Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

SEEM2460@2023
Introduction to Data Science
Final Report

Leveraging CBIS-DDSM Breast Cancer Image Dataset for Developing Data Analytics Approaches to Improve Breast Cancer Detection

Team Members:
Lam Kin Ho 1155158095
Leung Ka Ka 1155159753
Li Wing Lok 1155158024

Contents

1. Introduction

Breast cancer is one of the most common types of cancer among women worldwide. Early detection of breast cancer is crucial for successful treatment and improved patient outcomes. Medical imaging, particularly mammography, is a widely used screening tool for breast cancer detection. However, the interpretation of mammograms can be challenging, and there is a need for more accurate and efficient methods to aid in breast cancer detection.

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is a publicly available dataset that contains mammography images and associated clinical data. This dataset provides a valuable resource for developing data analytics approaches to improve breast cancer detection.

In this project, we aim to leverage the CBIS-DDSM dataset to develop data analytics approaches that can aid in breast cancer detection. We will explore various machine learning and deep learning techniques to analyze mammography images and associated clinical data. Our goal is to develop accurate and efficient methods for breast cancer detection that can assist radiologists in their clinical practice.

By leveraging the CBIS-DDSM dataset, we hope to contribute to the development of more effective and efficient methods for breast cancer detection, ultimately improving patient outcomes and reducing the burden of breast cancer on society.

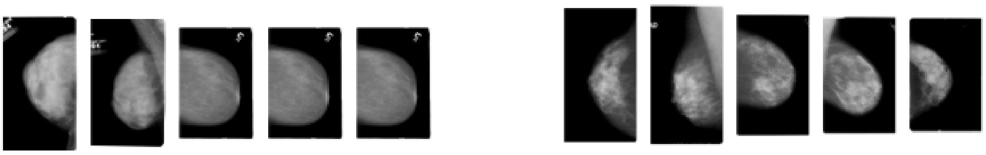


Figure 1: Benign(Left) and Malignant(Right) sample image from training set

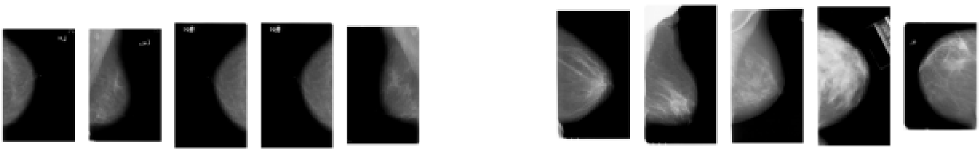


Figure 2: Benign(Left) and Malignant(Right) sample image from testing set

2. Data Overview

The CBIS-DDSM dataset is a publicly available dataset consisting of 10239 annotated mammography images collected from 2620 studies[1]. The dataset includes images from patients with benign and malignant findings, as well as normal cases. Each image in the dataset is accompanied by detailed annotations, including the type of abnormality, its location, and its size.

2.1 Data Description

The CBIS-DDSM dataset is a subset of the Digital Database for Screening Mammography (DDSM) and is curated to include only high-quality images with detailed annotations. The dataset includes images from patients with benign and malignant findings, as well as normal cases. The images are in DICOM format and have a resolution of 3000x4087 pixels.

2.2 Data Processing

The images in the CBIS-DDSM dataset are preprocessed to remove artifacts and enhance the contrast of the images. The preprocessing techniques include image resizing, normalization, and filtering. Feature extraction methods are then applied to extract relevant features from the images, such as texture, shape, and intensity. The dataset is then prepared by dividing it into training, validation, and testing subsets.

2.3 BI-RADS Classification System

The Breast Imaging Reporting and Data System (BI-RADS) is a standardized classification system developed by the American College of Radiology (ACR) to classify mammogram findings and standardize reporting of breast imaging results[2]. The system consists of seven categories (0-6) that indicate the level of suspicion for malignancy.[3]

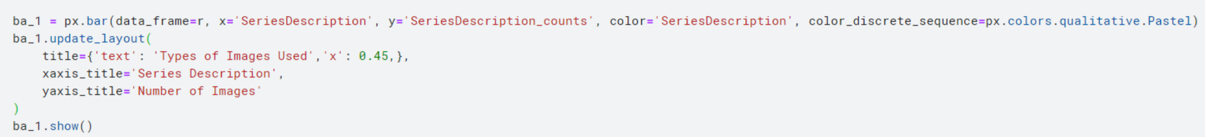
The CBIS-DDSM dataset has been annotated using the BI-RADS system. Each mammogram in the dataset is assigned a BI-RADS category by trained radiologists based on their interpretation of the images. The annotations in the dataset were derived from the BI-RADS categories, which allows for consistent classification of breast imaging findings and facilitates the development of automated classification models.

