Machine Learning Approaches for Breast Cancer Diagnosis

Fall 2021 Cohort | Team #6

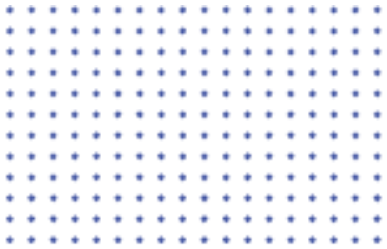
Team Members: Olohireme Ajayi, Sarah Darson, Weimin Liu, Halima Moncrieffe, Vertulie Pierre-Louis, Prastuti Singh

Acknowledgement

We appreciate the entire Data Science For All/Women team for this programme.
Special Thanks to our Mentor Elia and our Teaching Assistants Savannah and Paulene for their guidance and direction.

TABLE OF CONTENTS

INTRODUCTION



Problem Overview

In 2020, breast cancer accounted for 2.3 million positive diagnoses and took the lives of 685,000 women worldwide, becoming the most common cancer type in women both in the developed and less-developed world (BCRF, 2021). Cancer affects all demographics, however there are risk factor disparities that have an impact on healthcare outcomes. Ignoring these disparities when developing screening procedures and preventative measures can be fatal. There is a great need to adjust predictive algorithms to account for risk factors when assessing individual patients. Errors in patient care are caused by system or process failures, which may not take into account other risk factors when creating a patient care plan. It is important to adopt various process-improvement techniques to identify inefficiencies, ineffective care, and preventable errors to then influence changes associated with systems (Hughes, 2008).

Specific Issue

The risk factors for breast cancer survival are complex. In this study, we will evaluate the risk factors that contribute to 1 year survival. Once patients are identified at higher risk of breast cancer, accurate screening and early detection is important to save lives. An automated, accurate, and rapid screening tool to detect the presence of cancer with an easy-to-understand report to understand the findings would help providers and patients have confidence in their screening results.

Problem Importance

Breast cancer treatment accounts for 13% of all cancer treatment costs in the United States (CDC, 2021). Breast cancers that are treated at earlier stages have a lower cost (CDC, 2021). In the first 12 months after diagnosis, the average cost per patient was $60,637 when diagnosed with stage 0 breast cancer, compared to average costs of $134,682 when diagnosed at stage 4 (Blumen et al., 2016). Some healthcare systems may want to improve their big data tools but lack the necessary training or information staff members.

The medical industry has transformed over the last 50 years in terms of cancer outcomes. In 1970, of the total population diagnosed with cancer in the United States, approximately half would have been alive five years later. For those diagnosed in 2009, the figure was closer to 70 percent (Albrecht et al., 2021). Early and accurate diagnosis is a key tool in improving patient survival outcomes. Predictive models and classifiers can aid medical professionals in improving patient outcomes and forward progress.

Mammograms are an important way to screen patients for breast cancer; their accuracy can drop to 50% for women with dense breast tissue. A previous study estimated that 40% of women have dense breast tissue, which makes it hard to find tumours by mammogram alone (Rochman, 2015). In those circumstances, preventative ultrasound monitoring and histopathology are important tools for breast cancer diagnosis.

Ultrasound is non-invasive, and histopathology evaluates patient cells from a tissue biopsy. If we can understand the impact of breast cancer risk factors, medical help can be escalated for patients at greater risk.

Project Description

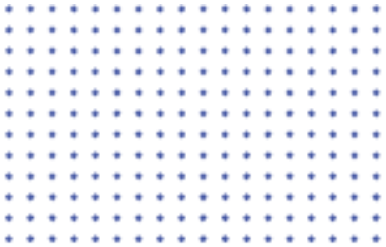
The project will address two issues:

- Analysing patient factors that may increase risk of breast cancer and predicting cancer risk
- Investigation of machine learning tools for diagnosing breast cancer

Through analysis of multiple risk factors, we will define the relative risk of key demographics on breast cancer risk. Patients with dense breast tissue are likely to need another test apart from mammograms. Understanding the factors that are associated with dense breasts will be investigated. More effectively identifying patients at risk and therefore improving screening accuracy is expected to lead to improved 1-year patient survival rates of breast cancer, which can have a significant effect on both patient outcome and overall costs.

In addition to the risk factor model, we examined the role that machine learning can play in diagnosing breast cancer using breast tissue microscopy data (histopathology) and breast ultrasound scans. The image classifier would serve as an analysis tool to aid trained radiographers, in addition to their clinical expertise, to analyse and form a more accurate diagnosis (American Cancer Society, 2019).

The long-term goal of an accurate risk factor analysis and image classifier is to provide resources that seamlessly integrate into varied medical health systems and provide both clinicians and patients with easy-to-understand results and improve screening methods for breast cancer. Through integration of machine learning and classifier tools, patient outreach providers, community groups, and payers can have renewed confidence in screening as a life-saving tool.



Methods

Datasets

The project includes three datasets:

- **Risk Factors Estimation Dataset** consists of risk factor data (e.g. age, race, age at first birth, etc.) from 280,660 screening mammograms from women in the Breast Cancer Surveillance Consortium. (Barlow et al., 2006). The dataset is composed of women that did not have a previous diagnosis of breast cancer and did not have any breast imaging in the nine months preceding the index screening mammogram. However, all women had undergone previous breast mammography in the prior five years (though not in the last nine months). Cancer registry and pathology data were linked to the mammography data and incident breast cancer (invasive or ductal carcinoma in situ) within one year following the index screening mammogram was assessed. The data have been aggregated by the cross-classification of risk factors and outcome, with a count indicating the frequency of each combination. The project aims to investigate risk factors, key demographic and interventional factors contributing to breast cancer diagnosis.
- **Breast Histopathology Dataset** consists of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. From that, 277,524 patches of size 50 pixels x 50 pixels were extracted (198,738 IDC negative and 78,786 IDC positive). Our target values are categorized as 'non-cancerous' and 'cancerous'. The project aims to predict Invasive Ductal Carcinoma (IDC) in breast histology images using a Convolutional Neural Network (CNN). (Janowczyk & Madabhushi, 2016).
- **Breast Ultrasound Dataset** consists of breast ultrasound images from 600 female patients between the ages of 25 and 75 years old. The dataset consists of 780 images with an average image size of 500 pixels x 500 pixels and each ultrasound is classified as 'normal', 'benign', or 'malignant'. The project aims to use a machine learning model to classify and detect early signs of masses or microcalcification in breasts from ultrasounds. (Al-Dhabyani, Gomaa, Khaled, & Fahmy, 2020).

Data Pre-processing and Cleaning Methodology

Risk Factors Estimation

For the data pre-processing, we removed the training feature because it did not represent demographic information, it was used to code if the data was used by the dataset authors for training or testing. We converted the cancer feature to the target, since it represented the presence of ductal carcinoma. We removed the invasive column because of the high correlation with cancer ~ 0.88; 'invasive' referred to the cancer spreading, so including this would result in data leakage. There were many unknown features within the risk factors estimation dataset outlined below:

Features	Percentage unknown
menopause	6.3
density	25
race	21
hispanic	29
bmi	40
age of first birth	35
number of 1st degree relatives	14
prev breast procedure	12
last mammogram result	38
surgical menopause	46
hormone therapy	35

- 0.4% of the data had > 9 unknown relevant features. These records were removed.
- Removing all records with at least one unknown relevant feature would result in loss of over 70% of the dataset so K-Nearest Neighbours Imputation was used to impute the missing data
- One-hot encoding on the race column because it has more than two classes

Breast Histopathology Dataset

- Import Library: We started by importing the pandas library
- Load Image: To load the image data, we import glob, part of the Python standard library.
- Visualize The Data: We import the Matplotlib library for visualizing our dataset and determining the number of images with or without cancer. Subsequently, we concatenate the images (image_cancer, non_image_cancer) into one dataset.
- Data Preprocessing: We reshaped the pixel values (between 0 and 255) to the [0, 1] interval of our dataset inputs and converted the labels into one-hot encoding binary vectors.
- Test Train Split: Using train_test_split from sklearn.model_selection, we split the data into the test set and training set.
- To build the model, we Import the Keras libraries:
    - from Keras.models import Sequential
    - from Keras.layers import layers-Conv2D, MaxPooling2D, Dropout, Flatten, Dense

Breast Ultrasound Dataset

The dataset consists of ultrasound images that were cropped to remove unimportant boundaries. Each ultrasound image includes a masked, segmented version (performed in MATLAB) that clearly outlines the area of interest and the original ultrasound. For project purposes, masked images are not included. All images were reshaped to 128 x 128 pixels.

