Detecting Breast Cancer Using CNN

By Anna Fenner

Why

Breast cancer, constituting over 10% of global cancer cases, is the most prevalent cancer worldwide and the leading cancer among women in the US. Approximately 1 in 8 women in the US will receive a breast cancer diagnosis in their lifetime (1, 2).

Invasive Ductal Carcinoma (IDC), the predominant form of breast cancer (80% of cases), originates in the milk ducts and infiltrates surrounding breast tissue (5). Its potential to spread to lymph nodes or the bloodstream escalates its severity (6).

Primary risk factors for breast cancer include gender and aging (2). A first-degree relative's breast cancer diagnosis doubles the risk, affecting 15% of women. Additionally, mutations in BRCA1 and BRCA2 genes elevate susceptibility (1).

Efficient breast cancer diagnosis is crucial. Automating the classification and categorization process, especially using Convolutional Neural Networks (CNNs), proves effective and can significantly save time and resources. CNNs have demonstrated high accuracy not only in breast cancer diagnosis but also in detecting other conditions like Alzheimer’s Disease (3).

Data

This project utilized the Kaggle dataset focusing on Breast Histopathology Images (4). The dataset comprises 277,524 patches. Each patch is 50 x 50 pixels that was extracted from 162 whole mount slides scanned at 40x magnification. Within these patches, 198,738 are IDC negative, while 78,786 are IDC positive.

The data is organized into 279 folders, each labeled by patient ID. Each patient folder contains two subfolders. These subfolders are categorized images as IDC negative (labeled 0) and IDC positive (labeled 1) (3).

Data Wrangling

During the data wrangling phase, I undertook the following steps:

1. Downloaded the dataset.
2. Constructed a data frame incorporating patient IDs, file pathways for photo access, and corresponding labels (0 for IDC negative, 1 for IDC positive).
3. Randomly sampled and retained one-fourth of the dataset for further analysis.

Exploratory Data Analysis

During the exploratory data analysis step, a bar graph was plotted to indicate the two classes. There were 198,738 IDC negative (0) images and 78,786 IDC positive (1) images, ensuring alignment with the intended classifications.

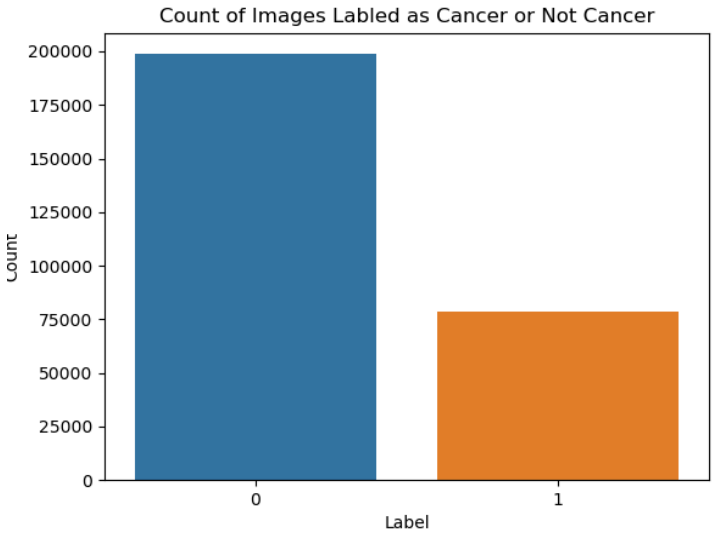

Figure 1: Bar plot to show the number of observations IDC negative (0) and IDC positive (1).

Subsequently, a bar plot depicted the image counts per patient. This unveiled variations where certain patients had over 2000 images, while others had fewer than 500.

