BI-RADS values in INBreast



(b) mass distribution w.r.t BI-RADS values in DDSM

3 Preprocessing

In this section we will shortly describe the preprocessing of image dataset we did. As a first step we needed to convert the files into png files which are well suited with deep learning. Then we create the pickle of the image dataset with all metadata such as view, image file name, unique ID etc. Because we wanted to associate all the views of a single case together and work on this as a batch in the training this association was necessary, as shown in Fig. 4.

```
{'horizontal_flip': 'NO', 'R-CC': ['20587994_024ee3569b2605dc_MG_R_CC_ANON
'], 'L-CC': ['20588020_024ee3569b2605dc_MG_L_CC_ANON'], 'R-MLO': ['2058804
6_024ee3569b2605dc_MG_R_ML_ANON'], 'L-MLO': ['20588072_024ee3569b2605dc_MG
_L_ML_ANON']}
```

Figure 4: association between four views of a single case

We crop the images by detecting the largest connected component in an image, which only contain the breast and removes the outer parts of the images. We save the windows sizes in pixel of the cropped image with respect to original image size which is helpful to find the correct segmentation maps if required later, as visualized in Fig. 5.[13] We have annotations for each of the image as well, we are interested in BI-RADS value. BIRADS<=3 is considered as benign whereas BIRADS>=4 is considered malignant[4]. We employ similar threshold to convert the annotations to the 0: Benign, 1: Malignant and treat it as a binary classification problem[13].

4 Proposed Methodology

Here we describe the methods used to compare and validate the results. We consider this as classification problem (in some cases Binary in Multiview it's Multiclass) and use cross entropy loss as objective function. We mostly used ADAM optimizer and tuned different hyperparameters for each of the methods separately.

```
{'horizontal_flip': 'NO',
  'R-CC': ['20587994_024ee3569b2605dc_MG_R_CC_ANON'],
  'L-CC': ['20588020_024ee3569b2605dc_MG_L_CC_ANON'],
  'R-MLO': ['20588046_024ee3569b2605dc_MG_R_ML_ANON'],
  'L-MLO': ['20588072_024ee3569b2605dc_MG_L_ML_ANON'],
  'window_location': {'L-CC': [(31, 3280, 5, 1494)],
   'R-CC': [(40, 3168, 1087, 2560)],
   'L-MLO': [(0, 3176, 0, 1608)],
   'R-MLO': [(0, 3184, 851, 2560)]},
  'rightmost_points': {'L-CC': [((1528, 1535), 1438)],
   'R-CC': [((1532, 1552), 1423)],
   'L-MLO': [((1671, 1741), 1557)],
   'R-MLO': [((1648, 1712), 1659)]},
  'bottommost_points': {'L-CC': [(3198, (100, 101))],
   'R-CC': [(3077, (101, 102))],
   'L-MLO': [(3125, (241, 263))],
   'R-MLO': [(3133, (357, 406))]},
  'distance_from_starting_side': {'L-CC': [5],
   'R-CC': [0],
   'L-MLO': [0],
   'R-MLO': [0]}
}
```


Figure 5: metadata for the segmentation

- Simple and Convolutional Neural Network: Single 2D layer, two 2D layers and five 2D layers. Each layer contains a convolutional2d layer, a max-pooling layer and dropout layer. There are five such layers in the Neural Network.
- Transfer learning: Pretrained models are usually trained on large amounts of data and using resources that aren't usually available to everyone. For example, models trained on the ImageNet dataset, which has 1.4 million images with 1000 classes cannot be trained in-house. Therefore, it would be really beneficial for us to use these models as a backbone for feature extraction to get lower layer features and add fully connected layers (randomly initialized) on top of it. In our cases, we choose some classical pre-trained models as below and add two more layers with SoftMax non-linearity to make the final classifications. Below Figure shows the architecture of pre-trained models. Fig 6. describes the complete architecture of a Transfer Learning Model.

