Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 13 (2020) Pp-21-27

BBRC
Bioscience Biotechnology Research Communications

Machine Learning Applications for Automated Breast Cancer Detection and Analysis

Priya B Bagewadi[1], Sujata N Patil[2], Parameshachari B. D[3] and Shweta Madiwalar[4]

[11]Department of Biomedical Engineering, KLE Dr. M.S. Sheshgiri College of Engineering and Technology, Belagavi, India
[2]Department of Electronics and Communication, KLE Dr. M.S. Sheshgiri College of Engineering and Technology, Belagavi, India
[3]Department of Electronics and Communications, GSSS Institute of Engineering & Technology for Women, India
[4]Department of Electronics and Communication, KLE Dr. M.S. Sheshgiri College of Engineering and Technology, Belagavi, India

ABSTRACT

Breast cancer is the most secretive and common cancer among women and rarely in men .It is a vital issue to get the faster and accurate diagnosis of the patient so that doctors can decide the treatment in due time. Across the globe around 10% of the people are affected in some stage of their lives. Frequently occurring cancers are present especially among women. Most of the challenges are faced when the carcinoma or the cancer is not detected correctly at the initial stage by experts for medication. In the proposed research work, different Machine Learning techniques have been tried to get the most suitable accuracy for the analysis of breast cancer. Generally the traditional methods of data classification in the diagnosis have been effective in the days so far. The classification techniques used are in the form of decision tree, K- nearest neighbors, XG Booster, Ada Booster, Naïve Bayes, Logistic Regression, SVM on Wisconsin Breast Cancer datasets, both before and after applying Principal Component Analysis. In this project supervised machine learning tool is used for detection of cancer.

KEY WORDS: ADA-BOOST, BENIGN, BREAST CANCER, MALIGNANT, SUPPORT VECTOR MACHINE.

INTRODUCTION

Through decades cancer has been the second largest cause of death worldwide. According to the World Health Organization (WHO), breast cancer is the major cause of death in women. In the year 2018, it was recorded that about 6, 27, 000 deaths have happened due to breast cancer, which corresponds to 15 percentage of all cancer mortality (Sujata Patil and Uday Wali; 2018). As per the records of ICMR department the death rate in women is higher than men due to cancer. Breast cancer is generally diagnosed using the imaging modalities like breast ultrasound, breast magnetic resonance imaging (MRI), breast computerized tomography (CT scans), breast mammography, biopsy and histology (Florin G, 2008).

Whenever the doctor observes any symptoms of cancer, the patient is informed to undergo various tests to confirm cancer, based on which the line of treatment is advised to the patient. The effective treatment of the patient depends on the efficiency of the diagnostic system to accurately detect the malignancy. Based on the type and stage of cancer, the therapy suggested may involve radiation therapy, chemotherapy or surgery. The recovery

ARTICLE INFORMATION

Received 11th Oct 2020 Accepted after revision 27th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.13/4

and survival of the patient is dependent on the accurate and early detection of cancer. The breast is made of milk ducts, adipose tissue, lobules, lobes and fatty tissues. Cancer usually starts from the area of the lobes, nipple and the milk duct. It starts as a cluster of abnormal cells in one area and then affects the normal and healthy tissues spreading to the other parts of the body.

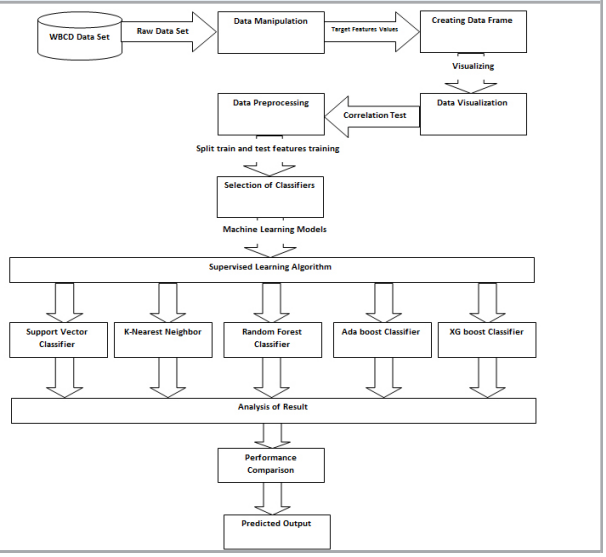
In some cases breast cancer does not show up any physiological signs at the initial stages, thus remains unattended.

