Mammography Classification Using Deep Learning Methods

Tianyu Ao
CS Dept, UIUC
Urbana, IL 61801
tianyua2@illinois.edu

Baohe Zhang
ECE Dept, UIUC
Champaign, IL 61820
baohez2@illinois.edu

Abstract

Mammography is the process of using low-energy X-rays to examine the human breast for diagnosis and screening. However, the rates of false-negative if the mammograms are relatively high by its manual nature. In this paper, we train multiple Convolution Neural Network-based classifiers to classify pre-segmented breast masses from mammograms as benign or malignant. Our dataset is created by extracting breast mass images from public available mammography dataset CBIS-DDSM. Our best model achieves accuracy of 0.969, recall of 0.947, and precision of 0.985, successfully beating all known prior works and human experts. This result is achieved by a ResNet-18 network with transfer learning. We also conduct visualization with saliency maps to explore how our models classify breast masses. Our result demonstrates that Deep Learning is a powerful tool for breast cancer diagnosis, and it can bring new insights into further clinical trials.

1 Introduction

1.1 Breast Cancer

Breast cancer is the most frequent cancer among women. It affects 2.1 million women each year and also causes the greatest number of cancer-related deaths among women. In 2018, it was estimated that 627,000 women died from breast cancer, which is approximately 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed regions, rates are increasing in nearly every region globally[1]. In the U.S., one in eight women is expected to develop invasive breast cancer over the course of her lifetime[2].

The first noticeable symptom of breast cancer is typically a lump that feels different from the rest of the breast tissue. More than 80% of breast cancer cases are discovered when the woman feels a lump[3].

Another typical symptom complex of breast cancer is Paget's disease of the breast[4]. This syndrome presents as skin changes resembling eczemas, such as redness, discoloration, or mild flaking of the nipple skin. As Paget's disease of the breast advances, symptoms may include tingling, itching, increased sensitivity, burning, and pain. There may also be discharge from the nipple. Approximately half the women diagnosed with Paget's disease of the breast also have a lump in the breast[5]

In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment[1]

Preprint. Under review.

1.2 Mammography

Mammography (also called mastography) is the process of using low-energy X-rays (usually around 30 kVp) to examine the human breast for diagnosis and screening. The goal of mammography is the early detection of breast cancer, typically through the detection of characteristic masses or microcalcifications[1].

It has been shown to reduce breast cancer mortality by approximately 20% in high-resource settings. WHO Position paper on mammography screening concluded that in well-resourced settings, women aged 50-69 should undergo organized, population-based mammography screening if pre-specified conditions on program implementation are met. In limited-resource settings with weak health systems, mammography is not cost-effective, and early detection should focus on reducing stage at diagnosis through improved awareness[1]. For women aged 40-49 years or 70-75 years, WHO recommends systematic mammography screening in women aged 40-49 years or 70-75 years only in the context of rigorous research and in well-resourced settings[6].

The goal of any screening procedure is to examine a large population of patients and find the small number most likely to have a serious condition. These patients are then referred for further, usually more invasive, testing. Thus a screening exam is not intended to be definitive; rather, it is intended to have sufficient sensitivity to detect a useful proportion of cancers. The cost of higher sensitivity is a larger number of results that would be regarded as suspicious in patients without the disease. This is true of mammography. The patients without disease who are called back for further testing from a screening session (about 7%) are sometimes referred to as "false positives." There is a trade-off between the number of patients with the disease found and the much larger number of patients without disease that must be re-screened[6].

Mammograms also have a rate of missed tumors, or "false negatives." Accurate data regarding the number of false negatives are very difficult to obtain because mastectomies cannot be performed on every woman who has had a mammogram to determine the false-negative rate. Estimates of the false-negative rate depend on a close follow-up of a large number of patients for many years. This is difficult in practice because many women do not return for regular mammography, making it impossible to know if they ever developed cancer. In his book The Politics of Cancer, Dr. Samuel S. Epstein claims that in women ages 40 to 49, 1 in 4 of cancer is missed at each mammography. Researchers have found that breast tissue is denser among younger women, making it difficult to detect tumors. For this reason, false negatives are twice as likely to occur in pre-menopausal mammograms (Prate). This is why the screening program in the UK does not start calling women for screening mammograms until age 50[6].

