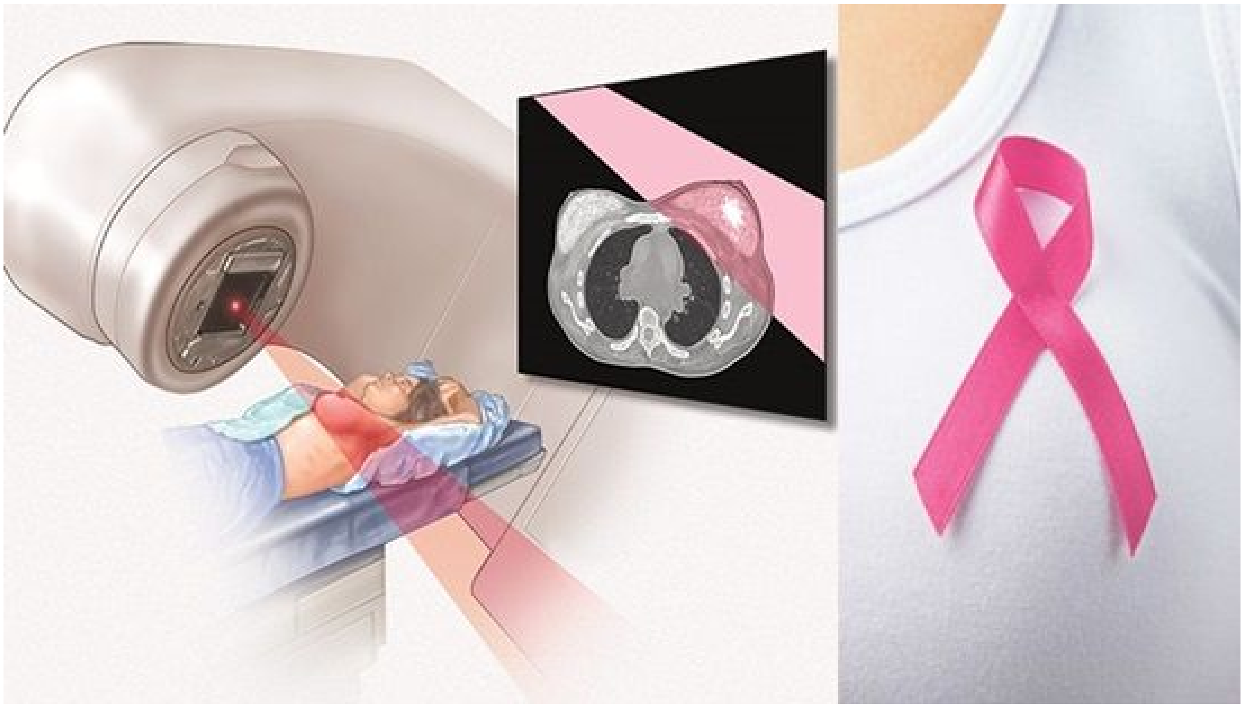

Breast Cancer (IDC) Detection

Using Digital Pathology Images.

25 November 2018

Animesh Seemendra
Machine Learning Engineer Advance Nanodegree
Capstone Project Proposal

Domain Background

Breast Cancer is the most common type of cancer in woman worldwide accounting for 20% of all cases. In 2012 it resulted in 1.68 million new cases and 522,000 deaths. One of the major problems is that women often neglect the symptoms, which could cause more adverse effects on them thus lowering down the survival chances. In developed countries the survival rate is although high, but it is an area of concern in the developing countries where the 5-year survival rates are poor. In India, there are about one million cases every year and the five-year survival of stage IV breast cancer is about 10%. Therefore it is very important to detect the signs as early as possible.

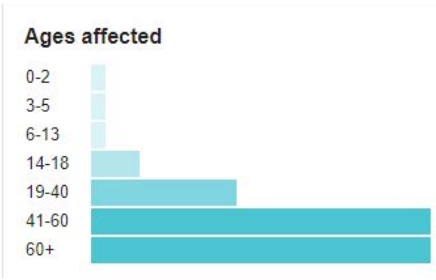

Fig 1

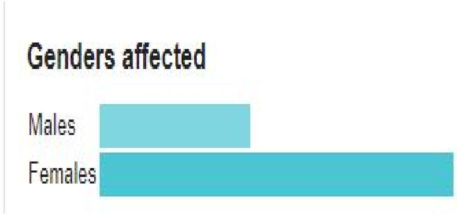

Fig 2

The data is on Breast cancer in India taken from various hospitals (Apollo and others)

Invasive ductal carcinoma (IDC) is the most common form of breast cancer. About 80% of all breast cancers are invasive ductal carcinomas. Doctors often do biopsy or a scan if they detect signs of IDC . The cost of testing for breast cancer sets one back with $5000, which is a very big amount for poor families and also manual identification of presence and extent of breast cancer by a pathologist is critical. Therefore automation of detection of breast cancer using Histopathology images could reduce cost and time as well as improve accuracy of the test. This is an active research field lot of research papers and articles are present online one that I like is :-https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5453426/ as they used deep learning approach to study on histology images and achieved sensitivity of 95% which is greater than many pathologist(~90%). This shows the power of automation and how it could help in detection of breast cancer. The motivation to work on this domain came from the lesson I studied in the nanodegree about skin cancer and how AI could help medical field and saves multiple lives therefore it inspired me to create my capstone project on topic related to cancer which is the most deadly form of all diseases and specially breast cancer which often get neglected.

