Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Performance Analysis of Accuracy of Machine Learning Classifier in the Classification of Breast Cancer Disease

[1] Dr. Vrinda Sachdeva, [2] Dr. Arun Kumar, [3] Dr. Vasudha Arora, [4] Dr Ashish Kumar, [5] Dr Seema Verma

[1] *Department of Computer Science Engineering, ITS Engineering College, Greater Noida, Uttar Pradesh, India*
ORCID ID:0009-0007-0110-0512

[2] *Department of Computer Science Engineering, ITS Engineering College, Greater Noida, Uttar Pradesh, India*
ORCID ID:0000-0002-5493-8059

[3] *Department of Computer Science Engineering, Sharda University, Greater Noida, Uttar Pradesh, India*
vasudharora6@gmail.com

[4] *Department of Computer Science Engineering , ITS Engineering College, Greater Noida , Uttar Pradesh, India*
ORCID ID:0000-0003-0828-9921

[5] *Department of Computer Science Engineering , Delhi Technical Campus, Greater Noida*
ORCID ID:0000-0002-7893-671X

Abstract

The Breast cancer is recognized as a highly prevalent and life-threatening condition affecting women. In various domains, ML techniques have shown to be reliable predictors. In this study, medical data from Wisconsin breast cancer dataset were used to examine and compare supervised machine learning algorithms for predicting Breast diseases. There are several classifiers with varying level of accuracy, which is the research challenge. This paper suggests a method for enhancing the efficiency and precision of four distinct classifiers: Random Forest Tree, K-nearest neighbor (KNN), Logistic Regression, and SVM. To assess the efficacy of different algorithms, the evaluation employed both the AUC score and confusion matrix. The decision to utilize an algorithm is determined by an AUC score exceeding 0.5, which serves as an assessment metric validating the algorithm's efficacy. The employment of a machine learning technique hinges on achieving this threshold AUC score. Among all the machine learning algorithms tested, the Random forest tree attained the highest AUC score in the trial, reaching 1.0.

Keywords: *SVM, WDBC Dataset, LR, KNN, ROC, Random Forest Tree, ML*

Introduction

This section will encompass the rationales driving AI adoption in the healthcare sector, the utilization of machine learning methods, and the various classifications of related research efforts. Artificial intelligence represents a model that operates with minimal human intervention. The scope of applications for machine learning in the healthcare system is virtually limitless. Healthcare systems are leveraging AI to optimize administrative processes, proactively handle infectious diseases, and tailor patient care. This technology finds application across diverse healthcare domains, encompassing the innovation of new medical methods, management of patient records and

mitigation of chronic illnesses. Machine learning techniques encompass supervised, unsupervised, semi-supervised and reinforcement learning algorithms. The organization of this paper is as follows. Section 2 explains about the existing literature survey. Section 3 discusses the algorithm for supervised learning.Section 4 explains about the data set. Section 5 describes all the performance matrixes and Section 6 compares the results. Finally section 7 concludes the paper.

A. Supervised Learning

It is a technique where a model learns to make predictions or decisions by training on a labeled dataset, where the correct answers are provided. It learns patterns and relationships between inputs and outputs to
generate precise forecasts for data that has not been previously encountered. This approach is commonly used for tasks like classification and regression.

B. Unsupervised learning

It is a technique where a model learns patterns and structures in data without explicit labels. It discovers hidden relationships, clusters, or representations within the data. Unlike supervised learning, there are no predefined correct answers.

C. Reinforcement Learning

It is a technique where an agent learns by taking actions in an environment to maximize rewards over time. It uses trial-and-error to figure out the best actions to achieve its goals, based on feedback from the environment. Knowledge is gained via experience only.

Literature Review

Machine learning offers a range of algorithms for the detection and prediction of cancer, with some of the widely employed ones including Naive Bayes, Support Vector Machines, Logistic Regression, and Decision Trees. To train and evaluate these models, researchers often turn to popular datasets like the Breast Cancer Dataset and the Wisconsin Breast Cancer Dataset, which are commonly utilized in the field of breast cancer diagnosis [18].

In [1], A technique has been presented to boost the performance of two distinct algorithms: Decision Tree (J48) and Naive Bayes (NB).These classifiers are employed to classify data with a 10-fold cross-validation process applied. The introduction of a resampling filter resulted in a substantial improvement in the J48 classifier's performance, attaining a precision level of 98.2%. The utilization of this resampling filter during the preprocessing stage is the effective method of enhancing the efficiency of all the classifiers.