However, mammography is still a manual process, quite prone to human error due to the variable shape and size of masses [2][7]. Our project, based on the previous researches, is to implement Convolutional Neural Network(CNN) methods to classify the breast digital mammograms as benign or malignant. By using different CNN methods and previous results to get more comprehensive comparisons on existed datasets and hope to explore new insights on further researches and clinical trials.

2 Related Work

Previous studies have hammered down the solid ground for our project scope.

In the paper by Jain and Levy, they implemented multiple Convolutional Neural Network (CNN) architectures to classify pre-segmented breast masses from mammograms as benign or malignant[2]. The dataset tested by their methods is the public Digital Database for Screening Mammography(DDSM). The best classification performance achieved on this dataset is an accuracy of 0.929, recall of 0.934, and precision of 0.924, successfully beating human performance. This result was achieved by modifying and fine-tuning the GoogleNet model from the ImageNet challenge[2]. They also visualized a number of masses and their interpretation by the models to glean better insights into how their models make their predictions. Lastly, they demonstrated the transfer learning from models pre-trained on the ImageNet data to a completely different domain such as mammogram images and yet achieve state-of-the-art results.

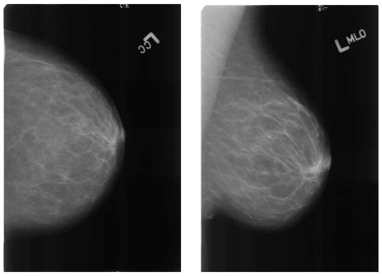
However, the size of pictures from DDSM is relatively small for training the model. In order to increase the amount of data, Scuccimarra proposed to extract the Regions of Interest (ROI) from each image, perform data augmentation, and then train ConvNets on the augmented data. The ConvNets were trained to predict whether a scan was normal or abnormal. The results based on DDSM have slightly improved by certain dataset groups. They use a system to output the probabilities rather than the predictions would allow such a system to provide additional information to radiologists rather than replacing them. In addition, the ability to adjust the decision threshold would allow radiologists to focus on more ambiguous scans while devoting less time to scans, which have very low probabilities[8].

In 2017, an end-to-end training algorithm for the detection and classification of breast cancer on digital mammograms was created by Li[9]. In the initial training stage, lesion annotations were used, but in subsequent stages, a whole image classifier was trained using only image-level labels, eliminating the reliance on rarely available lesion annotations. The simple all convolutional design provided superior performance in comparison with previous methods. The best single model achieved a per-image AUC of 0.88 on a holdout test set, and three-model averaging increased the AUC to 0.91. On an independent holdout set of images from the INbreast database, the best single model achieved a per-image AUC of 0.96. By analysis, their end-to-end approach may be particularly advantageous for training whole image classifiers for screening mammography exams that are often acquired using heterogeneous platforms and lack clinical ROI annotations in large datasets. Undoubtedly, their studies show the potential of deep learning algorithms to improve the accuracy of screening mammography[9].

Recently, Regab and Sharkas proposed a new computer-aided detection (CAD) system for classifying benign and malignant mass tumors in breast mammography images[8]. A well-known CNN architecture named AlexNet is used and is fine-tuned to classify two classes instead of 1,000 classes. The last fully connected (fc) layer is connected to the support vector machine (SVM) classifier to obtain better accuracy. The accuracy of the newly trained architecture is 71.01% . The highest area under the curve (AUC) achieved was 0.88 (88%) for the samples obtained from both segmentation techniques. Moreover, when using the samples obtained from the DDSM, accuracy is increased to 73.6%. Consequently, the SVM accuracy becomes 87.2% with an AUC equaling to 0.94 (94%). This is the highest AUC value compared to previous work using the same conditions[10].

