DOI: https://doi.org/10.52756/ijerr.2024.v46.028 *Int. J. Exp. Res. Rev.*, Vol. 46: 355-366 (2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

A Comparative Study on Detection of Breast Cancer by Applying Machine Learning Approaches

Pradip Chakraborty* and Bikash Kanti Sarkar

Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India-835215

E-mail/Orcid Id:

PC, phdcs10009.20@bitmesra.ac.in, https://orcid.org/0009-0005-9321-0764;
BKS, bksarkar@bitmesra.ac.in, https://orcid.org/0000-0002-3677-2649

Article History:

Received: 01st Jul., 2024
Accepted: 27th Dec., 2024
Published: 30th Dec., 2024

Keywords:

Breast Cancer, Ensemble, Feature Selection, Machine Learning, Medical Diagnostics

How to cite this Article:
Pradip Chakraborty and Bikash Kanti Sarkar (2024). A Comparative Study on Detection of Breast Cancer by Applying Machine Learning Approaches. *International Journal of Experimental Research and Review*, *46*, 355-366.
DOI: https://doi.org/10.52756/ijerr.2024.v46.028

Abstract: Cancer in breasts appears as a terrible malediction in society. It snitches huge human lives across the world and its peril is going to increase at a startling rate. Identification of this disease at the initial stages is indispensable. In many cases, traditional methods are prone to errors and protracted. Models applying machine learning approaches have been shown fruitful in this application area. There are large numbers of approaches in machine learning which demonstrate impressive results. This research strives to take out the short comings from the existing models and by resolving the underlying technical issues, deliver higher accuracy in end results. The research motivates and endeavours to make the patients' treatment processes more justified and cost-effective. The study works with WDBC dataset for breast cancer which is publicly accessible from the UCI research database. This study uses multiple individual learners namely, Support Vector Machines(SVM), Logistic Regression(LR), Random Forest(RF), Naive Bayes(NB), K-Nearest Neighbours(K-NN), Decision Tree(DT) and an ensemble learner called Gradient Boosting(GB) with multiple techniques of feature selection namely Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). The experimental techniques discern subtle patterns within the dataset. The proposed model evaluates the results and performances through metrics specificity, sensitivity and accuracy in a comparative structure. It succeeds with higher accuracy of 98%. The study highlights its potential as a significant tool in medical diagnostics.

Introduction

Cancer in the breast is a metastatic disorder that can spread over to other organs and therefore it is almost incurable, particularly in the advanced stages. Breast cancer caused 670000 deaths globally in 2022 (WHO, 2024). If the diagnosis can be established earlier, there is a good chance of a great prognosis and a higher survival percentage. Mammography is a widely used screening procedure for breast cancer that has been found to considerably reduce mortality. Other detection approaches have also been utilized and researched throughout the last decade. Gender, ageing, oestrogens, hereditary factors and genetic disorders are all considered as major risk factors for cancer. Cancer death is one of the most serious concerns confronting the healthcare system. This is one of the most dreaded diseases for women. Because of its physiological characteristics, breast tissues become dense with the peoples' age (Kononenko, 2001; Lai et al., 2018; Khuriwal and Mishra, 2018; Wang et al., 2020; Jabeen et al., 2022; Sharma et al., 2022; Rami et al., 2023; Vashist et al., 2024; Yadav et al., 2024). Among women, the most frequently available carcinoma is cancer in the breasts (Drukker et al., 2009; Bataineh, 2019; Ginsburg et al., 2020; Sung et al., 2021; Amethiya et al., 2022; Hassan et al., 2022). It accounts for 24.5% of all types of malignancies in women (WHO, 2020). A pie chart in this respect is depicted in Figure 1. The low survival rate is a result of the difficulty in diagnosing breast cancer and its late findings. Early detection can stop it from further spreading and lower the risk. The survival rate of this disease is increased by early identification and treatment.

*Corresponding Author: phdcs10009.20@bitmesra.ac.in

Early detection is more difficult as the symptoms are less indicative (Ahmed Medjahed et al., 2013). Computer-Aided Diagnostic (CAD) procedures are therefore absolutely necessary. CAD techniques may prove to be useful tools for radiologists. The necessity for automatic diagnosis of breast cancer emerges because manual diagnosis is challenging and time-consuming. Although present systems in healthcare show usefulness, they have several shortcomings and are sometimes erroneous. To categorize medical images in many categories and aid in early diagnosis and treatment, classification using CAD has become a useful and effective tool for medical images. In order to train computers effectively, build models for predictions and make intelligent decisions among the most useful methods, machine learning (ML) is leading and promising. It assists the professionals or experts in the process of diagnosis of malignancy in breasts in advance and analyses dimensions and sizes in order to determine the kind of cancer that exists in tumours. The most effective mechanism for favourably solving categorization and prediction problems are techniques with machine learning algorithms.

The application of these techniques for identifying cancer with forecasts of the existence or absence of tumours could be advantageous. Applying machine learning techniques, the malignancy of tumours can also be predicted.

In this study, six well-known individual classifiers are deployed, namely Logistics Regression (LR), K-Nearest Neighbours (K-NN), Support Vector Machines (SVM), Naïve-Bayes (NB), Random Forest (RF), Decision Tree (DT) and a hybrid classifier named Gradient Boosting (GB) to enhance the analysis of classification tasks. Two popular techniques, Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) have been used to select features for learning and testing purposes.

