AFJBS

African Journal of Biological Sciences

Machine LearningBased Breast Cancer Classification Prediction

Dr S Vasundhara[0000-0001-7682-8347]

Department of Humanities & Mathematics

vasucall123@gmail.com

Dr.M.Madhavilata[0000-0002-3211-3451]

Department of Humanities & Mathematics

madgnits@gmail.com

Dr.M.Aparna[0000-0002-2329-3067]

Department of Humanities & Mathematics

maparna@gnits.c.in

G Narayanamma Institute of Technology and Science
Shaikpet, Hyderabad – 500104

Article History
Volume 6, Issue 5, 2024
Received: 15 May 2024
Accepted: 02 Jun 2024
doi: ***10.48047/AFJBS.6.5.2024. 8799-8805***

Abstarct:

Breast cancer is the biggest cause of death among women worldwide. Early identification and diagnosis can significantly enhance breast cancer survival and treatment.Breast cancer mortality are increasing substantially each year. It is the most frequent type of cancer and the leading cause of death among women worldwide. Any progress in cancer detection and prognosis is critical for living a long and healthy life. As a result, having a high level of accuracy in cancer prognosis is crucial for updating patient survival standards and treatment options. As technology evolved, breast cancer prediction models were created utilising machine learning and artificial intelligence. The study attempts to forecast breast cancer using data visualisation and machine learning techniques. Data Visualisation Techniquesare used to investigate and analyse datasets, discovering critical factors that influence prediction performance. We utilise a permutation-based approach to calculate feature relevance and visually represent it with bar graphs.This study demonstrates that machine learning algorithms and data visualisation can accurately forecast breast cancer.Machine learning approaches have proven to be a potent technology, have become a research hotspot, and can make a substantial contribution to the early diagnosis and prediction of breast cancer. In this study, we used the Breast Cancer dataset from Kaggle to test five machine learning algorithms: Support Vector Machine (SVM), Discriminat, Logistic Regression, Naive Baye's (C4.5), and K-Nearest Neighbours.After receiving the results, the different classifiers' performance is reviewed and compared. The fundamental purpose of this research work is to discover the most efficient machine-learning algorithms for breast cancer diagnosis and prediction based on confusion matrix, accuracy, and precision. The Support Vector Machine is demonstrated to have achieved the The greatest accuracy is 97.6%, beating all other classifiers.This shows a lower level of performance than others. The outcomes can provide substantial information to healthcare practitioners, helping to improve diagnostic and therapeutic options for breast cancer patients.

Keywords:Breast Cancer prediction, ClassifierAlgorithms, Machine Learning Algorithms,SVM,KNN,CancerTreatment.

1.Introduction:In December 2020, the International Agency for Research on Cancer (IARC) revealed that breast cancer is now the most commonly diagnosed cancer in women globally, overtaking lung cancer.. Over the last two decades, the number of cancer diagnoses has nearly doubled from an estimated tenMillion in 2000 to 19.3 million by 2020 [1]. Cancer affects around one in every five persons worldwide. Cancer diagnoses are projected to increase by approximately 50% by 2040 compared to 2020. Cancer deaths have grown from 6.2 million in 2000 to 10 million in 2020. More than one in six.The cause of death is cancer. This highlights the need of investing in both cancer treatment and preventive efforts. Successfully implementing ICT in medical practice is crucial for transforming the healthcare system. and more specifically in cancer care. Big data has significantly increased data amount and value. Big data analysis of unstructured, heterogeneous, non-standard, and incomplete healthcare data has significantly impacted business intelligence (BI). Forecasting and decision-making are key benefits of this technology, which aims to improve patient care and minimise healthcare costs. Data mining algorithms were utilised.In the healthcare industry, AI plays a crucial role in disease prediction, diagnosis, cost reduction, and real-time decision-making, ultimately saving lives. The most typical data mining modelling aims are categorization and prediction, which uses a algorithms for predicting breast cancer. This paper compares the performance of five classifiers: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K Nearest Neighbours (KNN Network), which are considered among the most influential and top 10 data mining algorithms [2]. Our goal is to employ machine-learning algorithms to predict and detect breast cancer, and then rank the most effective approaches based on their performance. We compared each classifier's confusion matrix, accuracy, precision, and sensitivity. The rest of the paper is structured as follows. Section 2 describes the methods and conclusions of previous studies on breast cancer diagnosis. Section 3 explains the methodology for our work. Sections 3.1 and 3.2 present and elaborate on the experimental results.