3 Dataset

We use CBIS-DDSM[11] as the dataset. CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is an updated and standardized version of the Digital Database for Screening Mammography (DDSM) [12]. DDSM contains 2620 scanned film mammography studies. Each study consists of both mediolateral oblique (MLO) and craniocaudal (CC) views of each breast. DDSM includes normal, benign, and malignant cases with verified pathology information. CBIS-DDSM contains a subset of DDSM. Images of CBIS-DDSM are decompressed, and each image comes with the Regions of Interest (ROI) segmentation information.



(a) CC view (b) MLO view

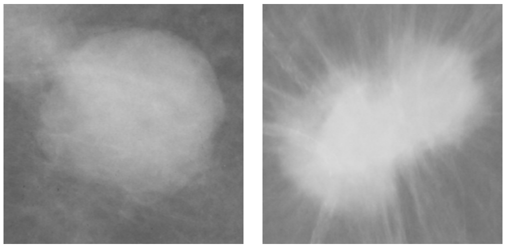
Figure 1: Mammography images

3.1 Data prepossessing

The purpose of this project is to identify whether a breast mass is benign or malignant, so breast mass images need to be extracted from the original raw mammography image. ROIs are extracted using a strategy similar to [2]. CBIS-DDSM contains mask information to obtain breast mass patches. The mask is a square box that defines the ROI. Some paddings are added to the mask to extract the context of the ROIs. 4725 breast mass images are extracted from raw images. Among these breast mass images, 2517 are benign, and 2208 are malignant. These breast mass images are randomly divided into the training set (80%) and the validation set (20%). The number of benign masses and malignant masses in the validation set are almost equal. Since the models we used in this project accept images with 3 channels, and the original image is grayscaled, the images will be converted to RGB by replicating pixel values across 3 channels. Images are also normalized at training time to achieve consistency for the dataset.

3.2 Data augmentation

In order to expand the size of the training set and avoid overfitting, we applied several techniques to augment data. Since the orientation of the mass images doesn't affect its pathology, we augment the image data through flip operation, rotate operation, and crop operation. For each breast mass image, we horizontal flip it. And for the original image and flipped image, we rotate them with a random angle. Finally, we randomly crop the image with a ratio (0.85 to 1.0) of the original size, and then resize the cropped image to 224*224 to match the model's input shape. Through data augmentation, our training set size is increased. The data augmentation also adds variety and randomness of the training set and helps avoid overfitting.



(a) Benign mass (b) Malignant mass

Figure 2: Breast masses

4 Methods

This section introduces the architectures and training strategies of the models we experimented with in this project. All models take images with shape 3*224*224 as inputs and use a softmax classifier with cross-entropy loss to predict the type of breast mass. Experiments are implemented with PyTorch on an NVIDIA Tesla K80 GPU hosted on Google Cloud Platform.

4.1 Baseline CNN

Neural Network is essentially a mathematical model to solve an optimization problem, and Convolutional Neural Network (CNN) is a powerful tool in image classification. The hidden layers of a CNN typically consist of convolution layers, pooling layers, fully connected layers, and normalization layers. We train a simple CNN model that contains 3 convolutional layers with detailed structure listed below as the baseline model to predict whether an input breast mass image is malignant or benign.

- Input layer
- Convolution (32 3 3 filters) - Batch Norm - ReLU - Max Pooling
- Convolution (32 3 3 filters) - Batch Norm - ReLU - Max Pooling
- Convolution (32 3 3 filters) - Batch Norm - ReLU - Max Pooling
- Fully-connected layer of dimension 512 - ReLU
- Fully-connected layer of dimension 128 - ReLU
- Fully-connected layer of dimension 2

