

Design of an early prediction approach BC-DEF for breast cancer using Ensemble Learning for Fisher Linear Discriminant feature selection approach

^[1] J. Jayapriya, ^[2] Dr. G. Ramachandran, ^[3] Dr.T.Priyaradhikadevi

^[1] Research Scholar, Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Tamilnadu, India

^[2] Associate Professor, Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Tamilnadu, India

^[3] Professor & Head, Dept. of Computer Science and Engineering
Mailam Engineering College
Mailam – 604304
Tindivanam
Tamilnadu, India

Email: ^[1] priya.jaya.83@gmail.com, ^[2] gmrama1975@gmail.com
^[3] prdcsehodmec@gmail.com

Abstract: This research work (BC-DEF) can be presented as a breast cancer prediction framework using diagnostic mammogram images and a clinical dataset based on biopsy. The proposed solution aims to provide an explainable classification model using strongly dependent quality of the extracted mammographic features of breast cancer affected patients. The classification approach uses BC-DEF as proposed framework which defines an Ensemble Learning for Fisher Linear Discriminant for Breast Cancer prediction. The proposed work BC-DEF focuses on improved performance metrics compared to traditional classifiers based breast cancer tumour pattern recognition algorithms. Observed accuracy is found to be 98.7 % better than other classifier models.

Keywords: Breast Cancer, Fisher Linear Discriminant, Ensemble Learning, Tumor

1. INTRODUCTION:

Research works [2][15] considers breast cancer to be a common disease prevalent in most countries. This disease can be defined as the development of breast cells unnaturally in size and uncontrollably, due to the development of tumours as mass density inside the breast. Medical experts [7] proved that cancer growth cells need to be addressed as lumps of mass present in the breast as maligns (M) and benign(B), else they can spread to other body regions including bones, liver, and lungs.

WHO [11] report suggests that the death ratio due to breast cancer is observed at 9.2 million people affected by malicious chances of cancer, where 1.7 million confirmed for breast cancer alone, while other cancer issues also carry importance. Researchers and medical experts admit that predicting breast cancer is a challenging task using diagnostic mammogram images alone; hence , clinical data analysis based on biopsy suggests a higher prediction rate[17]. Expert suggestions [3] on predicting cancer at an early stage consider clinical analysis of a lymph node or tumor as small and has not spread, being defined as Stage-1 and can be localized for treatment. Screening for cancer symptoms over regular periods of time is defined as a reliable method of early breast cancer detection. This will be a reliable method of predicting the prognosis of a woman with this disease. Normally, breast cancer diagnosis is clinical and biological in nature, performed by doctors. The development of more effective diagnostic techniques and improved treatment planning has attracted significant attention to breast cancer cases. The most common techniques [10] available for breast cancer

diagnosis are biopsy, mammography, ultrasound (US) imaging, magnetic resonance imaging (MRI) scans and experimental breast imaging.

BC-DEF adopts biopsy-based clinical data for predicting breast cancer to prevent the stage-3 approach of mastectomy[7]. Cancer detection pathologist [18] suggest supporting clinical biopsy results for decision making and to perform diagnosis based on the data collected by medical mammographic images. Various researches show that ML techniques [8][11] are helpful for decision making in breast cancer prediction. This paper works on various ML Classification techniques, such as Support vector machine(SVM), K-Nearest Neighbor (KNN), Random forest(RF), Logistic regression (LR) and Decision Tree(DT), as well as various ensemble learning techniques for verification and evaluation on breast cancer datasets and their different performance measure, are discussed in Section-4.

The need for understanding complex correlations arising from mammographic images and clinical datasets to suggest relationships between multiple variables and uncover hidden insights between resonant of the observed image is the motivation of this research work. This method enables us to develop a predictive model from a given dataset and make predictions regarding future results. The primary concept associated with machine learning is to identify patterns in breast cancer data that possess higher similarity with the aim of making accurate predictions[10]. Adopting machine learning (ML) algorithms with effective data analytics over automated tool support can provide an expert suggestion to work on medical diagnosis and make proper decisions, as well as differentiate between malignant and benign tumour[9].

