International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 5 Issue: 7
ISSN: 2454-4248
09 – 12

A Hybrid Machine Learning Approach for Breast Cancer Detection

Neha Kumari
Department of Computer Science
BIST, Bhopal, India
line 3: City, Country
nehakumariwitdbg@gmail.com

Khusboo Verma
Department of Computer Science
BIST, Bhopal, India
line 3: City, Country
sai.khushbu20@gmail.com

***Abstract*—** Now a day, cancer is one of most common and internecine disease among all disease present in the world. A cancer disease is classified into different types based on the body location like breast cancer, kidney cancer, liver cancer etc. but breast cancer is one of the most common cancer in woman and 8% of woman were diagnosed breast cancer in 2016. It is found that if the cancer is diagnosed in the early stage than the probability of the survival is higher. Now a day, Machine learning play a vital role in order to detect cancer in the early stage. Lots of work has been done previously which uses machine learning approach like support vector machine, Naïve Bayes, logistic regression etc. In this paper w proposed hybrid approach for detecting cancer in the early stage. This hybrid approach is the combination of Support vector machine and Naïve Bayes approach. In order to evaluate the performance of the proposed approach uses Wisconsin Breast Cancer (WBC) which is downloaded from the UCI machine learning repository. Performance of the proposed approach is measured in term of accuracy, F- value. Experiment results shows that proposed approach gives better result as compare to the competitive approach.

Keywords- *Breast cancer, Support Vector Machine, Naïve Bayes Approach, Accuracy, Machine Learning, UCI Repository*

*****

I. INTRODUCTION

As per the survey conducted by World Health Organization (WHO), death due to the cancer is higher as compare to the other disease and breast cancer is one of the common disease in the woman [1, 2]. Previous study shows that if the cancer diagnoses with in the three month then chance of the survival is higher [2,3]. One of the common test use to diagnose cancer is Mammograms. But main problem with this test is the accuracy and normally it has higher false positive rate. Due to high false positive rate this approach unnecessary diagnose the cell. Currently, ML algorithm heavily used in the medical domain in order to diagnose cancer in the early stage. Machine learning approaches can be categorized mainly in three types named as supervised learning, unsupervised learning and reinforcement learning. But cancer detection is a type of supervised learning where we need to classified cancer or not. There are several classifiers are available that can be used to detect cancer but most common classifier are Naïve Bayes classifier and Support vector machine (SVM). The process of supervised learning is shown in figure 1.

II. MACHINE LEARNING APPROACHES USED FOR THE CANCER DETECTION

Now a days use of machine learning approaches increase exponentially with time. There are several machine learning approaches available those can help to predict cancer in early stage. These approach are classified in two different types named supervised and unsupervised. In the case of medical domain one of the prime requirement is the accuracy. It is seem that supervised learning gives more accurate result as compare to the unsupervised learning. Most popular machine learning approaches are discussed below.

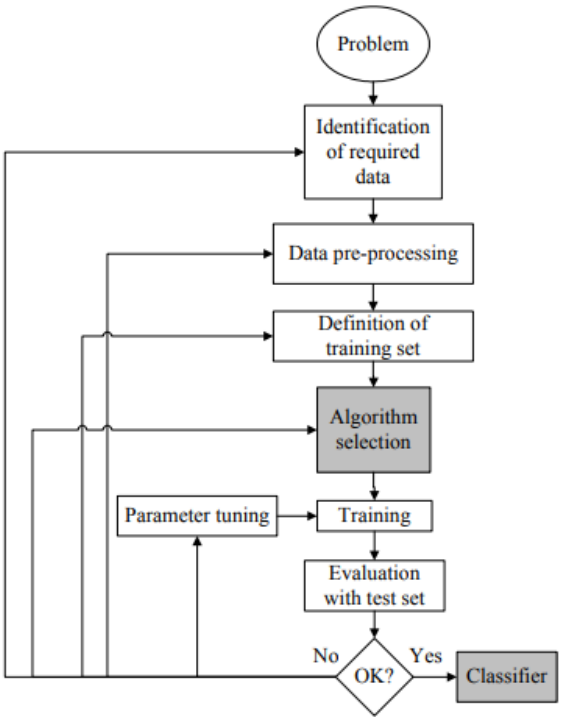

Figure 1. The process of supervised ML

A Bayesian method is a basic result in probabilities and statistics, it can be defined as a framework to model decisions. In NBC, variables are conditionally independent; NBC can be used on data that directly influence each other to determine a model.

