International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 471-475.
www.theijire.com ***ISSN No: 2582-8746***

Breast Cancer Detection Using Supervised Machine Learning Algorithms

K. Packiyalakshmi[1], A. Karmehala[2]
[1]Department of Computer Science, Sri Kaliswari College (Autonomous), Tamilnadu, India.
[2]Assistant professor, Department of Computer Sccience, Sri Kaliswari College (Autonomous), Tamilnadu, India.

How to cite this paper:
K. Packiyalakshmi[1], A. Karmehala[2], "Breast Cancer Detection Using Supervised Machine Learning Algorithms", IJIRE-V4I02-471-475.

***Abstract*:** *Breast cancer is one of the most widespread diseases among women in the India and worldwide. There is a chance of fifty percent for death in a case as one of two women diagnosed with breast cancer die in the cases of Indian women. This paper compares three of the largely popular machine learning algorithms and techniques commonly used for breast cancer prediction, namely Random Forest, Logistic Regression and kNN (k-Nearest-Neighbor). The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the three machine learning techniques in terms of keyparameters such as accuracy, precision and Recall. The results obtained are very competitive and can be used for detection and treatment.*

Key Word: *Breast Cancer, random forest, k-Nearest- Neighbor, Logistic Regression.*

I.INTRODUCTION

Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but has not spread.

1.1 Types of Breast Cancer

There are several different types of breast cancer, including:

There are multiple types of breast cancers, which are classified, based on how they look under a microscope.

- **Ductal carcinoma.** This is the most common type of breast cancer.
  - **Ductal carcinoma in situ (DCIS).** This is a non-invasive cancer (stage 0) that is located only in the duct and has not spread outside the duct.
  - **Invasive or infiltrating ductal carcinoma.** This is cancer that has spread outside of the ducts or lobules.
- **Invasive lobular carcinoma.** This is a less commonly occurring type of breast cancer that has spread outside of the ducts or lobules.

1.2 Symptoms

- Lump in the breast tissue
- Dimpline of skin
- Skin irritation
- Red scaly patch on skin
- Swollen Lymph Nodes
- Constant pain in breast and armpit area

1.3 Risk Factors

- Obesity
- Age
- Alcohol Consumption
- Hormone Replacement therapy
- Ionizing Radiation
- Having children late or not at all
- History of cancer
- Genetics (BRCA1, BRCA2, HER-2)

1.4 Prevention

- Control the weight
- Get plenty of physical activity
- Breast-feed

- Healthy diet
- Discontinue hormone therapy

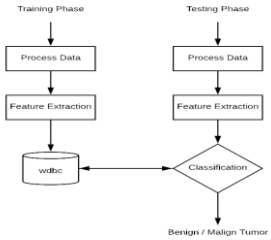

Figure 1: Proposed Breast Cancer Detection Model

II.RELATED WORK IN BREAST CANCER

Breast cancer detection using Machine Learning Algorithms, observed that each of the algorithm had an accuracy of more than 94%, to determine benign tumor or malignant tumor. It is found that kNN is the most effective in detection of the breast cancer as it had the best accuracy, precision and F1 score over the other algorithms.

Breast cancer detection using Relevance Vector Machine[3], obtained an accuracy of 97% using Wisconsin original dataset which has 699 instances and 11 attributes, while [4] allots distinct weights to different attributes with regard to their capabilities of prediction and yielded an accuracy of 92% working with the weighted naïve bayes method. [5] built a hybrid classifier of Support Vector Machines and decision trees in WEKA and obtained an accuracy of 91%.

[6] used Linear Discriminant Analysis for feature selection and trained the dataset by using one of the fuzzy inference method called Mamdani Fuzzy inference model and obtained an accuracy of 93%.

Various differentiation between multiple techniques has been provided through this manuscript [7] like Bayes Network, Pruned Tree, kNN algorithm using WEKA on breast cancer dataset, it has a total of 6291 data and a dimension of 699 rows and 9 columns. The highest accuracy is 89.71% which belongs to bayes network.[11][12][13]

III.MACHINE LEARNING ALGORITHMS

Machine learning(ML) may be defined as a subset of Artificial Intelligence that inculcates the ability of learning into a system on the basis of a data set used for the purposeof training in contrast to the normal approach of coding all possible outcomes beforehand. Multiple approaches andtechniques are present to making systems which can learn. Some of them are neural networks, decision trees and clustering.

A. ML is to be broadly categorised under three categories namely - reinforcement learning, supervised learning and unsupervised learning and.

1) Supervised Learning: generates a function predicting outputs based on input observations. The function is generated from the training data and guides the system to produce useful epiphanies for new data sets introduced to the system.