The study aims to exercise an experimental analysis to comprehensively compare these learners and feature selection techniques using the Wisconsin Breast Cancer Dataset (WDBC). By doing so, the study intends to substantiate the distinction between the performance and suitability of the said methods for breast cancer classification, contributing to advancing diagnostic tools for healthcare. We describe the dataset and the experimental setup and present our findings and discussions step by step as follows:

Literature Review

Naji et al. (2021) performed a comparison between different machine learning techniques on Wisconsin breast cancer data, which had 569 instances and 32 features. Among these 32 features, 30 features are useful for experiments. Different studies have found that SVM performs better on this dataset. Sharma et al. (2021) introduce Neural Network and Extra Trees (NN-ET) supported by empirical results and statistical analyses. Mohi Uddin et al. (2024) performed a comparative study between different ML-based classifiers. It was found that voting classifiers perform better and a web-based application is developed using these voting-based classifiers. Samieinasab et al. (2022) developed a novel ensemble approach based on Extra Tree classifier and this approach performs well in Wisconsin Breast Cancer Data. Solanki et al. (2021) employ methods using wrapper-based feature selection. Ghiasi and Zendehboudi (2021) introduced novel classification tools. Random Forest (RF) in conjunction with Extremely Randomized Trees (ET) approaches to distinguish between malignant and benign cytology from breasts. The RF and ET models exhibit superior diagnostic performance with an emphasis on factors like mitoses and uniformity in sizes of cells, highlighting their potential significance. Gopal et al.

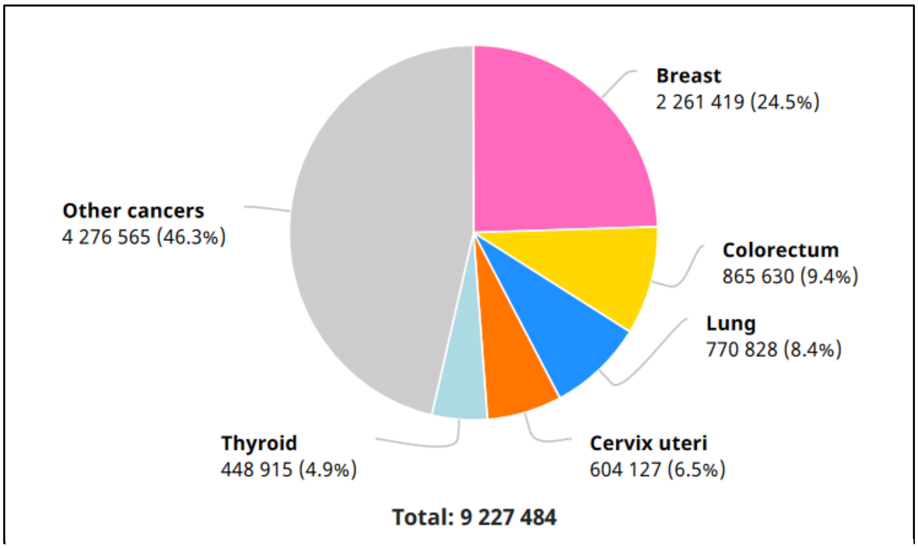

Figure 1. (Cancer Data 2020, WHO).

DOI: https://doi.org/10.52756/ijerr.2024.v46.028

(2021) incorporate IoT to apply at the early stages of this disease. It proposes a classifier achieving high performance with lower error rates. Rabiei (2022) utilizes demographical, laboratory and mammographic data. Random Forest (RF) outperforms other methods, achieving an accuracy of 80% and demonstrating the potential for early diagnosis and disease management by combining multiple risk factors. Kar and Sarkar (2022) developed a hybrid feature reduction approach based on the Co-relation coefficient and Information Gain. This feature reduction approach is suitable for reducing irrelevant features and selecting the important features and based on these important features, a disease can be identified accurately. Ibrahim et al. (2021) used variance assessment and correlation analysis followed by an ensemble approach using some popular classifiers. The proposed method surpasses trendsetting performance, attaining impressive accuracy, high precision and a recall rate. These results demonstrate its potential for improving personalized care and reducing cancer recurrence. Wu and Hicks (2021) analysed with RNA-Sequence data. The study finds that Support Vector Machines outperform other ML algorithms, demonstrating the potential of ML in efficiently distinguishing between these two breast cancer types. Wang et al. (2020) focuses on predicting breast cancer recurrence by incorporating NLP (Natural Language Processing) and machine learning techniques from King Abdullah University Hospital to EHR. It involves the creation of a medical dictionary for breast cancer and successful validation by physicians, demonstrating the potential of ML algorithms in aiding personalized medicine and clinical decision-making for breast cancer treatment.. Jabeen et al. (2022) introduced ultrasound images. The framework combines deep learning with feature selection techniques including data augmentation, transfer learning, and feature extraction. It achieves an impressive accuracy.

Research gaps

Various models for the detection of breast cancer exhibit impressive outcomes. The endeavour of the proposed study is to achieve higher accuracy by addressing the underlying issues in the selection of features and processing those with the most suitable learners.

Materials and methods

This section explains the technologies and procedures which are used in this research work. It includes a description of data, system configuration and techniques for selection of features, classification methods, performance measurement criteria and the tools used for model implementation. A step by step approach to the study has been depicted below in the form of a work flow diagram.

Materials

Description of the Dataset

The data are collected from the repository of Wisconsin Diagnostic Breast Cancer, often referred to as WDBC.

It is a popular and widely utilized dataset for research related to breast cancer. WDBC is primarily employed for classification and identifying malignant and benign tumours significantly in the case of breast cancers.

Dataset Size and Structure

Instances: The WDBC dataset 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 role in the process of classifying tumours, whether it is a kind of malignant or benign. No missing value found in the said data set.

Table 1. Wisconsin Diagnostic Breast Cancer (WDBC) Dataset [30].

