"Bodh", BPIT's International Journal of Technology & Management
BBIJTM, ISSN: 2454-8421, Volume 10, 2024

A Comparative Analysis of Machine Learning Classifiers for Breast Cancer Prediction

Anju Kaushik, Ayushi
Bhagwan Parshuram Institute of Technology, Delhi, India
BPIT, Delhi, India
anjukaushik@bpitindia.com, aayushisingh295@gmail.com

Abstract-This study examines the different machine learning models used for breast cancer classification. The effectiveness of various methods in differentiating between benign and malignant tumors is assessed through the use of medical data. The dataset used was acquired from the UCI Machine Learning Repository, The Wisconsin Breast Cancer Dataset is publicly available with 569 cases, out of which 357 are benign and 212 are malignant. Every machine learning model is evaluated on the basis of performance metrics like area under the curve (AUC), accuracy, precision, recall, F1-score, and computational efficiency. This project uses dual stage evaluation of all models, meaning evaluation is first done on baseline data and then on standardized data. This is done to illustrate how feature scaling affects the classification accuracy. Results show that the Support Vector Machine (SVM) model using a linear kernel consistently outperformed all the other algorithms after standardization was done and attained an accuracy of 98 percent. Hence, linear SVM is the best candidate for breast cancer detection because it successfully identifies patterns in structured medical data using the correct preprocessing.

Keywords: Breast Cancer Prediction, Machine Learning, Support Vector Machine (SVM), Logistic Regression, Feature Standardization, Classification Algorithms.

I. INTRODUCTION

Breast cancer is massive global health issue. It impacts millions of lives every year and is the leading cause of cancer-related mortality worldwide. [1-2] Therefore its necessary to timely and accurately diagnose breast cancer in order to improve patient prognosis and the survival rates. While traditional diagnostic methodologies, including various clinical examinations, medical imaging, and histopathological analysis have been the basis of diagnosis till now; these techniques often have certain limitations in like the invasiveness, subjectivity, and size of data required for any substantial evaluation. [3] As technology is advancing in this age, application of machine learning (ML) techniques has offered a vital aid, having the potential to develop non-invasive, automated, and highly accurate tools for all types of disease prediction and management. [4-5]

Breast cancer is primarily caused by aberrant growth in the breast's fatty and fibrous tissues. As cancer cells spread throughout the tumors, it leads to different stages of cancer. Breast cancer can also take different forms as the impacted cells and tissues spread in the entire body. [6]. First type is the non-invasive cancer or ductal carcinoma in situ (DCIS) is a form of breast cancer that develops when aberrant cells spread outside the breast. The second type is Invasive Ductal Carcinoma (IDC).

© *"BBIJTM 2024", All Rights Reserved*

[7] This type of cancer is typically found in men. The third type of breast cancer is called Invasive mammary breast cancer or Mixed Tumors Breast Cancer (MTBC)[8]. This type of cancer occurs in case of present abnormal duct and lobular cells. The fourth type of cancer is called Lobular Breast Cancer (LBC) this usually develops inside the lobule. Contracting this increases the risk of getting other invasive cancers. The fifth type of breast cancer is the Mucinous breast cancer (MBC) or the colloid breast cancer, this usually arises from invasive ductal cells [9]. This occurs when the duct is surrounded by aberrant tissues [10]. The final kind of breast cancer generally results in breast swelling and reddening. It is called inflammatory breast cancer (IBC). It first manifests when lymph vessels in the break the cell block and then it spreads rapidly [11].

The intersection of computer science and medicine has paved way for the utilization of sophisticated algorithms in medical treatments. These algorithms are capable of analyzing complicated patterns within complex biological datasets. Specifically looking at breast cancer, this means the analysis of a multitude of diagnostic features, these can be anything ranging from morphological characteristics of cell nuclei that are extracted from fine-needle aspirates or patterns that are identified in mammographic images and genetic markers. Due to advancements in the field of artificial intelligence and machine learning, computers can learn from data and perform better on tasks without any explicit programming. [12] Computers look for patterns in data using algorithms, then predictions or decisions are made using the information collected. [13] These algorithms can learn to identify features indicating malignancy that might escape natural vision or are hard to figure out from traditional methods. This is done by training the machine learning models on vast repositories of structured data. This method is promising for enhancing efficiency and accuracy of breast cancer screening and diagnosis. It ultimately leads to earlier diagnosis and improved patient outcomes.

This research endeavors to contribute to the already existing body of knowledge in this critical field by analyzing the comparative performance of several well-established supervised learning classification algorithms for breast cancer prediction. By accessing the widely used and publicly available breast cancer dataset, we aim to evaluate the efficacy of several machine learning models such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests in accurately distinguishing between benign and malignant breast tumors based on a set of diagnostic features. All the essential steps involved in a typical machine learning pipeline will be discussed in meticulous detail including the rigorous data pre processing techniques to ensure that the data is of quality, model is robust, careful model training methodologies and the vital step of hyperparameter tuning to optimize the prediction of each of the algorithms. [14-16]

The primary focus of this investigation will be the comprehensive and the comparative evaluation of the models that are being trained. [17] We will dive into nuanced assessment of model performance, beyond overall accuracy. We will employ a range of relevant evaluation metrics like precision, recall, and the F1-score. [18] These metrics provide an in depth understanding of the models' ability to accurately identify the malignant cases while minimizing false positives and false negatives, these are the critical considerations in any medical diagnostic context because both over-diagnosis and under-diagnosis can have significant consequences. We will also analyze confusion matrices, giving visual representation of the classification performance and provide strength and weakness for every model. [19]

This research aims to identify a machine learning model with highest promising prediction capability for breast cancer using the chosen dataset. Findings of this study will provide valuable insights for both researchers and practitioners in the field and will guide the selection of machine learning techniques used for developing effective breast cancer prediction tools. The ultimate goal is to contribute to the research of data driven approaches in healthcare.