2) Unsupervised Learning: Learning In this technique, the machine is forced to train from an unlabeled dataset and then differentiating it on the basis of some characters and allowing the algorithm to act on that information without external guidance.

3) Reinforcement Learning: The learning process continues from the environment in an iterative fashion. All possible system states are eventually learned by the system over a prolonged period of time.

B. Random Forest

It is a supervised learning algorithm. An ensemble of decision trees is created, the bagging method is used to trainthe system.

The ground methodology on which this technique is basedis recursion. A random sample of size N is picked from the data set in each instance of an iteration.

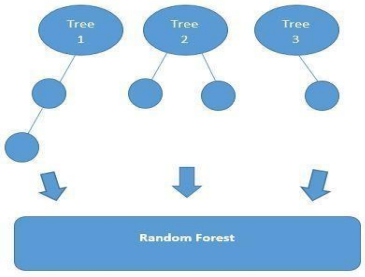

Figure 2: How Random Forest Works

The dataset has been divided into training and testing sets, there are 398 observations for training set and 171 observations for testing. The number of estimators are set to 72 thus it is ensured that every observation is predicted at least a few times.

The confusion matrix of random forest is quite promising. There are only five observations that are misclassified as Benign and four observations are misclassified as Malignant as represented in Table 1 and the accuracy equals 95.3%.

		Predicted	
		Benign	Malignant
Actual	Benign	103	s
	Malignant	4	59

Table 1: Random Forest Confusion Matrix

C. K-Nearest-Neighbor (kNN)

K may be seen as the representation of the data points for training in close proximity to the test data point which we are going to use to find the class. A k-nearest-neighbor may be defined as the algorithm used to determine where a data set belongs to on the basis of the other data sets present around it. The technique is a supervised learning approach used for regression and classification. To process a new data point, KNN gathers all the data points close by to it. Attributes which have a large degree of variation are key factors in determining the distance.

Given N training vectors in the Figure 3, kNN algorithm identifies the k nearest neighbors of regardless of labels.

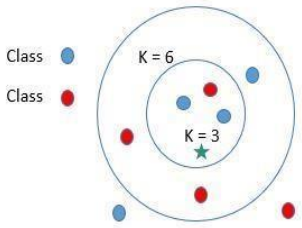

Figure 3: kNN Illustration

The accuracy of kNN is found to be 97.6% , there is eleven observation that is misclassified as Benign and two observations are misclassified as Malignant as represented in Table 2. The results are comparatively better than Random Forest algorithm.

		Predicted	
		Benign	Malignant
Actual	Benign	107	1
	Malignant	6	57

Table 2: kNN Confusion Matrix

D. Logistic Regression

Logistic Regression is one of the most popular Machine Learning algorithms, which comes the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It gives the probabilistic values which lie between 0 and 1. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

There are sixteen misclassified observations, eleven of them being benign and two of them are malignant.

The same 398 observations are used for training set and 171observations for testing and the accuracy equals to 92.3%.

		Predicted	
		Benign	Malignant
Actual	Benign	106	11
	Malignant	2	52

Table 3: Naïve Bayes Confusion Matrix

E. Comparison Among Proposed Algorithms

Each one of the three algorithm's – kNN, Logistic Regression and Random Forest have their advantage and disadvantage over each other in terms of performance, the type of problem they handle etc. As shown in Table 4: kNN test time is O(1)without preprocessing of training set [8], in the case ofNaïve Bayes: N is the number of training examples and d is the dimensionality of the features whereas for Random Forest [9]: N is the number of samples and K is the number of variables randomly drawn at each node.kNN, Logistic Regression and Random Forest can deal with classification as well as regression problems. In terms of accuracy both kNN and Random Forest can deliver high accuracy butLogistic Regression algorithm need large number of records in order to yield a better accuracy. Algorithms that simplifythe function to a known form are called parametric machine learning algorithms, Logistic Regression algorithm can be expressed as parametric as well as non-parametric model.

Parameter	KNN	Logistic Regression	Random Forest
Time Complexity (Training Phase)	O(1)	O(n*m)	0(MKNlog2N)
Problem Type	Classification & Regression	Classification & Regression	Classification & Regression
Accuracy	Provides high accuracy	For high accuracy it needs very large number of records	Provides high accuracy
Model Parameter	Non Parametric	Parametric/Non Parametric	Non Parametric

Table 4: Comparison among kNN, Naïve Bayes andRandom Forest

IV.PROPOSED METHODOLOGY

A. Dataset Description

The project is based on Wisconsin Diagnosis Breast Cancer data set.The data set has been obtained from the 'UCI ML' repo, it has 569 instances and 32 attributes and there are no missing values. The output variable is either benign (357 observations) or malignant (212 observations). The most influential variables are diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean etc. The positive class is used to for benin cases and the negative class is used in malignant cases. The k-fold cross-validationis utilised in which the presented data is divided into k eqully sized bits.