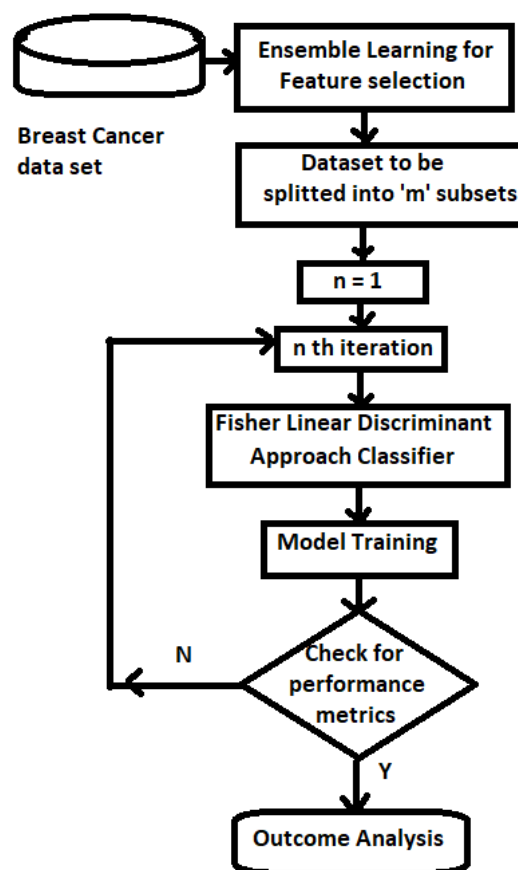


Fig-1: BC-DEF Architectural flow for Breast Cancer prediction

Numerous research works in the expert community are carried out to detect the causes of breast cancer and develop an early preventive measure. Accurate breast cancer detection and support for creating a novel approach for cancer treatment could be suggested as a preventive measure to minimize its mortality rate [12].

Diagnosis of breast cancer demands timely treatment and prediction at an early stage which requires an improved outcome associated with preventing the locoregional spread or avoiding distant metastasis.

The prediction performance of the ML models can be analysed using classification models and evaluating criteria for cross-validation such as accuracy, error and ROC. Pattern recognition techniques such as the Fisher Linear Discriminant (FLD), Probability Density Function (PDF) estimates were used to construct and evaluate a two class algorithm, malignant versus benign tissue areas. The paper is organized into five sections, with Section-1 introducing the need for research in Breast Cancer prediction as follows

1.1 Objectives of research work

The primary objective of this research work is of two fold

- a) To analyse, understand and identify primary features related to breast cancer among women and men.
- b) To propose feature reduction approach in order to support on vital features of breast cancer at an early stage with accuracy.

The proposed frame-work BC-DEF suggests the design of a multi-class approach to work on pre-processed, standard datasets for appropriate classification, feature minimization and prediction models. To predict the early chances of breast cancer, a dataset works on Ensemble Learning classifier to classify and suggest different similar groups to be established between sets of data, such that the classifier supports maximizing the chances of breast cancer prediction. Fisher Liner discrimination model is proposed to minimize the number of features required for prediction.

1.2 Mammography and its importance in predicting breast cancer

Among these techniques, mammography is the first modality for screening or detection of breast cancer[1]. In this technique, the doctor looking for the symptoms of breast cancer uses an X-ray image of the breast (a mammogram). Usually, symptoms like small white spots called calcifications, lumps or tumors and other suspicious areas indicate signs of breast cancer. However, the radiologist's decision may vary while interpreting the mammogram [5]. Moreover, mammography suffers from limitations such as false-negative results[4], false-positive results, etc. Ultrasound is often used to examine some changes in the breast that can be felt but not visualized on a mammogram. It may also be used for women with dense breast tissues. Further, it is also used to guide a biopsy needle into a suspicious area that can be taken out and tested for cancer[9]. However, the breast ultrasound imaging test solely depends on the skill and experience of the doctor interpreting the images. Breast biopsy is a technique usually followed after getting suspicious results from a screening test. Cells from a suspicious area are taken out by a surgeon and tested in the laboratory to detect the presence of cancer cells. There are different types of biopsy such as fine needle aspiration cytology (FNAC) biopsy, core needle biopsy, surgical biopsy and lymph node biopsy. FNAC is a standard work-up for patients with suspected breast cancer. A doctor takes the tissue sample, irrespective of the prescribed biopsy, and a pathologist for the final diagnosis examines it. The judgement given by the multidisciplinary oncology board consisting of a radiologist, oncologist and pathologist is the final decision in breast cancer diagnosis. However, machine learning techniques could assist them in improve their diagnostic capability. Fisher Linear Discriminant feature selection approach on feature reduction, which can reduce the errors committed in prediction due to inexperience or an unsupervised model.

1.3 Originality of work and its importance

Data considered for the research study is obtained from the UCLA repository [4] where 570 different ethnic women are gathered for analysis. Here 273 are obtained from malignant women (69.20%) and the balance belongs to benign (30.83%) women category. The primary work discussed in this research paper suggests an early prediction of breast cancer using the Fisher Linear Discriminant Selection approach for predicting dominant features and Ensemble Learning as a classifier model. The data set analysis includes mammographic data of both benign and malignant cases, and the medical dataset includes blood type, breast tumor size, colour and examining surrounding adipose tissues to analyse the extent of the individuals' exposure to metals such as zinc (Zn), lead (Pb), cadmium (Cd), iron (Fe), arsenic (As), chromium (Cr), and nickel (Ni).

2. LITERATURE SURVEY

A wide variety of methods based on machine learning, expert systems and soft computing have been proposed in the literature to solve the breast cancer diagnosis problem [2][6]. Earlier, data mining and machine learning techniques were integrated into a computer aided diagnosis (CAD) system for the classification of breast cancer. The successful execution of this approach transformed the breast cancer diagnosis into a 2-class (benign or malignant) classification problem [5]. The objective of breast cancer classification is to identify an unknown tumor sample (test sample) as benign or malignant by using labelled training samples. Almost all the methods consider WBCD taken from the UCI machine learning repository [12] for experiment.

