International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Hybrid Model for Detection of Breast Cancer through Efficient Feature Selection using Machine Learning Approaches

[1]Pradip Chakraborty, [2]Bikash Kanti Sarkar

Submitted:08/01/2025 **Revised**: 27/02/2025 **Accepted**: 10/03/2025

Abstract: Cancer in breasts is considered as one of the dreaded diseases. It causes huge loss of human lives throughout the world and its menace is spreading fast. Earlier detection of breast cancer significantly enhances treatment effectiveness and patient's prognosis. Traditional methods in many cases of diagnosis incur much expenses, time taking and prone to errors resulting demoralization and unsuccessful. Machine learning approaches have been showing promises in automating detection of cancers in breasts. There exist a number of approaches in machine learning which show good results. This research tries to find out the techniques from the existing models and by addressing and modifying the underlying technical issues towards attaining higher accuracy. This study works using three individual learners namely, 'Support Vector Machines, 'Logistic Regression' and 'Decision Trees'. Derives a hybrid learner from these three individual leaners .Using a comprehensive dataset obtained from clinical studies, available publicly online the proposed model applies Principal Component Analysis (PCA) for selecting features. This approach processes dataset, discern subtle patterns and enhances diagnostic accuracy reducing human errors. In the comparative analysis, it presents results of the model and evaluation of performance through metrics like accuracy, sensitivity and specificity. The model finally achieves 98.24% of accuracy in prediction which seems to be impressive in comparison to other existing models. The study upholds its potential as a significant tool in medical diagnostics.

Keywords: *Breast Cancer, Ensemble Learning, Feature Selection, Machine Learning, Medical Diagnostics.*

1. Introduction

Cancer in breast is a metastatic disorder that can spread over to other organs and therefore almost incurable particularly in the advanced stages. According to Global cancer data 2020, among women, most frequently available carcinoma is cancer in breasts .It accounts for 24.5% of all types of malignancies in women .A Pie-chart in this respect has been depicted in Figure1. If the diagnosis can be established earlier, there is a good chance of great prognosis and a higher survival percentage. Applications of this approach for identifying cancer with forecasts of existence or absence of tumours could be advantageous.

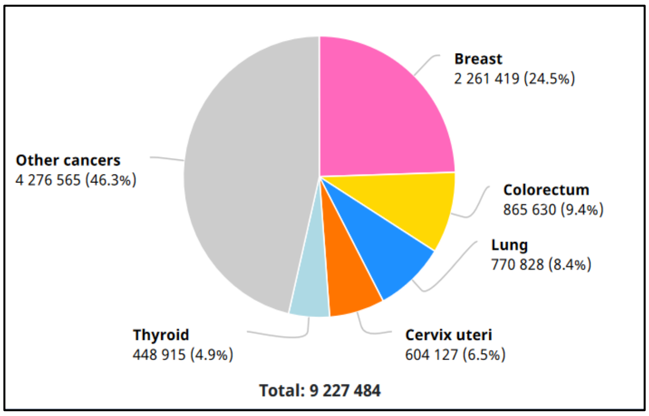

Figure 1: Cancer Data 2020 [1]

Detection of disease is fully based on available features of dataset of the relevant disease. It is found that there may exist some features which are not appropriate or with less or no significance in analytical processes. Sometimes these features may cause unnecessary complications in classifications. Therefore proper selection and utilization of features of the dataset are utmost important and key to success.

[1]*Department of Computer Science and Engineering*

Birla Institute of Technology, Mesra, Ranchi, India-835215

ORCID: 0009-0005-9321-0764

Email: phdcs10009.20@bitmesra.ac.in

[2]*Department of Computer Science and Engineering*

Birla Institute of Technology, Mesra, Ranchi, India-835215

ORCID: 0000-0002-3677-2649

Email: bksarkar@bitmesra.ac.in

**Corresponding Author*

Email: phdcs10009.20@bitmesra.ac.in

In this paper, Three well-known individual classifiers are deployed namely' Logistics Regression (LR)', 'Support Vector Machines (SVM)' and 'Decision Trees (DT)' and these learners are made to develop a hybrid classifier aiming to enhance the performance of classification tasks. Very popular technique, Principal Component Analysis (PCA) has been applied for selection of features for utilization of learning and testing mechanisms. The study runs the algorithms before and after selecting the features of the working data set. It is found that after selecting the most relevant features and then running the techniques the experiments yield better results. This can be considered as a novelty of the proposed study.