2.Related Works:

Breast cancer prediction and diagnosis can be accomplished using a wide range of machine learning methods. Some machine learning algorithms include Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbours (KNN Network). Many researchers realised research in Breast cancer was studied utilising multiple datasets, including SEER, mammogram pictures, Wisconsin, and hospital datasets. Using these datasets, authors extract and pick attributes to complete their research. This is some notable research. Sudarshan Nayak [3] compares various supervised machine learning algorithms for breast cancer classification using 3D pictures and concludes that SVM is the most effective overall performance. In contrast, B.M. Gayathri [4] conducted a comparative analysis using Relevance vector machine, which has a lower computational cost than other machine learning approaches utilised for breast cancer detection. RVM outperforms other machine learning algorithms for breast cancer diagnosis, with 97% accuracy even with limited variables. Hiba Asri [5] found that Support Vector Machine (SVM) is highly effective in predicting and diagnosing breast cancer, with a precision of 97.13% and minimal error rates. In a recent study, Youness Khoudfi and Mohamed Bahaj [6] compared machine learning algorithms and discovered that SVM outperformed others.This classifier outperforms K-NN, RF, and NB with an accuracy of 97.9%. It utilises Multilayer Perception with 5 layers and 10 times cross validation using MLP. The author, Latchoumiet TP [7] Found a classification value of 98.4%. For classification, we propose an optimisation weighting of the particle swarm (WPSO) using the SSVM.Ahmed Hamza Osman [8] provided a technique for diagnosing Wisconsin breast cancer (WBCD) with a 99.10% prediction rate using the SVM algorithm, which combines a clustering approach with an efficient method.Our research focused on this machine learning algorithms and comes with a conclusion of best methodology for breast cancer prediction and diagnosis.

3.Methodology:

Our experiment aimed to identify the most effective and predictive algorithms for detecting breast cancer using machine learning classifiers such as Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbours (KNN) on Breast Cancer Wisconsin Diagnostic.dataset and compare the results to determine which model has the highest accuracy.

