RSNA Screening Mammography Breast Cancer Detection

Find Breast Cancers in Screening Mammograms

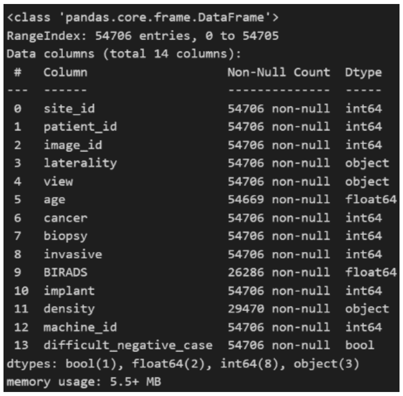
Framing the Problem

The goal of this project is to better identify breast cancer through regular early detection screenings. The model will assist radiologists with both the efficiency and accuracy in patient diagnosis and ultimately improve patient care.

The data for this model was provided by Kaggle and the Radiological Society of North America.

Collecting and Cleaning the Data

The data consisted of over 54000 mammogram dicom images from over 15,000 patients. The images were sorted in directories by patient. There was also a csv file with entries corresponding to the context of the pictures.

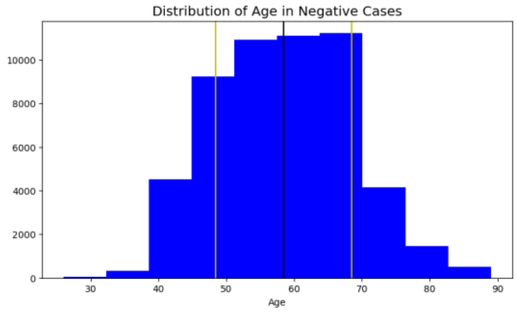


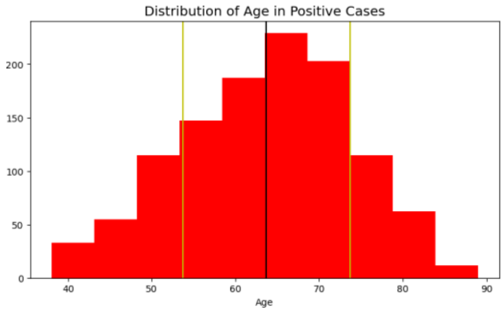
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54706 entries, 0 to 54705
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   site_id                  54706 non-null  int64
 1   patient_id               54706 non-null  int64
 2   image_id                 54706 non-null  int64
 3   laterality               54706 non-null  object
 4   view                     54706 non-null  object
 5   age                      54669 non-null  float64
 6   cancer                   54706 non-null  int64
 7   biopsy                   54706 non-null  int64
 8   invasive                 54706 non-null  int64
 9   BIRADS                   26286 non-null  float64
 10  implant                  54706 non-null  int64
 11  density                  29470 non-null  object
 12  machine_id               54706 non-null  int64
 13  difficult_negative_case  54706 non-null  bool
dtypes: bool(1), float64(2), int64(8), object(3)
memory usage: 5.5+ MB
```

The different columns of the pandas data frame guided the picture preprocessing and modification. As there was no column for the direct filename of the picture file within the patient directories, I constructed a column to point to where on the hard drive those files were located.

Additionally, the pictures were taken of both the left and the right, thereby introducing vertical symmetries that needed to be dealt with. These will all be analyzed and addressed during the EDA section of this documentation.

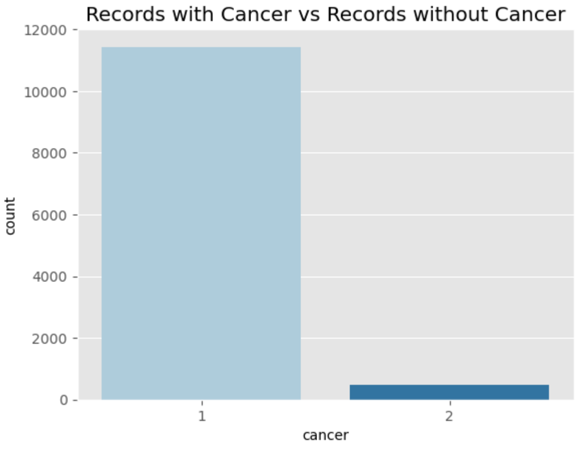
Exploratory Data Analysis



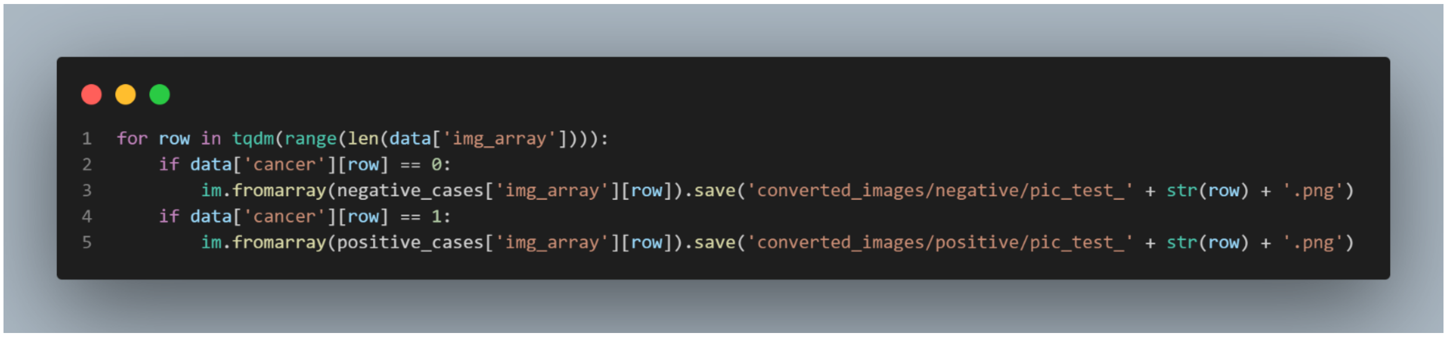


During exploratory analysis, some interesting trends were discovered. The distributions above indicate a 5-year difference between the median age of our cancer-positive population and our cancer-negative population. This indicates that age seems to play a part in the occurrence of breast cancer.

Upon exploration, it was discovered that the data is very imbalanced. While there were almost 12,000 negative cases to examine, there are only 486 positive cases to base our model off of. This imbalanced data will need to be dealt with through either oversampling the minority (cancer-positive cases) or under-sampling the majority (cancer-negative cases). This will need to be done prior to the construction of the model.

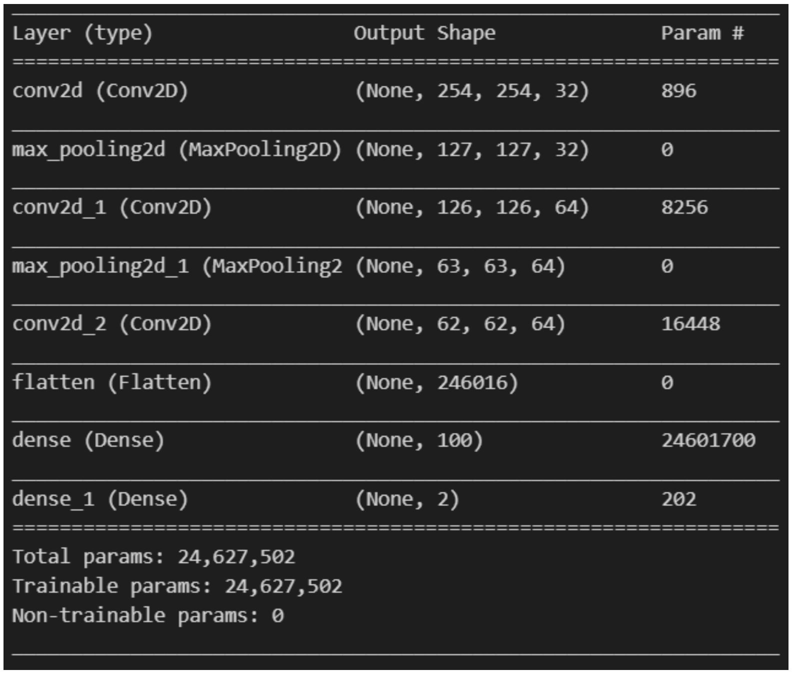


After examining the csv file through pandas and other python libraries, the problem of dicom analysis and conversion became apparent. Using the dicomdl library as well as a few other image decoders, I was able to convert the medical dicom images to a numpy array corresponding to a 256 by 256 pixel grayscale image. Reencoding the images to png files, I sorted them by binary label for processing through Keras.