2. Literature Review

Breast cancer is a very serious concern. Huge numbers of women are victims of this dreaded disease. Every year millions of women are losing their invaluable lives because of this disease. Abnormal growth of cells in an organ can form [2]. It can be cancerous or non-cancerous. The tumour carrying germs of cancer is dangerous and can lead to death with severe sufferings if it cannot be controlled clinically in time. However, no tumour is normal by nature [3,4]. There are different types of established systems in the society. With the advent of science and technology, many sophisticated methods are in use for treatment. Along with traditional practices, machine learning approaches have been incorporated in the process of diagnosis of this disease. The approaches of machine learning show its worthiness in the process of medical diagnostics [5]. These methods contribute to forecast future outcomes of a specific cancer type and its patterns.

Vrigazova et al.(2020) used ANOVA for feature selection and SVM as a classifier to classify malignant cells in breasts. Bootstrap procedure was applied for model evaluation [6]. Ed-daoudy et al. proposed a breast tumour classification method using Association Rule and SVM. Here, Association Rule is used to reduce the irrelevant features and SVM is used for classification approach [7]. IOT and ML based model was developed to identify the malignant cell by Nanda Gopal et al. The experimental work is performed on WDBC dataset and Feature Selection is performed by PCA and classification is done by Logistics Regression (LR), Random Forest (RF) and MLP Classifiers [8]. Six different classifiers namely Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), Naive Bayes (NB), and Random Forest (RF) were applied for diagnosis of this disease on the WDBC dataset. The experimental outcomes show superiority on the basis of performances. The study achieves an accuracy of 96.5% [9]. Another research worked with Kernel Neutrosophic C-means clustering and Bayesian classifier. That worked on WDBC dataset and obtained higher accuracy in prediction of this disease [10]. A study tried with five classifiers SVM, K-NN, DT, RF and LR on WDBC dataset and succeeded with an accuracy of 97.2%.It remarks SVM as the best performer among other participating classifiers [11]. A research developed a hybrid model using WDBC dataset .That combined SVM with Genetic algorithm and resulted 97.28% of accuracy [12] . Jabbar et al. [13] developed an ensemble approach in this purpose, they used Bayesian Network with Radial Basis Function and this approach provides 97.42% accuracy on Wisconsin Breast Cancer Original dataset.

3. Materials and Methods

3.1Materials

Description of the Dataset

Dataset Size and Structure:

Instances: The WDBC dataset [14] contains a total of 569 instances, which represent individual cases of breast cancer patients.

Features: It comprises a total of 32 features whereas 30 features are usable for experiments. These are attributes or characteristics associated with biopsy samples of each patient of breast cancer. These features play vital roles in the process of classifying tumours whether it is a kind of malignant or benign. There is no missing value found in the said data set.

Table 1: Wisconsin Diagnostic Breast Cancer Data Dataset [14]

Dataset	Number of Attributes	Number of instances	Number of Benign	Number of Malignant	Number of Classes
WDBC	32	569	357	212	2

3.2 Methodology

Splitting the dataset

The entire data set is split into 70:30 ratios. That is 70% for Training purpose and 30% for Test purpose.

3.2.1 Data pre-processing

3.2.2 Working with Features of the data set

Principal Component Analysis (PCA) is taken in the study for selecting the most relevant and effective features: The approach involved using Fisher score to rank the principal components obtained from PCA. The principal components with higher Fisher score were selected as features for modelling. The aim of this approach was to pull out the most significant factors towards better performance of the model.

3.2.2.1 Principal Component Analysis (PCA) An efficient mechanism to make less of the extent of the dataset. PCA helps converting the originally available features into a new lot of orthogonal features named principal components. They hold the maximum variance in the data, effectively reducing dimensionality while retaining as much information as possible.