4.2 AlexNet

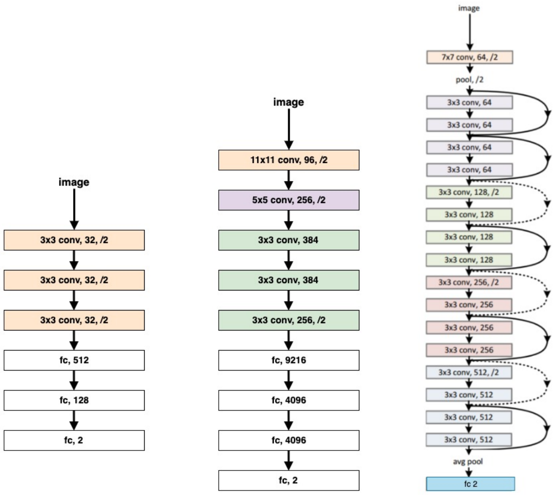
There are several popular deep learning models to do image classification, e.g., AlexNet[13], GoogleNet[14], VGG, ResNet[15], etc. We choose AlexNet as our first existing model to predict the type of breast mass due to its intuitive structure and reasonable computing resource requirement. AlexNet contains 5 convolutional layers and 3 fully connected layers. Relu is applied after every convolutional and fully connected layer. Dropout is applied before the first and the second fully connected layer. The last fully connected layer of AlexNet is modified to output 2 classes, benign or malignant, instead of 1000 classes. Figure 3 presents the architecture of our modified AlexNet. We train this AlexNet from scratch.

4.3 ResNet-18

Residual Neural Network(ResNet) is another famous deep learning model in image classification. It achieved a groundbreaking performance at ILSVRC'15. ResNet uses residual blocks to ensure that upstream gradients are propagated to lower network layers, aiding in optimization convergence in deep networks [15]. ResNet-18 is a variant of ResNet that contains 18 layers of convolutional neural network. We modified the last fully connected layer of ResNet-18 to make it available for breast mass diagnosis. We trained ResNet-18 from scratch, so this network can learn features only from breast mass images. We choose ResNet-18 for its lower complexity and shorter training time than other variants of ResNets. Figure 3 illustrates the architecture of our modified ResNet-18.

4.4 ResNet-18 with transfer learning

Transfer learning is an optimization technique that allows rapid progress or improved performance when modeling the second task [16]. The transfer learning technique is used frequently in the field of image classification. The model is pre-trained on a large corpus of images, and the model is required to predict a relatively large number of classes. So the pre-trained model can efficiently learn to extract features from different images. In our project, we use a ResNet-18 that is trained on more than a million images from the ImageNet dataset and can predict objects from 1000 categories with an extremely low error rate. Similar to previous models, we modify the last fully connected models to allow it to handle breast mass diagnosis job.



(a) Baseline CNN (b) AlexNet (c) ResNet-18

Figure 3: Architecture of Baseline CNN, AlexNet and ResNet-18, adapted from [15]

5 Results

5.1 Evaluation metric

We evaluated the performance of the models based on classification accuracy, recall, precision, and F1 scores. In breast mass image classification, we would weight more on recall. Because the false-negative cases (classifying the malignant breast mass as benign) would cause the patient untreated, while the false-positive cases (classifying the benign breast mass as malignant) only need an additional medical diagnosis to confirm the result.

Recall is the portion of actual positive cases that are identified correctly. It's defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Precision is the proportion of positive identifications that are actually correct. It's defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

F1 score is a weighted average of recall and precision, and is defined as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.2 Results analysis

After experimenting on different methods, we achieve the best validation accuracy of 0.969, along with the recall of 0.947, and precision of 0.985 when using ResNet-18 with transfer learning. Table 1 lists the highest accuracy, along with other evaluation metrics achieved by all methods mentioned in section 3. All models are overfitted after enough time of training. The performance of ResNet-18 with transfer learning outperforms other methods, shows that more convolutional layers and knowledge learned from a larger image corpus can significantly enhance the model's performance, and help the model converges at a shorter training time.

