An Experimental Evaluation of Deep Learning Networks for Automated Breast Cancer Detection

Partha Chakraborty*, Umme Aiman Jannat
Department of Computer Science and Engineering, Comilla University, Cumilla-3506, Bangladesh

***Abstract*—Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, where early and accurate diagnosis plays a vital role in improving survival rates. Recent advancements in deep learning have demonstrated significant potential in automating the analysis of medical images for cancer detection. This study presents a comprehensive comparative analysis of convolutional neural network (CNN)–based deep learning models for breast cancer classification using ultrasound and mammography images. Multiple architectures, including a baseline CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, were evaluated to classify breast lesions as benign or malignant. Experimental results reveal notable performance differences across imaging modalities. For ultrasound images, AlexNet achieved the highest accuracy of 89%, while DenseNet and the baseline CNN achieved 88% and 85%, respectively. In contrast, mammography-based classification yielded significantly higher performance, with the baseline CNN outperforming deeper architectures in terms of accuracy and F1-score, achieving 97%. The findings demonstrate that model complexity does not necessarily guarantee superior performance and that properly designed shallow CNNs can effectively outperform deeper networks on high-quality mammographic data. This study highlights the potential of deep learning–based computer-aided diagnosis systems to support radiologists in the early detection of breast cancer.**

***Keywords*—Breast cancer; deep learning; convolutional neural networks; mammography; ultrasound; medical image classification**

I. Introduction

Breast cancer is a major cause of death for women between the ages of 20 and 59. Breast cancer is the most common cancer in women worldwide. Bangladesh has one of the highest incidences of breast cancer in Asia. In Bangladesh, breast cancer is the most prevalent cancer among women aged 15 to 44 years, with an incidence of approximately 19.3 per 100,000 population compared to other cancer types. Early detection remains the most effective strategy to reduce mortality. Women under 40 are advised to perform regular breast self-examinations, while those over 40 are advised to undergo routine mammographic screening for early detection. Reducing death rates and increasing treatment outcomes depend heavily on early and precise identification. Despite their effectiveness, traditional diagnostic techniques are often limited by subjectivity and variability, making consistent and trustworthy diagnoses difficult. These constraints may be addressed by the revolutionary potential of machine learning (ML), which leverages enormous datasets and processing capacity to improve diagnostic accuracy and efficiency of diagnosis. Breast cancer happens when cells in the breast start growing in an abnormal and uncontrolled way, forming a lump called a tumor. If it is not treated in time, the cancer can spread from the breast to other parts of the body, such as the bones, liver, or lungs. It can affect both women and men, but it is far more common in women.

There are also early or precancerous conditions, often called precancerous or carcinoma in situ, in which abnormal cells are found in the milk ducts or lobules but have not yet spread into nearby tissues. Invasive breast cancer, on the other hand, begins to grow into the surrounding normal breast tissue and can potentially spread to other parts of the body [1]. The most common types include ductal carcinoma and lobular carcinoma. Other less common forms include Paget's disease of the breast and inflammatory breast cancer [2].

A prospective cross-sectional study was conducted at Ahsania Mission Cancer and General Hospital between July 2023 and December 2023 to investigate the frequency and causes of delayed presentation among patients with newly-diagnosed breast cancer. The study aimed to identify the extent of diagnostic delays, explore the underlying reasons, and examine associations with socio-demographic factors. The research involved 242 participants, primarily from low-income, uneducated backgrounds, with the majority (52.06%) aged between 41 and 60 years. Data was collected through face-to-face interviews and medical record reviews after obtaining informed consent. Diagnostic delay was defined as a period of 90 days or more between the onset of symptoms and the initiation of medical treatment by expert physicians. Findings revealed that nearly half of the patients (46.28%) experienced delays exceeding three months, with an average delay duration of 5.18 months. Stage II breast cancer was the most common diagnosis (56.6%). A statistically significant association was found between diagnostic delay and patients' socio-economic status as well as the stage at which cancer was diagnosed [3]. To address the critical issue of delayed breast cancer diagnosis, this study aims to enhance the performance of existing diagnostic approaches by validating deep learning–based models using real-world ultrasound and mammography datasets collected from multiple hospitals. This study utilizes datasets from mammography and ultrasound imaging to investigate the application of machine learning algorithms in breast cancer detection. Numerous machine learning models, including CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, were utilized and evaluated using key performance indicators such as accuracy, F1 score, weighted average, and macro average. The results show that CNN-based models are very successful, with mammography datasets outperforming ultrasound in terms of performance. In particular, CNN achieved the best accuracy for mammograms at 97%, followed by ResNet101 at 95% and ResNet50 at 94%.

*Corresponding author

With an accuracy of 89%, AlexNet outperformed other models for ultrasound data.

II. Related Works

Breast cancer detection has been extensively studied using machine learning (ML) and deep learning (DL) techniques, with numerous studies focusing on optimizing classification performance and early detection. Over the past decade, algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression, Naïve Bayes, and ensemble methods have been widely applied to both tabular and imaging datasets, demonstrating varying levels of effectiveness based on the nature of the data and preprocessing techniques employed [8], [20], [21].

Bazazeh and Shubair [9] conducted one of the earlier comparative studies on breast cancer detection using supervised ML techniques, including RF, SVM, and Bayesian Networks (BN). Their analysis highlighted the high predictive capability of BN, achieving a recall and precision of 97.1% and 97.2%, respectively, while RF exhibited optimal receiver operating characteristic (ROC) performance of 99.9%. The study emphasized the importance of leveraging historical labelled data for predictive modelling in clinical applications, demonstrating that even traditional ML techniques can provide high accuracy when appropriately tuned. Similarly, Mohammed et al. [22] and Bhise et al. [19] reviewed various ML and DL approaches, confirming the consistent use of SVM, KNN, RF, Logistic Regression, and Naïve Bayes in breast cancer detection. These studies collectively highlighted the strengths and weaknesses of each algorithm: SVM performs well in high-dimensional feature spaces but is less efficient for large or noisy datasets; KNN is conceptually simple and robust but computationally intensive for large datasets due to repeated distance calculations; RF effectively handles high-dimensional, categorical, and missing data, although hyperparameter tuning is essential to avoid overfitting; Logistic Regression is effective for binary classification but assumes linearity between independent and dependent variables; and Naïve Bayes is simple and fast, but its assumption of feature independence and the zero-frequency problem can limit real-world applicability.

