International Journal on Emerging Technologies (Special Issue NCETST-2017) **8**(1): 522-526(2017)
(Published by Research Trend, Website: www.researchtrend.net*)*

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer

Arpita Joshi *and* ***Dr. Ashish Mehta***

Department of Computer Science, S.S.J. Campus , Kumaun University, Almora, (U.K.), INDIA

ABSTRACT: Breast cancer is the most common reason of cancer death. Early detection and diagnosis is very important in the treatment of breast cancer. There are various Machine Learning techniques available for the purpose of diagnosis of Breast Cancer data. In this study, we have compared the classification results of different techniques i.e. Random Forest, Support Vector Machine (SVM),K Nearest Neighbor(KNN) and Decision Tree for classifying Breast Cancer dataset. We took Wisconsin Breast Cancer dataset available from UCI repository containing 699 instances. We compared the classification results obtained from these techniques i.e. KNN, SVM, Random Forest, Decision Tree (Recursive Partitioning and Conditional Inference Tree). The performance of each technique is evaluated using various performance measures. The classification results show that KNN gives better result

Keywords: Machine learning techniques, dataset.

I. INTRODUCTION

Cancer has been the most perilous disease ever. In the modern lifestyle, cancer has crept in the sound health of the world and making it vulnerable. The only chance to fight with it is to detect it in its early stage. Medical science can confront with cancer potently and triumph it if cancer is detected at its starting. Breast cancer is the top most common reason of cancer death. When abnormal cells in human body is divided into an uncontrolled way, it becomes cancer. There are many types of cancer occurs such as lung cancer, breast cancer, colon cancer etc. Breast cancer occurs when malignant tumour, which contain cancer cells form in the breast tissue and It can come about at any age. In 2012, there were 1.7 million cases of breast cancer diagnosed in world. It account for 12% of cancer cases and 25% of all cancers in women [25].The breast cancer occurs in almost all areas of the world. The disease occurs almost entirely in women, but men can get it, too. In 2015 ,estimated 234190 new cases (23,1840 women,2350 men) of invasive breast cancer are expected to be diagnosed and estimated 40,730 breast cancer deaths (40,290 women, 440 men) are expected in 2015 in US [23]. In breast cancer's earlier stage mammography technique has been very effective tool for the detection of breast cancer [11]. Early detection and diagnosis is very important in the treatment of breast cancer. So for diagnosis purpose there are various Machine Learning techniques are available. Machine Learning is an idea that hails under counterfeit consciousness empowers the framework with take starting with those existing information to accomplishing different assignments further more assessing execution[3].Machine Learning in keeps tabs on the configuration What's more advancement for situated for guidelines that could show themselves should develop and change when presented should new information. The principle reason for existing from claiming machine learning in is will produce a perfect model which will be utilized for performing different assignments. For example, such that classification, predication etc. There are different types of Machine Learning techniques like Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), Random Forest, Extreme Learning Machine and Decision Trees (DTs) etc would extensively connected done malignancy examine to the improvement about predictive models that might give powerful furthermore dependable choice making[15].The most common task in learning process is classification. The main task of classification is to classify the data into definite classes. For data classification, the classification algorithm should fit the training pattern properly and classify all features accurately [15]. There are several factors affecting the performance of Machine Learning techniques. These factors include various kinds of training provided by the Machine Learning technique, initial background knowledge that influences the data, feedback, various Machine Learning tech-

niques that extract relevant, useful, important information from training set [6].

In the present study, we will proposed an appropriate model for the classification of malignant and benign breast tumour that may be used for the diagnosis of breast cancer. The present study is organised as follows: Section 2 gives brief review of available literature.. The proposed system model is presented in section 3. Section 4 presents experimental results. Finally, concluding remarks are also drawn in section 4.