3.3 Classification models

3.3.1 Logistic Regression (LR)

It works for linear classification mainly for binary classification problems. By using a Logistic Function, it helps developing models to establish relationship between features and target values. It transforms linear combination of the features into a probability score. Lasso Regression (L1) for classification tasks is used here. The equation is derived from the straight line equation and is shown below:

$$LR=\frac{Y}{1-Y} \qquad (1)$$

Where Y is given by the equation below:

$$Y=B_1X_1+B_2X_2+\cdots+B_nX_n \qquad (2)$$

$$\text{Or} \quad Y=B_0+\sum_{i=1}^{n} B_i X_i \qquad (3)$$

This LR equation lies between 0and infinity to make it lie between 0 and 1 there is exponentiation of the equations then the function P(Y) becomes a probability function P(Y).

3.3.2 Decision Tree

Description: Decision Trees are non-linear classifiers that split the attribute-space recursively into regions on the basis of gains from the features to make classification decisions.

3.3.3 Support Vector Machine (SVM)

It identifies the hyper plane which may maximize the gap of separation between classes in the feature space. It addresses both the non-linear and linear classification tasks.

Strengths: SVM is effective in high-dimensional spaces, offers strong generalization capabilities, and can handle imbalanced datasets.

Use Case: SVM is used for recognition of images, classifications, biometrics interpretations etc. Different types of kernels of SVM were taken for experiments and the best performance was considered.

The hyper plane that is employed in n-dimensions is given by the following equation

$$\beta_0+\beta_1D_1+\beta_2D_2+\ldots\ldots+\beta_nD_n=0 \qquad (4)$$

Where $D_{1,}D_2$…………..D_n Denote the data points in a sample space of n dimension and β_0, β_1, β_2 up to β_n are hypothetical values.

3.3.4 Ensemble Learning

A very effective algorithm that integrates multiple learners to enhance the perfection of the final prediction. In this case, the learners ‘Decision Tree (DT)’,’ Logistic Regression (LR)’ and’ Support Vector Machines (SVM)’ participate for developing a more accurate and better performing model for identification of the disease.

The ensemble learning approach works by combining the predictions of the individual learners. The majority of the votes given by each learner are taken by the model for making final prediction. Fig. 2 represents the breast cancer prediction model graphically

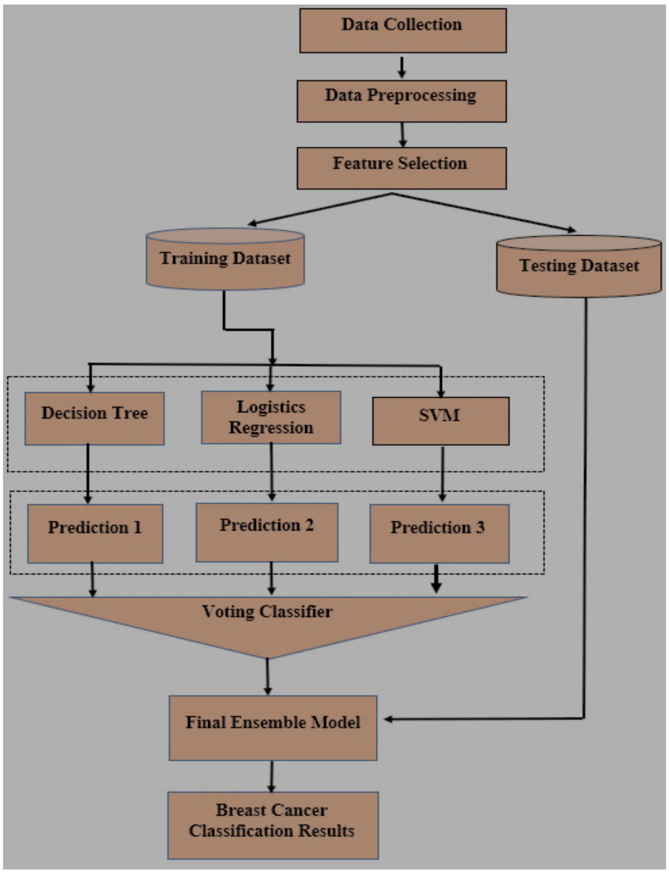

Figure2. A schematic diagram of Ensemble learning approach

3.3.5. Implementation

The experiments used Python Programming language,version 3.10.3 on Windows 10 operating system within a computer system having Intel i5 processor, 6 GB of pimary memory and 1TB of HDD .

4. Results

The ouput data obtained from the experiments are presented below in tablular forms..

Algorithms run in both the cases ,taking the data without having selection of features and after having selection of features. And the distincition in outcomes noticable significantly.

