Machine Learning Algorithms For Early Stage Breast Cancer Diagnosis

Amaad Khalil, Irfan Ahmed, Zawar Hussain Khan, Salman ilahi Siddiqui, Imran Ahmad

Abstract: Breast cancer is one of the dangerous diseases which is more common in female. It could be cured if diagnosed at early stages. This disease has some common symptoms which include the suspicion of harmful tumor, but accurate diagnosis requires different investigation modalities. Traditional diagnosis method of breast cancer consists of clinical, microscopic, and radiographic techniques. The traditional techniques have some limitations due to which proper assessment of the disease is not possible. In this work a machine learning based framework is proposed for automated intelligent assessment of breast cancer diagnosis. For this purpose, different machine learning techniques are compared based on the performance comparison of classifying malignant and benign tumor. The performance evaluation of experimented techniques shows the feasibility of different machine learning techniques to be used for stored time and real time breast cancer diagnosis.

Index Terms: Breast cancer detection, CART, Machine Learning, Random Forest, Supervised Learning, Tree Classifier, WEKA

1. INTRODUCTION

Breast cancer is a debilitating disease which is getting common in women; however, men can also become victim of this disease. If breast cancer is not diagnosed in time or during the early stage of tumor formation, it can seriously damage the health which ultimately leads to death. The early stage diagnosis of tumor development is indispensable for proper treatment. Although the disease has some symptoms which can help in the assessment of the disease, however most of the symptoms are common for the other various diseases, so the growth of the tumor remains unnoticed. The biologists had limited traditional ways for the diagnosis of the disease through microscopic methods, but such methods could not precisely assess the tumor involved in the breast cancer. The early detection of tumor is extremely important for the successful treatment of the disease and it is a challenging task to diagnose the tumor due to the lack of precise and accurate method of tumor diagnosis [6]. Machine learning algorithms has played a significant role in this regard and helping biologists to detect harmful tumor by obtaining the accurate information about the size and age of the tumor to get the information whether the tumor is curable or not [10] [7]. Some of the AI generated heat maps of breast cancer tumor are shown in Figure 1.

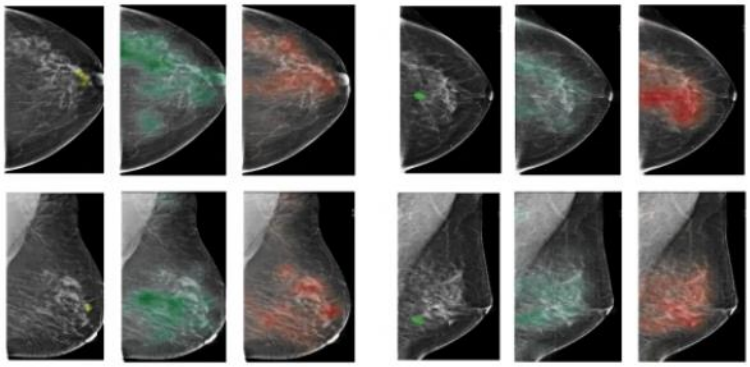

Figure 1: *Breast Cancer Heat Maps Generated By AI*

According to the survey conducted in 2014, there are 232,670 women and 2360 men in US who are suffering from the disease [6]. The earlier detection of malignant tumor is a challenging task. Recently machine learning models are trained to classify biomedical images as malignant or benign. Classification is a two steps process in which different objects or groups are classified into different labels or classes [5]. In the first step a model is developed by analyzing the training records of a database in which label of each class of training is already known. This mode of model training is known as supervised learning. In next step, test data is used to evaluate the classification performance of developed model. The efficient utilization of modern machine leaning methods in classification tasks is the prime motivation of their deployment in breast cancer detection task. The work done in this paper is aimed to propose efficient tree-based machine learning model for breast cancer detection. The prime contribution of this paper lies in training tree-based models with publicly available breast cancer dataset and performance of the models are compared on the basis of accuracy and decision time of a classification model. Machine learning algorithms have recently been used in biomedical engineering for breast cancer detection task. One of the most widely used machine learning techniques for classification tasks is a Decision tree [8]. Decision tree divides dataset of records recursively into sub parts by using one of the two artificial intelligence technique. First one is called depth first greedy approach and the other one is breadth first approach till the classification of data items. In this process knowledge is represented in the form of rules, making it easy for the users to understand the model. In the structure of a decision tree, a test on an attribute

