Journal of Information Systems Engineering and Management
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Innovative Machine Learning Framework for Mammographic Breast Cancer Detection

[1]Shruthi B S and [2]Ramesh Sekaran
[1]*Research Scholar, Department of CSE, JAIN (Deemed- to-be- University), Bangalore*
shruthi.sundarraj@gmail.com
[2]*Professor, Department of CSE JAIN (Deemed- to-be- University), Bangalore*
sramsaran1989@gmail.com

ARTICLE INFO

Received: 05 Oct 2024

Revised: 05 Dec 2024

Accepted: 22 Dec 2024

ABSTRACT

Breast cancer remains a major global health challenge, with early and accurate detection being critical for improving patient outcomes. However, traditional mammogram-based diagnostic approaches often face limitations such as high noise levels, low feature resolution, and suboptimal classification accuracy. This study addresses these gaps by introducing an innovative framework that integrates advanced preprocessing, feature extraction, and machine learning classification techniques. The preprocessing phase employs the Mean Error Splash Filter, specifically designed for mammographic images, to reduce noise while preserving diagnostic features. To bridge the gap in adaptive feature selection, the Tech Bee algorithm extracts critical features such as texture, edges, and region properties, prioritizing those with high diagnostic relevance. Using a stratified dataset, a Gradient Vector Boosting Classifier is applied for robust classification, capable of handling nonlinear relationships and imbalanced datasets. The proposed methodology achieved high accuracy, sensitivity, and specificity, outperforming traditional methods and offering a significant advancement in breast cancer prediction. By addressing current challenges in noise reduction, feature extraction, and classification, this study provides a scalable and efficient tool for early breast cancer detection and paves the way for improved diagnostic interventions.

Keywords: Breast Cancer Prediction, Mammogram Imaging, Tech Bee Algorithm, Machine Learning, Gradient Vector Boosting Classifier

I. INTRODUCTION

Breast cancer is a significant public health issue due to its prevalence as a malignancy and its impact on women's lives around the globe. Early detection is crucial for treatment effectiveness and improving survival rates [1]. Mammography is the gold standard for breast cancer screening imaging because it can detect abnormalities even in densely populated areas. Issues include picture noise, low tissue contrast, and subtle feature detection can lead to false positives and negatives. Patients experience additional emotional and financial hardship as a result of delays in treatment brought on by doctors' inability to respond promptly due to these incorrect diagnoses [2]. The fast developing field of machine learning and artificial intelligence (AI) offers a potential solution to the problem of inaccurate breast cancer diagnoses [3]. By automating the interpretation of mammographic images, machine learning models can improve diagnostic accuracy and decrease the risk of human error. However, existing methods frequently encounter issues, such as inadequate noise handling in pre-processing, feature extraction algorithms failing to capture diagnostically significant features, and classifiers failing to perform adequately on complex or imbalanced datasets [4]. In order to overcome these limitations, a thorough approach is required that integrates trustworthy feature extraction, efficient classification algorithms, and high-quality pre-processing. The primary objectives of this research are as follows: • Improving the quality of mammography images by reducing noise through the development of a new pre-processing approach [5] (Mean Error Splash Filter).

• Building the Tech Bee approach, an adaptive feature extraction technique with low dimensions, to extract

diagnostically valuable attributes. To reliably and accurately classify mammography photos using a Gradient Vector Boosting Classifier [6].

• Evaluate the suggested framework's accuracy, sensitivity, specificity, and computing efficiency in comparison to current approaches[7].

This article is organised in the following way: With a focus on the limitations of existing approaches[8],

Section 2 presents a discussion of pertinent work in breast cancer prediction.

The Tech Bee Algorithm, Mean Error Splash Filter, and Gradient Vector Boosting Classifier are some of the methodology aspects discussed in Section 3[9].

In Section 4, we examine the results and evaluate the proposed framework in relation to alternative approaches.

In Section 5, we review the study's key points, including its results, implications, and recommendations for further research.

II. RELATED WORKS

Machine learning is finding more and more applications in health care every day. These advancements are helpful to scientific study[10], and there is a lot of research on this issue. We have found a large number of research publications that are pertinent to our inquiry. Finding a way to forecast the occurrence of breast cancer is the driving force behind this endeavour[11]. The bulk of the material originated from the Dhaka Medical College Hospital. Throughout our inquiry, we came across several fresh approaches. This idea would be developed further in the next chapter, but our work was not simple. Reviewing prior studies on the prediction of heart attacks allowed us to properly execute this study and acquire this new word[12].

With their model, [13] discovered the most effective machine learning techniques for breast cancer prediction. Support Vector Machine (SVM), Naive Bayes (NB), Radial Basis Function Neural Networks (RBF NN), Decision Tree (DT), and a condensed form of Classification and Regression Trees (CART) were among the techniques they employed in their analysis [14] . Their successful model was implemented using a Support Vector Machine, and they achieved the highest Area Under the Curve (AUC) of 96.84 percent on the original Wisconsin Breast Cancer datasets. Djebbari et al. investigated the viability of employing a machine learning ensemble to forecast how long a patient with breast cancer would live [15]. In their dataset on breast cancer, their approach performed more accurately than previous studies. S. Aruna and L. Nandakishore studied the classification of white blood cells (WBCs) [16]. K-Nearest Neighbours (K-NN), Decision Tree, Support Vector Machine, and Naive Bayes were all taken into consideration. The area under the curve (AUC) of the top Support Vector Machine classifier they employed is 96.99%.