Exploratory Data Analysis

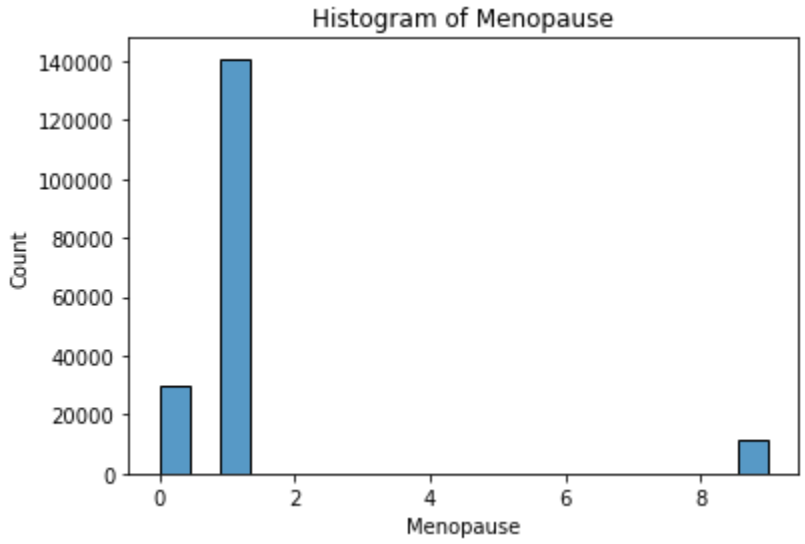
Risk estimation dataset

The risk estimation dataset includes 2,392,998 screening mammograms (called the "index mammogram") from women included in the Breast Cancer (Surveillance Consortium (BCSC). Based on data aggregation, the results were refined down to 280,660 screening mammograms The examined risk factors were quantified and measured with the corresponding values listed in the table below:

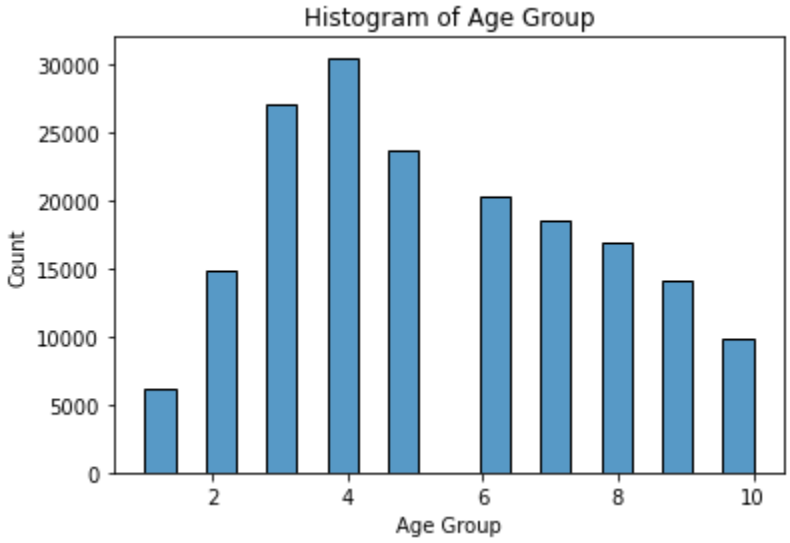
Full Name	Coding
Menopause	0 = premenopausal; 1 = postmenopausal or age>=55 ; 9 = unknown
Age Group	1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84
Breast Density	BI-RADS breast density codes 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense; 9 = Unknown or different measurement system
Race	1 = white; 2 = Asian/Pacific Islander; 3 = Black; 4 = Native American; 5 = other/mixed; 9 = unknown
Hispanic	0 = no; 1 = yes; 9 = unknown
Body Mass Index (BMI)	Body mass index: 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = unknown
Age at first birth	Age at first birth: 0 = Age < 30; 1 = Age 30 or greater; 2 = Nulliparous/no child; 9 = unknown
Number of first degree	Number of first degree relatives with breast cancer: 0 = zero; 1= one;

relatives with breast cancer	2 = 2 or more; 9 = unknown
Breast procedure	Previous breast procedure: 0 = no; 1 = yes; 9 = unknown
Last Mammogram	Result of last mammogram before the index mammogram: 0 = negative; 1 = false positive; 9 = unknown
Surgical menopause	Surgical menopause: 0 = natural; 1 = surgical; 9 = unknown or not menopausal (menopaus=0 or menopaus=9)
Current Hormone Therapy	Current hormone therapy: 0 = no; 1 = yes; 9 = unknown or not menopausal (menopaus=0 or menopaus=9)

The figures below show histograms of risk factors from the Risk Factor Estimation Dataset (Barlow et al., 2006).

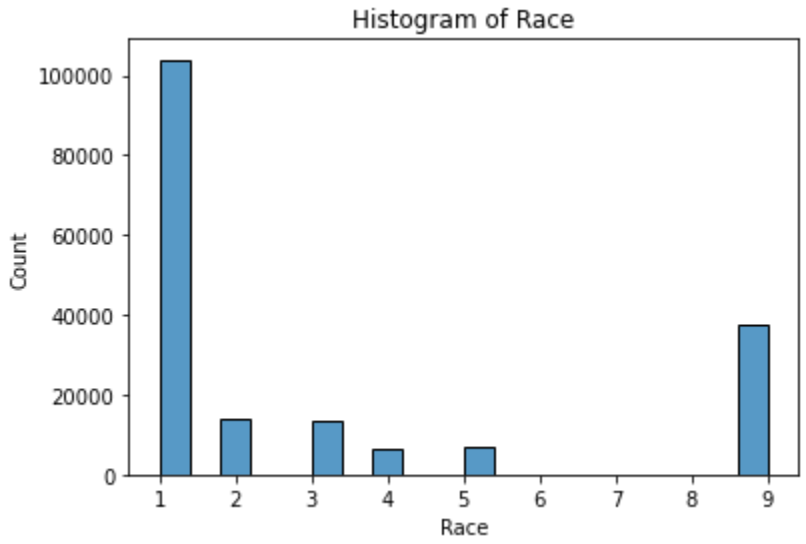


80.3% of the population in this dataset are post-menopausal, represented by the value (1). The patients that were pre-menopausal represented 14.1% of the sampled population and the remaining 5.6% were unknown.

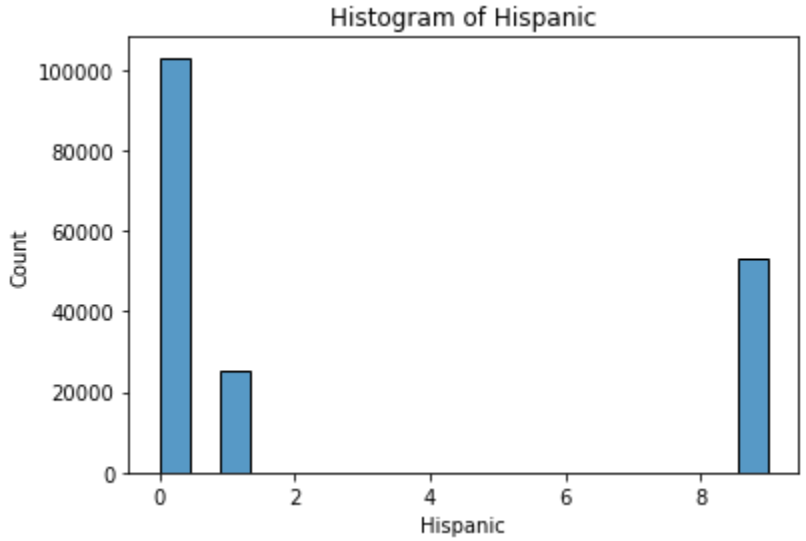


44.1% of the dataset population falls between the ages of 45 and 59, represented by the values (3, 4, 5). The population is right skewed with respect to age.

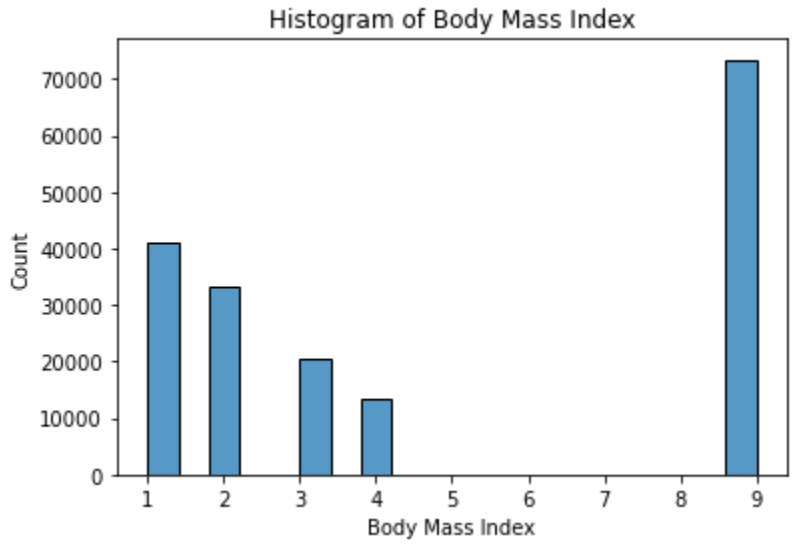
54% of the common breast densities observed are in the scattered fibroglandular densities (2) and heterogeneously dense (3), which are both a mix of fat and dense breast tissue. 25.2% of the population also has unknown breast density or values were noted in a different measurement system not compatible with the defined categories.



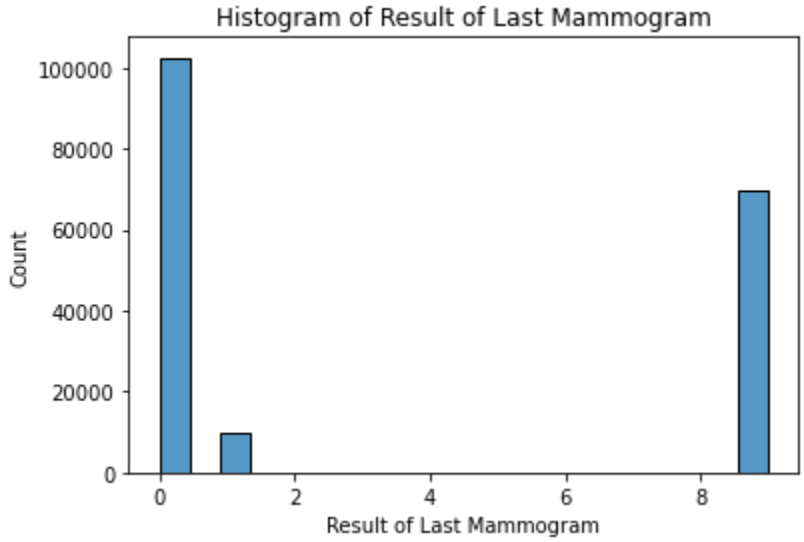
The dataset population consists of the following: 155,345 white (1), 23,013 Asian/Pacific Islander (2), 64,404 Black (3), 10,645 (4), 62,125 Native American (5), and 57,764 unknown race (9). 55.3% of the population are represented by white patients, with the second-highest proportion, 20.6%, identifying as unknown.



