Machine Learning techniques for Prediction from various Breast Cancer Datasets

Shalini M
Research Scholar, Dept. of Computer Science and Engineering,
Sathayabama Institute of Science and Technology
Chennai, India
shalini.mathi@gmail.com

Dr. S. Radhika
Assistant Professor, Dept. of Electricals and Electronics Engineering
Sathayabama Institute of Science and Technology
Chennai, India
radhikachandru79@gmail.com

***Abstract*— Cancer has become very common disease among Indians. Breast cancer occurs 14% of all cancers in women. One in every 28 women is getting affected by breast cancer [2]. Breast cancer detection and identification are done from gene expression and large datasets. Due to the increase in the generation of large datasets of medical data we can analyze specialized patterns that are hidden inside. To analyze the pattern from datasets machine learning techniques such as SVM, KNN, Decision tree are applied. In this paper, we tried to apply some Deep learning techniques to identify the patterns resides in a reoccurrence of breast cancer datasets, Mammogram datasets and to identify the patterns.**

Keywords — Machine Learning, Data Preprocessing, Decision tree, Random Forest.

I. INTRODUCTION

According to NICPR (National Institute of Cancer Prevention and Research), Cancer is become very common disease among Indians. Around 2.25 million of people are living with the disease and every year 1157294 lakhs people are newly register. Development of Cancer before 75 years for male and female are less than 10% and risk of dying before their 75 years for male is 7.34% and 6.28% for female. In the year 2018, Total death due to cancer in 2018 is 7, 84,821, in which 4, 13,519 male and 3,71,302 female[1].

TABLE I. CANCER TYPES

Top Five Cancer commonly occurring in Men and Women		
	Men	***Women***
1	Lip cancer , Oral Cavity Cancer	Breast Cancer
2	Lung Cancer	Lip cancer, Oral Cavity
3	Stomach Cancer	Cervix Cancer
4	Colorectal Cancer	Lung Cancer
5	Esophagus Cancer	Gastric Cancer
25% of men and women are affected by first two cancer [7]		

Breast cancer occurs 14% of all Cancers in women. From Globocan 2018 data, newly registered cases are 162468 and death due to breast cancer in 2018 is 87090.For Women the Breast cancer is the second most commonly affecting cancers. Even men's are affected by this cancer. Cause of Breast cancer may be due to age factor, BRCA1, BRAC2 and TP53 gene mutation in women's cell, recurrence of Breast cancer, Dense Breast, increase in estrogen due to early stage, obesity after menopause, intake of alcohol, treatments for Hormonal imbalance, etc.,

Breast cancer was ranked in terms of age adjusted rate, Regional wise and mortality. They referred it from PCBR report from (1982 – 2014) [2].

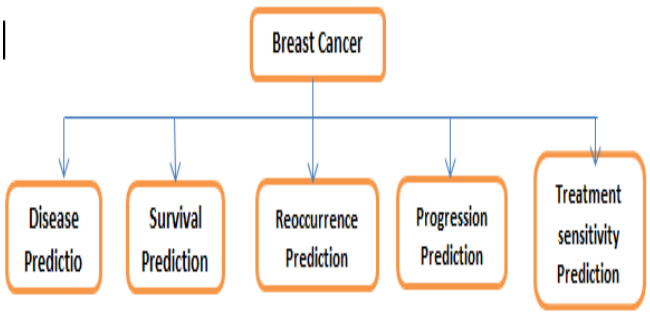
TABLE II. REGIONALLY AFFECTED WOMENS

Region	**Affected Women per 100000 Women**	**Age adjusted rate of PBCRs (1982 to 2014)**
Delhi	41%	1.44%
Chennai	37.9%	2.44%
Bangalore	34.4%	2.84%
Thiruvananthapuram	33.7%	
Breast cancer patients for India during 2020 will be as high as 1797900.		

Breast cancer detection and identification can be done from gene expression and large datasets. In gene expression, BRAC1, BRAC2 and TP53 are gene entail of breast cancer. Hereditary cancer is due to mutation of BRAC1 and BRAC2 gene. Gene mutation predictions are done Blood Sampling extraction, DNA extraction, Sample preparation from DNA or RNA extraction, Sequencing, Analysis of Sequence Data, Identification of any sequence mutation, perform the Cross checking and final result will generated. Disadvantages of gene expression for prediction is won't use to identify at early stage and difficult to differentiate benign from the Malignant tumor.

