ISSN 2723-6471

Breast Cancer Prediction Using Metrics-Based Classification

Sheeba Armoogum[1], Deshinta Arrova Dewi[2,*], Motean Kezhilen[3], Dedi Trinawarman[4]

[1]*Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia*

[2,3]*University of Mauritius, Reduit 80837, Mauritius*

[4]*Informatics Engineering Study Program, Tarumanagara University*

(Received: July 22, 2024; Revised: August 29, 2024; Accepted: September 13, 2024; Available online: September 23, 2024)

Abstract

Breast cancer remains the most prevalent form of cancer among women, with rising mortality rates worldwide. Early detection and accurate classification are crucial for improving patient outcomes, but manual detection methods are often time-consuming, complex, and prone to inaccuracies. This study aims to develop a machine learning (ML)-based desktop application to automate the detection and classification of breast cancer, thereby improving the efficiency and accuracy of diagnosis. Various ML algorithms, including Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, and K-nearest Neighbors, were employed to build classification models. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was used, and pre-processing techniques such as data cleaning, over-sampling, and feature selection were applied to optimize model performance. Experimental results demonstrate that the Random Forest classifier outperformed the other models, achieving an accuracy of 95.54%, precision of 96.72%, recall (sensitivity) of 95.16%, specificity of 96%, and an F1-score of 95.93%. These results highlight the potential of ML techniques in enhancing breast cancer diagnosis by offering a more reliable and efficient classification process. Future work could focus on improving feature selection techniques and applying the model to more diverse datasets for broader applicability.

Keywords: Breast Cancer Detection, Machine Learning Algorithms, Random Forest Classification, Medical Diagnosis Automation, Wisconsin Diagnostic Breast Cancer (WDBC) Dataset, Public Health

1. Introduction

Breast cancer is a multifaceted and complex malignancy, marked by the uncontrolled proliferation of cells within the breast tissue. Over recent decades, it has emerged as one of the most prevalent cancers affecting women worldwide. In 2020, the World Health Organization (WHO) [1] reported approximately 2.3 million new cases of breast cancer globally, resulting in 685,000 deaths. The disease disproportionately affects women, with over 99% of cases occurring in females, while men account for only 0.5-1% of total breast cancer incidences. The susceptibility of women to breast cancer is influenced by various factors, including age, hormonal influences, family history, genetic mutations such as BRCA1 and BRCA2, and lifestyle choices. Moreover, women with denser breast tissue are at higher risk due to greater amounts of glandular and connective tissue, which can obscure mammographic images and delay diagnosis.

In smaller nations like Mauritius, breast cancer remains a significant public health concern. In 2021, 591 out of the 1,681 reported cancer cases were related to breast cancer, and 257 of the 798 recorded cancer deaths were attributed to the disease [2]. These statistics highlight the urgent need for improved early detection methods and diagnostic technologies, as early diagnosis plays a critical role in reducing mortality rates associated with breast cancer.

Traditional diagnostic techniques, such as mammography, ultrasound, thermography, and biopsy, have proven essential in detecting breast cancer. However, manual interpretation of these tests is often limited by subjectivity, complexity, and time constraints. In cases involving dense breast tissue or subtle abnormalities, human error may lead to misdiagnosis or delayed diagnosis. To address these limitations, machine learning (ML) has emerged as a cutting-edge solution in the medical field, revolutionizing the way breast cancer is diagnosed. ML models have the ability to process vast amounts of data and detect intricate patterns that may be missed by human analysis, aiding medical professionals in making accurate, early diagnoses.

***Corresponding author: Deshinta Arrova Dewi (deshinta.ad@newinti.edu.my)**
DOI: https://doi.org/10.47738/jads.v5i3.351

ISSN 2723-6471

Despite the progress made in ML-assisted breast cancer diagnostics, several gaps remain in the existing research. While many studies have explored the application of ML algorithms, such as support vector machines, decision trees, and neural networks, for breast cancer classification, there is still limited understanding of which specific algorithms consistently offer superior performance across different datasets. Additionally, most of the current models focus on optimizing accuracy without considering the computational complexity or interpretability of the algorithms, which are crucial for real-world clinical applications. There is also a lack of comprehensive comparison studies that evaluate multiple ML techniques on standardized datasets like the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to determine the most effective approach for reliable classification.

Recent advancements in machine learning have introduced several state-of-the-art models that aim to address the complexities of breast cancer diagnosis. Deep learning models, particularly convolutional neural networks (CNNs), have shown impressive results in image-based diagnosis tasks such as mammogram analysis. Additionally, ensemble methods like Random Forest and Gradient Boosting have been employed for feature-based classification, yielding high accuracy in distinguishing between benign and malignant tumors. Furthermore, hybrid approaches that combine traditional diagnostic techniques with ML algorithms are being developed to improve diagnostic accuracy and reduce false positives and negatives. However, many of these models face challenges related to generalizability, model interpretability, and computational cost, which limit their integration into clinical practice.

