Behind skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. Every year, it accounts for nearly a third of all new cancer cases in U.S. women (and 12.5% of cancer cases overall). It is second to only lung cancer in the number of cancer related mortalities (with an estimated 43,700 deaths in 2023), and worldwide, female breast cancer is the fifth leading cause of death. In the U.S. approximately 1 in 8 women will develop invasive breast cancer during their life, and early detection of the disease is critical to positive treatment outcomes. In fact, better detection is one of the main reasons that breast cancer deaths have declined by 43% since 2020.

Early detection is so critical that according to the American Cancer Society, the five year survival rate for breast cancer detected in the localized SEER stage is 99%. That number drops to 86% when detected in the regional stage and just 31% when detected in the distant stage. Early detection also leads to less radical treatments, which can prolong the life of the person diagnosed, prevent some of the harmful side effects of cancer drugs, and lower treatment costs.

For women, having a mammogram performed regularly is one of the best ways to catch breast cancer in its earliest stages. A mammogram is a chest X-Ray that can detect tumors and calcifications before symptoms appear or you can even feel a mass. Mammograms are hugely beneficial tools and can detect abnormalities up to two years before a lump is large enough to feel. However, skillful interpretation of mammograms is necessary for them to be effective and for doctors to be able to localize and diagnose these types of abnormalities.

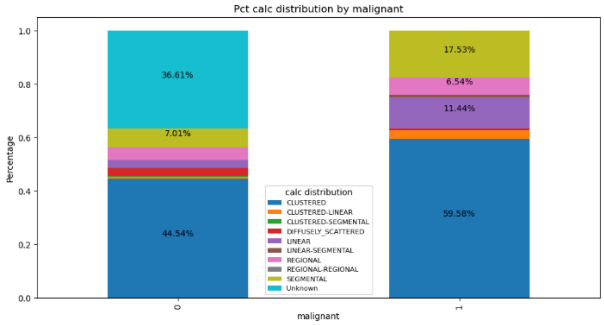
Current methods of interpreting mammograms are not perfect, and abnormalities can go undetected. On the other side, mammograms are notorious for their high false positive rate (about half of women getting annual screenings will experience a false positive over a ten year period), which often require extra testing and in some cases a biopsy. This is both costly (time and money) and can cause anxiety and physical discomfort for patients. With AI and data changing so many other industries, healthcare is no different, and doctors and leaders in the medical community are looking to machine learning models as a possible alternative to their current ways of doing things. The goal of this project was to use a convolutional neural network to detect malignant abnormalities in mammograms with a sensitivity of 87.5% (up to 1 in 8 cancers go undetected using current methods of interpreting the tests). Sensitivity measures how many of the malignant abnormalities are detected relative to the number of malignant abnormalities. I also wanted to avoid false positives, however, so I took the specificity of the model (number of benign mammograms detected relative to the number of benign mammograms) into consideration, as well, when picking a final model.

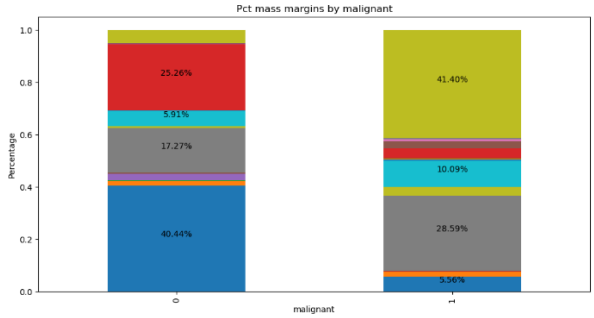
To do this, I used the CBIS-DDSM Dataset, an updated and standardized version of the National Cancer Institute's Digital Database for Screening Mammography (DDSM) dataset, which consists of 2,857 distinct full mammogram images and 3,566 distinct cropped mammogram images (of which 3,463 have associated region of interest (ROI) mask file). The data encompasses normal, benign, and malignant cases with verified pathology information from four hospitals (Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University of St Louis School of Medicine). It also has associated metadata files with information on the following:

- Abnormality Type: Mass or Calcification
- Laterality: Side mammogram was performed on
- Patient Orientation: MLO or CC

- Breast density: Category of breast density
- Abnormality Type: Category of abnormality
- Calcification Type (calcifications): Type of calcification
- Calcification Distribution (calcifications): Distribution of the calcifications
- Mass shape (masses): Shape of the mass
- Mass margins (masses): Feature that separates the mass from the adjacent breast
- Assessment: Assessment category
- Pathology: Malignant, benign, or benign without callback
- Subtlety: Subtlety category

