International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952278

Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction - A Review

Akshya Yadav[1], Imlikumla Jamir[1], Raj Rajeshwari Jain[1], Mayank Sohani[2]

[1]Computer Engineering Department, MPSTME, NMIMS, Shirpur, District: Dhule, Maharashtra, India

[2]Assistant Professor, Computer Engineering Department, MPSTME, NMIMS, Shirpur, District: Dhule, Maharashtra, India

ABSTRACT

Cancer has been characterized as one of the leading diseases that causes death in humans. Breast cancer being a subtype of cancer causes death in one out of every eight women worldwide. The solution to counter this is by conducting early and accurate diagnosis for faster treatment. To achieve such accuracy in a short span of time proves difficult with existing techniques. In this paper, different machine learning algorithms which can be used as tools by physicians for early and effective detection and prediction of cancerous cells have been studied and introduced. The different algorithms introduced here are ANN, DT, Random Forest (RF), Naïve Bayes Classifier (NBC), SVM and KNN. These algorithms are trained with a dataset that contain parameters describing the tumor of a person having breast cancer and are then used to classify and predict whether the cell is cancerous.

Keywords : ANN, DT, Random Forest (RF), Naïve Bayes Classifier (NBC), SVM and KNN

I. INTRODUCTION

One of the major cause of death of women around the world is Breast Cancer [7]. It is one of the most common type of cancer among women, accounting to 15% of all new cancer diagnoses with over 1 million being diagnosed every year.

Breast Cancer can be treated easily with fewer risks if it is detected in early reducing the mortality rate by 25%. The causes of breast cancer can be obesity, hormones, radiation therapy, reproductive factors or even family history [15]. Breast Cancer can be treated easily with fewer risks if it is detected in early reducing the mortality rate by 25%. These days various Machine Learning algorithms are used for diagnosing breast cancer in women because of its accuracy and prediction in the chances of recurrence. Because of ML the accuracy of cancer prediction is being increased by 15-20%. Machine learning usually has 4 four steps when it comes to classification of cancer, collection of data, selection the right model, training of model, and testing [15].

In this paper we have discussed various ML algorithms which can be used for Breast Cancer prediction like, Decision Tree, Support Vector Machine, Artificial neural Network, K-nearest Neighbor etc. The accuracy, precision and recall of all the algorithms are compared to find out which algorithm is most suited for Breast Cancer detection.

II. MACHINE LEARNING APPROACHES

2.1 RANDOM FOREST

The manner in which a jury in the court is assembled for making a decision in the court, this algorithm

collects many decision trees to assemble a forest of trees. RF uses a collection of trees rather than having a single decision tree as it provides either a very simple model or a specific one [9]. Compared to individual decision trees RF has a better stability. This points out that RF is unaffected by the noise present in the input data. RF has the feature of handling data minorities and so is used for cancer detection.

It also facilities anomalies, detection of errors and avoids outliers. It also includes specific outlines for the process of growing and combining trees, self-testing and post-processing.

Random forest algorithm:

The RF algorithm has a recursive approach, in this a sample is picked randomly from the data set whose size is denoted by N and is replaced, another sample is randomly picked from the predictors and is not replaced. The above step is done at every iteration after which the data that has been obtained is divided. The out-of-bag data is then dropped and depending on the number of trees needed the above steps are repeated. Lastly, a count is made on the number of trees that fall under both the categories. Then, classification is done depending on the majority of votes for the decision trees [10].

2.2 NEURAL NETWORK

ANN the abbreviation of artificial neural network is an algorithm that can be described as an information system used for processing. The design and working of ANN have been formulated in such a manner such that is similar to that of human brain. [13]. It detects and discovers relationships and common patterns in the raw data.

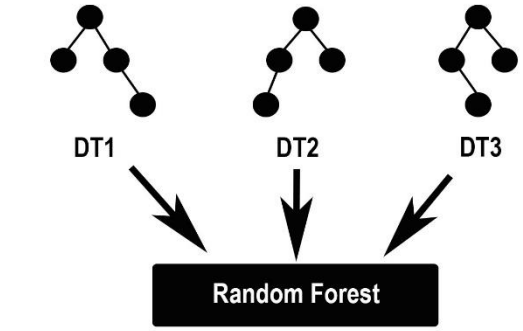

Fig 2.1.1: Diagram for the Random forest algorithm.