II. LITERATURE REVIEW

Baoyu Zheng, Wei Qian, and Laurence P. Clarke employed Back Propagation method with kalman filtering. They used Mixed Feature Neural Network (MFNN) technique They calculated the True Positive rate (TP) and False Positive rate(FP) for both raw(TP-.814,FP-.59) and enhanced images(TP-.901,FP-.71) with the help of MFNN.They experimented on 30 mammogram images[5].Chen Ren Dar,Chang Feng Ruey and Huang Len Yu reached 85.6% classification accuracy using 10 fold cross validation with self organizing map(neural network model) method for sonographic of breast masses cancer[7].

Hussein A. Abbass proposed evolutionary artificial neural network approach which was based on pareto-differential evolution algorithm. He used multi-propagation artificial neural network that gives better performance (98.12%).The author achieved better generalization and lower computation cost[13].Ubeyli Derya Elif compared the classification accuracies of Support Vector Machine (99.54%)(SVM) Combined (CNN) Neural network(98.15%),Probabilistic Neural Network (98.15%),Recurrent Neural Network(98.61%)and Multilayer Perceptron Neural Network (91.92%)methods and they found that Support Vector Machine outperforms other techniques. He performed an experiment on Wisconsin breast cancer data set [22].

Jihong Liu and Weina MaIn experimented on 120 breast cancer image using Support Vector Machine (SVM) and Back Propagation Neural Network.They resulted that SVM (96.12%) gives better result in comparison to Back Propagation Neural Network (93.33%)[14]. Dheeba J. and Tamil Selvi S. experimented on mammographic data with left and right breast images of 161 patients. The accuracy obtained by authors was 86.1% using Support Vector Machine with Gaussian Radial Basis Function kernel [9].

Hussain Muhammad, Wajid Kanwal Summrina, Ali Elzaart and Berbar Mohammed obtained and compared classification results of Support Vector Machine with different kernel functions(Polynomial, Radial basis function, Sigmoid, Mahalanobis)and Multi Layer Perceptron(MLP)method and discovered the affect of kernel function with feature subsets selected using genetic algorithm[12].They performed experiment on WBC dataset from Machine Learning Respository and result showed that Support Vector Machine with kernel is superior to MLP.Omar S. Soliman and Aboul Ella Hassanien combined Support Vector Machine with moment-based feature extraction obtaining highest classification accuracy (98.1%).They experimented on Ultra Sound Breast Cancer imaging[18].

Aboul Ella Hassanien, Nashwa El-Bendary, Milos Kudelka and Vaclav Snasel performed experiment on 120 real MRI images. They obtained 98% accuracy with Support Vector Machine and compared the result with existing rough set[1].Muhammad Shoaib B. Sehgal, Iqbal Gondal and Laurence Dooley found that Generalized Regression Neural Network performed better than Support Vector Machine in classification of genetic data[17].

Bao C. Q. Truong, H. D. Tuan, Anthony J. Fitzgerald and Incent P. Wallace used Bayesian Neural Network and Support Vector Machine for breast cancer Classification with Leave One Out Cross Validation (LOOCV) and Repeated Random Sub Sampling (RSS). They resulted that the best LOOCV was enhanced to 97.3% with the four-parameter combinations as compared to 93.2% with the Support Vector Machine [4].

III. PROPOSED SYSTEM MODEL

There are various Machine Learning techniques such as Decision Tree, Random Forest, KNN, SVM are used for the classification of breast cancer into benign and malignant with the higher accuracy and efficiency. In this study we use SVM, Decision Tree, Random Forest and K-Nearest Neighbor for the classification of breast cancer into benign and malignant. For this purpose R programming language is used. R is used to extract instances from a large data set, to create statistical software, graphics and data analysis. R is more user friendly as compared to other languages.

A. Proposed System Model

The data set available in the UCI Machine Learning Repository (Wisconsin Breast Cancer (WBC) will be used for this study. The data will be preprocessed and subsequently different Machine Learning (ML) algorithms such as Support Vector Machine, Decision Tree, KNN and Random Forest will be employed on the data comprising different features associated with the diagnosis of breast cancer. In order to evaluate the performance of Machine Learning techniques various performance measures like Accuracy, Precision, Sensitivity, Specificity are used. An appropriate model for the classification will be proposed and validated..There are various steps involved in proposed system model.