The early detection of cancer may be possible by medical screening or mammography. Some of the symptoms in the case of breast cancer are: tender nipple, lump formation near the nipple or in the underarm, swelling or shrinkage of the breast, bloody or milky discharge from the nipple et.al (Nuryanti M Harsa and Mohammad S; 2019). Breast cancer occurs in various types like ductal carcinoma in situ, inflammatory, invasive ductal carcinoma, metastatic and Lobular carcinoma in situ. Ductal Carcinoma in Situ is a non-invasive type of cancer. Here the cancer cells grow within the milk duct but don't spread beyond. It indicates the initial form of cancer and thus can be successfully treated. Inflammatory Breast Cancer involves the spread of cancerous cells through the skin and lymph vessels of the breast. In case of Invasive Ductal *Carcinoma*, the cells spread from the milk ducts to other parts of the body. Lobular Carcinoma in Situ involves non-cancerous cells but may be detected through biopsy. Metastasis is fourth stage of breast cancer.

where the cancerous cells have spread to other parts of the body like the brain, lungs, liver or bones. The cancer cells multiply and thus damage the other parts of the body, thus multiplying and forming tumors. Breast cancer can be diagnosed by various medical tests and based on the type and stage of the cancer the doctor can decide upon the treatment to be suggested. The diagnostic methods used are Physical Breast Examination, Mammogram, Ultrasound, MRI, Biopsy, Fine Needle Aspiration, Core Needle Biopsy and surgery. These diagnostic methods require some amount of time to obtain the results due to which there maybe delay in the therapy. Sometimes even some important features may be missed out that may affect the efficiency and accuracy of the results.

To avoid these limitations automated diagnostic systems are being introduced that provide early detection of cancer within the proper time frame and higher accuracy thus provided immediate treatment to the patient thus increasing their survival rate et.al (Yuan Jiao MA and Janghe R; 2020). The automated system is based on Artificial Intelligence and Machine Learning that apply the algorithms and statistical models onto the acquired images and thus differentiate between the benign and malignant tumors. Machine learning algorithms are widely used for the diagnosis and prognosis of breast cancer. The various models used for breast cancer detection are Artificial Neural Networks, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, Decision Tree, etc. ML techniques have shown their remarkable ability to improve classification & prediction accuracy.

Figure 1: Workflow diagram



MATERIAL AND METHODS

To apply the machine learning models appropriate dataset need to be collected. In the proposed research work, Wisconsin Breast Cancer Dataset (WBCD) is used . After collecting the dataset, the pre-processing methods are used to remove the undesirable observations or extraneous values from dataset as shown below in the workflow diagram (Figure 1).

Dataset: The Graphical computer program called XCYT is used which is capable of performing digital scan to perform the analysis of cytological features. Dr. Wolberg created dataset by taking fluid samples, solid breast masses of patients to compute ten features from each cell. In the sample, curve fitting algorithm is used by the computer program. For each feature of an image the mean value, extreme value and standard error is calculated et al (Hamhung Adi Nugroho and Umesh; 2018) .The dataset contains 212 cases of malignant breast cancer and 357 cases of benign breast cancer and 32 columns with first column starting with patient ID, followed by the standard deviation and mean of the worst measurements of ten features.

Attribute Information:

ID number

Diagnosis (M = malignant, B = benign) 3–32)

Data visualization:

Figure 2: Pair plot of all the features

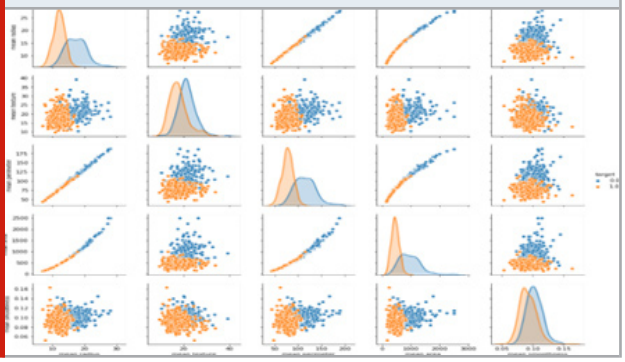


Figure 3: Total count of malignant and benign tumor patients in counter plot

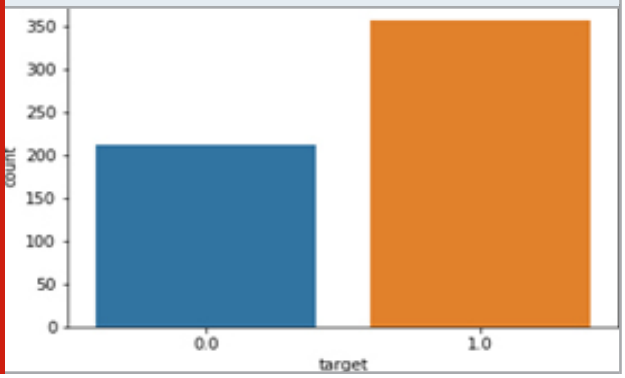
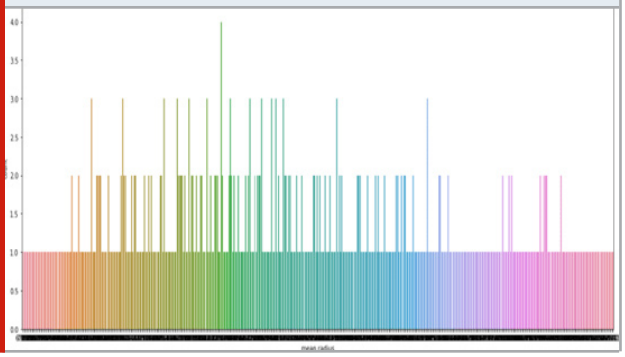


Figure 4: Counterplot max samples mean radius is equal to 1



Pair plot: Exploratory data analysis is the next step once we are ready with the required dataset. It is the process used to find the relationships or patterns for the analysis. There are many methods available to perform EDA; one of the most efficient tool is to make the pair plot (Claudio Manna ;2012). By using the pair plot tool we can observe the distribution of single variable and two variable relationships. As our data is in numerical format we have to take pair of plot which is already distributed in two categories, the malignant as 0 and benign as 1 and we can easily distribute it in blue and orange (Dr.R.H.Havaldar. 2016) as shown in (Figure 2).

Counter plot: The plot gives the information about how many patients are having benign and malignant tumors. The (Figure 3) shows the 212 dangerous malignant tumors of 38% and 357 benign tumors of remaining 62% of the class.

The counter plot (Figure 4), shows that the differentiation between the patients who are affected with cancer and who are not affected with cancer (Sagar Metri. 2018). From the counterplot graph we can see that mean radius value of the patients who are not suffering from cancer is less one and who are suffering from cancer mean radius is more than one.

Figure 5: Heat map Correlation Between All Variables

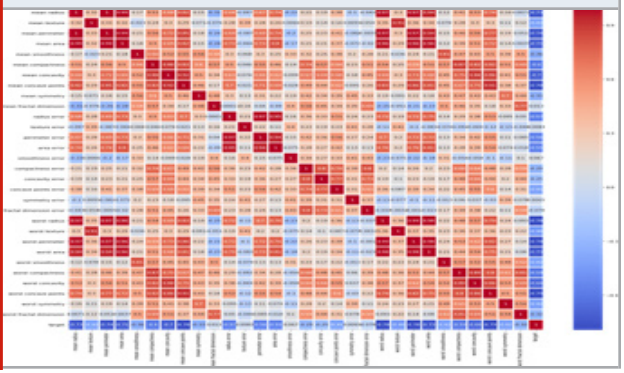
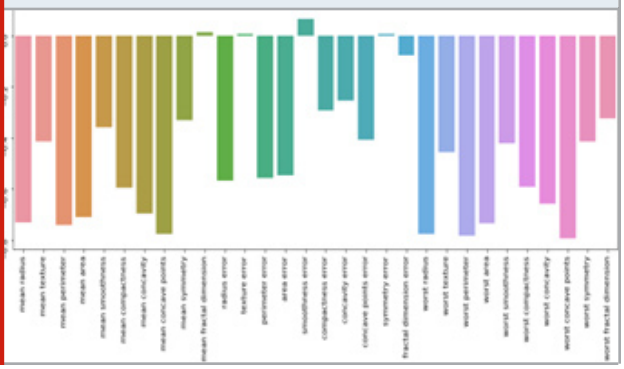