Given these gaps, the primary objective of this study is to build and compare multiple machine learning models for classifying breast cancer as either benign or malignant using the WDBC dataset. The study aims to evaluate several ML algorithms in terms of accuracy, computational efficiency, and model interpretability, addressing both the performance and practical applicability in real-world medical settings. By identifying the most effective model, this research seeks to contribute to the development of more reliable, efficient, and interpretable ML-based diagnostic tools for breast cancer.

The structure of the paper is as follows: Section II provides an extensive review of the literature, detailing previous research on ML applications for breast cancer diagnosis. Section III outlines the methodology used, including the selection of ML algorithms and evaluation metrics. Section IV presents the results, followed by the conclusions and future research directions in Section V.

2. Literature Review

Several studies have focused on the classification and prediction of breast cancer using machine learning (ML) techniques, reflecting the growing importance of computational approaches in medical diagnostics. As research in this field advances, a wide range of ML models have been applied to detect and classify breast cancer, yielding significant improvements in accuracy and efficiency.

Chawan et al. [3] introduced the use of supervised learning algorithms, including Decision Tree, Random Forest, and Logistic Regression, combined with Dimensionality Reduction techniques. Their study achieved accuracy rates of 95.8%, 98.6%, and 95.8% respectively, demonstrating the strong performance of Random Forest in particular. Similarly, another study [4] compared the performance of four classifiers—Multilayer Perceptron (MLP), Support Vector Machine (SVM), k-Nearest Neighbours (KNN), and Random Forest—on the Wisconsin Breast Cancer dataset. Among these, Random Forest exhibited the highest accuracy at 99.26%, followed by SVM and KNN, which achieved 97.78% and 97.04% respectively, with MLP showing the lowest accuracy at 94.07%.

Bazazeh and Shubair [5] further explored breast cancer classification by comparing the performance of SVM, Random Forest, and Bayesian Networks (BN). Their findings showed that Random Forest excelled in terms of Receiver Operating Characteristic (ROC) performance, indicating a higher capability to distinguish between malignant and benign cases. Hasan et al. [6], in a novel approach, employed two datasets: the WDBC and the SEER 2017 Breast Cancer Dataset. Using Principal Component Analysis (PCA) for feature extraction, they applied MLP and Convolutional Neural Networks (CNN) for classification. The MLP model achieved a remarkable 99.1% accuracy on the reduced WDBC dataset and 89.3% on the SEER dataset, while the CNN model reached 96.4% and 88.3% accuracy on these datasets, respectively.

Similarly, another study [7] utilized the WDBC dataset to address breast cancer diagnosis by employing a combination of SVM and Artificial Neural Networks (ANN) with feature selection. Their analysis revealed that ANN outperformed SVM, with an accuracy of 99% compared to SVM's 98%. Poornajaf and Yousefi [8] also examined the effectiveness of ML models, applying Logistic Regression and Extra Trees algorithms to multidimensional datasets. Both models achieved an accuracy of 99.14%, though their AUC ROC scores differed slightly at 99.6% and 99.1%, respectively.

In their research, Kiliç and Karakoyun [9] employed various ML algorithms alongside data quality enhancement techniques on the WDBC dataset, finding that KNN delivered the best performance, with a 99.3% accuracy, 98.9% precision, 100% recall, and a 99.4% F1-score. Chen et al. [10] focused on the use of XGBoost, Random Forest, Logistic Regression, and KNN, with recall as the primary evaluation metric. After applying processes such as data standardization, feature selection, and addressing class imbalance issues, XGBoost was found to outperform the other algorithms [11].

Mangukiya et al. [12] assessed the performance of several ML algorithms—SVM, Decision Tree, Naïve Bayes, KNN, AdaBoost, XGBoost, and Random Forest—on the WDBC dataset, using accuracy, precision, sensitivity, and specificity as evaluation metrics. XGBoost achieved the highest accuracy at 98.24%. In a similar vein, Sakib et al. [13] compared the performance of five ML algorithms (SVM, Decision Tree, Logistic Regression, Random Forest, and KNN) and a deep learning (DL) technique. Their results indicated that Random Forest was the best performer, achieving a 97.37% accuracy with default parameters and a 96.66% accuracy with tuned parameters. They also reported that Random Forest demonstrated the highest cross-validation accuracy at 96.84%, suggesting strong generalization to new datasets. They concluded that parameter tuning provided only slight improvements in performance.

Fatih Ak [14], in contrast to other studies, incorporated data visualization along with ML for breast cancer diagnosis. From the WDBC dataset, three distinct datasets were extracted, featuring independent, highly correlated, and low-correlated features. Logistic Regression was found to provide the best classification accuracy at 98.1%. Lastly, Hussain et al. [15] employed the WEKA tool to evaluate several algorithms, including Naïve Bayes, KNN, Decision Tree, Random Forest, SVM, and Logistic Regression. Their results showed that KNN and Random Forest performed better than other models in terms of accuracy, recall, and overall performance metrics.

3. Research Methodology

This section will provide further details on the methodology used to incorporate the use of Machine Learning for the early diagnosis of breast cancer. A flowchart is provided in figure 1, outlining the pipeline of the research work. Each process will be elaborated in subsections below.

