Detecting Breast Cancer Using CNN

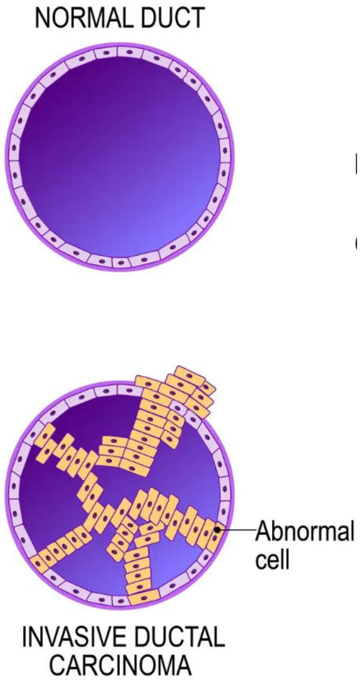
Anna Fenner

Why

https://www.karmanos.org/karmanos/news/you-may-have-heard-a-survivor-mention-these-breast-4733

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

Why



https://www.facs.org/for-patients/home-skills-for-patients/breast-cancer-surgery/breast-cancer-types/invasive-breast-cancers/

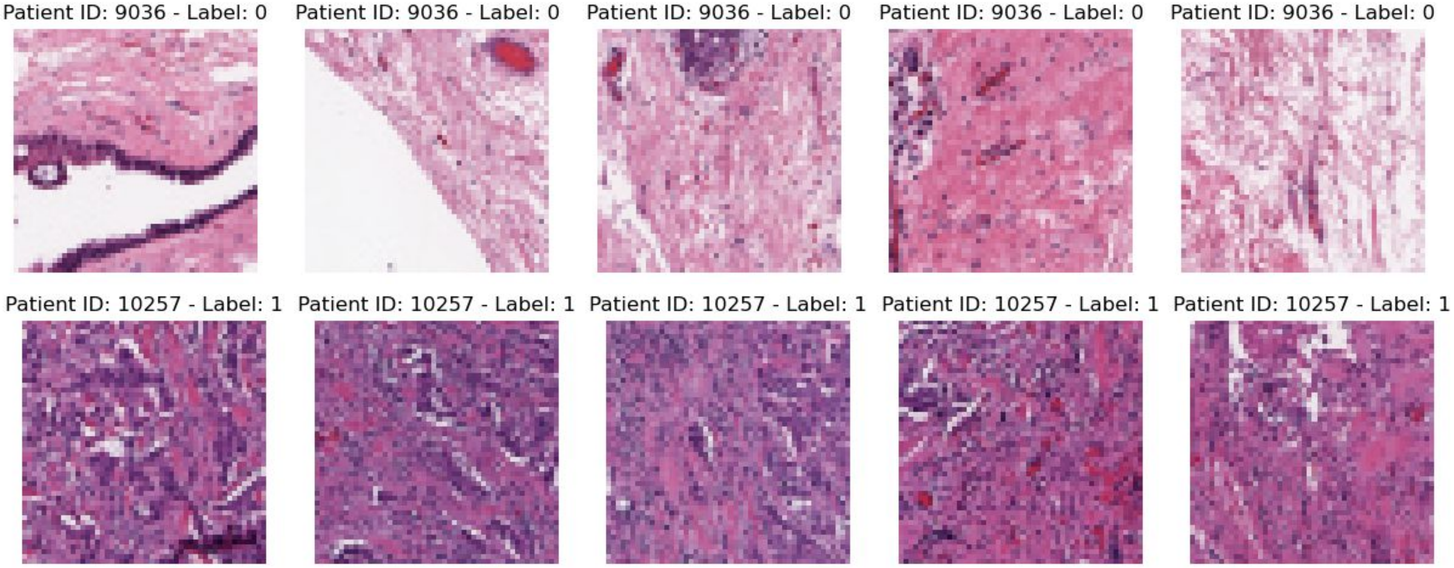
Invasive Ductal Carcinoma (IDC), the predominant form of breast cancer (80% of cases), originates in the milk ducts and infiltrates surrounding breast tissue (5). Its potential to spread to lymph nodes or the bloodstream escalates its severity (6).