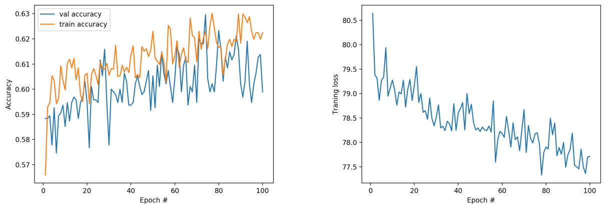
We evaluate the performance of models, mainly based on validation accuracy. And we find other evaluation metrics, including precision, recall, and F1 scores are closely related to validation accuracy. We believe the balanced validation set contributes to this consistency between validation accuracy and other metrics.

Table 1: Model performance evaluation summary

Model	Val acc	Train acc	Precision	Recall	F1	Epoch
Baseline CNN	0.629	0.623	0.681	0.387	0.494	74
AlexNet	0.879	0.960	0.872	0.868	0.870	476
ResNet trained from scratch	0.942	0.986	0.953	0.922	0.937	648
ResNet with transfer learning	0.969	0.996	0.985	0.947	0.966	230

5.2.1 Baseline CNN

Our baseline CNN model peaked at the highest validation accuracy after 74 epochs of training. Figure 4(a) illustrates the training, validation accuracy change during training. After many hyperparameter tweaks, we find the baseline CNN model performs best at a learning rate of 0.01. Figure 4(b) shows the loss change during training. The baseline CNN model is still under-fitted after 100 epochs of training. That suggests the naive CNN model cannot fit our dataset very fell due to its shallow structure. The validation accuracy of the baseline CNN model is significantly less than other complex models that suggest the baseline CNN model cannot learn enough knowledge from the training set due to its limited depth.

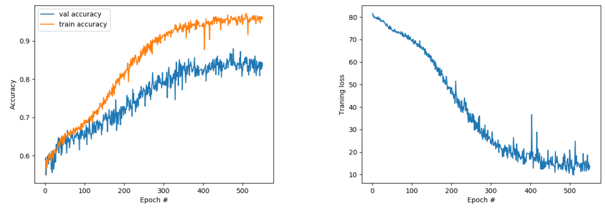


(a) Baseline CNN accuracies (b) Baseline CNN loss

Figure 4: Accuracies, training loss of Baseline CNN

5.2.2 AlexNet

After multiple experiments, the highest validation accuracy achieved by AlexNet is 0.879. AlexNet reaches the best performance at 476 epochs. That means the model is neither underfitting nor overfitting at this epoch. The AlexNet is trained with an SGD optimizer, and it achieves the best performance with a learning rate set to 0.01 and momentum set to 0.9. Figure 5[a] shows the train and validation accuracy achieved by AlexNet at each epoch. Figure 5[b] presents the training loss of AlexNet. AlexNet converges to a very low loss after 500 epochs.

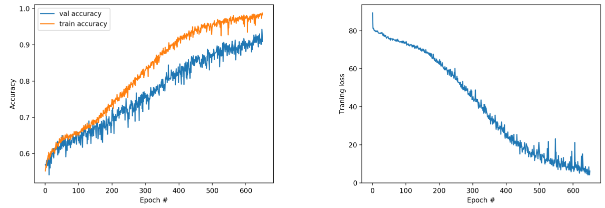


(a) AlexNet accuracies (b) AlexNet loss

Figure 5: Accuracies, training loss of AlexNet

5.2.3 ResNet-18

ResNet-18 has 18 convolutional layers, and it's the deepest model we have experimented with in this project. First, we train a ResNet-18 from scratch, so this model can extract features solely from the breast mass images. Figure 6[a] shows the train and validation accuracy during training. After training 648 epochs, ResNet-18 without transfer learning achieves the highest validation accuracy of 0.942. Compared to AlexNet, the validation accuracy and other evaluation metrics of ResNet-18 are better. That means, with deeper structure and shortcut connections, the ResNet-18 can extract more features from the breast mass images and make a better prediction.

