Data Mining and Principal Component Analysis on Coimbra Breast Cancer Dataset

Anupam Sen*

Department of Computer Science, Government General Degree College, Singur, University of Burdwan, India.

*Corresponding author

doi: https://doi.org/10.21467/proceedings.115.5

ABSTRACT

Machine Learning (ML) techniques play an important role in the medical field. Early diagnosis is required to improve the treatment of carcinoma. During this analysis Breast Cancer Coimbra dataset (BCCD) with ten predictors are analyzed to classify carcinoma. In this paper method for feature selection and Machine learning algorithms are applied to the dataset from the UCI repository. WEKA ("Waikato Environment for Knowledge Analysis") tool is used for machine learning techniques. In this paper Principal Component Analysis (PCA) is used for feature extraction. Different Machine Learning classification algorithms are applied through WEKA such as Glmnet, Gbm, ada Boosting, Adabag Boosting, C50, Cforest, DcSVM, fnn, Ksvm, Node Harvest compares the accuracy and also compare values such as Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE). Here the 10-fold cross validation method is used for training, testing and validation purposes.

Keywords: Glmnet, Node Harvest, Ksvm, MAE.

1 Introduction

According to the World Health Organization breast cancer is the most predominant cancer among women. The maximum number of cancer-related deaths among women were reported due to breast cancer causing 2.1 million deaths each year [1]. To detect early-stage breast cancer X-ray, mammography is used at present. In an asymptomatic population this method is very useful to detect breast cancer in a systematic way. Mammography images are used to differentiate smaller masses and microcalcifications to spot breast cancer in its starting phase [2]. At present, mammography is a widely used standard screening process for breast cancer. In breast cancer prediction, misclassification of mammograms remains one area that needs improvement. Still a challenge to develop a cheap and easily accessible method from those predictors. Parameters collected from blood samples may offer other ways to better diagnose breast cancer in females [3]. Good outcomes in treatment can be achieved by early diagnosis of breast cancer. More screening tools are required for healthy predictive models based on data which may be collected in blood analysis and routine consultation. Through routine blood analysis like Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1, Age and Body Mass Index (BMI) can be collected. In this work, try to assess how models based on data may be used to forecast the presence of breast cancer.

2 Literature Review

To classify breast cancer a large number of researches have already been conducted on the application of data mining and ML on different medical datasets. Remarkable accurateness is achieved in numerous studies in classification of breast cancer. Wisconsin breast cancer diagnosis (WBCD) dataset has been widely used. In

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

article [4], how blood related parameters are related to obesity-associated breast cancer. In article [5], various machine learning algorithms applied on breast cancer diagnosis and prognosis were discussed. Another study shows Metabolic Syndrome, specifically insulin resistance and abdominal fat women after menopause have a large possibility of breast cancer. "Subclinical insulin resistance, Homeostasis Model Assessment Insulin Resistance (HOMA-IR) can be used to identify patients. For high-risk patients this is important for prevention and testing" [6]. In the paper [7] Random Forest and Naive Bayes were used as feature selection method and rank the feature importance. In article [8], classifier model Deep Neural network (DNN) and Recursive Feature Elimination (RFE) for feature selection were used to obtain 98.62% accuracy. Another study shows that the optimal activation function is used to reduce the classification error by using fewer blocks. In another research study, the combination of age, body mass index (BMI), and metabolic parameters was determined as a potential reasonable and effective predictor for breast cancer [9]. In another study, a three-stage hybrid technique was used on the Coimbra dataset to detect the presence of breast cancer [10]. In another article, Fuzzy support vector machine (SVM) and principal component analysis (PCA) method was used for the diagnosis of breast cancer tumour [11].

3 Materials and Methods

Many Different techniques were used for the detection of breast cancer when related works were analyzed. There are several datasets available for the detection of breast cancer. In this article, Breast Cancer Coimbra dataset (BCCD) with 116 observations with 10 attributes and one of which is a class variable, i.e. (1 = Healthy, 2 = Patient) is taken from the UCI ML Repository [12]. Table 1 contains attribute information of the original dataset. Table. 2 contains seven features selected by Principal Component Analysis (PCA). In the proposed methodology, different ML classification algorithms are applied through WEKA such as Glmnet, Gbm, ada Boosting, Adabag Boosting, C50, Cforest, DcSVM, fnn, Ksvm, Node Harvest to build models. Here the 10-fold cross validation method is used for training, testing and validation purposes. The overall research methodology is depicted in fig. 1.