3. Data Preprocessing

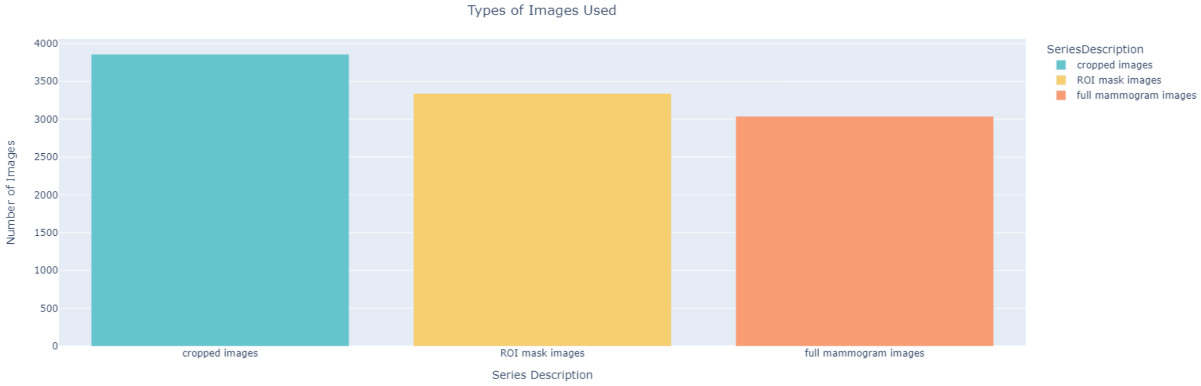
The CBIS-DDSM dataset is visualized to gain insights into the distribution of the images and the characteristics of the abnormalities. The images are visualized using various techniques, such as histograms, scatter plots, and heatmaps. The visualization of the dataset provides insights into the distribution of the abnormalities, the size and location of the abnormalities, and the characteristics of the normal cases.

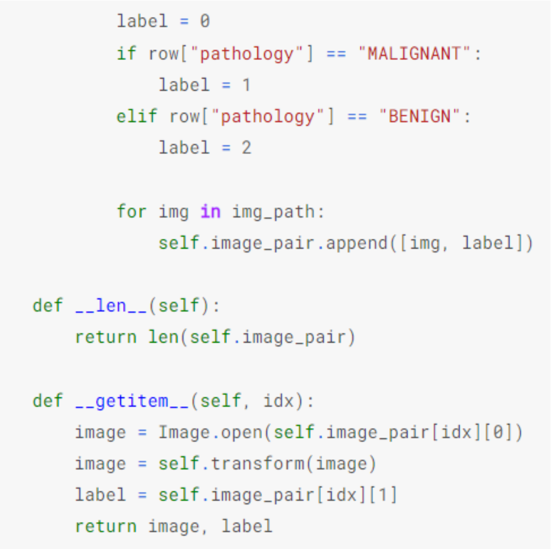
3.1 Image Preprocessing Techniques:

3.1.1 Data Loading and Organization



```
ba_1 = px.bar(data_frame=r, x='SeriesDescription', y='SeriesDescription_counts', color='SeriesDescription', color_discrete_sequence=px.colors.qualitative.Pastel)
ba_1.update_layout(
    title={'text': 'Types of Images Used','x': 0.45,},
    xaxis_title='Series Description',
    yaxis_title='Number of Images'
)
ba_1.show()
```





```
            label = 0
            if row["pathology"] == "MALIGNANT":
                label = 1
            elif row["pathology"] == "BENIGN":
                label = 2

            for img in img_path:
                self.image_pair.append([img, label])

    def __len__(self):
        return len(self.image_pair)

    def __getitem__(self, idx):
        image = Image.open(self.image_pair[idx][0])
        image = self.transform(image)
        label = self.image_pair[idx][1]
        return image, label
```

In this project, the image data was loaded into a custom dataset class called BreastDataset, which inherits from the PyTorch Dataset class. The BreastDataset class takes a pandas dataframe and an optional transform argument. The init method of the class processes the data frame to extract the file paths for the images and the corresponding labels (0 for normal, 1 for malignant, 2 for benign). The len method returns the number of image-label pairs in the dataset, and the getitem method loads an image-label pair from the dataset at the specified index. The transform argument allows for image transformations to be applied to the images such as resizing and converting to tensors.