Authors Tabrizchi et al [8] proposed a 10 – fold cross validation in conjunction with the C4.5 tree method. The author was silent about the effect of changing the training samples on the classification accuracy. [14] used a rule induction algorithm based on the approximate classification method to achieve an accuracy of 96%, which is low. [10] applied the supervised fuzzy clustering technique and obtained an accuracy of 95.57%. This is also a lower classification rate and thus, not fit for clinical usage. Arunkumar et al[3] employed a hybrid machine learning method for breast cancer diagnosis. The method involves two stages. Firstly, a fuzzy-artificial immune system was applied to reduce the dimension of the dataset. Secondly, k-NN performs the classification in the reduced space. In this manner, the processing time for classification is reduced. The method achieved a high classification accuracy of 99.14% via 10-fold cross validation. However, the authors did not investigate the effect of partitioning the training and test samples.

[14] suggested a technique that integrates wrapper-based and filter-based feature extraction methods. Such feature extraction methods are efficient enough to select a subset of features from the original feature space, without losing significant information. The method achieved an accuracy of 99.5%. The method suffers from computational complexity. Recently, [13] extended the idea of feature selection to cluster analysis techniques to improve their operation. The author proposed a hybrid intelligent technique for breast cancer diagnosis. [11] used rough set (RS) theory for feature selection. The authors implemented RS to select the significant optimal feature followed by support vector machine (SVM) for classification. They applied a subset that contains five optimal features. These authors worked in a high dimensional space using SVM[16].

[7] applied five different classifiers using neural networks for breast cancer diagnosis. The Levenberg-Marquardt algorithm is used to train the dataset for different classifiers. [19] proposed a hybrid method (K-means + SVM) and achieved a classification accuracy of 97.38% using 10-fold cross validation. The above authors used highly computationally complex methods for classification.

[16] implemented a back propagation neural network (BPNN), a binary coded genetic algorithm (GA) and a real coded GA for breast cancer diagnosis. [8] explained a novel artificial metaplasticity multilayer perceptron (MLP) for breast cancer diagnosis. [17] trained an MLP using an imperialistic competitive algorithm (ICA) and particle swarm optimization (PSO) for 40 iterations to achieve classification accuracy of 97.75% and 97.63% respectively. [9] implemented a rotation forest artificial neural network (ANN) using 9-classifiers and achieved a classification accuracy of 98.05%. [12] [18] proposed a novel technique for initializing and updating PSO for feature selection to achieve an accuracy of 94.74%. It is observed that most of the authors used neural networks for classification, which is inherently complex.

[9] proposed a genetically optimized neural network (GONN) algorithm to carry out the breast cancer classification. The method used GA to develop an optimized neural network architecture. The method implements new crossover and mutation operators for GA. However, the computational complexity is higher due to the use of GA and ANN. Further, the authors investigated the diagnosis problem on WBCD only. Moreover, to deal with real life problem, efficiency of methods for breast cancer represented by imaging techniques like mammography, US etc. needs to be evaluated. The imaging techniques define the diagnostic features based on the analysis of the images and are usually not very well correlated with a class. In this paper, our focus is on the evaluation of GNRBA on two different breast cancer datasets (WBCD and WDBC). The performance of GNRBA on different imaging modalities is beyond the scope of the paper and can be considered as our future work. [2] [7] proposed a knowledge-based system, that uses expectation maximization to cluster the data into similar groups. The method uses classification and regression trees to generate a set of fuzzy rules. Classification of breast cancer is done based on fuzzy rules. Further, to deal with the multi-collinearity problem, principal component analysis (PCA) is used. The method shows good classification accuracy on WBCD and

mammography mass datasets. However, the method is based on non-incremental data mining techniques and evaluated on small datasets. Sometimes, classification algorithms may be influenced by certain properties of the diagnostic images such as mammography. Early detection of breast cancer may depend on factors like differences in instrument setting or breast positioning by the operator. Taking these limitations into account, [6] proposed an adaptive artificial immune based system to investigate the properties of classification methods, applied to mammography images. The classification methods are trained to measure bilateral asymmetry between paired regions of the right and left breasts. Still, such a system requires proper investigation to identify focal and global asymmetric features. Additionally, attention is required to extract features to localize the source of asymmetry. [10] proposed an automatic quantitative image analysis method for the classification of breast cell nuclei as benign or malignant. The author uses breast cell histopathology (BCH) images. Segmentation of cell nuclei is done by using wavelet decomposition and multi-scale region growth. A double-strategy splitting model is used to partition the overlapping cells. A hybrid technique, which includes SVM and a chain-like genetic algorithm to extract optimal features (shape-based and texture-based features) for the classification of cell nuclei, is presented. The method shows a good classification accuracy of 96.19%, using 68 BCH images. However, to prove the robustness of the method, evaluation on a large dataset is preferred.

