Breast Cancer Detection Using Machine Learning Algorithms

[1]Shubham Sharma, [2]Archit Aggarwal, [3]Tanupriya Choudhury
[1]lftshubhamsharma@gmail.com,[2]archit.aggarwal1508@gmail.com, [3]tanupriya1986@gmail.com
[1,3] University of Petroleum & Energy Studies (UPES),Dept. of Informatics, School of Computer Science, Dehradun
[2] Amity University Uttar Pradesh

Abstract: The most frequently occuring cancer among Indian women is breast cancer. There is a chance of fifty percent for fatality in a case as one of two women diagnosed with breast cancer die in the cases of Indian women[1]. This paper aims to present comparison of the largely popular machine learning algorithms and techniques commonly used for breast cancer prediction, namely Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes. The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision. The results obtained are very competitive and can be used for detection and treatment.

***Keywords*— Breast Cancer, random forest, k-Nearest-Neighbor,
naive bayes**

I. Introduction

The most commonly occurring type of cancer is breast cancer. It is known to affect over two million women annually. For women diagnosed during 2010-14, five-year survival for breast cancer shows very heavy variation with changes in location. It is generally known to be above fifty 50% in most places. There are no prevention techniques for breast cancer but early detection and diagnosis is critical in determining the chances of survival.

During the early stages of the disease, the symptoms are not presented well and hence diagnosis is delayed. It is recommended by the NBCF (National Breast Cancer Foundation) that women over the age of forty years of age should get a mammogram once a year. A mammogram is an X-ray of the breast. It is a medical technique used for the detection of breast cancer in women without any side effects deeming the procedure as safe. Women who get regular mammograms have a higher survival rate as compared to women who do not. According to [2] in 2018, over six hundred thousand fatalities were caused by breast cancer. The number is approximately fifteen percent of the total deaths resulting from all types of cancer among women. The chances of contracting this particular type of cancer are usually higher in urban regions, however, the rate of contraction seems to be on an upward rising trend globally. The only current method of improving the results of breast cancer cases is early diagnosis and screening.

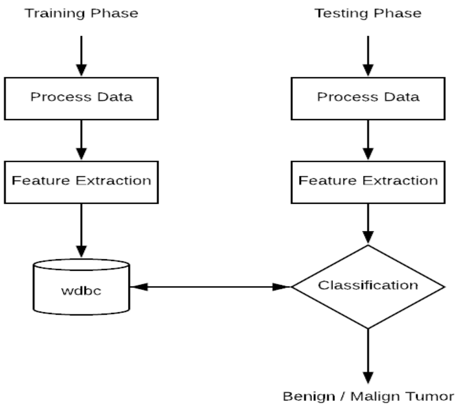


Figure 1: Proposed Breast Cancer Detection Model

II. RELATED WORK IN BREAST CANCER

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 purpose of training in contrast to the normal approach of coding all

[1]

978-1-5386-7709-4/18/$31.00 ©2018 IEEE

possible outcomes before hand. Multiple approaches and techniques 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 train the system.

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

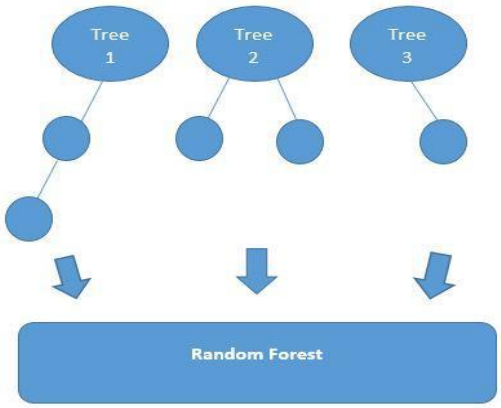
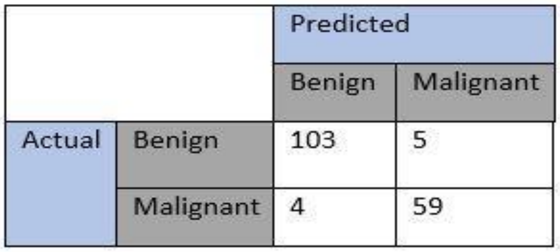


Figure 2: How Random Forest Works

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. It is obvious that diagnosis, radius_mean, texture_mean, perimeter_mean are influential variables, the other variables are of moderate influence but none of them can be neglected to increase the model accuracy.

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 and the accuracy equals 94.74%.