- *Amaad Khalil is currently pursuing PhD degree program in computer systems engineering in University of Engineering & Technology Peshawar, Pakistan. E-mail: amaadkhalil@uetpeshawar.edu.pk*
- *Irfan Ahmed is currently pursuing PhD degree program in computer systems engineering in University of Engineering & Technology Peshawar, Pakistan. E-mail: irfanahmed@uetpeshawar.edu.pk*
- *Zawar Hussain Khan is an Assistant Professor at Department of Electrical Engineering UET Peshawar Pakistan. zawarkhan@nwfpuet.edu.pk*
- *Salman Ilahi Siddiqui is a PhD student in Department of Electrical Engineering UET Peshawar Pakistan. salman.ilahi@uetpeshawar.edu.pk*
- *Imran Ahmad is an Assistant Professor at Department of Industrial Engineering UET Peshawar Pakistan. imranahmad@uetpeshawar.edu.pk*

is represented by node and its result makes the branch of the tree while leaf node represents a label or class. An unknown record is classified by testing the attributes in the tree from the root until a leaf path is defined. Although each node has an exclusive path from the root, but still many leaf nodes can make the similar classification [8] [17]. Alternating Decision Tree (ADTree) is another machine learning algorithm which is successfully used in this domain. In this algorithm, the decision tree classifier is combined with boosting techniques, thereby generating rules that are smaller in size and easy to interpret [9] [14]. In decision stump, the internal nodes also known as the root nodes are directly in contact with the terminal nodes, which are known as the leaves nodes [16]. The decision stump makes the prediction based on only single input feature. A decision stump has the following form:

$$g(x) = D(x_i > c) \quad (1)$$

Where, the value of argument is 1, if i^{th} element of x is greater than c and -1 otherwise.

Classification and Regression trees (CART) is also one of the tree based algorithms used for classification purposes. In 1984 this algorithm was introduced by Breiman [12]. This algorithm is based on both classification and regression trees. The serial implementation of this algorithm is also possible because it also supports Hunt's model of decision tree construction. In this method splitting attribute are selected using gini index splitting measure. Data pruning is carried out over some portion of training dataset. In CART both numerical and categorical attributes are used for developing the decision tree. CART has the capability to deal with missing attributes with the help of its in-built features [16] [12]. In CART we use information gain and entropy to find the overall gain. These can be calculated as:

$$I.G(\alpha,\beta) = \frac{-\alpha}{\alpha+\beta} log_2 \left(\frac{-\alpha}{\alpha+\beta}\right) - \frac{-\alpha}{\alpha+\beta} log_2 \left(\frac{-\alpha}{\alpha+\beta}\right) \quad (2)$$

$$Entropy\ (X) = \sum_{i=1}^{z} \frac{\alpha_i + \beta_i}{\alpha+\beta}\ (I.G(\alpha,\beta)) \quad (3)$$

$$Total\ Gain\ =\ I.G(\alpha,\beta) -\ Entropy\ (X) \quad (4)$$

Where α and β are class attributes and used to find the overall gain by entropy to construct the overall tree.