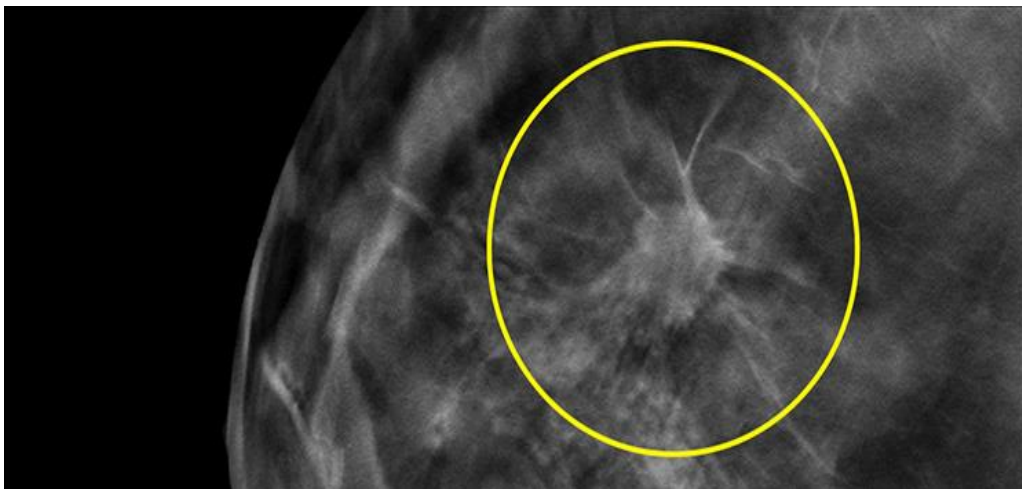


Fig-2 : Mammographic image (sample) with lymph nodes (encircled)

Research in the field of Breast Cancer detection has improved and supports computer-aided diagnosis (CAD) (Fig-2) systems which focus on gaining attention to the breast cancer classification problem for automatic diagnosis. [13] proposed a CAD system based on a deep-belief neural network. The network uses back propagation with Levenberg-Marquardt learning. The method uses WBCD for breast cancer classification. Nevertheless, such an algorithm has an inherent high computational complexity. Therefore, it is difficult to implement them on any commercially available hardware, to assist doctors in the early detection of breast cancer. [7] [14] proposed an adaptive CAD system for the classification of breast tumor. The CAD system utilizes tumor sizes for classification. US images of breast tumor detected during screening are used for classification. The CAD system uses quantitative morphological and texture features for classification. However, speckle noise and other artifacts that are inherent in the US image may degrade the performance of such a system. To overcome such problems, [14] proposed a super-resolution based CAD system. The system uses multiple US images of breasts instead of a single image. The CAD system uses four different steps like super-resolution computation, segmentation, feature extraction and classification to detect tumors and classify them as benign or malignant. The idea can be extended to other medical imaging modalities. Nowadays, a new ultrasound elastography imaging technique called shear wave elastography (SWE) is increasingly used for the classification of breast lesions. SWE provides the opportunity to evaluate the elasticity parameters of breast tissue, which is important information about cancerous tissue. [13] proposed a method for automatic characterization of malignant breast lesions. The method uses SWE to evaluate discrete wavelet coefficients at three different levels. Features like run length statistics and Hu's moments are extracted from the coefficients.

Significant features are then extracted by using sequential forward selection methods and ranked using the Relief F ranking technique. Different classifiers for the classification of benign and malignant lesions use ranked features. The method achieved a classification accuracy of 93.59%.

Data analysis and machine learning techniques were used by Hanker et al.[28] to create a paradigm for diagnosing autism in babies. Their structure also made extensive use of SVM-based training of data models and of data analysis in general. It was able to attain an accuracy of 89% but will need to be improved using DL and biomedical imaging in order to be more efficient. [12] used ML algorithms to diagnose ASD early enough to provide the most effective treatment for the condition. They used the UCI data set to test their own method. The technique can be used in conjunction with other diagnostic tools for autism spectrum disorder, such as EEG and MRI scans. For SVM and DL algorithms, in particular, an intelligence detective needs to be constructed.

Recently, sparse-based representation methods are being successfully implemented in pattern recognition for face recognition and verification [4][14]. It requires that a test face image be sparsely represented by the contributions of each class. A group of similar training face images represents a class. For each class, the contribution ability is represented by the weighted sum of the training images. The class contribution ability reflects the contribution of each class on a test image. Then the test face image is classified as the class having the maximum class contribution ability. Here, sparse indicates that the weighting coefficients of some of the training face images in a class are zero with respect to the matched class. The extent of sparsity of the weighting coefficients can be evaluated by using the l_1 -norm of the weighting coefficient vector [17]. A smaller norm indicates a stronger sparsity. In reality, conventional sparse-based representation methods tend to minimize the l_1 -norm of the coefficient vector. It has been found that both the decision tree and the XGBoost classifier have the highest accuracy 97% among other models and the highest AUC of 0.989 obtained for the XGBoost classifier.

3. BC-DEF PROPOSED WORK : FRAME WORK FOR EARLY PREDICTION OF BREAST CANCER

The proposed breast cancer prediction work BC-DEF works on pattern recognition techniques for dimensionality reduction and a feature selection approach applying Fisher Linear Discriminant (FLD)[6], along with Ensemble Learning model as a classifier to propose Probability Density Function (PDF) curves. The PDF curves were used to construct and evaluate a two-class algorithm, malignant versus benign tissue areas. This approach searches for a transformation matrix T which suggests mapping the original dataset to a space S with a low dimension while maximizing the separable spaces between classes. FLD is normally found to suffer from small sample size, bias and overfitting problems when dealing with high-dimensional breast image data.