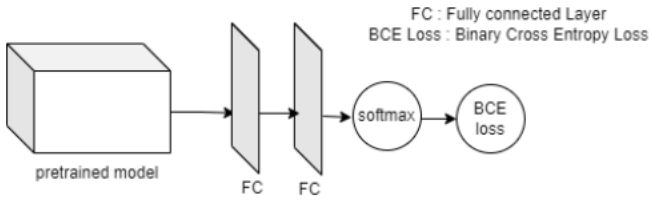


Figure 6: structure of a Transfer Learning model

1. InceptionV3: InceptionV3 has 189 layers with each factorization into smaller convolution layers, spatial factorization into asymmetric convolutions layers, auxiliary classifier and efficient grid size reduction. It achieved 77.9% accuracy on ImageNet.
2. ResNet50: Resnet50 has 50 layers with residual mechanism which contain 48 convolution layers along with max-pool and 1 average pool layer[8]. It can reach 74.9% accuracy on ImageNet.
3. MobileNet: It has 55 layers which contain depth wise separable convolution layers as opposed to standard convolution layers[9]. It achieved 70.9% accuracy on ImageNet.
4. VGG19/VGG16: models with 19 and 16 layers respectively, which contain convolutional layers and max-pool layers[10]. Both achieved similar accuracy of 70.1% on ImageNet.

- Attention based models
  1. Vision Transformer: Transformer architecture has become the standard for natural language processing tasks and Vision Transformation shows show that CNNs is not

necessary and transformer is applied directly to sequences of image patches can perform very good on image classification tasks[11].

2. External Attention Transformer: External attention is based on two external, small, learnable, shared memories, which can be implemented by using two linear layers and two normalization layers to replace self-attention in existing architectures. This model incorporates the multi-head mechanism into external attention to provide an all multi-layer perceptron architecture, External Attention Multilayer Perceptron (EAMLP) for image classification[12].

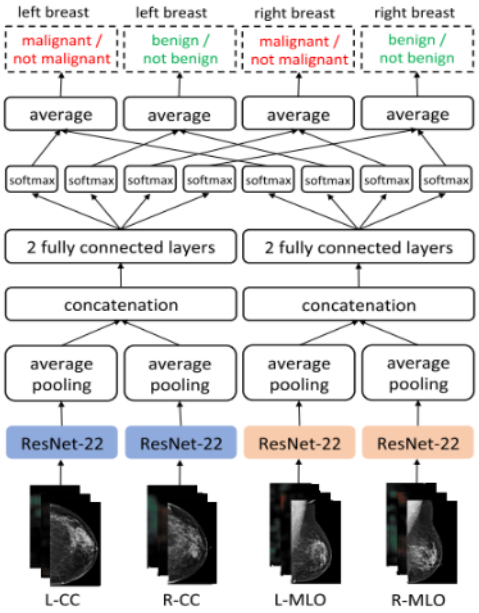


Figure 7: architecture of a Multi Stream view-wise model

- Multi Stream view-wise Model: This model is motivated by the idea of Meta-learning where multiple tasks are learned in parallel and collaboratively. Each task provides a cue to the learning of another task. Although we are interested in finding if the case is malignant, frequently both the finding are present at the same time. Authors propose solving this task with multiclass classification where they consider both the malignant and benign findings of both the breasts. The model is composed of four stream networks where each stream works on extracting the features of that view on both the breasts. For a given case each image is fed in parallel to respective streams where RESNET22 extracts the features. Average-pooled features are then concatenated and are fed to two fully connected layers. Four Non-linearities are applied at the end for each view denoting the probability of the breast being malignant or benign. Fig. 7 describes the architecture of this model. For left and right breasts, in each case total of four binary labels, and the model tries to predict these labels, $\hat{y}_{R,m}$, $\hat{y}_{L,m}$, $\hat{y}_{R,b}$, and $\hat{y}_{L,b}$ where R and L denote Right breast or Left breast respectively, and m and b denote malignant and benign.

$$\hat{y}_{R,m}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}, \mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}) = \frac{1}{2}\hat{y}^{\text{CC}}_{\text{R,m}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}) + \frac{1}{2}\hat{y}^{\text{MLO}}_{\text{R,m}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}),$$

$$\hat{y}_{R,b}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}, \mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}) = \frac{1}{2}\hat{y}^{\text{CC}}_{\text{R,b}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}) + \frac{1}{2}\hat{y}^{\text{MLO}}_{\text{R,b}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}),$$

$$\hat{y}_{L,m}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}, \mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}) = \frac{1}{2}\hat{y}^{\text{CC}}_{\text{L,m}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}) + \frac{1}{2}\hat{y}^{\text{MLO}}_{\text{L,m}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}),$$

$$\hat{y}_{L,b}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}, \mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}) = \frac{1}{2}\hat{y}^{\text{CC}}_{\text{L,b}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}) + \frac{1}{2}\hat{y}^{\text{MLO}}_{\text{L,b}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO})$$