		Predicted	
		Benign	Malignant
Actual	Benign	103	5
	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.

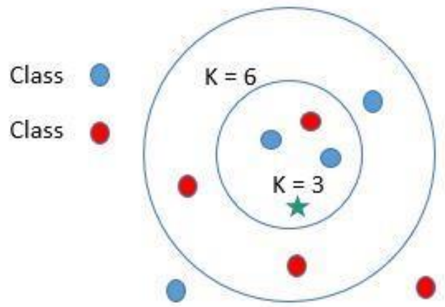
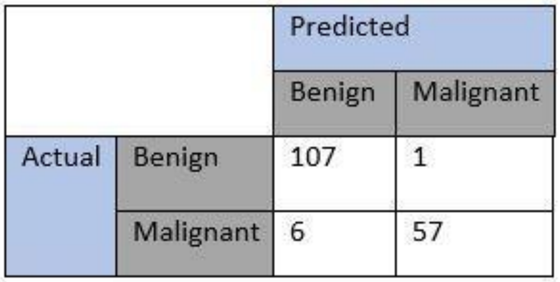


Figure 3: kNN Illustration

The accuracy of kNN is found to be 95.90% , there is only one observation that is misclassified as Benign and four 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

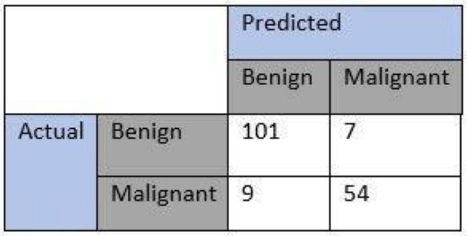
Table 2: kNN Confusion Matrix

D. Naïve Bayes

Classifiers which are probabilistic in nature, based on the application of Bayes theorem may be defined as Naive Bayes classifiers. It is naïve because it assumes that all features are independent from each other, this is generally not the case in real life scenarios, but still Naïve Bayes proves to be efficient for wide variety of machine learning problems.

There are sixteen misclassified observations, seven of them being benign and nine of them are malignant.

The same 398 observations are used for training set and 171 observations for testing and the accuracy equals to 94.47%.

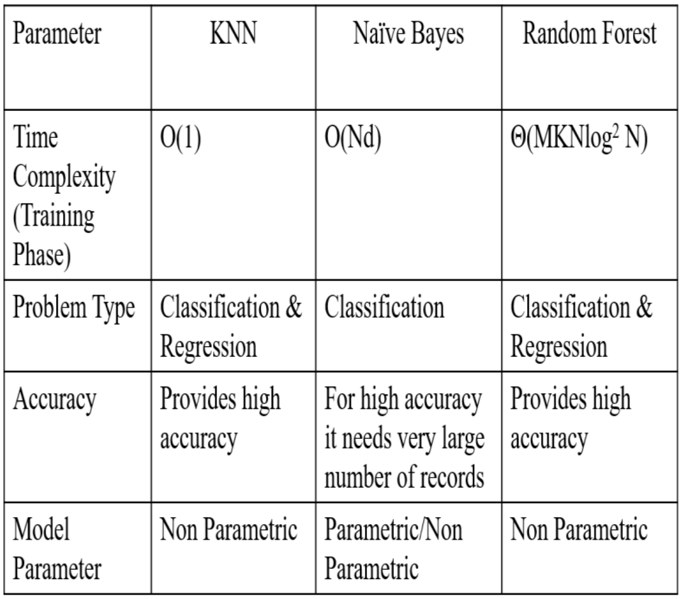


		Predicted	
		Benign	Malignant
Actual	Benign	101	7
	Malignant	9	54

Table 3: Naïve Bayes Confusion Matrix

E. Comparison Among Proposed Algorithms

Each one of the three algorithm's – kNN, Naïve Bayes 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 of Naïve 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. Naïve Bayes algorithm deal only with classification problems whereas both kNN 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 but Naïve Bayes algorithm need large number of records in order to yield a better accuracy. Algorithms that simplify the function to a known form are called parametric machine learning algorithms, Naïve Bayes algorithm can be expressed as parametric as well as non-parametric model.



Parameter	KNN	Naïve Bayes	Random Forest
Time Complexity (Training Phase)	O(1)	O(Nd)	$\Theta(MKN\log^2 N)$
Problem Type	Classification & Regression	Classification	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 and Random 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-validation is 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 that 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 of accuracy:

$$Accuracy = \frac{(TruePositive + TrueNegetive)}{(\text{TruePositive} + \text{FalsePositive} + \text{TrueNegative} + \text{False Negative})}$$

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.