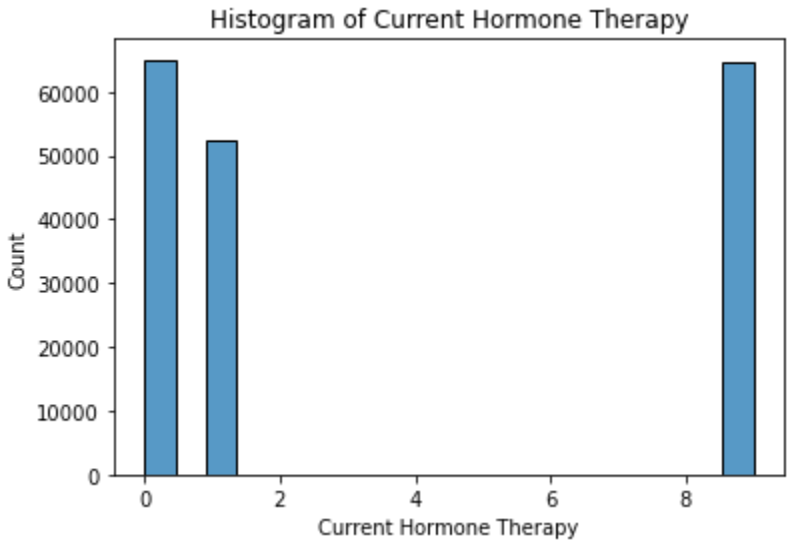
With respect to Hispanic origin, 154,832 of the population identified as not Hispanic (0), 40,370 identified as Hispanic (1), and 85,458 identified as unknown (9).



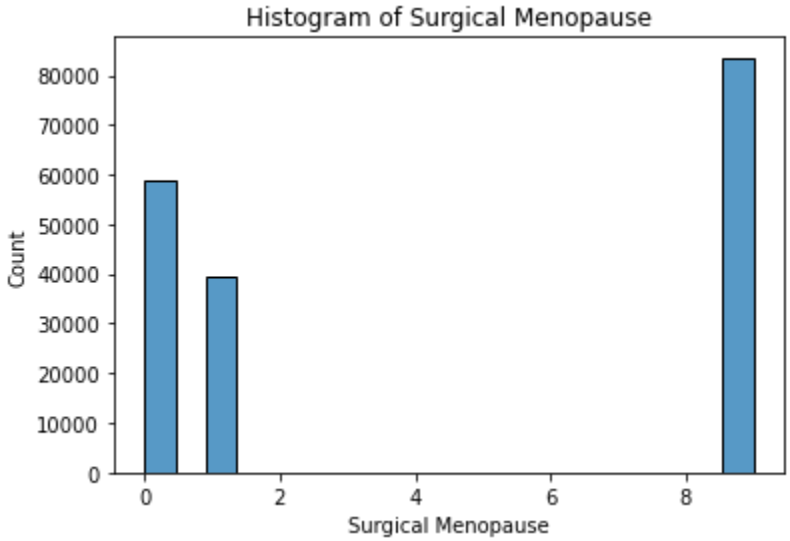
38.9% of the population identified as having unknown BMI (kg/m). Of the known BMIs, 41.4% of the results range between 10 and 29.99 for values (1) and (2), and 19.7% range between 30 to over 35, values (3) and (4).



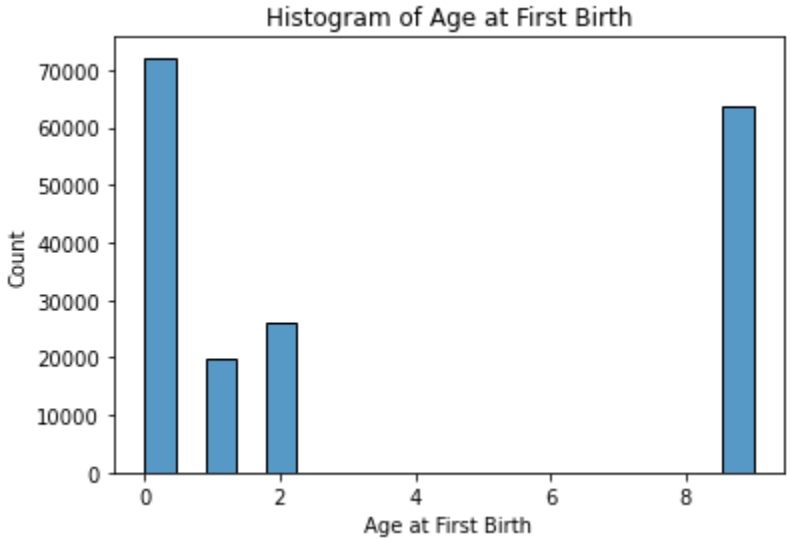
For results of the last mammogram before the index mammogram, 60.8% of mammogram results were negative for cancer (0) and 33% were unknown (9).



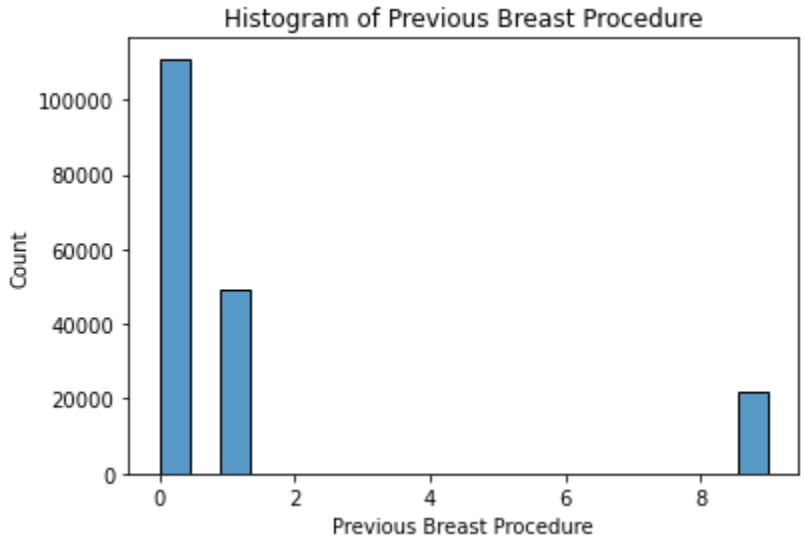
The mammograms were divided among the following current hormone therapy (HRT) statuses: 36% or 101,077 had no HRT (0), 30.5% or 89,659 were receiving HRT (1), and 33.5% or 93,885 were unknown or not menopausal.



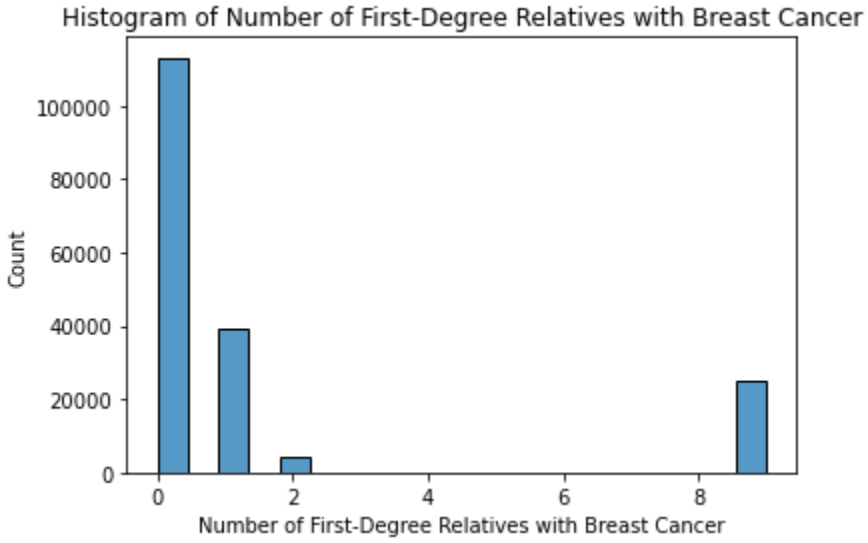
Of the participants, 33.6% or 94,435 experienced natural menopause (0), 23.1% or 64,823 underwent surgical menopause (1), and 43.3% or 121,402 had unknown status or not menopausal (9).



39.4% of the population had a first birth under the age of 30, represented by the value (0). The second-highest percentage, 32.6%, was unknown.



Most participants had not undergone a previous breast procedure (0), accounting for 58.4% of patients sampled. 28.8% had undergone a breast procedure and the remaining 12.9% were categorized as unknown.



Most of the dataset consists of participants who had no first degree relatives with breast cancer (0), accounting for 59.3% of participants.