Table 1: Dataset Attribute information

Attribute	**Data type**
Age	Numeric
BMI	Numeric
Glucose	Numeric
Insulin	Numeric
HOMA	Numeric
Leptin	Numeric
Adiponectin	Numeric
Resistin	Numeric
MCP.1	Numeric
Classification {1,2}	Nominal

Table 2: Attribute information Selected by PCA

Attribute	**Data type**
Age	Numeric
BMI	Numeric
Glucose	Numeric
HOMA	Numeric
Leptin	Numeric
Resistin	Numeric
MCP.1	Numeric

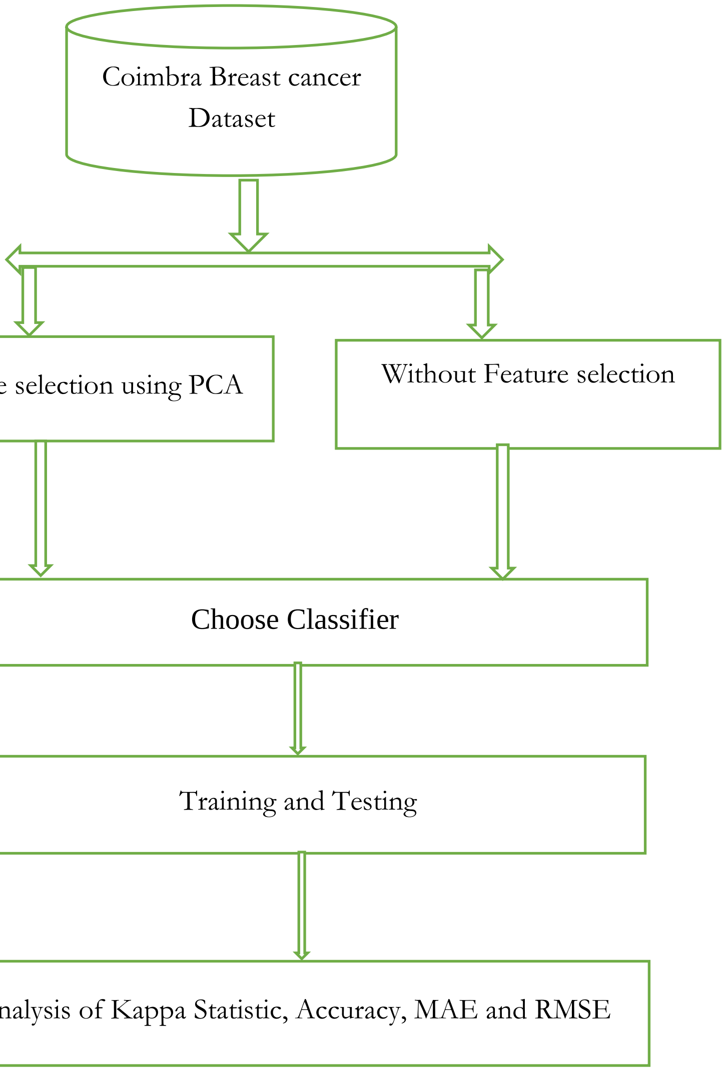

Coimbra Breast cancer Dataset
Feature selection using PCA
Without Feature selection
Choose Classifier
Training and Testing
Analysis of Kappa Statistic, Accuracy, MAE and RMSE