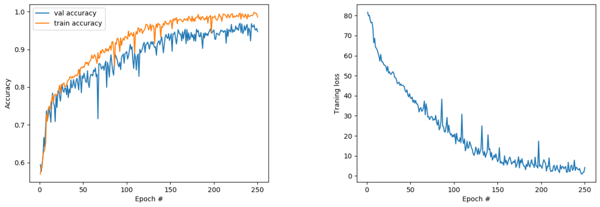


(a) ResNet-18 trained from scratch accuracies (b) ResNet-18 trained from scratch loss

Figure 6: Accuracies, training loss of ResNet-18 without transfer learning

5.2.4 ResNet-18 with transfer learning

The performance of ResNet-18 is further enhanced by leveraged transfer learning techniques. We trained a ResNet-18 that's already fine-tuned on the ImageNet dataset on breast mass images dataset. This model achieves an overall best validation accuracy of 0.969 after 230 epochs of training. We tuned hyperparameters several times, and we find setting the learning rate to 0.001 and momentum to 0.9 is most suitable. Figure [7] presents the accuracies and loss figure of ResNet-18 with transfer learning during training time. Compared to ResNet-18 trained from scratch, ResNet-18 with transfer learning achieves higher accuracy and converges with much shorter training time. The result states the knowledge learning from the ImageNet dataset is helpful in breast mass image identification.

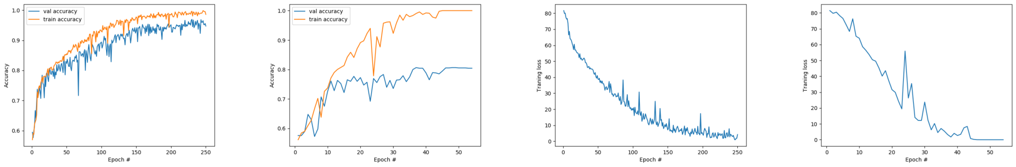


(a) ResNet-18 accuracies (b) ResNet-18 loss

Figure 7: Accuracies, training loss of ResNet-18 with transfer learning

5.3 Augmented dataset vs unaugmented dataset

Since the original dataset is too small, the complex deep learning models, like AlexNet and ResNet-18, can be easily overfitting after training tens of epochs, hence cannot achieve optimal performance. Obviously, the size of the dataset becomes a bottleneck that hinders us from achieving a better result. After we augment the data through the methods listed in section 3.2, AlexNet and ResNet-18 will need hundreds of epochs of training time to overfit the dataset. The performance improves a lot after adopting data augmentation. Figure[] compares the validation and training accuracy of ResNet-18 with transfer learning before and after data augmentation. The figure suggests ResNet-18 has an accuracy of 0.806 and converges at about 50 epochs with the unaugmented dataset and has an accuracy of 0.969 and converges at about 300 epochs with the augmented dataset.



(a) Accuracies with data augmentation (b) Accuracies without data augmentation (c) Loss with data augmentation (d) Loss without data augmentation

Figure 8: Accuracies, training loss of ResNet-18 with and without data augmentation

5.4 Visualization with saliency maps

Saliency map[17] offers a visualization of the pixels in the image that contribute the most to predictions by the model. It plots the gradient of the predicted outcome from the model with respect to pixel values. We implement saliency maps on breast mass images to explore what features of breast masses are used to make the classification, hence make our model more interpretable and explainable. From some saliency map samples listed in Figure 9 and Figure 10, we can find that pixels near the periphery of masses contribute most to the model's prediction.