3.1.2 Image resizing

Image resizing is used to standardize the size of the images in the CBIS-DDSM dataset. In this project, we resize the mammogram images to a size of 224x224 pixels using the PyTorch transform `transforms.Resize((224, 224))`. This ensures that the images are of a consistent size and suitable for feature extraction and analysis.

3.1.3 Image normalization

Image normalization is used to standardize the intensity values of the images. In this project, we used the PyTorch transform `transforms.Normalize()` to normalize the intensity values of the mammogram images. This transform subtracts the mean and divides by the standard deviation of the pixel values in each channel of the image.

3.1.4 Image filtering

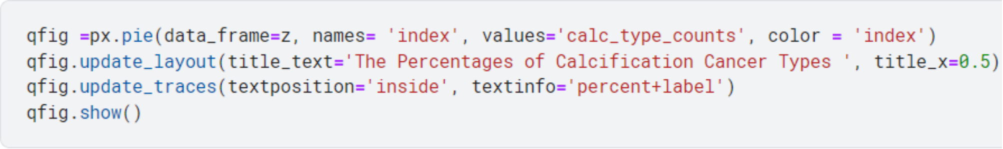
Image filtering is used to remove noise and artifacts from the images. In this project, we did not use a Gaussian filter for image filtering. Instead, we applied data augmentation techniques such as resizing and tensor conversion to the mammogram images as part of the BreastDataset class. This class also extracts the file paths for the images and their corresponding labels (1 for malignant, 2 for benign, and 0 for normal) from the data frames of the CBIS-DDSM dataset.

3.2 Dataset Preparation

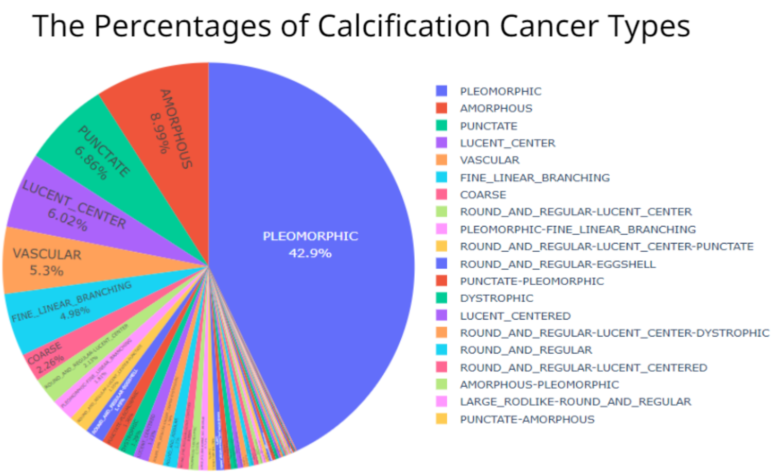
The CBIS-DDSM dataset is prepared by dividing it into training, validation and testing subsets. The training subset is used to train the machine learning models, the validation subset is used to tune the hyperparameters of the models and the testing subset is used to evaluate the performance of the models.