Dataset	No.of Attributes	No. of instances	No.of Benigns	No.of Malignants	No. of Classes
WDBC	32	569	357	212	2

Details of the features of the dataset are furnished in table 6, appendix-A

Methods

Data processing

The most crucial part of the process of building models in machine learning is choosing correct and relevant features. Most useful features (variables or attributes) from the original set are identified. This task is undertaken for several important reasons:

1. Reduction of Dimensionality: The most important reason for selecting features is to curtail the dimension of the original dataset. When datasets have a large number of features, they become computationally expensive and can lead to the inaccuracy of model. Large-dimensional dataset may cause make models prone to over fitting, decrease model interpretability and increase the time required for training and evaluation.

2. Improved Model Performance: By selecting only the best relevant attributes, feature selection can yield better model performance. Redundant and irrelevant features can intrude noise into the model, making it harder for the algorithm to discern meaningful patterns in

DOI: https://doi.org/10.52756/ijerr.2024.v46.028

the data. Removing such features can lead to simpler and more accurate models.

3. Enhanced Generalization: Feature selection helps models generalize better to unseen data. Suppose a model can be trained with a reduced number of features. In that case, it is less likely to memorize the training data (over fitting) and more likely to capture the underlying patterns applicable to new, unseen data.

4. Reduced Training Time: Fewer features mean faster training times for machine learning models. Training on a reduced set of features is computationally more efficient, which can be particularly important when dealing with large datasets or models with complex architectures.

5. Improved Model Interpretability: Simplifying the set of features can make the task easier. It becomes more straightforward to mark which features play the most important and significant role in impacting predictions which is important in fields where interpretability is critical, such as healthcare and finance.

6. Noise Reduction: Feature selection can help filter out noisy or irrelevant information that may be present in the data. This can lead to more robust models that are less sensitive to data variations and that do not contribute to predictive accuracy.

Working with Features of the data set

Several techniques are deployed step by step for selection of features for selecting the most relevant and effective features:

Recursive Feature Elimination (RFE)

RFE is an iterative method that repeatedly fits a model and eliminates the least important features. It ranks the features based on their relevance and selects the top or higher-ranked ones.

Principal Component Analysis (PCA)

This is one of the most popular and efficient mechanisms for reducing the dimension of the dataset. PCA helps convert the originally available features into many orthogonal features named principal components. They hold the maximum variance in the data, effectively reducing dimensionality while retaining as much information as possible.

Classification

Logistic Regression (LR)

Description: Logistic Regression works for linear classification, mainly for binary classification problems. Using a Logistic Function helps develop models to establish a relationship between the target variables and the features. It transforms a linear combination of the features into a probability score.

The logistic regression equation is derived from the straight-line equation and is shown below:

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

Where Y is given by the equation below

$$Y=B_1X_1+B_2X_2+B_nX_n \quad (2)$$

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

This LR equation lies between 0 and 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).

Decision Tree

Description: Decision trees are non-linear classifiers that split the attribute space recursively into regions based on features' values to make classification decisions.

Random Forest (RF)

Description: It is a classifier of ensemble nature that clubs several decision trees. It uses bagging (bootstrap aggregating) to train each tree on a different subset of the data and combines their predictions through majority voting (classification) or averaging (regression).

Support Vector Machine (SVM)

Description: SVM is considered to be a powerful and popular classifier. It identifies the hyper plane which maximizes the margin of separation between classes in the feature space. It addresses both the non-linear and linear classification tasks.

K-Nearest Neighbours (K-NN)

Description: It is a classifier that works on the basis of instances. It classifies data points based on the majority class among their neighbours which are called K-Nearest Neighbours within the feature space. The algorithm computes the distance between two points to find the nearest neighbours to K. It is called Euclidean distance.

Naive Bayes (NB)

Description: It works on probabilistic approaches assuming “Naïve” independent of features. It calculates the probabilities of a class given in the features and selects the class with the highest probability.

Gradient Boosting (Hybrid Classifier)

Description: Gradient Boosting (GB) is a sort of ensemble techniques. It builds decision trees that maintain a proper sequence; each one corrects the mistakes committed by the preceding ones. It concatenates weak learners to a strong learner.

In summary, the diverse set of classifiers employed in the study covers a wide spectrum of techniques, each with its own strengths and use cases. This diversity allows for a comprehensive evaluation of breast cancer classification performance, considering different modelling approaches and their respective advantages.