Figure 6: Correlation bar plot of all the features



Heat Map: The two dimensional illustration of data within which values are represented using colors are called as heat map, as shown in (figure 5). Direct visual outline of data is obtained from the heat map. We visualize heat map using correlation matrix to calculate the correlation between each target and each feature et.al. (Sujata N Patil and Uday Wali ; 2019). The correlation bar which is the indicator of the measure of SMOOTHENING VALUE for the error prediction is the only parameter strongly certainly correlated with the final targeted object than others. Other measures like MEAN FACTOR VALUE, ERROR WITH TEXTURE and SYMMETRY VALUE MEASUREMENT are hardly correlated or rarely positively correlated and rest of the remaining parameters are very rarely or weakly correlated as shown in (Figure 6).

Generally the predictive rate will be high if we use the machine learning techniques with single model. Now these techniques can be considered as better learning methods for comparative result analysis. These methods are the best solutions for the prediction of diagnostic results with validation from the experts. Here we have used the various Machine Learning Techniques to classify the mammographic images malignant or not using the dataset available. We different predictive methods of machine learning are applied to the early detection of Breast Cancer.

All the ML techniques are verified with the accuracy measurement value for various algorithms available. In the verified methods with ML techniques the support vector machine (SVM), XG Boost gave the comparative analysis relatively good as that of KNN, DT, Naïve Classifier, Xtra Tree etc. SVM and XGBoost algorithms have been used for early classification and analysis of cancer, and XGBoost algorithm has given better accuracy (98%) compared to SVM (96%). These results have been shown with the graphs (Figure (2-6)), indicating the accuracy with respect to the different Machine Learning (ML) Techniques. The following section briefly describes the machine learning algorithms:

1. LG_Logistic Regression
2. SVM_Support Vector Machines
3. KNN_K- Nearest Neighbor Classifier
4. Naïve_Based_Bayes_classifier
5. DT_Decision Tree Algorithm
6. RFC_Random Forest Classification
7. ADA_Booster_Classifier
8. XG Booster_Algorithm

Table 1. Sample Dataset of patients for malignancy prediction ,using ADA Boost algorithm

PID	VSS	NLD	EDC	CLWB
156	No	No	No	No
176	Yes	Yes	No	No
187	No	Yes	Yes	Yes
183	Yes	Yes	Yes	Yes
192	No	Yes	No	Yes
203	Yes	Yes	Yes	Yes
165	Yes	Yes	No	Yes

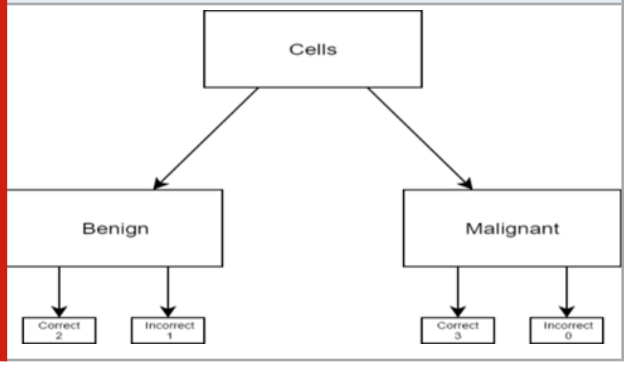
ADA Boost: The Yoav Freund and Robert Schapire formulate the first Meta algorithm in machine learning that is "Adaptive Boosting," in short it is called as Ada Boost.

$$F(x) = \text{sign}\left(\sum_{m=1}^{M} \theta_m f_m(x)\right) \quad (1)$$

Theories of weak classifiers are transformed into good classifiers by using Adaptive Boosting which will mainly focus on classification related issues. Adaptive Boosting is supervised techniques under the category of numerical prediction. It is especially used in over fitting on noisy datasets. In this case, the first step is to catalogue the cells onto healthy or malignant (Sujata N Patil, 2019). With a depth of 1 the decision tree for each feature is generated to identify the results it uses. Comparing the projections of each tree with the real marks is performed using training set .The next tree in the forest turn out to be the feature and associated with the tree that has done the best job of sorting the training samples.