Automating the diagnosis and categorization process for breast cancer is crucial. Convolutional Neural Networks (CNNs) has proven to be effective for this and can significantly save time and resources (3).

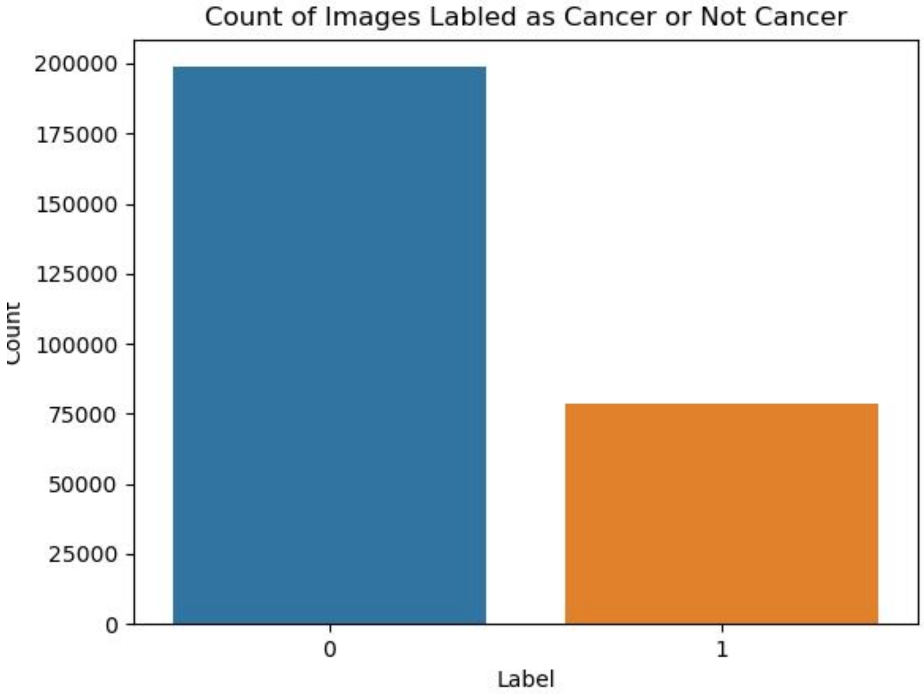
The Data

This project utilized the Kaggle dataset on Breast Histopathology Images (4).

- The dataset comprises of 277,524 patches: 198,738 patches were IDC negative (top 5 images) and 78,786 were IDC positive (bottom 5 images).
- Each sized 50 x 50 pixels
- Patches extracted from 162 whole mount slides scanned at 40x magnification