Data Collection. The data set (Wisconsin Breast Cancer) will be collected from UCI Machine Learning Repository to differentiate between malignant (cancerous) instances and benign (non-cancerous) instances. A brief description of the dataset is presented in table given as:

Dataset	Attributes	Instances	classes
Breast Cancer	11	699	2

Preprocessing. Firstly, we shall load the mlbench package using R tool, which contains Breast Cancer data and then load the data (Breast cancer). The data will be preprocessed to improve the quality of data. The raw data has many missing values. Missing value is a common property for large data sets. Some algorithms do not like missing values, so we can remove rows with missing values.

Machine Learning Techniques

(i) Support Vector Machine

Support vector machine (SVM) is supervised Machine Learning technique that is based on the concepts of decision planes that define decision boundaries between the data points of classes in high dimensional space. A decision plane is one that separates between a set of objects having different class memberships. SVM support both regression and classification tasks and handle multiple continuous and categorical variables. SVM provides flexibility in selecting a similarity function. It gives sparseness of solution when dealing with large data sets and having ability to handle large feature spaces. In SVM, the training involves the minimization of the error function:

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i$$

subject to the constraints:

$$y_i\left(w^T \phi(x_i)+b\right) \geq 1-\xi_i \text{ and } \xi_i \geq 0, i=1,\ldots,N$$

Where C is the capacity constant, w is the vector of coefficients, b is a constant, and ξ_i represents parameters for handling inputs .The index i labels the N training cases. $y \in \pm 1$ represents the class labels and xi represents the independent variables. The kernel ϕ is used to transform data from the input to the feature space [3].

In this study, to classify the breast cancer in to benign and malignant we will use the **tune** function to do a grid search over the supplied parameter (cost =1, gamma=0.01234568), using the train set. Number of support vectors are 174, SVM-Type is C- classification and SVM-Kernel is radial.

(ii) Decision Tree

Decision Tree is hierarchical in nature in which nodes represent certain conditions on a particular set of features, and branches split the decision towards the leaf nodes. Leaf determine the class labels. Decision Tree can be constructed either by using Recursive Partitioning or by Conditional Inference Tree. Recursive Partitioning is the step by step process by which a Decision Tree constructed by either splitting or not splitting each node. We can say that the tree is learned by splitting the source set into subsets based on an attribute value test. The recursion is completed when the subset at a node has all the same value of the target variable [2, 24].

Conditional Inference Tree is a statistical based approach that uses non parametric tests as splitting criteria that is corrected for multiple testing to avoid over fitting. This approach results in unbiased predictor selection and does not require pruning [10, 20].

(iii) Random Forest

Random Forest is a collection of Decision Trees. Random Forest is a supervised classifier created by Brieman. Random Forest is suitable for high dimensional data modeling as it can handle missing values, continuous, categorical and binary data.

(iv) K- Nearest Neighbor (KNN)

KNN is a simple technique that stores all available instances and classifies based on a similarity measure**.** It has been used in statistical estimation. For each row of the test set, the K nearest training set vectors are found and the classification is decided by majority vote, with ties broken at random. If there are ties for the kth nearest vector, all candidates are included in the vote [16,21]. The main drawback of KNN algorithm is expensive testing of each instance.

Model Evaluation. The performance of each Machine Learning technique can be evaluated using various performance measures like Accuracy, Sensitivity, Specificity, Precision [19].These measures are defined by four decisions: True Positive (TP), True Negative(TN),False Positive(FN) and False Negative(FN). TP decision occurs when malignant instances predicted rightly.TN decision benign instances predicted rightly. FP decision occurs when benign instances predicted as malignant. FN decision occurs when malignant instances predicted as benign [8].