In the context of image-based datasets, Charan et al. [4] utilized a large Kaggle breast cancer image dataset comprising 7,858 samples across four magnification levels. They applied ML classifiers, including SVM, KNN, RF, Logistic Regression, and Naïve Bayes, with feature extraction performed using convolutional and pooling operations. While competitive performance was achieved, SVM faced scalability issues on large, imbalanced image datasets. Selvathi and Poornila [24] further demonstrated that careful selection of classifiers and feature extraction methods is crucial when combining deep learning networks with traditional ML algorithms to optimize predictive performance.

Sharma et al. [7], [16] applied RF, KNN, Naïve Bayes, Logistic Regression, and Decision Tree classifiers on the Wisconsin Diagnostic Breast Cancer dataset. KNN achieved the highest accuracy of 95.90%, indicating its robustness for structured tabular data. Moreover, the study demonstrated that incorporating Principal Component Analysis (PCA) for dimensionality reduction and appropriate data preprocessing significantly enhances SVM performance. Specifically, SVM accuracy increased from 65.78% without preprocessing to 98.02% after PCA and Standard Scaling, and achieved 97.36% on test subsets after fine-tuning grid search parameters. These results underscore the importance of preprocessing and hyperparameter optimization in improving classifier performance.

Joshi and Mehta [10] applied image processing techniques along with ML classifiers on a set of 209 mammography images from 50 patients, using train-test-validation splits of 70:20:10. Relevance Vector Machine (RVM) achieved 97% accuracy, demonstrating that probabilistic sparse models can be highly effective even with reduced feature sets. However, the study noted that RVM is less frequently applied to breast cancer datasets, having been primarily utilized for the detection of lymphoma and leukemia [11].

Khuriwal and Mishra [12] combined deep learning with traditional classifiers (KNN, SVM, Decision Tree, and RF) on the mini-MIAS Mammographic Database. They applied comprehensive preprocessing steps including digitization, noise and artifact removal, background suppression, and pectoral muscle removal. Among the classifiers, RF achieved the highest accuracy of 98.89% in differentiating fatty and dense tissues, while SVM remained limited to binary classification, reflecting its well-known constraint in multi-class or high-dimensional imaging scenarios. Kumar et al. [13] further extended deep learning approaches by training a Convolutional Neural Network (CNN) on 70% of the MIAS dataset (322 images) using stochastic gradient descent with momentum (SGDM). Despite subdividing the dataset into seven classes (six abnormal), the model achieved only 65% accuracy for normal-versus-abnormal classification, highlighting the challenge of multiclass prediction in medical imaging.

Xiao et al. [14] proposed an unsupervised deep learning approach, employing feature extraction to reduce image dimensionality, followed by classification using stacked autoencoder–SVM (SAE-SVM). This approach demonstrated that combining unsupervised feature learning with supervised classification can improve predictive capability while reducing computational overhead. Halim et al. [15] implemented parallel AI models for early breast cancer detection, achieving faster results of simultaneously using multiple models. However, the study noted an increase in computational complexity due to model parallelism.

Ahmed et al. [17] analyzed the Wisconsin Breast Cancer dataset, containing 699 records (458 benign, 241 malignant), applying SVM, KNN, Naïve Bayes, RF, and Logistic Regression. All classifiers achieved over 95% accuracy, confirming the robustness of these algorithms for traditional tabular datasets. Vinayak et al. [18] applied ML techniques on cytological data from UCI, where Artificial Neural Networks (ANN) achieved 97.37% accuracy. Ensemble methods such as stacking and voting also produced competitive results, emphasizing the value of combining multiple classifiers for reliable predictions.

Ranjan et al. [6] highlighted the effectiveness of ensemble methods, including Gradient Boosting, AdaBoost, and XGBoost, with RF achieving perfect accuracy (100%), illustrating the advantages of ensemble learning while acknowledging limitations in real-world medical imaging applications. Easttom et al. [5] evaluated traditional ML models using the R

programming framework, finding that SVM achieved 94.71% accuracy but was constrained in high-dimensional imaging contexts.

The recent trend is the adoption of deep learning-based architectures such as CNNs, which automatically extract features from images without extensive preprocessing. CNNs can effectively differentiate between malignant and benign tumors by filtering and learning important image parameters, offering flexibility and robustness in image-based analysis. Contemporary implementations, often using Keras as a backend, have demonstrated superior accuracy and efficiency in automated breast cancer detection, forming the foundation of modern diagnostic systems [26], [27].

Overall, the reviewed literature demonstrates that both traditional machine learning and deep learning approaches have achieved promising performance in breast cancer diagnosis [23], [25]. Conventional classifiers, such as SVM, KNN, Random Forest, Logistic Regression, and Naïve Bayes, have demonstrated high accuracy on structured datasets, particularly when combined with effective preprocessing, feature selection, and dimensionality reduction techniques. However, these methods largely rely on handcrafted features and are sensitive to noise, class imbalance, and variations across imaging modalities. Recent studies increasingly favor deep learning models, especially Convolutional Neural Networks (CNNs), due to their superior capability in automatically extracting features and learning representations from complex medical images. Despite these advancements, most existing works are limited to a single imaging modality—primarily mammography or ultrasound—and focus mainly on classification, with limited attention to cross-modal analysis or comprehensive performance evaluation across diverse architectures. Furthermore, comparative studies involving multiple deep learning models under a unified experimental framework remain scarce. These limitations highlight the need for a robust and modality-aware diagnostic framework that systematically evaluates state-of-the-art deep learning architectures on both ultrasound and mammography data, while emphasizing reliable performance metrics beyond accuracy alone. Addressing these research gaps can significantly enhance diagnostic confidence and clinical applicability in automated breast cancer detection systems.