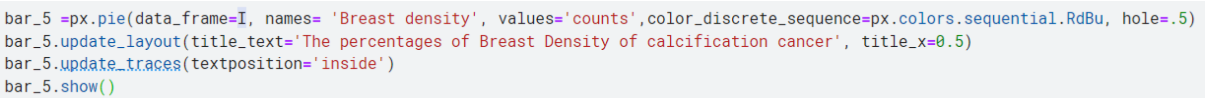
3.3 Visualization of the Dataset

3.3.1 Pie charts of breast calcification variables in cases of breast cancer

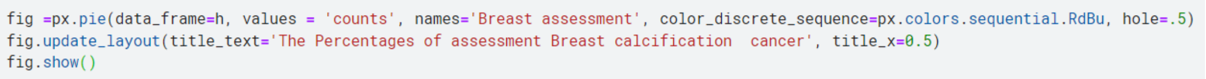


```
qfig =px.pie(data_frame=z, names= 'index', values='calc_type_counts', color = 'index')
qfig.update_layout(title_text='The Percentages of Calcification Cancer Types ', title_x=0.5)
qfig.update_traces(textposition='inside', textinfo='percent+label')
qfig.show()
```

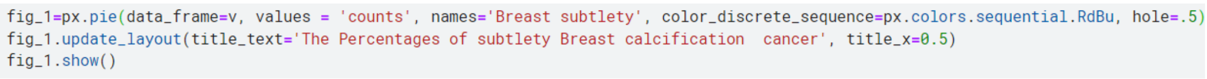




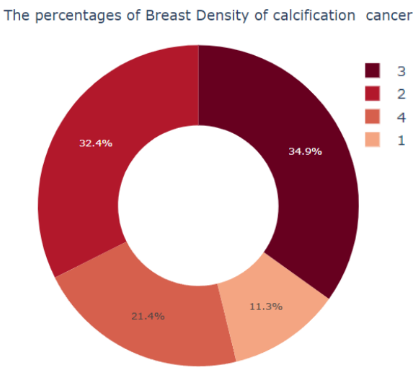
```
bar_5 =px.pie(data_frame=I, names= 'Breast density', values='counts',color_discrete_sequence=px.colors.sequential.RdBu, hole=.5)
bar_5.update_layout(title_text='The percentages of Breast Density of calcification cancer', title_x=0.5)
bar_5.update_traces(textposition='inside')
bar_5.show()
```

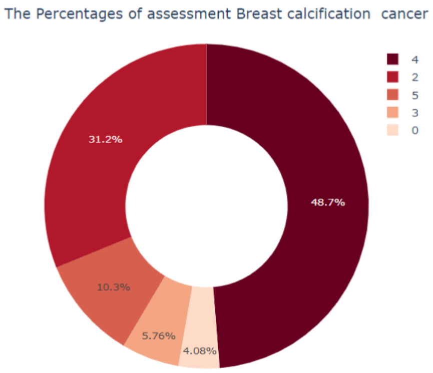


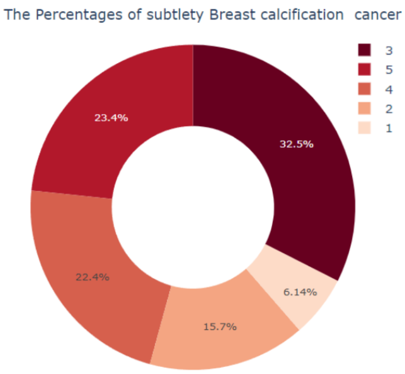
```
fig =px.pie(data_frame=h, values = 'counts', names='Breast assessment', color_discrete_sequence=px.colors.sequential.RdBu, hole=.5)
fig.update_layout(title_text='The Percentages of assessment Breast calcification  cancer', title_x=0.5)
fig.show()
```



```
fig_1=px.pie(data_frame=v, values = 'counts', names='Breast subtlety', color_discrete_sequence=px.colors.sequential.RdBu, hole=.5)
fig_1.update_layout(title_text='The Percentages of subtlety Breast calcification  cancer', title_x=0.5)
fig_1.show()
```

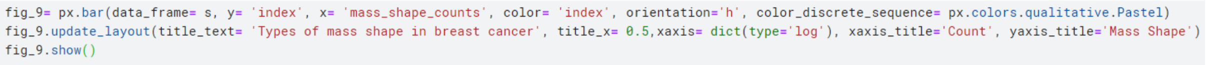




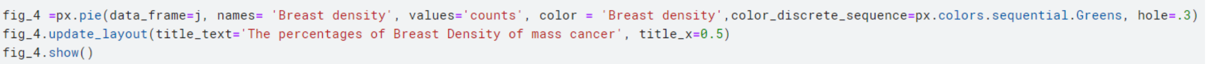


Based on the three pie charts analyzed, we can draw several conclusions about calcification cancer. Firstly, the majority of the breast masses in the dataset have a breast density of 3 or 2, which suggests that breast density may be a significant factor in the development of calcification breast cancer. Secondly, the majority of the breast masses have an assessment rating of 4, which indicates that the majority of the breast masses in the dataset are likely to be malignant. Finally, the majority of the breast masses have a subtlety rating of 3 or 5, which suggests that breast masses with a subtlety rating of 3 or 5 may be more difficult to detect and diagnose.