© *“BBIJTM 2024”, All Rights Reserved*

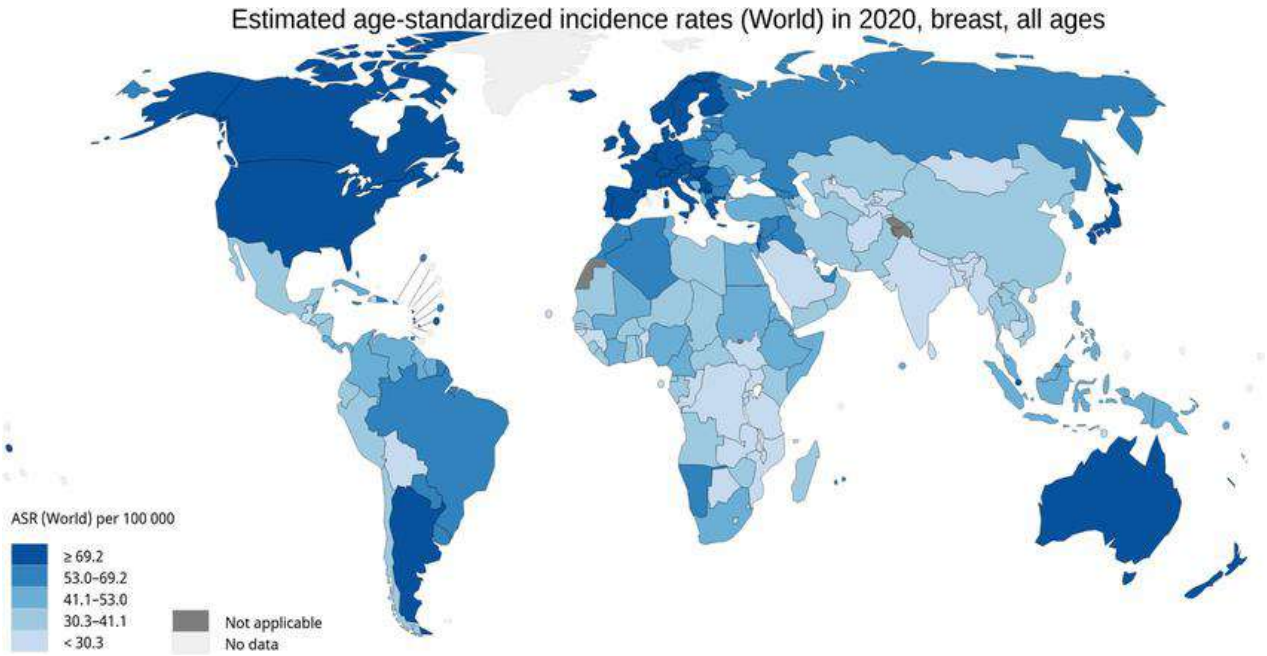


Fig. 1 Img src: Muhammad Idrees Khan, Abdelhakim Bouyahya, Naoufal E. L. Hachla (2021). "Anticancer properties of medicinal plants and their bioactive compounds against breast cancer: a review on recent investigations." Environmental Science and Pollution Research (2022) 29:24411–24444.

II. LITERATURE REVIEW

Some of the traditional techniques for breast cancer diagnosis involve a number of diagnostic modalities. These include biopsies, ultrasound and mammography. Despite their wide use in healthcare, all of these have their own drawbacks. [20-23]

Evidence suggests that there are a number of factors that significantly increase the risk of breast cancer, these include alcohol use, aging, high breast density, a family history of breast cancer, radiation exposure, obesity, and prior radiation therapy treatments [24]. While, a lower risk of breast cancer has been linked to routine physical activity, prolonged breastfeeding, limiting alcohol consumption, quitting smoking, and avoiding prolonged hormone therapy [25-26]. Breast cancer attributes to around 685,000 deaths around the world (according to data from 2020). This makes up to 24% of all cancer related fatalities. Recent improvements have lead to early detection and treatment which has decreased the global incidence and mortality rates but regional differences still exist [27]. For example, the incidence rate of breast cancer roughly is, 13 percent in India, 15 percent in Russia, 17 percent in the United Kingdom, 19 percent in the United States, 11 percent in Japan, and 16 percent in Australia. Global data shows that all continents are impacted but parts of Asia and Eastern Europe show especially high mortality rates. Hence, it is necessary to advance to better screening and prevention methods.

The subfield of machine learning, namely supervised learning techniques have advanced exponentially to improve breast cancer detection and classification. Numerous studies have been conducted that use various machine learning algorithms on imaging and clinical datasets for breast cancer diagnosis, these include artificial neural networks, support vector machines (SVMs), and ensemble methods. [28-29] Out of these, Support Vector Machines provide the most promising

© "BBIJTM 2024", All Rights Reserved

performance, they classify images, clinical and medical data related to breast cancer because they are highly efficient when handling high dimensional datasets which have fewer total observations. [30-31]

A Support Vector Machine operates by creating a hyperplane, this is a plane that divides data points from multiple classes in a higher dimensional space while maximizing the margin between them. Support vectors which are described as the data points closest to the hyperplane define each hyperplane's position. Different kernel functions enable non linear classification without using explicit transformation computation in high dimensions. Hence kernels play a huge role in optimizing an SVMs performance. [32-34]

Studies by Abbas et al. discovered that models like KNN, Logistic Regression, and SVM are extremely sensitive to data standardization. [35] They started by investigating the effect of feature scaling on different classifiers. Their results are well aligned with the two-phase assessment methodology employed in this project, wherein we test algorithms on standardized and unscaled data to look at how their stability and performance changes. Multiple review papers and meta analyses, namely the one by Chaurasia and Pal [36] have emphasized on the need to compare various machine learning classifiers to shed light on how different models function and which model has best performance. They employed the Waikato Environment for Knowledge Analysis (WEKA) version 3.6 for their investigation. Their results concluded that Naïve Bayes had the best accuracy of 97.36 percent which was significantly higher than the accuracy of 96.77 percent and 93.41 percent for the RBF network and J48 Decision Tree, respectively.

Yue et al. [37] in their paper, examined multiple machine learning methods to predict breast cancer using WBCD dataset, these included SVM, K-NN, ANN, and Decision Trees. They observed that the Artificial Neural Network architecture in conjunction with deep belief networks (DBNs) produced the best accuracy which was then closely followed by an SVM based method. The DBNs architecture had obtained 99.68% accuracy, however when using SVM with the two-step clustering has achieved 99.10% classification accuracy.

