Journal of Engineering Sciences

Vol 12, Issue 06, June /2021
ISSN NO: 0377-9254

ARTIFICIAL INTELLIGENCE APPROACH FOR BREAST CANCER CLASSIFICATION USING MACHINE LEARNING CLASSIFIERS

REKHA.B [1], DR. PRAVIN R KSHIRSAGAR [2], A.SUPRIYA [3], CH.SINDHUJA [4], A.RAJKUMAR [5]

1 Assistant Professor, Dept of ECE, AVN Institute of Engineering and Technology, Hyderabad.

2 Professor, HOD of Dept of ECE, AVN Institute of Engineering and Technology, Hyderabad.

3, 4, 5 B.Tech, Dept of ECE, AVN Institute of Engineering and Technology, Hyderabad.

ABSTRACT

In the field of assisted cancer diagnosis, it is expected that the involvement of machine learning in diseases will give doctors a second opinion and help them to make a faster / better determination. This article aims to evaluate the predictive models of machine learning classification regarding the accuracy, objectivity, and reproducible of the diagnosis of malignant neoplasm with fine needle aspiration. Also, we seek to add one more class for testing in this database as recommended in previous studies. We present four different classification methods: Decision Tree, Random Forest, Support Vector Machine and K-Nearest Neighbors (KNN) for evaluation. For this work, we used at University of Wisconsin Hospital database which is composed of thirty values which characterize the properties of the nucleus of the breast mass. According to our results, Random Forest 50 (95.02%) and 100 (95.42%) have the most significant area under the curve, being in the range of excellent prediction for the ROC curve metric, however, the Random Forest 100 area is larger than the Random Forest 50 area.

Keywords: Mammography, breast cancer, Decision Tree, Random Forest, Support Vector Machine and K-Nearest Neighbors(KNN).

1. INTRODUCTION

Breast cancer is one of the leading diseases that reflect an uncontrolled growth of abnormal cells in the breast. Due to the breast anomalies properties and the nature of the human visual perception, it is natural that, sometimes the abnormalities are missed or miss classified. As a result, unnecessary biopsies are taken. To mitigate this problem, computer aided diagnosis (CAD) system [1-2] has emerged. The proposed CAD system is implemented as the integrated system using image processing techniques and machine learning algorithms. CAD aims at the detection and localization of abnormalities at an early phase, which avoids the further spread of the abnormality. Breast cancer [3-4] is one of the leading diseases that reflect an uncontrolled growth of abnormal cells in the breast. Due to the breast anomalies properties and the nature of the human visual perception, it is natural that, sometimes the abnormalities are missed or miss classified. As a result, unnecessary biopsies are taken. In breast, normal cells [5-6] grow and divide at a particular time but in case of the cancerous cells, the cell growth is continuous and uncontrolled. Many researchers have addressed the issue of breast cancer detection from the past few years' later classification

approaches also discussed and presented by several authors. A new breast cancer diagnostic system by employing PSO-SVM framework is presented in [7], where the PSO aimed at mitigating the simplification ability of the SVM classification by concurrently tackle the essential kernel constraint set and recognizes the majority discriminative characteristic feature separation. In scheming categorization accurateness, the object utility quantity of SVs and quantity of characteristic features are concurrently followed as deliberation. Principally, during a sequence of observed experiments on standard database, PSO-SVM organization not merely exploits the simplification presentation but too choose the majority revealing characteristic features. Author in [8] reviewed supervised deep learning (SDL) area of research, conceptual groups and analyzes various techniques. They have proposed two unconventional combinations; the primary is founded on SDL representation mechanism utilized for feature drawing out, while the subsequent utilizes the imperative drawing out method. The analysis is followed by a comparative evaluation of the algorithms are relative performance as measured by several metrics. They have concluded by highlighting the potential research directions, such as the need for rule extraction methods. The recognition scheme for classification of tumor lesions appearing in mammographic X-ray images are addressed in [9].Genetic algorithm (GA) utilized FSS is defiant from clamor up to a definite stage and categorization rate is enhanced for GA used FSS method. FCM has separation the huge amount contour group clusters such that the level of alliance is burly for the features inside the similar groups and weedy for the features in dissimilar groups. GA explored the important contour features by concerning the magnificence of usual dispute. Utilizing three operatives similar to imitation, crossover and mutation, GA is able to choose important feature division. However, due to lack of accurate detection, efficient extraction of features and classification accuracy, conventional breast cancer diagnosis systems failed to produce acceptable outcome.

