Section A-Research Paper
ISSN 2063-5346

COMPARISON BASED ON EXAMINATION OF MACHINE LEARNING TECHNIQUES FOR MINIMIZING THE AMOUNT OF FEATURES UTILIZED IN BREAST CANCER FORECAST

[1] **Dr. Prashant Sharma**

Associate Professor, Department of Computer Science and Engineering

Pacific (PAHER) University, Udaipur, India

Email- prashant.sharma@pacific-it.ac.in

[2] Dr. Vaibhav Narawade
Professor, Department of Computer Engineering,
Ramrao Adik Institute of Technology,
D Y Patil Deemed to be University,
Nerul, India
Email- vaibhav.narawade@rait.ac.in

[3] Avantika Mahadik

Research Scholar, Pacific (PAHER) University, Udaipur, India

Email- avantika_mahadik@rediffmail.com

Abstract:

We explored several algorithms used in machine learning for predicting breast cancer in this research. Our primary objective is to find the optimum algorithm from the numerous machine learning algorithms which can be utilized for the prediction of breast cancer with fewer features. This work's principal objective is to see how feature selection contributes to the highest prediction accuracy.

In our research study, we examined six popular machine learning algorithms, namely, SVM[1] , KNN[2], RF[3], DT[4], MLP (Multi Layer Perceptron) and Linear Regression. Various measures are utilized to assess the model's quality or performance; these metrics are commonly referred to as evaluation or performance metrics. This allows us to fine-tune the hyper-parameters and enhance the model's performance. Through the performance matrix, we can understand the efficacy of the model in the performance of the given data. Reduction in features and focusing only on key features in the prediction of breast cancer is very extensively essential. In our study, we observed various methods which are used for eliminating the number of

[1] SVM – Support Vector Machine

[2] KNN- K nearest neighbor

[3] RF- Random Forest

[4] DT- Decision Tree

features. Confusion matrix, BCAM (Binary Classification Accuracy Method), F1 score, Precision, Mean Squared Error, Accuracy and Mean Absolute Error are some important methods. Our principal focus in this research work is doing comparative analysis of approaches which are used in limiting the features.

Keywords: Breast Cancer, features selection, Confusion matrix, machine learning algorithm, Performance matrix

Introduction:

Molecules that develop in the breast tissues are known as breast cancers. It is the greatest prevalent cancer type among women and also one of the major reasons for the increasing female death rate [1].

Building an accurate machine learning model involves several key processes, one of which is evaluating the model's performance. Several criteria are taken to assess the model's quality or performance; these metrics are also referred to as evaluation or performance metrics. We are able to better comprehend our model's performance for the information that was given using the aid of these performance metrics. This allows us to fine-tune the hyper-parameters to further improve the model's performance. Every machine learning model attempts to achieve good generalization on new or unseen data and performance measures assess the model's ability to do so. Performance metrics for classification include the confusion matrix, recall, accuracy, F-score, and AUC (area under the curve). However, for performance evaluation in the regression method, the R2 score, the MAE (Mean Absolute Error), adjusted R2 and the MSE (Mean Squared Error) are used.

Confusion Matrix- In situations where exact values are well-known, confusion matrix is used to measure the performance. Confusion matrix is a tabular representation of the predicted outcomes of any binary classifier. Implementing the confusion matrix is simple.

F-Score or F1 Score - A binary classification model's F1score also referred to as the F Score which is applied for evaluating it based on predictions made for the positive class. It is computed with precision and recall in mind. It is a particular kind of single score that combines both recall and precision.

AUC- Range which comes under the whole receiver operative characteristic curve, which is referred by ROC curve in two dimensions, is determined by AUC, as its name indicates. There are occasions when we need to see the classification model's performance on charts, in which case the AUC-ROC curve can be beneficial. It is a well-liked and significant statistic for assessing how well the classification model is functioning.

Mean absolute error (MAE) - One of the most basic metric MAE, estimates the complete variance between definite and forecast values. Absolute refers to taking a number as positive.

Mean squared error (MSE) – One of the most suitable metrics for evaluating regression is mean squared error, or MSE. MSE only takes non-negative values, most of which are positive and non-zero.

Our Research Techniques:

We conducted searches for relevant publications on Google Scholar considering specific phrases in order to steer our research project on the right path. We employed phrases like

“Prediction of breast cancer through machine learning algorithms”, “methods used in features selection”, “Techniques used for eliminating the features”. For our study we selected research papers which are published after 2017 onwards and in reputed journals like IEEE Access, Science Direct and Elsevier.

Related Work:

Naji M. A. and et al. (2021) in [13] ,used five machine learning algorithms to calculate the accuracy in predicting breast cancer and evaluated their proposed model on confusion matrix, sensitivity, precision and AUC. 97.20% accuracy achieved using a support vector machine, which was evaluated 97.20% precision among four performance evaluations.

According to Khourdifi Y and et al. (2018) in [8], Support vector machine gives 97.9% accuracy. The proposed system is further evaluated using various evaluation methods. Authors observed that among all mean absolute error, accuracy and precision are important evaluation criteria.