In [4], A CNN achieved an accuracy rate of approximately 88% when employed for the identification and categorization of invasive ductal carcinoma. Moreover, data mining is a widely adopted practice in the medical field for the purpose of predicting and classifying unusual occurrences, thereby enhancing our understanding of incurable diseases. The outcomes obtained from applying data mining to classify breast cancer diagnoses are indeed encouraging. Hence, this study incorporates the application of data mining.

[29] Support Vector Machines (SVMs) demonstrated better performance relative to other classifiers when applied to the WBCD dataset. The accuracy rates achieved for decision trees, artificial neural networks and SVMs were 93.6%, 94.7 %, and 95.7% respectively. The evaluation process involved conducting a 10-fold cross-validation.

Methodology

Supervised machine learning techniques utilize labeled input data to construct a predictive function that demonstrates strong performance when applied to unseen, unlabeled data. These methods are employed to address tasks related to classification and regression. In a classification problem, the outcome is a distinct value, whereas in regression problem the output represents a continuous real number. Decision Tree, logistic Regression, K-nearest neighbor (KNN) and SVM algorithms are used. This process will categorize our data into two distinct classes. Class 0 will denote benign tumors, while class 1 will indicate malignant tumors.

A. Random Forest

Random Forest algorithm uses bagging approach which generates a set of decision tree.it is based on the recursion approach.

B. Logistic Regression

Logistic regression employs a logistic function for classification tasks. When dealing with scenarios that involve only two potential results, it is specifically employed as binary logistic regression. It provides a probability

estimate for an input belonging to one of two classes, usually denoted as 0 or 1. By utilizing a sigmoid function on a weighted sum of input features, logistic regression models the relationship between features and probabilities, allowing for effective classification. The sigmoid function exhibits a curve that resembles the letter 'S' [2]. It's widely used in situations where you want to predict outcomes like yes/no, pass/fail, or true/false. As the input approaches positive infinity, the expected outcome tends to approach 1, while it tends to approach 0 as the input approaches negative infinity [2].

C. K- nearest neighbor (KNN)

K-nearest Neighbors (KNN) determines the label for a new data point by considering the labels of its closest neighbors within the training dataset. It calculates the distance between data points, select k nearest neighbors, and assigns the most common label among them to the new data point KNN is straightforward to grasp but becomes computationally demanding when dealing with extensive datasets.

D. SVM

The Support Vector Machine (SVM) is a robust machine learning technique employed for both classification and regression purposes. It identifies an optimal hyperplane to distinguish data into distinct categories, with the primary objective of maximizing the margin between these categories. The SVM algorithm seeks to discover the ideal boundary that maximizes the separation distance between the nearest data points of distinct classes and the decision boundary.SVM can also handles complex data distribution and perform well on high dimensional data. It uses a kernel trick to transform data into higher dimensions and capture nonlinear relationships.

Description of Dataset

The analysis utilized the freely available WBC Dataset [18], which serves research purposes. This dataset is characterized by numerous distinct statistical variables and involves the processing of multivariate numerical data, thus classifying it as multivariate. It comprises a total of 569 rows and 32 columns. In Figure 1, there are 32 attributes, including features like radius, smoothness, and texture. One of the primary challenges associated with this dataset is to predict whether a patient is afflicted with benign or malignant cancer based on the patient's given attributes.

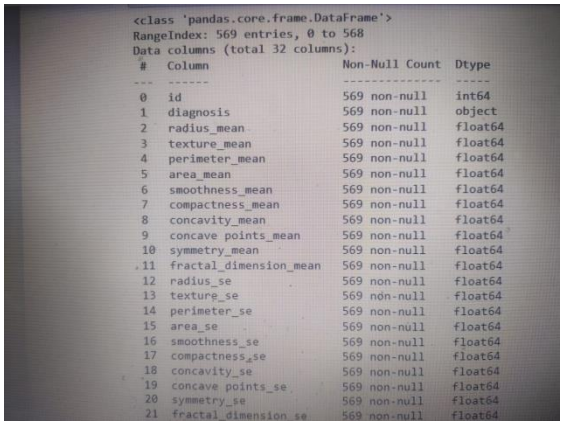

Fig. 1: Dataset for breast cancer disease

Parameter for Performance Analysis

A. Precision

It calculates the ratio of true positive instances (correctly predicted positive cases) in relation to the overall count of instances that were predicted as positive.

Precision= TP /(TP+FP)

B. Recall

It is alternatively referred to as sensitivity or the true positive rate.