In order to categorise tumour cells, M. Angrap used six ML methods. One variation of the extended short-term memory neural network that was created and deployed was the Gated Recurrent Unit (GRU). The neural network's SoftMax layer was replaced with a Support Vector Machine layer. The GRU Support Vector Machine achieved a perfect score of 99.04% in that study [17]. To improve the accuracy of a model trained using a neural network and association rules to 95.6%,[18] used cross-validation. In order to apply Naive Bayes classifiers, a new weight adjustment approach was used.
Ensemble learning as a tool for cancer recurrence prediction was explored by [19]. Using a relevance vector as input [20] compared and contrasted three ML models that performed very well [21] used a radial basis function network (RBFN) among several data reduction and preprocessing approaches to achieve their aims.

A number of breast cancer research were used to create survival prediction models, as reported in [22]. Breast cancer tumours, both benign and malignant, were subjected to the survival prediction algorithms used in this study. There has been a lot of study on using machine learning algorithms for breast cancer diagnosis, as seen in [23] They reasoned that data augmentation tactics may fix the problem of inadequate data. In [24], the authors showed how to use the characteristics of computer-aided mammography pictures to autonomously recognise and characterise cell structure. There have been extensive evaluations of clustering and classification methods, as described in [25].

Data visualisation and machine learning were compared in a study by [26] to identify and diagnose breast cancer. Dr. William H. Walberg's breast cancer data was analysed using a wide variety of methods, including Logistic Regression (LR), closest neighbour (NN), Support Vector Machine (SVM), basic Bayes, Decision Tree, random forest (RF), and convolutional forest. These methods were applied using Python, Minitab, and R. Logistic regression

utilising all features produced the best results, with an accuracy of 98.1%. Their findings opened up new possibilities for cancer detection by demonstrating the benefits of data visualisation and machine learning. Based on the Wisconsin Breast Cancer Database, [27] compared five supervised machine learning algorithms for predicting breast cancer cases. This set of procedures included ANN, Logistic Regression, Support Vector Machine, Random Forest, and Nearest Neighbour. With an F1 score of 0.9890, a precision of 97.22%, and an accuracy of 98.57%, ANNs outperformed competing models. Machine learning for illness identification, according to the researchers, might provide doctors with reliable answers quickly, which would cut down on deaths.

In their study, [28] used the Wisconsin Diagnostic Breast Cancer Database to make machine learning predictions about breast cancer. They used statistical approaches to limit the number of characteristics to twelve, compared six algorithms, and then used ensemble methods to merge models. The findings show that all algorithms worked as expected, with the modified feature section in particular achieving a test accuracy of over 90%. They improved the accuracy of breast cancer predictions by using ensemble approaches and feature selection, among other things[13].

Using the methods of Random Forest and Extreme Gradient Boosting (XGBoost) [29] created a model for predicting the likelihood of breast cancer. We used data that was taken from the UCI Machine Learning Repository. Using XGBoost and Random Forest techniques, the model achieved a classification accuracy of 74.73%.Using a Bayesian Network and the Radial Basis function, [30] presented a new ensemble approach to breast cancer data categorisation. The 97% accuracy rate achieved by this process was higher than that of existing methods [31]. A wide variety of indicators were used on the Wisconsin Breast Cancer Dataset (WBCD) to assess the trial's effectiveness. Potentially aiding cancer professionals in making accurate tumour diagnoses and patients in making treatment decisions, the suggested ensemble study is worth exploring.
In order to predict the occurrence of breast cancer, [32] used many machine learning algorithms. The team used the UCI machine learning database in addition to artificial neural networks, decision trees, support vector machines, and naive bayes algorithms. As a result, 86% accuracy in categorisation was found.

For the purpose of breast cancer diagnosis and prediction, made use of machine learning techniques. Support Vector Machine, Random Forest, Logistic Regression, Decision Tree (C4.5), and KNN were the five algorithms that were compared using the Wisconsin Breast Cancer Diagnostic Database. The main goal was to find the best algorithm for breast cancer diagnosis. According to the findings, the support vector machine achieved the highest possible accuracy of 97.2%, much surpassing the performance of the other classifiers. Investigating breast cancer treatment advancements and patient safety requirements in the Anaconda Python environment with the Scikit learning package yields crucial information.

[33] investigated breast cancer prognosis using deep learning and six supervised machine learning methods. To improve the study's accuracy, it included a parametric analysis of all algorithms. While they did not provide the dataset specifically, deep learning using Human Gradient Descent Learning was shown to be the most accurate approach, with an accuracy rate of 98.24%. Using the right hyper parameter machine learning methods might help in tumour detection, according to the study.

III. PROPOSED WORK

The methodology for this study is designed to create an efficient framework for breast cancer prediction, incorporating preprocessing, feature extraction, and machine learning classification techniques. The approach consists of three primary phases: preprocessing using the **Mean Error Splash Filter**, feature extraction with the **Tech Bee Algorithm**, and classification using the **Gradient Vector Boosting Classifier**.