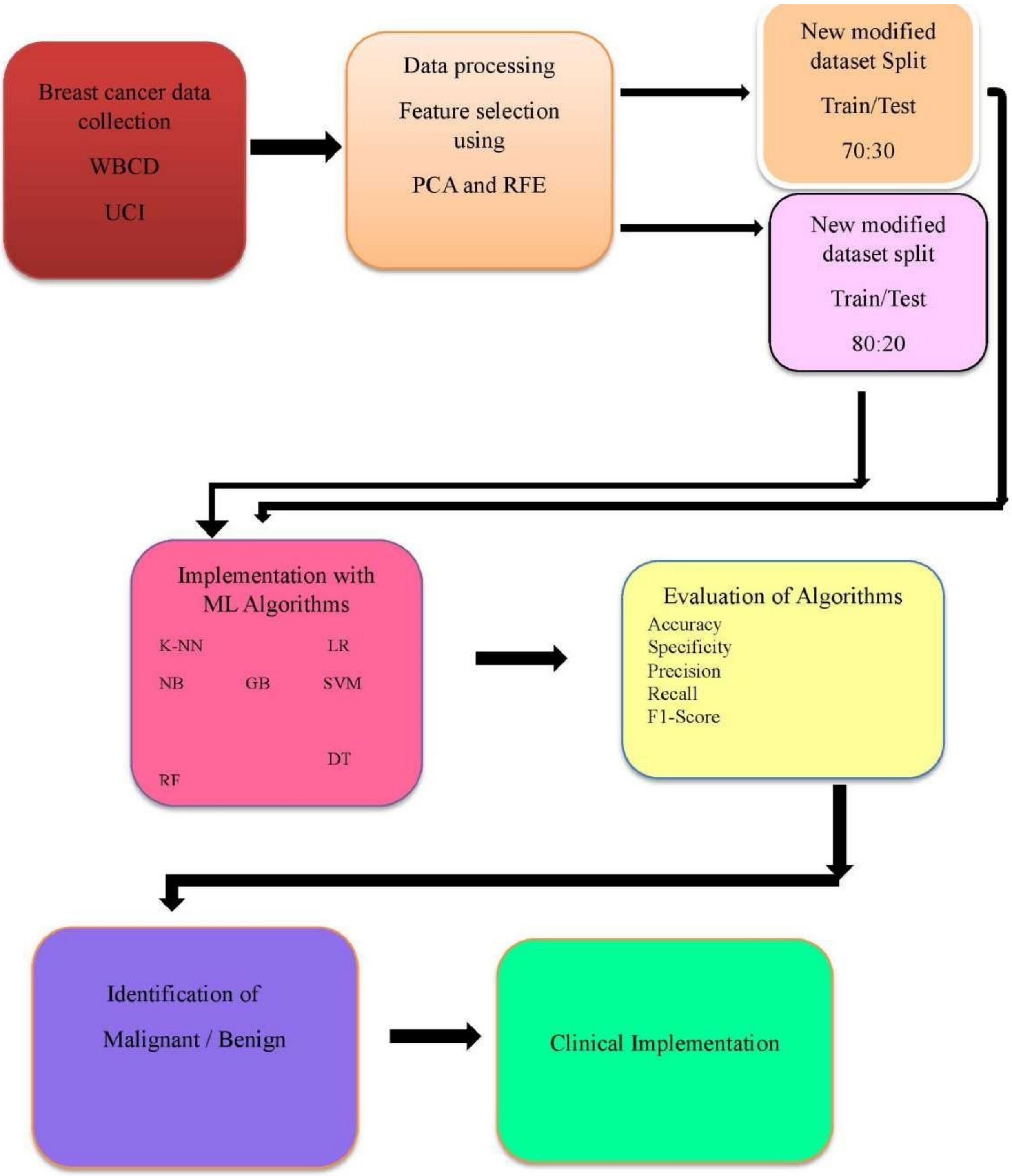

Figure 2. A schematic work-flow diagram of the study.

Experimental Setup and Implementations

The experiments used Python programming language, version 3.11.3, on a Windows 10 operating system with an Intel core i5 processor, 6 GB of RAM and 1TB of HDD, to mention a few.

Results

The outputs and outcomes of experiments were obtained using data set from an online data repository, Wisconsin Diagnostic Breast Cancer (WDBC).The performance of taken classifiers before and after feature selection using two different training and testing splits: (i) 70-30 split and (ii) 80-20 split. The programs of codes run for 10 times iteratively. Examine the effects of the selection of features and how it impacts on model. Assess performance by considering different numbers of selected features. The following classifiers are assessed: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) using different kernels K-Nearest Neighbours (K-NN), Naive Bayes (NB), and Gradient Boosting (Hybrid Classifier). Our experimental outputs and outcomes are shown below in tabular forms. Magnitudes of outputs are furnished in the following tables in percentages.

Table 2. Performance measuring metrics of different classifiers on 70-30 split (Before Feature Selection).

Classifier	Accuracy	Precision	Recall (TPR)	FPR
Logistics Regression (LR)	92%	88%	91%	9%
Decision Tree(DT)	88%	86%	87%	13%
SVM	91%	87%	91%	9%
K-NN	90%	86%	89%	11%
NB	88%	84%	88%	12%
RF	94%	91%	92%	8%
Gradient Boosting	**95%**	**93%**	**96%**	**6%**

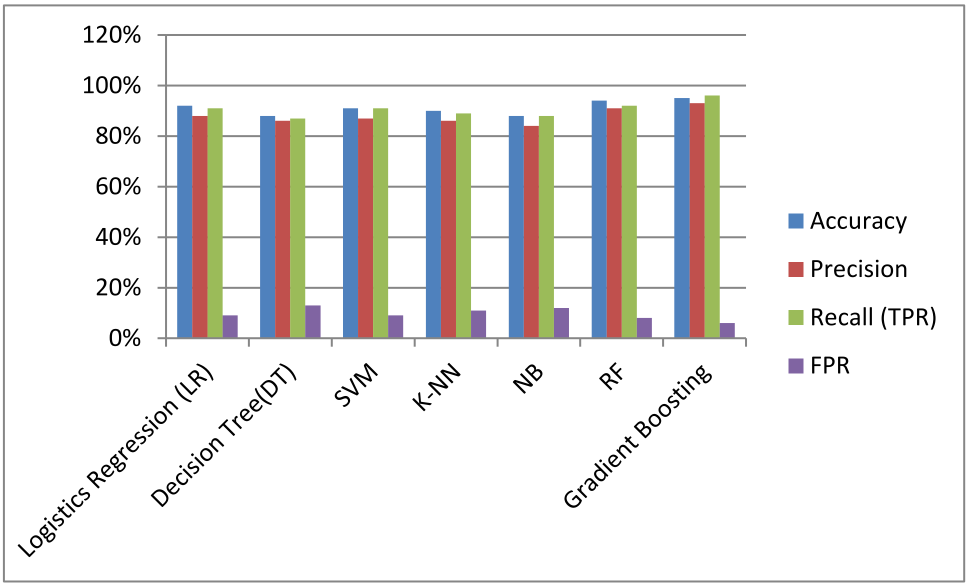

Figure 3.Graphical representation of the results (pre- feature selection) in 70:30 split as per table 2.

Table 3. Performance measuring metrics of different classifiers on 70-30 split (After Feature Selection).