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegetive)}$$

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.

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

4. F1 Score

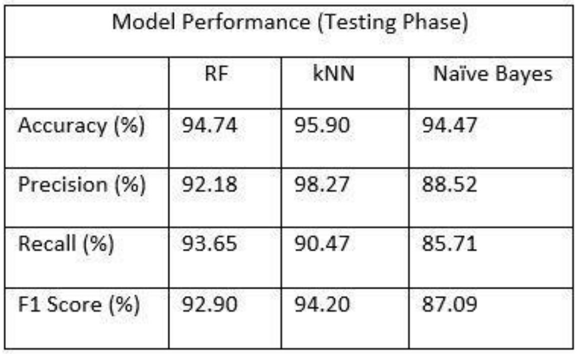
It is the weighted average of Precision and Recall. This measure hence, considers both type of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.

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

V. IMPLEMENTATION AND RESULT ANALYSIS

A comparative study using Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes algorithm which are implemented in a computer having configuration as Intel Core i7 with 16GigaBits 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 graphical representation of the performance metrics for the three illustrated algorithms are shown in Figure 4. The results presented in Table 6 shows that Random Forest's has the best *recall* performance measure but kNN has the best *accuracy, precision* and *F1 Score* over Naïve Bayes and Random Forest.



Model Performance (Testing Phase)			
	RF	kNN	Naïve Bayes
Accuracy (%)	94.74	95.90	94.47
Precision (%)	92.18	98.27	88.52
Recall (%)	93.65	90.47	85.71
F1 Score (%)	92.90	94.20	87.09

Table 6: Performance Measure Indices

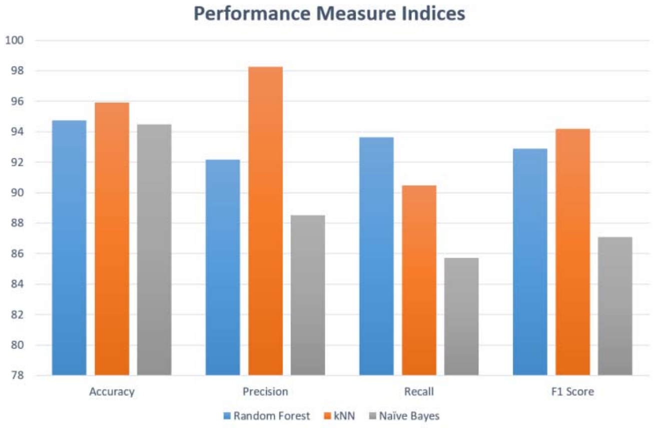


Figure 4: Graphical representation of Performance Measure Indices

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[10]. Thus, early detection of breast cancer can save a 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 the algorithm had an accuracy of more than 94%, 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] National Institute of Cancer Prevention and Research, cancer statistics [Online], Available: http://cancerindia.org.in/statistics/

[2] WHO breast cancer statistics [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[3] B.M. Gayathri, Dr. C.P. Sumathi, “Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer” 2016

[4] S Kharya and S Soni,”Weighted Naïve Bayes classifier –Predictive model for breast cancer detection”, January 2016

[5] Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model" 2015

[6] B.M.Gayathri and C.P.Sumathi,”Mamdani fuzzy inference system for breast cancer risk detection”, 2015.

[7] Mohd,F.,Thomas,M, “Comparison of different classification techniques using WEKA for Breast cancer” 2007.

[8] Time complexity and optimality of kNN [Online] Available: https://nlp.stanford.edu/IR-book/html/htmledition/time-complexity-and-optimality-of-knn-1.html

[9] Gilles Louppe, “Understanding Random Forests from theory to practice” 2015.

[10] U.S. Breast Cancer Statistics [Online] Available: https://www.breastcancer.org/symptoms/understand_bc/statistics

[11] T Choudhury, V Kumar, D Nigam ,An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter Way-International Journal of Computer Science and Mobile Communication (IJCSMC) 2014.

[12] T Choudhury, V Kumar, D Nigam, B Mandal ,Intelligent classification of lung & oral cancer through diverse data mining algorithms, International Conference on Micro-Electronics and Telecommunication Engineering 2016

[13] T Choudhury, V Kumar, D Nigam,Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm - International Journal of Advanced Research in Computer Science and Software Engineering, 2015