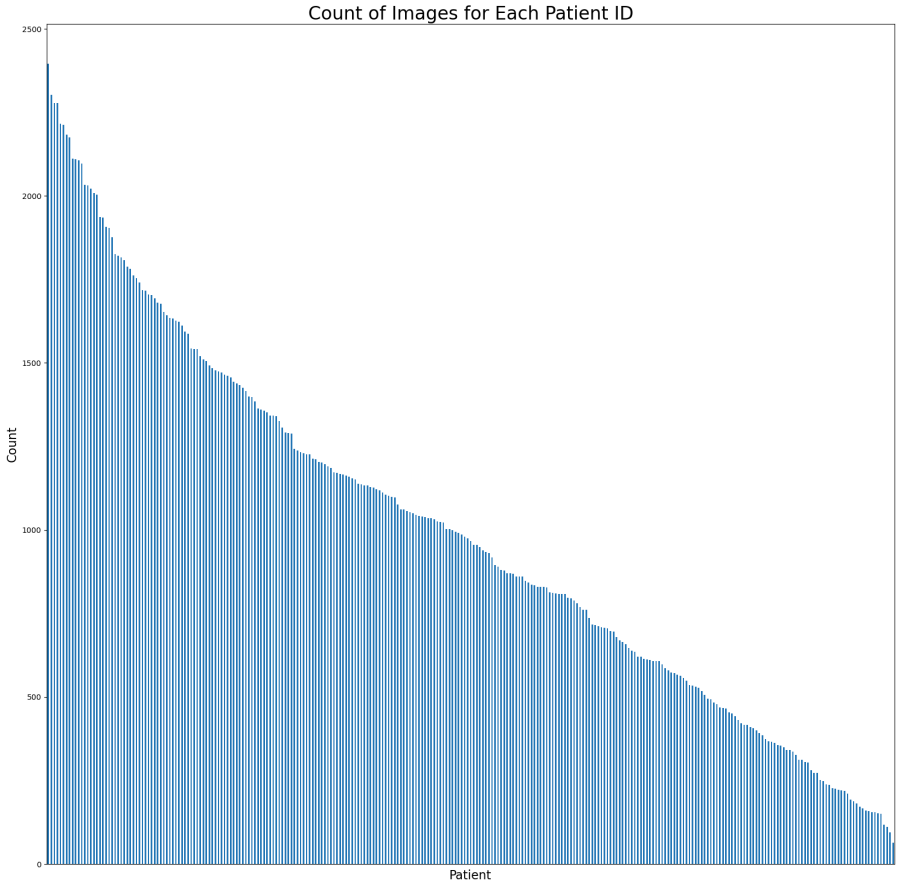

Figure 2: Bar plot to show the count of images for each patient.

Concluding the exploration, visual insights into IDC positive and IDC negative images were done to look at the differences from the human eye.

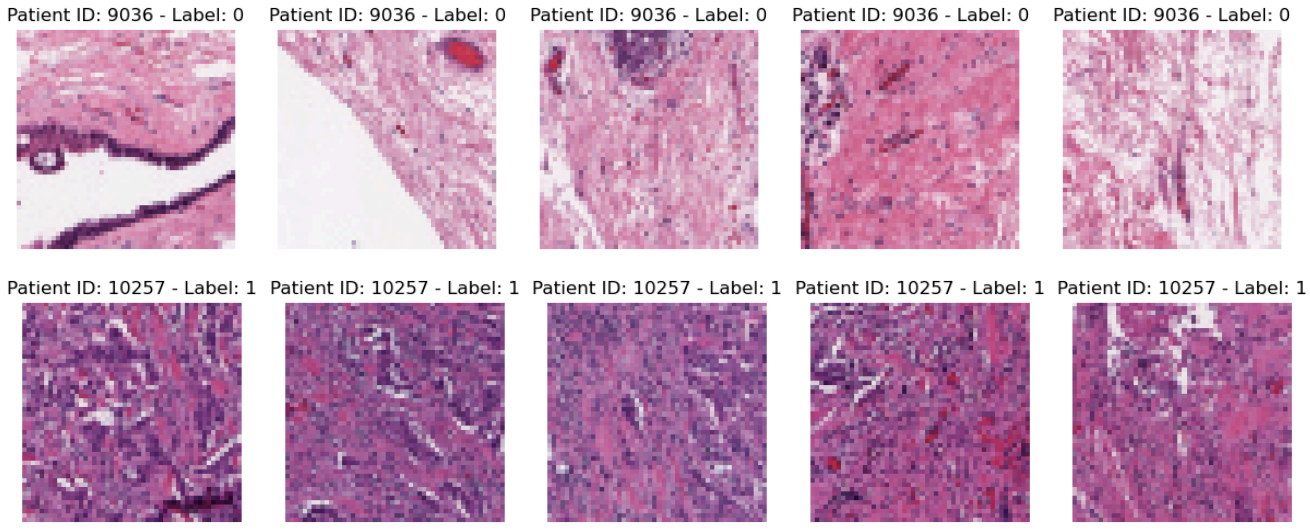

Figure 3: Images showing two different patients. First patient's images are IDC negative. Second patient's images are IDC positive.