Feature Selection Method	No. of Selected Features	Classifiers	Accuracy	Precision	TPR	FPR
RFE	09	LR	95%	90%	93%	7%
		DT	91%	88%	89%	11%
		RF	94%	93%	90%	10%
		SVM	95%	90%	92%	8%
		K-NN	92%	88%	91%	9%
		NB	90%	86%	91%	9%
		Gradient Boosting	**96**	**94**%	**94**%	**6**%
PCA	18	LR	96%	91%	93%	7%
		DT	91%	88%	89%	11%
		RF	94%	93%	90%	10%
		SVM	96%	90%	91%	9%
		K-NN	92%	88%	93%	7%
		NB	90%	86%	91%	9%
		Gradient Boosting	**97%**	**94%**	**94%**	**6%**

Table 4. Performance measuring metrics of classifiers, 80-20 split (Before Feature Selection).

Classifier	Accuracy	Precision	Recall (TPR)	FPR
Logistics Regression (LR)	96%	94%	93%	7%
Decision Tree	91%	86%	91%	9%
Random Forest	95%	91%	92%	8%
SVM	96%	93%	94%	6%
K-NN	92%	90%	92%	8%
Naïve Bayes	91%	89%	91%	9%
Gradient Boosting	**97%**	**94%**	**96%**	**4%**

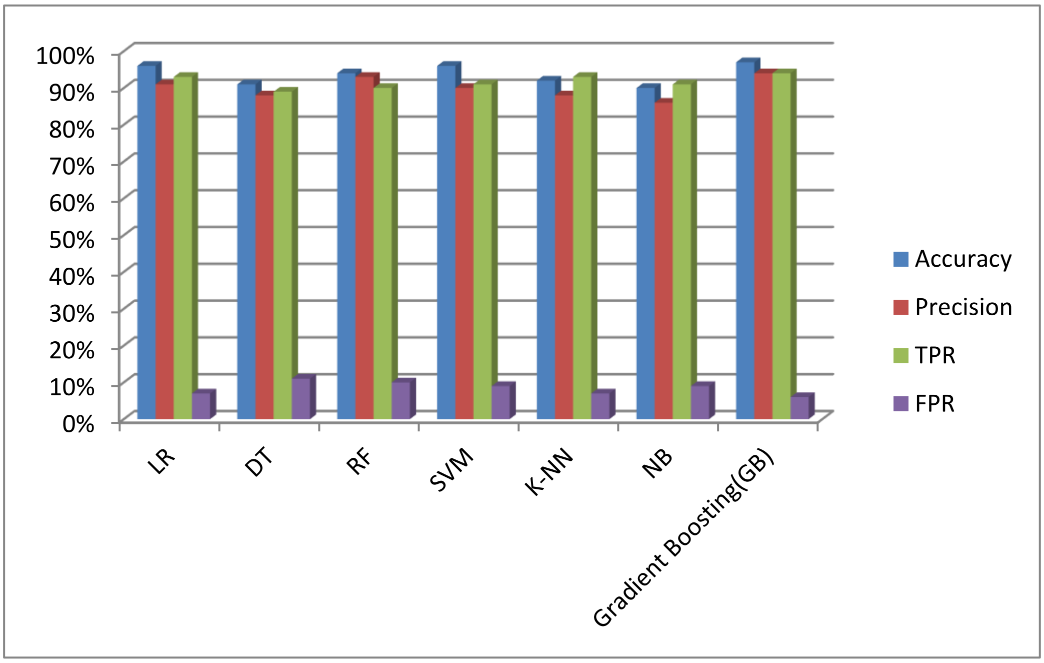

Figure 4. Graphical representation of the results after feature selection applying PCA in 70:30 split as per table 3.

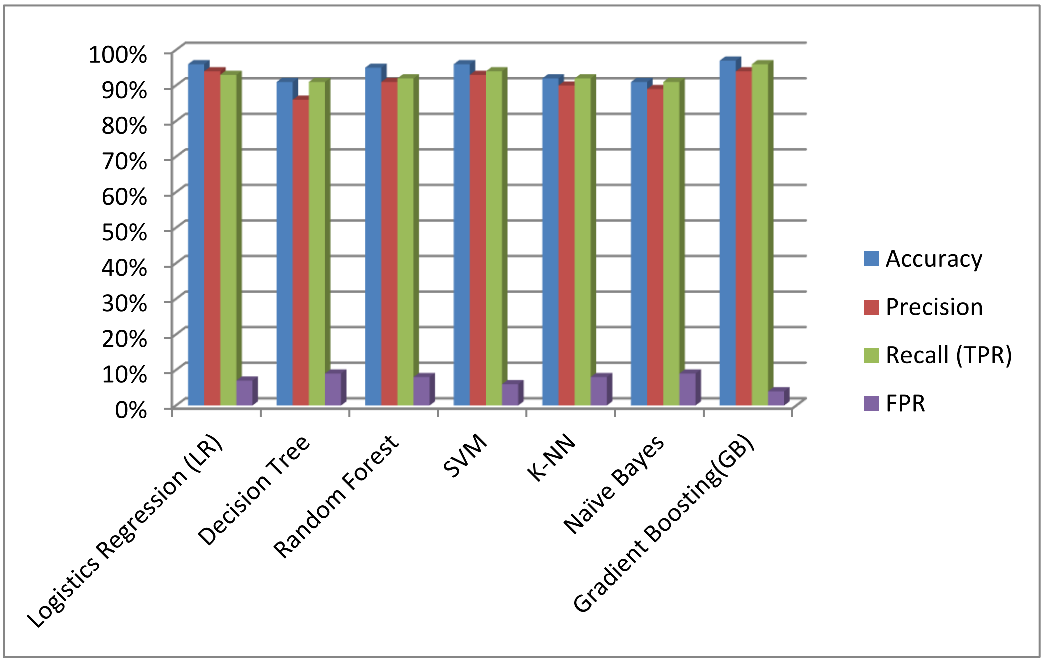

Figure 5. Graphical representation of results (pre-feature selection) in 80:20 split as per table 4.