Most of the research work is focused on using optimization techniques to develop a Classification [10] and Diagnosis [11] of breast cancer from Mammographic images. The detection & classification of irregularities in Mammographic images are considered for investigation in this paper. Poor noise-to-signal ratio is a drawback in Mammographic images. The anatomically distinct structures are often seen with a very low contrast. Reliable standard image processing technique [12] is needed for its computation. Modification in image content is done in a highly controlled and reliable way without any compromise in clinical decision-making, but the presence of artifacts leads to 10 – 25% of tumors being missed by radiologists. Basic noise removal filters [13] cannot be applied on Mammographic images as they are not able to remove the artifacts effectively. If we use those fundamental filters then, image get corrupted and enhancement operation will not work. Image denoising is one of the significant topics in image enhancement [14] that deals with noise contained imagery, which are need to be preprocessed using various approaches. Nowadays, medical imaging field and its equipment's are improved noticeably. The existing mammographic image segmentation approaches failed to perform well in terms of sensitivity, false positive rate, accuracy, specificity

and improved classification when processing the images generated from the advanced image generation sources. In order to overcome this issue, various classification and diagnosis of breast cancer approaches [15-18] are presented to process the images generated from mammographic images.

Rest of the paper is planned as following, Section 2 deals about the detailed architecture of proposed methodology with its block wise operation. Section 3 deals about results and discussion with comparison to the state of art methods using quantitative metrics. Section 4 deals about the conclusion and future enhancements of proposed methodology followed by bibliography.

2. PROPOSED METHOD

K-means model, that is an unsupervised learning method, presents satisfactory results in comparison to PCA (Principal Component Analysis). These models were combined with four models based on two dimensionality reduction methods, being evaluated through accuracy, sensitivity and specificity metrics [8].Other classifier's performance was tested to assess the efficiency and effectiveness of algorithms in cancer prediction. Support Vector Machine (SVM) (97.13%), present the highest accuracy in this study, compared to other models tested (Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN)) [3].Random forest presents a huge accuracy (98,77%) when compared to other methods. The random Forest model also presents great results when compared the other machine learning models (Decision Tree, Support Vector Machine, Neural Network, and Logistics Regression). In this study were used two different data sets combined, as well as a ROC Curve indicator as a measure for assessment [11]. From the previous work and their results, we understand how important to carry out a new study, gathering the techniques that presented the most consistent results and comparing them using accuracy metrics and ROC curve. In addition, we propose a discussion of the parameters used, analyzing from different points of view.

A. Data Set

The data set 1 is formed by ten features which are presented below II

Table I: DATASET ATTRIBUTES

Atributte	Range
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class	(B/M)

Table II DATA SET FEATURES

Item	Attribute
A)	radius (mean of distances from center to points on the perimeter)
B)	texture (standard deviation of gray-scale values)
C)	perimeter
D)	area
E)	smoothness (local variation in radius lengths)
F)	compactness
G)	concavity (severity of concave portions of the contour)
H)	concave points (number of concave portions of the contour)
I)	symmetry
J)	fractal dimension ("coastline approximation" - 1)

This work employed the WDBC1 data set2 , which attributes are summarized in Table II. All data are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Altogether, the data set composition contains 569 rows and 33 columns. Among the 33 columns, the first two are the identification number and diagnosis (M = malignant, B = benign). All resource values are recorded with

four significant digits. The sample class distribution is such that 357 are benign and 212 are malignant, which is not considered a severe situation. All ten attributes are described by three characteristics: mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

B. Pre-processing

The data set was provided in a CSV file, containing 837751 registers. It was performed from a data set of 569 women, being: The first column of the patient identification code, which is not being used in the training process. The second column is Diagnosis, where 1 indicates Malignant, and 0 indicates benign. The rest of the columns are 30 numeric values that show the measurements of the cell nucleus. The last column was deleted due to contained only NaN values. For the cell nucleus, the inclusion, texture, perimeter, area, softness, compactness, concavity, symmetry and large fractal are measured ten times. The significant error, default, and lower values are the properties calculated, resulting in 10 x 3, 30 columns of input data. In our feature selection/extraction, we opted for the cross validation method. Cross-Validation is a technique that aims to understand how your model generalizes, or how it behaves when you predict a data you have never seen. This metric creating different training‘s and testing sets, to make sure that the model is performing well. In this case, instead of using only one test set to validate our model, we will use N others from the same data [12].