Recall =TP/ (TP+FN)

C. Specificity

This metric quantifies the ratio of accurately predicted negative instances among all the actual negative instances.

Specificity = TN/ (TN+FP)

D. F-Score

The F-score is a metric that represents the harmonic mean of both recall and precision.

F-Score=2*Precision*Recall / (Precision + Recall)

E. ROC

Receiver operator characteristics curve (ROC) is used to measure how a binary classification can separate two classes using true positive rate and false positive rate.

Result and Discussion

An evaluation was conducted using various machine learning techniques, including Support Vector Machine (SVM), Random forest, Logistic Regression (LR) and K-Nearest Neighbors (KNN). The analysis was performed on Python, along with machine learning libraries such as NumPy, Pandas, and Scikit-learn utilized for this study. The computations were executed within the Jupyter Notebook web application, which is an open-source.

The research utilizes the publicly accessible WBC Dataset for its objectives. This dataset comprises 569 rows and 32 columns. The data is partitioned such that 29% is designated for testing, while the remaining 71% is allocated for training.

There are 166 in ytest ,403 in ytrain ,5252 in xtrain and 2145 in xtest.(Total 569 records x13=7397,71% used for training and 29 % used for testing)

A. Confusion Matrix

A performance matrix is an alternative term for a confusion matrix. It is used to understand how well an algorithm is performing when dealing with imbalanced data set. It is only employed to calculate the accuracy of classification models. The observations are mentioned below.

Fig 2: Confusion matrix for RFT

The cell in row 2, column 2 of the matrix represents the count of True Negatives, which corresponds to 59 instances. These instances were accurately classified by the model as not having benign breast cancer disease. In row 2, column 1, the cell contains the count of false positives, indicating that the model incorrectly predicted the presence of benign breast cancer disease for 2 records. Moving to row 1, column 2, the cell holds the count of false negatives. Lastly, the cell in row 1, column 1 represents the count of True Positives, which stands at 101. In these 101 cases, the model successfully predicted the presence of benign breast cancer disease.

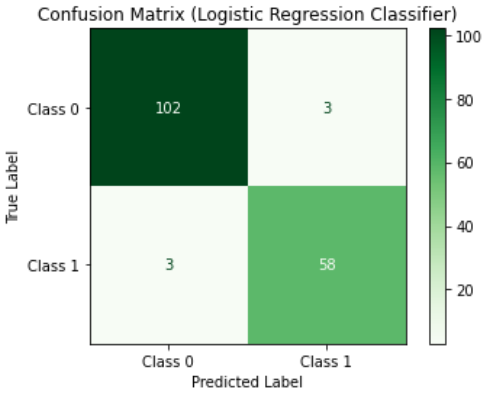

Fig 3: Confusion matrix for LR

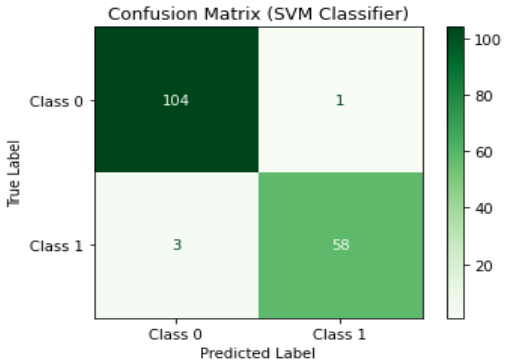

Fig 4: SVM's representation of confusion Matrix

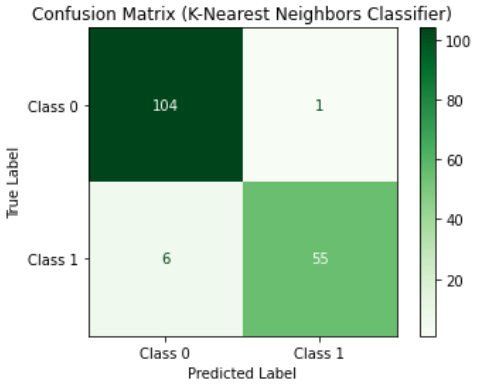

Fig. 5: Confusion Matrix for KNN

Table 1: classification reports for four distinct classifiers

Algorithm	Accuracy	Precision	Recall	F-Score
LR	98	0.50	0.53	0.23
RF	99.9	0.93	0.91	0.92
KNN	99.3	0.65	0.44	0.64
SVM	99.1	0.7	0.8	0.69

Figures 2, 3, 4, and 5 display the performance metrics, while the results in Table 1 reveal that KNN excels in accuracy, precision, and the F Score, whereas Random Forest demonstrates the highest performance in terms of recall.