while the training loss is computed as,

$$\mathcal{L}(y_{R,m}, y_{L,m}, y_{R,b}, y_{L,b}, \mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}, \mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}) = \ell(y_{R,m}, \hat{y}^{\text{CC}}_{\text{R,m}}((\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}))+$$
$$\ell(y_{R,m}, \hat{y}^{\text{MLO}}_{\text{R,m}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}))+$$
$$\ell(y_{R,b}, \hat{y}^{\text{CC}}_{\text{R,b}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}))+$$
$$\ell(y_{R,b}, \hat{y}^{\text{MLO}}_{\text{R,b}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}))+$$
$$\ell(y_{L,m}, \hat{y}^{\text{CC}}_{\text{L,m}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}))+$$
$$\ell(y_{L,m}, \hat{y}^{\text{MLO}}_{\text{L,m}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}))+$$
$$\ell(y_{L,b}, \hat{y}^{\text{CC}}_{\text{L,b}}(\mathbf{x}_{R-CC}, \mathbf{x}_{L-CC}))+$$
$$\ell(y + L, b, \hat{y}^{\text{MLO}}_{\text{L,b}}(\mathbf{x}_{R-MLO}, \mathbf{x}_{L-MLO}))$$

- Attention in Multi-stream view wise model: Since calcification and other abnormalities can be found from the local features, using attention to attend to these abnormalities is needed and it also promotes collaboration in Multi-Task Learning. We therefore add two attention blocks in the RESNET-22 model – one is spatial attention[15] and other is channel attention[15].
  1. Channel Attention
     - First feature map F is average pooled globally to get the channel vector of size Cx1x1.
     - Then is it sent to FC layer to have better interaction between different channels.
     - Batch Normalization is added at the end of it.

     Output of channel attention is given by, M_c(**F**)=BN(MLP(AvgPool(**F**)))
  2. Spatial Attention :
     - Input feature map F is passed through the chains of 1x1 convolution and 3x3 convolution with dilation of 4 is added which helps us to increase the area of the receptive field.
     - Finally its common to reduce the size of the feature map further by adding another convolution layer and then applying a batch-norm on it.

     Output of the spatial attention is given by,

$$M_s(\mathbf{F}) = BN(f^{1x1}_3(f^{3x3}_2(f^{3x3}_1(f^{1x1}_0(\mathbf{F})))))$$

Both the attentions are added together in block attention module as shown in Fig. 8, Output

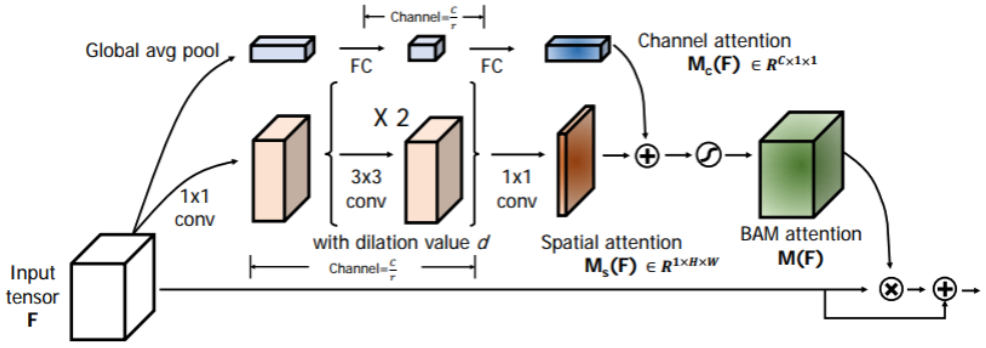


Figure 8: block Attention model

of the block attention module is given by, where **F'=F+F⊗M(F)**, where ⊗ is element wise multiplication. Original feature map is added in the output for the continuous gradient flow. Here, σ is a sigmoid function. Both outputs of and channel branch are resized to C×H×W before addition.

We add the variant of spatial and channel attentions from block attention module in RESNET-22 which gives us better accuracy as expected.

5 Challenges Faced

In this section, we discuss the challenges we faced while training. Peculiarly enough we found that the validation accuracies would remain the same over the training iterations despite the loss decrease. It appears that this is a common issue in the Medical Image analysis domain [16] and could be caused because of the fact that the variance of the train and validation dataset is different. Some of the solutions include – 1) adding regularization 2) adding dropout 3) adding data augmentations 4) adding more data to our dataset 5) Preprocessing the dataset. We did several trials on the first three techniques but it did not help. In Data augmentation we tried rotating the images by 40 degrees, height and width shifting 20% of images, and horizontal flipping. We resorted to the conclusion that increasing the dataset should be the same step, that's when we e decided to add the DDSM dataset.