3.1 Dataset

Dataset WDBC, which is maintained on the UCI repository, consists of 570 records of patient details with all patient information such as diagnosis of cancer, time of discharge and maintains as both categorical and numeric data. WDBC dataset [12] which possesses clinical data, determines whether a cancer is benign or malignant from the mammogram image of each patient. To extract features of the affine needle aspirate (FNA) of a mass in the breast, a digitized image can be collected and describe the properties of the nuclei in the image. This dataset contains 570 total records, where 213 are malignant (M) and 357 are benign (B) records. The feature of analysis includes patient details including Patient ID number, type of diagnosis, radius of breast, texture, perimeter, location or area of lymph, colour of lymph mass collected for analysis, compactness, concavity, concave points, symmetry, and fractal dimension, with 13 total features. Each image includes the observed mean, standard error, and worst or largest lymph as calculated features. Data is randomly selected for training and testing and split in the ratio of 80:20 respectively. Model effectiveness is verified and evaluated using the test data which can be trained with the help of training sets. Feature prediction analysis suggests the possibility of whether the patient can be affected or not.

3.2 Model and Approach

Fisher Linear Discriminant (FLD) is considered as a dimensionality reduction with feature extraction method among pattern recognition approaches. A total of 3 training sets are used; each set consists of 190 sub-

images, each comprising 95 abnormal and 95 normal sub-images, based on the domain of clusters. Patient's clinical and pathological characteristics [19] are obtained from a consistent medical record history. Breast Cancer features such as invasive ductal carcinoma (IDC), ductal carcinoma (DC) are gathered among the majority of cancer patients (86.1%). Diagnosis suggests that stage I, II and III breast cancer accounted for 14.6%, 55.2% and 32.2%, respectively.

Fisher linear discriminant is defined as a standard benchmark approach for linear discrimination between two classes C1 and C2 in the multidimensional space G.

$$G = \frac{\det(Y_t \cdot S_b Y)}{\det(Y_t \cdot S_w Y)} = [g_1, g_2, \dots, g_m] \quad (1)$$

G is considered the Fisher criterion which focuses on maximizing the observed values among the between-class (S_b) scatter while minimizes the within-class scatter (S_w). The transformation G is another projection into the Eigen space V (as defined in equation – 4)

3.2.2 Sub Image set

The dataset for analysis is defined as Testing set and training set being represented as $A_{mn} = [a_1, a_2, \dots, a_k]$ where $m = 2500$ (pixel size 50×50) and $n=1,2,\dots,95$. The number of sub set images determined by the pixel location of lymph nodes or the availability of tumour cells in the breast.

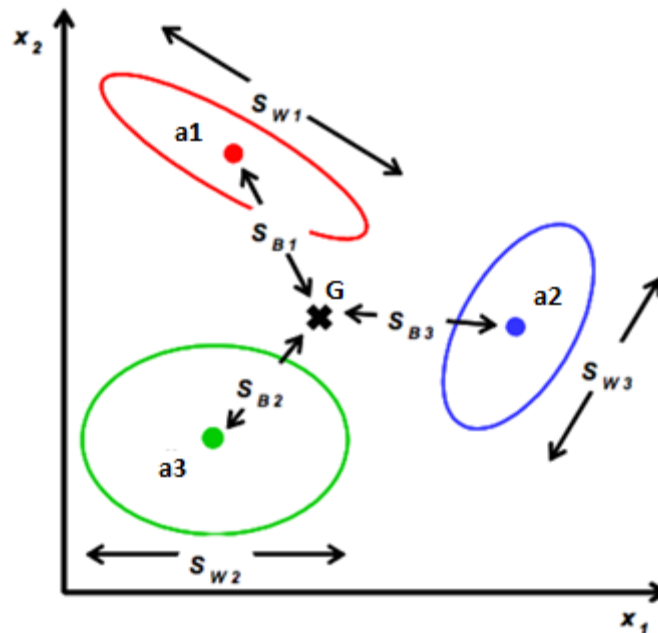


Fig-4: FLD approach for prediction of lymph node in mammographic image

3.2.3 Unsupervised Learning

Each sub-image A_{mn} of the training set is converted into a column vector g_i of dimension 2500×1 . Then, a training matrix A_{mn} is formed by placing the sub-images as columns. The column vector A_{mn} of dimension 2500×1 , and the row-wise mean of the matrix B_{mn} are calculated. The row wise matrix B_{mn} is formed by repeating the column vector A_{mn} a number of times equal to the number of sub-images in B_{mn} . Observed sub-image deviation from the row-wise mean of the sub-images is calculated as $C_{mn} = A_{mn} - B_{mn}$ ----- (2)

$$C_{mn} \text{ as covariance matrix is computed as, } C_{mn} = \frac{1}{r} \sum_{n=0}^r C_{mn} C_{mn}^T \quad (3)$$

Here r is the row of matrix A and B , n is the size of matrix as 95 sub-images in this study.