 ISSN: 2454-4248
09 – 12

From known training compounds, active (D) and inactive (H). A Naive Bayes classifier is a classifier which is based on Naïve hypothesis. It is a probabilistic model which assume two features are independents. Following equation is used for the Naïve classifier

$$p(c \mid v_1, v_2, ...v_n) = p(c \mid v_1, v_2...v_n) * \frac{p(v_1, v_2...v_n|c)}{p(v_1, v_2...v_n)}$$

Which can be written as

Posterior= prior* $\frac{liklihood}{evidence}$

One of the main advantage of the Naïve classifier is that it can be trained on small data set and gives result very fast. The main limitation of the this approach is that is gives less accuracy.

KNN is a supervised learning method which is used for diagnosing and classifying cancer [4]. In this method, the computer is trained in a specific field and new data is given to it. Additionally, similar data is used by the machine for detecting (K) hence, the machine starts finding KNN for the unknown data. It is recommended to choose a large dataset for training also K value must be an odd number.

Support vector machine (SVM) is a supervised pattern classification model which is used as a training algorithm for learning classification and regression rule from gathered data [5]. The purpose of this method is to separate data until a hyperplane with high minimum distance is found.

III. LITERATURE SURVEY

M. Amrane et al. [6], shows comparative analysis of different machine learning approaches like support vector machine, Naive Bayes (NB) classifier and knearest neighbor (KNN). There are several data sets are available for the breast cancer and they used Wisconsin breast cancer database for analyzing the performance of different machine learning approaches. These approaches are compared in term of accuracy and computation time and claim that KNN classification approach gives better result as compare to the support vector machine and Naïve Bayes.

M. Tahmooresi et al. [7], proposed a machine learning approach for detecting breast cancer in early stage. Most of the existing approach used for cancer detection use mammogram images to classified whether the patient is suffer by cancer or not. Main problem with this image is the false detection of cancer which will be very dangers for the patient. This paper introduced a hybrid machine learning approach which combined different machine learning approaches Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Decision Tree (DT) in order to correctly identified breast cancer. This approach may increase the accuracy but at the same time its required lots of data to train and will increase the training time.

Sweilam et. al. introduced least square support vector machine and active set strategy to show the classification on breast cancer dataset [8]. Khosravi et. al describes Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. This work is mainly focused on several computational methods based on convolutional neural networks(CNN) and build a stand-alone pipeline to effectively classify different histopathology images across different types of cancer [9] .

Hybrid machine learning method was applied by Sahan [10] in diagnosing breast cancer. The method hybridized a fuzzy artificial immune system with k-nearest neighbour algorithm. The hybrid method delivered good accuracy in Wisconsin Breast Cancer Dataset (WBCD). They believe it can also be tested in other breast cancer diagnosis problems.

David B.fogel et al. [11] has discussed the evolving neural networks for detecting breast cancer and the related works used for breast cancer diagnosis using back propagation method with multilayer perceptron. In contrast to back propagation David B.fogel et al. found that evolution computational method and algorithms were used often, outperform more classic optimization techniques.

In 2012, Z.Qinli et al. [12] has presented an article on, a approach to SVM and its application to breast cancer diagnosis. In this article, the authors have proposed a method for improving the performance of SVM classifier by modifying kernel functions. This is based on the differential approximation of metric. The method is to enlarge margin around separating hyper plane by modifying the kernel functions using a positive scalar functions so that the seperability is increased. It is observed that it is competent to reduce the generalization error and computational cost.