C. Classification Models

In the next sections, we did a briefly summarize of all classifiers used in this study.

K-Nearest Neighbor (k-NN)

KNN is a supervised learning technique that means the label of the data is identified before making predictions. Clustering and regression are two purposes to use it. K represents a numerical value for the nearest neighbors. KNN algorithm does not have a training phase. Predictions are made based on the Euclidean distance to k-nearest neighbors. This technique is applied to the prediction of breast cancer dataset since it already has labels such as malignant and benign. The label is classified according to the nearest neighbor to the class labels of its neighbors.

Decision Tree

Decision tree algorithms are considered an alternative for regressions and classifiers tasks. the Decision Tree Algorithms structure can be compared to a set of rules (If-then), classifying new samples and trying to develop an understandable and accurate model [14]. Thereby, the Decision Tree algorithm operates such as others Supervised Learning techniques, working with sets for training and tests. In their structure, the main idea of decision tree algorithms is in a first step, work with a classification problem (such as to determine whether or not something is alive) and establish features for each record (such weight and breath). All features are being divided into "nodes", each one with a decisive way (yes/no). Each yes or no decision could be considered a decision node, and at the end, you would have data that can’t be split further [15]. If a decision tree reaches an actual rank in a training set, could be not considered a good classifier, being more interesting a smaller tree that does not fit into all training data. A good accuracy would guide all the time to an improved decision tree [14].

Journal of Engineering Sciences
Vol 12, Issue 06, June /2021
ISSN NO: 0377-9254

Random Forest

Defined as an ensemble learner, Random Forest works creating multiple classifiers and regression trees, each one trained based on the subset of training examples and the subset of all given features at random [16]. Each decision tree, the input enters at the root of the tree and traverses down the tree according to the split decision at each node [3]. Although the Random Forest algorithms have many similar aspects of the Decision Tree process, this technique can handle high dimensional data and use a large number of trees in the ensemble. Also, there are some specific characteristics such as an effective method for estimating missing data, and method for balancing error in imbalanced data [16].

Support Vector Machine

Support Vector Machines (SVMs) is a supervised machine learning technique, having great theoretical foundations and excellent empirical successes [17]. The SVM has the constraint which makes the total weight for the positive class equal to that of the negative class. This kind of technique has been applied to different classification tasks such as text classification, object recognition, as well as prediction tasks. There are many advantages to support vector machines use, such as the number of dimensions that are greater than the number of samples, the subset uses of training points in the decision function (called support vectors), and also, be able to apply different kernel functions for decision function [18].

3. RESULTS AND DISCUSSIONS

3.1. Dataset

Towards the analysis of our algorithm, we used Jupyter Notebook, python modules (pandas, matplotlib, bumpy) and a scikit-learn framework to process ML algorithms. The following evaluated methods were:KNN, Decision Tree, Random Forest, and Support Vector. Random Forest was performed, which were determined 50 and 100 collections of trees. First, we performed training for 70 % of the dataset (398 randomized records), applying the cross-validation method verifying all metrics before cited. After, we ran two types of tests: without partition and with 50 % partition (one partition with 86 and one with 85 random registers) of the remaining 30 % of the dataset. The Table 4 shows the results of training and test(cross-validation).

Total 1000 mammographic images are adopted for this experiment analysis where 400 of malignant, 400 of benign and 200 of normal mammographic X-ray images with the consideration of patient mean age around differerent ages and ranging from 18 to 81. The breast grazes assortment from 2mm to 20mm in mass and several patients contain several grazes whereas some other patients might have merely one. Figure 1a shows the original input image datasets, Figure 1b shows the preprocessed noise free enhanced images using NSCT transform. Figure1c shows the detection of breast cancer using APFCM clustering method and classified using SVM methodology respectively.