Preprocessing the Data

In the preprocessing phase, the following steps were executed:

1. Normalized the images so pixel values fell within the range of 0 to 1.
2. Applied oversampling to the minority group (IDC negative) to balance representation to approximately 50% of the majority group.
3. Applied undersampling for the majority group (IDC positive) to achieve an equal number of instances in both groups– 44643 images each.
4. Transformed images into 4D-numpy arrays to prepare them for the CNN model.
5. Partitioned the data into training (80%) and testing (20%) sets for the model.

Modeling

Model 1:

For the modeling, applied a convolution neural net with exploration into trying different layers, optimizers, activation functions, and augmenting the images. In the best CNN model architecture, the sequence of layers is as follows:

First, Second, Third Convolutional Layers:

- Utilized 8 filters with a size of 2 x 2 and applied the rectified linear unit (ReLU) activation function to introduce non-linearity.
- Subsequently used a max-pooling layer (`MaxPooling2D`) for spatial down-sampling.

Flattening and Fully Connected Layers:

- Flattened the output from the convolutional layers into a one-dimensional array.
- Introduced a fully connected layer (`Dense`) with 256 units and ReLU activation for complex feature combination.
- The final output layer consists of a single unit with a sigmoid activation function, suitable for the binary classification task, predicting the probability of a sample belonging to the positive class (IDC positive).

Additionally:

- The ‘adam’ optimizer was chosen for efficient gradient-based optimization.
- BinaryCrossentropy served as the loss function, apt for binary classification tasks.
- The metric used for model evaluation during training is accuracy.

Training:

- The model was trained for 20 epochs.

- A validation set, comprising 10% of the training data, was utilized to monitor generalization performance.

Model 2: Pre-trained Transfer Learning - VGG16

Model Selection & Customization:

- Utilized a pre-trained VGG16 convolutional neural network (CNN) architecture, pre-loaded with weights trained on the ImageNet dataset.
- Excluded the top classification layers of the VGG16 model to replace them with custom layers tailored to the binary classification task.
- Froze all layers of the VGG16 model to prevent further training of their weights, ensuring that only the custom layers added on top would be trained.

Model Architecture:

- Flattened the output from the convolutional layers of VGG16 into a one-dimensional array.
- Added a fully connected layer with 256 units and ReLU activation to enable complex feature combinations.
- Defined the final output layer comprising a single unit with a sigmoid activation function.

Additionally:

- The ‘adam’ optimizer was chosen for efficient gradient-based optimization.
- BinaryCrossentropy served as the loss function.
- The metric used for model evaluation during training is accuracy.

Training:

- Trained the model for 20 epochs to iteratively update the model's weights based on the training data.
- Utilized a validation set comprising 10% of the training data to monitor the model's generalization performance and prevent overfitting.

Model 3: Pre-trained Transfer Learning - ResNet50

Model Selection & Customization:

- Utilized a pre-trained ResNet50 convolutional neural network (CNN) architecture, pre-loaded with weights trained on the ImageNet dataset, while excluding the top classification layers.
- Added custom classification layers on top of the ResNet50 base model to adapt it

for a specific binary classification task.
- Applied a global average pooling layer to reduce the spatial dimensions of the feature maps.
- Introduced a fully connected layer with 256 units and ReLU activation to facilitate complex feature combinations.
- Defined the final output layer with a single unit and sigmoid activation function, suitable for binary classification tasks.

Model Architecture:
- Created the transfer learning model by specifying the input and output layers.
- Froze all pre-trained layers of the ResNet50 base model to prevent further training and retain the learned feature representations.

Additionally:
- The ‘adam’ optimizer was chosen for efficient gradient-based optimization.
- BinaryCrossentropy served as the loss function.
- The metric used for model evaluation during training was accuracy.

Training:
- Trained the model for 20 epochs with a batch size of 32 to iteratively update the model's weights based on the training data.
- Utilized a validation set comprising 10% of the training data to monitor the model's generalization performance and prevent overfitting.

Results:

The CNN model, featuring a simple architecture of 3 convolutional layers, emerged as the top performer, showcasing consistent improvement in both training and validation accuracies. Both metrics reached an impressive accuracy of 83.36%.

Notably, significant reductions in training and validation losses were observed throughout training, settling at 39.21% and 39.09%, respectively.

In the final evaluation on the testing set, the model demonstrated strong performance with an accuracy of 83.9% and a corresponding test loss of 37.9%. Furthermore, the ROC curve analysis revealed an impressive AUC score of 0.915.