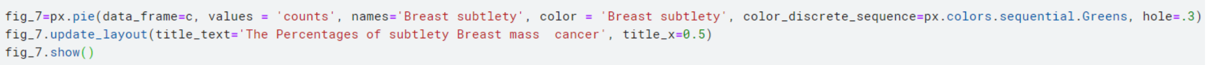
3.3.2 Pie charts of breast mass variables in cases of breast cancer



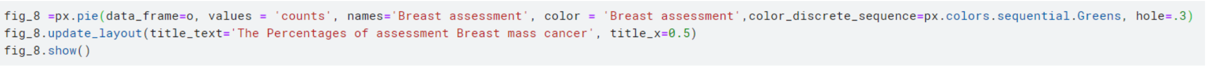
```
fig_9= px.bar(data_frame= s, y= 'index', x= 'mass_shape_counts', color= 'index', orientation='h', color_discrete_sequence= px.colors.qualitative.Pastel)
fig_9.update_layout(title_text= 'Types of mass shape in breast cancer', title_x= 0.5,xaxis= dict(type='log'), xaxis_title='Count', yaxis_title='Mass Shape')
fig_9.show()
```



```
fig_4 =px.pie(data_frame=j, names= 'Breast density', values='counts', color = 'Breast density',color_discrete_sequence=px.colors.sequential.Greens, hole=.3)
fig_4.update_layout(title_text='The percentages of Breast Density of mass cancer', title_x=0.5)
fig_4.show()
```



```
fig_7=px.pie(data_frame=c, values = 'counts', names='Breast subtlety', color = 'Breast subtlety', color_discrete_sequence=px.colors.sequential.Greens, hole=.3)
fig_7.update_layout(title_text='The Percentages of subtlety Breast mass  cancer', title_x=0.5)
fig_7.show()
```



```
fig_8 =px.pie(data_frame=o, values = 'counts', names='Breast assessment', color = 'Breast assessment',color_discrete_sequence=px.colors.sequential.Greens, hole=.3)
fig_8.update_layout(title_text='The Percentages of assessment Breast mass cancer', title_x=0.5)
fig_8.show()
```