Afzan Adam et al., [13] introduced a computerized breast cancer diagnosis by combining genetic algorithm and Back propagation neural network which was developed as faster classifier model to reduce the diagnose time as well as increasing the accuracy in classifying mass in breast to either benign or malignant. In these two different cleaning processes was carried out on the dataset. In Set A, it only eliminated records with missing values, while set B was trained with normal statistical cleaning process to identify any noisy or missing values. At last Set A gave 100% of highest accuracy percentage and set B gave 83.36% of accuracy. Hence the author has concluded that medical data are best kept in its original value as it gives high accuracy percentage as compared to altered data.

 ISSN: 2454-4248
09 – 12

IV. PROPOSED HYBRID APPROACH

The main advantage of the Naïve Bayes approach is that it is suitable for the high dimension data where size of the data set is high. It takes less training time but at the same time result of the Naïve Bayes classifier is not trusted. On the other side SVM gives high accuracy but it takes lots of training time. So in our approach we have proposed hybrid to take benefits of both approach. In our proposed approach first we apply the Naïve Bayes approach and then apply SVM to improve the accuracy of the prediction.

Algorithm for the Hybrid Approach

1) Collect dataset from UCI repository
2) Reduce the dimension of the dataset by removing the un-correlated attributes.
3) Pre-process the data by checking the null value. If any null value is find, then it is replacing by the median of previous and after null value.
4) Normalize the data before feeding in the Machine learning model
5) Apply Naïve Bayes approach
6) Apply SVM to output of the Naïve based approach
7) Tune C and Gamma Parameters of the support vector machine
8) Evaluate the proposed approach

V. PERFORMANCE MEASURES

In order to evaluate the performance of the proposed approach confusion matrix is used. Table 2 shows a confusion matrix for binary classification, where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative counts, respectively.

Table 1. General procedure to solve by confusion matrix for binary classification.

TP	FN
FP	TN

Classification Rate: **Classification** accuracy is the ratio of correct predictions to total predictions made

$$CR = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Sensitivity: It is used for identification of true positive cases

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

Specificity: It is used statistical analysis and also find binary classifier for true negative class

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

F-score is a measure of test accuracy. It considers both precision and there call to compute. These are calculated by

$$\text{Precision} = \frac{True\ Positive}{True\ Positive+False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive+False\ Negative}$$

$$F1 = 2 \times \frac{Precision*Recall}{Precision+Recall}$$

VI. RESULT ANALYSIS

In order to evaluate the performance of the proposed approach Wisconsin Breast Cancer (WBC) dataset has been downloaded from the UCI machine learning repository. It contains 569 attributes and target vector is labeled as either benign or Figure malignant. This dataset has 30 features and divided into two classes named 2 and 4 where 2 represents benign class and 4 represents malignant. Figure 2 shows the distribution of the different class. Here 0 represents benign and 1 represents malignant stage of the cancer.

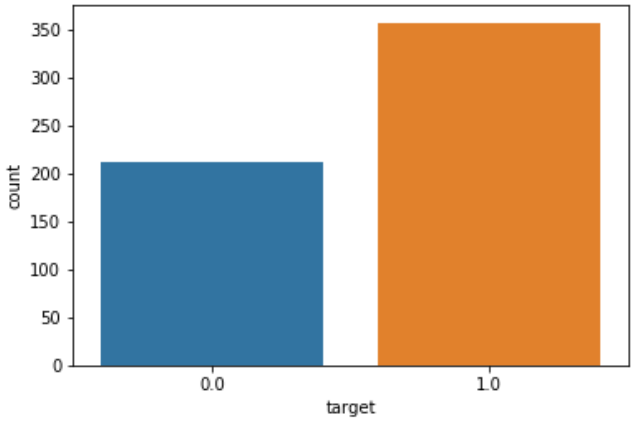

Figure 2: Distribution of Target Variable

To predict the accuracy of any classifier confusion table is used. Figure 3 shows the confusion table of the proposed approach.