In [12], Mahmood, A. M and et al. investigated their research work and proposed the model based on the novel heuristic function and randomized gini index for reducing dimensionality. Forty different kinds of large datasets were used for the study and they concluded that selection of features while building a model is a very important task.

In [9], Lu Y. and et al. observed that different forms of imaging, including histology, x-ray (mammography), and ultrasound imaging, must be utilized to diagnose breast cancer. Combining the detection data from all three imaging modalities is the most effective technique to develop the conclusion of breast cancer ailments.

Table No. 1 Summery shows evaluation methods and accuracy

Reference Number	Name of the machine learning algorithm used	Name of the feature evaluation method	Number of features	Experiments done on the dataset	Accuracy in prediction (%)
[4]	Support Vector Machine, KNN, Logistic Regression, Naïve Bayes and MLP	Accuracy , Sensitivity, Specificity and Error Rate	10 features	WBCD(Wisconsin breast cancer diagnostic)	SVM – 97.59%
[5]	KNN, Naïve Bayes and j48	Accuracy, Sensitivity, error rate, precision , F-Score and Specificity	32 features	WBCD(Wisconsin breast cancer diagnostic)	KNN- 98%
[13]	Support Vector Machine,	Confusion	11 features	WBCD(Wisconsin	SVM – 97.20%

	linear regression RF(Random Forest), DT(Decision Tree), LR(Linear Regression) and KNN(K Nearest Neighbour)	Matrix, Precision, AUC, sensitivity		breast cancer diagnostic)	

Conclusion:

Breast cancer is now becoming a prominent cause of increasing female mortality rates. A variety of approaches in machine learning are utilized in the construction of models that anticipate breast cancer in its earliest stages. X-rays and mammography are not sufficient in the prediction of breast cancer. While building models, it is very important to test the performance of the model. Using several kinds of performance matrix techniques, we can assess performance. We observed in our study that among all techniques, confusion matrix is commonly used. The confusion matrix is very simple and easily calculated with any machine learning algorithm.

References

[1] Fatima N., Liu L., Hong S. & Ahmed H. (2020). “Prediction of breast cancer, comparative review of machine learning techniques, and their analysis”, IEEE Access, 8, pp. -150360-150376.

[2] Israni P. (2019). “Breast cancer diagnosis (BCD) model using machine learning”, International Journal of Innovative Technology and Exploring Engineering, 8(10), pp.- 4456-4463.

[3] Keleş M. K. (2019). “Breast cancer prediction and detection using data mining classification algorithms: a comparative study”, Tehnički vjesnik, 26(1), pp. - 149-155.

[4] Kumar G. R., Ramachandra G. A. & Nagamani K.(2013). “An efficient prediction of breast cancer data using data mining techniques”. International Journal of Innovations in Engineering and Technology (IJIET), 2(4), 139.

[5] Maliha S. K. Ema R. R., Ghosh S. K., Ahmed H., Mollick M. R. J. & Islam T. (2019). “Cancer disease prediction using naive bayes, K-nearest neighbor and J48 algorithm”, In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-7, IEEE.

[6] Memon M. H., Li J. P., Haq A. U., Memon M. H. & Zhou W. (2019). “Breast cancer detection in the IOT health environment using modified recursive feature selection”, wireless communications and mobile computing, 2019, pp.- 1-19.

[7] Padmapriya B. & Velmurugan T. (2016). “Classification algorithm based analysis of breast cancer data”, International Journal of Data Mining Techniques and Applications, 5(1), pp. - 43-49.

[8] Khourdifi Y. & Bahaj M. (2018). “Applying best machine learning algorithms for breast cancer prediction and classification”, In 2018 International conference on electronics, control, optimization and computer science (ICECOCS), pp. 1-5, IEEE.

[9] Lu Y., Li J. Y., Su Y. T. & Liu A. A. (2018). “A review of breast cancer detection in medical images”, 2018 IEEE Visual Communications and Image Processing (VCIP), 1-4.

[10] Mahesh B. (2020). "Machine learning algorithms-a review", International Journal of Science and Research (IJSR). [Internet], 9(1), pp. - 381-386.

[11] Eltalhi S. & Kutrani H. (2019). "Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review", IOSR Journal of Dental and Medical Sciences, 18(4), pp. - 85-94.

[12] Mahmood A. M., Imran M., Satuluri N., Kuppa M. R. & Rajesh V. (2011). "An improved CART decision tree for datasets with irrelevant feature" In Swarm, Evolutionary, and Memetic Computing: Second International Conference, SEMCCO 2011, Visakhapatnam, Andhra Pradesh, India, December 19-21, 2011, Proceedings, Part I 2, pp.- 539-549.Springer Berlin Heidelberg.

[13] Naji M. A., El Filali S., Aarika K., Benlahmar E. H., Abdelouhahid R. A. & Debauche O. (2021). "Machine learning algorithms for breast cancer prediction and diagnosis", Procedia Computer Science, 191, pp. - 487-492.

[14] Visa S., Ramsay B., Ralescu A. L. & Van Der Knaap E. (2011). "Confusion matrix-based feature selection", Maics, 710(1), pp. -120-127.