As shown in (Figure 7), based on the results, the tree organizes cells as Benign and Malignant. But the decision tree mistakenly classified 1 cell as 'benign' based on existence situations. For all trees, we repeat the procedure and choose the one with the minimum number of wrong predictions. Where the cumulative error is the sum of the sample weights wrongly graded, Incorrect result = sum of result for incorrectly classified samples. Consider the errors made by the previous decision tree, then adjust the sample results for all the samples wrongly labelled by the current tree, use the following formula to raise their related results.

Figure 7: ADA Boost Detection



According to the new sample weight e to the power of the significance calculated in the previous step because it needs data

$$\text{New results} = \text{sample weight} * e^{-\Sigma} \quad (2)$$

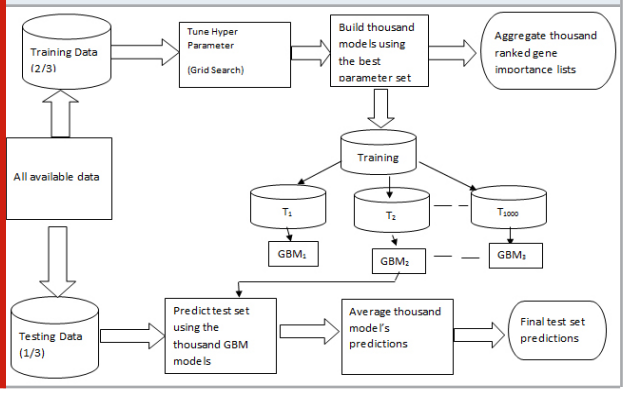
To increase exponentially, instead, it looks at the samples properly categorized by the tree and uses the following formula to decrease their related weights. The main result here is that the samples incorrectly classified by the previous stump should be associated with larger results, and those correctly classified with smaller sample weights should be associated with them (Sujata N Patil. 2019). Create the new dataset and empty the data corresponding to the original size. The envision a roulette table where each pocket fits a sample weight. Select the numbers random from 0 to 1.The position of the sample defines

where each number falls in the new dataset. Even though the samples are erroneously characterized higher weights relative to the others, there is greater likelihood that their distribution slice will fall below the random number. Therefore the dataset might contain the multiple copies of the samples as it was misclassified by the previous list (Sujata Patil. 2020).

Repeat the procedure until the number of iterations is equal to the number specified by the hyper parameter (i.e. number of estimators). To predict data outside the training set we can use the decision trees forest .By classifying each tree as a reference in the forest, the Ada Boost model does predictions as per their decisions and the trees are split into groups. For each group add the meaning of each tree within the group. The final classification made by the forest as a whole is decided by the category with the greatest amount. The number of iterations of Ada Boost is also a poorly set number of weak classifiers that can be calculated with cross-validation. These are some limitations of Ada Boost algorithm. But there are advantages also like it uses weak cascading classifiers. Various classification algorithms can be used as weak classifiers. Ada Boost has a high degree of precision.

XG Boost Classifier: XG boost was a gradient boosted decision tree. Rather than training all models, it trains through the sequence of the model and modifies errors made by the previous techniques with increase in efficiency of prediction. Ensemble Machine Learning algorithm, XG Boost is a Decision-tree- that uses a gradient boosting framework. XG Boost is a library of gradient boosting and used to get benefits to execute and perform better and are more used for regularized model formalization, to control over-fitting, which gives it better performance.