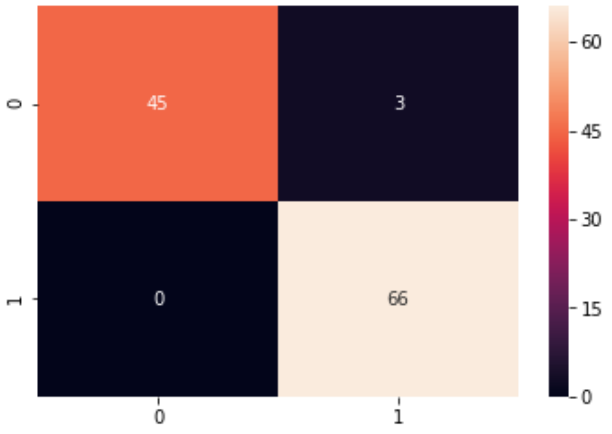

Figure 3: confusion Metrix

 ISSN: 2454-4248
09 – 12

To analyse the performance of the proposed approach it is compare with the Naïve Bayes classifier, Support vector Machine and Logistic regression in term of precision, recall, F1 score, accuracy and time required to train the machine learning model.

	Naïve Bayes	SVM	Logistic Regression	Proposed Approach
Precision	95	29	97	98
Recall	94.5	50	97	98
F1 Score	95	37	97	98
Support	57	57	57	57
Accuracy	94	57	96	98.9
Training Time (in Second)	.01	0.06	0.25	0.18

VII. CONCLUSION

Cancer is one of the serious diseases in all over world. In most of the cases cancer patient die. Main cause of this the late detection of the cancer. This paper discussed some machine learning approach that is used to detect cancer. Most of the work done in this domain says that accuracy is more important as compare to the diagnose time. A wrong prediction of diseases is more dangerous than time required to train machine learning model. In this paper we proposed a hybrid machine learning approach that combine two machine learning approach name support vector machine and Naïve Bayes to improve accuracy. Experiment results say that our proposed approach more accuracy as compare to the existing approach.

REFERENCES

[1] http://www.breastcancerindia.net/
[2] A. K. Dubey et al., "A Survey on Breast Cancer Scenario and Prediction Strategy", Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 pp 367-375.
[3] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, ―Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,‖Procedia Comput. Sci., vol. 83, no., pp. 1064–1069, 2016.
[4] M. F. Akay, ―Support vector machines combined with feature selection for breast cancer 50 diagnosis,‖ Expert Syst. Appl., 2009.
[5] E. D. Übeyli, ―Implementing automated diagnostic systems for breast cancer detection,‖ Expert Syst. Appl., 2007.
[6] M. Amrane et al., "Breast Cancer Classification Using Machine Learning", **IEEE conference on Electric Electronics, Computer Science, Biomedical Engineerings, pp. 1-4 , 2018.**
[7] M. Tahmooresi et al., "Early Detection of Breast Cancer Using Machine Learning Techniques", **Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, No. 3, 2018.**
[8] N. H. Sweilam, A. A. Tharwat, and N. K. A. Moniem, ―Support vector machine for diagnosis cancer disease : A comparative study,‖ Egypt. Informatics J., vol. 11, no. 2, pp. 81–92, 2010.
[9] P. Khosravi, E. Kazemi, and M. Imielinski, ―EBioMedicine Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images,‖ EBioMedicine, vol. 27, pp. 317–328, 2018.
[10] S. Şahan, K. Polat, H. Kodaz, and S. Güneş, ―A new hybrid method based on fuzzy artificial immune system and k-nn algorithm for breast cancer diagnosis,‖ Comput. Biol. Med., 2007.
[11] K. I. Al-Sulaiti et al., ―Research Methods for Business Students,‖ Int. Mark. Rev., 2010.
[12] W. Shitong, ―A Novel SVM and Its Application to Breast Cancer Diagnosis,‖ 2007.
[13] A. Adam and K. Omar, ―Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network.