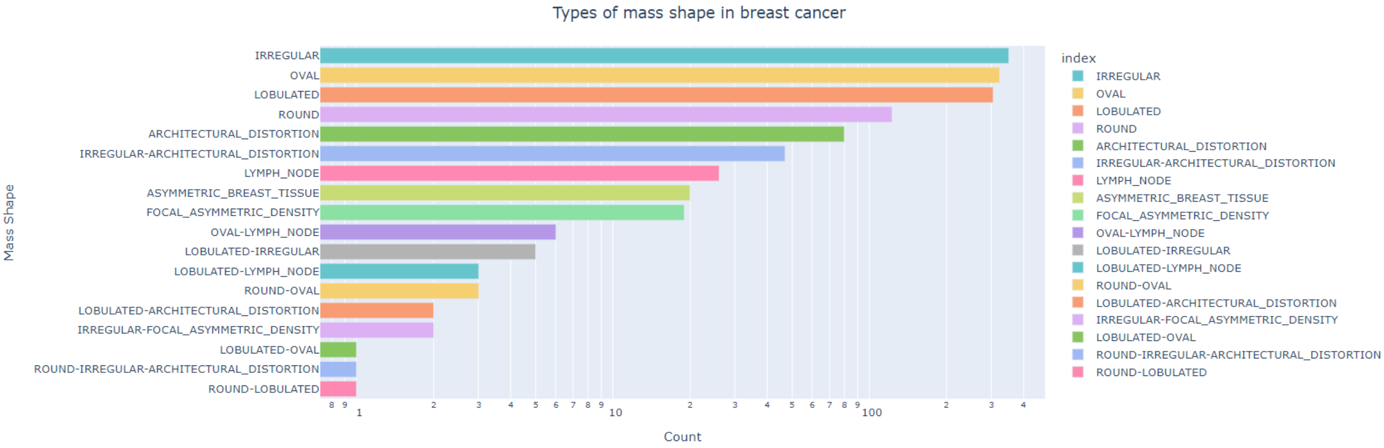


Fig_9

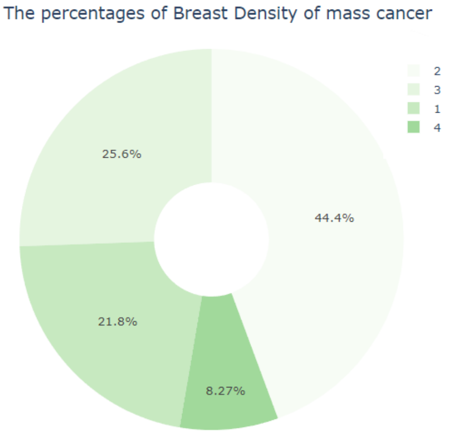


Fig_4

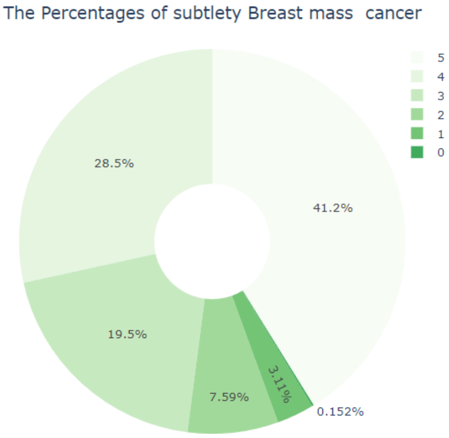


Fig_7

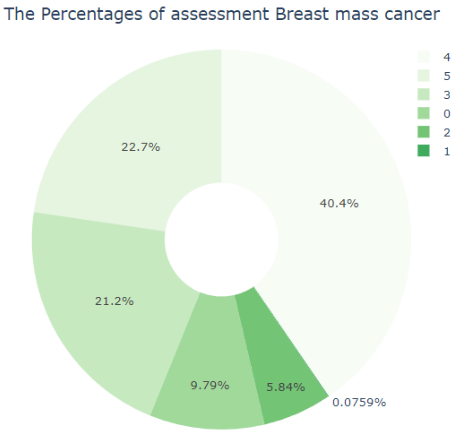


Fig_8

The three pie charts analyzed provide insights into the characteristics of breast masses in the dataset. The majority of the breast masses have a breast density of 2, a subtlety rating of 5, and an assessment rating of 4. These findings suggest that breast density may be a significant factor in the development of breast cancer, breast masses with a subtlety rating of 5 may be more difficult to detect and diagnose, and the majority of the breast masses in the dataset are likely to be malignant as assessment ratings of 4 and above are typically associated with a higher likelihood of malignancy.

4. Approach for Analysis

4.1 Baseline model

We use the same baseline model as Daniel and Arzav in the previous breast cancer detection model[12]. The model is made up of 3 convolutional blocks with 3 × 3 Convolutions - Batch Norm - ReLU - Max Pooling. The convolution layer having stride of 1 and each of 32, 32, 64 channels. While the all of the Max Pooling layers are configured as kernel size and stride 2. After convolutional blocks, the output of last Max Pooling layer is connected to 3 fully-connected layers with sizes of 128, 64, and 2. Finally the class is connected to a CrossEntropy loss function in Pytorch[14], which have a softmax layer before calculating the cross entropy. We use Adam optimizer[15] with a learning rate of 10−3 , batch size 64, and weight decay of 10−5 as a regulation term.

4.2 Transfer Learning

Although medical images differ from natural images. A pre-trained model based on a natural image dataset, ImageNet1K[16], is proven to have faster inference time compared to models trained from scratch[17]. Pretraining on natural images is also useful to increase accuracy on small medical dataset[18]. Considering the small dataset in our experiment (around 1500 images), fine-tuning on a pre-train model is incorporated.

