Machine Learning Based Expert System For Breast Cancer Prediction

Shivani, Nidhi, Yogesh, M.S.Bennet Praba, S.Deepa

Dept of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

Abstract: Breast cancer continues to be among the top causes of death in women globally. Detection at an early stage is fundamental in enhancing prognosis and survival rate. Conventional methods of detection, including biopsy and histopathological analysis, are invasive, time-consuming, and prone to human error. To overcome the above limitations, machine learning has proven to be a valuable application in medical imaging, improving efficiency and accuracy of breast cancer prediction. Machine Learning-Based Expert System for Breast Cancer Prediction, utilizes mammography for the detection of cancer at an early stage. Fuzzy C-means (FCM) clustering and Principal Component Analysis (PCA) are some advanced image processing techniques utilized for extracting meaningful texture features from the images of the breast. For the classification task, the K-Nearest Neighbours (KNN) algorithm is used with a staggering accuracy rate of 92%. The system not only categorizes tumour presence but also indicates tumor edges, enabling doctors to visualize the affected area clearly. The improvement in terms of accuracy and efficiency is considerably higher compared to existing work. Most existing research was based on simple statistical characteristics or single-level wavelet transform, which confined their capability in identifying complex textures of tumours. Additionally, classification methods such as Support Vector Machines and Artificial Neural Networks, though efficient, on occasion caused computational overhead. To address this, this system combines multi-level feature extraction algorithms with an optimized KNN classifier, resulting in higher accuracy and lower processing time. With its enhanced predictive value, this expert system assists radiologists in making quicker, more accurate diagnoses, thereby ensuring improved patient outcomes and increased survival rates. Future upgrades may include deep learning for even more precise diagnosis.

Keywords— Mammogram, KNN Classification, Benign, Malignant

I. INTRODUCTION

Digital image processing refines raw imaging data through pre-processing, enhancement, and information extraction. Content-based image retrieval (CBIR) improves retrieval accuracy and speed by assessing image similarity. The project utilizes classification techniques like Fuzzy C-Means (FCM) and K-Nearest Neighbors (KNN) for feature extraction. FCM clustering segments mammographic images, while Multi-level Discrete Wavelet Transform extracts edge details. Principal Component Analysis (PCA) and the Gray Level Co-occurrence Matrix (GLCM) analyze segmented data, yielding 13 key features for machine learning classification. Supervised algorithms categorize images as Benign, Malignant, or Normal based on tumor shape, size, and boundaries. The system highlights tumor areas, aiding in more effective breast cancer detection.

II. LITERATURE SURVEY

[1] Microarray breast cancer data using machine learning.Eight algorithms, including SVM, KNN, MLP, and Random Forest, were applied.Feature selection methods RFE and RLR reduced the dataset to 50 features. Classification performance was evaluated before and after feature selection.

[2] Deep learning approaches for breast cancer detection.The study emphasizes the role of Convolutional Neural Networks CNNs in mammogram analysis. It provides a comprehensive overview of existing research in this domain.Strengths and limitations of various deep learning models are discussed.The survey examines performance variations across different methodologies.Key insights are highlighted to guide future research in breast cancer diagnosis.Findings aim to assist healthcare professionals in adopting effective techniques.This study serves as a valuable reference for deep learning applications in medical imagin

[3] Breast cancer detection.The study explores classification algorithms like SVM, deision trees, and random forests.Strengths and limitations of these methods are discussed in detail. highlights the role of ML in improving diagnostic accuracy.Comparative analysis of various algorithms is provided.Key insights support further research in medical AI applications. Findings assist researchers and healthcare professionals in adopting ML techniques.This review serves as a valuable reference for breast cancer diagnosis.

[4] Breast cancer diagnosis.They analyzed SVMs, ANNs, and Decision Trees in medical applications.The study explores the strengths and limitations of each algorithm.Comparative insights highlight their effectiveness in diagnosis.Applications in improving diagnostic accuracy are discussed.Challenges in breast cancer detection are addressed. Findings support future research in medical AI.This review serves as a key resource for researchers and clinicians.

[5] Reviewed deep learning in medical image analysis.The study highlights the role of CNNs in breast cancer detection.It explores deep learning applications in disease diagnosis.Current advancements in medical imaging are discussed.Strengths and limitations of various techniques are analyzed. provides insights into AI-driven diagnostic improvements. Challenges in implementing deep learning in healthcare are addressed.This survey serves as a key reference for medical AI research.

