# DISEASE PREDICTION – BREAST CANCER IN WOMEN USING MACHINE LEARNING CLASSIFICATION TECHNIQUES

**Chaithra S[1], H M Harshitha[2], Kruthika S[3], Namratha N[4], Pradeep K R[5]**
*Department of Computer Science and Engineering,*
*K S Institute of Technology, Bengaluru, India*

***Abstract:*** *Among women worldwide, Breast cancer is one of the most common cancer cases. BC is characterized by the mutation of genes, constant pain, changes in the size, color (redness), skin texture of breasts. The objective is to classify Breast cancer into either Benign or Malignant tumor. Today, Machine Learning (ML) Techniques are being broadly used in the breast cancer classification problem. They provide high classification accuracy. In this paper, we present three different algorithms: Support Vector Machine (SVM), Artificial Neural Networks (ANNs) and Decision Tree for Breast cancer classification. We propose a comparison between the three implementations and evaluate their accuracy.*

***Index Terms-****Breast cancer, Decision tree, Support Vector Machine, Artificial Neural Networks, Diagnosis.*

## 1. INTRODUCTION

Breast Cancer's causes are multifactorial and involves family history, obesity, hormones, radiation therapy and even reproductive factors. Every year, one million women are newly diagnosed with Breast cancer, according to the report of the World Health Organization (WHO) half of them would die, because it's usually late when the doctors detect the cancer. Breast Cancer is caused by typo or mutation in a single cell, which can be shut down by the system or causes a reckless cell division. If the problem is not fixed after few months, masses are formed from cells containing wrong instructions. Malignant tumors expand to the neighboring cells, which can lead to metastatic tumor or reach other parts, whereas benign masses can't expand to other tissues, the expansion is then only limited to benign mass. Many previous studies confirm that detection of breast cancer in early stages significantly increase the chance of survival because it prevents the spreading of malignant cells throughout the entire body. The main contribution of this paper is to review the role of machine learning techniques in early detection of the breast cancer [1].

Artificial Intelligence (AI) can be applied to improve breast cancer detection and diagnosis. Combining AI and Machine Learning (ML) methods enables the prediction and empower accurate decision making. Machine learning is a set of tools utilized for the creation and evaluation of algorithms that facilitate prediction, pattern recognition, and classification. ML is based on four steps: Collecting data, picking the model, training the model, testing the model. The relation between BC and ML is not recent, it had been used for decades to classify tumors and other malignancies, predict sequences of genes responsible of cancer and determine the prognostic. The classification's aim is to put each observation in a category that it belongs to. In this study, we used three machine learning classifiers which are Support Vector Machine(SVM) Classifier, Artificial Neural Network(ANN) and Decision tree Classifier. The purpose is to determine whether a patient has a benign or malignant tumor. In this study, we customize three techniques of machine learning for classification of breast cancer. We use the Wisconsin breast cancer database. The purpose of this article is developing effective machine learning approaches for cancer classification using three classifiers on a data set. The performance of each classifier will be evaluated in terms of accuracy, training process and testing process.

## 2. BACKGROUND

In this section, we first introduce the breast cancer classification, then different machine learning techniques used in our cancer classification.

### 2.1 Breast Cancer Classification

BCC aims to determine the suitable treatment, which can be aggressive or less aggressive, depending on the class of the cancer. To make a good prognostic, breast cancer classification needs nine characteristics which are: 1.Determine the layered structures (Clump Thickness); 2. Evaluate the sample size and its consistency (Uniformity of Cell Size); 3. Estimate the equality of cell shapes and identifies marginal variances, because cancer cells tend to vary in shape (Uniformity of Cell Shape); 4. Cancer cells spread all over the organ and normal cells are connected to each other (Marginal Adhesion); 5. Measure of the uniformity, enlarged epithelial cells are a sign of malignancy (Single Epithelial Cell Size); 6. In benign tumors nuclei is not surrounded by cytoplasm (Bare Nuclei); 7. Describes the nucleus texture, in benign cells it has a uniform shape. The chromatin tends to be coarser in tumors (Bland Chromatin); 8. In normal cells, the nucleolus is usually invisible and very small. In cancer cells, there are more than one nucleoli and it becomes much more prominent, (Normal Nucleoli); 9. Estimate of the number of mitosis that has taken place. The larger the value, the greater is the chance of malignancy (Mitoses). In order to classify BC, pathologists assigned to each of these characteristics a number from 1 to 10. The likelihood of malignancy needs the nine criteria, even if one of them is very large[1].

