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.

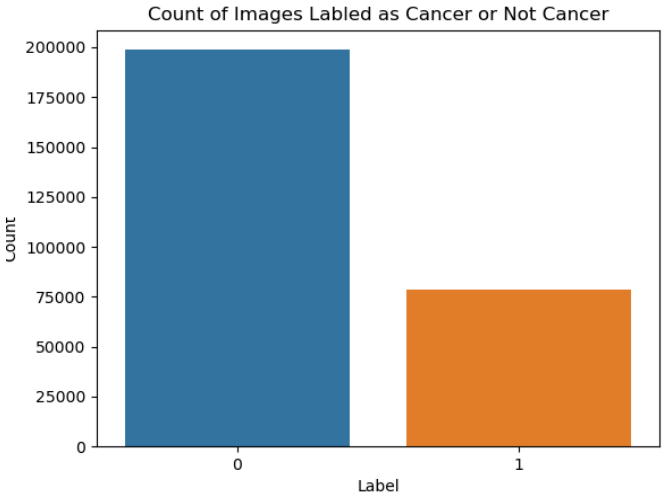

Figure 1: Bar plot to show the number of observations IDC negative (0) and IDC positive (1).

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.

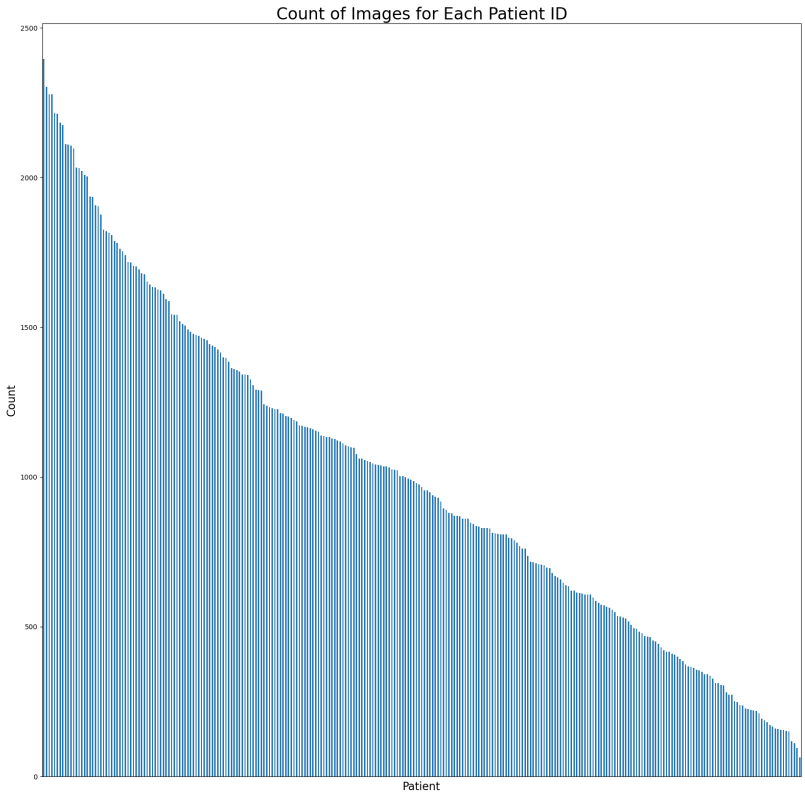

Figure 2: Bar plot to show the count of images for each patient.