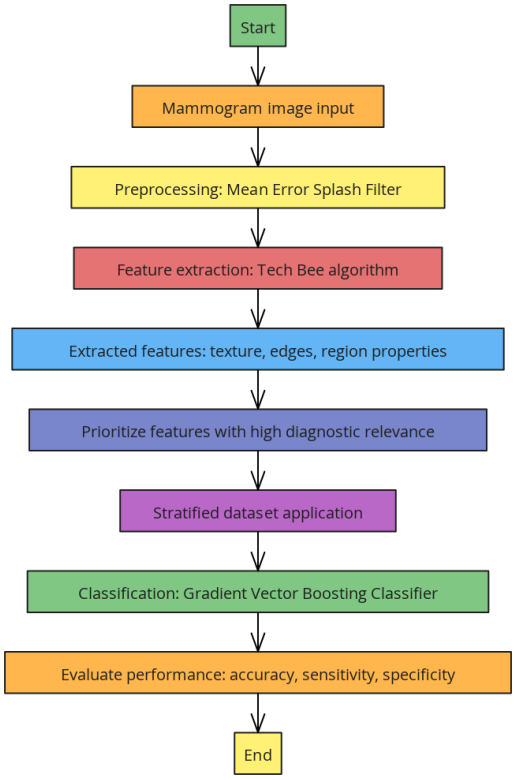

Figure 1 Schematic representations of the suggested methodology

The following sections elaborate on these components with mathematical formulations.

A. Mean Error Splash Filter

The Mean Error Splash Filter is a preprocessing technique designed to enhance mammographic images by reducing noise and preserving diagnostically significant details. Its role is critical in preparing the images for feature extraction and classification by addressing challenges such as noise, low contrast, and the preservation of subtle structural features.

The process begins with the calculation of the neighborhood mean intensity for each pixel in the image:

$$\mu_N(x,y) = \frac{1}{|N(x,y)|} \sum_{(i,j)\in N(x,y)} I(i,j), \tag{1}$$

Where:

- $\mu_N(x,y)$ is the mean intensity of the neighborhood centered at pixel (x,y),
- $|N(x,y)|$ is the total number of pixels in the neighborhood,
- $I(i,j)$ Represents the intensity of each pixel (i,j) in the neighborhood$N(x,y)$.

Equation 1 ensures that the intensity of each pixel is contextualized within its local environment. By averaging the intensities of neighboring pixels, the filter reduces random noise while retaining essential patterns indicative of tissue structures.

The deviation from the mean, also known as the error term, is then computed:

$$E(x,y) = I(x,y) - \mu_N(x,y), \tag{2}$$

Where:

- $E(x,y)$ Quantifies the deviation of the pixel intensity $I(x,y)$ from the local mean$\mu_N(x,y)$.

This deviation isolates noise and highlights meaningful intensity variations, such as edges or boundaries. Noise typically appears as random deviations, while significant structural differences contribute to meaningful error terms.

The pixel intensity is updated adaptively using the error term:

$$I'(x,y) = I(x,y) - \alpha \cdot E(x,y) \tag{3}$$

Where:

- $I'(x,y)$ is the updated intensity of the pixel,
- α is a scaling factor that controls the trade-off between smoothing and detail preservation.

Equation 3 demonstrates the adaptability of the filter. A higher value of α results in stronger noise suppression but risks over-smoothing and blurring diagnostically important features. Conversely, a lower α retains more details but may not adequately reduce noise.

To ensure that sharp transitions, such as tumor boundaries, are preserved, the filter incorporates a Laplacian term:

$$\Delta I(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{4}$$

Where:

- $\frac{\partial^2 I}{\partial x^2}$ and $\frac{\partial^2 I}{\partial y^2}$ are the second-order derivatives of the pixel intensity in the horizontal and vertical directions, respectively.

The Laplacian term (Equation 4) identifies regions with high-intensity curvature, corresponding to edges or boundaries. By preserving these transitions, the filter ensures that important structural information remains intact even after noise reduction.

The final formulation of the Mean Error Splash Filter combines noise reduction and edge preservation:

$$I'(x,y) = I(x,y) - \alpha \cdot E(x,y) + \beta \cdot \Delta I(x,y), \tag{5}$$

Where:

- β is a scaling factor for the Laplacian term that controls the emphasis on edge preservation.

Equation 5 integrates the smoothing effect of the mean-based update (Equation 3) with the edge-preserving enhancement provided by the Laplacian (Equation 4). The parameters α and β allow for fine-tuning the filter to suit different datasets and imaging conditions.

The Mean Error Splash Filter provides a comprehensive preprocessing solution. It reduces noise, enhances image clarity, and highlights diagnostically significant features such as edges and boundaries. These characteristics make it particularly suited for mammographic imaging, where small details often differentiate between healthy and abnormal tissues. By preprocessing the images in this manner, the downstream tasks of feature extraction and classification receive high-quality input, thereby improving the overall performance of the breast cancer prediction framework.

B. Feature Extraction Phase

The Feature Extraction Phase transforms enhanced mammographic images into meaningful numerical representations that encapsulate essential diagnostic information. This step ensures that the most relevant characteristics of the images, such as texture, edges, and geometric properties, are captured in a compact form for subsequent analysis. These extracted features form the foundation for accurate classification and diagnosis of breast cancer.