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Diagnosis Breast Cancer(WDBC)	32	569	2

Table 5: Description of WDBC Dataset

B. Performance Metrics

This sections describes the parameters hat are used for measuring performance of machine learning techniques.

A confusion matrix for actual and predicted class is derived comprising of the standard five values namely TruePositive,FalsePositive, TrueNegative and FalseNegative to evaluate the performance.

1. Accuracy

Accuracy is a good predictor for the degree of correctness in the training of the model and how it may perform generally. It may be defined as the measure of the correct prediction in correspondence to the wrong ones. Thus the equation presented can be used to calculate the value ofaccuracy:

$$\textbf{Accuracy} = \frac{\textbf{(Ture Positive + True Negetive)}}{\textbf{(Ture Positive + False Positive + True Negetive + False Negetive)}}$$

2. Recall

Recall known as sensitivity in general terms, may be defined as the ratio of rightfully determined positive instances to the all observations. Recall may be seen as a measure for the effectiveness of the system in predicting positives and determining costs.

$$\textbf{Recall} = \frac{\textbf{True Positive}}{\textbf{(True Positive + False Negetive)}}$$

3. Precision

The degree of correctness in determining the positive outcomes may be defined as precision. It is basically the ratio between true positives and the overall set of positives. This depicts the handling capacity of the system for positive values but does not provide insight into the negative values.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{(TP +FP)}}$$

4. F1 Score

It is the weighted average of Precision and Recall. Thismeasure hence, considers both type of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.

$$\text{F1 Score} = \frac{2*(\text{Precision}*\text{Recall})}{(\text{Precision} + \text{Recall})}$$

V.IMPLEMENTATION AND RESULT ANALYSIS

A comparative study using Random Forest, kNN (k-Nearest-Neighbor) and Logistic Rregression algorithm which are implemented in a computer having configuration as Intel Core i7 with 4GigaBits RAM has been proposed. We have used numpy, pandas and Scikit-learn which are open source machine learning libraries in Python. An open source web application named as Jupyter Notebook is used to run the program.

The classifier was tested using the k−fold cross validation method. We have utilized the 10 fold technique that is the data set segregated in ten different chunks. Nine out of the folds used in the system are used for training and the last set is used for the purposes of testing and analysis. We have utilized 398 observations for training set and 171 observations for testing out of 569 observations. The results presented in Table 6 shows that Random Forest's has the best *precision* performance measure but kNN has the best *accuracy*, Logistic Regression has the best *recall* and *F1 Score* over KNNand Random Forest.

Model Performance (Testing Phase)			
	Random Forest	**kNN**	**Logistic Regression**
Accuracy (%)	95.3	97.6	92.3
Precision (%)	95	92	91
Recall (%)	97	97	98
F1 Score (%)	96	96	94

Table 6: Performance Measure Indices

VI.CONCLUSION

The most frequently occurring type of across cancer is breast cancer. There is a chance of twelve percent for a women picked randomly to be diagnosed with the disease. Thus, early detection of breast cancer can savea lot of valuable life. The proposed model in this paper presents a comparative study of different machine learning algorithms, for the detection of breast cancer. Performance comparison of the machine learning algorithms techniques has been carried out using the Wisconsin Diagnosis Breast Cancer data set. It has been observed that each of thealgorithm had an accuracy of more than 97.6%, to determine benign tumor or malignant tumor. From Table 6, it is found that kNN is the most effective in detection of the breast cancer as it had the best accuracy, precision and F1 score over the other algorithms.

Thus supervised machine learning techniques will be very supportive in early diagnosis and prognosis of a cancer type in cancer research.

References

1. *Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury "Breast Cancer Detection using Machine Learning Algorithms"*
2. *National Institute of Cancer Prevention and Research, cancer statistics [Online], Available: http://cancerindia.org.in/statistics/*
3. *WHO breast cancer statistics [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breastcancer/en/*
4. *B.M.Gayathri and C.P.Sumathi,"Mamdani fuzzy inference system for breast cancer risk detection", 2015.*
5. *Mohd,F.,Thomas,M, "Comparison of different classification techniques using WEKA for Breast cancer" 2007.*
6. *Time complexity and optimality of kNN [Online] Available: https://nlp.stanford.edu/IR-book/html/htmledition/time-complexityand-optimality-of-knn-1.html*
7. *Gilles Louppe, "Understanding Random Forests from theory to practice" 2015.*