### 2.2 Machine Learning Approaches

Machine learning is branch of artificial intelligence; ML methods can employ statistics, probabilities, absolute conditionality, Boolean logic, and unconventional optimization strategies to classify patterns or to build prediction models. Machine learning can be divided into two categories: supervised learning (classification) and unsupervised learning. In this section, we will see three supervised learning classifiers.

#### 2.2.1 Decision Tree

Decision trees algorithm consists of two parts: nodes and rules (tests). The basic idea of this algorithm is to draw a flowchart diagram that contains a root node on top. All other (non-leaf) nodes represent a test to a single or multiple attributes until you reach a leaf node (final result). Decision tree algorithms have been widely used in data mining applications due to the fact that they are powerful classification tools [2]. Below are some important reasons that why decision trees are used in the area of data mining and classification:

- Decision trees create understandable rules: They are considered one of the friendliest algorithms to the end user in data mining. They initiate relationships among the dataset attributes in an easy-to-understand form.

- Decision trees provide a clear indication to important attributes: a major part of establishing rules between attributes is indicating the importance level of each one.

- Decision trees require less computation: They require less computation compared to other classification algorithms such as mathematical formulae. When implementing decision trees algorithm to detect breast cancer, leaf nodes are divided into two categories: Benign or Malignant. Rules will be established among the chosen data set attributes in order to determine if the tumor is benign or malignant [3].

#### 2.2.2 Support Vector Machine

A Support Vector Machine is a supervised learning algorithm. An SVM models the data into k categories, performing classification and forming an N-dimensional hyper plane. These models are very similar to neural networks. Consider a dataset of N dimensions. The SVM plots the training data into an N dimensioned space. The training data points are then divided into k different regions depending on their labels by hyper-planes of n different dimensions. After the testing phase is complete, the test points are plotted in the same N dimensioned plane. Depending on which region the points are located in, they are appropriately classified in that region[4,5].

### 2.2.3 Neural Network Model

Artificial neural network algorithm is a supervised learning method for multilayer feed-forward networks from the field of Artificial Neural Networks. Back propagation approach is to model a given function by modifying internal weightings of input signals to produce an expected output signal. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is presented to the system and used to modify its internal state [6,7].