[6] Explores CNNs in medical image analysis for breast cancer detection. It examines CNN-based methods for classification, segmentation, and feature extraction. This review serves as a valuable resource for medical imaging research.

[7] Both supervised and unsupervised learning approaches were analyzed.Supervised methods like SVMs and neural networks proved effective.Unsupervise techniques, including clustering, revealed data patterns.The study provides insights into ML applications in medical diagnosis.Comparative analysis highlights strengths and limitations of each method.Findings support advancements in AI-driven breast cancer detection.This survey serves as a key reference for ML-based diagnostics.

[8] CNN applications in mammogram classification. The study reviews CNN fundamentals for breast cancer detection.It highlights the role of representation learning in medical imaging.CNNs' potential for improving diagnostic accuracy is examined.Challenges and opportunities in mammogram analysis are discussed.Findings support advancements in AI-driven breast cancer diagnosis.The study informs the development of more effective CNN-based systems.

[9] Invasive breast cancer detection.The study focuses on whole-slide pathology image analysis using deep learning.It demonstrates the effectiveness of AI in identifying breast cancer.Diagnostic accuracy and reproducibility are key areas of improvement.Findings highlight machine learning's impact on clinical decision-making..

[10] Study explores pattern recognition techniques for breast cancer classification.Naive Bayes and SVMs are analyzed in medical imaging applications. provides an overview of methods for diagnostic model development.Pattern recognition approaches are evaluated for accuracy and efficiency.Findings highlight their potential in improving breast cancer detection.Applications in healthcare and medical imaging are discussed.The study offers insights into AI-driven diagnostic advancements.It serves as a valuable resource for researchers in pattern recognition

[11] Study examines machine learning algorithms for breast cancer detection. Mammogram analysis is conducted using various classification methods.Decision trees and SVMs are evaluated for tumor identification.The study compares classifier performance in cancer detection.Findings highlight the most effective method for diagnosis.Machine learning's role in improving accuracy is emphasized.Challenges and advancements in medical AI are discussed.This research supports the development of better diagnostic models.

[12] Study Logistic regression and random forests are applied to digital mammography. Digital mammography captures detailed breast tissue images for diagnosis.Classifier performance in cancer detection is thoroughly evaluated.Strengths and limitations of different models are analyzed.Findings highlight machine learning's role in improving diagnostics.The study contributes to developing more accurate detection methods.It serves as a valuable resource for medical AI

[13] Provides an overview of machine learning techniques in breast cancer detection.Various algorithms, including random forests, gradient boosting, and neural networks, are discussed.The study highlights machine learning's potential to enhance accuracy.

[14] Applied machine learning to predict breast cancer recurrence.Decision trees and logistic regression were used for analysis.Decision trees classified patients based on key characteristics.Logistic regression assessed variable relationships and recurrence probability.Models were trained on patient data, tumor features, and outcomes.Combining both methodsimproved predictive accuracy.Findings support personalized medicine and clinical decision-making.This study enhances AI-driven cancer prognosis research.

[15] Feature extraction and SVM.Mammogram images were analyzed for texture and statistical patterns. Extracted features enhanced classification accuracy.SVM effectively distinguished between benign and malignant cases.Its binary classification capability proved highly effective.Results showed SVM outperformed other machine learning models.Feature extraction played a crucial role in improving diagnosis.This study supports AI-driven advancements in medical imaging.

[16] Explores pattern recognition techniques for breast cancer classification.Naive Bayes and SVMs are analyzed in medical imaging applications.A comprehensive overview classification methods is presented.These techniques aid in developing accuracy.

[17] Explored instance-based learning for breast cancer prediction.Classification is based on the majority vote of nearest neighbors.K-nearest neighbors (KNN) is highlighted for its effectiveness. KNN utilizes local training data to improve prediction accuracy.The study demonstrates its applicability in medical diagnosis.Instance-based learning shows promise for breast cancer classification.Findings support its role in enhancing diagnostic precision.This research contributes toAI-driven advancements in healthcare.

[18] study examines machine learning algorithms for breast cancer detection.Mammogram analysis is conducted using various classification methods.Decision trees and SVMs are evaluated for tumor identification.Classifier performance is compared to determine the most effective approach. Findings highlight machine learning's role in improving diagnostic accuracy.Strengths and limitations of different models are discussed.The study contributes to AI-driven advancements in cancer detection.