III. Methodology

The overall architecture of the proposed system is illustrated in Fig. 1. The framework begins with dataset acquisition, followed by image collection and preprocessing to enhance image quality and standardize input dimensions. The preprocessed images are then split into training and testing subsets. Deep learning models are trained using the training dataset, while the testing dataset is used to evaluate generalization performance. Model effectiveness is assessed using performance validation metrics, training and validation loss curves, and visual inspection of classification outcomes. Ultimately, the system generates a binary classification output indicating whether a case is benign or malignant. Fig. 1 provides a high-level representation of the complete workflow, highlighting the interaction between data preprocessing, model training, and evaluation components within the proposed framework.

While Fig. 1 presents the generalized architecture of the proposed system, it does not explicitly illustrate how different

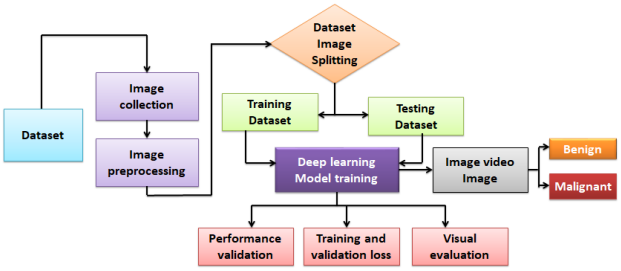


Fig. 1. System architecture

imaging modalities are handled independently. To address this, a more detailed experimental workflow is introduced in the next stage, focusing on modality-wise dataset splitting, model selection, and evaluation strategies for ultrasound and mammography images.

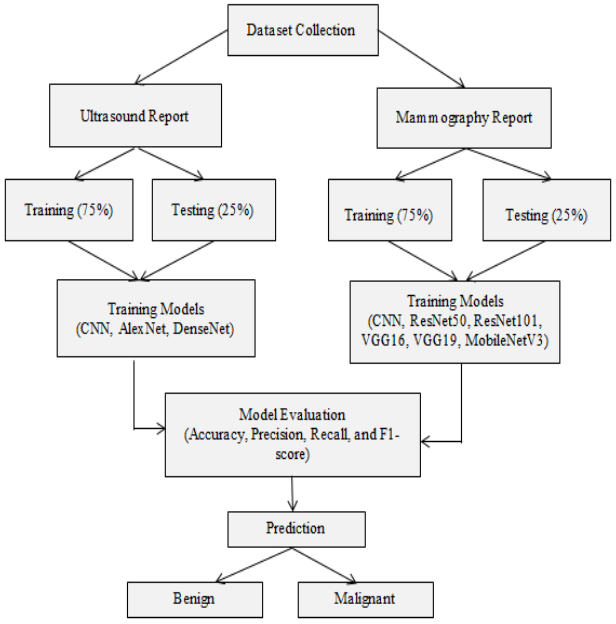


Fig. 2. Dataset-specific training and evaluation pipeline for ultrasound and mammography-based breast cancer classification.

The detailed experimental pipeline is illustrated in Fig. 2, outlining the dataset-specific model development process. The collected data are categorized into two imaging modalities: ultrasound reports and mammography reports. Each dataset is divided into 75% training and 25% testing subsets to ensure sufficient learning and reliable evaluation. For the ultrasound dataset, three models—CNN, AlexNet, and DenseNet—are trained to capture texture-based and low-contrast features commonly present in ultrasound images. In contrast, the mammography dataset is evaluated using CNN, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, enabling an in-depth comparison of shallow, deep, and lightweight architectures on high-resolution mammographic data.

A. Data Collection

The dataset used in this study was compiled from a combination of publicly available online repositories and offline clinical sources to ensure data diversity and reliability. The dataset consists of two primary imaging modalities: ultrasound and mammography. Ultrasound image data were collected from open-source platforms, including Kaggle, Hugging Face, and the UCI Machine Learning Repository.

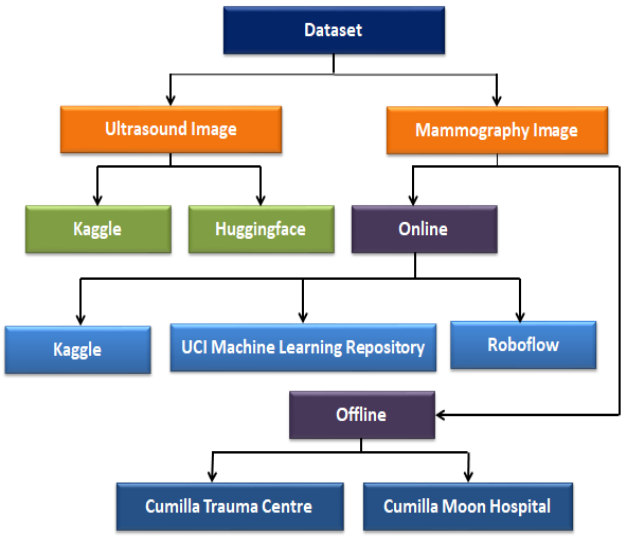


Fig. 3. Dataset collection process

Mammography images were obtained from online repositories, including the UCI Machine Learning Repository and Roboflow. To enhance clinical relevance, additional imaging data were collected from offline healthcare facilities, namely Cumilla Trauma Centre and Cumilla Moon Hospital. All offline data were acquired in compliance with institutional ethical guidelines, and patient-identifiable information was removed to maintain privacy and confidentiality. In this study, we have collected data from various sources, including Kaggle, UCI Machine Learning Repository, Roboflow, Hugging Face, and others.