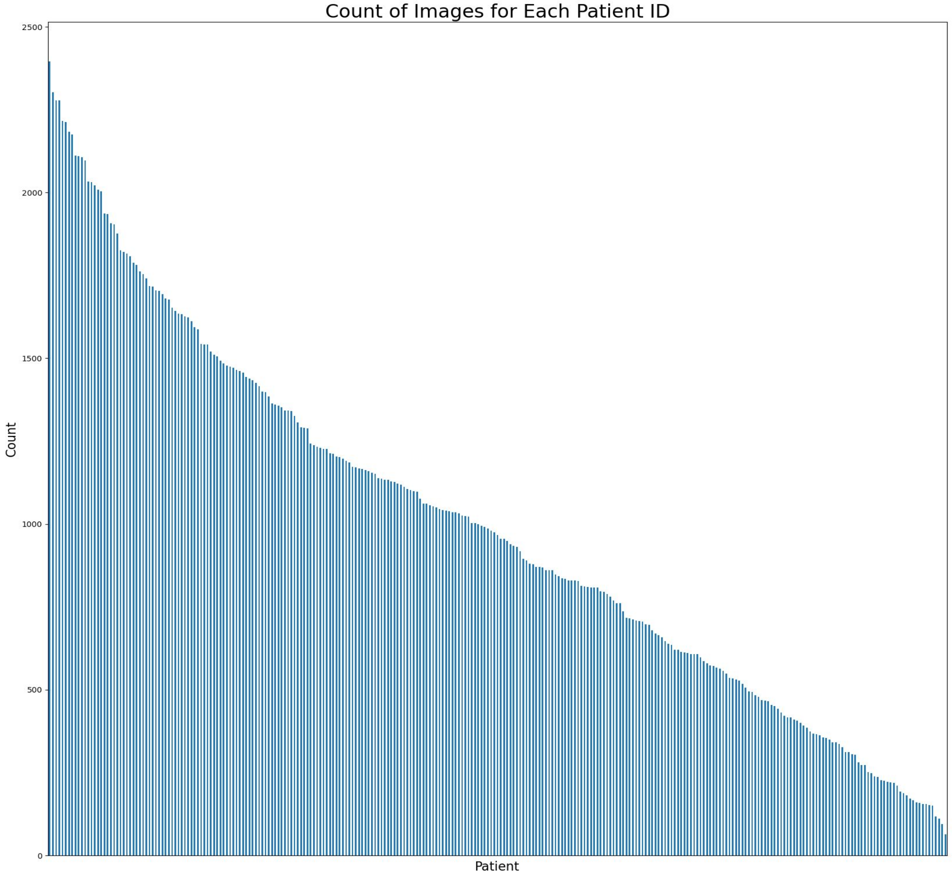

Patient ID: 9036 - Label: 0
Patient ID: 9036 - Label: 0
Patient ID: 9036 - Label: 0
Patient ID: 9036 - Label: 0
Patient ID: 9036 - Label: 0
Patient ID: 10257 - Label: 1
Patient ID: 10257 - Label: 1
Patient ID: 10257 - Label: 1
Patient ID: 10257 - Label: 1
Patient ID: 10257 - Label: 1

Insights from Exploring the Data



Ensured that the dataset did indeed have 277,524 patches, 198,738 patches were IDC negative and 78,786 were IDC positive.

Insights from Exploring the Data



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

3 Models Explored for Image Classification:

1. **CNN Model with Simple Architecture:** Utilized a straightforward architecture with 3 simple convolutional layers.
2. **VGG16:** Leveraged VGG16, a CNN model, for transfer learning.
3. **ResNet50:** Employed ResNet50 model for transfer learning approach.

CNN Best Model Architecture

First, Second, and Third Convolutional Layers:

- Utilized 8 filters with a size of 2 x 2.
- Applied the rectified linear unit (ReLU) activation function.
- Subsequently used a max-pooling layer (`MaxPooling2D`).

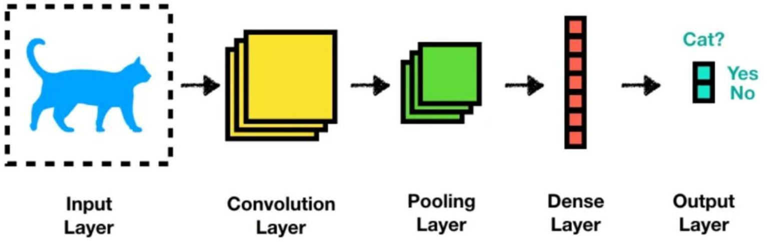
Flattening and Fully Connected Layers:

- 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:

