Breast Cancer Prediction using Deep Learning

Anandhavalli D[1], Sridhar C R[2], Naga Subramanian S[3]

[1]*Assisstant Professor, Velammal College of Engineering & Technology, Madurai*
[2&3]*UG Students, Velammal College of Engineering & Technology, Madurai*

---***---

Abstract - *Breast cancer is the most frequent cancer among women, impacting over 1.5 million women each year, and also causes the greatest number of cancer-related deaths among women. We can predict the occurrence of breast cancer, based on the several factors like tumor cell data, nuclei information, cell division status, physical dimensions, etc. It helps doctors to decide the critical conditions of the patients, which is difficult to classify the tumor to be malignant or a benign one. Neural network is a deep learning algorithm that predicts/classifies the output variable from a set of dependent variables, which can be used in our problem. New Principal Components are developed from the existing data, which removes the bias, noise and inconsistencies. The neural network model is developed to fit the dataset accurately. After building the model, we need to train it with our training dataset, so that it can construct a learned model. After training the model, we test them with the testing dataset to get classification results. Also, the cross-fold validation gives a wonderful insight about the accuracy. The objective of the project is to maximize the accuracy as much as possible. We improve the model based on the error rate and what causes it. By analyzing the user inputs with the trained neural network model, it will classify the nature of the tumor.*

Key Words: Neural networks, PCA.

1. INTRODUCTION

A. Breast Cancer

Breast cancer is the malignant tumor (a tumor with the potential to invade other tissues or spread to other parts of the body) that starts in the cells of the breast. It occurs both in men and women.

One of the main ways breast cancer spreads is through the lymphatic system. Lymph vessels carry a clear fluid called lymph which drains into lymph nodes. Lymph nodes are small bean-shaped structures which contain cells that fight infections (immune system cells). Lymph vessels from the breast drain into the axillary lymph nodes and supraclavicular lymph nodes.

Breast cancer has ranked number one cancer among Indian females with age adjusted rate as high as 25.8 per 100,000 women and mortality 12.7 per 100,000 women. Data reports from various latest national cancer registries were compared for incidence, mortality rates. According to study published in Asia-Pacific Journal of Clinical Oncology, breast was found as high as 41 per 100,000 women for Delhi, followed by Chennai (37.9), Bangalore (34.4) and Thiruvananthapuram District (33.7) in 2017. According to this study number of cases of Breast cancer will become almost double (17,97,900) by 2020.

According to health ministry of India breast cancer ranks as the number one cancer among Indian females with rate as high as 25.8 per 100,000 women and mortality of 12.7 per 100,000 women. India continues to have a low survival rate for breast cancer, with only 66.1% women diagnosed with the disease between 2010 and 2014 surviving, a Lancet study found. The US and Australia had survival rates as high as 90%, according to the study. Globally, about 10% of breast cancer is genetic or due to an inherited DNA mutation. But a recent study suggests that there may be a greater occurrence of genetically-linked breast cancer among Indian women.

To overcome this, A few machine learning techniques will be explored. In this exercise, Support Vector Machine is being implemented with 96% accuracy with the help of Machine learning and deep learning.

ML techniques have been widely used in intelligent healthcare systems, especially for breast cancer (BC) diagnosis and prognosis. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. However, machine learning is not a simple process. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, when you provide a model with an input, you will be given an output. Machine learning techniques are required to improve the accuracy of predictive models.

B. Deep Learning

Deep Learning is a technique for implementing Machine Learning. Deep learning is especially useful to learn patterns from unstructured data. Deep learning complex neural networks are designed to emulate how the human brain works, so computers can be trained to deal with poorly defined abstractions and problems. The average five-year-old child can easily recognize the difference between his teacher's face and the face of the crossing guard. In contrast, the computer must do a lot of work to figure out who is who. Neural networks and deep learning are often used in image recognition, speech, and computer vision applications.

2. Related Work

Various machine learning algorithms are employed in order to derive the accurate solution, but each of them has its own advantages and disadvantages. The algorithm can only change input to output and nothing more, we have to provide the input in the best way, which the algorithm can change them to output with full accuracy.

The clinical results of the tumor analysis data is used for the computation. The data are thoroughly studied and analyzed for this problem-solution. It is highly useful in order to group the patients according to the data.

3. Proposed Work

Deep learning projects are highly iterative; as you progress through the lifecycle, you'll find yourself iterating on a section until reaching a satisfactory level of performance, then proceeding forward to the next task (which may be circling back to an even earlier step). Moreover, a project isn't complete after you ship the first version; you get feedback from real-world interactions and redefine the goals for the next iteration of deployment.

1. Data collection and labeling
2. Model exploration
3. Model refinement
4. Testing and evaluation
5. Model deployment
6. Ongoing model maintenance

3.1. System Architecture

The system base architectural flow is shown here. The dataset is considered and a model is created, trained and to be made to fit with the data, so that it can be ready to classify the user input data.