(i) Accuracy can be calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

(ii) Sensitivity can be calculated as:

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

(iii) Specificity can be calculated as:

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

(iv) Precision can be calculated as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

The confusion matrix for the data set is then computed using these values into above equations to find Accuracy, Sensitivity, Specificity and Precision.

IV. RESULT AND CONCLUSION

The dataset is randomly divided into two subsets, one with about 70% of the instances to training, and another with around the remaining 30% of instances to testing. The performance of each Machine Learning algorithm is evaluated using four measures: Classification Accuracy, Sensitivity, Specificity and Precision. Table 4.1 shows the classification results of breast cancer in to benign and malignant were displayed by using confusion matrix of different Machine Learning techniques. In Table 4.2, the results are reported for different Machine Learning techniques for the breast cancer dataset in terms of different performance measures.We compared the classification results of SVM, KNN, and Random Forest, Decision Tree (using Recursive Partitioning and Conditional Inference Tree)..According to the classification results the KNN gives better results.

Table 1. Confusion Matrix.

(i) KNN

	Test outcomes	
Predictions	Benign	Malignant
Benign	161(TP)	0(FN)
Malignant	0(FP)	44(TN)

(ii) Decision Tree (Recursive Partitioning)

	Testoutcomes	
Predictions	Benign	Malignant
Benign	125(TP)	7(FN)
Malignant	8(FP)	71(TN)

(iii) Decision Tree (Conditional Inference Tree)

	Testoutcomes	
Predictions	Benign	Malignant
Benign	125(TP)	5(FN)
Malignant	8(FP)	73(TN)

(v) Support Vector Machine

	Testoutcomes	
Predictins	Benign	Malignant
Benign	136(TP)	8(FN)
Malignant	4(FP)	79(TN)

(iv) Random Forest

	Testoutcomes	
Predictions	Benign	Malignant
Benign	126(TP)	8(FN)
Malignant	7(FP)	70(TN)

	Support Vector Machine(SVM)	**KNN**	**Random Forest**	**Decision Tree using Recursive Partitioning**	**Decision Tree Using Conditional Inference tree**
Accuracy (%)	94.714	100	92.891	92.891	93.839
Sensitivity (%)	94.444	100	94.030	94.697	96.154
Specificity (%)	95.181	100	90.909	89.873	90.123
Precision (%)	97.143	100	94.737	93.985	93.985

CONCLUSION

The present study compares the performance measures of various Machine Learning techniques such as SVM, Decision Tree, Random Forest and KNN. The objective of this study is to find a best classifier. All the classifiers are used to classify the breast cancer dataset (Wisconsin Breast Cancer).The experiments were conducted in R tool. The results are compared and found that KNN outperforms other classifiers with respect to accuracy, sensitivity, specificity and precision. In future work we propose to analyze the Bayesian Network, Gradient Descent and Artificial Neural Network with and without dimensionality reduction techniques.

REFERENCES

[1] Aboul Ella Hassanien, Nashwa El-Bendary, Milos Kudelka and Vaclav Snasel (2013),*Breast Cancer Detection and Classification Using Support Vector Machines and Pulse Coupled Neural Network*, Springer Proceedings of the Third International Conference on Intelligent Human Computer Interaction, Page(s)-269-279.

[2] Alan Julian Izenman, *Modern Multivariate Statistical Techniques part of the series Springer Texts in Statistics*, Page(s):281-314.

[3] Ashfaq Ahmed K and Sultan Aljahdali, Nisar Hundewale and Ishthaq Ahmed K (2012), *Cancer Disease Prediction with Support Vector Machine and Random Forest Classification Techniques*, Computational Intelligence and Cybernetics (Cybernetics Com), IEEE, Print ISBN: 978-1-4673-0891-5, Page(s):16 – 19.

[4] Bao C. Q. Truong, H. D. Tuan, Anthony J. Fitzgerald and Incent P. Wallace (2015),*Breast Cancer Classification Using Extracted Parameters from a Terahertz Dielectric Model of Human Breast Tissue*, IEEE.