This project expands upon the body of existing literature by implementing a comparative evaluation of multiple machine learning classifiers using particular attention to preprocessing and kernel function analysis in Support Vector Machine. It utilizes the dual phase evaluation strategy to research and learn the best machine learning pipeline for structured dataset based breast cancer detection. The Wisconsin Breast Cancer Dataset remains a dependable benchmark that contributes to reproducibility and comparability of findings when doing such exploratory work. [38]

III. METHODOLOGY

Dataset Interpretation

This study utilizes the Wisconsin Breast Cancer Diagnostic dataset that is available from the UCI Machine Learning Repository. This dataset consists of 569 instances; these are characterized by 30 features; each computed from the digitized images of fine needle aspirates (FNAs) of breast mass. These features describe the characteristics of cell nuclei present in images, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The target variable is binary, indicating whether the diagnosis is malignant or benign.

© "BBIJTM 2024", All Rights Reserved

To summarize the dataset used here, it has 357 instances that are labeled as benign and 212 instances labeled as malignant.

Initial Data Analysis & Visualization

We started by inspecting the data using data.head(), df.tail(), df.info(), removing null values, etc. Then we thoroughly performed Exploratory Data Analysis to summarize:

1. Descriptive Statistics: Descriptive statistics refers to the process of summarizing numerical and categorical data in a concise and informative manner like mean, standard deviation and correlation etc.
2. Data Visualizations: To observe general trends in instances and their features we first used Unimodal Data Visualizations using frequency graph, histograms, density plots and box plots as shown below

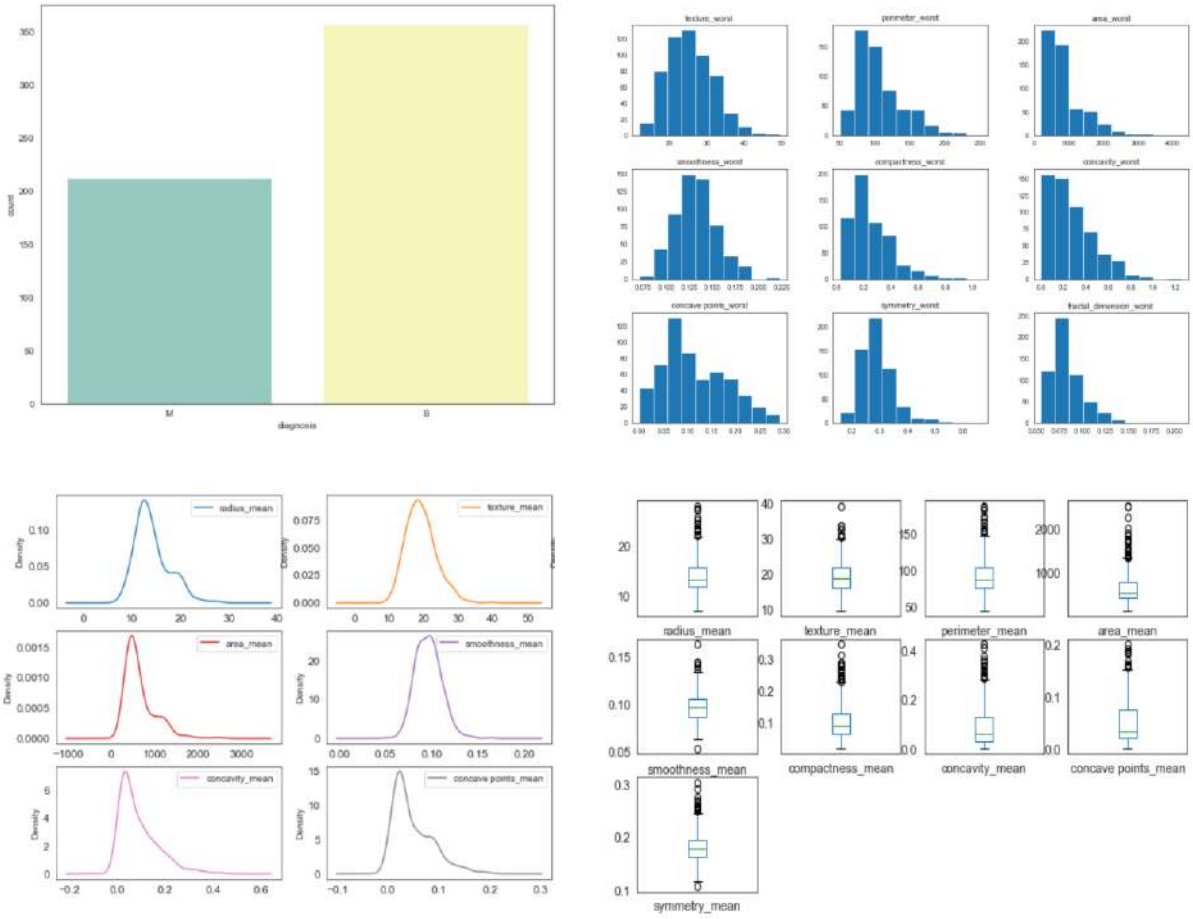


Fig. 2 Data Visualization

After this we proceed to Multimodal Data Visualization using Correlation Matrix and Scree Plot. This helps us check how data features correlate with each other. In a Correlation Matrix, we assign values from 0 to 1 to signify the level of correlation between features.

© "BBIJTM 2024", All Rights Reserved

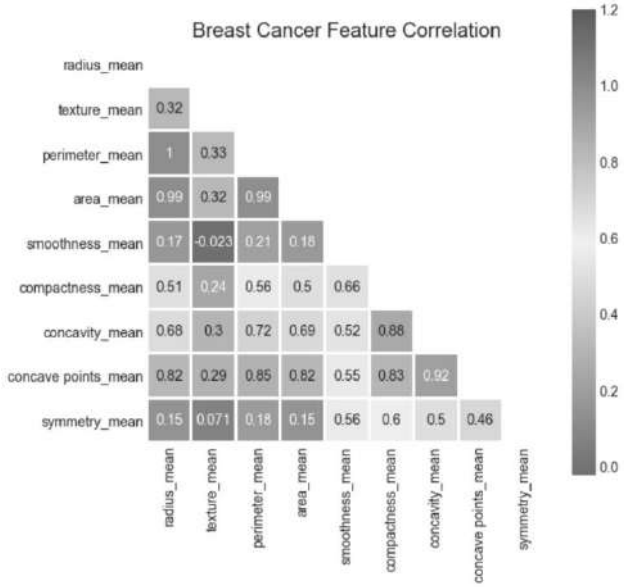


Fig. 3 Breast Cancer Feature Correlation

A strong positive relationship exists with mean values parameters.