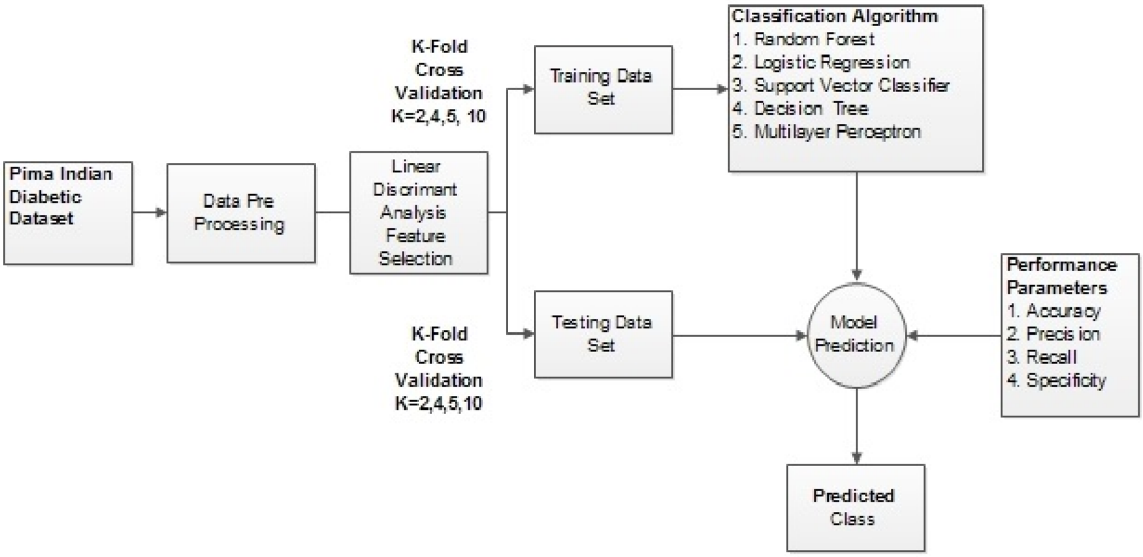


Figure 1

Our methodology starts with data collecting and pre-processing, which includes four steps: cleaning, selecting attributes, setting target roles, and extracting features. The prepared data is utilised to create machine learning algorithms that forecast breast cancer using new metrics. To evaluate the algorithmsWe show the model new data that we have labelled. To split labelled data into two pieces, we use the Train_test_split technique. Our machine learning model is built using 75% of the data, often known as the training data or set. 25% of the data will be utilised to evaluate the model's performance, known as the test data or test set. After testing,We evaluate data to select the most accurate and predictive algorithm for detecting breast cancer.

3.1. Machine Learning Algorithms.

Our study focuses on predictive analysis of machine learning techniques. Machine learning. In our study, we employed the Support Vector Machine (SVM) algorithm, which divides datasets into classes and identifies the maximum marginal hyper plane (MMH) by analysing neighbouring data points.

Random forests, also known as random decision trees, are an ensemble method used for classification, regression, and other applications. During training, they generate a huge number of decision trees and output the average of the individual trees (classification or regression). Random decision forests address the tendency of decision trees to overfit their training sets.
• K-Nearest Neighbours (K-NN) is a supervised classification algorithm. The algorithm uses previously identified points to educate itself how to label new ones. Label a new The algorithm considers the labels.points closest to the new point as its nearest neighbours, and votes on their behalf [10].Mohammed Amine Naji et al., Procedia Computer Science 191 (2021), 487–492. 4Naji Mohammed Amine/Procedia Computer Science 00 (2021) 000–000

• Logistic regression is a powerful modelling technique that builds on linear regression [11]. Logistic regression evaluates the likelihood of a disease or health condition using risk factors and covariates. Simple and multiple logistic regression models look at the link between independent variables (Xi), also known as exposure or predictor variables, and a binary dependent variable (Y), also known as the outcome or response variable. It is generally used for predicting binary or multiclass outcomes. Dependent variables. • The Decision Tree C4.5 is a versatile predictive modelling tool. An algorithmic method can partition a dataset based on a various conditions.

3.2 Data Set description:

The data set used in this investigation is from the Kaggle website, specifically the Breast Cancer Diagnostic dataset [13]. The dataset's characteristics are calculated using a digitised image of a breast cancer sample obtained via fine-needle aspiration (FNA). These parameters determine the appearance of the cell nuclei in the photograph. Breast Cancer Wisconsin Diagnostic has 569 cases (357 benign and 212 malignant), two classes

(62.74% benign and 37.26% malignant), and 11 integer-valued attributes (Id, Diagnosis, Radius, Texture, Area, Compactness, Concavity, and Concave points). -Symmetry (fractional dimension).

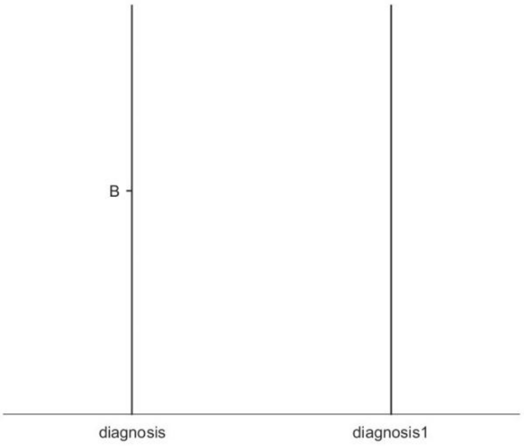
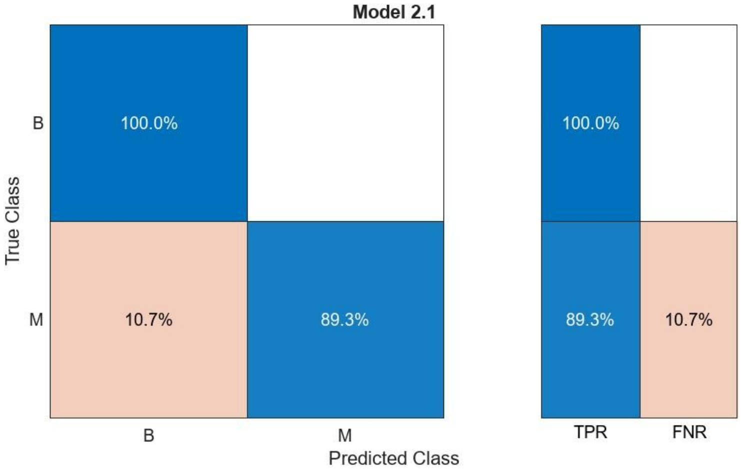