```
for row in tqdm(range(len(data['img_array']))):
    if data['cancer'][row] == 0:
        im.fromarray(negative_cases['img_array'][row]).save('converted_images/negative/pic_test_' + str(row) + '.png')
    if data['cancer'][row] == 1:
        im.fromarray(positive_cases['img_array'][row]).save('converted_images/positive/pic_test_' + str(row) + '.png')
```

Model Building and Deployment



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 254, 254, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 127, 127, 32)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 126, 126, 64)      8256
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 63, 63, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 62, 62, 64)        16448
_________________________________________________________________
flatten (Flatten)            (None, 246016)            0
_________________________________________________________________
dense (Dense)                (None, 100)               24601700
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 202
=================================================================
Total params: 24,627,502
Trainable params: 24,627,502
Non-trainable params: 0
_________________________________________________________________
```

In building a model, I decided to use a Convoluted Neural Network. This perceptron network will analyze the individual pixels of the image and determine important features in order to determine which visual properties correlate strongly with the occurrence of cancer.

In the model, I used the Rectified Linear Activation Function (ReLU) for all the hidden layers except the output. For t he output layer, I used the Softmax activation function in order to get a solidified classification determination. The loss function I used was the Categorical Cross Entropy Loss Function in order to compute the cross-entropy loss between true labels and predicted labels. I ran it through a maximum of 20 epochs until the model achieved a minimum loss function for both the training and validation sets. The model was then saved and reran for functional verification.

Results

The results of this exercise indicate that, with a more balanced data-set, CNNs can approach the problem of breast cancer diagnostics in a faster and more efficient way. This can assist radiologists in the determination of whether cancer is present and where it is based on the features our model has presented. This model can be further tuned to detect smaller and less detectable cancers in the future.