Problem Statement

The idea is to use pathology test images and classify them as IDC(+) and IDC(-). Accurately identifying and categorizing breast cancer subtypes is an important clinical task, and automated methods can be used to save time and reduce error. The pathological tests include images of the tissues, the task is to train a computer to use these images and respond on whether the person is IDC(+) or IDC(-). Since it is a medical field problem it is important that sensitivity of the output should be high.

Datasets and Inputs

The problem statement involves using images that are obtained during pathology tests to detect whether the patient is IDC positive or negative. Histopathology images are the images of tissues that are obtained during pathology tests, therefore, these images will act as inputs for the problem.

The dataset that contains histopathology images for breast cancer is present on kaggle at : https://www.kaggle.com/paultimothymooney/breast-histopathology-images . The dataset contains 198,738 of negative IDC and 78,786 of positive IDC, therefore, it is a good dataset with enough data for our task. The original dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. However, the data that I have selected contains images that are cropped from the original dataset i.e it contains patches of regions where the IDC occurs, making it more specific to our problem. Each patch's file name is of the format: u_xX_yY_classC.png — > example 10253_idx5_x1351_y1101_class0.png . Where u is the patient ID (10253_idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC.

Solution Statement

Our data involves images with the classes written on data file name, therefore, we would need to extract the class name from it and create a column to store them. We also need to split the dataset into the training set, validation set and testing set. Testing set for checking how good the model works on completely unseen data and validation set to check and avoid underfit or overfit, the will also help to select the best model. One hot encoding will be done in classes column so that it could work better with our model. Image processing step is also required to reduce the pixel range from 0-250 to 0-1. After it CNN model is to be used to predict the class, CNN creates an effective architecture the 2D structure of the image, therefore, it would be the best to use, considering that we are working with the images.

Benchmark Model

The breast cancer detection using pathology images is a critical task even for pathologists. The accuracy of the model present in the paper that was mentioned is 74.1 %. The various website mentioned different accuracy that a pathologist achieves while detecting IDC but all were above 90% and sensitivity mentioned was also above 90. So 90% could be a good benchmark to achieve but there is no authenticity about the source. Considering the fact that the data mentioned in the paper has more classes and is different from what I will be using, therefore, I will consider 70% accuracy as a benchmark result and would try to achieve it using CNN structure built from scratch (this would be our benchmark model) and if still their accuracy is not achieved as expected I would try to add transfer learning and Image argumentation to achieve the expected score.

Evaluation Metrics

The performance of the model will be evaluated using ROC curve and confusion matrix. A receiver operating characteristic curve, i.e., ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity). It uses the concept of true positive, true negative, false positive and false negative.

Sensitivity = $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$	Recall = $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
Specificity = $\frac{\text{True Negative}}{\text{True Negative + False Positive}}$	Precision = $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

The perfect classification has the area under the ROC curve equal to 1. Therefore closer the area of our ROC curve to 1 better would be our model. Third is confusion matrix, it is a two by two table that contains four outcomes produced by a binary classifier. Various measures, such as error-rate, accuracy, specificity, sensitivity, and precision, are derived from the confusion matrix. Sensitivity and accuracy (ACC) can be calculated from the confusion matrix , these two are considered for the benchmark therefore confusion matrix can help in calculating them . Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR.

- $\mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$

The ROC curve and confusion matrix would be a good evaluation matrices because they both are used for binary classification and our data is also based on binary classification. Also the benchmark and the solution we have decided, we require accuracy and sensitivity to be calculated which can be easily obtained through these two matrices along with visual correctness of the model.

Project Design

The workflow of solving the problem would be in the following order :

- Import Modules
- Explore and Visualizing Data
- Extraction of classes from the file path
- Preprocessing the Data
    - Converting categorical data into one hot encoding.
    - Image processing
        - Image resize to 50X50
        - Image pixel scaling to (0-1) from (0-250)
- Training, validation and testing data split.
- Model using CNN from scratch
    - Model Building
    - Model evaluation
- Model using Transfer Learning
    - Model Building
    - Model Evaluation
- Model using transfer learning and Image argumentation
    - Model building
    - Model evaluation
- Comparison of models
- Evaluation of best model using confusion matrix and ROC curve
- An algorithm to direct supply the image to the model.
- Final Conclusion

Refrences

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5453426/
- https://www.breastcancer.org/research-news/study-on-accuracy-of-biopsy-results
- https://www.cbsnews.com/news/breast-biopsies-often-get-it-wrong/
- https://en.wikipedia.org/wiki/Breast_cancer
- http://www.breastcancerindia.net/statistics/trends.html
- https://classeval.wordpress.com/introduction/basic-evaluation-measures/
- https://en.wikipedia.org/wiki/Receiver_operating_characteristic