[19] Study examines machine learning algorithms for breast cancer detection. Logistic regression and random forests are applied to digital mammography.Digital mammography captures detailed breast tissue images for diagnosis. Classifier performance is evaluated to assess detection accuracy. Strengths and limitations of different models are analyzed. Findings highlight machine learning's role in improving diagnostics.The study contributes to developing more reliable detection methods. It serves as a valuable resource for AI-driven medical imaging research.

[20] Reviews feature selection techniques in microarray data analysis.Feature selection is crucial in bioinformatics for DNA microarray processing.Biogeography-Based Optimization (BBO) models species migration for optimization.Particle Swarm Optimization (PSO) simulates particle movement in a search space.Gene selection is categorized into independent, half-dependent, and dependent features in bioinformatics.This review serves as a key resource.

III. PROPOSED WORK

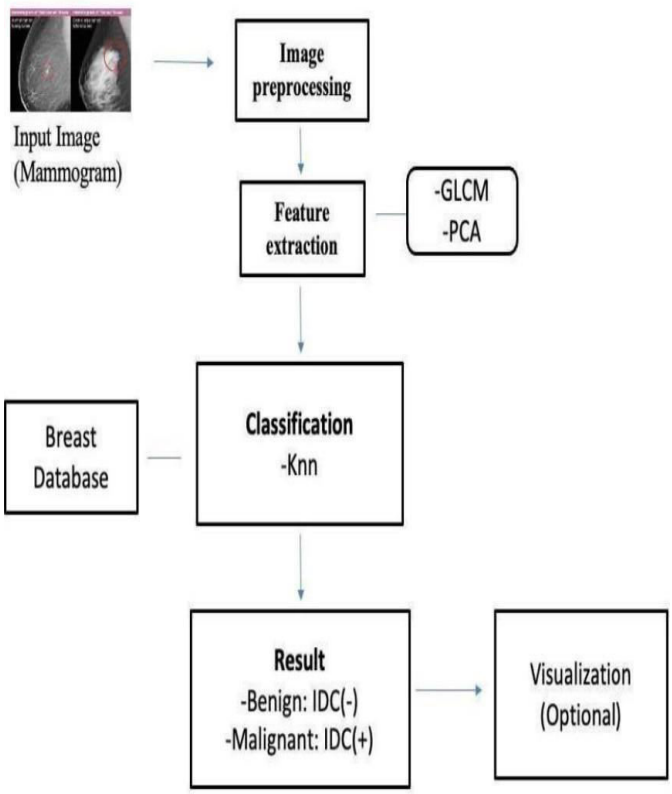

Fig 1. Architecture diagram for the working of the system

A. Image Preprocessing – The system begins by acquiring and refining input data. Once the image is obtained, it is converted into a grayscale format to enhance processing.

B. Image Segmentation – Suitable segmentation techniques are applied to extract significant objects within the image. Clustering, a powerful segmentation method, groups data points into clusters. The system employs Fuzzy C-Means (FCM) clustering for this purpose.

C. Feature Extraction – The segmented region is thoroughly analyzed using Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray Level Co-occurrence Matrix (GLCM). A total of 13 features, including mean, variance, and entropy, are extracted, with their pixel values stored in a database as a matrix.

D. Classification – The extracted features are compared with a dataset stored in a .mat file and later classified as Benign, Malignant, or Normal. Morphological operations are performed to determine region properties such as area, eccentricity, and Euler number. If cancerous cells are detected, the system computes and highlights the tumor area along with the detected boundary.

E. Results – The final output categorizes the image as either Cancerous (IDC+) or Non-Cancerous (IDC-).

MODULES

1. Image Preprocessing

Input data is gathered and cleaned.

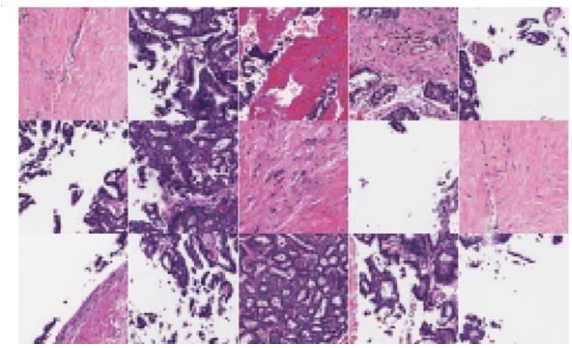

Fig 2. Sample images from the dataset

2. Image Segmentation

Pseudocode:

Clustering which is based on FCM algorithm

$$Jm = \sum(i = 1\ to\ N) \sum(j = 1\ to\ C)\ (uij)^{m} * ||x(i) - c(j)||^{2}$$

Equ-(1)

Where
Jm is the objective function,
N is the number of data points,
C is the number of clusters,
uij is the membership degree of xi in cluster j,
m is the fuzziness parameter (m>1),
xi is a data point, cj is a cluster center,
| | xi−cj| | ^2 is the squared Euclidean distance.