Rapid tree algorithm is a fast decision tree, and it has the roots from its predecessor C4.5 machine learning algorithm. It is used to build decision or regression models [16] [15]. In this algorithm, decision tree is built by gaining the information from splitting the corresponding instances and reduced error pruning is used for pruning. However, REPTree algorithm has achieved lower accuracy according to some studies [3]. J48 is a statistical decision tree for the classification purposes and is developed by Ross Quinlan. It is based on either pruned or UNN pruned C4.5 machine learning algorithm. In this algorithm, tree is generated from the input data using entropy concept [1] [13]. The training data sets have pre-classified data with some additional information. In decision tree effective attribute is selected at every node decision of the input data and splits it into different classes [13]. Random Forest is also a type of decision tree in which a statistical ensemble learning algorithm developed on the basis of bagging is used for the classification of input data [2]. It is effective in producing better accuracy in sample space of huge dimensionality and high variance inhibits the decision tree capability to produce better results [11]. Steps in Random Forest algorithm are as follow.

- Bootstrap samples are generated for 'n' numbers of trees, which are originated from the input data set i.e., samples X= x_1, x_2… x_m and response labels Y=y_1, y_2 ,….., y_m.
- After the generation of bootstrap samples, Bootstrap aggregation is done by connecting the learners in parallel. For a set of 'm' training examples, regression tree is trained by selecting samples with replacement 'n' times.
- In the last step by using either mean prediction or classification prediction is made. Test sample $\tilde{x}$ is predicted by the following mathematical relationship:

$$\hat{g} = \frac{1}{n} \sum_{i=1}^{n} g_i(\tilde{x}) \quad (5)$$

The idea of Random trees is developed from both no pruning classification and regression criteria and is categorized as statistical classification technique [3]. In random tress the input data samples are classified using each instance of the tree. The result of the classification process is the class having majority votes. In this algorithm regression problem is eliminated by noticing average response of the tree [15]. In FT trees, linear function of the attributes is combined with standard decision with the help of linear regression techniques tree; such as C4.5 [12]. Unlike univariate DT where simple value tests are carried out on single attributes in a node, FT is able to combine different attributes of a single node or of the whole leaf linearly. In the constructive phase new attributes are developed by the function mapping of earlier ones [15]. LAD is the short form of Logical Analysis of Data. In this method logical expression is developed from the classification of binary target variables. Binary variables can be classified either as positive or negative variables in a data set [11]. The elementary concept of LAD model is that if a single binary value is surrounded by just some positive patterns and having no traces of any negative pattern around it is classified as positive and vice versa.

2 OUR CONTRIBUTION

The basic objective of this research work is to sort out the best machine learning classifier that can accurately detect malignant tumor in breast cancer diagnosis task. The tool used for the experiment purposes is WEKA, which is termed as Waikato Environment for Knowledge Analysis. The devised methodology is based on the application of supervised classification techniques for the analysis of machine learning techniques for breast cancer dataset in WEKA. In these techniques, tree based machine learning algorithms [8] are applied to carry out performance evaluation of applied methods. In general, doctor assesses the disease by symptoms and other tests but such information may not be sufficient and sometime misleading. Therefore, the proposed work automates the task of cancer diagnosis by proposing efficient machine learning techniques for the improvement of breast cancer tumor classification results. The objective of the work is to improve current breast cancer diagnosis techniques which may revolutionize the field of telemedicine and tele-monitoring.

3 EXPERIMENTS

The framework used in this research work is called WEKA. It is a freely available open source package [4] [16] under GNU public general license. It is basically a package of different supervised and unsupervised machine learning algorithms, which are widely used for data mining and classification tasks. These algorithms can also be experimented with external datasets; otherwise Java code also supports external datasets provision to the inbuilt algorithms. WEKA provides various data analysis techniques such as clustering, association rules, data visualization and feature selection. WEKA is also a useful platform for developing new and updated versions of existing machine learning algorithms. The classification accuracy of a machine learning technique is a key parameter in this research. Accuracy is measured in percentage which provides quantitative measure of correctly classified instances by a classifier.

$$\text{Accuracy } (\%) = \frac{\text{Correctly Classified Instances}}{\text{Total Intances}} * 100 \quad (6)$$

$$\text{Error}(\%) = \left(1 - \frac{\text{Accuracy } (\%)}{100}\right) * 100 \quad (7)$$