Figure 8: XG Boost Classifier



New results = results * e^{Σ} (3)

Above (Figure 8), shows the schematic work flow diagram of XG Boost. GBM stands for gradient boosting machine and T stands for tree. The two oval boxes on the right depict the outputs from XG Boost. The various measurement parameters used to evaluate the accuracy, precision parameter, recall and the value with respect to confusion matrix. The preciseness of the classifier can be measured with the Accuracy of the prediction classifier and it generally gives the information about the number of missed samples and the samples wrongly classified. These measurement values will be the number of accurate true samples denoted by TP, the prediction with false identification FP, the negative samples identified as positive TN and samples which are negative identified wrongly FN. These can be written with the equations as below:

The other metrics derived from a confusion matrix are defined as follows:

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} \quad (1)$$

TP, TN, FP and FN are the number of true positives, true negatives, and false positives and false negatives, respectively, when the classifier is predicted.

CS derived from a confusion matrix are defined as follows:

$$\text{RECALL} = \frac{TP}{(TP+FN)} \quad (2)$$

$$\text{PRECISION} = \frac{TP}{(TP+FP)} \quad (3)$$

$$\text{F1\_Score} = 2 * \frac{(PRECISION * RECALL)}{(PRECISION + RECALL)} \quad (4)$$

The measure of the receiver sensitivity parameter and the ability of the classifier to remember the learning rate can be measured with the recall rate of the classifier. This recall rate depends on the receipt of the sample rate with respect to the specificity value. The minimal number of error values is considered for the analysis of the classifier. The linear prediction models applied for the all the ML techniques are given the results as depicted in the Table 2 and the accurate classifier will give the boundary covering precisely. Based on the region to be considered for the prediction of the malignancy detection and analysis performance is verified with all the available threshold values. The accuracy of the characteristic or the performance curve can be checked to predict the model correctly shown in (Figures 5) and compare the performances of the nine computational models. Here in this proposed method we analyzed the predictive values and found to be with region characteristic curve to be 77%.

Table 2. Showing the different accuracy values

Techniques	Accuracy without Standard Scale	Accuracy with Standard Scale
SVM	57%	96%
KNN	93%	57%
RF	97%	75%
Adaboost	94%	94%
XGboost	98%	98%

RESULTS

We analyze and compare among all these algorithms that we have used for our research work. This comparison is based on some core characteristics like, what is the average predictive accuracy, how fast the classifiers train and make predictions, what happens when there is a small dataset. Average predictive accuracy of Logistic Regression, Naive Bayes, K- Nearest Neighbors and Decision Tree are comparatively lower. On the contrary, Support Vector Machine, Random Forest and Ada Boost Tree have higher accuracy.

In our research work we also noticed the upper scenario as SVM and XG Boost gave the highest accuracy. Training speed also defers for different algorithms. Training speed of Logistic Regression, Naive Bayes and Decision Tree are faster and rest classifiers training speed is much slower. K- Nearest Neighbor classifier doesn't need any training. Most of the classifiers predict Fast but Random Forest classifiers predict at an average speed. On the other hand, prediction speed of K- Nearest Neighbor classifier is slower. Logistic Regression, SVM, Naive Bayes, K-Nearest neighbors performs well with a small number of observations whereas Random Forest, Ada Boost Tree and Decision tree need a comparatively big number of observations to make a satisfactory result.

Cross Validation: Cross-validation is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. Cross-validation is largely used in settings where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive model. The mean accuracy value of cross-validation is 96.24% and XG Boost model accuracy is 98.24%. It shows XG Boost is slightly over fitted but when training data more it will be generalized model.

CONCLUSION

Breast cancer is the most secretive and common cancer among women and rarely in men .It is a vital issue to get the faster and accurate diagnosis of the patient so that doctors can decide the treatment in due time. The mean accuracy value of cross-validation is 96.24% and XG Boost model accuracy is 98.24%. We analyzed the results of all algorithms and tried to find out the best possible one. It shows XG Boost is slightly over fitted but when training data more it will be generalized model. We have used standard dataset which is widely renowned dataset. In near future, we will try to enhance our work by managing a comparatively big dataset and adding some more functionality like the stage detection of breast cancer, treatment predictions and so on. The XG Boost algorithm is giving a better result with respect to the prediction accuracy, further we have considered these algorithms for the detection of cancer malignant.

ACKNOWLEDGEMENTS

The proposed research work is under the collaboration with our own KLE hospital and research center which is supporting with the number of samples. All the images are digitized and preprocessed with the standardization procedure of the digitization so as to have the normal sized images. We thank our supporting team of doctors from hospital for the subjective assessment of the images. We also thank to the KLE Dr. Prabhakar Kore Hospital and Medical Research Center for the ethical clearance, support with the images and help us to understanding clinical aspects of the digital images.