1. Randomly initialize the cluster membership values, µij.
2. Calculate the cluster centers:

$$cj = \sum(i = 1\ to\ N) uijm \cdot xi / \sum(i = 1\ to\ N) uijm$$

Equ-(2)

3. Update µij according to the following:

$$uij = 1 / (\sum (||x\_j - c\_k||^2 / ||x\_j - c\_i||^2)^{(m-1)})$$

Equ-(3)

4. Calculate Jm which is the objective function

5. Repeat the steps 2-4 until Jm improves by less than a specified minimum threshold or until after a specified maximum number of iterations.

Equation (1) represents the formula for Clustering based on FCM
Equation (2) is used to calculate the cluster centers
Equation (3) calculates the membership degree of clusters

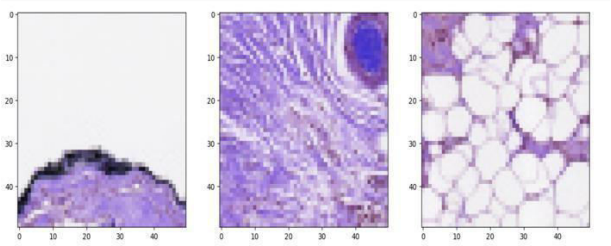

Fig 3. Segmented images

3. Feature Extraction

Pseudocode:
1. Dimensionality reduction
2. Wavelet Transform

$$F(a,b) = int - infty\ ^{\wedge}\ infty\ f(x)\ \psi,(a,b)\ ^{\wedge} * (x) dx$$

Equ-(4)

Where
F(a,b) is the wavelet transform coefficient,
f(x) is the original function,
ψ a,b∗(x) is the conjugate of the wavelet function,
ψ a,b(x)=1|a|ψ(x−ba), a is the scale, b is the translation.

3. Discrete Wavelet Transform (DWT)

$$\phi(x) = k \sum hk\ \phi(2x - k)$$

Equ-(5)

4. Principal component analysis (PCA)
5. Grey Level Co-occurrence Matrix (GLCM)

```
Total number of images: 90000
Number of IDC(-) Images: 61191
Number of IDC(+) Images: 28809
Percentage of positive images: 32.01%
Image shape (Width, Height, Channels): (50, 50, 3)
```

Fig 4. Data from feature extraction

Equation (4) represents the wavelet transform coefficient
Equation (5) gives the wavelet function

4. Classification
Pseudocode:
1. Read dataset and dataset has number of rows “r” and number of columns “m”
2. For (i=0;i=r; i++) /// selection of centroid point
For (j=0; j=m; j++)

Select k=data (i, j); End

3. Calculation of Euclidian distance
4. For(i=0;i=r;i++)
   For(j=9;j=m;j++)
   A(i)=data(i);
   B(i)=data(j);
   Distance = sqrt[(A(i+1)-A(i)^2) –(B(j+1)-B(j)^2);
   End
   End