Nearly 13k data have been collected. The aggregation of raw data from multiple online sources is used in the study (see Fig. 4). It showcases four publicly available platforms: Kaggle, UCI Dataset, Roboflow, and Hugging Face, each serving as an individual data source. All these sources are connected through a common pipeline that converges into a single repository labelled Raw Dataset. The study utilizes datasets collected from multiple public repositories and clinical sources, which may introduce variability due to differences in imaging protocols, acquisition devices, and annotation practices. Such source-specific biases and domain shifts can affect model generalization across different data distributions. While the proposed models demonstrate strong performance on the available datasets, further investigation using domain adaptation techniques and standardized datasets is necessary to improve robustness in real-world clinical settings. This diagram represents the initial stage of the data collection process, in which heterogeneous datasets from different platforms are combined into a unified raw dataset prior to further preprocessing, cleaning, and analysis.

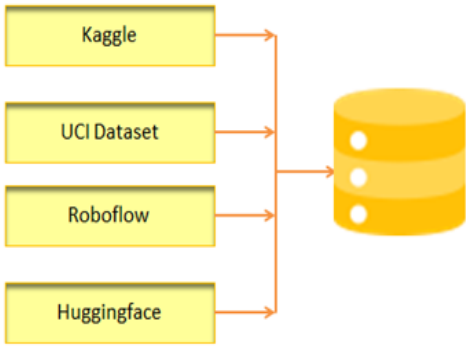


Fig. 4. Aggregation of raw dataset from multiple online sources.

The dataset used in this research consists of labelled mammographic images categorized into two clinically significant classes: Benign and Malignant, as illustrated in the figure. Benign images represent non-cancerous abnormalities, and malignant images correspond to confirmed cancerous cases. Fig. 3 describes the whole process of dataset collection. The mammograms exhibit variations in shape, texture, and tissue density, which are critical features for effective breast cancer detection. This diversity makes the dataset well-suited for training and evaluating machine learning models, as it enables them to learn discriminative patterns associated with different breast conditions. The dataset plays a crucial role in improving the robustness and accuracy of the proposed ML-based breast cancer detection system. The dataset used in this study was collected from two medical institutions in Bangladesh: Cumilla Moon Hospital and Cumilla Trauma Centre. A total of over 1,000 mammographic images were obtained with proper authorization and anonymization to ensure patient privacy. The collected data include cases classified as benign or malignant based on clinical diagnoses provided by experienced medical professionals. These real-world clinical images reflect a wide range of breast tissue characteristics, imaging conditions, and pathological variations, making the dataset representative of practical diagnostic scenarios (see Fig. 5).

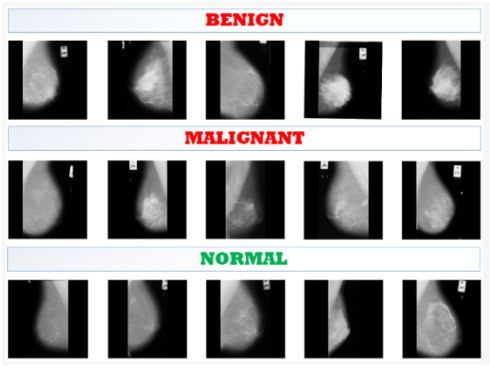


Fig. 5. Mammogram samples: Benign, malignant, and normal [4].

This dataset forms the foundation for training, validation, and performance evaluation of the proposed machine learning–based breast cancer detection models.

B. Image Preprocessing

Several essential preprocessing operations were performed internally within the convolutional neural network (CNN) models as part of the learning pipeline. The input images were first resized to match the fixed input dimensions required by the network architectures. Pixel intensity scaling and normalization were handled automatically within the model to stabilize gradient propagation and accelerate convergence during training. Convolutional layers learned hierarchical feature representations directly from the raw images, progressively capturing low-level features such as edges and textures, followed by higher-level semantic patterns.

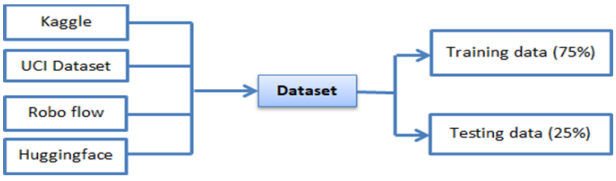


Fig. 6. Data preparation

Fig. 6 depicts the data preparation steps of this system, where, after data collection, a dataset was created. Then, the dataset splitting was done. The processed data fit the model. Here, data is gathered from different sources, which include ultrasound reports of the breast, and it is a very common report in breast cancer detection, where the patient's report is evaluated. Data is also collected from a medical-based website. This dataset was partitioned into two subsets for model development and evaluation. Specifically, 75% of the data was allocated for training, while the remaining 25% was reserved for testing. This data splitting strategy was employed to ensure effective model learning while enabling unbiased performance evaluation on unseen data.

Additionally, normalization layers embedded within the networks, such as batch normalization, mitigated internal covariate shift by standardizing intermediate feature distributions during training. Regularization mechanisms, including dropout layers, were employed to reduce overfitting and improve model generalization. Where applicable, data augmentation operations, such as random rotation, flipping, scaling, and intensity variation, were incorporated into the training pipeline and applied dynamically to input images. By integrating all preprocessing-related operations within the CNN framework, the proposed approach enables end-to-end learning, minimizes manual intervention, and enhances robustness across diverse imaging conditions.

C. Model Development

In this research, a convolutional neural network (CNN) is used as a baseline model to automatically classify breast images as benign or malignant, using both ultrasound and mammography datasets. CNNs are especially well-suited for medical images because they can learn hierarchical features directly from raw pixels, reducing the need for manual feature extraction. When a breast image is fed into the CNN, the network assigns learnable weights to small local regions, enabling it to capture important patterns such as mass boundaries, texture changes, and variations in tissue density. Compared to traditional machine learning methods, CNNs require less preprocessing and are highly effective at learning discriminative features from complex medical images.

