# Review on Comparative Study of Machine Learning Classifiers on Breast Cancer Prediction

**Arvind Lal[1], Nayan Kumar Sinha[2], Menuka Khulal[3], Manzil Gurung[4]**
[1, 2, 3, 4] Dept of Computer Science and Technology
[1, 2, 3, 4] CCCT

***Abstract-*** *In today's world cancer is the most common diseases which lead to greatest number of death. Cancer can involve in any tissue of the body and have many different forms and in each body part. Breast Cancer is a grim disease and it is the only type of cancer that is widespread among women worldwide. If it does not identify in the early-stage then the result will be the death of the patient. It is a common cancer in women worldwide, near about 12% of women affected by breast cancer and the number is still increasing. It manually takes long hours so there is a need to develop an automatic diagnosis system for early detection of breast cancer.*

***Keywords***- WBCD, Support Vector Machine, K-Nearest Neighbor, Random Forest, Adaboost Classifier and XGboost Classifier.

## I. INTRODUCTION

Breast cancer has become one of the most common diseases among women that lead to death. Breast cancer can be diagnosed by classifying tumors. There are two different types of tumors such as malignant and benign tumors. A physician needs a reliable diagnosis procedure to distinguish between these tumors. But generally it is very difficult to distinguish tumors even by the experts. Therefore automation of diagnostic system is needed for diagnosing tumors. As the most prevalent cancer in women, breast cancer has always had a high incidence rate and mortality rate. According to the latest cancer statistics, breast cancer is expected to account for 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide. In case of any sign or symptom, usually people visit doctor immediately, who may refer to an oncologist, if required. The oncologist can diagnose by undertaking thorough medical history, physical examination and also check for the swelling or hardening of any lymph nodes in the breast area or armpit.

## II. MACHINE LEARNING CLASSIFIERS

This is the most important phase where machine learning algorithm selection is done for the developing a system where Data Scientists use various types of Machine Learning algorithms which can be classified as: Supervised, Unsupervised and Reinforcement learning.

2.1 Supervised Learning:

The supervised learning algorithm learns from the training data, which helps you to predict the outcomes for unpredicted data. It helps you to optimize performance criteria using experience also helps you to solve various types of real-world computation problems and such classifiers that are used mostly briefly explained below.

I. Support Vector Machine (SVM)

It is one of the most popularized Supervised Learning algorithm, which is used for Classification as well as Regression problems. However, basically, it is used for Classification problems in Machine Learning scenario. The intent of the SVM algorithm is to create the best decision boundary that can segregate n-dimensional space into classes so that it can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

II. K - Nearest Neighbor (K-NN)

It is one of the simplest Machine Learning algorithms based on Supervised Learning technique. And assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity and easily classified into a well suite category by using K- NN algorithm.

III. Random Forest Classifier

Random Forest classifier is a learning method that operates by constructing multiple decision trees and the final decision is made based on the majority of the trees and is chosen by the random forest. It is a tree-shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, instance, or reaction.Using of Random Forest Algorithm is one of the main advantages is

that it reduces the risk of over fitting and the required training time. Additionally, it also offers a high level of accuracy. It runs efficiently in large databases and produces almost accurate predictions by approximating missing data.

2.2 Unsupervised Learning

In unsupervised learning there would be no labels, only the real time data in the dataset. Here everything are learnt physically so whatever the machine learns, the data will save automatically and create a new dataset and various machine learning algorithms are used in that real time dataset and examine the resultant output accuracy in percentile and also look over which machine learning models is giving the best accuracy and also make it as a good predicted output. Cluster analysis method is used mostly to find hidden patterns or grouping data.

2.3 Reinforcement Learning

Here the machine learning models are used to make a sequence of decisions. It is connected with various software's and machines to get best possible behavior or output. Reinforcement learning is different from the supervised learning as the training data has the answer key with it and the machine learning model have the correct answer itself on the contrary reinforcement learning doesn't have any answer but it decides what do to perform in the given task. In the absence of training dataset, it is necessarily learn from its experience.

## III. MAMMOGRAPHY

It is an x-ray picture of the breast. Mammogram can be used if you have a lump or other sign of breast cancer in women who have no signs or symptoms of the disease. It can help to reduce the number of deaths from breast cancer among women ages of 40 to 70. The radiologist will inspectcarefully in search of highcompact regions or areas of unusual contour that look different from normal tissue. These areas could represent many different types of abnormalities, including cancerous tumors, non-cancerous masses called benign tumors, fibro adenomas, or complex cysts. Radiologists will examine the parameters of an abnormal region, as well as the appearance of the edges or margins of such an area, all of which can indicate the possibility of malignancy (i.e. cancer).