Displayed in Figure 6, the ROC curve derives from the model's probability and within the range of 0 to 1. The ROC curve illustrates how the False Positive Rate (FPR) and the True Positive Rate (TPR) are interconnected across different thresholds. This represents the balance between sensitivity and specificity. Classifiers exhibit superior performance as their curves approach the upper-left part of the graph. Enhanced models encompass larger areas beneath their curves.

In the context of Figure 6, the area under the curve (AUC) is noticeably more substantial for random forest than for SVM, KNN and LR. It's crucial for a good model to possess high AUC score. Fortunately, this criterion is met for our classifiers, as their AUC scores exceed 0.5.

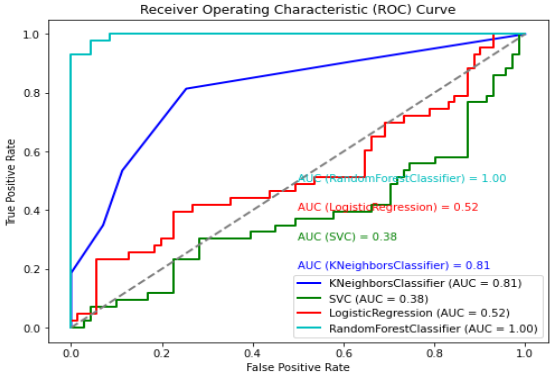

Fig. 6: Comparing the ROC curve of various classifiers

At the time of comparison of different classifiers, it is valuable to summarize their performance by computing AUC. The AUC indicates the level of separability, where instances identified as positive are more likely to receive higher rankings in terms of probability compared to those identified as negative.

It serves as a measure of how effectively a model can distinguish between classes. A greater AUC value indicates that the model is more accurate at distinguishing between 0s and 1s when making classifications.

Similarly, the model's ability to distinguish between patients with benign and malignant tumors can be assessed. Figure 6 presents the AUC scores of an alternative classifier, clearly demonstrating that the Random Forest classifier outperforms the LR, SVM, and KNN classifiers.

Conclusion

Machine learning techniques have proven their effectiveness as predictive tools in various domains. Early detection of breast cancer can significantly save lives. In this particular study, an examination and comparative analysis were conducted on four widely recognized algorithms used for predicting breast cancer. The study employs the WDBC dataset to assess and compare the performance of various classification algorithms.

As per table 1 ,the K-Nearest Neighbors (KNN) algorithm emerges as the most effective choice, outperforming other methods in accuracy, precision, and F-score.

References

[1] Sachin Upadhyay, Anjali Dwivedi, Ashutosh Verma, Vatsya Tiwari. "Heart Disease Prediction Model using various Supervised Learning Algorithm" , 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), 2023,DOI: 10.1109/CSNT57126.2023.10134595

[2] S. Shukla, D. L. Gupta, and B. R. Prasad, "Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques, "International Journal of Database Theory and Application"2016,DOI:10.14257/ijdta.2016.9.9.10

[3] Alghodhaifi, H., Alghodhaifi, A., Alghodhaifi, and M.: Predicting Invasive Ductal Carcinoma in breast histology images using Convolutional Neural Network. In: 2019 IEEE National Aerospace and Electronics Conference (NAECON), pp. 374–378 (2019), DOI: 10.1109/NAECON46414.2019.9057822

[4] Noreen Fatima; Li Liu; Sha Hong; Haroon Ahmed ,Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis in IEEE Access Vol. 8, pp150360 – 150376, Electronic ISSN: 2169- 3536 INSPEC Accession number :19974665 ,DOI: 10.1109/ACCESS.2020.3016715 in 2020.

[5] Hiba Asria ,Hajar Mousannifb ,Hassan Al Moatassime c ,Thomas Noeld, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016) DOI:10.1016/j.procs.2016.04.224

[6] Mohammed Abdullah Hassan Al-Hagery , Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques In International Journal of Advanced Biotechnology and Research (IJBR) ISSN 0976-2612, Online ISSN 2278–599X, Vol-7, Issue-2, 2016, pp760-772

[7] Chagpar, A.B.; Coccia, M. Factors associated with breast cancer mortality-per-incident case in low-to-middle income countries (LMICs),J. Clin. Oncol. 2019, DOI: 10.1200/jco.2019.37.15_suppl.1566