Figure 2

3.3 Results and Discussions:

After running Machine Learning algorithms on the Breast Cancer dataset. To establish the best algorithm for breast cancer prediction, we assessed and compared models using measures such as Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, and AUC. The confusion matrix is a technique for evaluating the performance of a classification problem having two or more classes as output. A confusion matrix is a table with two dimensions: "Actual" and "Predicted", as well as "True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)". Classification algorithms are frequently tested on their accuracy. It is defined as the proportion of correct forecasts to all predictions made. The document demonstrates precision. Retrievals are the number of correct documents retrieved by our machine learning model.Sensitivity is defined as the number of positive results returned by the ML model.F1 ratings represent the weighted average of precision and sensitivity.



Algorithm	Training	Testing
Discriminant	94.21965	94.69027
SVM	97.68786	97.34513
Efficient Logistic Regression	93.64162	90.26549

KNN	97.68786	95.57522
Naive Bayes	94.79769	90.26549
Kernel	98.84393	94.69027

Table1

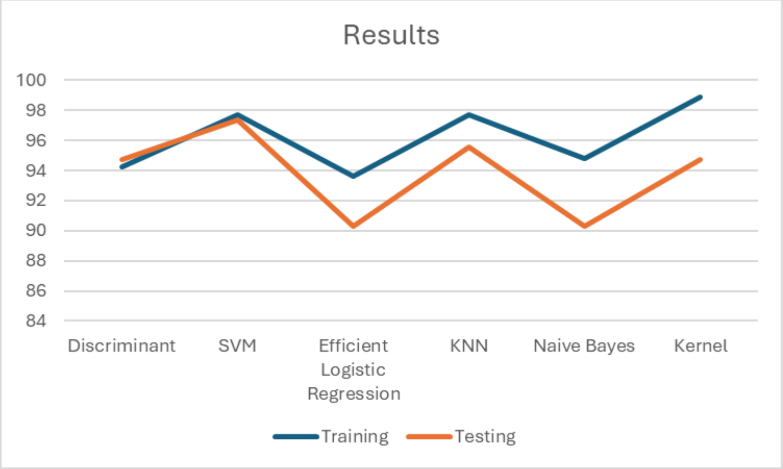


Figure 3

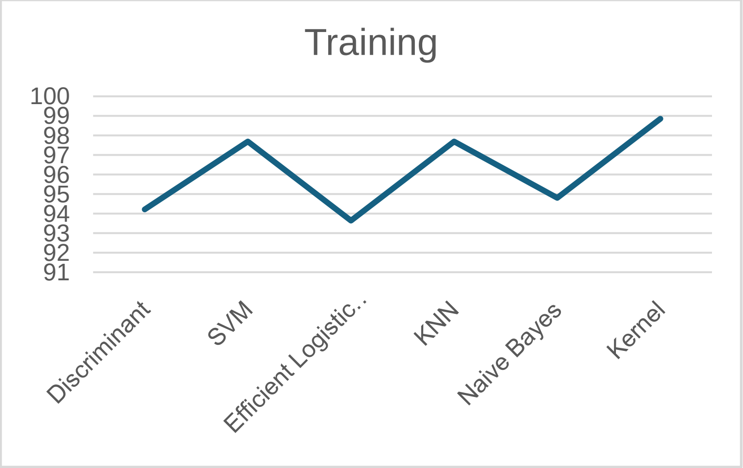


Figure 4

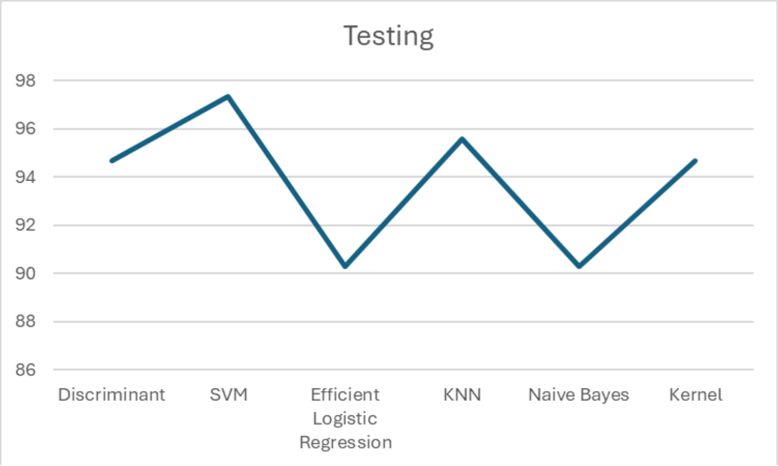


Figure 5

Row Labels	**Sum of Testing**	**Sum of Training**
Discriminant	94.69026549	94.21965318
Efficient Logistic Regression	90.26548673	93.6416185
Kernel	94.69026549	98.84393064
KNN	95.57522124	97.68786127
Naive Bayes	90.26548673	94.79768786
SVM	97.34513274	97.68786127
Grand Total	**562.8318584**	**576.8786127**

Table 2

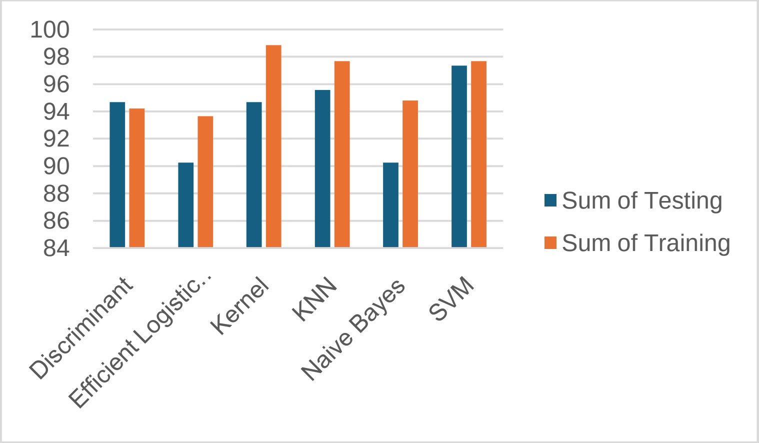


Table 3

Conclusion:

We investigated the Breast Cancer Diagnostic dataset from the Kaggle website using five algorithms: SVM, Random Forests, Logistic Regression, Decision Tree, and K-NN. To find the best machine learning model, the results were evaluated using the confusion matrix, accuracy, sensitivity, precision, and AUC metrics. Algorithms that are accurate, trustworthy, and produce higher degrees of accuracy. The algorithms were created in Python with the scikit-learn module in the Anaconda environment. Our investigation found that the Support Vector Machine beat all other approaches, with 97.2% efficiency, 97.5% precision, and 96.6% AUC. Support Vector Machine beats other methods in predicting and detecting breast cancer due to its high accuracy and precision.

References

[1] ‘WHO | Breast cancer’, WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[3] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[4] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, ‘Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis’, Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[6] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 4225- 2/18/$31.00 ©2018 IEEE.

[7] L. Latchoumi, T. P., & Parthiban, “Abnormality detection using weighed particle swarm optimization and smooth support vector machine,” Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[8] A. H. Osman, “An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[9] Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.

[10] Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004. [11] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag;2001.

[12] Quinlan JR. C4.5: Programs for Machine Learning.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[13] “https://www.kaggle.com/datasets/bittupanchal/breast-cancer-detection-dataset

[14] Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.