- Parameters such as concavity and area, concavity and perimeter etc., have a strong connection (r between 0.5 and 0.75).
- Mean area of the tissue nucleus has a high positive correlation with mean values of radius and parameter.
- Similarly, we find an inverse relationship between fractal dimension and the mean values of radius, texture, and parameter.

Data Preprocessing

The dataset went through a number of pre-processing procedures before any machine learning models were trained on it. The dataset was found to have no missing values. The features were then normalized using the library scipy. After this we encode labels to classes M (malignant) and B (benign).

Additionally, to balance model training and assessment, the processed data was the separated into a training set (70%) and a testing set (30%) using sklearn. After which we applied feature standardization using Standard Scaler ().

Machine Learning Models

The following classification algorithms were implemented and evaluated:

1. Linear Regression: Linear Regression is an algorithm used in supervised learning. It is used to predict continuous numerical values. It looks at a regression line or a best fitting straight line to reduce the discrepancy between actual and the expected results. In this project we use linear regression to compare rather than for regression problem. The independent and dependent variables are assumed to have a linear relationship in the model.
2. Linear Discriminant Analysis (LDA): Linear Discriminant Analysis is a classification method that maximized class separability by projecting features onto a lower dimensional

© "BBIJTM 2024", All Rights Reserved

space. It figures out the linear feature combination that would best divide two or mode classes. LDA preassumes that each class’s data has the same covariance and is normally distributed. It performs the best with small datasets and well separated classes.

3. Support Vector Machine (SVM): For classification tasks, supervised learning has Support Vector Machine that performs exceptionally well with complicated data. It uses support vectors which are a small number of crucial data points, to determine the best hyperplane that could be used to separate classes with the largest margin possible. Support Vector Machines can be modified for non-linear classification with the help of various kernel functions and are highly effective when used in high-dimensional spaces.
4. K-Nearest Neighbors (KNN) Classifier: KNN is a straightforward instance-based learning algorithm that uses the majority label among its "k" nearest neighbors to categorize a new data point. It doesn’t create a model during the training phase, rather learns the training dataset by heart. The selection of “k” as well as the distance metric (such as Euclidean distance) significantly impacts the performance of KNN. It can be computationally costly for large datasets, but it performs well for small datasets.
5. Decision Tree Classifier: In a decision tree classifier, a decision based on a feature is represented by each internal node. This is a tree-structured model. Whereas a class label is represented by each leaf node. It uses criteria such as entropy or Gini impurity to divide the dataset into subsets according to the feature values. Decision trees can handle both categorical and numerical data. However, they have the limitation of being easily overfit.
6. Gaussian Naïve Bayes (GNB) Classifier: Gaussian Naïve Bayes is a probabilistic classifier based on Bayes’ theorem; It assumes that features are unrelated to one another. Additionally, it also assumes that the feature distributions are either normal, or Gaussian. GNB works well when there is a small datasets and is well suited for high-dimensional data.

Model Building and Implementation

To ensure that the results are fair, dependable and scientifically correct, the methodology that has been used in this study for developing, implementing, optimizing and assessing machine learning models is following a methodical and extracting pipeline. To make sure that the model comparisons were impacted by the appropriate data preprocessing, hyperparameter optimization, and multi-metric evaluation techniques in addition to raw performance, a systematic approach was required.

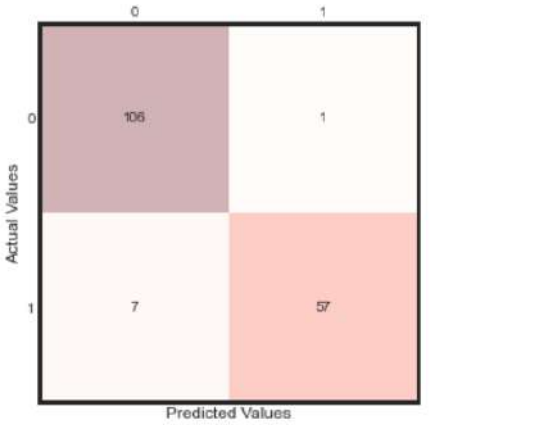
Model Formation and Selection:

Six supervised learning algorithms were accessed for this investigation based on their efficacy in classification tasks, especially in medical datasets. These were Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN) Classifier, Decision Tree Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Logistic Regression (LR). Every algorithm has unique learning biases, strengths, and presumptions. Like, Support Vector Machines can handle both linear and non-linear boundaries using kernel methods, whereas for linearly separable data, Linear Regression and Linear Discriminant Analysis are more appropriate models. KNN and decision trees are non-parametric methods that can identify complicated patterns without making significant assumptions about data.

First we generated a predictive model using Support Vector Machine (SVM) and used 3 fold cross validation to get accuracy score of 0.97%. Then we evaluate it using a confusion matrix. Using this

© *“BBIJTM 2024”, All Rights Reserved*

matrix we generate an ROC curve (Receiver Operating Characteristic Curve) to evaluate model accuracy for binary classification problems.



	precision	recall	f1-score	support
0	0.94	0.99	0.96	107
1	0.98	0.89	0.93	64
accuracy			0.95	171
macro avg	0.96	0.94	0.95	171
weighted avg	0.95	0.95	0.95	171

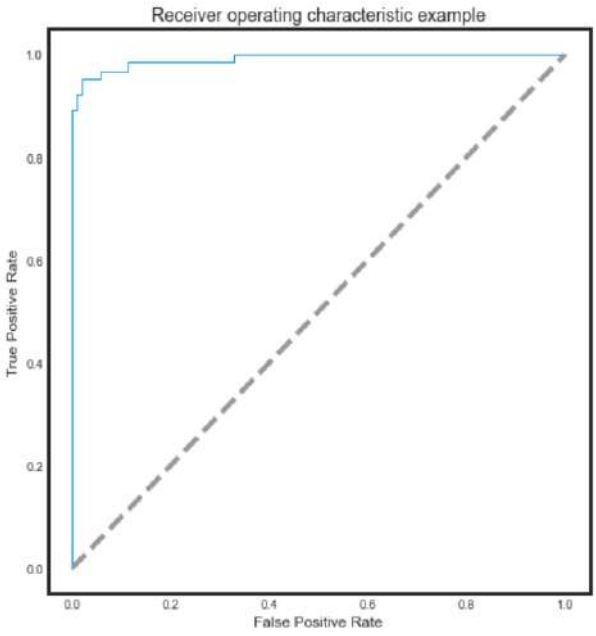


Fig. 4 Confusion Matrix for SVM

Result: For the points above the diagonal, tpr > fpr, and the model says that you are in a zone where you are performing better than random. For example, assume tpr = 0.99 and fpr = 0.01, Then, the probability of being in the true positive group is 99%.