Machine learning algorithms are used in prediction of Breast cancer. Machine learning is a technique where model are created from given input and predict new input using that model. The Machine learning techniques are categorized into supervised learning, Unsupervised Learning and Reinforcement Learning.

In Supervised learning, the input and what to derive from the input is known formerly. Examples of Supervised learning are classification and regression. In Classification techniques, the given input is classified into given labels (either belongs to malignant or benign). In regression technique, the output derived from model will be a real number (from weight derive a height of a person).



In Unsupervised learning, input is known but don't know the label to classify. Examples of Unsupervised learning are Clustering, Anomaly detection, Neural networks. In Clustering, the given inputs are grouped according to their properties. In Anomaly detection, identification of unusual events occurrence from the given data is done. In neural network, it is a collection of connected nodes and has input layer and Output layer, in between one or more hidden layers to process the classified output.

II. DATA COLLECTION

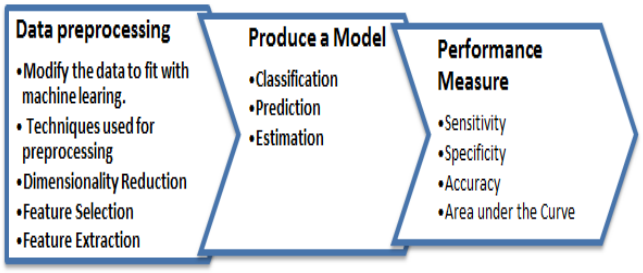
A. Data samples

Datasets for breast cancer are available as open source. Datasets can be collected from TCGA, UCI Machine Learning Repository. Breast Cancer Wisconsin data set of UCI machine Learning repository is available with 10 attributes of 699 instances. In *Wisconsin datasets* focus on information about affected cell structures such as Thickness of cancer cell, consistency of Cell Size and Cell Shape, ticking property, Single Epithelial Cell Size, cell without cytoplasm coverage, rate of Bland Chromatin, check visibility of Nucleoli, Mitoses- cell production activity[10]. In *Mammographic Mass Data Set from* UCI machine Learning repository contains 6 attributes such as mammogram assessment (BI_RADS), age of the patient, shape of the cancer cell, mass margin, mass density, and identify whether benign or malignant of 961 instances are available[10]. In UCI Machine Learning repository, *Breast Cancer Coimbra Data Set* are freely available with 10 attributes such as Age of the patient , Body Mass Index, Glucose , Insulin level ,HOMA, Leptin, Adiponectin, Resistin, MCP-1 of 116 instances[10]. *Breast Tissue datasets* of UCI Machine Learning repository, have 10 attributes of I0 Impedivity ,y, PA500 phase angle(500hz), HFS high-frequency slope of phase angle, DA impedance distance between spectral ends, AREA area under spectrum, A/DA area normalized by DA, MAX IP maximum of the spectrum, DR distance between I0 and real part of the maximum frequency point, P length of the spectral curve, Class carcinoma, fibro-adenoma, mastopathy, glandular, connective, adipose of 106 instances. *Breast Cancer datasets of UCI Machine Learning repository*, have no-recurrence and recurrence of breast cancer was collected. Have 9 attributes such as age of the patient, menopause stage, size of the tumor, auxillary lymph nodes, diffusion of lymph node, degree of malignancy, which breast is affected, Affected breast quadrant, irradiation information of 286 instances.

B. Data Preprocessing

Initially, Data samples are collected with several features and different values. The collected data can have many issues such as noisy data, outliers, missing data, duplicate data and biased data. To overcome these data related issues data preprocessing should be done. The Data cleaning process include remove or reduce noise information and missing data. This missing data can be avoid by deleting that tuples itself, type the missing values, if it is numeric use fill with attribute mean and attribute mean of same class[4].

Data preprocessing techniques such as Dimensionality Reduction, Feature Selection and Feature Extraction, modifies collected data such that it can fit with Machine Learning techniques. Dimensionality reduction rejects unrelated features and reduces noisy data. In feature selection, after dimensionality reduction is done the new features are selected from the old features. Embedded, Filter and wrapper are some techniques used for feature selection. The disadvantage of feature selection possibly will leads to variations in the prediction features. In feature extraction, select the data/ feature that can extract knowledge.