1. Normalization ()
2. For (k=0;k=data++)
3. Swap k(i+1) and k(i); End

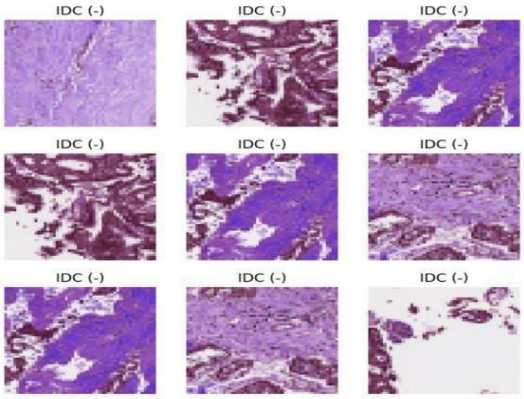

Fig 5. IDC(-) and IDC(+) classifications of dataset

IV. RESULTS AND DISCUSSION

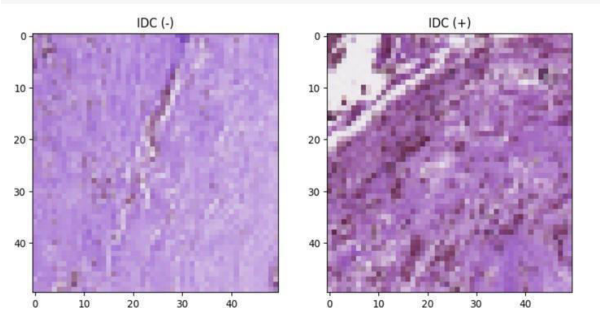

Fig 6. IDC(-) and IDC(+) results

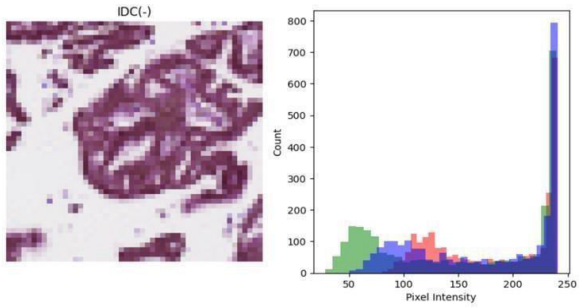

Fig 7. IDC(-) image with histogram of pixel intensity

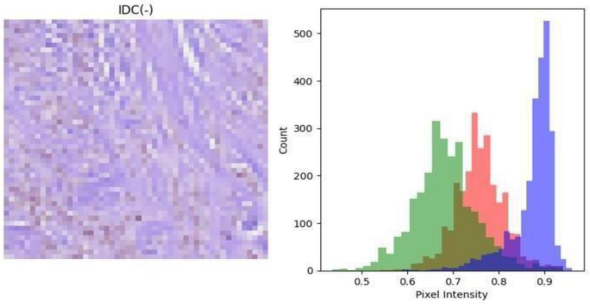

Fig 8. IDC(-) image with histogram of pixel intensity

The physician uploads the mammogram of the patient to the device that is subjected to the process of image segmentation . Using image processing technique, the segmented image is pre- processed. Extraction methods to extract necessary features are applied to the image. The classifier model is given extracted features, then the test image classification process is performed with respect to the training data present in the database. The test picture is marked as either cancerous or non- cancerous. Unless the test image is marked as cancerous, the tumor region will be measured, and the findings will be shown to the doctor along with the observed boundary image.

In this project the accuracy is calculated by using the formula below . Let us take x as the testing data true label and the prediction labels for it would be p. Here we are taking 70 datasets for testing and calculating the true classified sets with its total number of the test data.

Accuracy= sum(x which is the true label)/ sum (total of all the test data) × 100.

Using this formula the accuracy is computed and gives a good efficiency. The use of Multi-Level Wavelet Conversion strategy coupled with PCA with 13 features extracted and subsequently classified provides an average accuracy of almost 92%.

Table 1. Shows existing work and the proposed work

EXISTING WORK	PROPOSED WORK
Traditional thresholding, region growing, or manual segmentation	Fuzzy C-Means (FCM) Clustering for automated and soft clustering-based image segmentation
Single or basic methods (e.g., GLCM only, texture or shape features)	Multi-level DWT, PCA, and GLCM — giving a richer 13-feature representation
Often not used or basic feature selection methods	Principal Component Analysis (PCA) applied for effective dimensionality reduction
SVM, Random Forest, or basic ANN	K-Nearest Neighbors (KNN) used for classification based on cell shape

Table 1 shows the major differences between existing work and the proposed work based on the algorithms and techniques used.

V. CONCLUSIONS

Breast cancer remains a leading cause of death worldwide, with limited accurate prognostic and predictive factors currently in clinical use. A proposed system utilizing Clustering with a Level Set approach aims to improve detection accuracy of affected cell shapes by marking detected contours. Fuzzy-C-means (FCM) clustering is applied for image segmentation, and features such as Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray Level Co-occurrence Matrix (GLCM) are extracted, resulting in 13 features stored in a database. The KNN classifier then classifies images as Benign, Malignant, or Normal, depending on cell shape. Morphological operations also improve boundary detection and calculation of tumor area. The system has an average accuracy of 92%, and future enhancements can include medicine and treatment suggestions depending on disease severity to aid in diagnosis and treatment.