Cancer Incidence Rates by Features

Race	Incidence Rate %
White	4.31
Asian/Pacific Islander	1.76
Black	2.73
Native American	0.93
Other/Mixed	0.94
Unknown	2.83
All	3.45

Age Group	Incidence Rate %
35- 39	1.24
40-44	2.96
45-49	2.81
50-54	3.12
55-59	3.75
60- 64	3.90
65-69	4.18
70-74	4.17
75-79	4.00
80-84	3.27
All	3.45

Age at First Birth	Incidence Rate %
< 30	2.93
30 or >	2.19
Nulliparous (Has not given birth before)	2.39
Unknown	4.87
All	3.45

Body Mass Index	Incidence Rate %
10 - 24.99	3.32
25 - 29.99	2.96
30 - 34.99	2.12
35 or >	1.79
Unknown	4.42
All	3.45

The incidence rate represents the ratio of cancer to non-cancer in a particular value of a feature. All incidence rates calculated were less than 5%. The incidence rate i.e the occurrence of cancer in women less than 30 years of age was slightly higher than the incidence rate in women over 30. This was different from our expectations. From the race incidence rates table, we can see that black participants had a higher incidence rate among the minority classes.

Feature Correlations

Positive Correlations with Invasive or in situ Ductal Carcinoma

- Cancer [1.0000]
- Invasive [0.8814] : Highest correlation coefficient, misleading because this is a label
- Age of First Birth [0.0542] : Highest Non-label correlation coefficient
- Body Mass Index [0.0372]
- Training [0.0303]
- Breast Density [0.0278]
- Age Group [0.0268]
- Surgical Menopause [0.0002]

Negative Correlations with Invasive or in situ Ductal Carcinoma

- Hispanic [-0.0375]
- Race [-0.0331] :
- Count [-0.0293]
- Previous Breast Procedure [-0.0372]
- Menopause [-0.0053]
- Current Hormone Therapy [-0.0019]

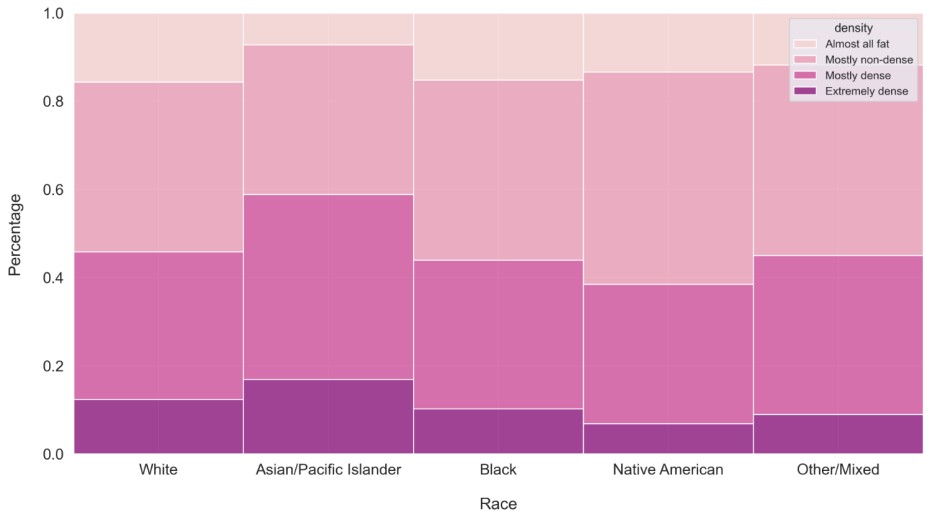
- Result of Last Mammogram [-0.0017]
- Number of first degree relatives with breast cancer[-0.0011]: Lowest correlation Coefficient

Observations

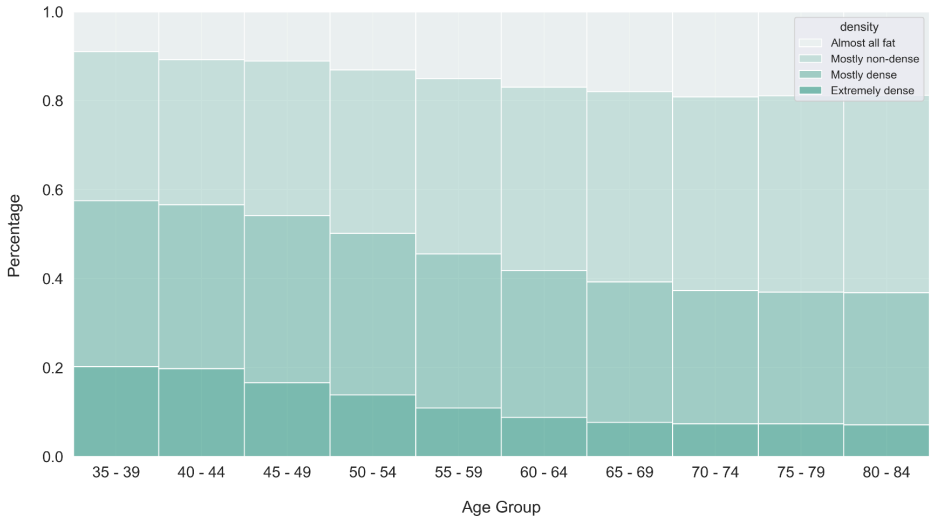
1. Most of the correlation coefficients are insignificant. (<0.1)
2. Perhaps Correlation Coefficients were low because there were a lot of unknowns

Breast Density

Breast tissue density vs race



Breast density tissue vs age group



These stacked barcharts show the variation in breast tissue density across different races and age groups. All samples with 'Unknown' 'density' or 'race' values were removed in these plots. In general, ~40% of women have breast tissue under the labels 'Mostly dense' or 'Extremely dense'. For 'Asian/Pacific Islander', this proportion can be as high as 60%. Across age groups, ~40% of older women have breast tissue that can be classified as 'Mostly dense' or 'Extremely dense' and this percentage rises to nearly 60% of the population age group for younger women. As mentioned before, mammograms are known to be imperfect diagnosis tools for breast cancer in dense breast tissue.

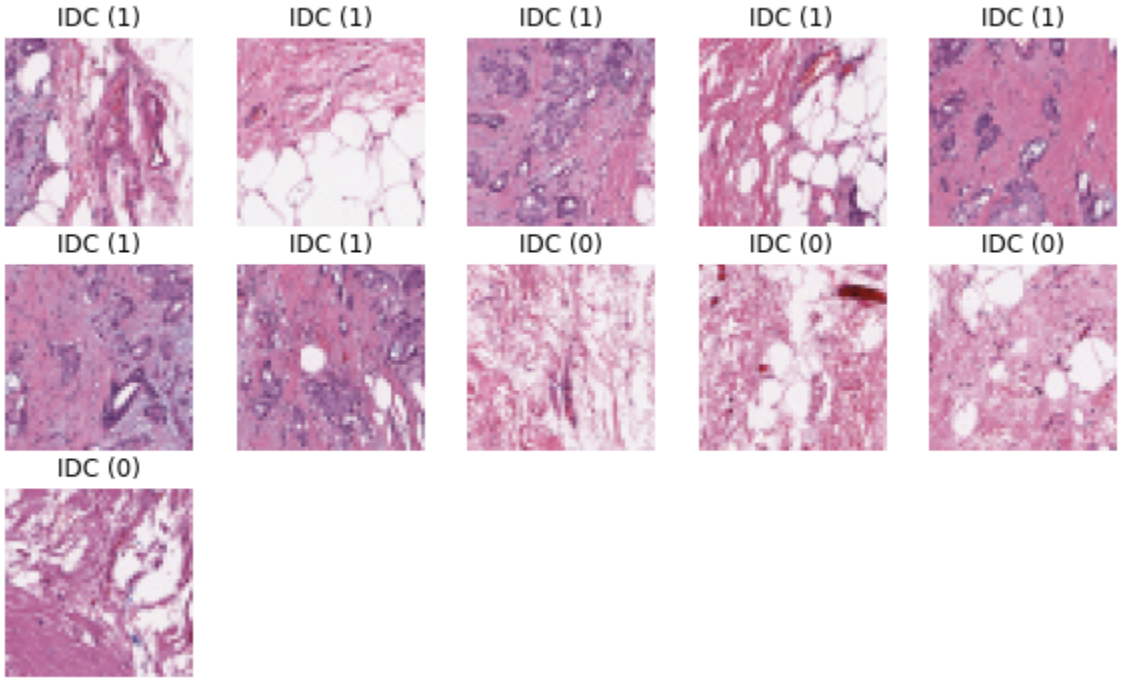
Other Feature Observations

Invasive is the only feature with an absolute correlation coefficient > 0.1. It is also very highly correlated with IDC. This is likely because Invasive Breast Cancer usually occurs means IDC or DC in situ has spread, thus this feature is a label in its own right and was not included in model building. Interestingly, 4912 patients were diagnosed with both invasive breast cancer and IDC. That is, invasive and cancer were both coded as 1. There was no patient that was diagnosed with only Invasive Breast Cancer, this provides stronger evidence that IDC occurs before Invasive Breast Cancer.

Another interesting feature is that there were 10083 non-white Hispanic patients in the dataset. Thus, removing Hispanic would lead to data loss, even though it is more an ethnicity and not a race.

