Study on Breast Cancer Prediction using Fine Needle Aspiration (FNA)

1st Reuben Abraham Andrews
Computer Science and Engineering
PES University
Bangalore, India
reubenandrews2000@gmail.com

2nd Zia Ur Rehman
Computer Science and Engineering
PES University
Bangalore, India
ziaprivate11@gmail.com

3rd Nikita Shesha Patgar
Computer Science and Engineering
PES University
Bangalore, India
nikkipatgar024@gmail.com

Abstract

The goal of this project is to create an understanding of how one can use statistical techniques on Fine Needle Aspiration results to detect Breast Cancer, while also classifying tumors as malignant or benign.
In the first stage of this report, we wish to present and demonstrate the synopsis of the problem statement, explaining the dataset and it's relation to the problem statement using an Exploratory data analysis (EDA). The problem statement is manipulated and created taking into consideration various angles using other existing papers and kaggle EDAs.
Our objectives and goals moving further down the line will be to perform a few more EDA/Visualizations to greater understand the dataset and our problem statement to refine and improve the model to be more forgiving, accurate and informative for a larger scope of use and analysis. These adaptations will help us to forcast with an even higher accuracy. We also plan to experiment with various models in order to select the most optimal one.

Glossary: FNA - Fine Needle Aspiration

1 Introduction

Breast cancer is the most common cancer in women in developed countries, being the primary cause of suffering in young women. Survival from breast cancer has significantly improved, and the potential late effects of treatment and the impact on quality of life have become increasingly important. Young women constitute a minority of breast cancer patients, but commonly have distinct concerns and issues compared with older women, including queries regarding fertility, contraception and pregnancy. Further, they are more likely than older women to have questions regarding potential side effects of therapy and risk of relapse or a new primary. In addition, many will have symptoms associated with treatment and they present a management challenge. Reproductive medicine specialists and gynaecologists commonly see these women either shortly after initial diagnosis or following adjuvant therapy and should be aware of current management of breast cancer, the options for women at increased genetic risk, the prognosis of patients with early stage breast cancer and how adjuvant systemic treatments may impact reproductive function.
For our project, We'll be aiming to catch these carcinogenic tumors and lumps in young women using Fine Needle Aspiration along with simple regression techniques. Our aim to to statistically classify the bodies.

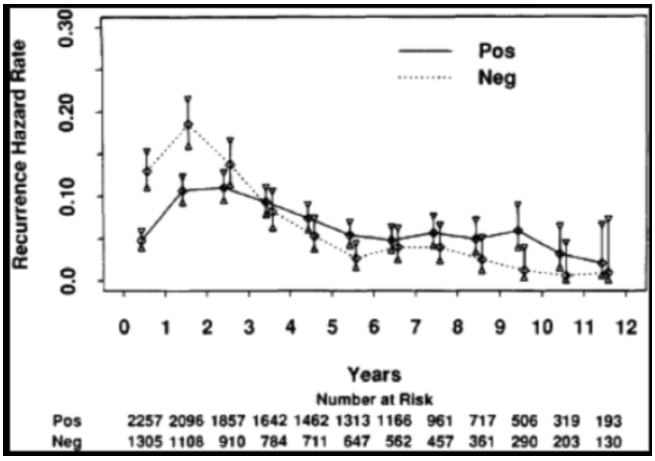

Figure 1: Annual hazard of recurrence of 3563 patients separated by ER status. ER, estrogen receptor.

By using simple correlation between standardized scans and obtained scans we can use various regression techniques to catch these outliers through the findings of Fine Needle Aspiration, thus increasing the chances of catching the cancer before it becomes deadly.

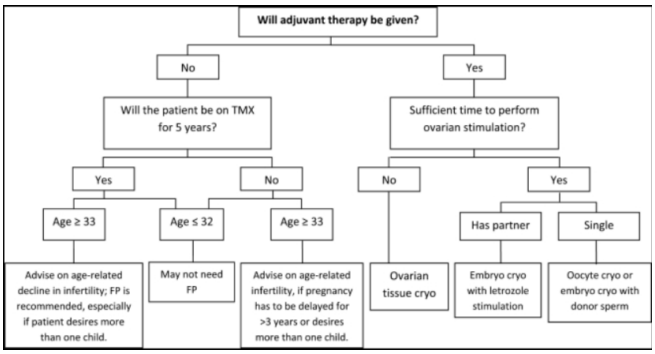

Figure 2: A proposed algorithmic approach to decision-making for fertility preservation in breast cancer patients.