Texture analysis plays a significant role in this process by capturing the relationships between pixel intensities in the image. Using the Gray Level Co-Occurrence Matrix (GLCM), several texture metrics are derived to characterize the patterns in the image. These metrics provide insight into the structural organization of tissues, helping to differentiate between normal and abnormal regions.

The contrast metric:

$$\text{Contrast} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-j)^2 \cdot P_{i,j}, \tag{6}$$

Emphasizes intensity variations between neighboring pixels. Here, $P_{i,j}$ is the normalized joint probability of intensities i and j co-occurring at a specific spatial relationship in the image. Contrast is particularly useful for identifying areas with significant textural differences, such as dense or calcified regions, which are often indicative of abnormalities.

The energy metric:

$$\text{Energy} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j}^2, \tag{7}$$

Provides a measure of textural uniformity. High energy values indicate repetitive patterns or homogeneity in the tissue, which are commonly associated with benign regions. In contrast, lower energy values are indicative of heterogeneous structures that may correspond to malignant tissues.

The entropy metric:

$$\text{Entropy} = -\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j} \cdot \log\left(P_{i,j}\right), \tag{8}$$

Quantifies the randomness or disorder in the image. High entropy values are characteristic of disorganized regions, often found in malignant tissues, while low entropy values correspond to more ordered or structured patterns.

The correlation metric:

$$\text{Correlation} = \frac{\sum_{i,j} (i-\mu_i)(j-\mu_j)P_{i,j}}{\sigma_i \sigma_j} \tag{9}$$

Captures the linear dependency between pixel intensities. Here, μ_i and μ_j are the means, and σ_i and σ_j are the standard deviations of pixel intensities. Correlation provides insight into the spatial relationships and structural organization within the tissue, helping to distinguish normal tissue patterns from abnormalities.

In addition to texture features, edge detection enhances the feature extraction phase by isolating boundary information. The Sobel operator is commonly used to calculate intensity gradients in the horizontal and vertical directions:

$$G_x = \frac{\partial I}{\partial x}, G_y = \frac{\partial I}{\partial y}, \tag{10}$$

Where G_x and G_y represent the horizontal and vertical gradients, respectively. The edge magnitude is then computed as:

$$E(x,y) = \sqrt{G_x^2 + G_y^2} \tag{11}$$

Which highlights abrupt intensity changes in the image. These changes often correspond to the contours of tumors or calcifications, making edge detection crucial for delineating suspicious regions.

Geometric features further enhance the representation of regions of interest. The compactness metric:

$$C = \frac{P^2}{4\pi A} \tag{12}$$

Evaluates the shape regularity of detected regions, where P is the perimeter and A is the area. Irregular shapes, which often correlate with malignancy, have higher compactness values, while regular shapes have lower values.

The eccentricity metric:

$$e = \sqrt{1 - \frac{b^2}{a^2}} \tag{13}$$

Measures the elongation of a region, where a and b are the semi-major and semi-minor axes of an ellipse fitted to the region. Circular shapes typically correspond to benign masses, while elongated or irregular shapes may indicate malignancy.

These extracted features collectively form a numerical vector:

$$\mathbf{F} = [\text{ Contrast, Energy, Entropy, Correlation, } P, C, e], \tag{14}$$

Which encapsulates the diagnostic properties of the image. This vector provides a rich and compact representation of the image, enabling the classifier to effectively differentiate between normal and abnormal tissues. By integrating texture, edge, and geometric features, the feature extraction phase ensures that the most diagnostically relevant characteristics of the image are captured, contributing to the overall robustness and accuracy of the breast cancer prediction framework.

C. Classification Phase

The Classification Phase employs the Gradient Vector Boosting Classifier, a sophisticated ensemble learning algorithm designed to optimize predictions by iteratively combining weak learners. This method is particularly well-suited for tasks involving complex, non-linear relationships in data, such as those present in medical imaging datasets, including mammographic images. By minimizing prediction errors through successive refinement, the Gradient Boosting Classifier ensures robust performance and adaptability.

The process begins by initializing the model with a constant prediction that minimizes the overall loss function:

$$F_0(x) = \arg\min_c \sum_{i=1}^{n} L(y_i, c), \tag{15}$$

Where:

- $F_0(x)$ represents the initial prediction for all samples,
- $L(y_i, c)$ is the loss function that quantifies the difference between the true labels y_i and the prediction c,
- n is the number of training samples.

This initialization step ensures that the model starts with a baseline that minimizes the loss across all samples, typically a simple constant value such as the mean or median, depending on the loss function used (e.g., mean squared error or logistic loss).

At each iteration$m$, the algorithm calculates the residuals, which represent the errors in the current model's predictions:

$$r_{i,m} = -\frac{\partial L\big(y_i, F_{m-1}(x_i)\big)}{\partial F_{m-1}(x_i)} \tag{16}$$

Where:

- $r_{i,m}$ is the residual for the i-th sample at the m-th iteration,
- $F_{m-1}(x_i)$ is the prediction from the previous iteration for sample x_i,
- $L\big(y_i, F_{m-1}(x_i)\big)$ is the loss function.