4.3 Augmentation

Data augmentation is usually included in experiments of small datasets. We apply random rotation, horizontal flipping, and cropping to increase the size effectively of our relatively small dataset. Each image is first resize to 256 × 256 image, following by rotation from −5 to +5 degree. A mirroring transformations is applied with probability of 0.5. We also included a random resize crop to 224 × 224.

4.4 Model Selection

We conduct the experiment using a pre-trained model available on Pytorch[14], including AlexNet[6],VGG13 with batch normalization(VGG13BN)[19], and ResNet18[20]. The pretrained model weight on ImgeNet1K is also available on PyTorch[14, 16]. We train all the models, including the baseline model, under the same hyper-parameter and compare the result on present data augmentation work.

4.5 Implementation Detail

We train all of the models with Adam update algorithm[15], base learning rate of 10−4 , batch size 64, and weight decay of 10−5 . A cosine annealing learning rate scheduler[21] and warm up[22] strategy is used. The implementation of the warm up scheduler is available at github[23]. We set the total number of epochs to 200 with patients of 20 epochs. If the valid accuracy has not improved after 20 epochs, we early stop our training process. All of our experiments are conducted on a single Nvidia RTX 3090 Ti.

4.6 Enhancement

4.6.1 Ensemble Learning with Stochastic Gradient Descent (SGD)

To further improve the performance of the breast cancer image diagnosis model, we propose using ensemble learning with stochastic gradient descent (SGD) optimization. Ensemble learning involves combining multiple models to make predictions, while SGD is an alternative optimization method that updates the model parameters using a small batch of random samples at a time.

Ensemble learning is a technique that can enhance the accuracy and reliability of a model by merging multiple models that capture distinct features of the data. To achieve this, we can train several convolutional neural network (CNN) models with different initializations, hyperparameters, or architectures, and then combine their predictions using methods such as majority voting, weighted voting, or stacking. [9]

SGD optimization can improve the convergence speed and performance of the model by updating the parameters more frequently and adapting to the data more efficiently than the current Adam optimizer. [10]

5. Results and Analysis

We compare our results with respect to the baseline architecture. Then conduct a quantitative analysis of our best model. All of the experiment is implemented using Pytorch[11]. We also create a validation set using 20% data from the training set. The number of mammography of training, validation, testing set is 1237, 309, 326 respectively.
The result is shown here

Table 1: Result of Experiment on different model

Model	Final Train Acc	Final Valid Acc	Best Valid Acc	Best Epoch
Without Augmentation				
Baseline	97.993%	65.210%	66.958%	15
AlexNet	78.496%	74.110%	77.670%	17
VGG13BN	98.6719%	75.407%	78.101%	11
ResNet18	99.218%	78.890%	81.604%	19
With Augmentation				
Baseline	99.084%	62.238%	62.937%	26
AlexNet	62.975%	72.168%	74.757%	49
VGG13BN	84.293%	68.706%	70.280%	25
ResNet18	90.622%	74.340%	74.416%	23

5.1 Analysis

5.1.1 Transfer Learning

Transfer Learning provides a more efficient way to train the model and overcome the difficulty of a relatively small dataset. In our experiment, we train the base model from scratch and fine-tuning on the pretrain model. The pretrain model tends to obtain higher accuracy in smaller epochs.

5.1.2 Data augmentation

We would like to overcome the common limitation of most medical dataset, relatively small number of samples by data augmentation. However in our experiment, models without augmentation tend to perform better compared to the same model with augmentation. The reason behind may be the augmentation method is not suitable for whole mammography. Such augmentation methods may work for cropped or segmented image data only[12].

5.2 Performance

We choose ResNet18 without data augmentation to be our final model as it outperform all others model in both final and best validation accuracy. The performance of our model is concluded in Table 2. Confusion matrix and receiver operating characteristic(ROC) curve is provided in Fig.3to visualize the performance as well.