## IV. MATLAB

It is a high-performance language for technical computing. It integrates computation, envision and programming with an easy-to-use environment where every problems and solutions are expressed in mathematical notation. Here the basic data element is an array that does not require dimensioning and also allows you to solve many technical computing problems, especially those with matrix and vector formulations.

## V. DATA MINING

It is the analysis of huge data to discover a meaningful rules and patterns. It analyzes the data patterns in huge collection of data using one or more software also known as Knowledge Discovery in Data (KDD). It will study for the hidden, valid, and potentially useful patterns in huge data sets also aims to predict future outcomes. Now data mining techniques are used to build the Machine Learning (ML) models that boost the Artificial Intelligence (AI) applications such as database marketing, credit risk management, sentiment analysis, fraud detection etc.

## VI. BAYESIAN NETWORK

It is a type of probabilistic graphical model that uses Bayesian inference for probability computation aims to model conditional dependence by edges via directed graph (DAG) model; it is the family of probability distributions that accepts a compact parameterization that can described by it i.e. directed graph. This model captures both conditionally dependent and conditionally independent relationships between random variables and then used for inference to estimate the probabilities for causal or subsequent events.

## VII. THERAGNOSIS

Theragnosis is a treatment strategy for various diseases that combine therapeutic and diagnosis. Aim of sensitive and early detection of disease also increases the efficacy of therapeutic agents. By combining therapeutic and diagnostic ability into one single objective, the new protocol is anticipated to tailor a treatment based on the test results, thereby providing more specific and efficient systems for curing diseases.

## VIII. FINE NEEDLE ASPIRATION

Fine needle aspiration (FNAs) is a type of biopsy procedure. In Fine Needle Aspiration (FNA) Biopsy is a simple procedure that involves passing a thin needle through the skin to sample fluid or tissue from a cyst or solid mass, as with other types of biopsies, the sample collected during fine needle aspiration can help make a diagnosis or rule out conditions such as cancer. It is generally considered a safe procedure. Complications are rare without a biopsy; it's usually hard for a doctor to confirm what these abnormal areas

contain off. The most common reason to do theFNAs is basically to test for cancer. These tests are done on these areas such as breast, thyroid gland, lymph nodes in the neck, or armpit.

## IX. LITERATURE SURVEY

[1] Ch. Shravya, K. Pravalika, ShaikSubhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques International", Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-6, April 2019.

This paper is a relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) is done on the dataset taken from the UCI repository. With respect to the results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm is measured and compared. These techniques are coded in python and executed in Spyder, the Scientific Python Development Environment.

[2] MamtaJadhav[1], ZeelThakkar[2], Prof. Pramila M. Chawan[3], "Breast Cancer Prediction using Supervised Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET)Volume: 06 Issue: 10 Oct 2019.

This system mainly focuses on prediction of breast cancer where it uses different machine learning algorithms for creating models like decision tree, logistic regression, random forest which are applied on pre-processed data which suspects greater accuracy for prediction.

[3] R. Chtihrakkannan, P. Kavitha, T. Mangayarkarasi, R. Karthikeyan, "Breast Cancer Detection using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-11, September 2019.

The main objective of our paper is to enhance an image processing algorithm for earlier finding of breast cancer. X-ray mammogram images which have been acquired are used as input Images. The pre-processing of input images are carried out by applying Gaussian Filter and Edge detection techniques to enhance image quality. Wavelet Transform is useful to identified first order features and GLCM based second order features are extracted from the Pre-processed images. The statistical parameters are then used for classification using DNN a Multilayer supervised classifier. Dataset images are created from the training phase. In testing Phase the acquired image from a patient is given as input to the classifier after completing the image processing steps such as Pre-processing and feature extraction. The output of the classifier consists of two classes, normal and abnormal respectively. The entire algorithm is developed in Python language. The Processing time for testing and conformation of Positive cases is very minimum.

[4] MandeepRana[1], PoojaChandorkar[2], AlishibaDsouza[3], NikahatKazi[4], "Breast Cancer Diagnosis and Recurrence Prediction using Machine Learning techniques", IJRET: International Journal of Research in Engineering and Technology Volume: 04 Issue: 04 Apr-2015.

In this our aim is to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period. To achieve this we have used machine learning techniques such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes. These techniques are coded in MATLAB using UCI machine learning depository.

[5] Varsha J. Gaikwad, "Detection of Breast Cancer in Mammogram using Support Vector Machine", International Journal of Scientific Engineering and Research (IJSER) Volume 3 Issue 2, February 2015.