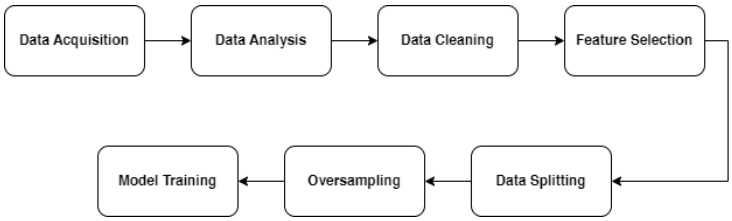

Figure 1. Metric-based Classification pipeline.

3.1. Dataset Description

In this research work, the Wisconsin Diagnostic Breast Cancer dataset (WDBC) is used and was normally provided by Dr. William of the Clinical Medicine Research Institute of the University of Wisconsin [11]. Features that are computed from digitized images of fine needle aspirates (FNA) of a breast mass describe the cell nuclei characteristics. Among the 569 experimental samples of the dataset, 212 are malignant cases while 357 refers to benign ones. Each sample is associated with 10 nucleus features, including radius, perimeter, smoothness, area, compactness, concavity, symmetry, texture, concave points, and fractal dimension. Furthermore, for each image, these features are calculated as the mean, standard and worst, thus resulting in a total of 30 features. The classification label (benign/malignant) is also provided for each sample as shown in figure 2.


```
id                        smoothness_se
diagnosis                 compactness_se
radius_mean               concavity_se
texture_mean              concave points_se
perimeter_mean            symmetry_se
area_mean                 fractal_dimension_se
smoothness_mean           radius_worst
compactness_mean          texture_worst
concavity_mean            perimeter_worst
concave points_mean       area_worst
symmetry_mean             smoothness_worst
fractal_dimension_mean    compactness_worst
radius_se                 concavity_worst
texture_se                concave points_worst
perimeter_se              symmetry_worst
area_se                   fractal_dimension_worst
```

Figure 2. WDBC independent and target variables.

3.2. Dataset Analysis and Preprocessing

This section presents a comprehensive analysis of the dataset to better understand its structure and ensure the quality of the data. Initially, unnecessary columns are removed through data cleaning to enhance the dataset's reliability and clarity [16]. Once the data is cleaned, visualization techniques are applied to explore trends, patterns, and relationships within the dataset. First, the distribution of the target variable is analyzed to check for any potential imbalance, as illustrated in figure 3. This step is crucial to ensure that the data is balanced, as an imbalance could affect the performance of the machine learning models.

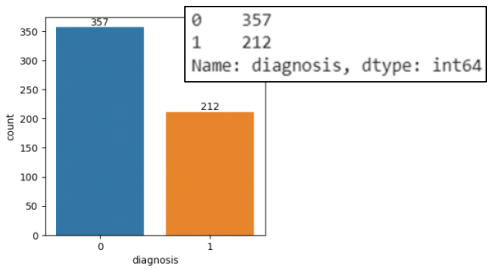

Figure 3. Class Distribution of target variable.

Next, the distribution of values within each feature is examined, providing insights into the spread and variance of the data in different columns, which is visualized in figure 4. This exploration helps to identify any outliers or irregularities in the data that may need further attention.

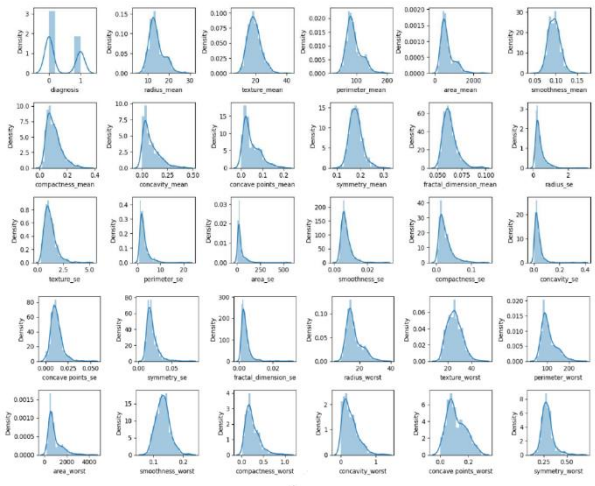

Figure 4. Spread of variables.

 ISSN 2723-6471

Additionally, potential relationships between variables are identified using a correlation matrix, as shown in figure 5. This matrix helps to reveal any significant correlations between features, which could inform feature selection or further preprocessing steps.