Table2. Performances of classifiers before feature selection (in percentage)

Learner	Recall	Precision	F1 Score	Accuracy
Support Vector Machines	90.12	92.23	91.16	94.54
Logistic Regression	90.05	92.07	91.04	94.12
Decision Tree	90.32	90.12	90.21	90.46
Proposed Hybrid Model	**90.69**	**92.76**	**91.71**	**95.17**

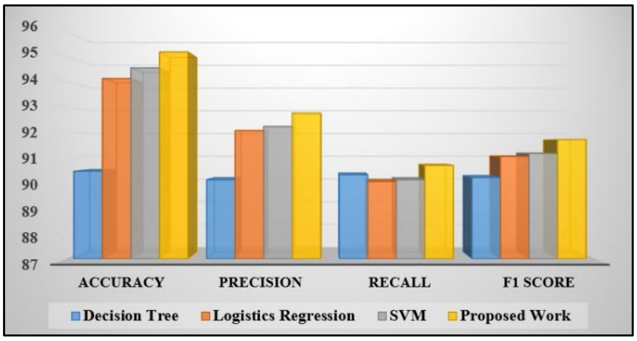

Figure3. Graphical representation of the hybrid model before feature selection

After selecting the features, the data fed into the classifiers and were re-trained, and their performances are evaluated using the same metrics. Table 3 shows the performances of the classifiers after feature selection. As shown in the table, the performances of all classifiers improved after feature selection. The proposed model got to attain the highest recall, precision, F1 score and accuracy values, indicating that the modified PCA with Fisher score feature selection methods and ensemble classification technique using Decision Tree, Logistics Regression, and SVM with voting classifier is effective in predicting breast cancer.

Table3. Performances of classifiers after feature selection (in percentage)

Learner	Recall	Precision	F1 Score	Accuracy
Support Vector Machines	94.43	94.21	94.30	95.87
Logistic Regression	93.89	93.78	93.82	95.14
Decision Trees	91.23	91.42	91.32	92.65
Proposed Hybrid Model	97.72	97.35	97.53	98.24

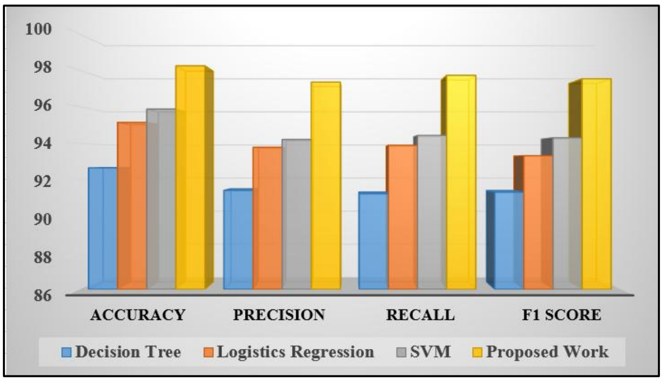

Figure4. Graphical representation of the hybrid model after feature selection

4.1 Parameters for evaluation of performances.

Performance measuring metrics for a breast cancer prediction model are employed to assess the worth of the accuracy and robustness of the predictions of models. The following are some commonly used metrics:

Accuracy: It is a quantification of correct predictions of models.

Accuracy= $\frac{(AP+AN)}{(AP+WP+AN+WN)}$ (5)

Mathematical equation is

Precision= $\frac{AP}{AP+WP}$ (6)

Recall= $\frac{AP}{AP+WN}$ (7) F1 score= 2* $\frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$ (8)

Where

Actual Positive (AP): The count of positive cases which are correctly predicted by the model.

Wrong Positive (WP): The count of negative cases which are incorrectly predicted as positive by the model

Actual Negative (AN): The count of negative cases which are correctly predicted by the model.

Wrong Negative (WN): The count of positive cases which are incorrectly predicted as negative by the model.

In summary, performance measuring metrics for a breast cancer prediction model include precision, F1 score, recall, and accuracy. These measures help in assessment of performances of the models and ensure that it provides reliable and accurate predictions.

4.2. Discussion

The study applies modified PCA with Fisher score in the process of selecting best features for the model. This method has shown to be effective in selecting relevant features as well as for reduction of dimensionality of the data set. Another contribution is implementation of a suitable ensemble approach .This technique combines the strengths of above mentioned individual classifiers and procured higher accuracy and robustness of the model.