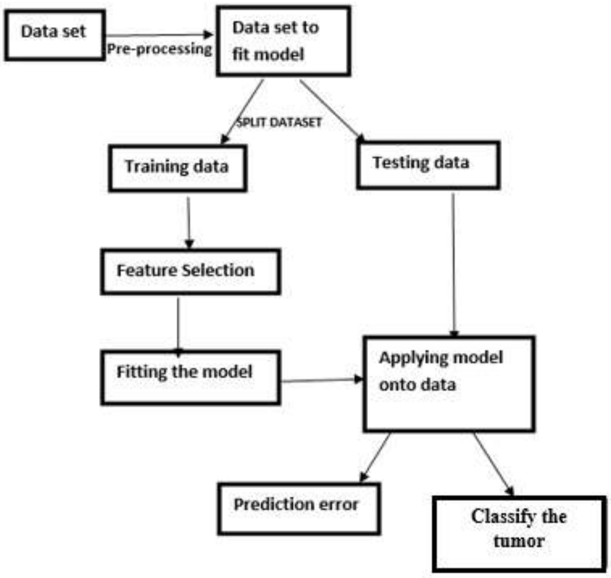

Fig:1 Base Architecture

3.2. Data Preprocessing

Data is typically messy and often consists of missing values, useless values (e.g., NA), outliers, and so on. Prior to modeling and analysis, raw data needs to be parsed, cleaned, transformed, and pre-processed. This is typically referred to a data munging or data wrangling. For missing data, data is often imputed, which is a technique used to fill in, or substitute for missing values, and is very similar conceptually to interpolation.

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.

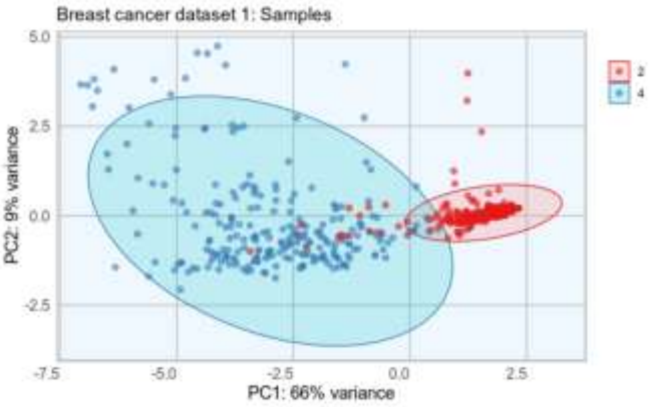

Fig:2 Class diagram

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data.

3.3 Class Diagram

The class diagram shows the overview of the Object-Oriented structure in this project. The data fields are specified in class Patient. The data is being analyzed with the model and the results are provided to the UI.

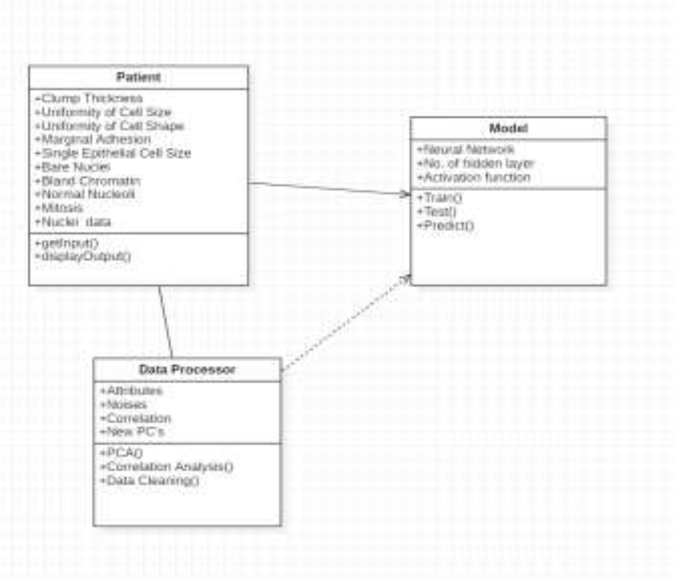

Fig:3 Class diagram

3.2. Dataset

The dataset has the following features.

Attributes	Datatype	Range
Clump Thickness	Numeric	1-10
Uniformity of Cell Size	Numeric	1-10
Uniformity of Cell Shape	Numeric	1-10
Marginal Adhesion	Numeric	1-10
Single Epithelial Cell Size	Numeric	1-10
Bare Nuclei	Numeric	1-10
Bland Chromatin	Numeric	1-10
Normal Nucleoli	Numeric	1-10
Mitoses	Numeric	1-10
Class	Numeric	1-10

The above data set is perfectly designed to remove the problems of feature scaling(normalization). It is taken from the Wisconsin University.