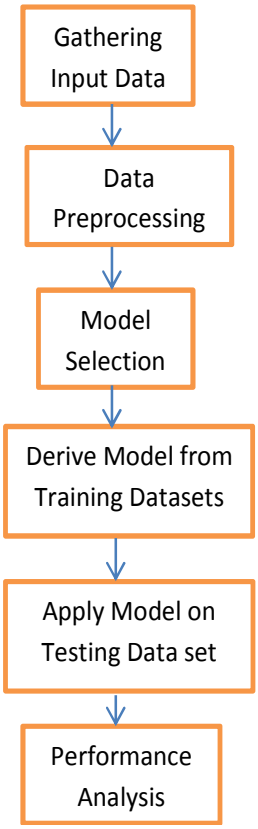


In the recurrence breast cancer dataset there is no missing values and noise.so directly can be used for model prediction.

In the Mammogram dataset, 2 missing values in BI_RADS assessment, 5 missing values in age, 31 in shape, 48 in margin, 76 in density. Missing values are replaced with mean of attribute values of same class.

TABLE III. MISSING VALUE DERIVATION

Class	BI_RADS assessment	Age	Shape	Margin	Density
0	3.957364	49.71318	2.065891	1.943798	2.899225
1	4.782022	61.55955	3.458427	3.723596	2.939326



MODEL FORMATION

A. *Introduction*

After the data preprocessing techniques, derive the model using any machine learning techniques such as classification, prediction and estimation. In Machine learning techniques, before deriving model, the datasets are separated as training sets and test sets. The models are created using training sets and test datasets are used to predict error of derived model. The various methods are used for prediction of Breast Cancer such as Artificial Neural Network, Decision tree, SVM and Bayesian networks.

Deep Learning techniques used for predictions are Unsupervised Pre- Training Networks, Convolutional Neural Networks, Recurrent Neural Networks, and Recursive Neural Networks. CNN used on images to identify Objects, and classify images. RNNs contains feedback loop used to process Languages. The various methods are used for evaluating the performance of a classifier. Some of them are Holdout method, Random sampling, Cross validation and Bootstrapping.

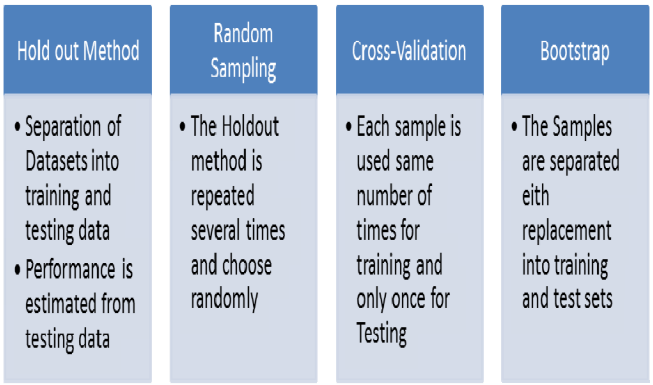


TABLE IV. MACHINE LEARNING TECHNIQUES

Methodology	**Description**	**Advantage / Disadvantage**
ANN	Output is generated from the combination of input layer and hidden layers.	Generic layered structure causes time consuming and poor performance
Decision Tree	A tree structured classification contains nodes(Variables) and leaves (Decision Outcomes)	Easy to infer and quick to learn
SVM	It works on high dimensional feature space and identifies many hyper plane and choose best hyper plane that classify the data points into two classes.	Feasible for smaller datasets and cannot handle larger datasets.
Bayesian Network	Probability estimations rather than prediction	Expensive

B. *Apply Model on Mammogram datasets*

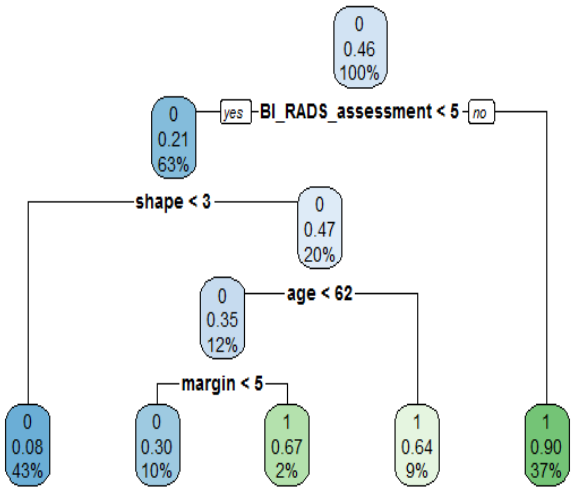
By applying decision tree on mammogram datasets, predict the severity of mammogram based on BI_RADS_assessment, age of the patients, shape of the tumor, margin of the tumor, and density of the tumor. From 961 instances we took 80% of data as training data (786) and rest of them are testing data (193). In this Mammogram datasets, severity of patients not having breast cancer is 53% and severity of patients having breast cancer is 46%.