Fig 1: Research Methodology

4 Experimental Results

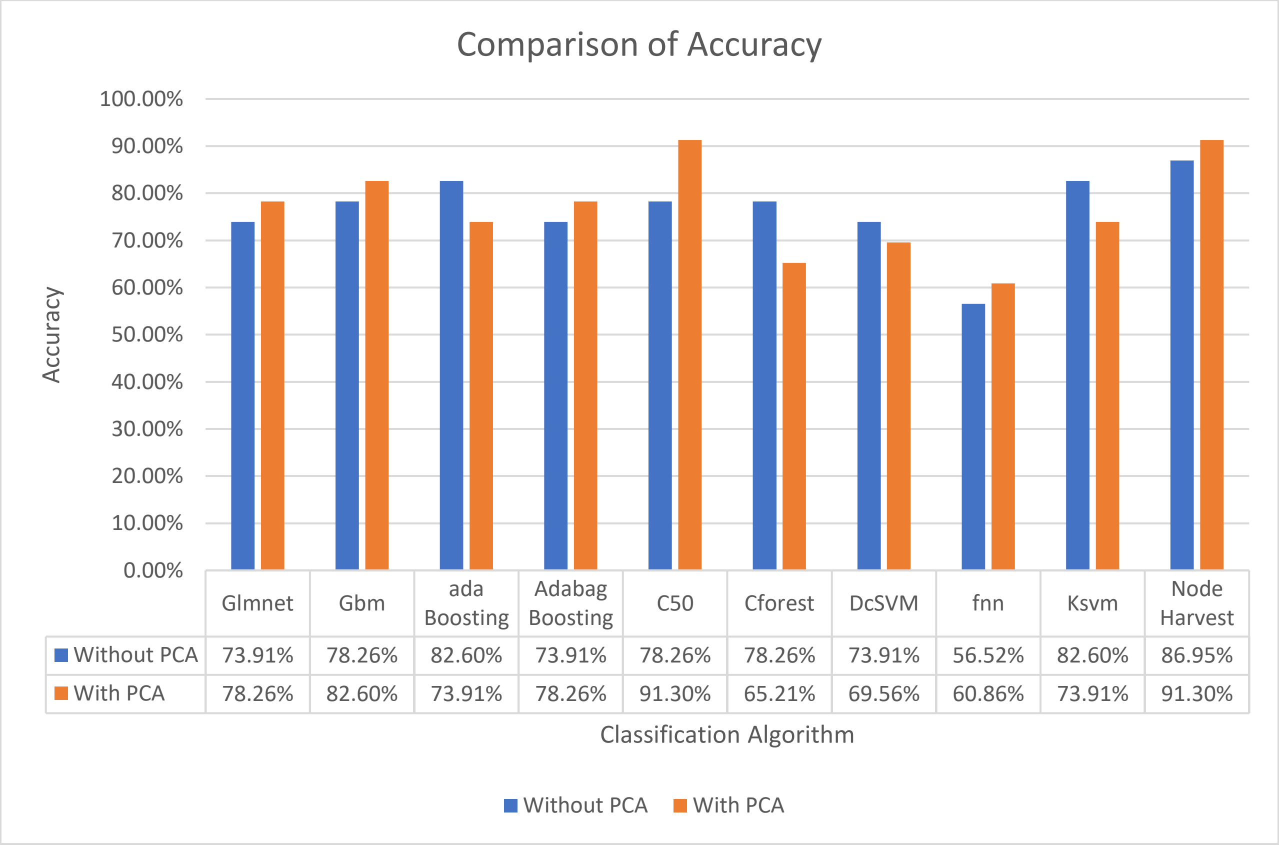
In this research work dataset is taken from UCI repository [12]. Ten ML classifiers have been used to build models without selecting features. The value in Table 3 compares accuracy, kappa statistic, Mean Absolute Error, Root Mean Square Error of different classifier algorithms without selecting features. Here the accuracy is 86.95% and Kappa Statistic is 0.7039 for Node Harvest classifier outperforms all other classifiers. Mean Absolute Error (MAE) is minimized for C50 classifier and Root Mean Square is minimized for Node Harvest Classifier. Table 4 represents information about the accuracy and Kappa Statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) of the different classifier algorithms based on Principal Component Analysis (PCA). Here the accuracy is 91.30 % for C50 and Node Harvest classification algorithm performs better and Kappa statistic is also high. Mean Absolute Error and Root Mean Square Error is minimized for C50 classification algorithm. Comparison of accuracy, Kappa statistic, Mean Absolute Error, Root Mean Square Error without feature selection and with feature selection through PCA of different classification algorithms are shown in fig. 2, fig. 3, fig. 4, fig.5 respectively. Fig 2 represents that the accuracy of the Glmnet, Gbm, Adabag Boosting, C50, fnn and Node Harvest classification algorithms lead to better results in the proposed work. Fig 3 represents that the Kappa Statistics of the Glmnet, Gbm, Adabag Boosting, C50, fnn and Node Harvest classification algorithms perform better results in the proposed work. Fig 4 demonstrates that the Mean Absolute Error (MAE) is minimized for the Gbm, ada Boosting, Adabag Boosting, C50, Cforest and Node Harvest classification algorithms. Fig 5 shows that the Root Mean Square Error (RMSE) is minimized for Glmnet, Gbm, ada Boosting, Adabag Boosting, C50, Cforest, fnn and Node Harvest classification algorithms in the suggested work.