In next step we optimize SVM classifier using 5 fold cross validation and tuning the two key parameters, the value of C and the type of Kernel. We use skit-learn library and GridSearchCV() to perform the parameter tuning.

We get the following confusion matrix with improved results after running this sequence:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	107
1	1.00	0.92	0.96	64
accuracy			0.97	171
macro avg	0.98	0.96	0.97	171
weighted avg	0.97	0.97	0.97	171

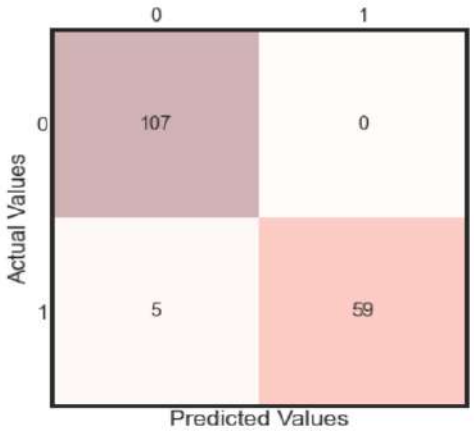


Fig. 5 Optimized Confusion Matrix

© "BBIJTM 2024", All Rights Reserved

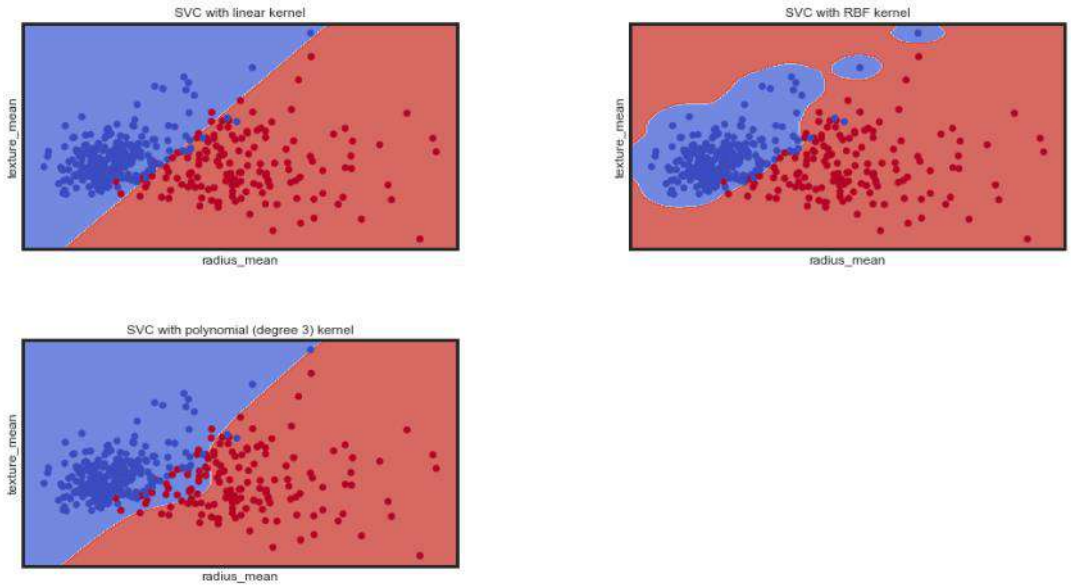


Fig. 6 SVC Decision Boundary with Different Kernels

Result: From GridSearchCV() output we get that the best parameters are {'C': 0.1, 'gamma': 0.001, 'kernel': 'linear'} with a score of 0.98. And Linear kernel is the best kernel.

Model Implementation and Training:

After initial preprocessing, the dataset was divided into training (70 percent) and testing (30 percent) sets using a stratified sampling technique to ensure a balanced distribution of dataset. The model was first trained with default hyperparameters, to create a baseline performance. Then the performance was assessed using the model's ability to distinguish between benign and malignant tumors. During training, we first evaluate algorithms on raw data to understand their default behavior. This is called Baseline Performance. We setup the test harness and implement 10 fold cross validation and then build 5 different models. The representation of the distribution of accuracy values is calculated graphically using box and whisker plots as shown below.