VI. REFERENCES

[1] M. K. Bashar, et al. (2019). "Deep Learning Techniques for Breast Cancer Detection and Diagnosis: A Comprehensive Survey." This survey reviews deep learning methods, particularly convolutional neural networks (CNNs), used in breast cancer detection from mammogram images.

[2] S. Khan, A. Hussain, et al. (2018). "Automated Breast Cancer Detection Using Machine Learning Techniques: A Review." Discusses the use of classification algorithms, such as support vector machines (SVM), decision trees, and random forests in breast cancer detection.

[3] E. Şahin, et al. (2019). "A Comprehensive Review of Machine Learning Methods for Breast Cancer Diagnosis." Focuses on SVMs, artificial neural networks (ANNs), and decision trees for breast cancer diagnosis.

[4] S. Al-Shaikhli, et al. (2021). "Breast Cancer Prediction Models Using Machine Learning Techniques: A Review." Reviews machine learning algorithms for breast cancer prediction, such as KNN, SVM, and neural networks.

[5] G. Litjens, et al. (2017). "A Survey on Deep Learning in Medical Image Analysis." This paper highlights the use of deep learning techniques, such as CNNs, in analyzing breast cancer images.

[6] A. Qayyum, et al. (2020). "Medical Image Analysis Using Convolutional Neural Networks: A Review." Reviews CNN-based methods for detecting breast cancer through mammogram and histopathology image analysis.

[7] L. Zhang, et al. (2019). "Breast Cancer Diagnosis Using Machine Learning: A Survey." Discusses various machine learning approaches for breast cancer diagnosis, including supervised and unsupervised learning.

[8] J. Arevalo, et al. (2016). "Representation Learning for Mammogram Classification Using Convolutional Neural Networks." Reviews CNNs and their application in classifying mammograms for breast cancer detection.

[9] S. Cruz-Roa, et al. (2014). "Accurate and Reproducible Invasive Breast Cancer Detection in Whole-slide Images." Discusses the use of machine learning, especially deep learning, in detecting invasive breast cancer from whole-slide pathology images.

[10] R. O. Duda, P. E. Hart, and D. G. Stork (2012). "Pattern Classification." This paper covers pattern recognition techniques like Naive Bayes and SVMs for classifying breast cancer images.

[11] D. W. Aha, et al. (2016). "Instance-Based Learning Algorithms for Breast Cancer Prediction." Focuses on instance-based learning approaches such as K-nearest neighbors (KNN) for classifying breast cancer datasets.

[12] M. Elter, A. Horsch (2012). "Detection of Cancerous Tumors in Mammograms Using Machine Learning Techniques." Reviews different machine learning classifiers such as decision trees and SVMs for detecting breast cancer from mammograms.

[13] J. Shan, L. Alam, and M. Garra (2017). "Breast Cancer Detection Using Machine Learning Algorithms and Digital Mammography." Analyzes machine learning classifiers like logistic regression and random forests for breast cancer detection.

[14] A. K. Ghosh and S. Guha (2015). "Machine Learning for Breast Cancer Diagnosis: A Survey." Highlights the application of machine learning techniques, such as random forests, gradient boosting, and neural networks in breast cancer detection.

[15] H. Park, et al. (2018). "Using Machine Learning to Predict Breast Cancer Recurrence." Discusses how decision trees and logistic regression models are used to predict the recurrence of breast cancer.

[16] K. Li, et al. (2017). "Breast Cancer Diagnosis Using Feature Extraction and Support Vector Machines." Focuses on feature extraction and SVM- based classification for breast cancer diagnosis from mammogram images.

[17] W. Cai, F. Xue, and D. Feng (2013). "Content-Based Image Retrieval by Feature Adaptation in Medical Image Analysis." Discusses the use of CBIR methods in mammogram image retrieval and breast cancer prediction.

[18] S. Yasmin, R. B. Saeed, and M. A. Khan (2020). "AReview of Machine Learning Techniques for Breast Cancer Detection." Provides an overview of classification models, including decision trees, neural networks, and SVMs, for breast cancer detection.

[19] R. D. Lins, A. S. Junior (2019). "Artificial Intelligence Techniques Applied to Breast Cancer Diagnosis Using Image Processing." Surveys the use of AI algorithms for image-based breast cancer detection.

[20] J. R. Quinlan (2014). "Decision Trees for Breast Cancer Diagnosis." This paper surveys decision tree techniques for classifying breast cancer based on mammogram data.