After applying Decision tree to the mammogram datasets, if BI_rads_assessment is greater than 5 then there is a 37% possiblity of Breast cancer. If BI_rads_assesment is less than 5 and age of that person is greater than 62 then there is 9% of possibilities. Last condition is if BI_rads_assessment is less than 5, age less than 62 but if margin is greater than 5 then 2% possible of breast cancer

TABLE V. CONFUSION MATRIX FOR MAMMOGRAM DATASETS

	False	**True**
False	99	13
True	14	67

From Confusion Matrix, 99 members are exactly identified as not having cancer and 67 members are predicted as cancer patients. 13 patients are wrongly classified as cancer patient. 14 members are wrongly classified as not having cancer. Accuracy of the above DT is 86%.



.

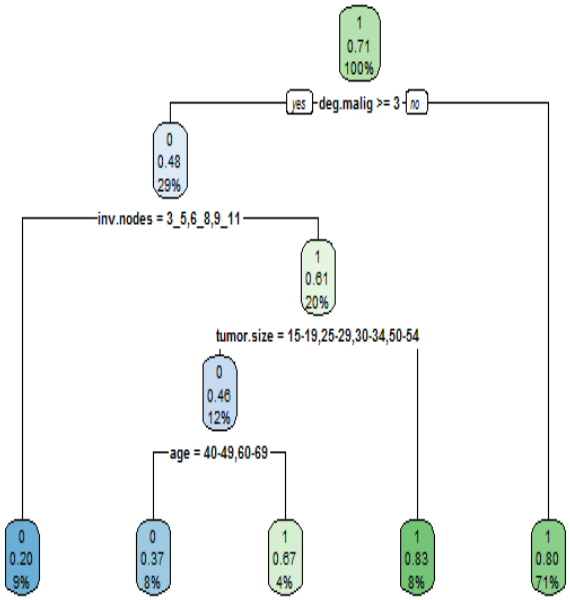
.

Random Forest Regression and Neural Network algorithms are applied on Mammogram datasets using COLAB. Severity in mammogram datasets can be predicted using Random forest algorithm and Neural Network algorithm both the algorithm had 0.41 and 0.38 mean squared prediction error.

```
0.41058297925970183
Train on 672 samples, validate on 288 samples
Epoch 1/10:loss: 0.4656 - mean_squared_error:
0.4656 - val_loss: 0.2690
val_mean_squared_error: 0.2690
Epoch 2/10 :loss: 0.2086 - mean_squared_error:
0.2086 - val_loss: 0.1573 -
val_mean_squared_error: 0.1573
Epoch 3/10 :loss: 0.1700 - mean_squared_error:
0.1700 - val_loss: 0.1580 -
val_mean_squared_error: 0.1580
Epoch 4/10 :loss: 0.1517 - mean_squared_error:
0.1517 - val_loss: 0.1713 -
val_mean_squared_error: 0.1713
Epoch 5/10: loss: 0.1541 - mean_squared_error:
0.1541 - val_loss: 0.1561 -
val_mean_squared_error: 0.1561
Epoch 6/10 :loss: 0.1489 - mean_squared_error:
```

C. *Apply Model on Recurrence datasets*

Decision tree is applied on Recurrence datasets, to predict re-occurrence of breast cancer based on age of the patient, menopause stage, size of the tumor, auxiliary lymph nodes, diffusion of lymph node, degree of malignancy, which breast is affected, Affected breast quadrant, irradiation information. The missing values are altered with frequent values. 80% of random data act as training data and 20 % act as testing data.



After applying Decision tree algorithm on Recurrence Datasets, Patterns derived are if degree of Malign is less than 3 then 71% of patient can have reoccurrence of Breast cancer. Suppose degree of malignant greater than or equal to 3 and inv_nodes are 3-5,6-8,9-11 and tumor size is 15-19,25-29,30-34,50-54 then possibility of recurrence is 8%.