[8] Turkki, R.; Byckhov, D.; Lundin, M.; Isola, J.; Nordling, S.; Kovanen, P.E.; Verrill, C.; von Smitten, K.; Joensuu, H.; Lundin, J.; et al. Breast cancer outcome prediction with tumour tissue images and machine learning. Breast Cancer Res. Treat. 2019, 177, 41–52. DOI: 10.1007/s10549-019-05281-1

[9] Guo, Y.; Shang, X.; Li, Z. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing 2019, 324, 20–30.DOI:10.1016/j.neucom.2018.03.072

[10] Golden, J.A. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. JAMA 2017, DOI: 10.1001/jama.2017.14580

[11] Li L., Pan X., Zhang L.,Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. Multimed. Tools Appl. 2020,DOI: 10.1007/s11042-018-6970-9

[12] Zhu, Z.; Albadawy, E.; Saha, A.; Zhang, J.; Harowicz, M.R.; Mazurowski, M.A. Deep learning for identifying radiogenomic associations in breast cancer. Comput. Biol. Med. 2019, 109, 85–90. DOI: 10.1016/j.compbiomed.2019.04.018

[13] Bejnordi, B.E.; Veta, M.; Van Diest, P.J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.A.; Hermsen, M.; Manson, Q.F.; Balkenhol, M.; et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017, 318, 2199–2210. doi:10.1001/jama.2017.14585

[14] Bi, W.L.; Hosny, A.; Schabath, M.B.; Giger, M.L.; Birkbak, N.J.; Mehrtash, A.; Allison, T.; Arnaout, O.; Abbosh, C.; Dunn, I.F.; et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J. Clin. 2019, 69, 127–157. DOI: 10.3322/caac.21552

[15] Ahmad, L.G.; Eshlaghy, A.; Poorebrahimi, A.; Ebrahimi, M.; Razavi, A. Using three machine learning techniques for predicting breast cancer recurrence. J. Health Med. Inf. 2013,DOI: 10.4172/2157-7420.1000124

[16] Bi W.L., Hosny A., Schabath M.B., Giger M.L., Birkbak N.J., Mehrtash A., Allison T., Arnaout O., Abbosh C., Dunn I.F., et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J. Clin.* 2019; 69:127–157. doi: 10.3322/caac.21552.

[17] V. Sachdeva and V. Arora, "Comparative Analysis of Accuracy of Supervised Learning Classifier for breast cancer classification," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 395-399, doi: 10.1109/ICAC3N56670.2022.10074298.

[18] Sachdeva V., Gupta S., "Vulnerability Assessment For Advanced Injection Attacks Against Mongodb," -Journal of Mechanics of Continua and Mathematical Sciences - Web of Science ,vol. 14, no. 1, pp. 402– 413, January-February (2019).ISSN (Online):2454-7190 ISSN (Print):0973-8975 https://doi.org/10.26782/jmcms.2019.02.00028

[19] Khan, S.; Islam, N.; Jan, Z.; Din, I.U.; Rodrigues, J.J.C. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition. Lett. 2019, 125, 1–6. DOI: 10.1016/j.patrec.2019.03.022

[20] V. Sachdeva and S. Gupta, "Basic NOSQL Injection Analysis and Detection On MongoDB," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-5, DOI: 10.1109/ICACAT.2018.8933707.

[21] Naiwrita Borah, Udayan Baruah, TR Mahesh, V Vinoth Kumar, D. Ramya Dorai, Jonnakuti Rajkumar Annad. "Efficient Assamese Word Recognition for Societal Empowerment: A Comparative Feature-Based Analysis.",IEEE Access, 2023 DOI: 10.1109/ACCESS.2023.3301564

[22] Shen, L.; Margolies, L.R.; Rothstein, J.H.; Fluder, E.; McBride, R.; Sieh, W. Deep learning to improve breast cancer detection on screening mammography. Sci. Rep. 2019, 9, 1–12. DOI: 10.1038/s41598-019-48995-4

[23] Coccia, M., Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence, Technology in Society, 2020, DOI: 10.1016/j.techsoc.2019.101198

[24] Lamy, J.B.; Sekar, B.; Guezennec, G.; Bouaud, J.; Séroussi, B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artif. Intell. Med. 2019, 94, 42–53.

[25] DOI:10.1016/j.artmed.2019.01.001

[26] Kumar, A., & Panda, S. P. Performance analysis of neuro linguistic programming techniques using confusion matrix. Indonesian Journal of Electrical Engineering and Computer Science, 25(3), pp.1696-1702,2022.