## 3. METHODOLOGY

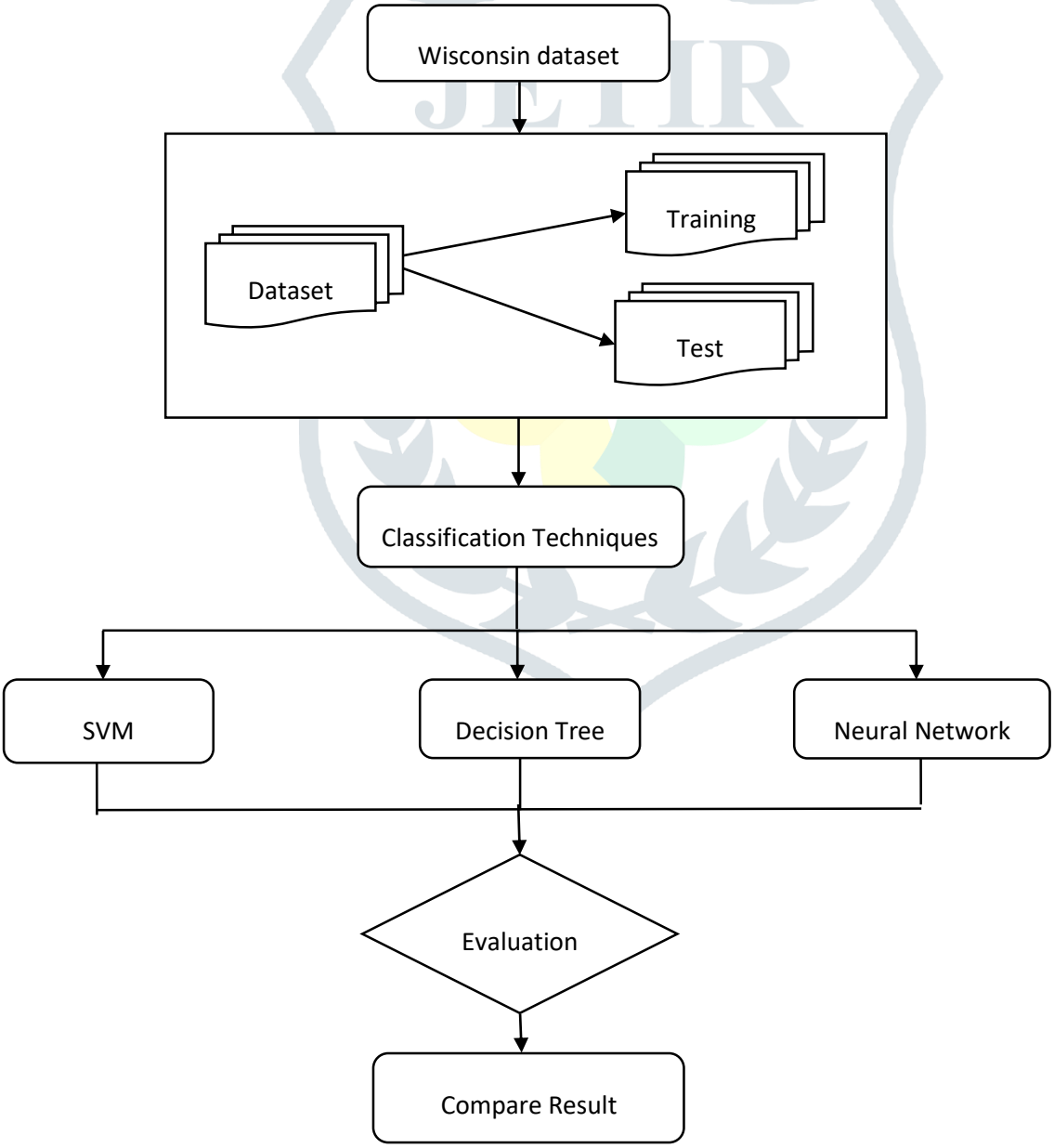


Fig 1: Architecture Diagram

The proposed work is implemented in Python 3.6.4 with libraries Keras, Tensorflow, scikit-learn, pandas, matplotlib and other mandatory libraries. We downloaded Wisconsin dataset from uci.edu. The data downloaded contains 569 instances with two different classes of benign and malignant. The whole dataset is split into train set and test set. Machine learning algorithm is applied such as decision tree and SVM and neural networks.

### 3.1 DATASET

The Breast Cancer Dataset (BCD) that we used is donated to the University of California, Irvine (UCI). There are 11 attributes and the first one is ID that we will remove (it is not a feature we actually want to feed in our classification). The nine criterions are as discussed earlier in breast cancer classification section, they are meant to determine if a tumor is benign or malign, the last feature contains a binary value (2 for benign tumor and 4 for malign tumor).

Table 1: Data set description

| Number | Attribute | Description | Value |
|---|---|---|---|
| 1. | Sample code number | Unique key | ID Number |
| 2. | Clump thickness | Cancerous cell is grouped often in multilayers, while benign cells are grouped in monolayers | (1-10) |
| 3. | Uniformity of cell size | Cancer cells vary in size and shape | (1-10) |
| 4. | Uniformity of cell shape | Cancer cells vary in size and shape | (1-10) |
| 5. | Marginal adhesion | Normal cells tend to stick together, while cancer cells fail to do that | (1-10) |
| 6. | Single epithelial cell size | Epithelial cells that are enlarged may be a malignant cell | (1-10) |
| 7. | Bare nuclei | In benign tumours, nuclei is often not surrounded by the rest of the cell | (1-10) |
| 8. | Bland chromatin | The texture of nucleus in benign cells | (1-10) |
| 9. | Normal nucleoli | Nucleus small structures that are barely visible in normal cell | (1-10) |
| 10. | Mitoses | The process of cell division | (1-10) |
| 11. | Class | Indication of tumour category | 2- Benign<br>4- Malignant |

### 3.2 METHODOLOGY PROPOSED

- Data pre-processing
- Training model
- Prediction

The figure 1 represents architecture of proposed system, in which all modules of the work are represented. User gives input dataset collection, training model and prediction is mentioned.

## 4. RESULTS AND DISCUSSION

The proposed work is implemented in Python 3.6.4 with libraries keras, tensorflow, scikit-learn, pandas, matplotlib and other mandatory libraries. The Wisconsin dataset is considered for study. Machine learning algorithm is applied such as decision tree, SVM and Neural networks. We used these machine learning algorithms and predicted breast cancer as Benign or Malignant. The result shows that breast cancer prediction is efficient using Decision Tree algorithm. Decision Tree achieves 98.5% accuracy, SVM achieves around 99% accuracy, Neural Network achieves 99.5% accuracy.