3.4 Model

The data set is studied and visualized to know the hidden proportionalities with the target variable. Also, the missing data were imputed with the **mice** package in r.

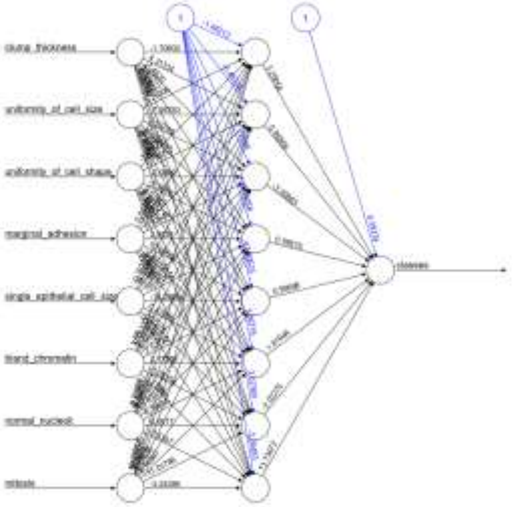

Fig:4 Neural Network model

1. The input layers are defined by the number of attributes that are provided to get the classification results.
2. There is only one hidden layer, since it is a simple neural network and it has 10 nodes in it. The number of hidden nodes is defined based on the below formula.

$$N_h = \frac{N_s}{\alpha^* \, (N_i + N_o)}$$

Ni = number of input neurons.
No = number of output neurons.
Ns = number of samples in training data set.
α = an arbitrary scaling factor usually 2-10.

3. The output layer has single node, which is the best way to classify the binary classification problem. Also, we can have two nodes with other activation function.

3.5 Tuning and Improvement

The flow is a loop, as discussed in the design phase. Once the model is developed, it will be tested to check the accuracy. Then again rigorous data analysis, we will uncover more insights about the data. In this case, as a first step the data set is directly applied to the neural network. Then the feature extraction and selection techniques and data preprocessing were done and applied to the neural network model. As we saw drastic improvements, since the data analysis plays a vital role in deep learning projects.

4. RESULTS

The below screen shows the classification process, it shows the results based on the user input.

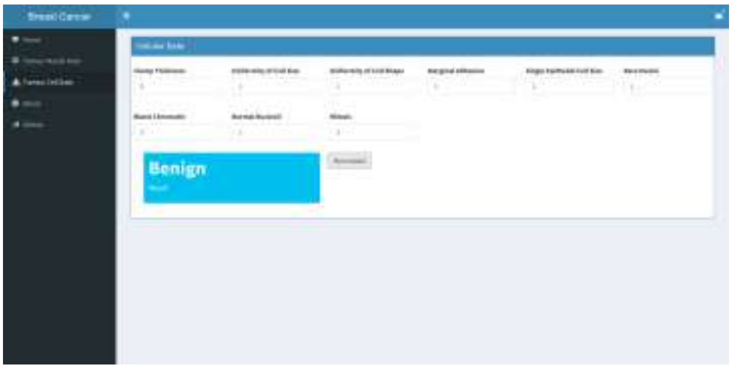

Fig:5 Execution (Sample for Benign Tumor)

It can be either Benign or Malignant. The values are classified using the Logistic function as similar to the Logistic Regression. Here activation function is logistic so it can easily derive the binary classification.

5. CONCLUSIONS

This project helps the patients to clear their doubts on their case, about the tumor whether it is safe or not. It improves the results from the manual method, which is prone to human errors. The model is perfectly balanced, hence won't lie on any subjective feature. In other words, there won't be a biased model, which relies on the single attribute. The error rates have been improved, but not to 100%. There is always a small percent of possibility that it will give wrong results.

The model can be generalized so that it can predict multiple types of cancer. The accuracy can be raised to 100%, so that it can be more accurate than the doctor's diagnosis. It can improve the domain of oncology, the survival rates can be increased by the early detection and advanced treatments. As we know that, the early stage cancer can be easily treated.

6. References

[1] Borges, Lucas Rodrigues, "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection": XI Workshop de Visão Computacional (October 05th-07th, 2015)

[2] G. Ravi Kumar, Dr. G. A. Ramachandra, K. Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques": International Journal of Innovations in Engineering and Technology (IJIET)

[3] Vikas Chaurasia1, Saurabh Pa "Data mining techniques: To resolve breast cancer survivability": International Journal of Computer Science and Mobile Computing

[4] National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data (1973-2008). Cancer Statistics Branch; 2011.

[5] Madhu Kumaria, Vijendra Singh, "Breast Cancer Prediction": IN: International Conference on Computational Intelligence and Data Science (2018)

[6] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians (2005 Jan 1)

[7] Polat K, Güneş S, "Breast cancer diagnosis using least

[8] square support vector machine": Digital Signal Processing(2007 Jul 1)

[9] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques

[10] Delen D, Walker G, Kadam A. "Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine" (2005 June 2)