The network includes connections which has its respective weights and nodes. Hidden, middle, input and output are the four main layers in the neural network. Each of these layers inside the neural network are connected by a connector called the weight connector. The neural network used in this paper is a multilayer perceptron in which back propagation is mostly used [12]. These networks (Back propagation) contains three layers (input, hidden, output), through which a signal traverses through these layers in one direction in such a manner that it doesn't return back to its source after the signal has travelled to the output neuron from the input neuron.

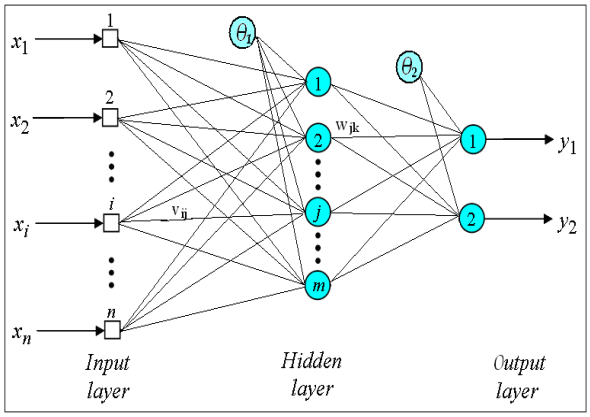

Figure 2.2.1 : Diagram for neural network algorithm.

2.3 SUPPORT VECTOR MACHINE

It is the supervised machine learning classification technique that divides the dataset into classes using a suitable maximal margin hyperplane i.e. the optimized decision boundary. This technique is widely used in the field of medicine for diagnosis of the disease. As dataset may have many such hyperplanes, SVM algorithm performs margin maximization which means that it tries to create maximum gap between different classes [2]. In this dataset we observe the obscurity in classes. A simple line is unable to divide the Wisconsin breast cancer dataset into desired classes. The obscurity is seen in the figure 1.

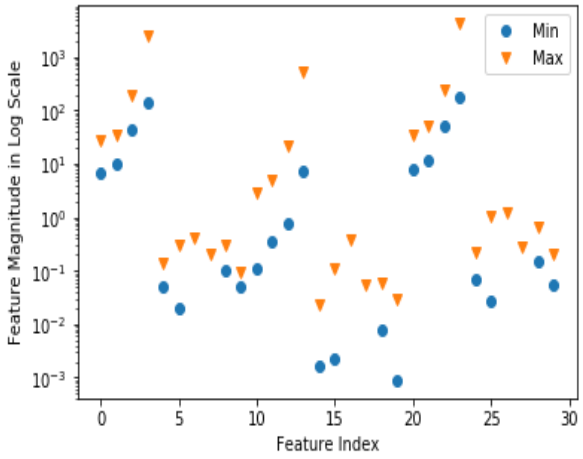

Figure 2.3.1: Class Distribution in Wisconsin Dataset.

To overcome this complexity, we apply transformation and add one more dimension as we call it z-axis. Now if the dataset is plotted in Z-axis, the clear distinction between the classes is clearly visible. This transformation is done using kernels. Polynomial and exponential kernels calculate separation line in higher dimension [3].In the figure 2 given below we can see how kernels have played a major role in obtaining distinction in the obscure dataset.

$$X = (x_1, x_2, x_3, ..., x_n)$$

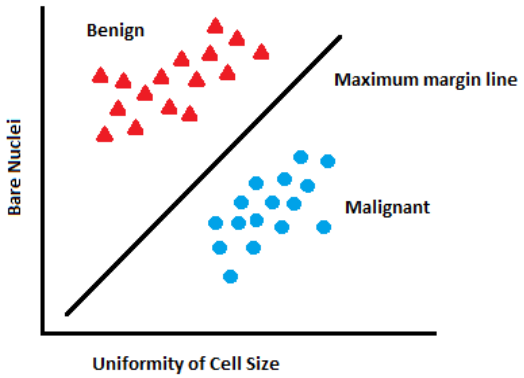

Figure 2.3.2 : classification using kernel.