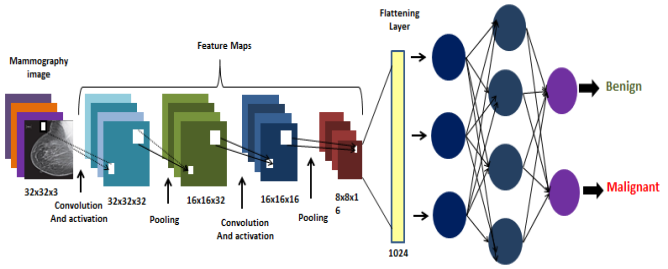


Fig. 7. Layer-wise architecture of the proposed CNN for mammography image classification. The network comprises three convolutional blocks with 32, 64, and 128 filters, followed by batch normalization, max-pooling, flattening, and fully connected layers with dropout regularization. The model was trained on a learning rate of 0.001 and 10 epochs. The final output layer performs binary classification into benign and malignant classes.

The CNN architecture used in this study follows a clear workflow, as shown in Fig. 7. It starts with an input layer that accepts preprocessed ultrasound or mammography images. These images then pass through several convolutional layers with learnable kernels, which automatically extract features ranging from low-level patterns like edges and contours to higher-level details such as lesion-specific textures. Nonlinear activation functions are applied after each convolution to give the network more flexibility in representing complex patterns. Pooling layers follow, reducing the size of the feature maps while keeping the important information, which helps lower computational costs and makes the model more robust to shifts or small changes in the images. The resulting feature maps are then flattened and passed through fully connected layers, which combine the learned features to make predictions. Finally, an output layer with a softmax activation produces probability scores for the benign and malignant classes.

On the ultrasound dataset, the baseline CNN achieved an accuracy of 85%, indicating that it can effectively capture basic texture and structural information. For mammography images, the CNN performed even better, reaching 97% accuracy, with an F1-score of 97%, macro-average of 96%, and weighted-average of 97%. These results highlight that a well-designed shallow CNN can handle high-quality mammograms, in which lesion boundaries and tissue density differences are more pronounced.

To further evaluate performance, AlexNet was also tested on these medical images. Thanks to its large convolutional kernels and deep fully connected layers, AlexNet is particularly good at capturing coarse texture features. On ultrasound images, AlexNet achieved the highest accuracy of 89%, outperforming both the baseline CNN and DenseNet. This suggests that AlexNet's design is well-suited for ultrasound data, where speckle noise and low-contrast patterns make feature extraction more challenging.

AlexNet is one of the early deep convolutional neural networks, with five convolutional layers and three fully connected layers. It is particularly good at capturing coarse texture features in medical images, especially ultrasound scans. Its large

kernels, ReLU activations, dropout layers, local response normalization, and data augmentation all work together to make learning more robust, even in the presence of speckle noise and low-contrast patterns. Designed to take full advantage of GPU processing, AlexNet outperformed baseline CNNs and DenseNet on ultrasound images, achieving 89% accuracy. This success comes from its deeper architecture, faster convergence, strong regularization, and ability to efficiently handle large datasets.

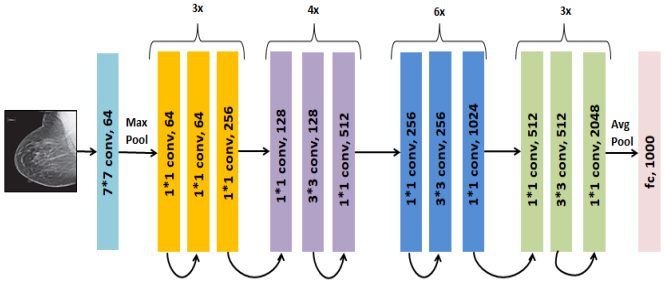


Fig. 8. Customized ResNet-50 architecture employed for mammographic image classification. The model consists of an initial 7×7 convolutional layer with 64 filters followed by max pooling, and four residual stages containing 3, 4, 6, and 3 bottleneck blocks, respectively. Each stage uses 1×1 and 3×3 convolutions with increasing feature dimensions (64–2048) to extract hierarchical features from mammograms. Global average pooling and a fully connected layer with 1000 neurons are used to produce the final classification output.

ResNet50 uses residual learning to make training deep networks more stable and efficient. On mammography images, it achieved an accuracy of 94%, with an F1-score of 93%, a macro-average of 93%, and a weighted-average of 94%. While its residual connections improved feature learning compared to VGG-based models, ResNet50 did not outperform the baseline CNN, suggesting that adding more depth isn't always necessary for mammogram classification.

The ResNet50 architecture, shown in Fig. 8, is structured into four main parts: convolutional layers, identity blocks, convolutional blocks, and fully connected layers. The convolutional layers first extract features from the input image. The identity and convolutional blocks then process and refine these features, while the fully connected layers make the final classification.

ResNet101 takes this a step further by increasing the network's depth to capture more complex patterns. It achieved slightly better results on mammography images, with an accuracy of 95% and F1 and average metrics around 95–96%. However, the performance gain over ResNet50 was marginal, suggesting that simply increasing depth can introduce redundancy without significantly improving diagnostic accuracy.