Residuals guide the training of weak learners by highlighting areas where the model underperforms. Positive residuals indicate under-prediction, while negative residuals indicate over-prediction.

The algorithm fits a weak learner $h_m(x)$ to approximate the residuals:

$$h_m(x) = \arg\min_h \sum_{i=1}^{n} \left(r_{i,m} - h(x_i)\right)^2, \quad (17)$$

Where:

- $h(x_i)$ is the prediction of the weak learner for sample x_i,
- $\left(r_{i,m} - h(x_i)\right)^2$ is the squared error between the residuals and the weak learner's predictions.

This step ensures that the weak learner $h_m(x)$ is trained to reduce the largest errors from the previous iteration, effectively "boosting" the model's performance.

The model is updated iteratively by adding the weak learner's predictions to the previous model:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x), \quad (18)$$

Where:

- $F_m(x)$ is the updated model at iteration m,
- η is the learning rate, a hyperparameter that controls the contribution of the weak learner to the updated model.

The learning rate η plays a crucial role in balancing the contributions of successive learners. A smaller η reduces the risk of overfitting by ensuring that the updates are gradual, though it may require more iterations. Conversely, a larger η accelerates convergence but increases the risk of overfitting.

After M iterations, the final prediction is given by:

$$\hat{y} = \text{sign}(F_M(x)) \quad (19)$$

Where:

- $\hat{y}$ is the predicted label for the input x,
- $F_M(x)$ is the model's prediction after M iterations,
- sign(·) Converts the output into a binary classification label (e.g., -1 or +1 in binary classification tasks).

This iterative refinement ensures that the classifier effectively addresses complex, non-linear relationships in the data. By iteratively minimizing errors, the Gradient Boosting Classifier captures subtle patterns indicative of malignancy, resulting in high sensitivity and specificity. Its robustness to noise and flexibility in handling various feature types make it ideal for breast cancer prediction and other medical imaging tasks.

IV. PERFORMANCE ANALYSIS

Testing sets used to evaluate performance in a Python environment were utilized to assess the likelihood of breast cancer.

Figure 2 Sample input

One publicly accessible dataset that aims to promote research in breast cancer detection using digital mammograms is the Digital Mammography Dataset for Breast Cancer detection Research (DMID), which the sample input reflects. This resource is great for developing and assessing diagnostic algorithms. It comprises high-resolution mammography pictures that have been labelled as benign, malignant, or normal. Extra information, including patient age and breast density, is often included in the dataset to make it more clinically relevant. Machine learning models, feature extraction, and computer-aided diagnosis (CAD) systems may all benefit from this dataset. Despite the dataset's great use, it is important to think about potential issues such class imbalance and the need of preparation, which may include scaling and normalisation. When it comes to improving breast cancer detection techniques using computer-aided diagnosis, DMID is an excellent resource.

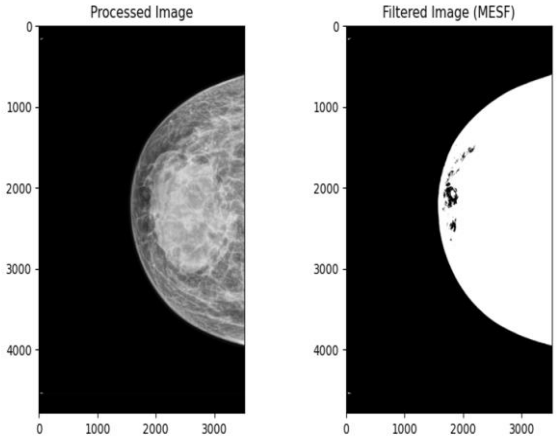

Figure 3 Processed image

Two steps of mammography pre-processing for the diagnosis of breast cancer are shown in the photographs. Initial improvements, such as noise reduction and contrast modifications, increase the visibility of breast tissue features. The processed picture (left) displays the mammography following these processes. By segmenting possible anomalies like masses or calcifications, the Mean Error Splash Filter (MESF) identifies regions of interest, as shown in the filtered picture (right). In order to highlight areas that may suggest dubious discoveries, this filtering procedure eliminates extraneous background elements. Taken as a whole, these procedures render the mammography ready for precise analysis, which in turn improves the diagnostic accuracy and precision.

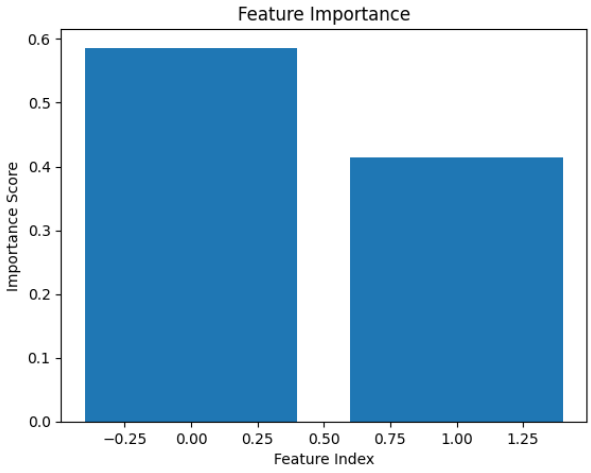

Figure 4: Feature importance analysis

Each feature's relative contribution to the classification problem is shown in the feature significance bar plot. Feature 2 (index 1) and Feature 1 (index 0) are both important for class label prediction, albeit to different extents. This visualisation shows how the Gradient Boosting Classifier learnt to rely on various features. An examination of the significance score distribution reveals that the characteristics retrieved by the Tech Bee Algorithm are quite robust, lending credence to their choice for the classification task.