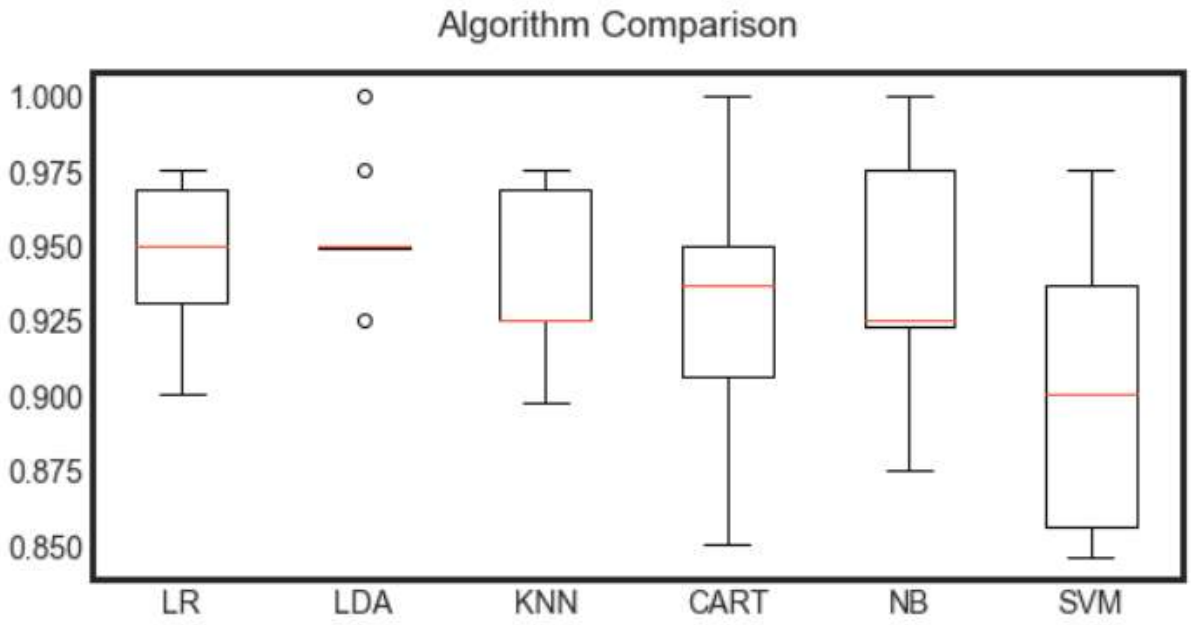


Fig. 7 Mean Accuracy Values for Baseline Algorithm Comparison

© "BBIJTM 2024", All Rights Reserved

Results show a tight distribution for all classifiers except Support Vector Machine. This suggests low variance.

Feature Scaling and Standardization:

Many machine learning models, especially SVM, Logistic Regression, and KNN, are sensitive to the scale of input features hence the dataset was standardized using scikit-learn StandardScaler. Standardization converts features to have zero mean and unit variance to prevent models from becoming skewed toward variables with larger scales.
Models were assessed before and after standardization to give a comprehensive review of the effects that scaling has and to enable more nuanced evaluation of how model behaves under various data conditions.
The results post standardization are attached below.

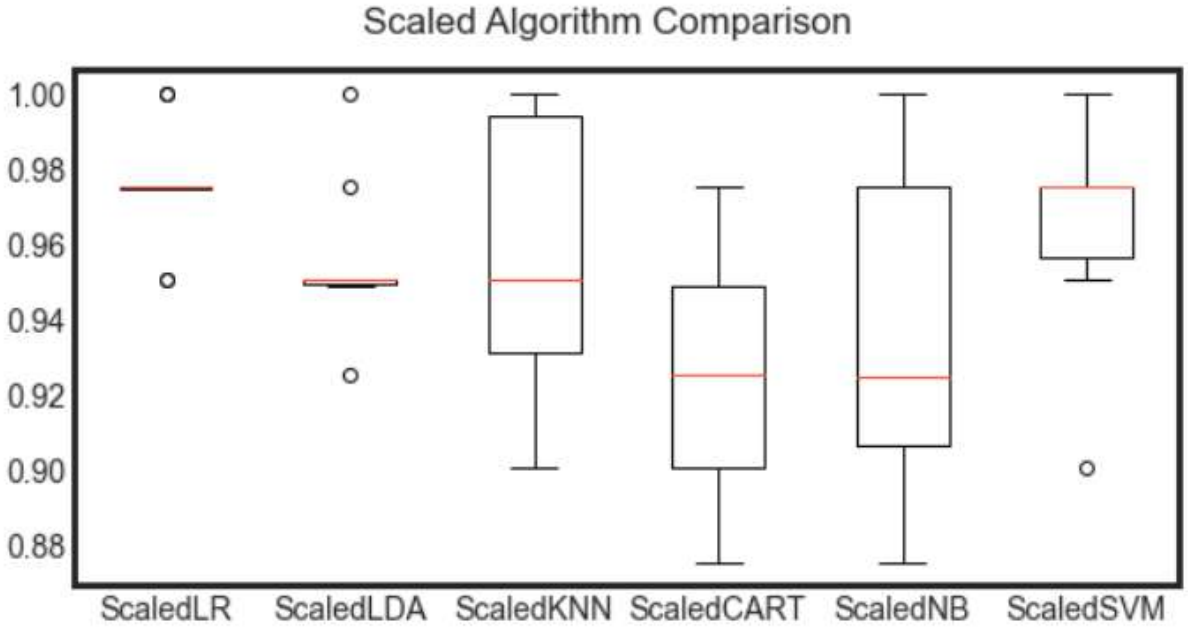


Fig. 8 Mean Accuracy Values for Standardized Algorithm Comparison

Given box plot shows the distribution of cross validation accuracy for different machine learning models after scaling.

- Scaled SVM and Scaled Logistic Regression have narrow interquartile ranges (IQRs) and high medians, hence they perform the bests, most reliable.
- Scaled SVM in particular exhibits high accuracy with little spread.
- Although Scaled Linear Discriminant Analysis has some outliers, and a slightly wider spread, it still performs well.
- On the other hand, Scaled KNN, CART, and NB are least stable because they exhibit high variability and low median accuracies.

All things considered, SVM is notable for its high accuracy and consistency, this explains bolsters its choice as the chosen model.

Hyperparameter Tuning and Optimization:

A Grid Search Cross-Validation (GridSearchCV) technique was applied to reduce the chances of overfitting as well as to optimize model performance. GridSearchCV methodically investigated

© "BBIJTM 2024", All Rights Reserved

different combinations of hyperparameter and presented with the following results for Support Vector Machine:

```
➔ Model Training Accuracy: 0.940 +/- 0.034
➔ Tuned Parameters Best Score:  0.9446794871794871
➔ Best Parameters:
 {'clf__C': 1.0, 'clf__kernel': 'linear'}
```

Evaluation Metrics

Each of the selected models was trained on the training dataset. To optimize the performance of the algorithms, hyperparameter tuning was performed using GridSearchCV with k-fold cross-validation. The trained models were then evaluated using the following metrics:

- Accuracy: Accuracy is the number of accurate predictions made out of all the cases predicted. Accuracy indicates that every result, both good and bad, was accurate. It is easy to comprehend, calculate, and operate effectively when the classes are balanced. [39]

$$Accuracy = \frac{Nmber\ of\ correct\ predictions}{Total\ Number\ of\ Predictions} = \frac{TP + TN}{TP + TN + FN + FP}$$

- Precision: The proportion of correctly identified positive cases out of all instances predicted as positive.[40]

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall (Sensitivity): The sensitivity (Recall) is the degree of effectiveness in a classification algorithm that classifies data points into a positive class. [41]

$$\text{Sensitivity(Recall)} = \frac{TP}{TP + FN}$$

- F-scores: Taking a positive real number and setting β to 1, 2, and 3 yields the F1, F2, and F3 scores. Precision and recall are two distinct metrics that are combined in the F1-Score matrix. These are the most widely used performance metrics in diagnosis and classification-related medical research.[42]

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Confusion Matrix: A table summarizing the performance of a classification model by showing the counts of true positives, true negatives, false positives, and false negatives.