Table 3: Classifier performance without PCA

Classification Algorithm	Accuracy	Kappa Statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Glmnet	73.91%	0.4692	0.3765	0.4452
Gbm	78.26%	0.5064	0.3368	0.4074
ada Boosting	82.60 %	0.6198	0.3318	0.3896
Adabag Boosting	73.91%	0.4692	0.4195	0.4358
C50	78.26 %	0.5418	0.2515	0.4324
Cforest	78.26 %	0.5418	0.3814	0.4197
DcSVM	73.91%	0.4298	0.2609	0.5108
fnn	56.52 %	0.4348	0.3913	0.6594
Ksvm	82.60%	0.5893	0.3506	0.3997
Node Harvest	86.95%	0.7039	0.3507	0.3754

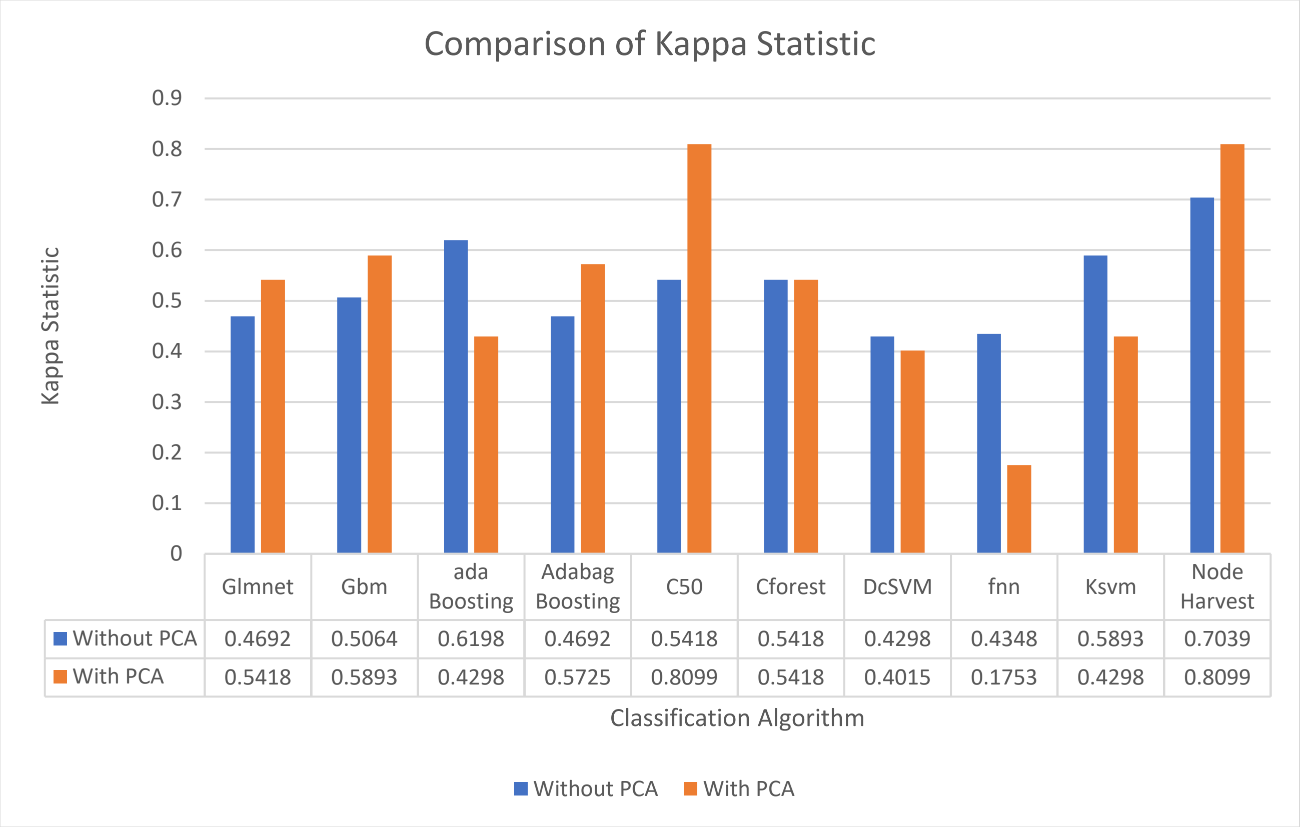
Table 4: Classifier performance with PCA