Fig. 9 shows the architecture of a customized ResNet-101–based CNN used for classifying mammography images. The process starts with a 7×7 convolutional layer with 64 filters, followed by max pooling to capture low-level features and reduce the spatial size of the image. The network then goes through four residual stages made up of identity blocks, which allow the deep architecture to train efficiently. The Conv2, Conv3, Conv4, and Conv5 stages contain 3, 4, 23, and 3 blocks, with channel depths increasing from 256 to

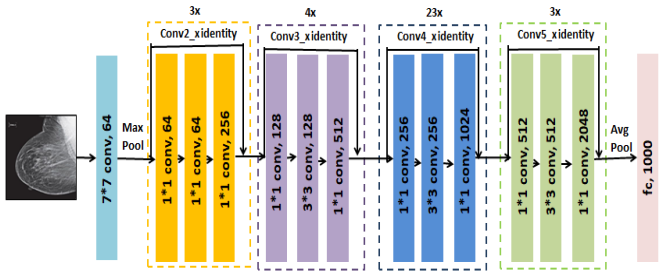


Fig. 9. Architecture of the customized ResNet-101–based CNN used for mammography image classification. The network begins with a 7×7 convolution and max-pooling, followed by four residual stages comprising Conv2 (3×), Conv3 (4×), Conv4 (23×), and Conv5 (3×) identity blocks with increasing channel depths of 256, 512, 1024, and 2048, respectively. Global average pooling and a fully connected layer with 1000 units are used for final classification.

2048. This structure helps the network progressively learn richer and more complex feature representations. After these stages, global average pooling condenses the information into a compact feature vector, which is then passed to a fully connected layer with 1000 units for classification.

MobileNetV3, on the other hand, is a lightweight model designed for efficiency. On mammography images, it achieved 92% accuracy, an F1-score of 91%, and balanced macro and weighted averages of 92%. While its performance is slightly lower than that of deeper residual models, MobileNetV3 offers a good balance between accuracy and computational cost, making it ideal for resource-constrained settings.

For experimental consistency, all deep learning models were trained using the Adam optimizer for 10 epochs with a learning rate of 0.001. The experiments revealed interesting trends across imaging types. Ultrasound images benefit from models that capture texture-rich features, such as AlexNet and DenseNet. Mammography images, with their clearer structures and higher contrast, consistently achieved better classification performance. Interestingly, the baseline CNN outperformed some of the deeper architectures on mammograms, suggesting that adding more depth doesn't always yield better results. Finding the optimal model complexity is key to balancing accuracy and efficiency.

D. Training and Evaluation

The training and evaluation process was designed to ensure robust learning, fair comparison among models, and reliable performance assessment across different imaging modalities. The collected dataset consists of ultrasound reports and mammography images, which were handled independently to preserve modality-specific characteristics. Each dataset was partitioned into training and testing subsets using a 75:25 split, ensuring sufficient data for model learning while retaining an unbiased test set for evaluation.

Prior to training, all images were resized to a uniform input dimension compatible with the selected network architectures. Pixel intensity values were normalized to improve convergence and stabilize the learning process. Labels were assigned according to clinical diagnosis, categorizing images into benign

and malignant classes. Data shuffling was applied to minimize ordering bias during training.

For the ultrasound dataset, deep learning models, including CNN, AlexNet, and DenseNet, were trained. In contrast, the mammography dataset was evaluated using various architectures, including CNN, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3. Transfer learning was employed where applicable by initializing networks with pre-trained weights and fine-tuning the higher layers to adapt to breast cancer imaging features. This strategy reduced training time and improved generalization, particularly for datasets with limited samples.

Model training was carried out using the Adam optimizer, with categorical cross-entropy as the loss function. The learning rate was empirically tuned to achieve stable convergence, and early stopping was applied to prevent overfitting by monitoring validation loss. Batch size and number of epochs were selected based on computational efficiency and model performance during preliminary experiments:

$$P = \frac{TP}{TP + FP} \tag{1}$$

To evaluate classification effectiveness, multiple performance metrics were employed, including accuracy, precision, recall, and F1-score, which collectively provide a balanced assessment of model reliability, especially in the presence of class imbalance. Accuracy measures overall correctness, while precision and recall quantify false-positive and false-negative behavior, respectively. where TP and FP represent true positives and false positives, respectively. The recall R evaluates the ratio of correctly detected damaged regions to the total number of actual damaged regions:

$$R = \frac{TP}{TP + FN} \tag{2}$$

where, FN denotes false negatives.

The F1-score, which is the harmonic mean of precision and Recall, provided a balanced assessment of detection:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{3}$$

The F1-score is a harmonic mean of precision and recall, providing a comprehensive indicator of diagnostic performance. Comparative evaluation was conducted across all trained models to identify the most effective architecture for each imaging modality. The final prediction stage classifies input images as either benign or malignant, thereby supporting clinical decision-making with consistent and interpretable results. The evaluation framework ensures that the proposed system is both robust and generalizable, demonstrating its suitability for real-world breast cancer diagnosis applications.

IV. Results and Discussion

This section analyzes the performance of the proposed models using both quantitative evaluation metrics and qualitative visual inspection. Quantitative results highlight numerical performance differences, while qualitative analysis, as shown in confusion matrices, provides insight into classification consistency and misclassification patterns.

A. Result Analysis

1) Quantitative performance analysis: The quantitative evaluation of the framework is presented through performance metrics obtained from mammography and ultrasound datasets, as summarized in Table I and Table II, respectively. The models were assessed using accuracy, precision, recall, and F1-score to ensure a comprehensive evaluation under class imbalance conditions.

TABLE I. Performance Comparison of Deep Learning Models on the Mammography Dataset (Precision, Recall, F1-Score are Written as Benign/Malignant Format).

Model	Precision	Recall	F1-Score	Accuracy
CNN	97% / 96%	95% / 97%	96% / 97%	97%
ResNet50	92% / 95%	94% / 93%	93% / 94%	94%
ResNet101	95% / 95%	94% / 96%	95% / 96%	95%
VGG16	91% / 90%	87% / 93%	89% / 92%	90%
VGG19	91% / 90%	87% / 93%	89% / 91%	90%
MobileNetV3	88% / 95%	94% / 90%	91% / 92%	92%

For the mammography dataset (Table I), the CNN model achieved the highest overall performance with an accuracy of 97%, demonstrating balanced precision and recall for both benign and malignant classes (precision: 97%/96%, recall: 95%/97%). This indicates strong discriminatory capability and stable generalization. ResNet101 performed closely, achieving 95% accuracy of 95%, benefiting from residual learning to capture complex mammographic patterns. ResNet50 achieved a slightly lower accuracy of 94%, suggesting that deeper residual connections in ResNet101 contribute to improved feature representation.