[5] Baoyu Zheng, Wei Qian, and Laurence P. Clarke (1996),*Digital Mammography: Mixed Feature Neural Network with Spectral Entropy Decision for Detection of Microcalcifications*, Transaction medical imaging, Volume 15, No 5.

[6] Chao W.L. and Ding J. J. (2011), *Integrated Machine Learning Algorithms for Human Age Estimation*, NTU.

[7] Chen Ren Dar, Chang Feng Ruey and Huang Len Yu (2000),*Breast Cancer Diagnosis using self organizing map for Sonography*, Elsevier Ultrasound in Medical and Biology,Volume-26, Page(s)-405-411.

[8] Data Mining Algorithms in R.

[9] Dheeba J. and Selvi Tamil S.(2011),Classification of Malignant and Benign Microcalcification Using SVM Classifier, IEEE Proceeding of ICETECT, Page(s)-686-690.

[10] Hothorn, T. and Hornik, k. (2006), *Unbiased Recursive Partitioning: A Conditional Inference Framework*, Journal of Computional and Graphical Statistics, Page(s): 651-674.

[11] Htet Thazin, Tike Thein and Khin Mo Tun (2015), *An Approach For Breast Cancer Diagnosis Classification Using Neural Network*: *Advanced Computing*, An International Journal (ACIJ), Vol.6, No.1.

[12] Hussain Muhammad, Wajid Kanwal Summrina, Ali Elzaart and Berbar Mohammed (2011),*A Comparison of SVM Kernel Functions for Breast Cancer Detection*, IEEE Eighth International Conference Computer Graphics, Imaging and Visualization, Page(s)-145-150.

[13] Hussein A. Abbass (2002),*An Evolutionary Artificial Neural Networks Approach for Breast Cancer Diagnosis*, Elsevier Artificial Intelligence in Medicine 25, Page(s)-265-281.

[14] Jihong Liu and Weina Ma (2007),*An Effective Recognition Method of Breast Cancer Based on PCA and SVM Algorithm*, Springer ICMB, Page(s) - 57-64.

[15] Konstantina Kourou and Themis P.Exarchos (2014), *Machine learning applications in cancer prognosis and prediction,* ComputStructBiotechnol.J, Page(s):8-17.

[16] Mucherino Antonio, Papajorgji J. Petraq and Pardalos M. Panos (2009), *Data Mining in Agriculture volume 34 of the series springer* optimization and its applications, Page(s): 83-106.

[17] Muhammad Shoaib B. Sehgal, Iqbal Gondal and Laurence Dooley (2014),*Support Vector Machine and Generalized Regression Neural Network Based Classification Fusion Models for Cancer Diagnosis*, IEEE Proceedings of the Fourth International Conference on Hybrid Intelligent Systems.

[18] Omar S. Soliman and Aboul Ella Hassanien (2012), *A Computer Aided Diagnosis System for Breast Cancer using Support Vector Machine*, Springer, Page(s)-106-115.

[19] S. Syed Shajahaan, S. Shanthi and V. ManoChitra (2013), *Application of Data Mining Techniques to Model Breast Cancer Data,* ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 11, Page(s): 362-364.

[20] Strobl, C. and Malley,J(2009), *An Introduction to recursive Partitioning Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests,* Psychological Methods, Page(s):323-348.

[21] Tipping M.E. (2001), *Sparse Bayesian learning and the relevance vector machine*,Journal of Machine Learning Research,Vol. 1, Page(s):211-244.

[22] Ubeyli Derya Elif.(2007), *Implementing Automated Diagnostic Systems for Breast Cancer Detection*, Elsevier Expert Systems with Applications 33, Page(s)-1054-1062.

[23] Website: www.cancer.org(American Cancer Society), Date: 1/02/2017, Time: 11:25 am.

[24] Website: www.StatSoft Inc, Support Vector Machine, Date: 20/02/2017, Time: 11:30 pm.

[25]Website:www.wcrf.org,Date:20/02/2017,Time:11:00 am.