Classification Algorithm	Accuracy	Kappa Statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Glmnet	78.26%	0.5418	0.3801	0.4447
Gbm	82.60%	0.5893	0.3263	0.4
ada Boosting	73.91 %	0.4298	0.3168	0.3644
Adabag Boosting	78.26%	0.5725	0.3949	0.4114
C50	91.30 %	0.8099	0.1896	0.293
Cforest	65.21 %	0.5418	0.3762	0.4165
DcSVM	69.56%	0.4015	0.3043	0.5517
fnn	60.86 %	0.1753	0.3913	0.6255
Ksvm	73.91 %	0.4298	0.3608	0.4056
Node Harvest	91.30%	0.8099	0.3374	0.3583



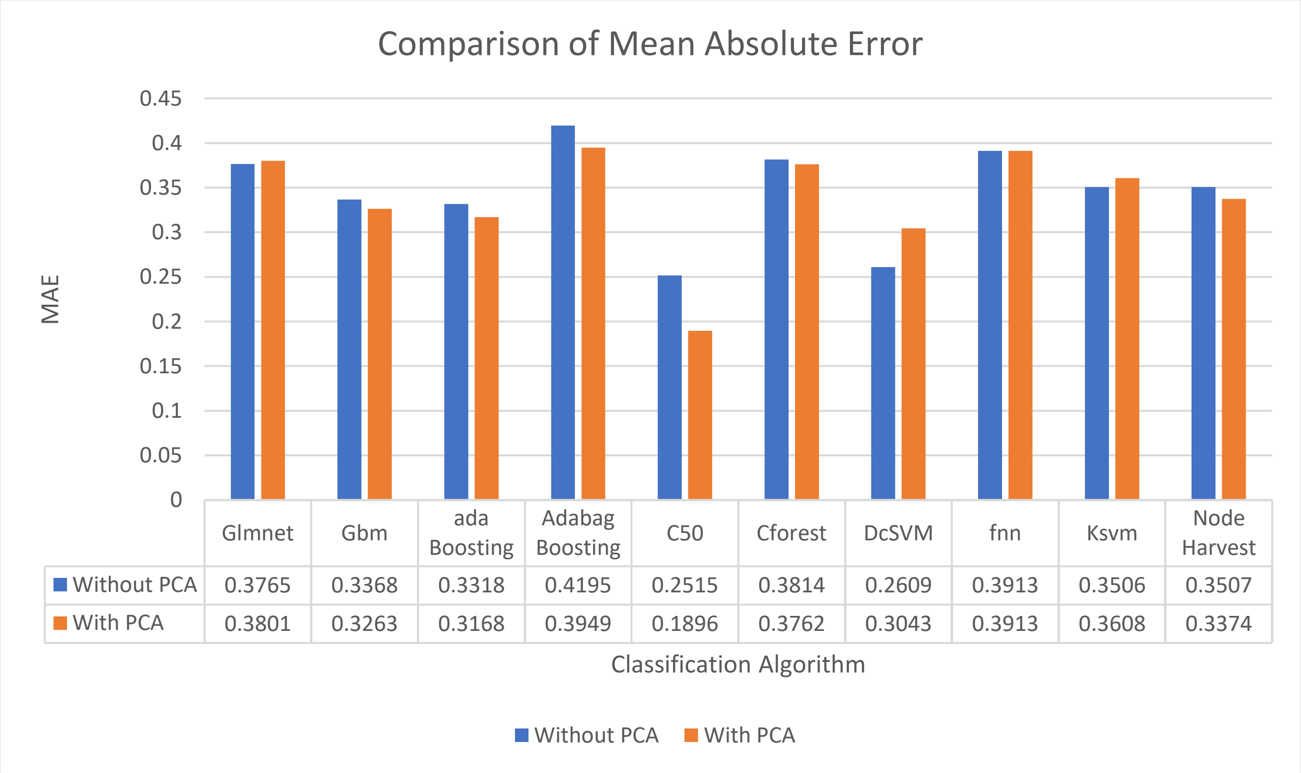
	Glmnet	Gbm	ada Boosting	Adabag Boosting	C50	Cforest	DcSVM	fnn	Ksvm	Node Harvest
🟦 Without PCA	73.91%	78.26%	82.60%	73.91%	78.26%	78.26%	73.91%	56.52%	82.60%	86.95%
🟧 With PCA	78.26%	82.60%	73.91%	78.26%	91.30%	65.21%	69.56%	60.86%	73.91%	91.30%