DOI: https://doi.org/10.52756/ijerr.2024.v46.028

Table 5. Post feature selection performance measuring metrics of classifiers on 80-20 split.

Feature Selection Method	No. of Selected Features	Classifier	Accuracy	Precision	TPR	FPR
RFE	09	LR	96%	92%	93%	7%
		DT	93%	91%	89%	11%
		RF	94%	93%	91%	9%
		SVM	96%	92%	93%	7%
		K-NN	93%	91%	91%	9%
		NB	93%	91%	91%	9%
		Gradient Boosting	**97**%	**94**%	**94**%	**6**%
PCA	18	LR	96%	93%	93%	7%
		DT	92%	91%	89%	11%
		RF	95%	93%	90%	10%
		SVM	97%	94%	92%	8%
		K-NN	92%	89%	91%	9%
		NB	91%	89%	92%	8%
		Gradient Boosting	**98**%	**94**%	**94**%	**6**%

This table summarizes the feature selection results, classifier performance for different algorithms and the number of selected features after feature selection. It allows for a concise comparison of performance under various configurations.

Performance Evaluation Metrics

Certainly, let's discuss these performance metrics in detail, including their equations:

Accuracy

Definition: It quantifies the overall correctness of predictions.

Equation:

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \qquad (5)$$

Where TP= True Positive, TN=True Negative, FP=False Positive and FN=False Negative

Precision:

Definition: It indicates the proportion of true positive predictions amongst all positive predictions. It determines the efficiency of the model.

Equation:

$$\mathbf{P=TP/(TP+FP)} \qquad (6)$$

Where P= Precision, TP=True Positive and FP= False

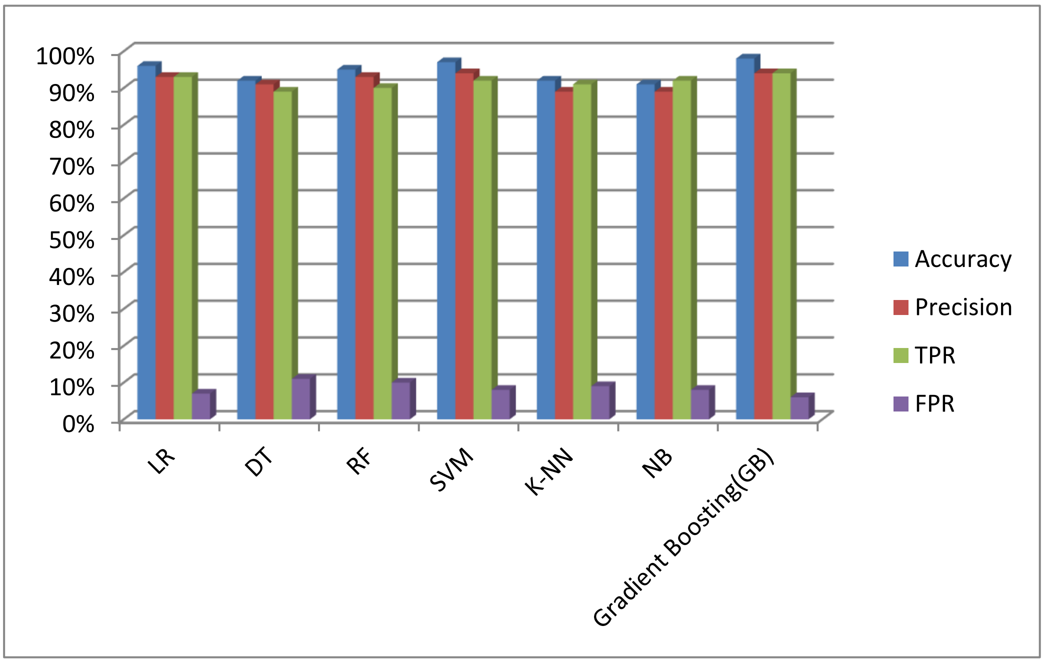

Figure 6. Graphical representation of results after feature selection applying PCA in 80:20 split as per table 5.

DOI: https://doi.org/10.52756/ijerr.2024.v46.028

Positive.

Recall

Definition: It quantifies the ratio of predictions which are literally true instances. It evaluates the capability to indicate or mark out the cases with all positives correctly.

Equation:

Recall=TP/(TP+FN) (7)

True Positive Rate (TPR) (Same as Recall)

Definition: TPR, also known as Sensitivity or Recall

Equation:

TPR= TP / (TP + FN) (8)

False Positive Rate (FPR)

Definition: It measures the ratio of false positives with all actual negative instances. It indicates how often the model incorrectly classifies negative instances as positive.

Equation:

FPR = FP / (FP + TN) (9)

These performance metrics are fundamental in assessing the quality of classification models. It supplies insights into various aspects of model and its accuracy and efficiency with the capability of correctly identifying positive cases (Recall/TPR) and negative cases (FPR).The most relevant metric(s) for evaluating our classifier can be chosen according to the requirements of the specified problem and types of errors.

These metrics collectively provide insights into the classifiers' effectiveness in identifying breast cancer cases.

Discussion

This section delves into the implications and insights derived from the results of classification experiments on the WDBC dataset. Feature selection and its impact, ratios for splitting the entire dataset and the performances of different classifiers are analysed.