While training the pretrained models one of the other issues that we faced was resource exhausted error on google colab pro+, due to RAM was full. We think that this is because of the fact that the image size in INBreast dataset is much bigger as compared to other datasets, we had to get the results by running it on the local GPUs with resized images and very low batch size. While training on the DDSM dataset, we also faced the similar issue of constant validation accuracy, however, after adding regularization term, a dropout layer, batch normalization and hyperparameter tuning, this issue was resolved unlike in case of INBreast dataset.

6 Results

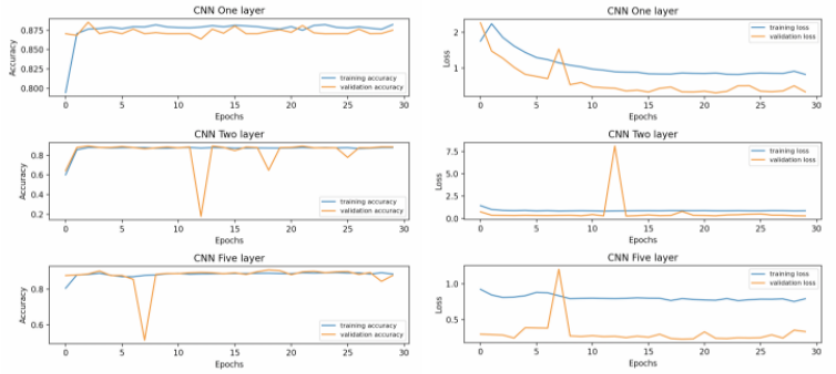


Figure 9: accuracies of CNN layers 1,2,5(left),loss of CNN layers 1,2,5(right)

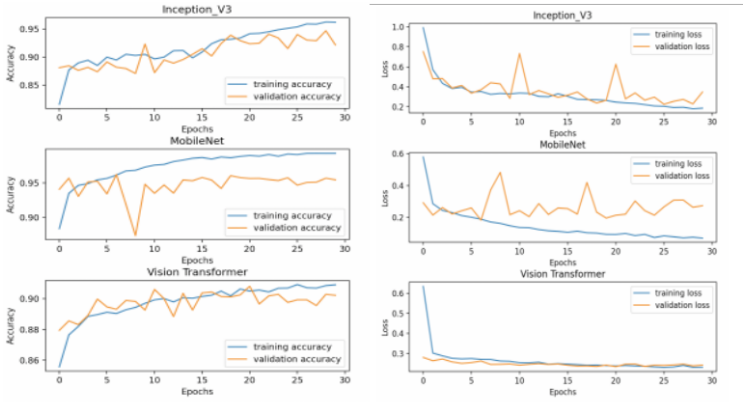


Figure 10: for IncerptionV3, MobileNet, Vision Transformer- accuracies(left),loss(right)

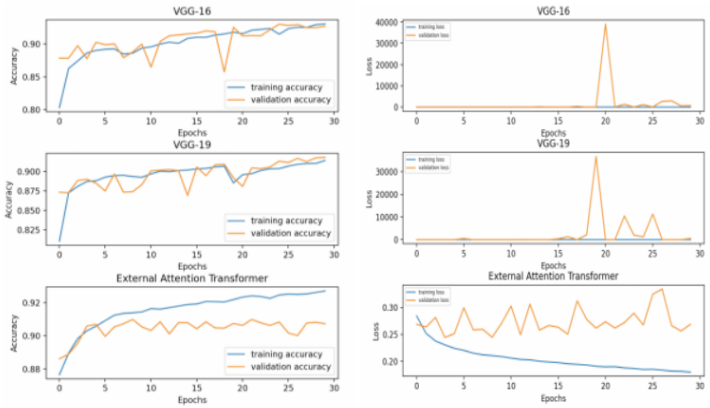


Figure 11: for VGG16, VGG19, External Attention Transformer- accuracies(left),loss(right)

Here we describe the results of our experimentation on both the datasets. All the graphs in the figures 9-11 are captured on the DDSM dataset. From the loss and accuracy curves it is clear that pretrained MobileNet performs the best in non-attention based methods. Although the difference is marginal, the possible explanation is that the number of parameters in the MobileNet is lower as compared to other models – having larger model might result in the overfitting if not taken care of. As can be seen from the graphs, even simple CNN model is able to achieve the good accuracy. Appendix,Table I annotates the performance of all the models. In the attention based models both the transformers perform good as expected. In multi-view model after adding the attention methods we are able to see the improvement as expected quantitatively. Comparing the masks of abnormalities from annotations with the attention map should be done as a qualitative analysis.