Fig 2: Comparison of Accuracy without PCA and with PCA



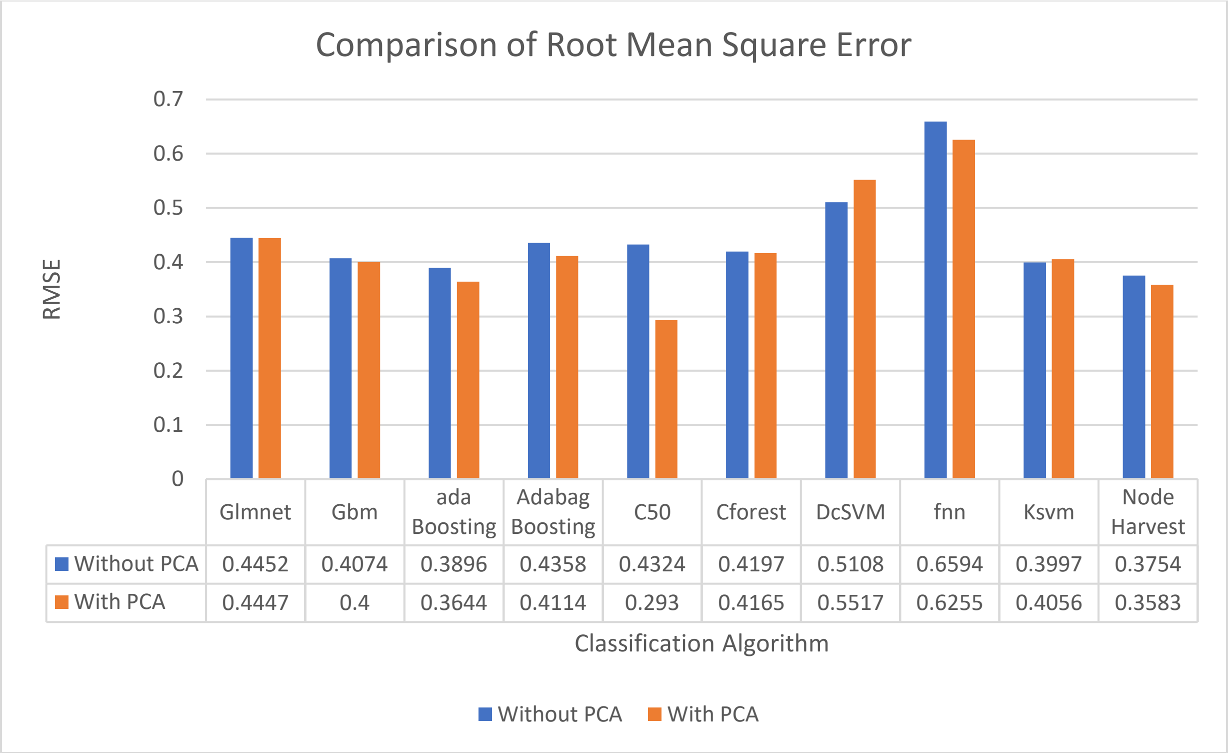
	Glmnet	Gbm	ada Boosting	Adabag Boosting	C50	Cforest	DcSVM	fnn	Ksvm	Node Harvest
🟦 Without PCA	0.4692	0.5064	0.6198	0.4692	0.5418	0.5418	0.4298	0.4348	0.5893	0.7039
🟧 With PCA	0.5418	0.5893	0.4298	0.5725	0.8099	0.5418	0.4015	0.1753	0.4298	0.8099