Due to under fitting the accuracy achieved using SVM was 95.1%. On further processing, by increasing the value of the regularization parameter (C), the accuracy was finally raised up to 97.2%.

2.4 NAIVE BAYES CLASSIFIER

Naïve Bayes is a simple technique which is used to assign class to input instances. Naïve Bayes classifier can be defined as the framework that is used to model the decisions. The variables used in this classifier are conditionally independent [4]. Bayesian classifiers uses Bayes' theorem and is especially preferred when the dimensionality of the input is high. The mathematical equation of Bayes' theorem is given as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (1)$$

Where, y is known as the class variable and X is the evidence i.e. the dependent variable, P(y) is the class' prior probability i.e. probability of event prior evidence, P(X|y) is the likelihood which is the predictor probability, P(y|X) is the class' posterior probability, i.e. event probability post evidence and P(X) is the prior predictor probability.

As said earlier X a dependent feature vector of size n is given as [5]:

Now equation 1 can be expressed as:

$$P(y|x_1,...,x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Which can be further expressed as follows:

$$P(y|x_1,...,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

In the above equation, the denominator can be considered as constant for a given input. After simplifying the equation, it can be deducted that the numerators are directly proportional to each other [5].

Multinomial, Gaussian and Bernoulli are the three different types of Naïve Bayes classifier. The classifier used in this paper to calculate the accuracy using Wisconsin Breast Cancer dataset is the Gaussian Naïve Bayes classifier as the predictors in the dataset hold continuous value.

2.5 KNN

KNN is one of the most central classification techniques in machine language. This classification algorithm is a non-parametric lazy learning algorithm, that means it doesn't make any assumptions on the basis of underlying data [8]. KNN can be used for classification as well as regression. In classification, each test data point will have a K nearest training data point. All the training data points are divided into classes and the most frequently occurring class is assigned as the test data. Therefore, K is used to represent number of training data points lying in proximity. Let us understand KNN algorithm with an example with respect to Breast Cancer diagnosis.

The figure given above represents the KNN structure where K=6. The blue circle in the middle is the test the blue circle in the middle is the test sample, pink square represents malignancy and the green triangle represents benign.

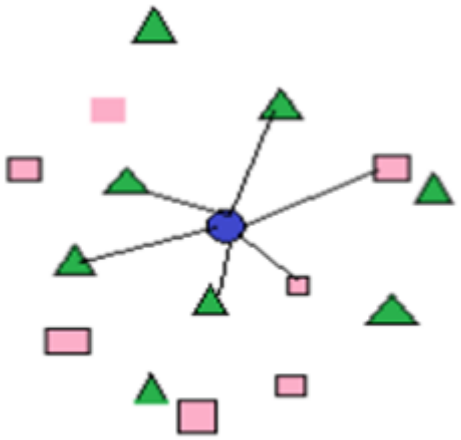

Figure 2.5.1: A 6NN classification algorithm.

2.6 DECISION TREE

It is a predictive model which is used for both classification and regression problems. Decision tree mimics the human level thinking making the understanding and interpretation of data easier.

It is a tree like classifier which classifies all the possible outcomes of data recursively into classes [8]. Every derived subset of a class recursively goes through the same process, this process of partitioning the subset again and again is called recursive partitioning. This process usually ends when the subset of a node has the same value as the target variable, or when any cannot be added to the prediction. The top most node of the tree is called root node. Every node in DT is used for representing a feature, Every link for a decision, and every leaf for an output. It can process both numerical and categorical data. There are various types of DT methods such as, C4.5, and classification and regression tree (CART), Iterative Dichotomiser 3 (ID3).

To understand the working of Decision tree in terms of Breast Cancer diagnosis, there is an example given above of a decision tree structure, which is used to determine whether a person has cancer or not [8].