Equation (1) can be expanded as follows :

S_w is within-class scatter matrix, representing the scatter of a single class and S_b is between-class scatter matrix, representing the scatter of different classes as shown in Fig-4.

$$(4) \quad S_b = \sum_{i=1}^C \sum_{Y_i \in n} (Y_k - \rho) (Y_k - \rho)^T$$

$$(5) \text{-----} S_w = \sum_{i=0}^C (\rho_i - \rho) (\rho_j - \rho)^T$$

Here C is no of classes represented, ρ is mean of classes C1 and C2 and their images in training set.

The matrix transformation G is projection of sub-image of classes into Eigen space (V), where

$$S_b g_i = \alpha_i S_w \cdot g_i, \text{ where } i = 1, 2, \dots, m \text{-----}(6)$$

3.2.4 Classification Module

This module works on the effective euclidean distance ‘d’ between the testing matrix T and each column of the Fisher linear space Fwg which is analysed using equation (6) and eigen value (V). FLD approach identifies the nearest neighbor node to the test sample, which is classified in relation to calculated distances. Then, the test sample is assigned to the class of the nearest neighbor

General characteristics of study population

Characteristics	N	Percentage
Age (years)		
<45	78	51.7%
≥45	73	48.3%
Size of mass		
T1	37	24.5%
T2	93	61.6%
≥T3	21	13.9%
Node status		
N0	69	47.9%
≥N1	82	55.7%
Cancer Intensity		Stage
I	20	13.2%
II	85	56.3%
III	46	30.5%
Pathological type		
Ductal	130	86.1%
IDC	21	13.9%
Lymph Node Shape		
Color	26	17.2%
Blood Type	T	
Country Origin	T	

Based on the data preparation achieved, data training is carried out with various machine models such as SVM, KNN, and Decision Tree used to train the data. Based on the trained data obtained, the proposed model system BC-DEF performance is evaluated and verified on test data. Multiple variable ML algorithms are verified and applied over a learning ensemble classifier. The goal of ensemble learning is to produce a single superior predictive model by combining the results of multiple learning algorithms. This approach is carried out by combining the Random forest bagging technique [], Adaboost [] and XGBoost boosting techniques [] as supervised learners in order to boost the predictive ability of the model. Following different Classification

Algorithms and three ensemble learning techniques are applied to a pre-processed breast cancer dataset to decide whether it is malignant or benign.

KNN: K-Nearest Neighbourhood approach adopts the training data to classify new data points based on the similarity update. To assign a label to new data point it calculates its distance from a different label point and finally finds nearest neighbour.

- SVM: Support Vector Machine is a classification technique that can be defined as the requirement of prior distribution knowledge for the classification method. This method uses hyperplanes to classify data points.
- Decision Tree: In this technique instances are classified using feature value. For splitting it uses the Gini Index, Information Gain. Leaf node indicates the label.
- Random Forest: This approach is defined as an ensemble technique, where multiple DT are created on multiple samples of data whose majority votes are verified and classification is analysed
- Early detection of cancer is critical to improving breast cancer survival and reducing the high mortality rate in BC.1.Despite early detection and the advent of new treatments, about 50% of patients will develop distant metastases during their follow-up time.⁸ Therefore, a precise and reliable system is required for the early diagnosis of tumors. Results suggest that patients of ages are vital such that more than 1.29% of breast cancer patients are under 45 years old.

4. RESULTS AND PERFORMANCE ANALYSIS

On the Wisconsin Breast Cancer dataset, various models are used, and their performance is measured using Accuracy and AUC, precision, and recall. Performance of other models is compared based on accuracy measure and the ROC curve. The results are verified with standard ML algorithms and ensemble techniques. Analysis is carried out based on validation with actual and predicted labels as well as True Negative (TN), False Negative (FN) True Positive (TP) and False Positive (FP). Precision is defined as the number of positive class predictions that are actually positive class predictions, as shown in equation 1. The recall is the number of correct positive class predictions made out of all correct positive examples in the dataset and it is calculated as shown in equation 2. The F1 Score is derived from the weighted average of Precision and Recall; it is calculated as shown in equation 3. Accuracy is calculated by using the confusion matrix as shown in equation 4; it tells how the tuples in training and testing data are correctly classified.

Table 1. Classifier Performance for proposed Ensemble against other approaches

Classification Parameter	Breast Cancer Analysis	Precision	Recall	F1- Score	Accuracy
Learning Ensemble	Benign	0.96	0.94	0.92	0.98
K-Nearest Neighbor	Benign	0.93	1	0.89	0.94
SVM	Malignant	0.97	0.92	0.96	0.96
Decision Tree	Maligant	0.97	0.98	0.96	0.97
NaiveBayes	Benign	0.92	0.91	0.92	0.92
Regression	malignant	0.88	0.89	0.97	0.95

The performance analysis of the experimental study indicates that the proposed BC-DEF approach edges out other research adopted and traditional classification algorithms in terms of classification accuracy. Fig-5 shows a generated Ensemble Learning classifier with relatively small sizes, and the proposed algorithm runs faster than most multivariate decision trees and its computing time increases linearly with data size, indicating that the algorithm is scalable to large datasets.