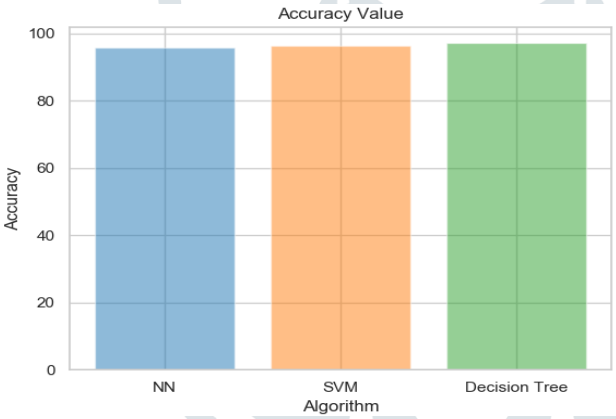


Fig 3:Comparision of Algorithms used

## 5. CONCLUSION

The paper provides a comprehensive review of the state of the art of predictive data field. From the exhaustive review of work carried out during last 10 years, ANN is found to be most widely used predictive technique in medical prediction as compare to traditional methods like Decision Tree, Regression Tree etc. Owing to the fact that ANN technique provides Robust solution to real time prediction problem till date they have invaded almost all the realm of medical prediction process.

## 6. FUTURE SCOPE

In further extended work, we are interested to implement deep learning concepts such as LSTM model or recurrent neural networks. In future work, we are plan for more attribute addition on instead binary classification we are interested in disease severity stages classification.

## 7. ACKNOWLEDGEMENT

The authors thank their management and institution K. S. Institute of Technology for providing them the resources and platform for showcasing their idea. They would also like to thank their HOD and all the lecturers and professors for motivating them into framing this research.

## REFERENCES

[1] L.A. Altonen, R. Saalovra., P. Kristo, F. Canzian, A. Hemminki, Peltomaki P, R. Chadwik, A. De La Chapelle, "Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease", N Engl J Med, Vol. 337, pp. 1481–1487, 1998.

[2] Shrivastava, Shiv, Anjali Sant, and Ramesh Aharwa. "An Overview on Data Mining Approach on Breast Cancer Data." International Journal of Advanced Computer Research (2013): n. pag. Web.

[3] Autsuo Higa Toyohashi University of Technology, Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms, International Journal of Computer Applications Technology and Research Volume 7–Issue 01, 23-27, 2018, ISSN:-2319–8656

[4] Shlomi Laufer and Boris Rubinsky "Tissue Characterization with an Electrical Spectroscopy SVM Classifier" IEEE transactions on biomedical engineering, vol. 56, no. 2, February 2009 pp 525-528.

[5] SUDHIR D. SAWARKAR,GHATOL,Neural network aided Breast Cancer Detection & diagnosis using Support Vector Machine,Proceeding of 7th WSEAS International conference on Neural Network,June 12-14,2006(pp 158-163)

[6] Florin gurunescu, Marina gurunescu, smaranda gurunescu and Elia El-Darzi "a stastical evaluation of neural computing approaches to predict recurrent events in breast cancer", 4th IEEE International Conference on Intelligent Systems, 2008, pp 38-43.

[7] Nuryanti Mohd Salleh , Harsa Amylia Mat Sakim and Nor Hayati Othman "Neural Networks to Evaluate Morphological Features for Breast Cells Classification" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.9, September 2008, pp 51-58.

[8] Anna N. Karahaliou, Ioannis S. Boniatis, Spyros G. Skiadopoulos, Filippos N. Sakellaropoulos, Nikolaos S. Arikidis, Eleni A. Likaki, George S. Panayiotakis, and Lena I. Costaridou "Breast Cancer Diagnosis: Analyzing Texture of Tissue Surrounding Microcalcifications" IEEE transactions on information technology in biomedicine, vol. 12, no. 6, November 2008 , pp 731-738.

[9] Al Mutaz M, Abdalla , Safaai Deris, Nazar Zaki and Doaa M. Ghoneim " Breast Cancer Detection Based on Statistical Textural Features Classification" 2008 IEEE , pp 728-730.

[10] Mohammad Sameti, Rabab Kreidieh Ward, Jacqueline Morgan- Parkes and Branko Palcic "Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer" IEEE journal of selected topics in signal processing, vol. 3, no. 1, February 2009, pp 46-52.