Fig 3: Comparison of Kappa Statistic without PCA and with PCA



	Glmnet	Gbm	ada Boosting	Adabag Boosting	C50	Cforest	DcSVM	fnn	Ksvm	Node Harvest
🟦 Without PCA	0.3765	0.3368	0.3318	0.4195	0.2515	0.3814	0.2609	0.3913	0.3506	0.3507
🟧 With PCA	0.3801	0.3263	0.3168	0.3949	0.1896	0.3762	0.3043	0.3913	0.3608	0.3374

Fig 4: Comparison of Mean Absolute Error without PCA and with PCA



	Glmnet	Gbm	ada Boosting	Adabag Boosting	C50	Cforest	DcSVM	fnn	Ksvm	Node Harvest
🟦 Without PCA	0.4452	0.4074	0.3896	0.4358	0.4324	0.4197	0.5108	0.6594	0.3997	0.3754
🟧 With PCA	0.4447	0.4	0.3644	0.4114	0.293	0.4165	0.5517	0.6255	0.4056	0.3583

Fig 5: Comparison of Root Mean Square Error without PCA and with PCA

5 Conclusion and Future Work

In this proposed work, try to improve the accuracy, kappa statistic of the different classifiers to more accurately identify the early diagnosis of breast cancer. In this proposed model, better accuracy and kappa statistics are obtained for Glmnet, Gbm, Adabag Boosting, C50, fnn and Node Harvest classification algorithms. To obtain more accurate results, a large dataset is needed. It is concluded that feature extraction and machine learning algorithms play an essential role in identifying the early diagnosis of breast cancer to reduce cost and time. Different feature selection methods and newer algorithms can be applied to get better results.

References

[1] "Breast cancer", World Health Organization, 2018. [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. [Accessed: 24- Sep- 2018].

[2] P. Mora et al., “Improvement of early detection of breast cancer through collaborative multi-country efforts: Medical physics component,” Phys. Medica, vol. 48, no. December 2017, pp. 127– 134, 2018.

[3] S. Y. Loke and A. S. G. Lee, “The future of blood-based biomarkers for the early detection of breast cancer,” Eur. J. Cancer, vol. 92, pp. 54–68, 2018.

[4] Crisóstomo J, Matafome P, Santos-Silva D, Gomes AL, Gomes M, Patrício M, Letra L, Sarmento-Ribeiro AB, Santos L, Seiça R. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. Endocrine. 2016 Aug;53(2):433-42. doi: 10.1007/s12020-016-0893-x. Epub 2016 Feb 18. PMID: 26892376.

[5] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis,” Designs, vol. 2, no. 2, p. 13, 2018.

[6] I. Capasso et al., “Homeostasis model assessment to detect insulin resistance and identify patients at high risk of breast cancer development: National Cancer Institute of Naples experience,” J. Exp. Clin. Cancer Res., vol. 32, no. 1, p. 1, 2013.

[7] P. Suryachandra, and P. V. S. Reddy, “Comparison of machine learning algorithms for breast cancer.” IEEE International Conference on Inventive Computation Technologies (ICICT), pp. 1- 6, 2016.

[8] Karthik S., Srinivasa Perumal R., Chandra Mouli P.V.S.S.R. (2018) Breast Cancer Classification Using Deep Neural Networks. In: Margret Anouncia S., Wiil U. (eds) Knowledge Computing and Its Applications. Springer, Singapore. https://doi.org/10.1007/978-981-10-6680-1_12

[9] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," BMC cancer, vol. 18, no. 1, pp. 29, 2018.

[10] K. Polat and U. Sentürk, "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier," 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2018, pp. 1-4, doi: 10.1109/ISMSIT.2018.8567245.

[11] Z. Luo, X. Wu, S. Guo and B. Ye, "Diagnosis of Breast Cancer Tumor Based on PCA and Fuzzy Support Vector Machine Classifier," 2008 Fourth International Conference on Natural Computation, Jinan, China, 2008, pp. 363-367, doi: 10.1109/ICNC.2008.932.

[12] UCI "Machine Learning Repository" https://archive.ics.uci.edu/ml/index.php.