In this work, kappa statistics is used for the assessment of the accuracy of any specific measuring case, so collected data's reliability and its validity must be properly distinguished and is mandatory. In this work, the following mathematical equations are used for analyzing classification based on kappa score:

$$k = \frac{a_o + a_e}{1 - a_e} = 1 - \left(\frac{a_o + a_e}{1 - a_e}\right) \quad (8)$$

Where for N total samples, the a_e score between two samples can be calculated as:

$$a_e = \frac{1}{N^2} \sum_m n_{m1}\, n_{m2} \quad (9)$$

We performed all experiments by using libraries and built-in datasets of the framework. The dataset used in this research work contains 10 attributes and 276 number of instances for each attribute. The last value in the attribute table is the class, which can be recurrence symptoms or no recurrence symptoms. The detailed information related to attributes of dataset is given in Table 1.

Table 1 *Attributes of Dataset*

S.NO	ATTRIBUTE	EXPLANATION
1	Age	Age of Patient
2	Menopause	Menopause status of patient
3	Tumor_size	Size of the Tumor in (mm)
4	Inv_nodes	axillary lymph nodes showing breast cancer
5	Node_caps	Penetration of tumor
6	Deg_malig	Degree of Breast Cancer
7	Breast	Either Breast cancer or not
8	Breast_quad	Position of Breast nipple
9	Irradiat	X-ray Radiations therapy
10	Class	Recurrence or no recurrence symptoms

4 RESULTS AND DISCUSSION

The obtained results in the form of kappa score ranges between 0.6 and 0.7 from the selected algorithms. Based on the kappa statistic criteria, substantial part of these methods is classification accuracy. An algorithm with the lowest rate of error is appreciated, as it possesses more powerful classification capability of breast tumor. Out of 276 total numbers of instances, 200 instances are correctly classified, with the highest number is of 211 instances and the lowest of 196 instances as shown in Figure 2. It is clear from Table 2 that j48 tree classifier has achieved the highest accuracy level of 75% while Random Tree algorithm has the lowest accuracy level of 66 %. Other algorithms on an average have gained the accuracy level up to 70%. Apart from the accurately classified instances, model preparation time is also an important parameter during the comparison of different classification algorithms. As shown in Table 3, the minimum classification time is of ADTree algorithm, which is around 0.02 seconds as compared to other algorithms. The LMT algorithm takes longest time among all the algorithms of 1.2 seconds. K-Nearest Neighbor is applied to the dataset and produces 1.7% better results than other techniques.

Table 2*. Accuracy of Correctly Classified*

CLASSIFIER	CORRECTLY CLASSIFIED	INCORRECTLY CLASSIFIED
ADTree	73.776 %	26.223%
BFTree	67.832%	32.167%
Decision Stump	68.532%	31.469%
FT	70.279%	29.720%
J48	75.525%	24.475%
LADTree	70.629%	29.371%
LMT	74.175%	24.825%
NBTree	70.979%	29.021%
RandomForest	67.832%	32.168%
RandomTree	66.783%	33.217%
REPTree	70.629%	29.370%
SimpleCart	69.231%	30.769%

The comparison graph is also shown below:

Table 3*. Time Required for Model Development*

CLASSIFIER	TIME (S)
ADTree	0.02
BFTree	0.27
Decision Stump	0.1
FT	0.55
J48	0.02
LADTree	0.14
LMT	1.22
NBTree	0.86

RandomForest	0.05
RandomTree	0.01
REPTree	0.02
SimpleCart	0.41

The comparison graph is also shown in the form of Figure 2.