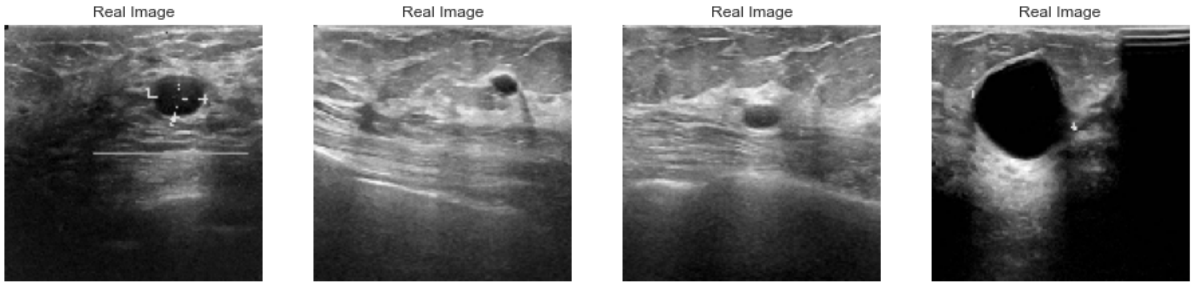
Breast Histopathology Dataset

Sample images from this dataset are shown below. The (1) indicates a positive (cancerous) label whereas the (0) indicates a negative (non-cancerous) label.



Breast Ultrasound

Examples of the ultrasound images are shown below:



The dataset contains 891 'benign' images, 266 'normal' images and 422 'malignant' images.

Model Architecture

Risk Estimation

Three models were used to predict the presence of Invasive or in situ Ductal Carcinoma from 12 risk factors - Menopausal, Age Group, Breast Density, Race, Hispanic, Body Mass Index, Age of First Birth, Number of first degree relatives with cancer, Previous Breast procedure, Result of Last Mammography, Surgical Menopause and Current Hormone Therapy.

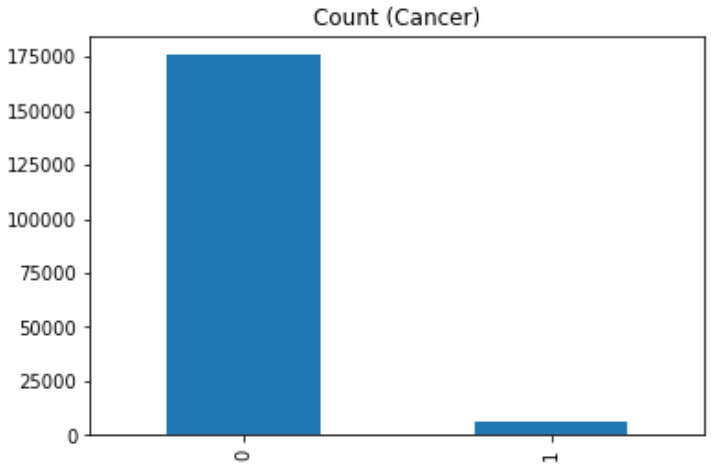
Models Used:

- **Logistic Regression**: Logistic Regression is used to model the probability of a certain class. In its basic form it uses a logistic function to model a binary dependent variable. It is a good starting point for classification tasks on datasets with binary targets like the risk estimation dataset. Being a relatively simple model it serves as a good benchmark for more complex models.
- **Linear Support Vector Machines: Linear SVM is another linear supervised learning technique. It is effective in high dimensional spaces. It is more complex and usually outperforms logistic regression. It is useful for comparison, in the event that there is not sufficient improvement, we can choose a simpler model, in this case - logistic regression.**
- **Gradient Boosting**: Gradient Boosting (or Gradient Boosted Decision Trees (GBDT) belongs to a class of machine learning algorithms called ensemble methods. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator. Boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a

powerful ensemble. Each estimator can be likened to a Decision Tree, varying the number of estimators or trees used to tune performance. GBDT is time-consuming but quite robust and accurate.

Feature Engineering

The original ratio of Cancer to No-Cancer samples in the dataset is **3:100** i.e 3% of the data has a label of 1. Thus, the ratio of minority class to majority is 0.03. Our dataset is highly imbalanced (see below), most of our targets are zero so a model can achieve high accuracy (up to 97%) by predicting 0 (No cancer) every time. To account for this data imbalance we employ **SMOTE - Synthetic Minority Oversampling Technique.** SMOTE is an oversampling technique where synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together (Satpathy, 2020)**.** Other approaches such as undersampling the majority and cross validation can be used. However, Undersampling may lead to loss of information and cross validation will expose the test data to the model.



For oversampling we used four ratios of the minority to majority class: [0.25, 0.3, 0.4 and 0.5] to test our models.

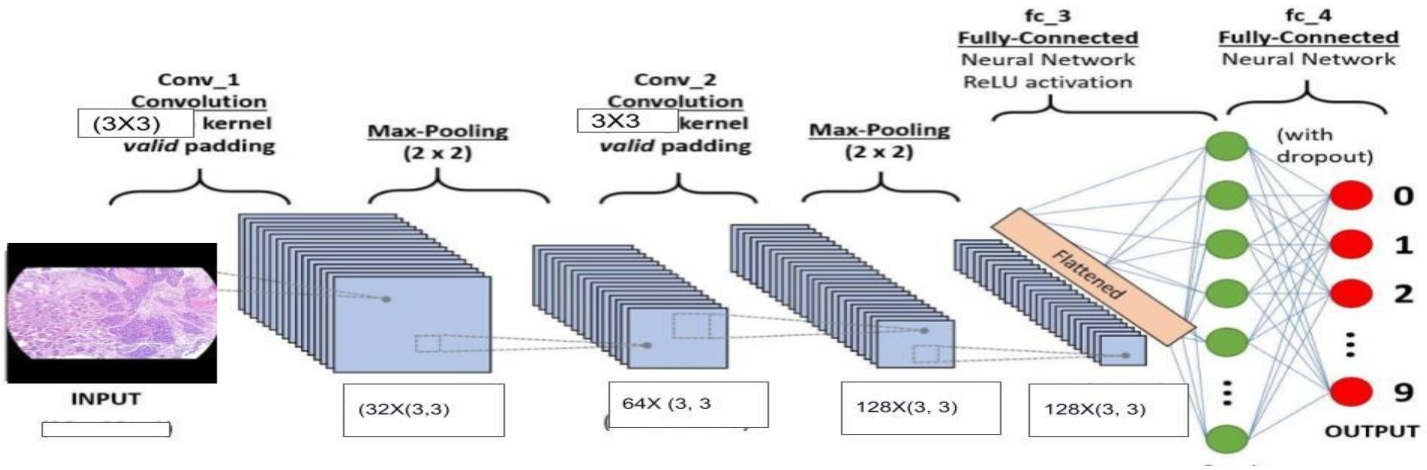
Breast Histopathology/Breast Ultrasound

ConvNet/CNN

The Convolutional Neural Networks (CNNs) is the most popular neural network model used for image classification problems because of its high accuracy. It has proven to be successful in many different real-life case studies and applications, including object detection, segmentation, face recognition, and so forth. The practical benefit of using CNN for our model prediction is that having fewer parameters greatly improves the time it takes to learn and reduces the amount of data required to train the model.

To design the CNN architecture, we use Four main types of layer: INPUT - CONV - RELU - MaxPOOL - FC

Our model consists of 4 Convolutional Layers, 4 MaxPooling2D Layers, 1 Flatten, and 2-Fully-Connected Layers. We added the layers in a sequential order. Each sequential layer performs some computation on the input it receives; then, it communicates the output information to the next layer.



Advantages

- High Accuracy
- Compared to its predecessors, the main advantage of **ConvNet/CNN** is that it automatically detects the important features without any human supervision.

Disadvantages

- Significantly slower due to max pool.
- If the CNN has several layers, the training process takes a lot of time if the computer doesn't have a good GPU.
- A ConvNet requires a large Dataset to process and train the neural network.

VGG-16

The other model that we used is VGG-16. VGG-16 is a type of convolutional neural network architecture proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition". The model layers are depicted in the figure below. The number of parameters are reduced because of its optimal architecture which ultimately reduces the training time. In addition, by pre-loading weights for VGG-16 trained on ImageNet and only fine-tuning the final layers (technique known as transfer learning), we can benefit from training on the 14 million images in ImageNet and reduce training time. This technique was also used on both the Breast Histopathology and Breast Ultrasound dataset. In both cases, the first 12 layers of the model were frozen (no tuning) while the last four layers were fine-tuned.

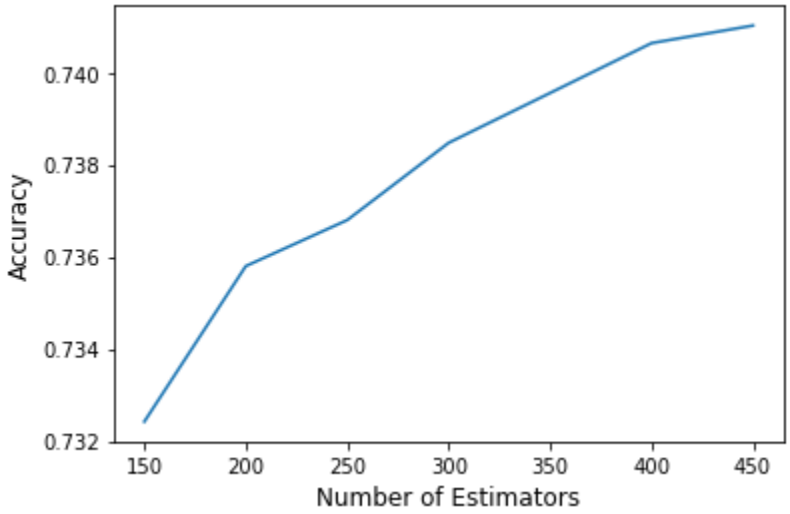
Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	224 x 224 x 3	-	-	-
1	2 X Convolution	64	224 x 224 x 64	3x3	1	relu
	Max Pooling	64	112 x 112 x 64	3x3	2	relu
3	2 X Convolution	128	112 x 112 x 128	3x3	1	relu
	Max Pooling	128	56 x 56 x 128	3x3	2	relu
5	2 X Convolution	256	56 x 56 x 256	3x3	1	relu
	Max Pooling	256	28 x 28 x 256	3x3	2	relu
7	3 X Convolution	512	28 x 28 x 512	3x3	1	relu
	Max Pooling	512	14 x 14 x 512	3x3	2	relu
10	3 X Convolution	512	14 x 14 x 512	3x3	1	relu
	Max Pooling	512	7 x 7 x 512	3x3	2	relu
13	FC	-	25088	-	-	relu
14	FC	-	4096	-	-	relu
15	FC	-	4096	-	-	relu
Output	FC	-	1000	-	-	Softmax

(Varshney, P. (2020). *VGGNet-16 architecture: A complete guide*. Kaggle. https://www.kaggle.com/blurredmachine/vggnet-16-architecture-a-complete-guide.