7 Future Work

Since this project only focus on image classification about mammography, object detection may be a good way to extend the benefits for healthcare professionals in the future. Such as, importing head and backbone architecture of YOLO family[17] and customized feature extraction will help us determine precision and location in the same time. Furthermore, bidirectional encoder representation from Image transformers may also be a good direction that we could take the implementation towards. Qualitative analysis of attention based methods could give us hints to improve the model further. Reinforcement based methods could also be employed for region proposal [14]. Although some of the state of the art methods traditional methods perform better, they could be deemed unuseful in real world because they do not take into account the images from multiple view for a single exam. Therefore, more authoritative validation from the domain experts is needed for traditional methods of classification.

8 Conclusion

We have compared the performance of traditional image classification methods for breast classification. We identified that using a model which incorporates images from multiple views is more useful in the real world. Therefore, models such as Multi stream view wise model make sense. Adding attention mechanism in multi stream view wise model attends on collaborative information available from multiple views and helps the multitask learning and further improves the performance as expected.

9 References

[1] Susan G Komen Foundation `https://www.komen.org/breast-cancer/facts-statistics/breast-cancer-statistics/`

[2] National Cancer Institute NIH https://www.cancer.gov/about-cancer/screening/research/what-screening-statistics-mean

[3] Li Shen (2017) End-to-end Training for Whole Image Breast Cancer Diagnosis using An All Convolutional Design

[4] Inês C Moreira et.al (2011). INbreast: toward a full-field digital mammographic database.

[5] Rebecca L et al. (2017) A curated mammography data set for use in computer-aided detection and diagnosis research

[6] Christian Szegedy et al. (2015), Rethinking the Inception Architecture for Computer Vision.

[7] Rose, C., Turi, D., Williams, A., Wolstencroft, K., Taylor, C. (2006). Web Services for the DDSM and Digital Mammography Research. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (eds) Digital Mammography. IWDM 2006. Lecture Notes in Computer Science, vol 4046. Springer, Berlin, Heidelberg.https://doi.org/10.1007/11783237_51

[8] Kaiming He et al.(2015), Deep Residual Learning for Image Recognition.

[9] Andrew G. Howard et al. (2017), MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[10] Karen Simonyan et al. (2015), VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

[11] Alexey Dosovitskiy et al.(2020), An image is worth 16X16 words: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

[12] Meng-Hao Guo et. al(2021), Beyond Self-attention: External Attention using Two Linear Layers for Visual Tasks

[13] Nan Wu et al. IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 39, NO. 4, APRIL 2020, Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening.

[14] Aleksis Pirinen et.al, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Deep Reinforcement Learning of Region Proposal Networks for Object Detection.

[15] Jongchan Park et al. British Machine Vision Conference 2018, Bottleneck Attention Module.

[16]https://stackoverflow.com/questions/52356068/validation-accuracy-constant-in-keras-cnn-for-

[17] Yali Nie et. al, Automatic Detection of Melanoma with Yolo Deep Convolutional Neural Networks

10 Appendix

Architecture	Dataset	Train Accuracy	Val Accuracy	Test Accuracy
Single Convolution Layer	INBreast	0.7293	0.8276	0.65853
Two Convolution layers	INBreast	0.739	0.8276	0.65853
Pretrained Mobilenet on Imagement+ two FCs	INBreast	0.9582	0.7927	0.73170
Preatrained RESNET50 on Imagement+ two FCs	INBreast	0.7944	0.7927	0.70731
Single Convolution layers	DDSM	0.882	0.8751	0.8687
Two Convolution layers	DDSM	0.8796	0.8857	0.8852
Five Convolution layers	DDSM	0.884	0.8764	0.8691
Preatrained Inception on Imagement+ two FCs	DDSM	0.9622	0.9219	0.9207
Pretrained VGG19 on Imagement+ two FCs	DDSM	0.9264	0.9224	0.9253
Pretrained RESNET50 on Imagement+ two FCs	DDSM	0.992	0.856	0.8433
Pretrained VGG16 on Imagement+ two FCs	DDSM	0.9349	0.9224	0.9243
Pretrained Mobilenet on Imagement+ two FCs	DDSM	0.9951	0.9569	0.9546
Vision transformer	DDSM	0.9091	0.9022	0.9014
External attention Transformer	DDSM	0.9271	0.9073	0.9102

Table 1: accuracies for different classifiers