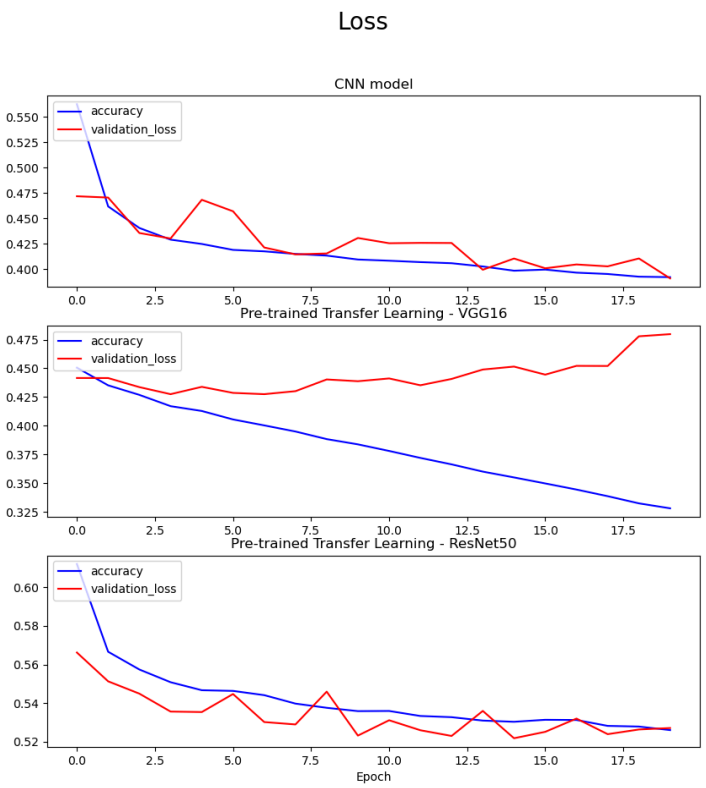

Figure 4: Graphs for the three models showing the training and validation loss across 20 epochs.

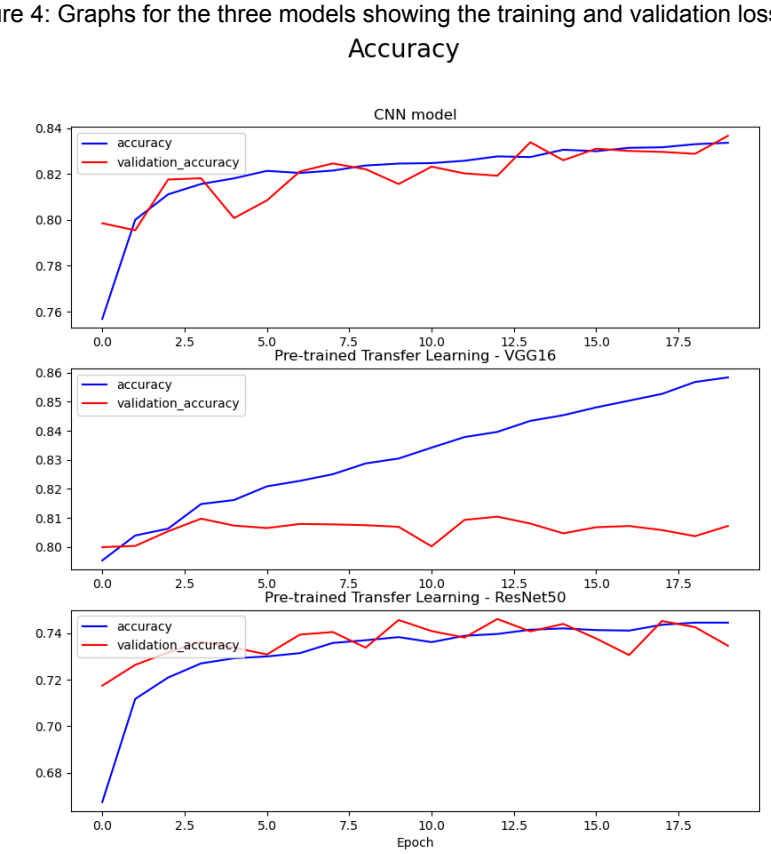

Figure 5: Graphs for the three models showing the training and validation accuracy across 20 epochs.