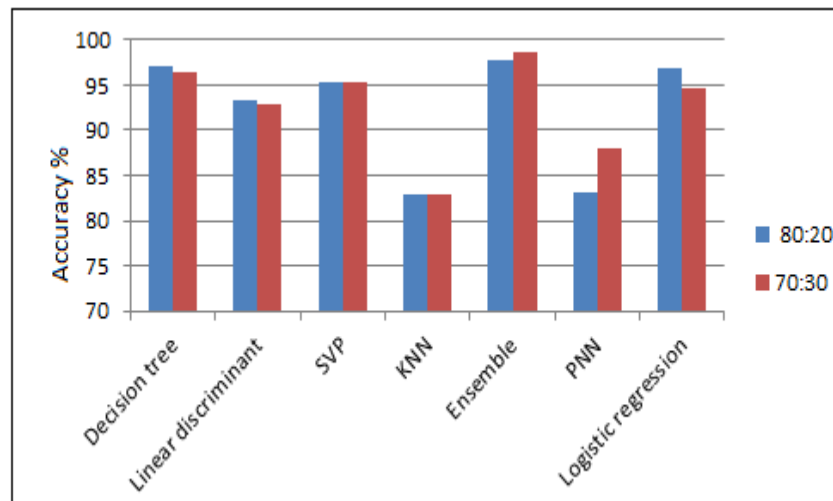


Fig-5: Accuracy observed for Learning Ensemble Approach

The performance accuracy of BC-DEF work depends on F1 score, precision, Recall and related metrics which improved performance compared to existing methods such as KNN, RF, SVM and ANN approaches. The selection of an appropriate classifier as FLD with Ensemble Learning suggests on its performance as 98.6% in comparison to SVM whose performance is 96% and Decision Tree which suggests 97% as shown in Fig-5. Analysis is carried out for both the testing and training datasets in ratio of 80:20 and 70:30 respectively.

The observed Precision and Recall values for the analysis of Breast Cancer over other approaches are compared and discussed in Fig-6. BC-DEF performance for Ensemble Learning performs better than other models.

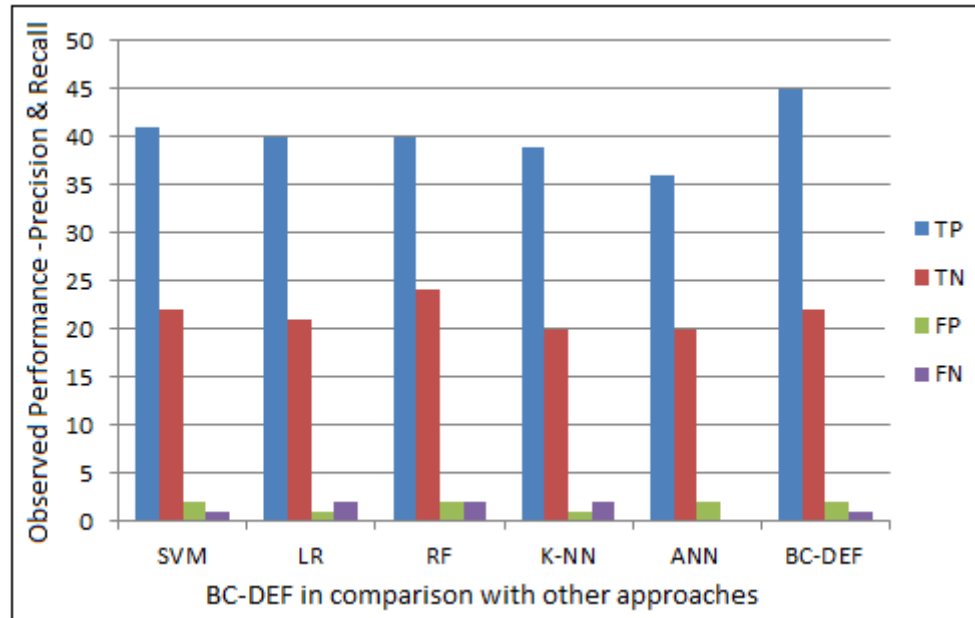


Fig-6 : Proposed BC-DEF in comparison with other approaches

To check on the observed performance in terms of Sensitivity, specificity, in relation to observed accuracy, Positive Predict value and Negative Predict value the performance is shown in Fig-7

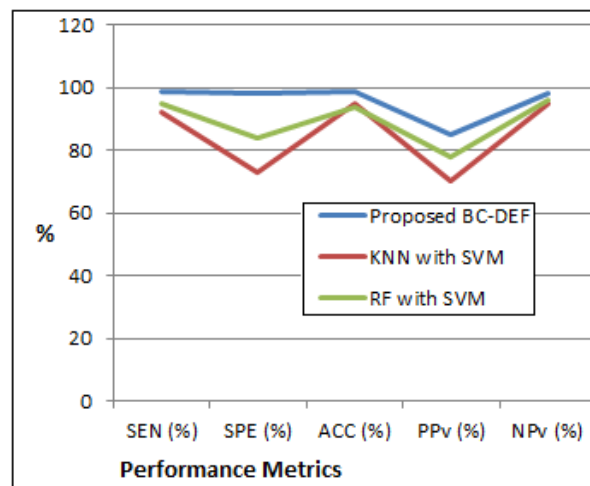


Fig-7: Analysis of Sensitivity, Specificity, Accuracy, Positive Predict Value and Negative Predict Value

Proposed BC-DEF model shows improved performance metrics for accuracy and other metrics as similar to the hybrid model of KNN with SVM and RF with SVM approach.