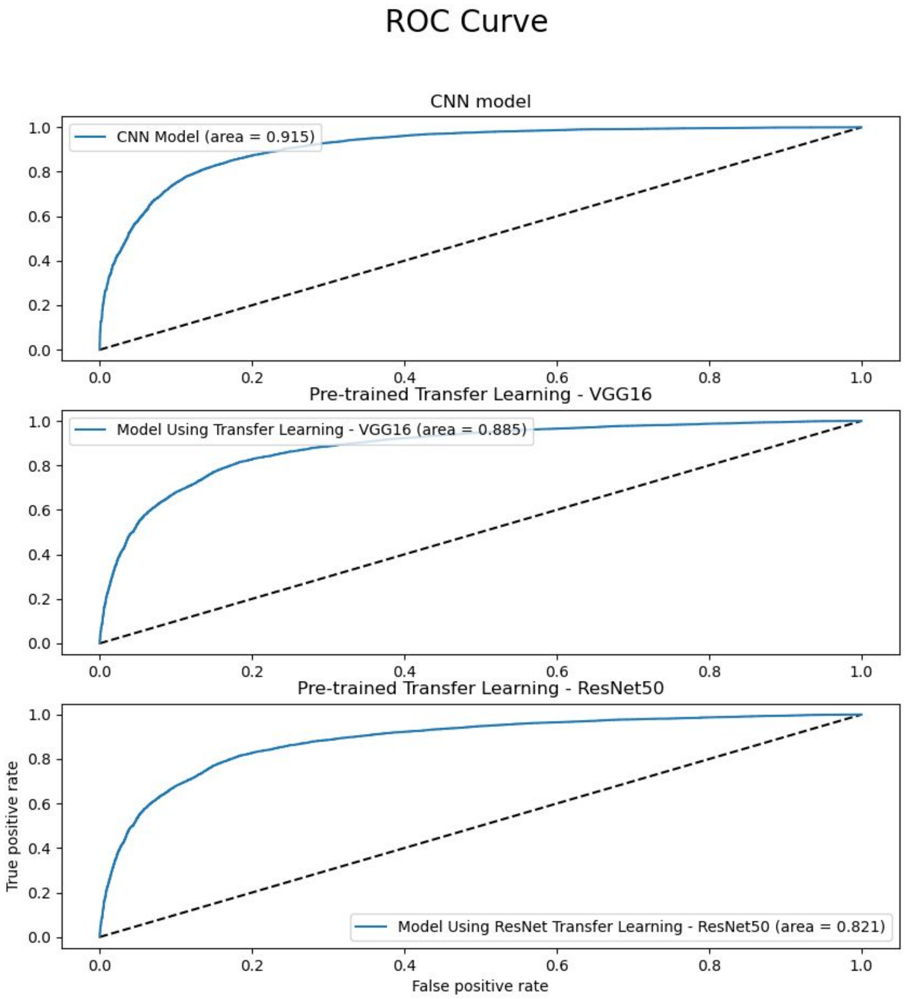
- The ‘adam’ optimizer was chosen for efficient gradient-based optimization.
- BinaryCrossentropy served as the loss function, apt for binary classification tasks.
- The metric used for model evaluation during training is accuracy.