The comparison with other standard works shows that the proposed model earns higher accuracy, precision, recall, F1 score than other studies. This indicates that this model can potentially be implemented as a reliable technique for breast cancer prediction.

5.1Conclusion

The study fulfils the objectives to a good extent. The main objective was to secure higher accuracy through efficient inclusion of relevant features and exclusion of correlated attributes which have less or no significant roles so far through its performance. There are number of researches in this field have been conducted to make the detection of breast cancer accurate. Every study more or less adds some inputs to researches .This study succeeds with 98.24 % of accuracy which is impressive in comparison with other similar studies.

The study contributes to the development of effective breast cancer prediction models and can potentially benefit healthcare practitioners and patients in early diagnosis of breast cancer.

5.2 **Future Scope**

There are still rooms for improvements. Further research may be conducted for exploration of other feature selection techniques to be executed during processing of data .Another potential direction is to investigate other ensemble classification techniques and compare their performances with the proposed model. Additionally, it would be interesting to appraise this model on larger datasets .May be tried in different clinical settings. The approach can be extended to apply on other medical fields also. The development of accurate prediction models can benefit patients and healthcare practitioners in improving the diagnosis and treatment of various diseases.

References:

[1] CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4.

[2] Farrukh Khan, Muhammad Adnan Khan, Sagheer Abbas, Atifa Athar, Shahan Yamin Siddiqui, Abdul Hannan Khan, Muhammad Anwaar Saeed, Muhammad Hussain.

https://doi.org/10.1155/2020/8017496

[3] Bisoyi P. A brief tour guide to cancer disease. Understanding cancer (Elsevier) (2022). p. 1–20, DOI: 10.1016/B978-0-323-99883-3.00006-8

[4] T. Sathyapriya, Dr. T. Ramaprabha Deep learning algorithems for breast cancer image classification. - 2020 (volume 8 - issue 03),doi : 10.17577/IJERTCONV8IS03011

[5] Ayer , T. et al., 2010. Breast cancer risk estimation with artificial neural networks revisited. *Cancer,* 2010 Jul 15;116(14):3310-21, doi: 10.1002/cncr.25081.

[6] Vrigazova, B.P.,"Detection of Malignant and Benign Breast Cancer Using the ANOVA-BOOTSTRAP- SVM", Journal of Data and Information Science, vol.5, no.2, 2020, pp.62-75.

[7] Ed-daoudy, A., Maalmi, K. Breast cancer classification with reduced feature set using association rules and support vector machine. *Netw Model Anal Health Inform Bioinforma* ,34(2020), https://doi.org/10.1007/s13721-020-00237-8

[8] V.Nanda Gopal, Fadi Al-Turjman, R. Kumar, L.Anand, M. Rajesh,Feature selection and classification in breast cancer prediction using IoT and machine learning,Measurement,Volume178,2021,109442,ISSN:0263-2241, https://doi.org/10.1016/j.measurement.2021.109442

[9]Ara S, Das A, Dey A. (2021). Malignant and benign breast cancer classification using machine learning algorithms, in: 2021 International Conference on Artificial Intelligence (ICAI). (Islamabad, Pakistan: IEEE).

[10] Kumar P, Nair GG. An efficient classification framework for breast cancer using hyper parameter tuned random decision forest classifier and Bayesian optimization. Biomed Signal Process Control (2021) 68:102682. doi: 10.1016/j.bspc.2021.102682

[11] Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Proc

Comput Sci (2021) 191:487–92. doi: 10.1016/j.procs.2021.07.062 (3):1–4.

[12] Phan AV, Nguyen ML, Bui LT. Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems. Appl Intell (2017) 46(2):455–69. doi: 10.1007/s10489-016-0843-6

[13] Jabbar, M. A. . (2021). Breast Cancer Data Classification Using Ensemble Machine Learning. *Engineering and Applied Science Research*, *48*(1), 65–72. Retrieved from https://ph01.tci-haijo.org/index.php/easr/article/view/234959

[14] Dataset URL: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

Statement on availability of data

The study works with the data which are openly available in UCI machine learning repository https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

Declaration on conflict of interest

The authors do hereby declare that there is no conflict of interest in respect of this research, funding or publication. No fund has taken from any person or organization in any way for conduction of this research work.