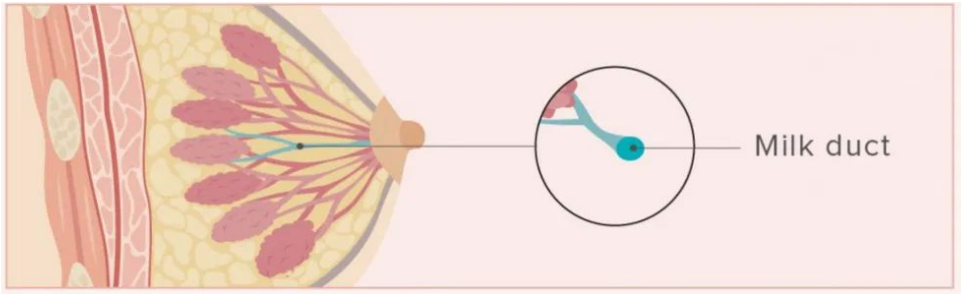
Breast Cancer Detection

Alexis Babb, Taizhou Qing, Ashish Roy, Zhiyu Zhou

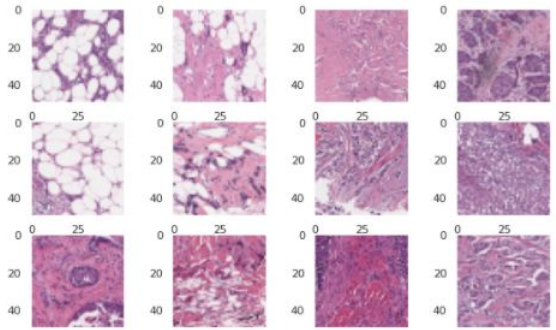
Problem - Breast Cancer Diagnosis is Time Consuming & Painful

Invasive ductal carcinoma (IDC) is - with ~ 80 % of cases - is one of the most common types of breast cancer. It can be very malicious by forming metastasis. Diagnosis requires a biopsy with pathological analysis, which is invasive and requires significant resources (time, equipment, etc).



Project Goal - Build a Tool that Can Predict If a Lump Is Cancerous

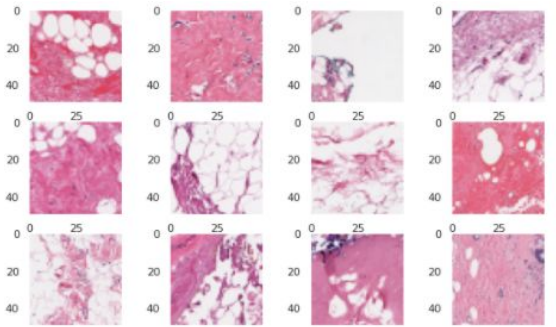
For this project, we built a tool - specifically, a deep learning model - that can predict if a lump is cancerous or not based on images. While there are many variations, cancerous patches generally tend to look more violet and crowded than healthy ones. Data is sourced from Kaggle.



← Cancerous (Left)

vs.

Healthy (Right) →



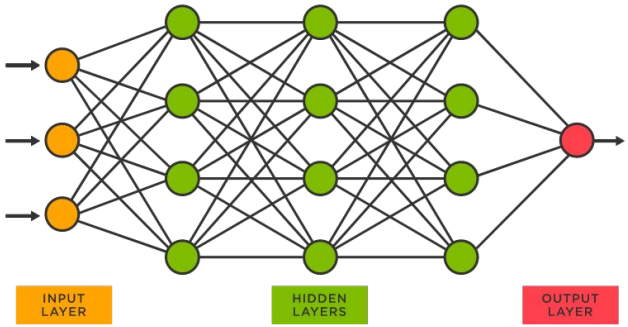
Solution Overview - Optimize & Evaluate 3 Deep Learning Models

We used 3 steps to create our model:

1. Exploration & Pre-Processing

2. Model Creation & Tuning
   a. 3 models - CNN, RNN, ANN
   b. Tuned Hyperparameters (Batch Size, Epochs)

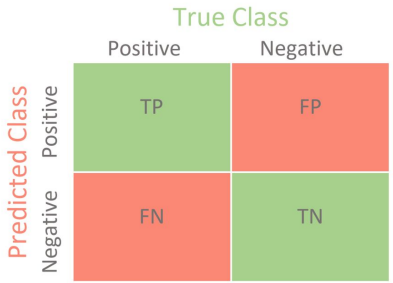
3. Model Evaluation
   a. Confusion Matrix
   b. AUC-ROC

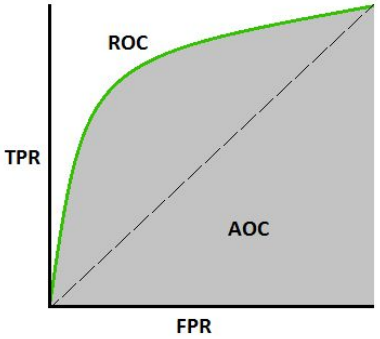


Evaluation Details - Confusion Matrix & AUC-ROC

The confusion matrix was selected so we could see how many predictions were correct and how many were incorrect (both false pos/neg).

The ROC curve was selected because it allows for visualization of model performance at different threshold values. The AUC was selected because it aggregates performance into a single, easy-to-analyze value (from 0 to 1, with 1 being a perfect fit).



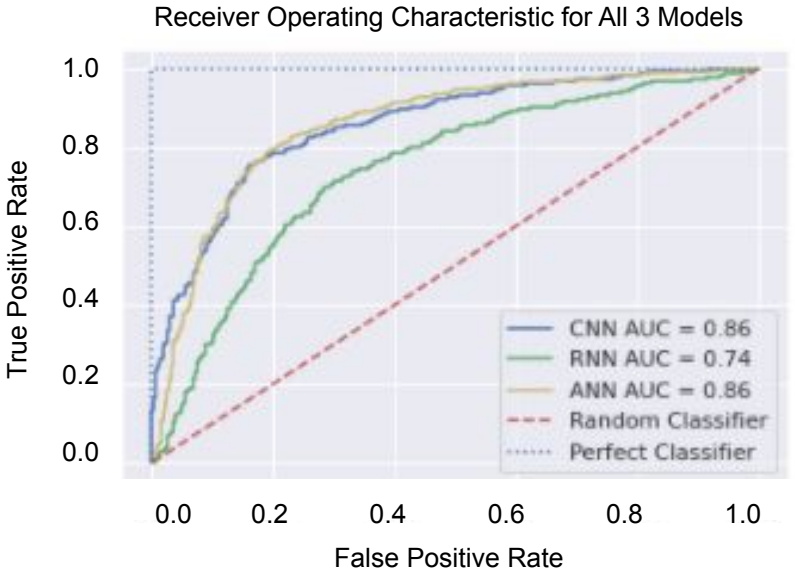


Solution Development- Leverage Keras and Sci-Kit Learn Packages

The Keras package (from Tensorflow) was used to transform the images into arrays for processing and build the layers of the models. KerasClassifier wrappers were used so GridSearchCV (from sklearn) could be leveraged for hyperparameter tuning of each model. Sklearn was also leveraged for building the AUC-ROC plots. Supporting packages for data manipulation and plotting include numpy, pandas, and matplotlib.

Conclusion - ANN is the Best Model for the Breast Cancer Prediction

The AUC for both CNN and ANN were found to be the best at ~0.86*, indicating that the models have similar performance. Also, the training and testing ROC curves for each model are similar, indicating that the models are well-fit (not over-fit). However, the ANN model requires significantly less run-time. As breast cancer diagnoses are time-sensitive, **we recommend the ANN model**.



(*Values may vary slightly from run to run due to randomness)