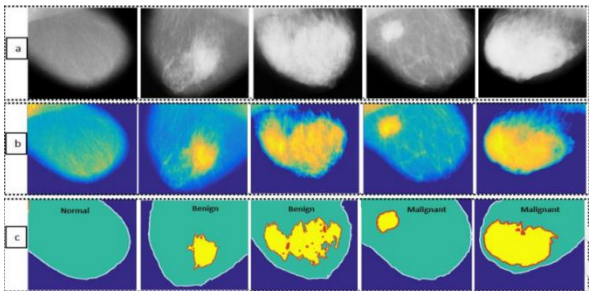

Fig.1. detection and classification procedure of proposed method
(a) Original images, (b) Preprocessed images, (c) segmented and classified image

Journal of Engineering Sciences
Vol 12, Issue 06, June /2021
ISSN NO: 0377-9254

4.2. Evaluation criteria

For valuation of classification outcomes, we utilized three qualitative metrics such as specificity, accuracy and sensitivity. The accuracy can be defined as out of certain random test cases, how many outcomes give the perfect classification output. The sensitivity is defined as individual classification accuracy, how much the method is sensitive towards the malignant and benign cancers. And specificity is defined as the how much accurately the location of tumor is recognized.

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

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

where TP conveys the amount of test cases properly recognized as malignant, FP conveys the amount of test cases improperly recognized as malignant, TN conveys the amount of test cases properly recognized as benign and FN is conveys the amount of test cases improperly recognized as benign.

Table 3. Performance of quality metrics using existing and proposed hybrid SVM model

Method	Accuracy (in %)	Specificity (in %)	Sensitivity (in %)	Recall (in %)
D Tree [16]	76.09	75.51	77.40	78.40
Perceptron [17]	80.01	79.18	80.72	81.72
RF 100 [1]	80.42	80.18	81.81	83.81
RF 50 [13]	90.11	89.28	90.36	94.36
SVM	95.91	95.81	96.34	97.34

In the training procedure, network limits were attuned by the preparation slaughter and after that the justification dataset would be utilized to check the matching amount of the attuned system. The matching curvatures of system depend on network testing slaughter and training loss slaughter. In order to additionally calculate the planned technique, we contrasted it with pair of NN-contained methods utilized in [14], [15]. For the categorization, we adopted D Tree [16], RF 100 [1], RF 50 [13], SVM classifiers from the literature for comparison with the proposed hybrid SVM classifier model. Table 1demonstrates that quality evaluation criteria of existing and proposed classifiers, where proposed hybrid SVM classifier outperforms the conventional classifiers to distinguish the benign and malignant from the mammographic X-ray images.

Table 4 TRAIN AND TEST(CROSS-VALIDATION) RESULTS

	Test		Train	
Classifier	Mean	Std	Mean	Std
D Tree	0.833231	0.024752	0.974431	0.013073
Perceptron	0.833649	0.052749	0.847188	0.060723
RF 100	0.871742	0.017536	0.977251	0.013075
RF 50	0.867416	0.018830	0.976589	0.012746
SVM	0.831416	0.064201	0.885632	0.063091

Table V BASELINE PERFORMANCE

	AUC		Accuracy		Average		
Classifier	Mean	Std	Mean	Std	Precision	Recall	F1 Score
D Tree	0.8700	0.0372	0.8531	0.0283	0.8442	0.8400	0.8403
Perceptron	0.9248	0.0198	0.8156	0.1146	0.7490	0.7622	0.7415
RF 100	0.9464	0.0129	0.8913	0.0251	0.8909	0.8744	0.8805
RF 50	0.9457	0.0133	0.8895	0.0262	0.8884	0.8730	0.8786
SVM	0.8857	0.0307	0.8162	0.0913	0.7852	0.7781	0.7692

Table VI ROC SCALE

Classifier	AUC	Roc Scale
D Tree	83.24 %	Good Prediction
Perceptron	91.07 %	Excellent Prediction
RF 100	95.45 %	Excellent Prediction
RF 50	95.02 %	Excellent Prediction
SVM	83.24 %	Good Prediction

5. CONCLUSION

Our study presented a set of classification models, trying to find the best model to classify Breast Cancer according to our data set (WDBC). For this proposal, we selected four different techniques of machine learning, which were considered in other studies with similar proposals. Random Forest was divided between two models: 50 and 100 trees collections. Furthermore, we use a group of metrics to evaluate all results. In this sense, we gave special attention to accuracy and ROC curve measures, proposing a comparison and discussion between these metrics. The outcomes obtained from experiments have been analyzed across, data tables and charts. Regarding our results, Random forest models presented the best results for the accuracy and the Roc curve. Which model has the highest accuracy, objectivity, and reproducibility. It is not so easy to see if one algorithm is better than another only by looking at the error - rate and accuracy values, since there is no classification algorithm for all the challenges to be overcome. It is important to understand the power and limitations of different classifiers, and there is a scale for the challenge/community to use it in the best possible way in order to compare the models in question. Breast Cancer has provided many studies in recent years, through different approaches as computing vision, classification, and prediction. As a future work, we considered an improvement in predictions, testing approaches in databases containing images.