Table 1: Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

At last we finalize the model using the best parameters and get general final scores and evaluation metrics as follows

© "BBIJTM 2024", All Rights Reserved

Table 2: Final Score

Metric	Value
Final Model Training Accuracy	0.940 ± 0.034
Final Accuracy on Test Set	0.94737

Table 3: Evaluation Matrix

	Predicted Positive	Predicted Negative
Actual Positive	113 (TP)	3 (FN)
Actual Negative	6 (FP)	49 (TN)

Class	Precision	Recall	F1-Score	Support
B (Benign)	0.95	0.97	0.96	116
M (Malignant)	0.94	0.89	0.92	55
Accuracy			**0.95**	**171**
Macro Avg	0.95	0.93	0.94	171
Weighted Avg	0.95	0.95	0.95	171

1. RESULTS AND DISCUSSION

Ater comparing six different machine learning models, namely Gaussian Naive Bayes, K-Nearest Neighbors, Decision Tree Classifier, Support Vector Machine, Linear Discriminant Analysis, and Logistic Regression we got the results. Accuracy, precision, recall, F1-score, and AUC-ROC scores were the performance metrics used to evaluate models on test set after training them, hyperparameter tuning and evaluating. Gaussian Naive Bayes, Decision Trees, and Logistic Regression performed much better after feature standardization, than they did while evaluating the baseline, or unstandardized data. After scaling, Support Vector Machine showed noticeable improvement and achieved high precision and recall scores.

The standardized and unstandardized datasets, both showed high performance from Random Forest Classifier. Support Vector Machine model performed the best overall with standardized data, according to the test outcome. Random Forest and Logistic Regression were second and third respectively. K Nearest Neighbor classifiers and Decision Trees are less ideal because of their lower recall values, especially in a medical setting where failing to detect a malignant case, or false negative can have fatal repercussions. Gaussian Naïve Bayes, even though straightforward, it

© "BBIJTM 2024", All Rights Reserved

produced good results, although the assumption of feature independence held back the results. The model's performance is summarized below in a table:

Table 4: Model Performance

Model	Accuracy (%)	Notes
Logistic Regression (LR)	98.24	High, stable performance
Linear Discriminant Analysis (LDA)	97.64	Very strong, slightly below LR
K-Nearest Neighbors (KNN)	96.47	Very good, but sensitive to k
Decision Tree Classifier (CART)	91.76	Lower, prone to overfitting
Gaussian Naive Bayes (NB)	94.71	Decent, simple probabilistic model
Support Vector Machine (SVM)	98.24	Equal to Logistic Regression after scaling

V. CONCLUSION

This project use several machine learning models for assessment and then diagnose features to predict the classification of breast tumors. Results showed that Support Vector Machine and Logistic Regression had the highest accuracy amongst all models after preprocessing, feature scaling, model optimization and data standardization. Whereas Linear Discriminant Analysis also showed strong performance. Evaluation was thoroughly done emphasizing scaling, model tuning, using metrics like precision, recall, and F1 score. After concluding, the study shows that when used in conjunction with appropriate preprocessing and evaluation, machine learning techniques can offer trustworthy support for diagnosis of breast tumors

REFERENCES

1. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
2. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.
3. Mangasarian, O. L., & Street, W. N. (1995). Machine learning techniques for diagnosing cancer. Computerized Medical Imaging and Graphics, 19(5), 473-493.
4. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In IS&T/SPIE 1993 International Symposium on Electronic Imaging:6 Science and Technology (pp. 861-870).
5. N. Azamjah, Y. Soltan-Zadeh, and F. Zayeri, "Global Trend of Breast Cancer Mortality Rate: A 25-Year Study," Asian Pacific Journal of Cancer Prevention, vol. 20, no. 7, pp. 2015–2020, 2019.
6. T. D. Ellington et al., "Trends in Breast Cancer Mortality by Race/Ethnicity, Age, and U.S. Census Region, United States (1999–2020)," Cancer, vol. 129, no. 3, pp. 454–466, 2023.
7. S. A. Feig, G. S. Shaber, A. Patchefsky, G. F. Schwartz, J. Edeiken, H. I. Libshitz, R. Nerlinger, R. F. Curley, and J. D. Wallace, "Advantages and limitations of physical examination and mammography in breast cancer screening," American Journal of Roentgenology, vol. 128, no. 3, pp. 403–408, Mar. 1977