Feature Selection Enhances Performance of the Model

One of the key findings of the research is the significant influence on the selection of features on performance of the model. It is observed that after applying feature selection techniques, the performance of classifiers improved or remained relatively stable across various metrics. Feature selection serves a dual purpose by enhancing model's performance and reducing dimensionality. Two techniques for the selection of features have been deployed namely; Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). Each algorithm selects a different subset of features, leading to model performance variations. Identifying and choosing the feature selection method depends on the trade-offs between the number of selected features and classification accuracy. Researchers and practitioners may need to experiment with different algorithms and selected feature counts to optimize their models for specific applications.

Impact of Training/Testing Split Ratios

The proposed study considers two common training/testing split ratios: 70-30 and 80-20. These splits are employed both before and after feature selection. The choice of split ratio can influence model's performance and generalization. It is observed that classifiers generally perform slightly better on the 70-30 split, likely due to the larger training dataset which provides more opportunities for learning the model.

However, it is essential to maintain a good balance between training and testing data to prevent over fitting. The 80-20 split shows slightly higher accuracy in the experiments. The choice of split ratio should align with the availability of data and the goals of the classification tasks.

Performances of the Classifiers

The study evaluates seven classifiers including traditional models and ensemble method for breast cancer diagnosis. It is found that Gradient Boosting consistently demonstrates high performance across various scenarios, even after feature selection. This ensemble method leverages the diversity of base classifiers to improve accuracy and robustness. However, it is essential to consider that choosing the best classifier depends upon the demand and needs of the particular application. Logistic Regression (LR) and Random Forest (RF) also exhibit strong performances and are known for their interpretability and ease of implementation. Naive Bayes (NB) demonstrates competitive results, leveraging probabilistic reasoning and independence in the assumption of attributes.

Support Vector Machines (SVM) and Decision Tree (DT) offer satisfactory performances that may be suitable choices when interpretability is not the primary concern. K-Nearest Neighbours (K-NN) shows moderate performance,highlighting its instance-based classification approach.

Contributions of the Proposed Research Work

Firstly, the study helps adopt the most appropriate methods for the selection of features. Split the entire data set into two phases, taking the ratios 70:30 (Training: Testing) and 80:20 (Training: Testing) and it appears from the results of experiments that splitting the data set into 80:20(Training: Testing) is more advantageous for achieving higher accuracy of predictions. Instead of taking individual learners, the experiment chooses a hybrid learner with appropriate algorithms. Comparing

DOI: https://doi.org/10.52756/ijerr.2024.v46.028

different approaches and locating the ensemble approach, this research may help researchers and practitioners choose the right algorithm for better results. Additionally, incorporating deep learning and neural networks in the classification of breast cancer may extend opportunities for even higher accuracy.

Table 6. Containing data from WDBC dataset (Appendix-A).

SL. NO.	FEATURES	SL.NO.	FEATURES
1	Id. Number	17	Area2
2	Diagnosis	18	Smoothness2
3	Radius1	19	Compactness2
4	Texture1	20	Concavity2
5	Perimeter1	21	Concave_Points2
6	Area1	22	Symmetry2
7	Smoothness1	23	Fractal Dimension2
8	Smoothness	24	Radius3
9	Compactness1	25	Texture3
10	Concavity1	26	Perimeter3
11	Concave_ points1	27	Area3
12	Symmetry1	28	Smoothness3
13	Fractal_Dimension1	29	Compactness3
14	Radius2	30	Concavity3
15	Texture2	31	Concave_Points3
16	Perimeter2	32	Diagnosis(M=Malignant B=Benign)

Conclusion

In conclusion, this study underscores the importance of feature selection, classifier choice and training/testing split ratios in the identification of breast cancer using the WDBC dataset. The insights gained may provide valuable guidance for concerned professionals, researchers, and data scientists working in the field of oncology. Further research, including the exploration of advanced machine learning techniques and larger datasets, holds promise for even more accurate and efficient breast cancer classification systems. Collective efforts are vital in the on-going battle against breast cancer, ultimately aiming for early detection, precise diagnosis, and improved patient outcomes.

The findings of this study hold substantial clinical implications in the domain of breast cancer diagnosis. Feature selection facilitates identifying the most informative factors, potentially reducing the need for extensive and costly data collection. Moreover, the applications of ensemble techniques, such as Gradient Boosting, can exhibit diagnostic accuracy and provide more reliable predictions.

These results facilitate the professionals supplying thoughtful insights for healthcare services. Researchers working on this disease can be benefitted a lot, particularly in classification tasks. The ability to accurately distinguish between malignant and benign tumours can lead to early detection and timely interventions, ultimately improving patient outcomes.

Future research in this area can look into the hyper parameters and explore additional methods for selection of features and tuning the performances of the model.

Statement of availability of data

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

Declaration on conflict of interest

The authors hereby declare that there is no conflict of interest with respect to this research, funding or publication.

References

Ahmed-Medjahed, S., Ait Saadi, T., & Benyettou, A. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications*, *62*(1), 1–5. https://doi.org/10.5120/10041-4635

Amethiya, Y., Pipariya, P., Patel, S., & Shah, M. (2022). Comparative analysis of breast cancer detection using machine learning and biosensors. *Intelligent Medicine*, *2*(2), 69–81. https://doi.org/10.1016/j.imed.2021.08.004

Bataineh, A. A. (2019). A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection. *International Journal of Machine Learning and Computing*, *9*(3), 248–254. https://doi.org/10.18178/ijmlc.2019.9.3.794

Drukker, K., Sennett, C. A., & Giger, M. L. (2009). Automated Method for Improving System Performance of Computer-Aided Diagnosis in

Breast Ultrasound. *IEEE Transactions on Medical Imaging, 28*(1), 122–128. https://doi.org/10.1109/tmi.2008.928178

Ghiasi, M. M., & Zendehboudi, S. (2021). Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine, 128,* 104089. https://doi.org/10.1016/j.compbiomed.2020.104089