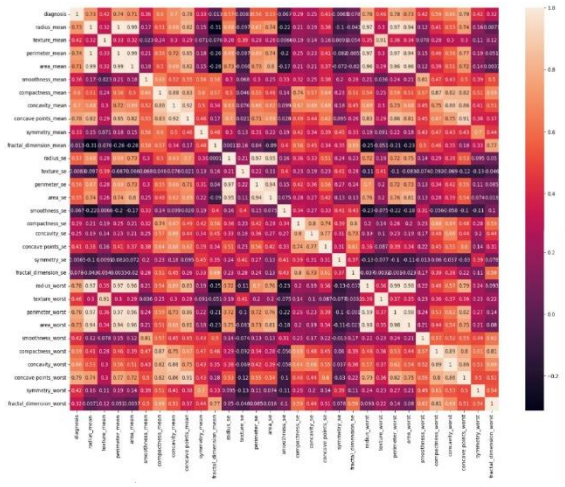

Figure 5. Correlation matrix.

Finally, label encoding is performed, with benign samples assigned a value of 0 and malignant samples a value of 1. This step prepares the target variable for machine learning algorithms that require numerical input, ensuring consistency in the dataset for model training.

3.2.1. Feature Selection

When a model is built using data features, the model tries to learn from them which means that features that are irrelevant or partially irrelevant can affect the performance of the model. Ergo, feature selection is used to select the relevant features thus eliminating data ambiguity and complexity [17]. The most common types of feature selection methods are wrapper methods, filter methods and embedded methods. However, filter methods will be used due to its algorithm independence. The selected features are shown in figure 6.

```
Index([
'radius_mean', 'perimeter_mean', 'area_mean', 'concavity_mean',
'concave points_mean', 'radius_worst', 'perimeter_worst', 'area_worst',
'concavity_worst', 'concave points_worst'], dtype='object')
```

Figure 6. Columns (Target and Independent Variables) used for training.

3.3. Data-Splitting

Before initiating the data split for training and testing, five samples from each category (benign and malignant) were set aside for independent testing within the desktop application, ensuring that these samples would not influence the training process and allowing for an unbiased evaluation. Consequently, the training dataset now consists of 559 rows, slightly reduced from the original dataset [18]. In this project, 80% of the available data is allocated for training, while 20% is reserved for testing, providing a sufficient balance between training the model and evaluating its generalization capability. This careful separation of the dataset is critical to avoid overfitting and to ensure that the model's performance is tested on previously unseen data, offering a reliable measure of its effectiveness. The data composition for both training and testing is illustrated in figure 7, emphasizing the methodical approach taken to maintain the integrity and reliability of the evaluation process.

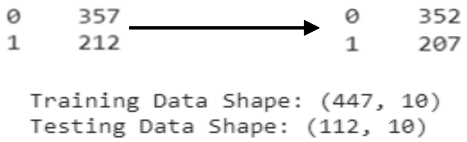

Figure 7. Split Data Composition

 ISSN 2723-6471

3.4. Over-Sampling

To address the issue of class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to generate synthetic samples for the minority class (malignant). Class imbalance, where benign cases significantly outnumber malignant ones, can lead to biased models that underperform on detecting malignant samples [19]. SMOTE enhances the training dataset by interpolating between existing malignant samples and their nearest neighbors, creating new synthetic instances that help the model better learn the characteristics of the minority class. Importantly, over-sampling is applied only to the training data, as applying it to the entire dataset, including the test set, could introduce bias and artificially inflate the model's performance. By restricting over-sampling to the training set, the test data remains reflective of real-world conditions, allowing for a fair evaluation of the model's ability to generalize to unseen cases. Figure 8 illustrates the effect of SMOTE in balancing the training dataset and resolving the initial class imbalance issue.

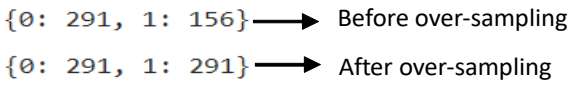

Figure 8. Over-samppling example

3.5. Model Training

Following the over-sampling process, the data is input into Machine Learning algorithms for training. The study employs six distinct Machine Learning algorithms: Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, and K-nearest Neighbour. Post-training, a comparative analysis will be conducted to evaluate the performance of each ML algorithm. Figure 9 shows the pipeline of the training process.

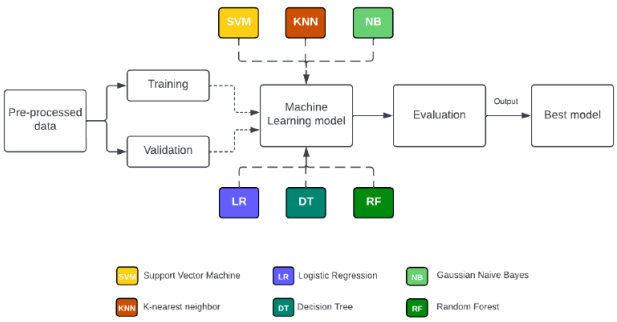

Figure 9. Model-Training Pipeline.

3.6. Evaluation Metrics

In this study, the evaluation criteria include accuracy, precision, recall, and F1 score metrics, calculated using equations. (1), (2), (3), and (4) respectively [20]. The subsequent equations depict the metric calculations based on the confusion matrix extracted in table 1.