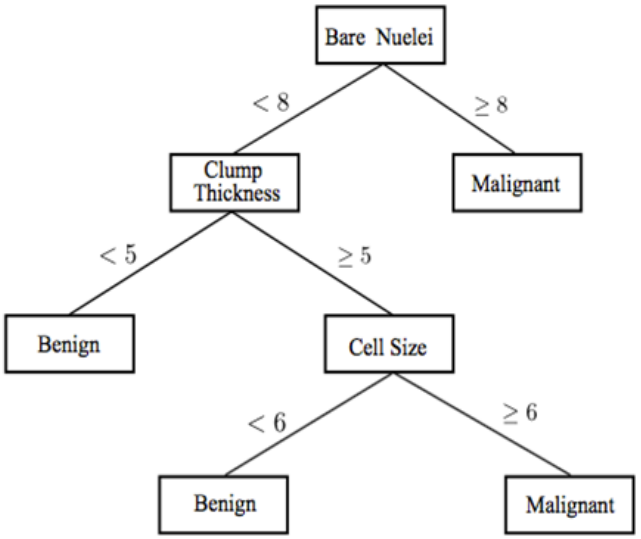

Figure 2.6.1: An example of how Decision tree is used.

III. DATASET

The training data set that has been used in this paper has been taken from the Wisconsin Breast Cancer Data. This dataset is present in the open source repository called the UCI Machine Learning [14]. This data set is multivariate and has over 569 instances. There are over ten features that describe the cell nuclei whose digital image is taken to classify it as malignant or benign. The Ten attributes are:

i) Area
ii) Compactness: (p*p/a-1)

where, p stands for perimeter and a stands for area.

iii) Concave points: number of concave portions of the contour
iv) Concavity: extremity of concave portions of the contour
v) Fractal dimension: it is referred to as approximation of the coastline.
vi) Perimeter
vii) Radius: mean of d, where d is the distance between the center and perimeter.
viii) Smoothness: local disparity in the length of radius.
ix) Symmetry
x) Texture: S.D of gray-scale values, where S.D stands for Standard Deviation.

3.1 SIMULATION SOFTWARE

In this paper, the Anaconda software was used as the machine learning tool. Anaconda is a python based open source tool that was first released to the public in 2012 under the New BSD License. It provides several machine learning algorithms and techniques including the algorithms that are being studied in this paper.

Anaconda is a free and open-source [5] distribution of the Python and R programming languages for scientific computing .The features of this software also include data science, large-scale data processing and predictive analytics.

IV. DISCUSSION

This section describes the parameters and sets forth the result of six different machine learning techniques that are being explored in this paper.

A. Accuracy

Accuracy is also known as the recognition rate. The formula of accuracy is given as the total number of cases identified as correct divided by the total number of cases in the dataset. It should be noted that accuracy can change for different sets and is exceedingly dependent upon the threshold used in the classifier [6].

The accuracy can be calculated as:

Accuracy = (TP+TN)/(T+P) ……..(5)

Where, TP stands for True Positive and TN stands for True Negative. In the similar manner, the full form of P is positive and it signifies the cancerous cells i.e. malignant whereas N stands for Negative and represents the non-cancerous cells i.e. benign.

B. Recall

The definition of recall is given as the rate of True positive instances and False Negative instances that have been classified as True Positive. This measure is used in the medical field as it gives knowledge about the number of cases that are correctly identified either as malignant or as benign. It is the ability of the model to find all the relevant cases in the dataset.

The recall can be calculated using the equation:
Recall = TP/(TP+FN) .....(6)

C. Precision

Precision is also called confidence. The definition of precision is given as the rate of True Positive instances and False Positive instances that have been classified as True Positive. Precision shows the ability of the classifier to handle the positive instances; it does not say anything about the negative instances. Recall and precision share an inversely proportional relationship with each other [7]. This parameter can be calculated using the equation:

Precision = TP/(TP+FP) .....(7)

The evaluation and comparison of the recall, precision and accuracy of the six machine learning techniques using the Wisconsin dataset is shown in the table below:

Table 4.1.1: Comparison of Classification Algorithms Performance on Wisconsin Breast Cancer Dataset

	Accuracy	Precision	Recall
SVM	97.2%	97.0%	97.0%
NBC	95.7%	97.2.%	97.1%
ANN	95.8%	97.0%	97.0%
RF	97.2%	97.0%	97.0%
KNN	93.7%	96.0%	96.0%
DT	93.0%	93.0%	93.0%

V. CONCLUSION

Cancer analysis and treatment techniques have been improved since the last few decades [6]. With advances in Medical sciences, technologies like Machine Learning has been widely used to diagnose cancer and has served physician in analyzing the given data and improving the accuracy in cancer prediction.