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

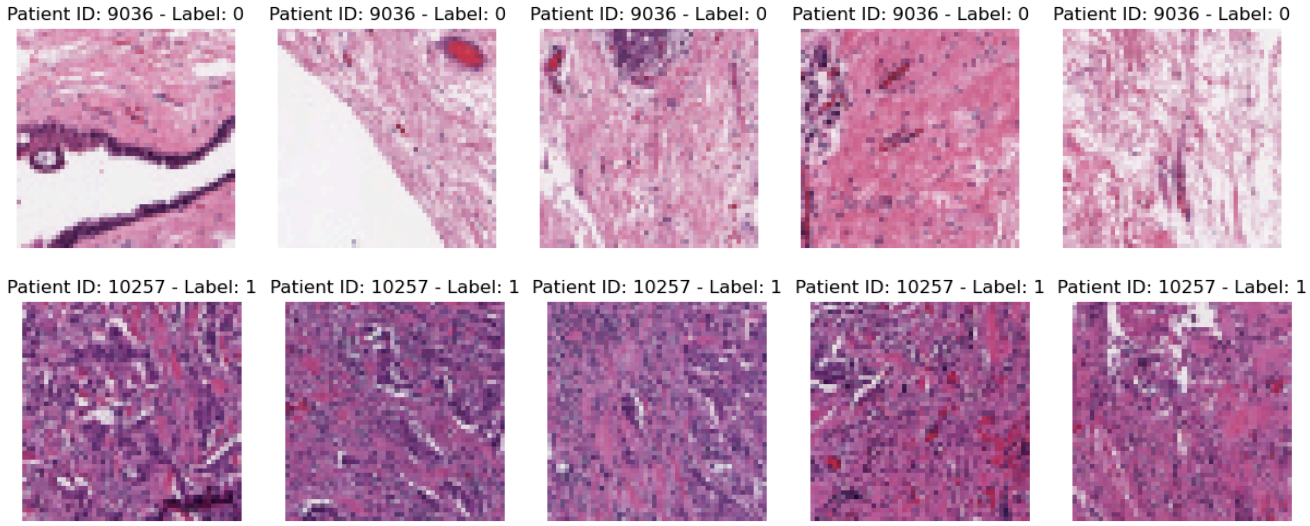

Figure 3: Images showing two different patients. First patient's images are IDC negative. Second patient's images are IDC positive.

Preprocessing the Data

In the preprocessing phase, I executed the following steps to optimize the dataset:

1. Normalized the images, ensuring 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.
4. Transformed images into numpy arrays, preparing them for the CNN model.
5. Partitioned the data into training (80%) and testing (20%) sets for model.

Modeling

In the CNN model architecture, the sequence of layers is as follows:

1. First Convolutional Layer:
   - Utilized 8 filters with a size of 2 x 2.
   - Applied the rectified linear unit (ReLU) activation function to introduce non-linearity.
   - Subsequently used a max-pooling layer (`MaxPooling2D`) for spatial down-sampling, aiding in preserving essential features.

2. Second Convolutional Layer:
   - Employed 16 filters of size 2 x 2 with ReLU activation for further feature extraction.
   - Followed by another max-pooling layer to reduce dimensionality while retaining crucial information.

3. Third Convolutional Layer:
   - Used 8 filters of size 2 x 2 with ReLU activation.
   - Followed by a max-pooling layer.

4. 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.

The model was trained for 20 epochs. A validation set, comprising 10% of the training data, was utilized to monitor generalization performance.

Results

Throughout the 20 epochs of training, the model consistently demonstrated improvement in accuracy in the training dataset. It culminated to 86.80%. On the validation set, the accuracy reached 83.37% after epoch 20. The validation accuracy had minor fluctuations observed throughout.

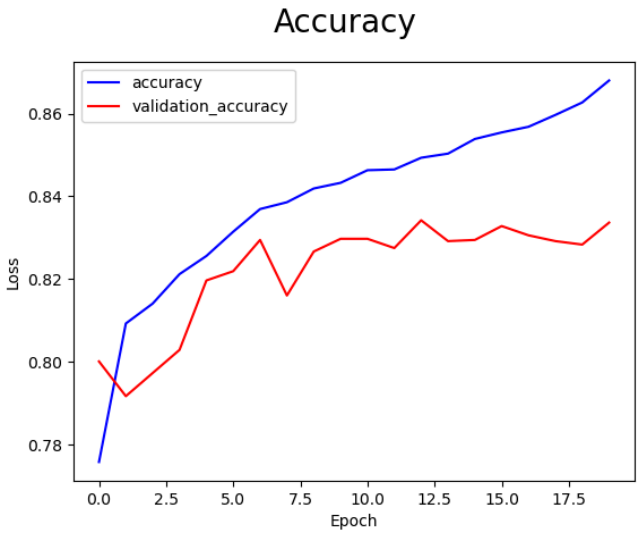

Figure 4: Graph showing the training and validation accuracy across 20 epochs.