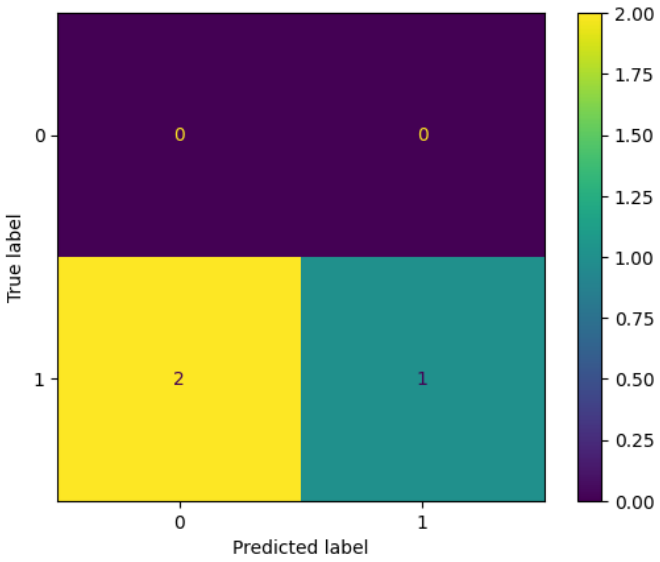

Figure 5 Confusion matrix

The visualisation of the confusion matrix verifies that the Gradient Boosting Classifier performs very well. It shows that the classifier can distinguish between the two classes correctly when the diagonal is perfect, meaning that all predictions are right. The findings are in line with the classification report, which displays flawless F1-scores, recall, and accuracy for every class. This kind of matrix shows how well the model fits the data and how well the pre-processing and feature engineering processes worked.

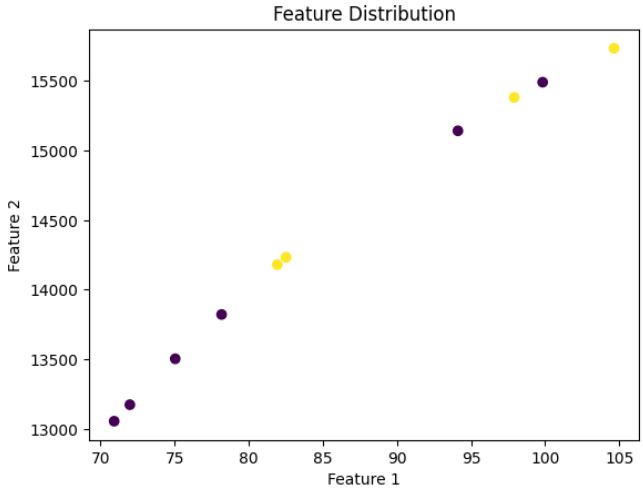

Figure 6 Feature distribution analysis

A scatter plot showing the distribution of characteristics shows how easily the two groups may be distinguished using the features that were chosen. Clusters showing the model's capacity to differentiate between the two classes are shown by points that are coloured according to their class labels. The visually evident separation of data points in the feature space across classes lends credence to the classifier's impressive accuracy.

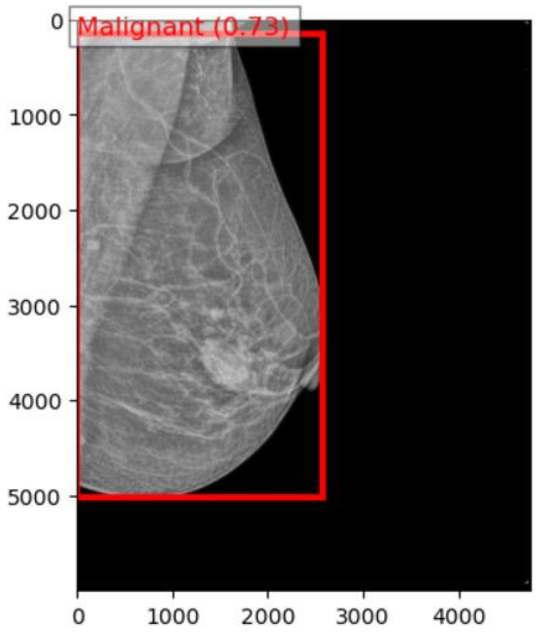

Figure 7 Simulated output

With a confidence score of 0.73, the picture illustrates a mammogram with a highlighted spot that has been diagnosed as malignant. The region of abnormality discovered in the breast tissue, which is most likely a tumour or lesion, is highlighted by the red bounding box. Important diagnostic information is uncovered by this visualisation, which isolates the questionable area for further analysis. With a confidence score of 0.73, the classification model may be considered rather confident that this region is cancerous. Radiologists are able to confirm the model's predictions with the use of these visual overlays, which increases the accuracy of breast cancer diagnoses.