VGG-based architectures (VGG16 and VGG19) exhibited comparatively lower accuracy (90%), primarily due to reduced recall for benign cases (87%), indicating higher false-negative rates. MobileNetV3 achieved an accuracy of 92%, offering a favorable trade-off between performance and computational efficiency, particularly with high malignant recall (94%), which is clinically significant for cancer detection.

TABLE II. Accuracy Comparison of Models on the Ultrasound Dataset.

Model	Accuracy
CNN	85%
AlexNet	89%
DenseNet	88%

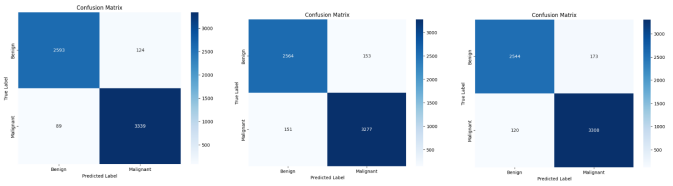
For the ultrasound dataset (Table II), overall classification performance was lower compared to mammography, reflecting the inherent challenges of ultrasound imaging, such as noise, speckle artifacts, and low contrast. Among the evaluated models, AlexNet achieved the highest accuracy of 89%, followed by DenseNet at 88% and CNN at 85%. The improved performance of AlexNet and DenseNet indicates their effectiveness in capturing texture-based features from ultrasound images.

Overall, the quantitative results confirm that mammography-based classification outperforms ultrasound-based classification across all evaluated models. The superior

performance of CNN and ResNet101 on mammographic images highlight the importance of deep hierarchical feature learning for high-resolution medical imaging. The performance reporting in this study is not fully symmetric across modalities. For the mammography dataset, detailed evaluation metrics including precision, recall, F1-score, and accuracy are reported, whereas for the ultrasound dataset, only accuracy values are presented. This difference arises from the original experimental setup and the available evaluation outputs. Although this limits a strictly uniform cross-modal comparison, the results still provide meaningful insights into model performance trends across different imaging modalities. Future work will focus on extending the evaluation to include comprehensive metrics for all datasets to enable a fully balanced comparison. These findings demonstrate the robustness and clinical relevance of the proposed deep learning framework for automated breast cancer diagnosis.

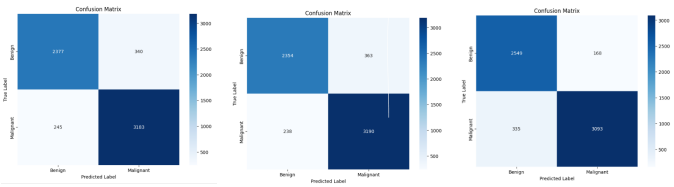
2) Qualitative performance analysis: Qualitative analysis is performed to visually evaluate the classification behavior of the proposed deep learning models beyond numerical performance metrics. Confusion matrices provide intuitive insight into how well each model distinguishes between benign and malignant mammographic images, highlighting both correct predictions and patterns of misclassification.

As illustrated in Fig. 10 and Fig. 11, the confusion matrices of all six models exhibit strong diagonal dominance, indicating a high number of correctly classified samples for both benign and malignant classes. The CNN-based model shows a relatively balanced distribution of true positives and true negatives, with only a limited number of false classifications. This suggests that the model effectively captures essential features required for mammogram classification.



(a) CNN (b) ResNet50 (c) ResNet101

Fig. 10. Confusion matrix for CNN, ResNet50 and ResNet 101 on the mammography image.



(a) VGG16 (b) VGG19 (c) MobileNetv3

Fig. 11. Confusion matrix for VGG16, VGG19 and MobileNetv3 on the mammography image.

The MobileNetV3 confusion matrix demonstrates improved detection of malignant cases, as reflected by a higher count of true positives, although a small number of benign cases are miResNet-based architectures, these models exhibit a slightly increased misclassification of benign cases, which may be attributed to their architectural complexity and the achieves the most consistent separation between the two classes, highlighting the advantage of deeper residual learning in minimizing classification errors.

The VGG16 and VGG19 models also present strong performance, with a high proportion of malignant samples correctly identified. However, compared to ResNet-based architectures, these models exhibit slightly higher misclassification rates of benign cases, which may be attributed to their greater architectural complexity and the absence of residual connections.

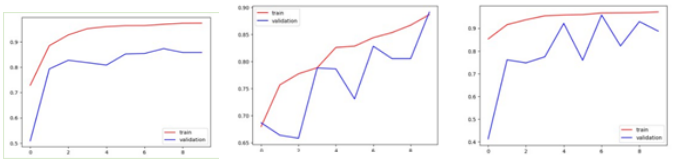
Overall, the visual analysis confirms that all evaluated models are capable of effective mammogram classification, with ResNet101 and CNN providing the most balanced and reliable predictions. The qualitative observations are consistent with the quantitative results, reinforcing the effectiveness of deep learning approaches for breast cancer detection.

B. Discussion

This study presents a comparative analysis of multiple CNN-based architectures for breast cancer classification, utilizing both ultrasound and mammography imaging modalities. A clear performance gap is observed between the two imaging modalities. Overall, mammography-based models consistently outperform ultrasound-based models, achieving higher accuracy, F1-score, macro-average, and weighted-average values. Although the proposed framework achieves strong classification performance, the current study does not evaluate computational complexity, training overhead, inference latency, memory consumption, or deployment cost. These factors are important for real-world clinical deployment, where resource efficiency and scalability are critical. Therefore, a comprehensive evaluation of these aspects is considered as a potential direction for future work. This outcome can be attributed to the higher contrast, clearer structural details, and more consistent appearance of lesions in mammography images compared to ultrasound images, which often suffer from speckle noise and low contrast.