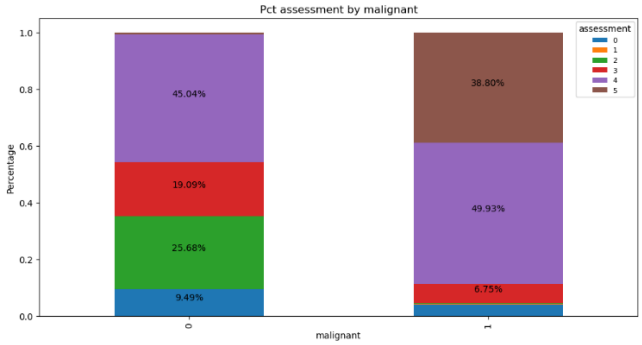
Since the goal of this project was to predict malignant cases, the pathology variable was ultimately categorized as either “Benign” or “Malignant”, and performance metrics grouped the “Benign” and “Benign Without Callback” labels together. Exploratory data analysis was conducted in this way, and differences in the distributions between the “Benign” and “Malignant” cases were observed. Most notably, these were for the Calcification Distribution, Mass Margins, and Assessment variables. For Calcification Distribution, many of the benign observations were missing data (which was imputed with a value of “Unknown”). This was not the case for the malignant cases, which led me to believe it was intentionally left blank. For the Mass Margins variable, the distribution is markedly different between the two groups, and for the Assessment variable, a large percentage of benign cases have an Assessment of 2, whereas the malignant cases have a large percentage of 5’s (while the other class has hardly any of these).





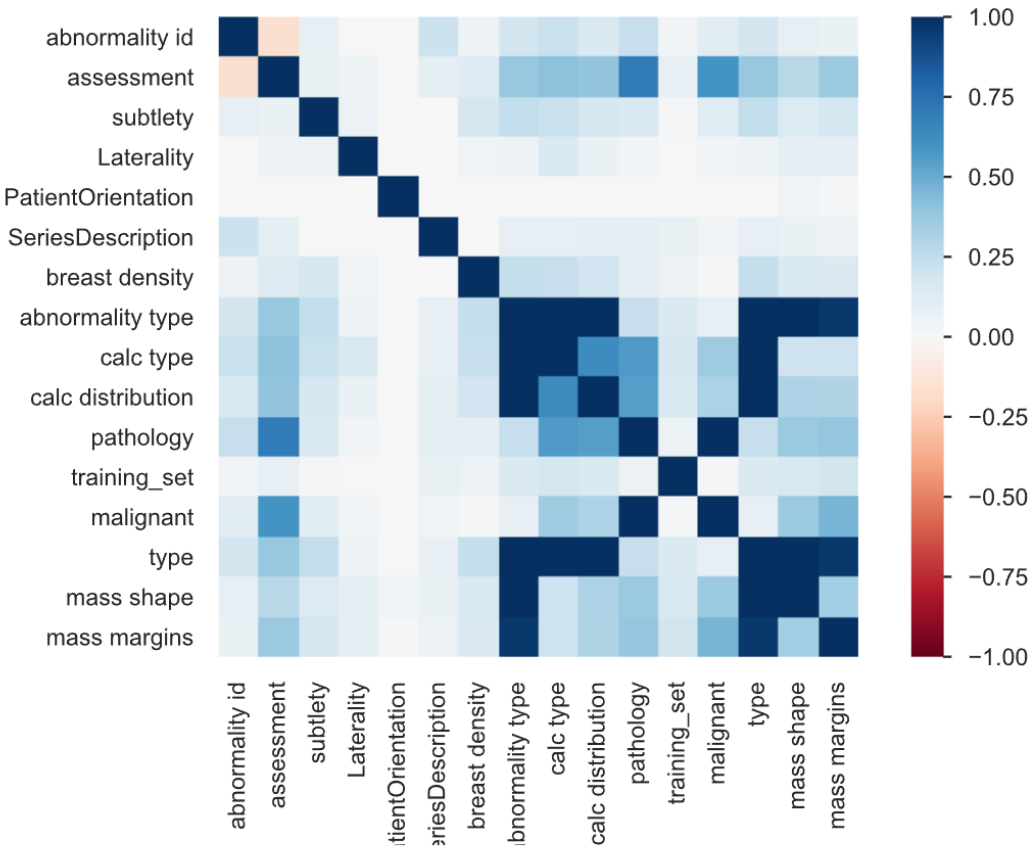
1. Calcification Distribution by Malignant (left)

2. Mass Margins Distribution by Malignant (right)



3. Assessment Distribution by Malignant

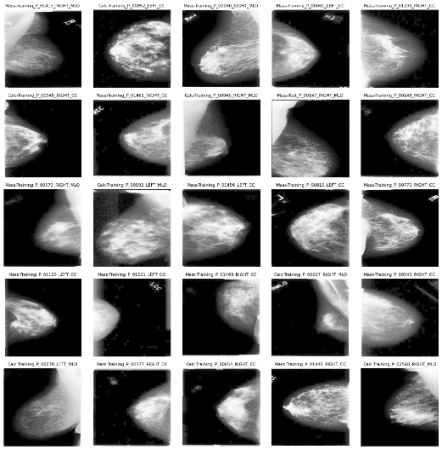
The heatmap confirmed these observations, with Calcification Type, Calcification Distribution, Mass Shape, Mass Margins, and Assessment being the predictive features most correlated with Malignant.