In terms of loss metrics, the training loss experienced a notable decrease, settling at 32.17%, while the validation loss also exhibited a reduction, recording 38.05%.

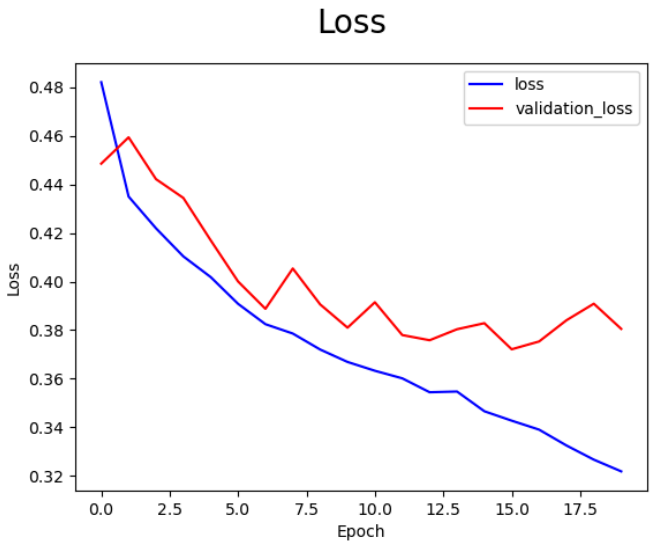

Figure 5: Graph showing the training and validation loss across 20 epochs.

For the final evaluation on the testing set, the model delivered an accuracy of 84.76% and a corresponding test loss of 36.98%.

The ROC curve was plotted and had a AUC score of 0.918.

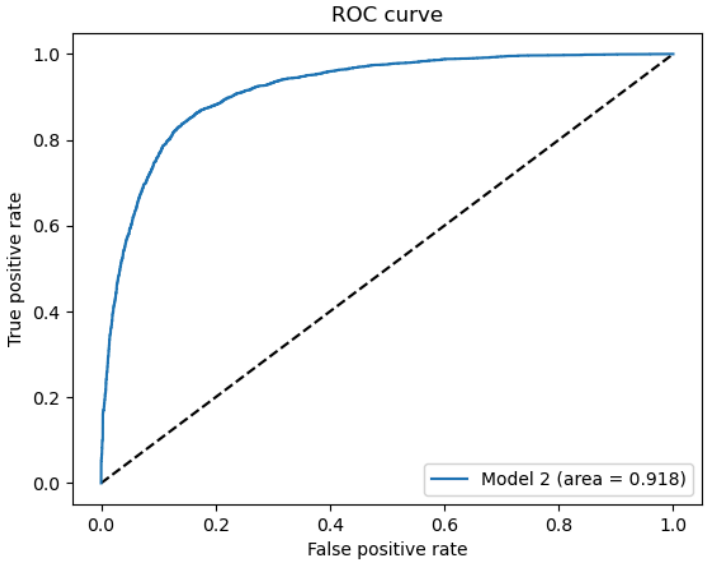

Figure 6: Plotting ROC curve.

Conclusion

The model maintains simplicity with just three convolution layers, employing 8 or 16 filters and utilizing ReLU activation functions for each. Demonstrating a robust

performance, the AUC score of 0.918 highlights the model's effectiveness in detecting true positive cases, distinguishing IDC-positive images, and accurately identifying IDC-negative images.

Attempts to introduce additional layers, filters, and alternative activation functions resulted in overfitting, causing a decline in accuracy across both validation and testing datasets, accompanied by significant increases in loss scores. Notably, despite using a reduced dataset of around 44,000 patches compared to 277,524, the model achieved commendable testing accuracy, reaching 84.76% in detecting IDC positivity.

Future Improvements

Future work involves leveraging cloud services for enhancing processing power. Subsequently, expanding the dataset with additional images and delving into fine-tuning the convolutional layer architecture—potentially by adding more filters and layers—will be pivotal directions for further exploration.

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