In this paper we proposed the technique to detect the cancer in the mammogram. The proposed method has been implemented in four stages preprocessing, segmentation for ROI extraction, feature extraction and classification. The proposed method was evaluated with Mammogram Image Analysis Society (MIAS) database.

[6] SusmithaUddaraju[1], M. R. Narasingarao[2], "A Survey of Machine Learning Techniques Applied for Breast Cancer Prediction", International Journal of Pure and Applied Mathematics (IJPAM) Volume 117 No. 19 2017.

This paper focuses on the learning perspectives of different systems, that provides not only an overview of the most common techniques encountered in disease diagnosis, but also manages to classify each paper in terms of solutions in predicting the disease.

[7] RajkamalkaurGrewalBabitaPandey, "Two Level Diagnosis of Breast Cancer Using Data Mining", International Journal of Computer Applications (IJCA) Volume 89 – No 18, March 2014.

In this paper evaluate the performance based on correct and incorrect element of data classification using J48

classification algorithm. The experiment result shows that classification accuracy, sensitivity and specificity of J48 is good.

[8] Priyanka Gupta, Prof. Shalini L, "Analysis of Machine Learning Techniques for Breast Cancer Prediction", International Journal Of Engineering And Computer Science (IJECS) Volume 7 Issue 5 May 2018.

In this paper, machine learning techniques are explored in order to increase the accuracy of diagnosis. Methods such as CART, Random Forest, K-Nearest Neighbors are compared. The dataset used is obtained from UCI Machine Learning Repository. The obtained accuracy prediction performances are proportionate to existing methods. However it is found that KNN algorithm has much better performance than the other techniques used in comparison.

[9] Ravi Aavula, R. Bhramaramba, "An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability", International Journal of Engineering and Advanced Technology (IJEAT) Volume-8 Issue-5, June 2019.

In this paper we propose an Extensible Breast Cancer Prognosis Framework (XBPF) for breast cancer prognosis which includes susceptibility or risk assessment, recurrence or redevelopment of the cancer after resolution, and survivability. We proposed a representative feature subset selection (RFSS) algorithm that is used along with SVM to improve efficiency in prognosis. SEER dataset is used to have experiments.

[10] Dania Abed Aljawad1, Ebtesam Alqahtani2, Ghaidaa AL-Kuhaili3, Nada Qamhan4, Noof Alghamdi5, Saleh Alrashed6, Jamal Alhiyafi7, Sunday O. Olatunji8, "Breast Cancer Surgery Survivability Prediction Using Bayesian Network and Support Vector Machines", 978-1-4673-8765-1/17/$31.00 ©2017 IEEE

In this paper we evaluate the performance of support vector machine (SVM) and Bayesian network (BN) in predicting the survival state of breast cancer patients after having a surgery. The experiments on both techniques have been carried out using Weka software package. Empirical results from simulations showed that support vector machine outperformed Bayesian network in this task.

[11] Mehrdad J. Gangeh, Senior Member, IEEE, Simon Liu, HadiTadayyon, and Gregory J. Czarnota, "Computer Aided Theragnosis Based on TumourVolumetric Information in Breast Cancer", DOI 10.1109/TUFFC.2018.2839714, IEEE.

A computer-assisted technology has recently been proposed for the assessment of therapeutic responses to neoadjuvant chemotherapy in patients with locally advanced breast cancer (LABC). The system, however, extracted features from individual scans in a tumorirrespective of its relation to the other scans of the same patient, ignoring the volumetric information. This study addresses this problem by introducing a novel engineered texton-based method in order to account for volumetric information in the design of textural descriptors to represent tumour scans. A noninvasive computer-aided-theragnosis (CAT) system was developed by employing multi-parametric quantitative ultrasound spectral and backscatter coefficient maps.

[12] Madhuri Gupta1, Bharat Gupta2, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques", 978-1-5386-3452-3/18/$31.00 ©2018 IEEE.

The research work presented an overview of evolve the machine learning techniques in cancer disease by applying learning algorithms on breast cancer Wisconsin data –Linear regression, Random Forest, Multi-layer Perceptron and Decision Trees (DT). The result outcome shows that Multilayer perceptron performs better than other techniques.

[13] AfsanehJalalian, BabakKarasfi, "Machine Learning Techniques for Challenging Tumor Detection and Classification in Breast Cancer", 978-1-7281-2842-9/18/$31.00 ©2018 IEEE.

This paper provide comparison of two supervised classification techniques for angiogenesis detection in computed tomography laser mammography image which is the major sign of high hemoglobin concentration and breast cancer.