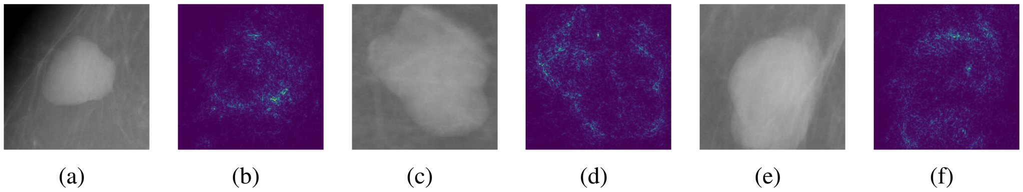


Figure 9: Saliency maps of benign masses with our best model

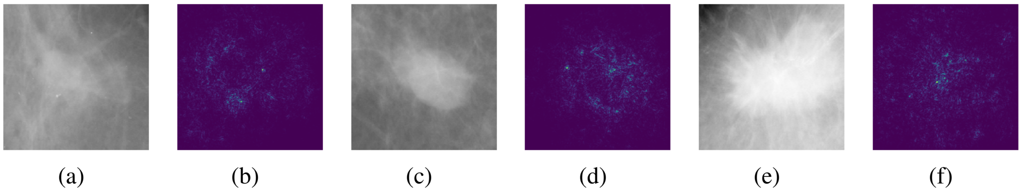


Figure 10: Saliency maps of malignant masses with our best model

6 Conclusion

In this project, we experiment with various deep learning methods to classify breast mass as benign or malignant. Our dataset consists of 4725 breast mass images, which are collected by extracting ROI from a public available mammography database CBIS-DDSM. We achieve state-of-the-art performance with the accuracy of 0.969 in this task with a fine-tuning ResNet-18 model. We train a variety of models, ranging from a simple CNN to the famous ResNet-18. The results show the complexity of the model can significantly enhance accuracy since the baseline CNN model only achieves an accuracy of 0.629. We find using transfer learning can not only further enhance accuracy but will help the model converge quicker because the ResNet-18 trained from scratch can only reach an accuracy of 0.942 after 648 epochs of training. We also find that data augmentation can significantly increase performance. Without data augmentation, a ResNet-18 will converge very soon and can merely achieve an accuracy of 0.806. We visualize our model by implementing saliency maps on breast mass images, and we find the pixels near the periphery of masses contribute most to the model's prediction.

References

[1] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer

[2] D. Levy, A. Jain, Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks, arXiv:1612.00542v1, 2016

[3] Merck Manual of Diagnosis and Therapy (February 2003). "Breast Disorders: Breast Cancer". Archived from the original on 2 October 2011. Retrieved 5 February 2008.

[4] Ashikari, Roy; Park, Keun; Huvos, Andrew G.; Urban, Jerome A. (1970). "Paget's disease of the breast". Cancer. 26 (3): 680–685. doi:10.1002/1097-0142(197009)26:33.0.CO;2-P(inactive 7 December 2019). ISSN 1097-0142.

[5] Carson, William E.; Edge, Stephen B.; Varanasi, Jay S.; Kollmorgen, Daniel R. (1 August 1998). "Paget's disease of the breast: a 33-year experience". Journal of the American College of Surgeons. 187 (2): 171–177. doi:10.1016/S1072-7515(98)00143-4. ISSN 1072-7515.

[6] https://en.wikipedia.org/wiki/Mammography

[7] Ball, J., Bruce, L. Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In: EMBS 2007. 29th Annual International Conference of the IEEE, IEEE (2007) 49734978.

[8] https://medium.com/@ericscuccimarra/convnets-for-classifying-ddsm-mammograms-1739e0fe8028

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

[10] Ragab, D. A., Sharkas, M., Marshall, S., Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ, 7, e6201.

[11] Lee, Rebecca Sawyer, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. “A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research.” Scientific Data 4, no. 1 (2017). https://doi.org/10.1038/sdata.2017.177.

[12] Heath, M., K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran. “Current Status of the Digital Database for Screening Mammography.” Computational Imaging and Vision Digital Mammography, 1998, 457–60. https://doi.org/10.1007/978-94-011-5318-8_75.

[13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” Communications of the ACM 60, no. 6 (2017): 84–90. https://doi.org/10.1145/3065386.

[14] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going Deeper with Convolutions.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. https://doi.org/10.1109/cvpr.2015.7298594.

[15] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. https://doi.org/10.1109/cvpr.2016.90.

[16] Pan, Sinno Jialin, and Qiang Yang. “A Survey on Transfer Learning.” IEEE Transactions on Knowledge and Data Engineering 22, no. 10 (2010): 1345–59. https://doi.org/10.1109/tkde.2009.191.

[17] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013).