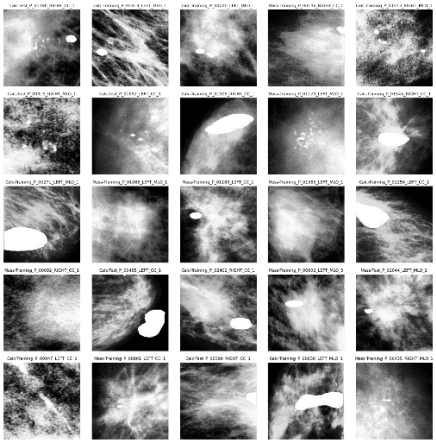


4. Heatmap of metadata features

Prior to training the models, the images needed to be preprocessed. First, the ROI mask images were overlayed over the cropped images in cases where they existed. Images were then resized to a height of 224 pixels and a width of 224 pixels with three color channels (the models I used required three color channels even though the images were in black and white). Contrast was then increased in

the images using histogram equalization. Samples of each of the output image types is shown below:





5. Malignant Cases - Full Mammograms (left)

6. Malignant Cases - Cropped Images w/ ROI Overlay (right)

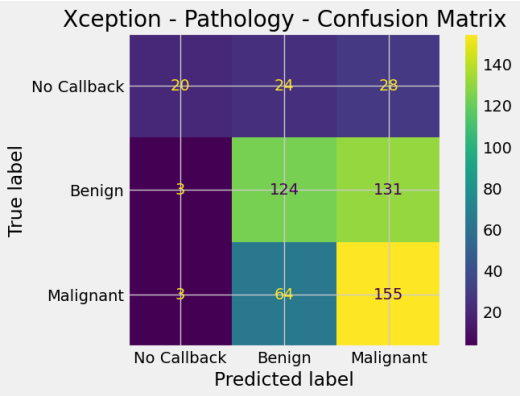
The data was already split into training (80%) and testing sets (20%) whose metadata was approximately matched; however, the testing set was randomly split in half to form a validation set for training. This left three datasets: training (80%), validation (10%), and testing (10%) that were used for the modeling portion of the project.

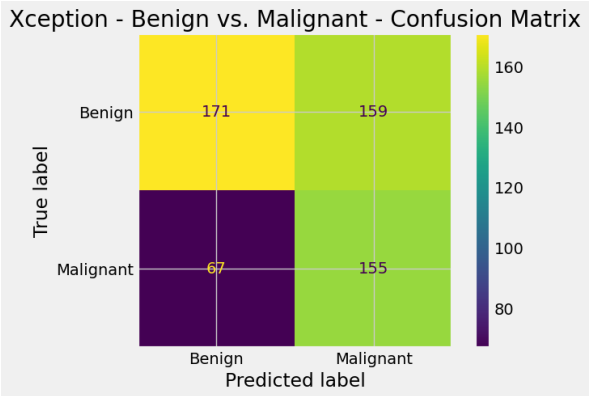
In the modeling portion of the project, three different models were evaluated and tested before a final model was selected. Further, each was trained on two sets of classes (Malignant vs. Benign) and (Malignant vs. Benign vs. Benign Without Callback). The latter of these were then assigned a label (Malignant vs. Benign) based upon the classification that the model gave (Malignant, if Malignant, otherwise Benign).

All three models were pretrained models, and transfer learning was used to classify the images. The three models were DenseNet169, Xception, and EfficientNetV2S and were chosen given their relatively small size vs. accuracy on image classification tasks. All three models used Adam as the optimizer, Categorical Crossentropy as the loss function, a learning rate of .001, and weights from imagenet (a large visual database designed for use in visual object recognition software research). Training was done using a batch size of 32 and epochs were set to 30; however, early stopping prevented any of the models from using more than nine epochs. The images were passed through the DenseNet preprocessing algorithm prior to training, and images were augmented by applying random contrast, rotation, and zoom adjustments prior to training. For the pretrained models, the last layer was generally left trainable, the top layer dropped, and an average pooling layer then applied. A dropout layer with a rate of 0.2 was then applied followed by a two or three unit Dense layer with softmax activation for classifying the images.