Table 1. Confusion-Matrix

Predicted	**Benign**	**TP**	**FP**
	Malignant	FN	TN
		Benign	Malignant
		Actual	

$$Accuracy = TP+TN \,/\, TP+FP+TN+FN \quad (1)$$

$$Precision = TP \,/\, TP+FP \quad (2)$$

$$Recall = TP \,/\, TP+FN \quad (3)$$

$$f1 - score = 2 * (precision*recall \,/\, precision+recall) \quad (4)$$

 ISSN 2723-6471

A Receiver Operating Characteristics (ROC) graph serves as a visual representation of a classifier's performance by plotting the model's true positive rate against its false positive rate. Normally, The Area under the ROC graph quantifies the classifier's performance, calculated by dividing the area under the plot by the total graph area and values closer to 1 indicate higher classifier performance.

4. Results and Discussion

This section provides a detailed analysis of the performance of six distinct classifier models, each evaluated based on key statistical metrics derived from their respective confusion matrices, as depicted in figure 10. These metrics include accuracy, precision, recall, and F1 score, which collectively offer a comprehensive understanding of the models' performance in predicting breast cancer cases.

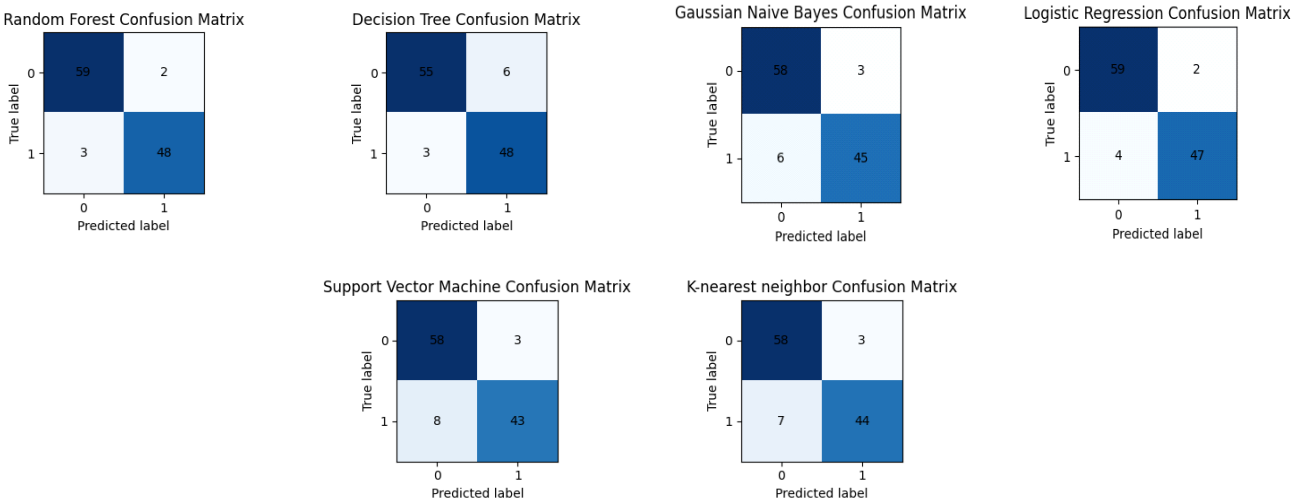

Figure 10. Confusion Matrix of 6 classifiers

Table 2 provides a detailed summary of the key performance metrics—accuracy, precision, recall, and F1 score—calculated from the confusion matrices of the six classifier models. These metrics offer a more nuanced understanding of each model's performance, going beyond accuracy to include how well each model handles the minority (malignant) class. Precision indicates how many of the predicted malignant cases are correct, while recall measures the model's ability to identify all actual malignant cases, both of which are crucial in medical diagnostics where minimizing false positives and negatives is important. The F1 score combines precision and recall, offering a balanced view of the model's overall effectiveness. To complement this, figure 11 visually represents these metrics, allowing for a clearer and more intuitive comparison of the classifiers' strengths and weaknesses. This visualization helps to highlight areas where certain models, such as Random Forest or Logistic Regression, perform better in terms of precision or recall, making it easier to select the most appropriate model based on the specific goals of the classification task.

Table 2. Results

Algorithm	Accuracy	Precision	Recall	f-1 score
Random Forest	0.9554	0.9672	0.9516	0.9593
Decision Tree	0.9464	0.9344	0.9661	0.95
Logistic Regression	0.9464	0.9672	0.9365	0.9516
Naïve Bayes	0.9196	0.9508	0.9063	0.928
SVM	0.9018	0.9508	0.8788	0.9134

As shown in figure 11, the Random Forest model demonstrated the strongest performance with the highest accuracy (95.54%) and a well-balanced combination of precision (96.72%) and recall (95.16%). Its robust performance across multiple evaluation metrics indicates that it is highly effective at both correctly identifying malignant cases (high recall) and maintaining precision, which reduces false positives.