Ginsburg, O., Yip, C., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., Gyawali, B., McCormack, V., McLaughlin de Anderson, M., Mehrotra, R., Mohar, A., Murillo, R., Pace, L. E., Paskett, E. D., Romanoff, A., Rositch, A. F., Scheel, J. R., Schneidman, M., Unger‐Saldaña, K., … Anderson, B. O. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer, 126*(S10), 2379–2393. Portico. https://doi.org/10.1002/cncr.32887

Gopal, V. N., Al-Turjman, F., Kumar, R., Anand, L., & Rajesh, M. (2021). Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement, 178*, 109442. https://doi.org/10.1016/j.measurement.2021.109442

Hassan, N. M., Hamad, S., & Mahar, K. (2022). Mammogram breast cancer CAD systems for mass detection and classification: a review. *Multimedia Tools and Applications, 81*(14), 20043–20075. https://doi.org/10.1007/s11042-022-12332-1

Ibrahim, S., Nazir, S., & Velastin, S. A. (2021). Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis. *Journal of Imaging, 7*(11), 225. https://doi.org/10.3390/jimaging7110225

Jabeen, K., Khan, M. A., Alhaisoni, M., Tariq, U., Zhang, Y.-D., Hamza, A., Mickus, A., & Damaševičius, R. (2022). Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion. *Sensors, 22*(3), 807. https://doi.org/10.3390/s22030807

Kar, B., & Sarkar, B. K. (2022). A Hybrid Feature Reduction Approach for Medical Decision Support System. *Mathematical Problems in Engineering, 2022*, 1–20. https://doi.org/10.1155/2022/3984082

Khuriwal, N., & Mishra, N. (2018). Breast Cancer Diagnosis Using Deep Learning Algorithm. *2018 International Conference on Advances in Computing, Communication Control and Networking* (ICACCCN), 98–103. https://doi.org/10.1109/icacccn.2018.8748777

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine, 23*(1), 89–109. https://doi.org/10.1016/s0933-3657(01)00077-x

Lai, Z., & Deng, H. (2018). Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer Perceptron. *Computational Intelligence and Neuroscience, 2018*, 1–13. https://doi.org/10.1155/2018/2061516

Mohi Uddin, K. M., Sikder, I. A., & Hasan, Md. N. (2024). A Comparative Study on Machine Learning Classifiers for Cervical Cancer Prediction: A Predictive Analytic Approach. *EAI Endorsed Transactions on Internet of Things, 11*. https://doi.org/10.4108/eetiot.6223

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms for Breast Cancer Prediction And Diagnosis. *Procedia Computer Science, 191*, 487–492. https://doi.org/10.1016/j.procs.2021.07.062

Rabiei, R. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *Journal of Biomedical Physics and Engineering, 12*(3). https://doi.org/10.31661/jbpe.v0i0.2109-1403

Rami, N., Kulkarni, B., Chibber, S., Jhala, D., Parmar, N., & Trivedi, K. (2023). In vitro antioxidant and anticancer potential of *Annona squamosa* L. Extracts against breast cancer. *Int. J. Exp. Res. Rev., 30*, 264-275. https://doi.org/10.52756/ijerr.2023.v30.024

Samieinasab, M., Torabzadeh, S. A., Behnam, A., Aghsami, A., & Jolai, F. (2022). Meta-Health Stack: A new approach for breast cancer prediction. *Healthcare Analytics, 2*, 100010. https://doi.org/10.1016/j.health.2021.100010

Sharma, D., Kumar, R., & Jain, A. (2021). A Systematic Review of Risk Factors and Risk Assessment Models for Breast Cancer. In: Marriwala, N., Tripathi, C.C., Kumar, D., Jain, S. (eds) Mobile Radio Communications and 5G Networks. Lecture Notes in Networks and Systems, vol 140. Springer, Singapore. https://doi.org/10.1007/978-981-15-7130-5_41

Sharma, D., Kumar, R., & Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. *Measurement: Sensors, 24*, 100560. https://doi.org/10.1016/j.measen.2022.100560

Solanki, Y. S., Chakrabarti, P., Jasinski, M., Leonowicz, Z., Bolshev, V., Vinogradov, A., Jasinska, E., Gono, R., & Nami, M. (2021). A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. *Electronics*, *10*(6), 699. https://doi.org/10.3390/electronics10060699

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.*, *71*(3), 209-249. https://doi.org/10.3322/caac.21660

Vashist, A., Sagar, A., & Goyal, A. (2024). Correlation of Prognostic Factors of Invasive Lobular Carcinoma and Invasive Ductal Carcinoma. *International Journal of Experimental Research and Review*, *42*, 50-59. https://doi.org/10.52756/ijerr.2024.v42.005

Wang, H., Li, Y., Khan, S.A., & Luo, Y. (2020). Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artif Intell Med.*, *110*, 101977. https://doi.org/10.1016/j.artmed.2020.101977.

WHO (2020). Cancer. https://www.who.int/health-topics/cancer#tab=tab_1

WHO (2024). Breast Cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Wu. J., & Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of Personalized Medicine*, *11*(2), 61. https://doi.org/10.3390/jpm11020061

Yadav, P., Bhargava, C., Gupta, D., Kumari, J., Acharya, A., & Dubey, M. (2024). Breast Cancer Disease Prediction Using Random Forest Regression and Gradient Boosting Regression. *International Journal of Experimental Research and Review*, *38*, 132-146. https://doi.org/10.52756/ijerr.2024.v38.012

How to cite this Article:

Pradip Chakraborty and Bikash Kanti Sarkar (2024). A Comparative Study on Detection of Breast Cancer by Applying Machine Learning Approaches. *International Journal of Experimental Research and Review*, *46*, 355-366.

DOI : https://doi.org/10.52756/ijerr.2024.v46.028