Evaluation and Results

Risk Estimation

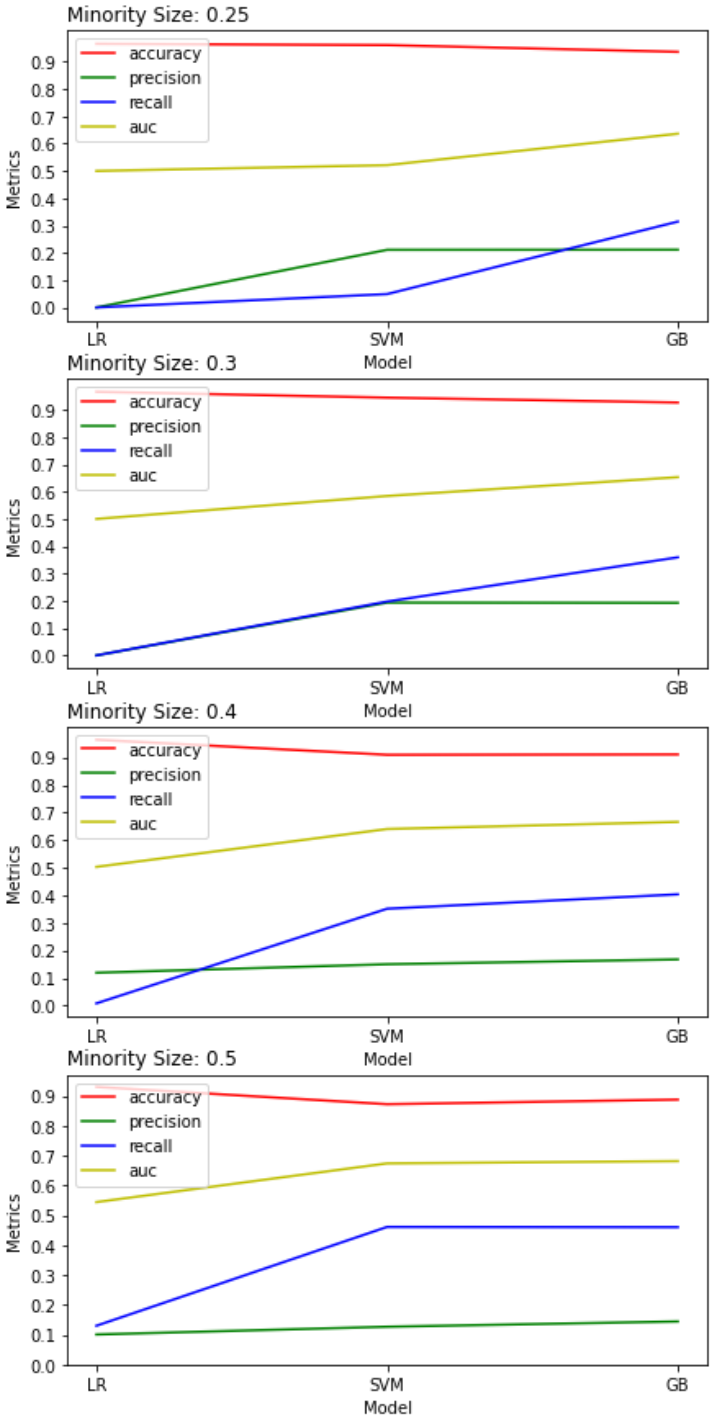
We evaluated all three models on the data without oversampling. This resulted in high accuracies (>96%) but recall and precision values of zero. The number of estimators (trees) in this ensemble was varied on the data (without oversampling) using Grid Search to find the best performance. Estimator Params - [150, 200, 250, 300, 400 and 450] 450 estimators gave the best accuracy. There was a linear increase in performance and number of estimators. Increasing the estimators even further may have led to increased performance but also an increase in training time. Thus, we chose 450 estimators for further model fitting tasks.



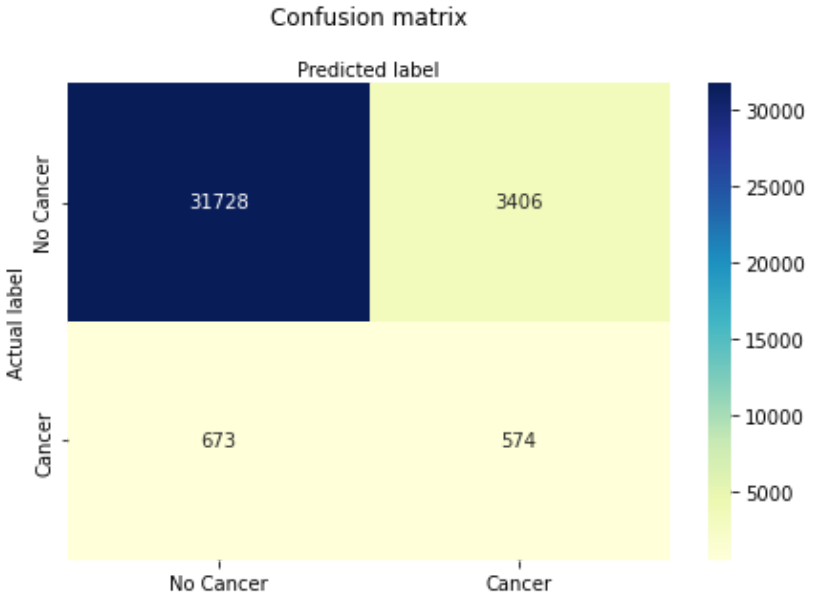
Plot of Training Accuracy against Number of Estimators for Gradient Boosting

We oversampled the minority class (no cancer) using four ratios of the minority to majority class: [0.25, 0.3, 0.4 and 0.5]. For each ratio, we extracted our 12 features and target(cancer) and fit the features and targets on all three models. This resulted in 12 models with varying parameters.

The plots below show the model accuracy, recall, precision and auc for each tested model (LR for Logistic Regression, SVM for Support Vector Machines and GB for Gradient Boosting) for each minority ratio. We found that the Gradient Boosting performed best because it had better values of accuracy, precision and recall. From the plots below, we can see that oversampling the training and validation sets with a higher minority ratio of 0.4 or 0.5 is more desirable.



The confusion matrix for the validation set using the best model (GB with 0.5 minority ratio) is shown below, along with the calculated accuracy, recall, precision and F1-score.



True Positives (TP) = 574
True Negatives (TN) = 31728
False Positives (FP) = 3406
False Negatives (FN) = 673

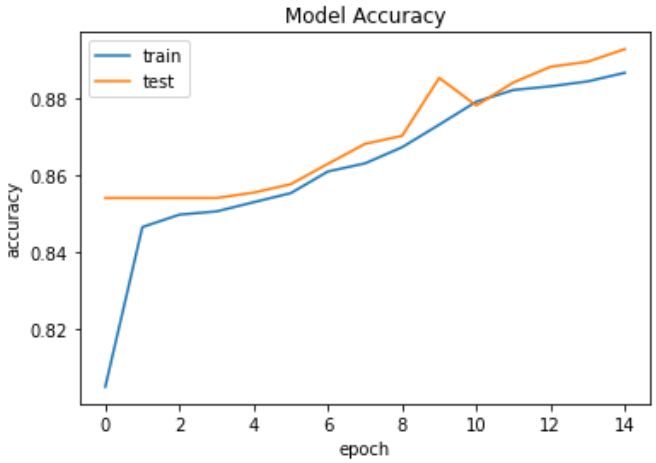
Best Accuracy : 89%
Best Recall: 46%
Best Precision: 14%
Best F1 Score: 21%

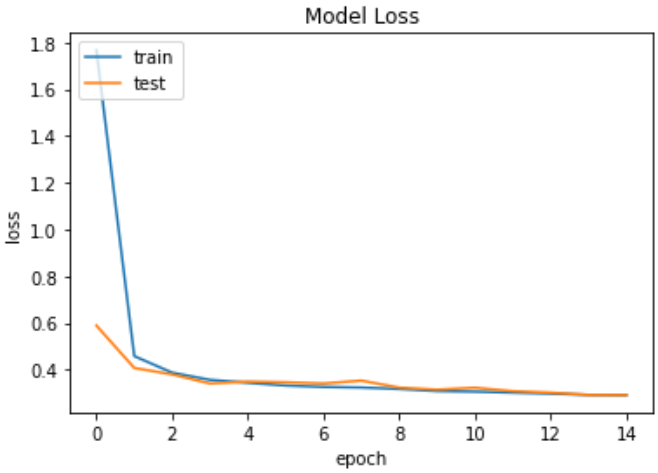
Risk Estimation Results Summary

The model accuracy is high, but most of the performance is explained by precision, recall, and F1-score, which are low. It seems that the model misclassified the images. The confusion Matrix shows many false positives because the model incorrectly predicts the *positive* class of cancer diagnosis where there is no cancer. The Low Precision emerges because very few positive predictions are true, and the Low Recall because most positive values are never predicted. The low F1 score is an indication of both poor precision and poor recall. Despite the feature engineering efforts and multiple models the evaluation metrics are not good. Perhaps these features are not enough to adequately predict Breast Cancer. For better performance, we may try advanced feature selection techniques, more complex models (such as Long Short Term Memory Networks) or a more robust dataset.)