REFERENCES

[1] Kaymak, Sertan, Abdulkader Helwan, and DilberUzun. "Breast cancer image classification using artificial neural networks." Procedia computer science 120 (2017): 126-131.

[2] Singh, Anuj Kumar, and Bhupendra Gupta. "A novel approach for breast cancer detection and segmentation in a mammogram." Procedia Computer Science 54 (2015): 676-682.

[3] Danala, Gopichandh, et al. "Classification of breast masses using a computer-aided diagnosis scheme of contrast enhanced digital mammograms." SVMals of biomedical engineering 46.9 (2018): 1419-1431.

[4] Jouni, Hassan, et al. "Neural Network architecture for breast cancer detection and classification." 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET). IEEE, 2016.

[5] Abdel-Ilah, Layla, and Hana Šahinbegović. "Using machine learning tool in classification of breast cancer." CMBEBIH 2017. Springer, Singapore, 2017. 3-8.

[6] Li, Xingyu, et al. "Discriminative pattern mining for breast cancer histopathology image classification via fully convolutional autoencoder." IEEE Access 7 (2019): 36433-36445.

[7] Wang, Zhiqiong, et al. "Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features." IEEE Access 7 (2019): 105146-105158.

[8] Sebai, Meriem, Tianjiang Wang, and Saad Ali Al-Fadhli. "PartMitosis: A Partially Supervised Deep Learning Framework for Mitosis

Detection in Breast Cancer Histopathology Images." IEEE Access 8 (2020): 45133-45147.

[9] Saha, Monjoy, and Chandan Chakraborty. "Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation." IEEE Transactions on Image Processing 27.5 (2018): 2189-2200.

[10] Zhang, Xiaofei, et al. "Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks." IEEE transactions on nanobioscience 17.3 (2018): 237-242.

[11] Eltrass, Ahmed S., and Mohamed S. Salama. "Fully automated scheme for computer-aided detection and breast cancer diagnosis using digitized mammograms." IET Image Processing 14.3 (2019): 495-505.

[12] Das, Asha, Madhu S. Nair, and S. David Peter. "Sparse representation over learned dictionaries on the riemSVMian manifold for automated grading of nuclear pleomorphism in breast cancer." IEEE Transactions on Image Processing 28.3 (2018): 1248-1260.

[13] Khan, Hasan Nasir, et al. "Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network." IEEE Access 7 (2019): 165724-165733.

[14] K. Sohn et al., "Learning and selecting features jointly with point-wise gated BoltzmSVM machines," in International Conference on International Conference on Machine Learning, 2013, pp. II–217.

[15] Q. Zhang et al., "Deep learning based classification of breast tumors with shear-wave elastography," Ultrasonics, vol. 72, pp. 150–157, 2016.

[16] S. Z. Erdogan and T. T. Bilgin, "A data mining approach for fall detection by using k-nearest neighbour algorithm on wireless sensor network data," IET Communications, vol. 6, no. 18, pp. 3281–3287, 2013.

[17] E. Karthikeyan and S. Venkatakrishnan, "Beast Cancer Classification using SVM Classifier", International Journal of Recent Technology and Engineering, vol. 8, no. 4, pp. 527-529, Nov. 2019.

[18] P. B. Chandra and S. K. Sarkar, "Detection and classification technique of breast cancer using multi kernel SVM classifier approach", In Proc. of International Conference on Applied Signal Processing, Kolkata, India, IEEE, Jul. 2019.

Authors profile:

A.SUPRIYA
Final Year ECE
ambatisupriya550@gmail.com

CH.SINDHUJA
Final Year ECE
ch.sindhuja1712@gmail.com

Journal of Engineering Sciences

Vol 12, Issue 06, June /2021
ISSN NO: 0377-9254

A.RAJKUMAR
Final Year ECE
raj.aed7@gmail.com

Ms.Rekha.B,
rekha20.ec@gmail.com
M.Tech. Assistant Professor