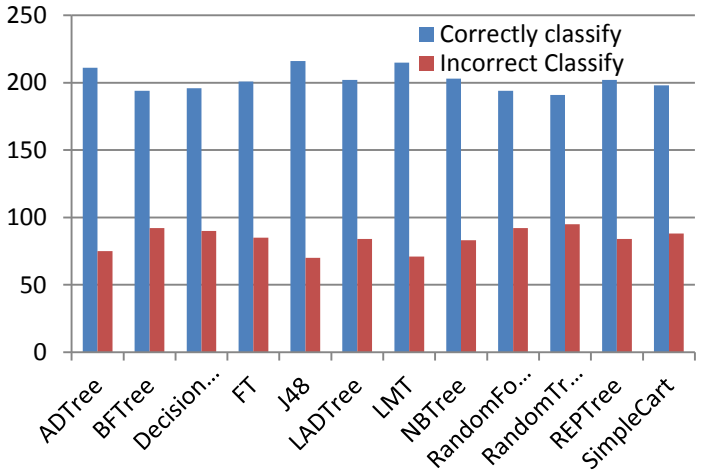

Figure 2: *Correctly Classified instances Graph*

5 CONCLUSION

This paper analyzes classification performance of widely used tree based machine learning techniques for breast cancer diagnosis task. The obtained results of various classifiers, it can be concluded that the classification performance of J48, LMT, and ADTree is better among various classifiers which were used for classification purposes. As far as model development time is concerned, ADTree and J48 take lesser time as compared to other classifiers. Since the numbers of correctly classified instances are greater than that of incorrectly classified ones, some of these classifiers can be used in developing real time diagnosis system for breast cancer. Despite the usefulness of some classifiers none of them has achieved the optimum level of accuracy, so future work is intended to introduce algorithms that can generate better results.

REFERENCES

[1] Neeraj Bhargava, Girja Sharma, Ritu Bhargava, and Manish Mathuria. “Decision tree analysis on j48 algorithm for data mining”. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), 2013.

[2] Leo Breiman. “Random forests”. Machine learning, 45(1):5{32, 2001.

[3] Na Chen, Guochu Shou, YiHong Hu, and ZhiGang Guo. “An experimental research of tra_c identification algorithms in broadband network.” In 2009 International Symposium on Computer Network and Multimedia Technology, pages 1{4. IEEE, 2009.

[4] Eibe Frank, Mark Hall, Geo_rey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H Witten, and Len Trigg. “Weka-a machine learning workbench for data mining.” In Data mining and knowledge discovery handbook, pages 1269{1277. Springer, 2009.

[5] Jiawei Han, Jian Pei, and Micheline Kamber. “Data mining: concepts and techniques.” Elsevier, 2011.

[6] Jahanvi Joshi, Rinal Doshi, and Jigar Patel. “Diagnosis of breast cancer using clustering data mining approach”. International Journal of Computer Applications, 101(10):13{17, 2014.

[7] Abraham Karplus. “Machine learning algorithms for cancer diagnosis”. Santa Cruz County Science Fair, 2012.

[8] Hian Chye Koh, Gerald Tan, et al. “Data mining applications in healthcare”. Journal of healthcare information management, 19(2):65, 2011.

[9] Yishay Mansour. “Pessimistic decision tree pruning based on tree size.” In MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-, pages 195{201. Citeseer, 1997.

[10] Tracey McCready, Dot Littlewood, and Jane Jenkinson. “Breast selfexamination and breast awareness: a literature review.” Journal of Clinical Nursing, 14(5):570{578, 2005.

[11] Nicolai Meinshausen. “Quantile regression forests”. Journal of Machine Learning Research, 7(Jun):983{999, 2006.

[12] J Ross Quinlan. “C4. 5: programs for machine learning”. Elsevier, 2014.

[13] JR Quinlan. “The morgan kaufmann series in machine learning”. San Mateo, 1993.

[14] Haijian Shi. “Best-_rst decision tree learning.” PhD thesis, The University of Waikato, 2007.

[15] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. “Data Mining: Practical machine learning tools and techniques.” Morgan Kaufmann, 2016.

[16] Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geo_rey Holmes, and Sally Jo Cunningham. “Weka: Practical machine learning tools and techniques with java implementations”. 1999.

[17] ZOU Yuan. “Data mining algorithm based on decision tree application and research” [j]. Science Technology and Engineering, 18, 2010.