TABLE VI. CONFUSION MATRIX FOR RECURRENCE DATASETS

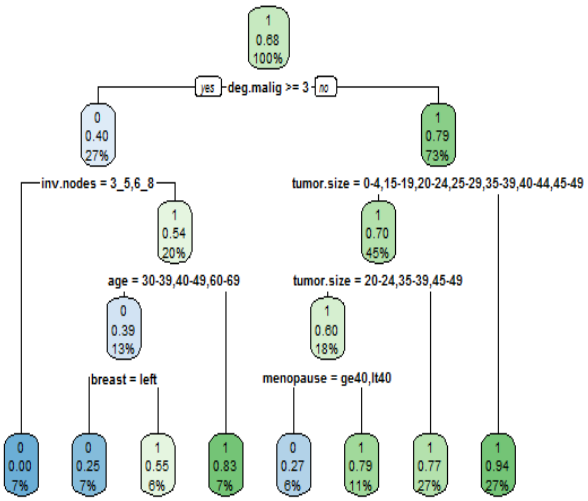
	False	**True**
False	8	11
True	5	34

From Confusion Matrix, 8 members are exactly identified as not having cancer and 34 members are predicted as recurrence of cancer. 8 patients are wrongly classified as cancer patient. 34 members are wrongly classified as not having cancer. Accuracy of the above DT is 72%.

Percentage of training and testing data taken for performing DT are 60% and 40%, in both the cases the common prediction is if degree of malignant is less than 3 then more possible of recurrence of Breast Cancer.

TABLE VII. PERFORMANCE

	Mammogram	**Recurrence**
Sensitivity	83.75%	75%
Specificity	87.61%	61%
Accuracy	86%	72%



III. PERFORMANCE

A. *Performance Measure*

The errors are classified into two such as training error and generalization error. Training error causes misclassification, this can be reduced by increases the complexity of Model because whenever the model complexity increases then training error rate decreases. The generalization error is testing error; this can be reduced using Bias – Variance Decomposition (Bias + Variance). Whenever Training error rate decreases test error rates increases is called as over fitting. Performance of each classification methodology can be measured using sensitivity, specificity, accuracy and Area under the curve (AUC).

TABLE VIII. PERFORMANCE

Methodology	Description
Sensitivity	The Proportion of which are exactly predicted as disease present with correctly observed both positive and negative results
Specificity	The proportion of which no. of patient exactly predicted as not having disease with total no. of exact both positive and negative prediction.
AUC	A measure of the models performance which is based on the ROC curve that plots the tradeoffs between Sensitivity and (1 – Specificity)
Accuracy	A measure related to the total no. of correct prediction

IV. CONCLUSION

There are tremendous amount of medical datasets are available now. Analysis of data can give new pattern that can be used as new clinical traits, drug discovery, early prediction of disease, personalized medicine, side effects of any treatment and change food intake. Gathering of data from health care will leads to some security issues and need some ethical clearance to access those data. After Collection of data, Data preprocessing techniques are used to clean noisy data and avoid redundant data. Machine Learning techniques are used to identify the hidden patterns. Deep Learning techniques will have automatic feature selection. Prediction of Breast cancer can be done using machine learning and deep learning techniques.

[1] Shreshtha Malvia, et.al., "Epidemology of breast cancer in Indian Women: Breast Cancer epidemiology" in Asian-Pacific Journal of Clinical Oncology 2017;13:289 – 295

[2] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.)

[3] M. Elter, R. Schulz-Wendtland and T. Wittenberg (2007), "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process".

[4] Medical Physics 34(11), pp. 4164-4172 I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] Suad. A. Alasadi and Wesam S. Bhaya, " Review of Data preprocessing Techniques in Datamining", Journal of Engineering and Applied Science 12(16); 4102 – 4107, 2017.

[6] ByMichael K. K. Leung, Andrew Delong, Babak Alipanahi, and Brendan J. Frey,"Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets", IEEE. Translations and content mining are permitted for academic research only,

[7] [http://cancerindia.org.in/]

[8] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael, Algorithms and Genome Sequencing: Identifying Driver Pathways in Cancer, © IEEE, 0018-9162/12 (2012), 39-46.

[9] Boiculese LV1, Dimitriu G., Multi-valued logic in breast cancer detection, Rev Med Chir Soc Med Nat Iasi. 2003 Apr-Jun;107(2):425-8.

[10] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science