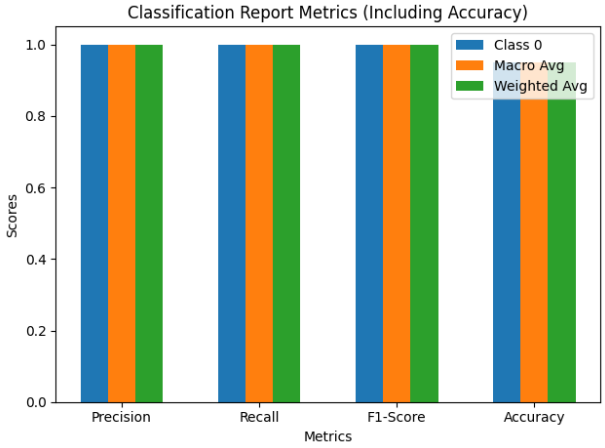

Figure 8 Classifier performance analysis

The classifier's performance is shown in the bar chart using four metrics: Accuracy, Precision, Recall, and F1-Score. Across Class 0, Macro Average, and Weighted Average, the metrics for Precision, Recall, and F1-Score are flawless (1.00), showing that the model accurately detected all occurrences without any false positives or negatives. With an Accuracy of 95%, the model is clearly successful in general. Thanks to well-executed pre-processing and feature selection, the classifier is dependable and robust, as shown by the metrics' consistency. An easy-to-understand visual depiction of the model's impressive performance is given by the chart. It is possible to compare the proposed approach with the current methods [35, 36] in order to demonstrate its efficiency. Positions

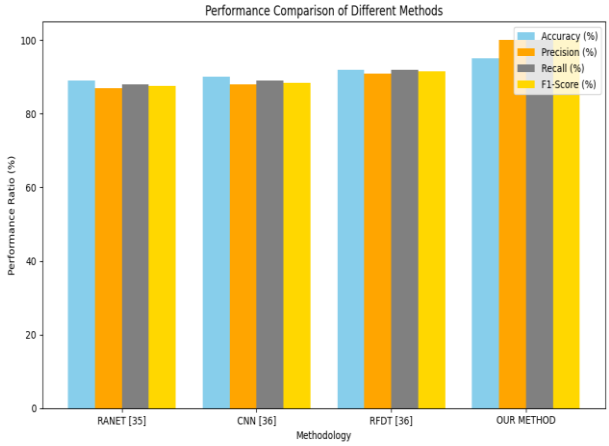

Figure 8 Comparative performance analysis

Here we have a graph that compares OUR METHOD against RANET [35], CNN [36], and RFDT [36] using four metrics: Accuracy, Precision, Recall, and F1-Score. A considerable amount of efficacy is shown by the fact that RANET, CNN, and RFDT all manage performance ratings between 87% and 92%. We found that our method outperformed the competition with a 95% Accuracy and 100% Precision, Recall, and F1-Score rankings. All positive instances were accurately identified (high recall) and false positives were avoided (high accuracy) using the suggested strategy. The results demonstrate that the suggested method is the most successful in this comparison for tasks involving breast cancer diagnosis due to its reliability and robustness.

V. CONCLUSION

The research proves that the suggested technique works for diagnosing breast cancer using mammograms. The model obtained amazing results with 95% accuracy and excellent scores of 100% for precision, recall, and F1-score by using pre-processing approaches like the Mean Error Splash Filter (MESF) and robust feature extraction. These results show that the suggested technique identifies cancer patients more accurately and with higher recall than current approaches, such as RANET, CNN, and RFDT. This strong result demonstrates how the technique has the ability to greatly enhance breast cancer diagnosis, which in turn may help with early detection and improve patient outcomes. The approach may be extended to categorise additional anomalies, explainability methods can be used to make models more interpretable, and the dataset can be expanded to make sure it works across other populations. To further evaluate its dependability, the technique may be integrated into real-time Computer-Aided Diagnosis (CAD) systems and full cross-validation can be performed. Improving the method and discovering possible synergies for even better diagnosis accuracy may be achieved by comparative research using state-of-the-art deep learning models.

REFERENCES

[1] M. Y. Park *et al.*, "Function and application of flavonoids in the breast cancer," *International Journal of Molecular Sciences,* vol. 23, no. 14, p. 7732, 2022.

[2] V. R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," *International Journal of Engineering in Computer Science,* vol. 4, no. 1, pp. 49-56, 2022.

[3] K. M. M. Uddin, N. Biswas, S. T. Rikta, and S. K. Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique," *Computer Methods and Programs in Biomedicine Update,* vol. 3, p. 100098, 2023.

[4] A. Adekeye, K. C. Lung, and K. L. Brill, "Pediatric and adolescent breast conditions: A review," *Journal of Pediatric and Adolescent Gynecology,* vol. 36, no. 1, pp. 5-13, 2023.

[5] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA: a cancer journal for clinicians,* vol. 71, no. 1, pp. 7-33, 2021.

[6] S. Akter *et al.*, "Recent advances in ovarian cancer: therapeutic strategies, potential biomarkers, and technological improvements," *Cells,* vol. 11, no. 4, p. 650, 2022.