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

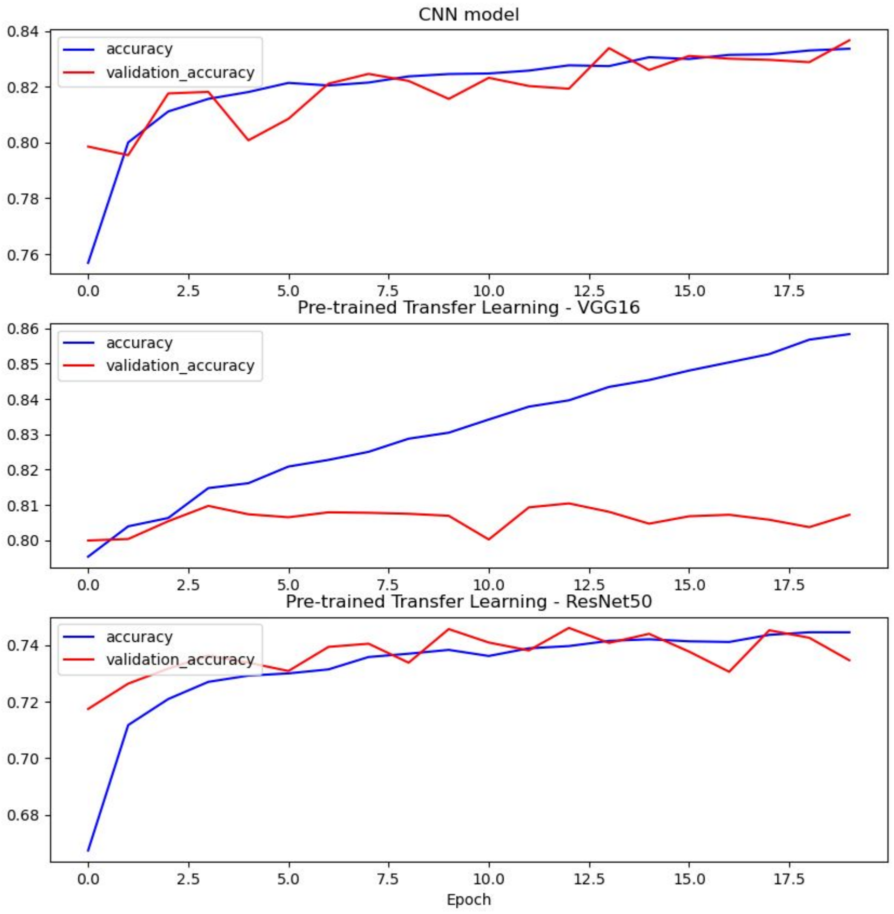


https://towardsdatascience.com/convolutional-neural-network-a-step-by-step-guide-a8b4c88d6943