2 BackGround

2.1 Computer Aided Detection (CAD)

The term CAD for breast cancer can be associated with two methods: Computer-assisted detection (CADe) or Computer-assisted diagnosis (CADx). The first focuses on the detection of suspicious regions in medical images, while the second is responsible for diagnosing the suspicious region. CAD systems generally have the following parts :-

1. The region of interest (ROI): The area of the lesion is selected for segmentation manually, semi-automatically or fully automatically.
2. Image preprocessing: The image is enhanced, eliminating noise and increasing contrast.
3. Segmentation: The lesion is segmented from the area of interest, selected in the first process, by identifying the contour or by region of pixels.
4. Feature extraction and selection: Measurable features are extracted from the segmented region and grouped into a feature vector. Examples of these characteristics are: Perimeter of the lesion, size, area and elongation among others. The characteristics selection process is performed using filter, wrapper, or hybrid techniques.
5. Automatic classification: This last step is essential for CADx systems, as it determines the class of the lesion using the characteristics provided. The classes generally correspond to the type of lesion or malignant or benign condition.

2.2 Breast Cancer

Breast cancer (breast) is a malignant tumor that originates in the cells of the breast. A malignant tumor is a group of cancer cells that may grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body.

This disease occurs almost entirely in women, but men can develop it as well. The validation dataset for breast cancer that was tested contains 357 Benign cases (Noncancerous) and 212 Malignant cases (Cancerous). The data tables for our analysis can be consulted on the project's **Github repository**

2.3 Fine Needle Aspiration

The fine needle aspiration biopsy technique is the most cost-effective test for the first-line diagnosis of breast lesions, palpable and non-palpable, in a simple and easy way. It can be performed in the office, with an immediate evaluation, using a needle with syringe and suction gun.

Fine needle aspiration biopsy of the breast can be performed on any palpable mass independent of imaging findings, especially in young women whose masses persist for more than two months and do not decrease in size or disappear with neither ovulatory nor menstrual phase. It is a routine and safe clinically usable technique. By combining medicine with statistical analysis and rudimentary methods, We can detect carcinogenic bodies.

The FNA test involves extracting fluid from the breast tissue, and a visual examination of this sample under a microscope. In smart detection, at first, this image turns to a gray level (it does not require color info). Then the related software defines the cell core boundary based on the image processing techniques, and calculates the features for each core such as radius, texture, perimeter, area, compaction (square perimeter divided by area), flat (mean difference in length of lines radially adjacent), concavity, symmetry and fractal (border kernel of approximation coastline. Finally, it calculates the mean square error and the mean of the three largest achieved values.

2.4 Pattern Recognition

Feature selection is the choice of features that have maximum power at output prediction. To solve the problems which depend on optimal subset, feature selection algorithms are divided into two major categories. If feature selection is done independent of any type of learning algorithm, it will be called Filter method, in which, the result of the selected features are determined before processing. If the assessment process is associated with a classified algorithm, the method of feature selection will be called Wrapper or Closed loop. This is a common pattern recognition system which consists of 4 sections: Feature extraction and selection; Designing and training classifier; and Testing.

3 Dataset

The dataset used for analysis and reference is the Breast Cancer Wisconsin (Diagnostic) Data Set which can be found on Kaggle. This dataset tells us about features that are computed from a digitized image of a Fine Needle Aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus which include radius, texture, perimeter, area and so on. The dataset can be viewed here :-
`https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?sort=votes`

4 Exploratory Data Analysis(Initial Insights)

In order to understand the nature of the data, some of the attributes that were presenting us with null values were dropped. We have used the required packages and libraries to obtain and present us with a readable statistical analysis. Correlations were made and visualised. Training and testing datasets were taken to further help our search. Three models were taking into consideration :-

1. Logistic Regression Model
2. Decision Tree Classifier
3. Random Forest Classifier

From our initial EDA, we have seen that the Random Forest Classifier gave us the most accurate results in distinguishing benign bodies from malignant ones. The EDA can be seen **here**.

5 Acknowledgement

We would like to show our profound gratitude and sincere thanks to Dr. Jyothi R for providing us with this opportunity, and guiding us along the way. We would also like to thank the Computer Science and Engineering department at PES University, for always inspiring us to conduct frequent research and inculcating a problem-solving discipline in us. We would also like to acknowledge our assistant professors and the teaching assistants who have helped in the making and material of the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.

6 References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4209568/

2. https://pubmed.ncbi.nlm.nih.gov/19174449/

3. https://ieeexplore.ieee.org/document/9254891

4. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01506-y