 ISSN 2723-6471

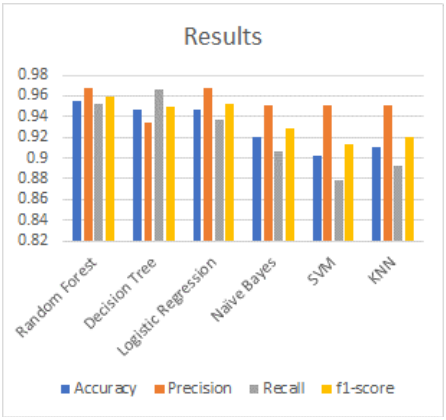

Figure 11. Results of all the classifiers.

Decision Tree and Logistic Regression also performed admirably, each achieving an accuracy of 94.64%. Both models displayed a good balance between precision and recall, with the Decision Tree excelling slightly in recall (96.61%) and Logistic Regression showing higher precision (96.72%). These results suggest that both models are reliable alternatives, depending on the specific priorities of the classification task. The Naïve Bayes and SVM models exhibited more moderate performance. Naïve Bayes prioritized precision (95.08%) but had a lower recall (90.63%), making it better suited for tasks where minimizing false positives is critical. Similarly, SVM focused on precision (95.08%) but had a slightly reduced recall (87.88%), indicating that while it is strong at predicting benign cases, it may underperform in correctly identifying malignant ones.

K-Nearest Neighbors (KNN) was also included in the initial analysis, though its results are not shown in the table, it demonstrated reasonable performance but did not outperform the other models. In choosing the most suitable model for breast cancer classification, it is essential to align the model's strengths with the specific objectives of the task. Given its superior performance across all metrics, Random Forest stands out as a strong candidate, offering a high level of accuracy and a good trade-off between precision and recall. This makes it highly effective in both identifying malignant cases and minimizing false positives, which is critical in medical diagnostics.

To further evaluate the models, a Receiver Operating Characteristic (ROC) curve was generated, as shown in figure 12. The ROC curve provides insight into each model's ability to balance the true positive rate (TPR) against the false positive rate (FPR). The Area Under the Curve (AUC) metric, derived from the ROC curve, serves as a significant indicator of overall model performance, with a higher AUC value (closer to 1) signifying better discrimination between the positive and negative classes.

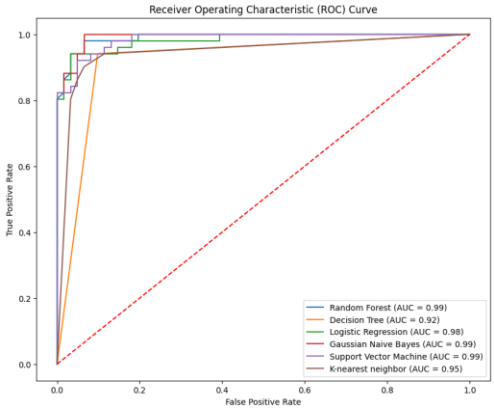

Figure 12. ROC Curve of classifiers.

As illustrated by the ROC curves in figure 12, the Random Forest classifier outperformed the other models, showcasing superior predictive ability. Its Area Under the Curve (AUC) score, which approaches 1, reflects its strong capacity to

 ISSN 2723-6471

balance true positive and false positive rates. This high AUC value, along with its consistently robust performance across various evaluation metrics, confirms that Random Forest is the most reliable model for this classification task.

The classification report in figure 13 provides a more granular evaluation of the Random Forest model's performance. The high precision, recall, and F1 scores across both the benign and malignant classes demonstrate the model's accuracy in distinguishing between the two categories. Specifically, the high precision reflects the model's ability to minimize false positives, while the high recall indicates its success in reducing false negatives. These combined metrics suggest that the Random Forest model is both reliable and effective in identifying breast cancer cases with a minimal margin for error, making it an excellent tool for medical diagnostics where accuracy is crucial.

	precision	recall	f1-score	support
0	0.95	0.97	0.96	61
1	0.96	0.94	0.95	51
accuracy			0.96	112
macro avg	0.96	0.95	0.95	112
weighted avg	0.96	0.96	0.96	112

Figure 13. Classification Report of Random Forest.

Once the model was confirmed as suitable for deployment, it was downloaded for local use in the desktop application. In practical application, when a user inputs the necessary features or metrics for classification, the pre-trained Random Forest model will be loaded and employed to classify the input as either benign or malignant. This integration provides a seamless and efficient workflow for breast cancer diagnosis, as illustrated in the simple pipeline shown in figure 14.

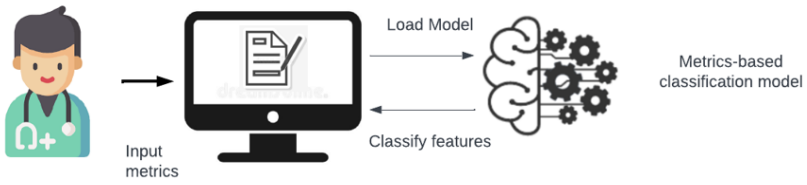

Figure 14. Classification Process.