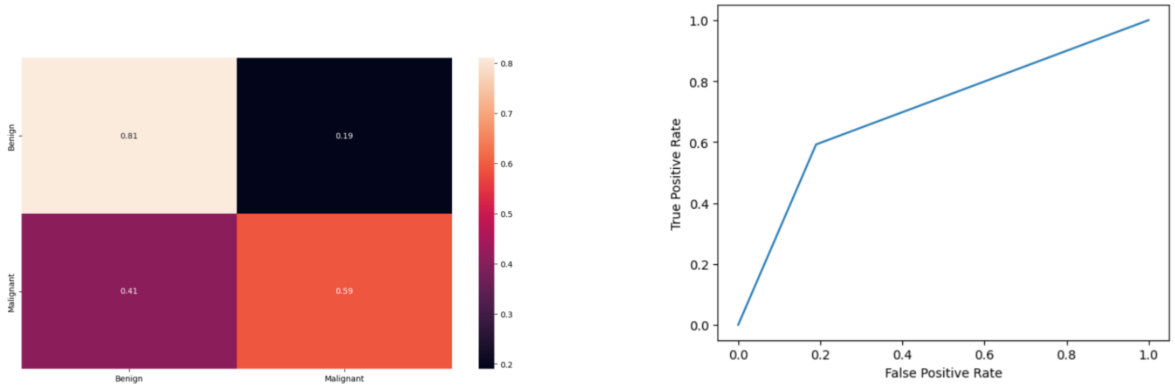


Figure 3: Confusion Matrix(Left) and ROC(Right) of ResNet18 on testing set

6 Conclusion

In our work, we proposed a deep learning model to early detected the breast cancer in the calcification abnormality. Transfer learning of pretrained model on natural image and is included in our work to show effectively to tackle the issue of small dataset. Which is typical for medical data.

Our approach requires minimal human annotation, only a single mammography is needed to feed into the model. There is no requirement to mark a interest area. The result of our work enables an easy and quick adoption in the real clinical data.

Future work includes experiments on other CNN architectures, and integration of the latest attention mechanism. A pipeline including segmentation to mark interest area before feeding the mammography into the classifier may also be feasible work to provide a more concrete result.

7. References

[1] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of digital database for screening mammography (cbis-ddsm)," 2016. [Online]. Available: https://wiki.cancerimagingarchive.net/x/lZNXAQ. [Accessed: May 8, 2023].

[2] H. Barazi and M. Gunduru, "Mammography BI RADS Grading," in StatPearls, updated Aug. 1, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK539816/. [Accessed: May 10, 2023].

[3] Breast Cancer Surveillance Consortium, "BCSC Data Definitions," [PDF] 2018. [Online]. Available: https://www.bcsc-research.org/application/files/5915/4904/2425/BCSC_data_definitions_2018.pdf. [Accessed: May 10, 2023].

[4] J. Brownlee, "Random oversampling and undersampling for imbalanced classification," MachineLearningMastery, [Online]. Available: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/. [Accessed: May 10, 2023].

[5] J. Brownlee, "Ensemble learning methods for deep learning neural networks," MachineLearningMastery, [Online]. Available: https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/. [Accessed: May 10, 2023].

[6] A. Obinguar, "Does batch sizes affect the performance of the INCEPTIONV3 model for image classification?," DEV Community, [Online]. Available: https://dev.to/armlynobinguar/does-batch-sizes-affect-the-performance-of-the-inceptionv3-model-for-image-classification-4c9m. [Accessed: May 10, 2023].

[7] J. Brownlee, "How to configure the learning rate when training deep learning neural networks," MachineLearningMastery, [Online]. Available: https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks/. [Accessed: May 10, 2023].

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, p. 60, May 2019, doi: 10.1186/s40537-019-0197-0.

[9] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," Journal of imaging, vol. 6, no. 6, p. 39, May 2020, doi: 10.3390/jimaging6060039.

[10] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, "The effect of choosing optimizer algorithms to improve computer vision tasks: A comparative study," Multimedia Tools and Applications, vol. 82, no. 11, pp. 16591–16633, 2022. doi:10.1007/s11042-022-13820-0

[11] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[12] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," CoRR, vol. abs/1612.00542, 2016. [Online]. Available: http://arxiv.org/abs/1612.00542

[14] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[17] Y. Xie and D. Richmond, "Pre-training on grayscale imagenet improves medical image classification," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops,September 2018.

[18] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," 2019.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[21] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2017.

[22] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large mini batch sgd: Training imagenet in 1 hour," 2018.

[23] Tony-Y, "pytorch_warmup," https://github.com/Tony-Y/pytorch_warmup, 2022.