Model Performance



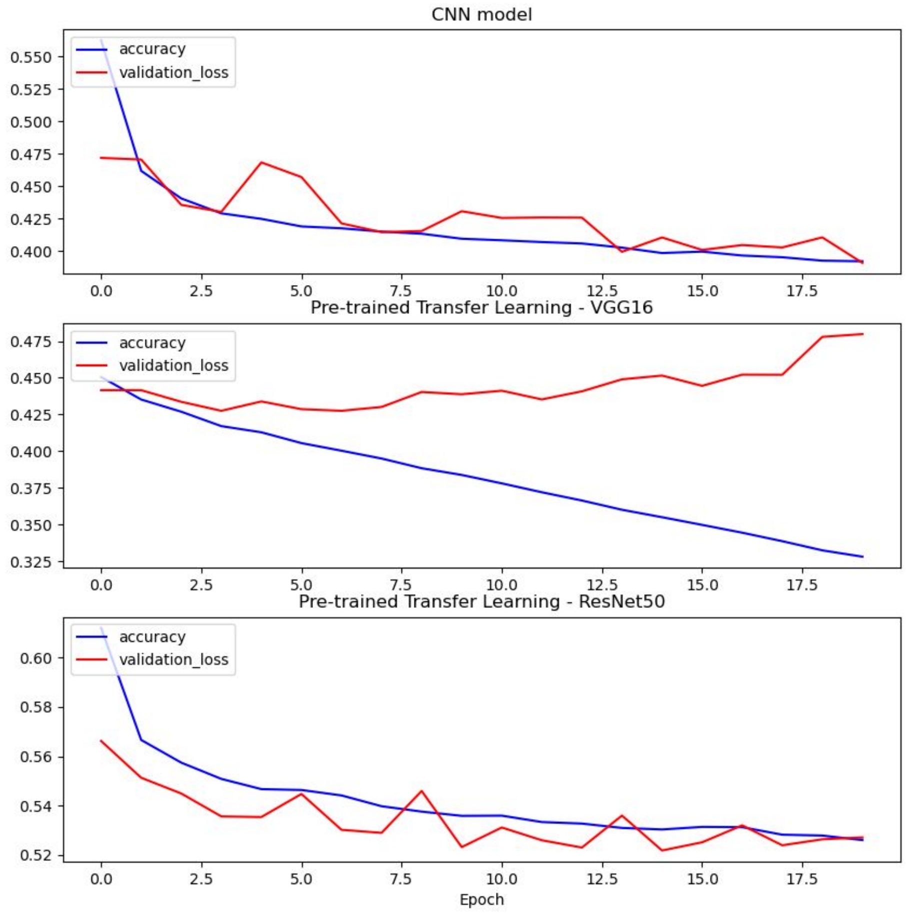
Accuracy


CNN model

Pre-trained Transfer Learning - VGG16

Pre-trained Transfer Learning - ResNet50

Loss


CNN model

Pre-trained Transfer Learning - VGG16

Pre-trained Transfer Learning - ResNet50

Model Performance

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

CONCLUSION

Model Exploration:

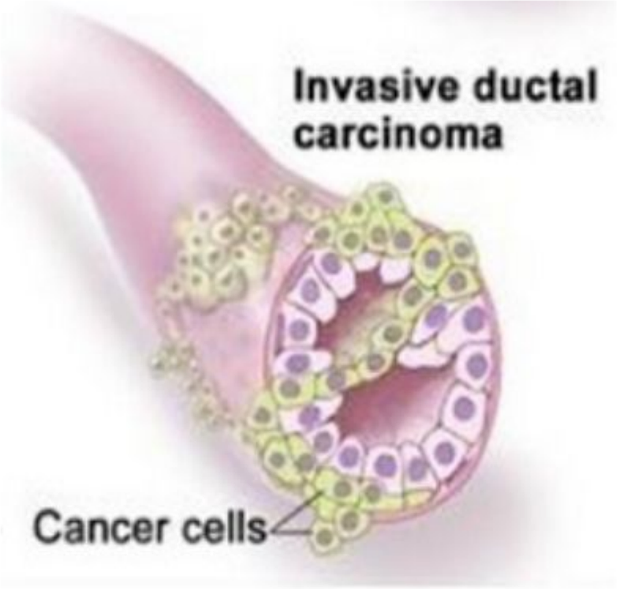
Three models were explored for image classification:

1. A custom built CNN
2. VGG16 for transfer learning
3. ResNet50 for transfer learning.

Performance Overview:

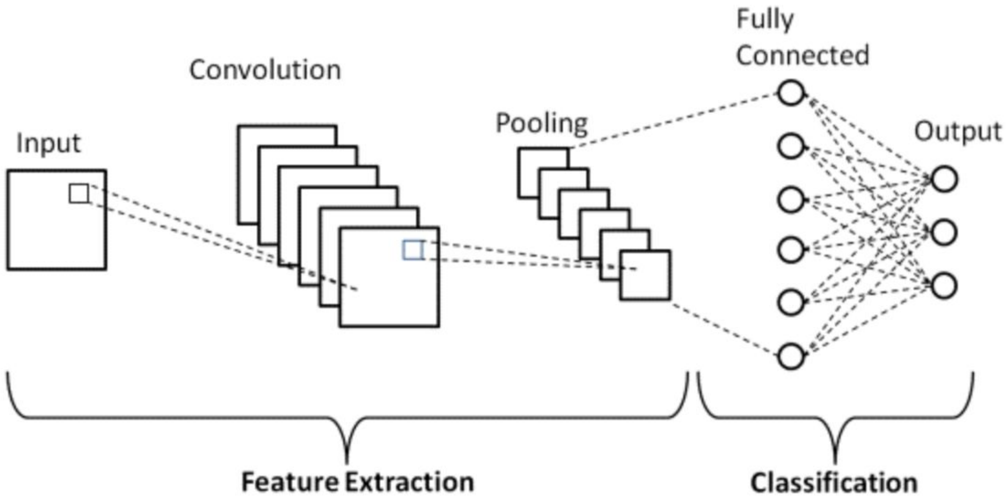
CNN Model performed the best.

- Achieved an AUC score of 0.915.



Future Work

Future endeavors will involve harnessing cloud services to increase 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=Cj0KCQiAhomtBhDgARIsABcaYylH3OZmhgmy4s9zyUghSblIYPOGdo97UYLI_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