The desktop application was developed entirely using Python, utilizing its powerful libraries for both machine learning and graphical user interface (GUI) development. Python's flexibility made it an ideal choice, as it allows seamless integration of machine learning models with GUI frameworks like Tkinter for creating a user-friendly interface. The application relies on libraries such as pandas, scikit-learn, and numpy for data processing and loading the pre-trained Random Forest model. Designed with simplicity in mind, the interface, shown in figure 15, enables users to input necessary diagnostic features and receive immediate classification results, predicting whether a case is benign or malignant. This application not only provides fast and accurate results but also ensures data security by being deployed locally, making it a practical and accessible tool for medical professionals in clinical settings.

 ISSN 2723-6471

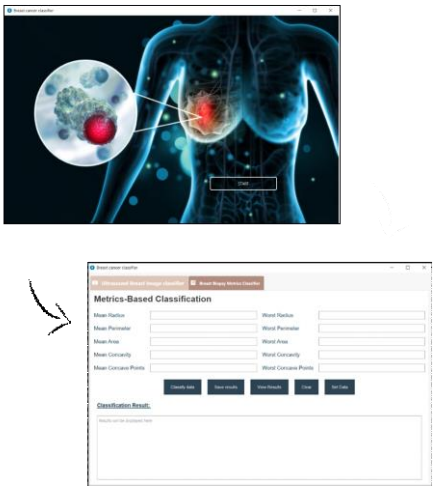

Figure 15. App UI.

The final evaluation step involved thoroughly assessing the model's performance after integrating it into the desktop application. Prior to training, five rows of data from each class (benign and malignant) were set aside from the original dataset to ensure an unbiased evaluation of the model's accuracy. Upon testing, the model exhibited excellent performance, accurately classifying all 10 instances, resulting in a 100% accuracy for this specific subset of data. However, when evaluated on the full test dataset consisting of 112 rows, the model did misclassify a few instances, as highlighted by the confusion matrix in figure 10. These misclassifications suggest that while the model performs well, there is still room for improvement in certain areas, particularly in refining its ability to handle more complex or borderline cases. An example of the classification process, as implemented in the desktop application, is shown in figure 16, demonstrating the practical application of the model in a real-world setting.

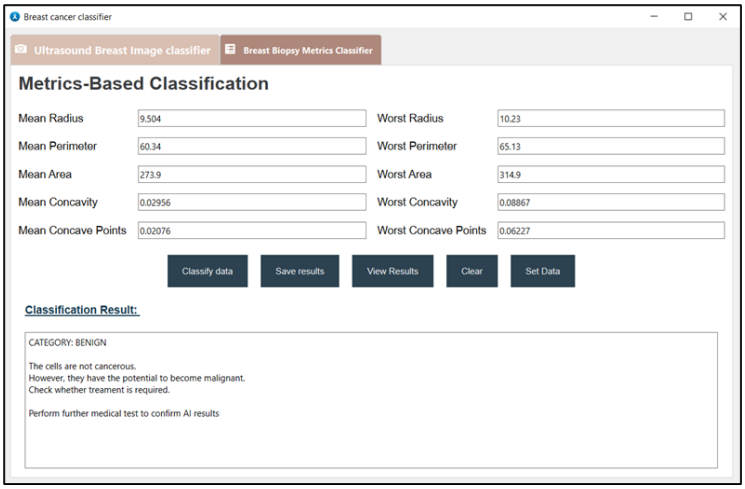

Figure 16. Example of metric-based classification.

5. Conclusion

In this research, the objective was to enhance the detection and classification of breast cancer using machine learning algorithms, aiming to improve the accuracy and efficiency of early diagnosis. Several classifiers, including Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, and K-nearest Neighbors, were applied to the Wisconsin Diagnostic Breast Cancer dataset. Through pre-processing steps, including over-sampling and feature selection, the models were trained and evaluated to determine their effectiveness. The results revealed that the Random Forest classifier outperformed the other models, achieving an accuracy of 95.54%, precision of 96.72%, recall of 95.16%, specificity of 96%, and an F1-score of 95.93%, showcasing its robustness in correctly classifying benign and malignant cases. These findings underscore the potential of Random Forest as an effective tool for breast cancer diagnosis, offering a balance between precision and recall that is critical in medical applications were

minimizing both false positives and false negatives is essential. The integration of machine learning can significantly reduce diagnostic time and cost, making it an impactful solution in healthcare.

However, the study has certain limitations. The dataset, though widely used, could benefit from more diverse data sources to generalize the model's effectiveness across different populations. Additionally, while over-sampling improved performance, it can introduce synthetic bias, and further exploration of advanced techniques could mitigate this. Future research should focus on enhancing feature selection techniques by incorporating wrapper or embedded methods, which may improve model accuracy and reduce computational overhead. Furthermore, expanding the dataset and testing models on more varied data would increase generalizability and robustness. This study demonstrates the effectiveness of machine learning, particularly Random Forest, in classifying breast cancer cases. The promising results suggest that continued exploration of more sophisticated techniques and larger datasets could further improve early detection, ultimately contributing to better healthcare outcomes.