This paper represented a brief study of six Machine Learning techniques which are commonly used for breast cancer detection, namely, Artificial Neural Network, Naïve Bayes Classifier, Random Forest, DT, SVM and KNN. The accuracy, precision, and recall of all the six ML techniques were compared. With the help of Wisconsin dataset, we have compared the performance of the aforementioned algorithms. We noticed that SVM and Random Forest achieved the highest accuracy of 97.2%, Naive Bayes Classifier has highest precision and recall of 97.2% & 97.1%.

VI. REFERENCES

[1]. G. Williams, “Descriptive and Predictive Analytics”, Data Min. with Ratt. R Art,Excav. Data Knowl. Discov. Use R, pp. 193-203, 2011.

[2]. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis,“Machine learning applications in cancer prognosis and prediction,” Comput. Struct.,Biotechnol. J., vol. 13, pp. 8-17, 2015.

[3]. K. Kourou etal Computational and Structural Biotechnology Journalxxx (2014).

[4]. B. Networks, F. Faltin, and R. Kenett, “Bayesian Networks,” Encycl. Stat. Qual.,Reliab., vol. 1, no. 1, p. 4, 2007.

[5]. S.Kanta Sarkar, A.N., "Identifying patients at risk of breast cancer through decision trees", International Journal of Advanced Research in Computer Science.,Vol. 08, pp. 88-96, 2017.

[6]. 2014 IEEE 10th International Colloquium on Signal Processing & its Applications,(CSPA2014), 7 - 9 Mac. 2014, Kuala Lumpur, Malaysia

[7]. Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis Dana Bazazeh1 and Raed Shubair 1,2 1Electrical & Computer Engineering Department, Khalifa University, UAE 2Research Laboratory of Electronics, Massachusetts Institute of Technology, USA.

[8]. Wenbin Yue, Zidong Wang, Hongwei Chen, Machine Learning with Application In Breast Cancer Diagnosis and Prognosis, MDPI Journals, Designs 2018

[9]. I. Kononenko, “Machine learning for medical diagnosis: history , state of the art and perspective,” vol. 23, 2001.

[10]. Y. Yasui and X. Wang, Statistical Learning from a Regression Perspec- tive by BERK, R. A., vol. 65, no. 4. 2009.

[11]. Muhammad Sufyian Bin Mohd Azmi,Zaihisma Che Cob,”Breast Cancer Prediction Based On Backpropagation Algorithm ”,Proceedings of 2010 IEEE Student Conference on Research and Development (SCOReD 2010), 13 - 14 Dec 2010,Putrajaya, Malaysia.

[12]. Burke, H.B., Goodman, P.H., Rosen, D.B., Henson, D.E., Weinstein, J.N., Harrell,F.E., Marks, J.R., Winchester, D.P & Bostwick, D.G, “Artificial neutral network improve the accuracy of cancer survival prediction,” Cancer, vol.79, 1997, pp.857-862.

[13]. Caudill M. and Butler C, "Understanding Neural Networks,” Volume 1: Basic Networks, The MIT press, Cambridge, Massachusetts, London, England 1992.

[14]. M. Lichman, UCI Machine Learning Repositry, 2013. Online. Available:https://archive.ics.uci.edu/.

[15]. Meriem Amrane, Saliha Oukid, Breat Cancer Clasification,Using Machine Learning.

[16]. Boulehmi Hela, Mahersia Hela, Hamrouni Kamel, Breast Cancer Detection ,AReview On Mammograms Analysis Techniques, 2013 10th International Multi-Conference on Systems, Signals & Devices (SSD) Hammamet, Tunisia.

Cite this article as :

Akshya Yadav, Imlikumla Jamir, Raj Rajeshwari Jain, Mayank Sohani, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction - A Review", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 979-985, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952278 Journal URL : http://ijsrcseit.com/CSEIT1952278