© "BBIJTM 2024", All Rights Reserved

8. A. S. G. Curley, "Applications of Computational Methods in Biomedical Breast Cancer Imaging Diagnostics: A Review," Journal of Imaging, vol. 6, no. 10, p. 105, Oct. 2020.
9. Y. Khourdifi and M. Bahaj, "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.
10. F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on f ine needle aspiration biopsy data using random forest classifier," in 2013 13th International Conference on Intellient Systems Design and Applications, pp. 121–125, IEEE.
11. A. R. Chaudhury, R. Iyer, K. K. Iychettira, and A. Sreedevi, "Diagnosis of invasive ductal carcinoma using image processing techniques," in 2011 International Conference on Image Information Processing, pp. 1 6, IEEE.
12. A. B. Tuck, F. P. O'Malley, H. Singhal, and K. S. Tonkin, "Osteopontin and p53 expression are associated with tumor progression in a case of synchronous, bilateral, invasive mammary carcinomas," Archives of pathology and laboratory medicine, vol. 121, no. 6, p. 578, 1997.
13. S. Masciari et al., "Germline e-cadherin mutations in familial lobular breast cancer," Journal of medical genetics, vol. 44, no. 11, pp. 726–731, 2007.
14. A. Gradilone et al., "Circulating tumor cells (ctcs) in metastatic breast cancer (mbc): prognosis, drug resistance and phenotypic characteriza tion," Annals of Oncology, vol. 22, no. 1, pp. 86–92, 2010.
15. F. M. Robertson et al., "Inflammatory breast cancer: the disease, the biology, the treatment," CA: a cancer journal for clinicians, vol. 60, no. 6, pp. 351–375, 2010.
16. T. M. Mitchell, "The discipline of machine learning," Carnegie Mellon University, Machine Learning Department, Tech. Rep. CMU-ML-06-108, 2006.
17. P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, Oct. 2012.
18. J. Quist, L. Taylor, J. Staaf, and A. Grigoriadis, "Random Forest Modelling of High-Dimensional Mixed-Type Data for Breast Cancer Classification," Cancers, vol. 13, no. 5, p. 991, 2021.
19. B. Kurian and V. L. Jyothi, "Support Vector Machine based Breast Cancer Classification using Next Generation Sequences," in Proceedings of the First International Conference on Advanced Scientific Innovation in Science, Engineering and Technology (ICASISET), 2020.
20. S. S. Ja, P. K. S. Cb, and T. A. Assegie, "A cost-sensitive logistic regression model for breast cancer detection," The Imaging Science Journal, vol. 70, no. 1, pp. 10–18, 2023.
21. M. A. Alzubi, "Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer
22. A. S. Alhussaini et al., "Performance Assessment of Hybrid Machine Learning Approaches for Breast Cancer and Recurrence Prediction," Scientific Reports, vol. 13, no. 1, p. 11293701, 2023.
23. A. S. Alhussaini et al., "Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study," Journal of Imaging, vol. 3, no. 3, p. 42, 2023.
24. L. J. Grimm and M. A. Mazurowski. Breast cancer radiogenomics: Current status and future directions. Academic Radiology, vol. 27, no. 1, pp. 39-46, 2020.
25. A. S. Rana, J. Rafique and H. Riffat. Advances in Breast Ultrasound Imaging: Enhancing Diagnostic Precision and Clinical Utility. IntechOpen, London, 2024.
26. 3. J. H. Catani, R. Matsumoto, F. Horigome, T. Tucunduva, M. Costenaro and N. de Barros. A Pictorial Review of Breast Biopsy Complications. In: 2017: European Congress of Radiology-ECR, 2017.
27. J. Chen, Z. Gandomkar and W. M. Reed. Investigating the impact of cognitive biases in radiologists' image interpretation: A scoping review. European Journal of Radiology, vol. 166, p. 111013, 2023.
28. A. Alzu'bi, H. Najadat, W. Doulat, O. Al-Shari, and L. Zhou, "Predicting the recurrence of breast cancer using machine learning algorithms," Multimed Tools Appl, vol. 80, no. 9, pp. 13787–13800, Apr. 2021, doi: 10.1007/S11042-020-10448-W/METRICS.

© *"BBIJTM 2024", All Rights Reserved*

29. S. Rani, M. Kaur, and M. Kumar, "Recommender system: prediction/diagnosis of breast cancer using hybrid machine learning algorithm," Multimed Tools Appl, vol. 81, no. 7, pp. 9939–9948, Mar.2022, doi: 10.1007/S11042-022-12144-3/METRICS.
30. N. Arya and S. Saha, "Multi-modal advanced deep learning architectures for breast cancer survival prediction," Knowl Based Syst, vol. 221, p.106965, Jun. 2021, doi: 10.1016/J.KNOSYS.2021.106965.
31. S. Lei et al., "Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020," Cancer Commun, vol. 41, no. 11, p. 1183, Nov. 2021, doi:10.1002/CAC2.12207.
32. R. M. Al-Tam and S. M. Narangale. Breast cancer detection and diagnosis using machine learning: A survey. Journal of Scientific Research, vol. 65, no. 5, pp. 265-285, 2021.
33. S. A. Khan, et al., "Breast Cancer Detection and Prevention Using Machine Learning," Journal of Healthcare Engineering, vol. 2022, Article ID 10572157, 2022.
34. A. B. Yusuf, R. M. Dima and S. K. Aina. Optimized breast cancer classification using feature selection and outliers' detection. Journal of the Nigerian Society of Physical Sciences, vol. 3, pp. 298-307, 2021.
35. S. M. Islam, et al., "Machine learning-based diagnosis of breast cancer utilizing feature scaling and optimization with hyper-parameter tuning," Informatics in Medicine Unlocked, vol. 30, p. 100007, 2023.
36. M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke and C. F. Tsai. SVM and SVM ensembles in breast cancer prediction. PLoS One, vol. 12, no. 1, p. e0161501, 2017.
37. M. A. Khan, et al., "Breast cancer diagnosis using support vector machine optimized by intelligent dynamic differential evolution algorithm," Scientific Reports, vol. 12, no. 1, p. 61322, 2022.
38. D. Valkenborg, A. J. Rousseau, M. Geubbelmans and T. Burzykowski. Support vector machines. American Journal of Orthodontics and Dentofacial Orthopedics, vol. 164, no. 5, pp. 754-757, 2023.
39. M. Abbas, et al., "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," ResearchGate, Jul. 2021.
40. V. Chaurasia and S. Pal, "Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer," SN Computer Science, vol. 1, no. 6, pp. 1–6, 2020.
41. Yue W, et al. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018;2(2):13.
42. A. Sharma, et al., "Performance evaluation of machine learning for breast cancer prediction," Data Science for Healthcare, vol. 27, pp. 152–160, 2022.
43. D. Valero-Carreras, J. Alcaraz and M. Landete. Comparing two SVM models through different metrics based on the confusion matrix. Computers & Operations Research, vol. 152, p. 106131, 2023
44. J. Miao and W. Zhu. Precision-recall curve (PRC) classification trees. Evolutionary Intelligence, vol. 15, no. 3, pp. 1545-1569, 2022
45. A. Kovács, P. Palásti, D. Veréb, B. Bozsik, A. Palkó and Z. T. Kincses. The sensitivity and specificity of chest CT in the diagnosis of COVID-19. European Radiology, vol. 31, pp. 2819-2824, 2021.
46. A. S. Assiri, S. Nazir and S. A. Velastin. Breast tumor classification using an ensemble machine learning method. Journal of Imaging, vol. 6, no. 6, p. 39, 2020.

© *"BBIJTM 2024", All Rights Reserved*