Breast Histopathology

First, we trained a CNN model for 15 epochs with a batch size of 50 using a binary_crossentropy loss function and an Adam optimizer. After every learning cycle, our network evaluates its performance on the training and validation sets. Typically, the training and validation loss will both decrease after each epoch. The model accuracy plot shows that the model has comparable performance on both train and test datasets, indicating little to no overfitting. The CNN model has a validation accuracy of 88%.





When we increase the number of epochs, both the training and validation loss keep on decreasing, as can be seen in the figure above.

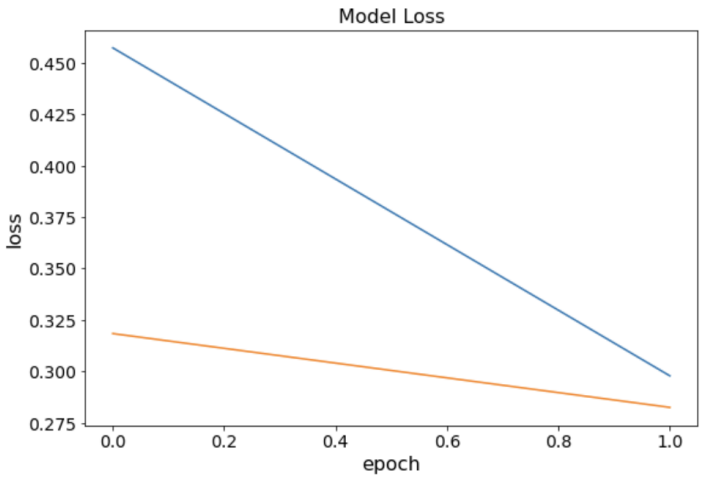
Confusion Matrix

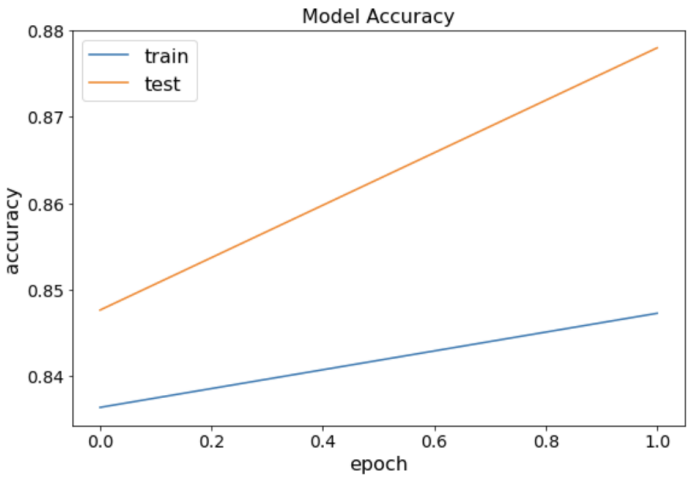
We calculated a confusion matrix from our predictions. The testing data contains 15,000 samples, of which 2367 belong to the "no cancer" class, and the remaining 12,633 belong to the "cancer" class. The confusion matrix shows that our model correctly classified 12374 samples belonging to the "Cancer" class as "Cancer "(True Positives). Also, 938 images belonging to the "No cancer " class were correctly classified as "No Cancer" (True Negatives). 259 samples belonging to the " cancer" class were incorrectly classified as "No Cancer " (False Negative). Another 1429 samples

belonging to the "No cancer " class were incorrectly classified as "Cancer" (False Positives). The precision score of our model is 89.6%, the recall of our model is 98%, and the F1 score is 93.6% on Testing data.

VGG-16 Results for Breast Histopathology

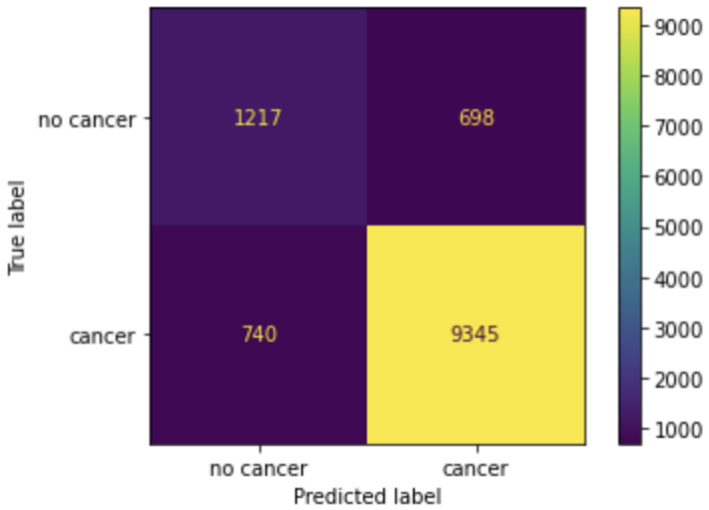
We also implement the VGG-16 model for Breast Histopathology. VGG-16 is a well-known model known for being computationally efficient for image-based tasks. We used it on the Breast Histopathology dataset as a comparison with the traditional CNN presented above. We chose not to proceed to VGG-19 since VGG-19 adds little performance gain but requires far more memory.





The VGG-16 here was trained with a batch size of 32 and a learning rate of 5e-5. The model was trained for 2 epochs.

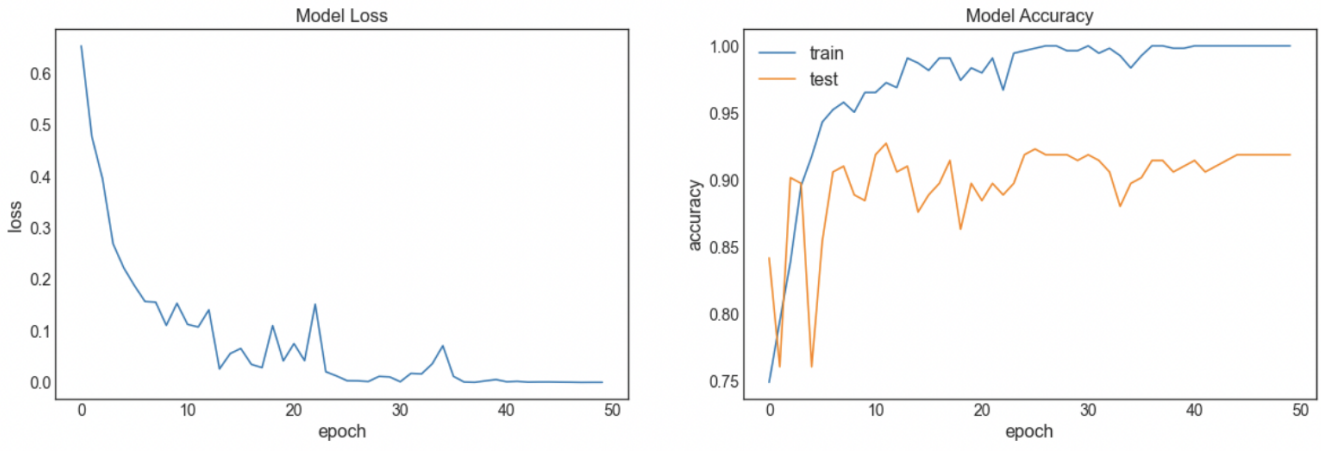
For evaluating the model performance, again we utilize a confusion matrix. We chose to use precision, recall, and F1-scores for model evaluation comparisons between the CNN and VGG-16 models. We evaluated our model on the testing data using a confusion matrix. The testing data contains 12,000 samples, of which 1957 belong to the "no cancer" class, and the remaining 10,043 belong to the "cancer" class. The confusion matrix shows that our model correctly classified 9345 samples belonging to the "Cancer" class as "Cancer "(True Positives). Also, 1217 images belonging to the "No cancer " class were correctly classified as "No Cancer" (True Negatives). 740 samples belonging to the " cancer" class were incorrectly classified as "No Cancer " (False Negative). Another 698 samples belonging to the "No cancer " class were incorrectly classified as "Cancer " (False Positives).