5. CONCLUSION

To predict the early chances of breast cancer among women of different age groups, this paper works Ensemble Learning classifier to classify and suggest predicting similar groups of breast cancer based on mammogram images and clinical pathological values being that are established between sets of data. This research suggests adopting training data and relating it to testing data for an appropriate classifier model to maximizing the chances of breast cancer prediction. Fisher Liner discrimination model is proposed to minimize the number of features required for prediction. The obtained data from standard repositories suggests a robust and novel ensemble machine-learning framework as Ensemble Learning Classifier and a Fisher Linear Discriminant based classification approach are proposed to predict breast cancer in women. The proposed solution aims to provide an explainable classification model using the strongly dependent quality of the extracted mammographic features of breast cancer affected patients. The research work has inferred from various ensemble learning models and suggests using corresponding classifier approach to work on balance or imbalance classes, such that predicting breast cancer requires minimal iterations to execute and has higher accuracy. The proposed framework outperforms other machine-learning techniques, achieving the highest accuracy of 98.17% with other models. The work in the future can be extended to support many variant types of breast cancer approaches for diagnostic and expert suggestions.

REFERENCES

- [1] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare*, vol. 8, no. 2, p. 111, 2020.
- [2] A. Aloraini, "Different machine learning algorithms for breast cancer diagnosis," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 6, p. 21, 2012.
- [3] AK Sangaiah, AK Thangavelu, XZ Gao, N Anbazhagan, MS Durai, An ANFIS approach for evaluation of team-level service climate in GSD projects using Taguchi-genetic learning algorithm, *Applied Soft Computing* 30, 628-63m 2013
- [4] 4.AK Sangaiah, AK Thangavelu, *Cognitive Computing for Big Data Systems Over IoT: Frameworks, Tools and Applications*, January, Springer Publications, 2018
- [5] H. Benbrahim, H. Hachimi, and A. Amine, "Comparative study of machine learning algorithms using the breast cancer dataset," in *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, pp. 83–91, Springer, Marrakech, Morocco, July 2019.
- [6] Domingo, M.J.; Gerardo, B.D.; Medina, R.P. Fuzzy decision tree for breast cancer prediction. *ACM Int. Conf. Proc. Ser.* 2019, 12, 316–328.

- [7] Islam, T., Akhi, A.B., Akter, F., Hasan, M.N., & Lata, M.A. (2023). Prediction of Breast Cancer using Traditional and Ensemble Technique: A Machine Learning Approach. *International Journal of Advanced Computer Science and Applications*.
- [8] B. McAree, M. E. O'Donnell, A. Spence, T. F. Lioe, D. T. McManus, and R. A. J. Spence, "Breast cancer in women under 40 years of age: a series of 57 cases from Northern Ireland," *The Breast*, vol. 19, no. 2, pp. 97–104, 2010.
- [9] D. Oyewola, D. Hakimi, K. Adeboye, and M. D. Shehu, "Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis," *International Journal of Engineering Technologies IJET*, vol. 2, no. 4, pp. 142–145, 2017
- [10] Rose, S. Intersections of machine learning and epidemiological methods for health services research. *Int. J. Epidemiol.* 2021, 49, 1763–1770.
- [11] Roy, S.; Kumar, R.; Mittal, V.; Gupta, D. Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning. *Sci. Rep.* 2020, 10, 4113–4126.
- [12] Saba, T. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *J. Infect. Public Health* 2020, 13, 1274–1289.
- [13] Saoud, H.; Ghadi, A.; Ghailani, M.; Boudhir, A.A. Application of data mining classification algorithms for breast cancer diagnosis. *ACM Int. Conf. Proc. Ser.* 2018, 20, 34–46.
- [14] Subhani, Shaik Mahabu. "Prediction of Breast Cancer Using Supervised Machine Learning Techniques." (2019).
- [15] Tabrizchi, H.; Tabrizchi, M. Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree. *SN Appl. Sci.* 2020, 2, 1–19.
- [16] M Venkatesan, Arunkumar Thangavelu, P Prabhavathy, An improved Bayesian classification data mining method for early warning landslide susceptibility model using GIS Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012) Volume 2, pg 277-288, 2013
- [17] Wu, J.; Mamidi, T.K.K.; Zhang, L.; Hicks, C. Unraveling the Genomic-Epigenomic Interaction Landscape in Triple Negative and Non-Triple Negative Breast Cancer. *Cancers*, 12, 15-20, 2020
- [18] Zhang, J.; Chen, L.; Abid, F. Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method. *J. Healthc. Eng.* 2019, 2019.
- [19] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.