CNN Model	VGG16	ResNet50
AUC score: 0.915	AUC score: 0.885	AUC score: 0.821
Test loss: 0.379 Test accuracy: 0.839	Test loss: 0.466 Test accuracy: 0.814	Test loss: 0.530 Test accuracy: 0.739

Table 1: Table showing AUC score, testing loss and testing accuracy across the three models.

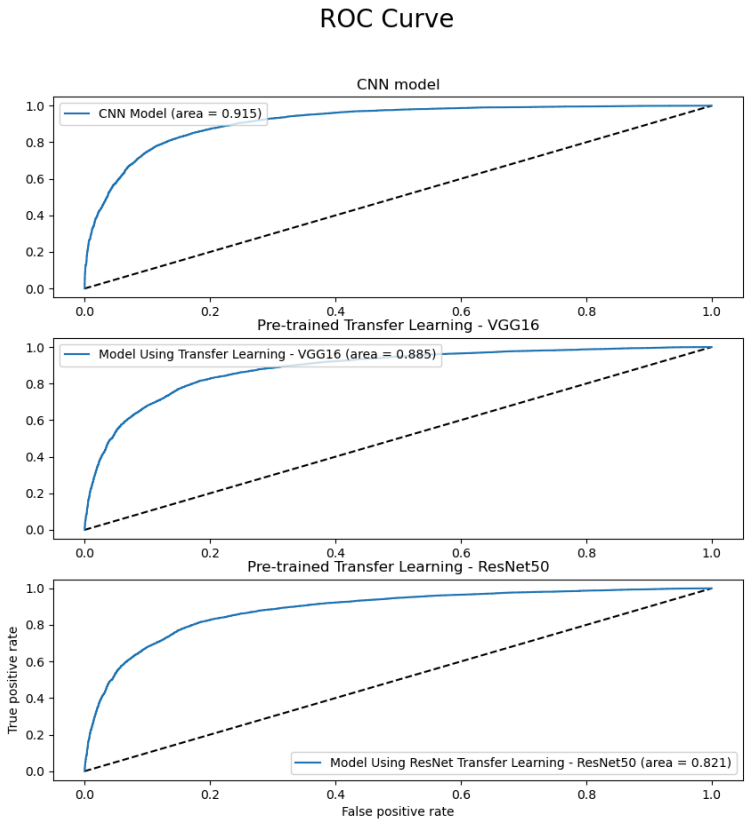

Figure 6: ROC graph for the three models.

Conclusion

Three models were explored. A simple CNN model with three convolution layers, VGG16 used for transfer learning, and ResNet50 used for transfer learning. Further, efforts to enhance model performance through additional layers, filters, activation functions, image augmentation, and optimizer variations yielded suboptimal results. Notably, only half of the available images were utilized across all models.

The CNN model that outperformed the others maintained simplicity with three convolutional layers, each employing 8 filters and utilizing ReLU activation functions. With an AUC score of 0.915, the model demonstrated commendable effectiveness in

detecting both IDC-positive and IDC-negative images.

Future Improvements

Future endeavors will involve harnessing cloud services to bolster processing power and train models over 50 epochs or more. Additionally, expanding the dataset by incorporating previously unused images will be crucial for further exploration and model refinement.

References

1. *Breast cancer facts and statistics 2024*. (n.d.). https://www.breastcancer.org/facts-statistics?gad_source=1&gclid=Cj0KCQiAhomtBhDgARIsABcaYylH3OZmhgmy4s9zyUghSbIIYPOGdo97UYLI_ej-FjqAxDrJrADTM7caArEAEALw_wcB
2. *Basic information about breast cancer*. (2023, July 27). Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/index.htm#:~:text=Each%20year%20in%20the%20United,cancer%20than%20all%20other%20women
3. Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A study of CNN and transfer learning in Medical imaging: Advantages, challenges, future scope. *Sustainability*, *15*(7), 5930. https://doi.org/10.3390/su15075930
4. *Breast histopathology images*. (2017, December 19). Kaggle. https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images
5. *Invasive Ductal Carcinoma (IDC)*. Pennmedicine.org. (n.d.). https://www.pennmedicine.org/cancer/types-of-cancer/breast-cancer/types-of-breast-cancer/invasive-ductal-carcinoma#:~:text=Invasive%20ductal%20carcinoma%20(IDC)%2C,other%20areas%20of%20the%20body
6. *Invasive ductal carcinoma (IDC)*. (2023, March 21). Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/invasive-ductal-carcinoma-idc