REFERENCES

Claudio Manna, Loris Nanni, Alessandra Lumini. October 2012. Artificial intelligence techniques for embryo and oocyte classification. Article in reproductive Biomedicine online.

Dr.R.H.Havaldar, Prof.S.S.Ittannavar.2016. Comparative study of Mammogram Enhancement Techniques for Early Detection of Breast Cancer. International Journal of Technology and Science (online) 2350-1111.Volume IX. Issue1. Pp.5-8.

Florin G, Marina G, Smaranda G, Elia El-Darzi. 2008. A statistical evaluation of neural computing approaches to predict recurrent events in breast cancer. 4th IEEE International Conference on Intelligent Systems. pp 38-43.

Hamhung Adi Nugroho, Faisal N, Indah Soesanti. 2014. Analysis of Computer Aided Diagnosis on Digital Mammogram Images. International Conference on Computer, Control, Informatics and Its Application Bandung, Indonesia.

Janghe R ,Anupam S , Ritu Tiwari3 , Rahul Kala Breast Cancer Diagnosis using Artificial Neural Network Models 4 1,2,3,4 Indian Institute of Information Technology and Management, Gwalior, India.

Mohammad S, Rabab K, Ward, Jacqueline Morgan Parkes, Branko P. 2009. Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer. IEEE journal of selected topics in signal processing. vol. (3). pp 46-52.

Nuryanti M, Harsa A Mat Skim, Nor H O. 2008. Neural Networks to Evaluate Morphological Features for Breast Cells Classification. IJCSNS International Journal of Computer Science and Network Security. VOL. (8). pp 51-58.

Rastghalam, R., & Pourghassem, H. 2016. Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images. Pattern Recognition. 51. 176-186.

Sagar Metri and Asha T. 2018. Patch Based Wiener filter for Image DE noising. International Conference on Computational Techniques, Electronics and Mechanical Systems .CTEMS.

Sagar Metri and Raghuvamsa GH .2012. Performance improvement in WiMAX networks with femto cells. International Conference on Computing, Electronics and Electrical Technologies. ICCEET.

Sujata N Patil, Uday V Wali , M K Swamy. 2019. Selection of Single Potential Embryo to Improve the Success Rate of Implantation in IVF Procedure using Machine Learning Techniques. International Conference on Communication and Signal Processing (ICCSP). Tamilnadu.0881-0886 IEEE.

Sujata N Patil, Uday V Wali, M K Swamy. 2016 IEEE. Application of vessel enhancement filtering for automated classification of human In-Vitro fertilized (IVF) images. International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT). 27978-1-5090-4697-3/ 16/.

Sujata N Patil, Uday V Wali, M K Swamy. 2018. Deep Learning Techniques for Automatic Classification and Analysis of Human in Vitro Fertilized (IVF) embryos. Volume (5). Issue (2). Journal of Emerging Technologies and Innovative Research (JETIR).

Sujata N Patil, Uday V Wali, M K Swamy. Performance Analysis And Classification of Human In Vitro Fertilized (Ivf) Embryos Using Vesselness Filters and Hough Transform Algorithm. Int J Recent Sci Res. 9(1). pp. 23475-23479.

Sujata Patil, Shweta Madiwalar V M Aparanji .March 2020. Artificial Intelligence for early Detection of Breast Cancer and Classification of Mammographic Masses. Volume 8. Issue 6. International Journal of Recent Technology and Engineering (IJRTE).

Umesh, B Ramachandra. 2015. Association rule mining based predicting breast cancer recurrence on SEER breast cancer data. International Conference on Emerging Research in Electronics, Computer Science and Technology – Mandya. India

Yuan Jiao MA, Ziwu WANG, Jeffrey Lian LU, Gang WANG, Peng LI Tianjin MA, YinfuXIE ,Zhijie ZHENG. 2006. Extracting Microcalcification Clusters on Mammograms for Early Breast Cancer Detection. IEEE International Conference on Information Acquisition. August 20 – 23. Weihai, Shandong. China. Pp 499-504.