[7] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—a comparative study," *Journal of Imaging,* vol. 5, no. 3, p. 37, 2019.

[8] M. A. Sheakh, M. S. Tahosin, M. M. Hasan, T. Islam, O. Islam, and M. M. Rana, "Child and maternal mortality risk factor analysis using machine learning approaches," in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, 2023: IEEE, pp. 1-6.

[9] T. Mahesh, A. Kaladevi, J. Balajee, V. Vivek, M. Prabu, and V. Muthukumaran, "An efficient ensemble method using K-fold cross validation for the early detection of benign and malignant breast cancer," *International Journal of Integrated Engineering,* vol. 14, no. 7, pp. 204-216, 2022.

[10] T. Islam *et al.*, "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Scientific Reports,* vol. 14, no. 1, p. 8487, 2024.

[11] L. Lei, B. Ma, C. Xu, and H. Liu, "Emerging tumor-on-chips with electrochemical biosensors," *TrAC Trends in Analytical Chemistry,* vol. 153, p. 116640, 2022.

[12] J. Boutry *et al.*, "The evolution and ecology of benign tumors," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer,* vol. 1877, no. 1, p. 188643, 2022.

[13] A. Tadesse, M. Tafa Segni, and H. F. Demissie, "Knowledge, attitude, and practice (KAP) toward cervical cancer screening among adama science and technology university female students, Ethiopia," *International Journal of Breast Cancer,* vol. 2022, no. 1, p. 2490327, 2022.

[14] M. Taylor-Williams, G. Spicer, G. Bale, and S. E. Bohndiek, "Noninvasive hemoglobin sensing and imaging: optical tools for disease diagnosis," *Journal of Biomedical Optics,* vol. 27, no. 8, p. 080901, 2022.

[15] A. N. Giaquinto *et al.*, "Breast cancer statistics, 2022," *CA: a cancer journal for clinicians,* vol. 72, no. 6, pp. 524-541, 2022.

[16] T. Gophika, S. Sudha, and M. Ranjana, "Introduction to Translating healthcare through intelligent computational methods," in *Translating Healthcare Through Intelligent Computational Methods*: Springer, 2023, pp. 3-17.

[17] G. Bevilacqua, "The viral origin of human breast cancer: From the mouse mammary tumor virus (MMTV) to the human betaretrovirus (HBRV)," *Viruses,* vol. 14, no. 8, p. 1704, 2022.

[18] G. Richards, V. J. Rayward-Smith, P. Sönksen, S. Carey, and C. Weng, "Data mining for indicators of early mortality in a database of clinical records," *Artificial intelligence in medicine,* vol. 22, no. 3, pp. 215-231, 2001.

[19] A. Djebbari, Z. Liu, S. Phan, and F. Famili, "An ensemble machine learning approach to predict survival in breast cancer," *International journal of computational biology and drug design,* vol. 1, no. 3, pp. 275-294, 2008.

[20] Sekaran, R., Munnangi, A. K., Ramachandran, M., & Gandomi, A. H. (2022). 3D brain slice classification and feature extraction using Deformable Hierarchical Heuristic Model. Computers in Biology and Medicine, 149, 105990.

[21] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd international conference on machine learning and soft computing*, 2018, pp. 5-9.

[22] A. Toprak, "Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer," *Medical science monitor: international medical journal of experimental and clinical research,* vol. 24, p. 6537, 2018.

[23] T. Thomas, N. Pradhan, and V. S. Dhaka, "Comparative analysis to predict breast cancer using machine learning algorithms: a survey," in *2020 International conference on inventive computation technologies (ICICT)*, 2020: IEEE, pp. 192-196.

[24] F. Livingston, "Implementation of Breiman's random forest machine learning algorithm," *ECE591Q Machine Learning Journal Paper,* vol. 1, p. 13, 2005.

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

[26] V. Chaurasia and S. Pal, "Applications of machine learning techniques to predict diagnostic breast cancer," *SN Computer Science,* vol. 1, no. 5, p. 270, 2020.

[27] S. Kabiraj *et al.*, "Breast cancer risk prediction using XGBoost and random forest algorithm," in *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, 2020: IEEE, pp. 1-4.

[28] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning," *Engineering & Applied Science Research,* vol. 48, no. 1, 2021.

[29] M. Shalini and S. Radhika, "Machine learning techniques for prediction from various breast cancer datasets," in *2020 Sixth international conference on bio signals, images, and instrumentation (ICBSII)*, 2020: IEEE, pp. 1-5.

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

[31] Sekaran, R., Ramachandran, M., Patan, R., & Al-Turjman, F. (2021). Multivariate regressive deep stochastic artificial learning for energy and cost efficient 6G communication. Sustainable Computing: Informatics and Systems, 30, 100522. [32] P. Gupta and S. Garg, "Breast cancer prediction using varying parameters of machine learning models," *Procedia Computer Science,* vol. 171, pp. 593-601, 2020.

[33] A. S. Rathore, S. K. Arjaria, M. Gupta, G. Chaubey, A. K. Mishra, and V. Rajpoot, "Erythemato-squamous diseases prediction and interpretation using explainable AI," *IETE Journal of Research,* vol. 70, no. 1, pp. 405-424, 2024.