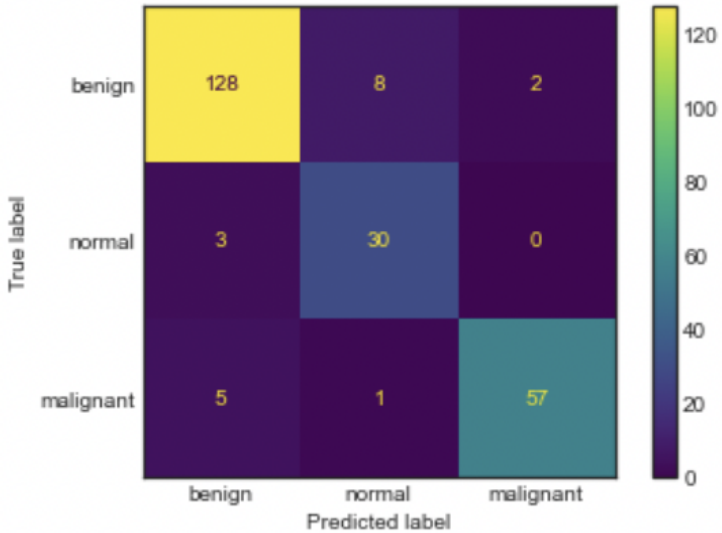
The precision score of our model is 93.0%, the recall of our model is 92.7%, and the F1 score is 92.9% on Testing data.

Breast Ultrasound

VGG-16 was also trained on the Breast Ultrasound dataset in order to classify images as 'normal', 'benign' or 'malignant'. For model fitting, we used a batch size of 16 and a learning rate of 5e-5. The model was trained for 50 epochs. The loss and accuracy curves for the fine-tuned VGG model are shown below:



We observed some overfitting on the training set, which is not a surprise given the small size of the dataset. Nevertheless, the overall validation accuracy was 92.7%. The confusion matrix for the results of this model is shown here:



Based on the confusion matrix, we can calculate the following metrics for this classifier:

	Benign	Normal	Malignant	Weighted Average
Precision	94.1%	76.9%	96.6%	92.3%
Recall	92.8%	90.9%	90.4%	91.9%
F1-score	93.4%	83.3%	93.4%	92.0%

The precision and recall for all labels are quite high. We chose to use a weighted average to quantify overall model precision, recall and F1 scores because of the large class imbalance, especially in the 'normal' class. Overall, all metrics are ~0.92, which would indicate a good model. However, the model is severely overfitting, which means that it may not generalize well in a clinical setting.

Given the nature of this work, one of our primary concerns is false negative rates (do not want to miss true cancer label). The false negative rate for this classifier for 'malignant' is ~8%, which is quite low. In contrast, the false negative rate in mammograms can be 10-20% (Newton, 2019), so this classifier is on par with the lower end estimates for false-negatives in mammograms. One way to improve this model would be working with a larger dataset, which would hopefully avoid the overfitting we observe in the accuracy plot. In particular, the dataset could benefit from more 'normal' labels. In addition, it would be useful to have demographic data for the patients along with their ultrasounds. Age, race and/or breast tissue density data could help check for model accuracy across different groups. Mammograms are known to be less effective for women with dense breast tissue, which can vary with age and race. As a result, mammograms can be highly inaccurate for younger women and women of different races, and there is a real need for accurate and easily accessible diagnosis methods that could be filled with ultrasounds.

Image Classification Results Summary

Both the CNN model and the VGG-16 model perform comparably in terms of precision and recall. The precision of VGG-16 is higher while the recall of CNN is better. Therefore, we consider which one is more important in pharmaceutical and clinical practice. For a person, is it false positive (meaning a person who is identified to have cancer but he/she has) or false negative (meaning a person who is not found to have cancer but he/she has) more serious? It is obviously the false negative. That being said, a high recall should be considered more important! This suggests that the traditional CNN is the winner. However, we should also see that VGG only uses 2 epoches to present a great result, that is a surprisingly short time and more efficient in comparison with traditional CNN.

In general, both the CNN model and the models that were fine-tuned versions of VGG-16 showed great performance in identifying cancer with high recall and precision values. The VGG-16 models benefitted from having initial layers that were pre-trained on ImageNet and the deep nature of the model, which can be beneficial for image

classification tasks. In addition, the pre-training was especially important for the Breast Ultrasound dataset, which only contained 1500 samples. However, the deep nature of the model and the large number of parameters also led to overfitting on the training set. Nevertheless, the model exhibited 92.7% accuracy on the test set as well as high recall and precision rates. In order to reduce overfitting, we can consider adding regularization, augmenting the variety of the data, and collecting a larger, more diverse dataset.

Conclusion and Future Work

This was a preliminary analysis to explore the factors that contribute to breast cancer and investigate how image classification could be used to improve accuracy of cancer detection.

In the first part of this study, the factors that correlate with ductal carcinoma, i.e. breast cancer were evaluated. Initial analyses of risk factors from over 180,000 individuals indicate that breast cancer incidence generally increases with age. Over the age of 40, incidence remains low overall but doubles from 1.2% to 2.9% compared to 35-39 year olds. This matches The United States Preventive Services Task Force (CDC, 2021b) recommends women aged between 50 and 74 at average risk should have mammograms every two years. The analysis above shows it could be beneficial for women aged over 40 to seek medical advice for their screening. There is the risk of false positives, so simply recommending additional mammograms for the younger age group is not advisable.

The second part of this study was to understand the factors that result in uninterpretable mammograms, and to improve the accuracy of screening results for secondary methods: ultrasound and histopathology. High breast tissue density can obscure the presence of tumors on standard mammograms. In this study, up to 60% of women had mostly or extremely dense breast tissue. The highest being in younger age groups, which strengthens the conclusion that additional mammograms for women in their 40s is not the full answer to early cancer detection. There are a number of secondary screens available to women that provide an opportunity for early diagnosis even if there is an inconclusive mammogram. Using neural network models for ultrasound and histopathology had mixed results. Ultrasound image classification resulted in more false positives, which again would result in more diagnostics for the patient. Ultrasound is non-invasive, so a negative result would be of value.
The histopathology neural network had high precision and recall. This was a highly trained model which is not recommended for use in real clinical settings, but provides a glimpse into the potential for this approach.

This initial work provides opportunities for interesting further work. For example, more detailed statistical analyses of the risk factors, investigating different model settings and testing the model on new screens. This area is actively investigated, so working with oncologists and patients to understand the areas of unmet need would be an important step to leveraging the power of data analytics to improve clinical practice and ultimately save lives. Expanding and using larger, more diverse datasets that contain little to no unknown data would help to expand the scope of the testing models and provide more robust results.

References

Albrecht, B., Alfano, S., Keane, H., & Yang, G. (2021). Delivering innovation: 2020 oncology market outlook. McKinsey & Company; Company. Retrieved from https://www.mckinsey.com/industries/life-sciences/our-insights/delivering-innovation-2020-oncology-market-outlook.

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. Data in Brief, 28, 104863. https://doi.org/10.1016/j.dib.2019.104863

American Cancer Society. (2019). Limitations of mammograms: How often are mammograms wrong? American Cancer Society. Retrieved, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html.

Barlow, W.E., White, E., Ballard-Barbash, R., Vacek, P.M., Titus-Ernstoff, L., Carney, P.A., Tice, J.A., Buist, D.S.M, Geller, B.M., Rosenberg, R., Yankaskas, B.C., Kerlikowske, K. (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst. 2006; 98:1204-1214.

Blumen, H., Fitch, K., Polkus, V. (2016) Comparison of Treatment Costs for Breast Cancer, by Tumor Stage and Type of Service. Am Health Drug Benefits. 2016 Feb; 9(1): 23–32.

Breast Cancer Research Foundation (BCRF). (2021). *Breast cancer statistics and resources*. Breast Cancer Statistics And Resources. Retrieved from https://www.bcrf.org/breast-cancer-statistics-and-resources/.

Brownlee, J. (2020, August 14). *What is a confusion matrix in machine learning*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/confusion-matrix-machine-learning/

Center for Chronic Disease Prevention and Health Promotion (CDC) 2021. Cost-Effectiveness of Breast Cancer Interventions. Retrieved from https://www.cdc.gov/chronicdisease/programs-impact/pop/breast-cancer.htm.

Center for Chronic Disease Prevention and Health Promotion (CDC) 2021b. What Is Breast Cancer Screening? Retrieved from https://www.cdc.gov/cancer/breast/basic_info/screening.htm

Hughes, R. G. (2008). *Tools and strategies for quality improvement and patient safety*. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Retrieved October 18, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK2682/.

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of Pathology Informatics, 7(1), 29. https://doi.org/10.4103/2153-3539.186902

Keras. (2021). *Keras Documentation: Accuracy metrics*. Keras. Retrieved from https://keras.io/api/metrics/accuracy_metrics.

Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high?. *The British journal of radiology*, *89*(1065), 20160045. https://doi.org/10.1259/bjr.20160045

Newton, E. (2019). What is the prevalence of false-positive and false-negative mammography results in breast cancer screening? https://www.medscape.com/answers/1945498-167946/what-is-the-prevalence-of-false-positive-and-false-negative-mammography-results-in-breast-cancer-screening

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Rochman, S. (2015) Study Finds Black Women Have Denser Breast Tissue Than White Women, JNCI: Journal of the National Cancer Institute, Volume 107, Issue 10, October 2015, djv296, https://doi.org/10.1093/jnci/djv296

Satpathy, S. (2020) Overcoming Class Imbalance Using SMOTE Techniques, Analytics Vidhya, https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR, abs/1409.1556*..

Priyanshi, Sharma (2021), Disadvantages of CNN models, https://iq.opengenus.org/disadvantages-of-cnn/

University of Texas. (2021). Cancer treatment algorithms. MD Anderson Cancer Center. Retrieved from https://www.mdanderson.org/for-physicians/clinical-tools-resources/clinical-practice-algorithms/cancer-treatment-algorithms.html.

Varshney, P. (2020, July 28). *VGGNet-16 architecture: A complete guide*. Kaggle. Retrieved from https://www.kaggle.com/blurredmachine/vggnet-16-architecture-a-complete-guide.