6. Declarations

6.1. Author Contributions

Conceptualization: D.A.D., S.A., M.K., and D.T.; Methodology: S.A.; Software: D.A.D. and D.T.; Validation: D.A.D., S.A., and M.K.; Formal Analysis: D.A.D. and M.K.; Investigation: D.A.D.; Resources: S.A.; Data Curation: S.A. and D.T.; Writing Original Draft Preparation: D.A.D. and M.K.; Writing Review and Editing: S.A. and D.A.D.; Visualization: D.A.D.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K. W. Goh, A. Stephen, Y. S. Wu, M. S. Sim, K. Batumalaie, S. C. B. Gopinath, R. M. Guad, A. Kumar, M. Sekar, V. Subramaniyan, N. K. Fuloria, S. Fuloria, A. Velaga, Md. M. R. Sarker, "Molecular targets of aptamers in gastrointestinal cancers: Cancer detection, therapeutic applications, and associated mechanisms," *Journal of Cancer*, vol. 14, no. 13, pp. 2491–2516, 2023. doi:10.7150/jca.85260

[2] Mauritius National Cancer Registry, "Ministry of Health and Wellness African Cancer Registry Network World Health Organization CANCER IN THE REPUBLIC OF MAURITIUS Incidence and Mortality Study for 2021 Report of the National Cancer Registry," 2023.

[3] P. Chawan, Z. Thakkar, and M. Jadhav, "Breast cancer prediction using supervised machine learning algorithms," *International Research Journal of Engineering and Technology*, vol. 7, no. 4, pp. 1020–1026, Apr. 2020.

[4] H. N. Iqbal, A. B. Nassif, and I. Shahin, "Classifications of breast cancer diagnosis using machine learning," *International Journal of Computers*, vol. 14, no. 1, pp. 86–92, 2020.

[5] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," *in 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), IEEE*, vol. 5, no. 1, pp. 1–4, 2016.

[6] Md. M. Hasan, Md. R. Haque, and M. Md. J. Kabir, "Breast cancer diagnosis models using PCA and different neural network architectures," *in 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, vol. 2019, no. Jul., pp. 1–6, 2019.

[7] M. Divyavani and G. Kalpana, "An analysis on SVM and ANN using breast cancer dataset," *Aegaeum Journal*, vol. 2021, no. 1, pp. 369–379, 2021.

[8] M. Poornajaf and S. Yousefi, "Improvement of the performance of machine learning algorithms in predicting breast cancer," *Frontiers in Health Informatics*, vol. 12, no. 1, pp. 132–139, 2023.

[9] A. Kiliç and M. Karakoyun, "Breast cancer detection using machine learning algorithms," *International Journal of Advanced Natural Sciences and Engineering Researches*, vol. 2023, no. 3, pp. 91–95, Mar. 2023.

[10] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification prediction of breast cancer based on machine learning," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, pp. 1–12, Jan. 2023.

[11] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," *University of Wisconsin-Madison Department of Computer Sciences*, vol. 1990, no. 1, pp. 1–10, 1990.

[12] M. Mangukiya, A. Vaghani, and M. Savani, "Breast cancer detection with machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 2, pp. 141–145, Feb. 2022.

[13] S. Sakib, N. Yasmin, A. K. Tanzeem, F. Shorna, K. Md. Hasib, and S. B. Alam, "Breast cancer detection and classification: A comparative analysis using machine learning algorithms," *in Proceedings of the Third International Conference on Communication, Computing and Electronics Systems*, vol. 2022, no. 1, pp. 156–162, 2022.

[14] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare*, vol. 8, no. 2, p. 111, Apr. 2020.

[15] M. Hussain, U. I. Haq, M. Azeem, and S. Bashir, "Breast cancer detection: A comparative study using machine learning models," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 8, pp. 234–241, 2023.

[16] D. Sugianto and A. R. Hananto, "Geospatial Analysis of Virtual Property Prices Distributions and Clustering," *Int. J. Res. Metav.*, vol. 1, no. 2, pp. 127-141, 2024.

[17] B. H. Hayadi and I. M. M. El Emary, "Enhancing Security and Efficiency in Decentralized Smart Applications through Blockchain Machine Learning Integration", *J. Curr. Res. Blockchain.*, vol. 1, no. 2, pp. 139–154, Sep. 2024.

[18] A. R. Yadulla, G. S. Nadella, M. H. Maturi, H. Gonaygunta, "Evaluating Behavioral Intention and Financial Stability in Cryptocurrency Exchange App: Analyzing System Quality, Perceived Trust, and Digital Currency in Indonesia," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 2, pp. 103-124, 2024

[19] A. Imakura, M. Kihira, Y. Okada, and T. Sakurai, "Another use of smote for interpretable data collaboration analysis," *Expert Systems with Applications*, vol. 228, no. Oct., pp. 120385–120398, Oct. 2023. doi:10.1016/j.eswa.2023.120385

[20] Y. Wang, Y. Jia, Y. Tian, and J. Xiao, "Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring," *Expert Systems with Applications*, vol. 200, no. Aug., pp. 117013–117025, Aug. 2022. doi:10.1016/j.eswa.2022.117013