[14] U. Karthik Kumar1, M.B. Sai Nikhil2 and K. Sumangali3, "Prediction of Breast Cancer using Voting Classifier Technique", 978-1-5090-5905-8/17/$31.00 ©2017 IEEE.

The main objective of this paper is to compare the results of supervised learning classification algorithms and combination of these algorithms using voting classifier technique. Voting is one of the ensemble approaches where we can combine multiple models for the better classification. The dataset is taken from Wisconsin University database.

[15] Xingyui Li1 (Member, IEEE), Marko Radulovic2, Ksenija Kanjer2, and Konstantinos N. Plataniotis1, "Discriminative Pattern Mining for Breast Cancer

Histopathology Image Classification via Fully Convolutional Auto-encoder ", (Fellow, IEEE)DOI 10.1109/ACCESS.2019.2904245, IEEE Access

In this paper, we propose a practical and self-interpretable invasive cancer diagnosis solution with minimum annotation information; the proposed method mines contrast patterns between normal and malignant images in weak supervised manner and generates a probability map of abnormalities to verify its reasoning.

## REFERENCES


[1] Ch. Shravya, K. Pravalika, ShaikSubhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques International", Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-6, April 2019.

[2] MamtaJadhav[1], ZeelThakkar[2], Prof. Pramila M. Chawan[3], "Breast Cancer Prediction using Supervised Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET)Volume: 06 Issue: 10 Oct 2019.

[3] R. Chtihrakkannan, P. Kavitha, T. Mangayarkarasi, R. Karthikeyan, "Breast Cancer Detection using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-11, September 2019.

[4] MandeepRana[1], PoojaChandorkar[2], AlishibaDsouza[3], NikahatKazi[4], "Breast Cancer Diagnosis and Recurrence Prediction using Machine Learning techniques", IJRET: International Journal of Research in Engineering and Technology Volume: 04 Issue: 04 Apr-2015.

[5] Varsha J. Gaikwad, "Detection of Breast Cancer in Mammogram using Support Vector Machine", International Journal of Scientific Engineering and Research (IJSER) Volume 3 Issue 2, February 2015.

[6] SusmithaUddaraju[1], M. R. Narasingarao[2], "A Survey of Machine Learning Techniques Applied for Breast Cancer Prediction", International Journal of Pure and Applied Mathematics (IJPAM) Volume 117 No. 19 2017.

[7] RajkamalkaurGrewalBabitaPandey, "Two Level Diagnosis of Breast Cancer Using Data Mining", International Journal of Computer Applications (IJCA) Volume 89 – No 18, March 2014

[8] Priyanka Gupta, Prof. Shalini L, "Analysis of Machine Learning Techniques for Breast Cancer Prediction", International Journal Of Engineering And Computer Science (IJECS) Volume 7 Issue 5 May 2018.

[9] Ravi Aavula, R. Bhramaramba, "An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability", International Journal of Engineering and Advanced Technology (IJEAT) Volume-8 Issue-5, June 2019.

[10] Dania Abed Aljawad1, Ebtesam Alqahtani2, Ghaidaa AL-Kuhaili3, Nada Qamhan4, Noof Alghamdi5, Saleh Alrashed6, Jamal Alhiyafi7, Sunday O. Olatunji8, "Breast Cancer Surgery Survivability Prediction Using Bayesian Network and Support Vector Machines", 978-1-4673-8765-1/17/$31.00 ©2017 IEEE.

[11] Mehrdad J. Gangeh, Senior Member, IEEE, Simon Liu, HadiTadayyon, and Gregory J. Czarnota, "Computer Aided Theragnosis Based on TumourVolumetric Information in Breast Cancer", DOI 10.1109/TUFFC.2018.2839714, IEEE.

[12] Madhuri Gupta1, Bharat Gupta2, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques", 978-1-5386-3452-3/18/$31.00 ©2018 IEEE.

[13] AfsanehJalalian, BabakKarasfi, "Machine Learning Techniques for Challenging Tumor Detection and Classification in Breast Cancer", 978-1-7281-2842-9/18/$31.00 ©2018 IEEE.

[14] U. Karthik Kumar1, M.B. Sai Nikhil2 and K. Sumangali3, "Prediction of Breast Cancer using Voting Classifier Technique", 978-1-5090-5905-8/17/$31.00 ©2017 IEEE.

[15] Xingyui Li1 (Member, IEEE), Marko Radulovic2, Ksenija Kanjer2, and Konstantinos N. Plataniotis1, "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Auto-encoder ", (Fellow, IEEE)DOI 10.1109/ACCESS.2019.2904245, IEEE Access.