The model with the highest sensitivity on the testing set that also had a higher specificity than the Dummy Classifier that randomly assigned labels to the observations (sensitivity = 0.50; specificity = 0.51) was chosen as the best model. This was the Xception model trained on the pathology labels (sensitivity = 0.70; specificity = 0.52). The Xception model trained on the malignant labels nearly had a sensitivity of 87.5% (85%); however, its specificity was only 0.29, so the other Xception model was deemed the better model. The Xception model trained on the pathology labels also had the highest F1

score of any of the models and tied for the highest accuracy. Most of the errors were made mixing up benign and malignant cases, and the model seemed to group the images into these two categories naturally.

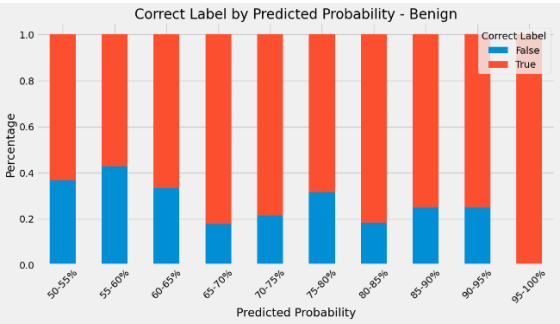


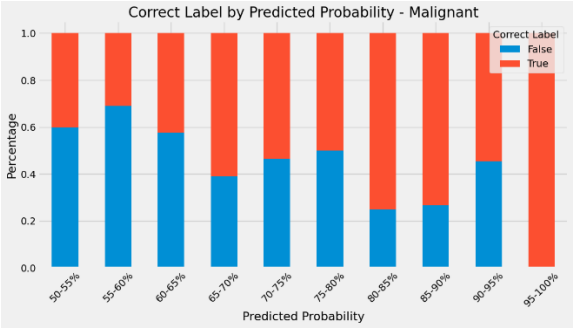


7. Xception - Pathology Model (Pathology Confusion Matrix) (left)

8. Xception - Pathology Model (Malignant Confusion Matrix) (right)

As expected, the model performed better the more confident it was in a prediction. Just as it's harder for someone reading a mammogram to identify fringe cases, it's also harder for the ML model. It seems like there are clear cases when an image is benign and malignant and then cases where it is uncertain. Since this model did not perform well enough to outperform traditional ways of interpreting mammograms, perhaps it could be used to reduce the number that need to be reviewed. Clear cases could be identified by the model, and it could flag ones that needed to be reviewed by a mammographer.





9. Benign Cases - Correct Label by Predicted Probability of Class (left)

10. Malignant Cases - Correct Label by Predicted Probability of Class (right)

Overall, the model was doing a fairly good job at predicting whether a mammogram image contained a malignant abnormality; however, it did not meet the initial goal of achieving over 87.5% sensitivity. Although one of the models approached the 87.5% goal, it was not specific enough in its predictions and had too many false positives. For the problem at hand, this is problematic because false positives require extra testing and in some cases a biopsy. This is both costly to the hospital (time and money) and to the patient, who could experience unnecessary levels of anxiety from the result.

While the final model (Xception trained on pathology labels) had a sensitivity of 0.70, this is not good enough to be used in a clinical setting. Further, it also had a high false positive rate which only narrowly outperformed the Dummy Classifier. The model was very accurate, however, when it was very confident in a prediction, so this model could potentially be used to reduce the number of mammograms that have to be manually reviewed. This could save mammographers time and be another tool at their disposal for diagnosing cancer.

Future research should train the model using a higher number of images (if available) and should revisit the hyperparameters used in this project. Other CNNs might want to be tested, as well, to see if they perform better than the ones tested here. By improving upon these results and increasing the ability to detect breast cancer in its early stages, millions of dollars and thousands of lives can be saved. This project attempted to do that and future projects can expand on its work to make further improvements. Early detection is the key to successful outcomes, and CNNs offer a promising way to get there.

References

Breast cancer - statistics. Cancer.Net. (2023, February 23). https://www.cancer.net/cancer-types/breast-cancer/statistics.

Breast cancer early detection. National Breast Cancer Foundation.. https://www.nationalbreastcancer.org/early-detection-of-breast-cancer/.

Breast Cancer Facts and Statistics. Breast cancer facts and statistics 2024. https://www.breastcancer.org/facts-statistics.

CBIS-DDSM. The Cancer Imaging Archive (TCIA). https://www.cancerimagingarchive.net/collection/cbis-ddsm/.

Limitations of mammograms: How accurate are mammograms?. How Accurate Are Mammograms? | American Cancer Society. https://www.cancer.org/cancer/types/ breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html.

Mammograms: Why early detection matters. Mammograms: Why Early Detection Matters | Houston Methodist. https://www.houstonmethodist.org/blog/articles/2019/sep/mammo grams-why-early-detection-matters/.

Our history. American Cancer Society. https://www.cancer.org/about-us/who-we-are/our-history.html.