1) Ultrasound-based model performance: For the ultrasound dataset, AlexNet achieved the highest accuracy of 89%, followed closely by DenseNet at 88%, while the baseline CNN achieved 85% accuracy. As shown in Fig. 12(b) and Fig. 12(c), these models demonstrate faster convergence during training but exhibit noticeable fluctuations in validation accuracy. Such instability reflects the inherent variability and noise present in ultrasound images.

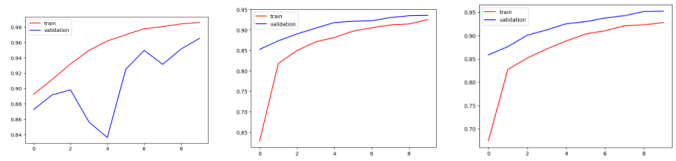
AlexNet's superior performance suggests that architectures with larger convolutional kernels and strong texture-extraction capabilities are more effective for ultrasound imaging. DenseNet also performed competitively due to its dense feature reuse mechanism; however, the increased architectural complexity did not translate into significant performance gains. The baseline CNN, while simpler, showed limited capability in modeling the complex texture patterns characteristic of ultrasound images [see Fig. 12(a)].



(a) CNN (b) AlexNet (c) DenseNet

Fig. 12. Training and validation accuracy curves for CNN, AlexNet, and DenseNet on the ultrasound dataset.

2) Mammography-based model performance: In contrast, mammography-based experiments yielded substantially higher and more stable performance across all models. Notably, the baseline CNN achieved the best overall performance, with an accuracy of 97%, an F1-score of 97%, a macro-average of 96%, and a weighted-average of 97%. As illustrated in Fig. 13(a), the CNN model exhibits smooth convergence and minimal divergence between training and validation curves, indicating strong generalization and low overfitting.



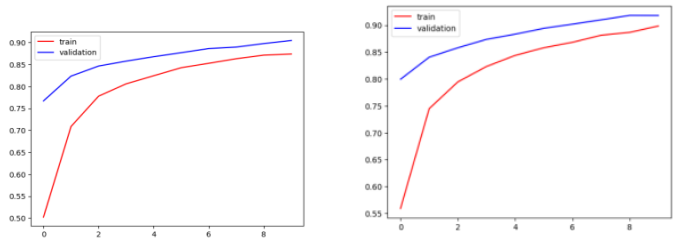
(a) CNN (b) ResNet50 (c) ResNet101

Fig. 13. Training and validation accuracy curves for CNN, ResNet50, and ResNet101 on the mammography dataset.

Among deeper architectures, ResNet101 achieved 95% accuracy, marginally outperforming ResNet50 (94%). Fig. 13(b) and Fig. 13(c) show that residual networks benefit from stable training dynamics due to skip connections, yet the additional depth of ResNet101 provides only limited improvement over ResNet50. This suggests diminishing returns from increased network depth when applied to high-quality mammographic data.

3) Performance of VGG and mobilenet architectures: VGG16 and VGG19 achieved moderate performance with accuracies of 90%, as shown in Fig. 14(a) and corresponding results in the table. Although VGG architectures are effective feature extractors, their large parameter count and lack of residual connections may contribute to overfitting and reduced generalization in medical imaging tasks.

MobileNetV3 achieved an accuracy of 92%, offering a favorable balance between performance and computational efficiency. As shown in Fig. 14(b), MobileNetV3 demonstrates stable learning behavior, making it a practical choice for deployment in resource-constrained clinical environments, despite slightly lower accuracy compared to deeper residual models. These findings emphasize that architecture selection should be guided by data characteristics rather than depth alone, particularly in medical imaging applications. The strong performance and stable generalization of the baseline CNN on mammography images highlight its potential for integration into computer-aided diagnosis (CAD) systems. Such systems



(a) VGG16 (b) MobileNetv3

Fig. 14. Training and validation accuracy curves for VGG16 and MobileNetv3 on the mammography dataset.

can assist radiologists by providing reliable second opinions, improving early detection rates, and ultimately enhancing patient outcomes.

4) Comparative analysis with existing studies: Previous papers on breast cancer detection have predominantly relied on traditional machine learning techniques, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forests, and Logistic Regression, often using handcrafted features and benchmark datasets, including the Wisconsin Breast Cancer Dataset or limited histopathological image collections. While these approaches have demonstrated reasonable classification accuracy, their effectiveness is constrained by manual feature engineering, limited scalability to large image datasets, and reduced robustness when handling complex visual patterns and class imbalance.

In contrast, the proposed study adopts a deep learning–based approach that eliminates the need for handcrafted feature extraction by enabling automatic hierarchical feature learning through convolutional neural networks. Unlike earlier works that focus on a single imaging modality, this research performs a comparative analysis of both ultrasound and mammography images, providing a more comprehensive evaluation of modality-specific diagnostic performance.

Moreover, several prior studies report results based on a single or limited number of classifiers, whereas the proposed framework evaluates multiple CNN architectures, including CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3. This comprehensive model comparison enables a deeper understanding of how network depth, residual learning, and lightweight architectures impact breast cancer classification accuracy.

Another key limitation of earlier research is the reliance on public or small-scale datasets, which may not accurately reflect real-world clinical scenarios. The proposed study addresses this gap by validating models using real-world ultrasound and mammography datasets, thereby enhancing the clinical relevance and generalizability of the findings.

Performance-wise, the proposed framework achieves superior accuracy and F1-score, particularly in mammography-based classification, where the CNN and ResNet101 models outperform traditional machine learning methods reported in prior literature. Additionally, the use of macro and weighted-average metrics ensures a more reliable evaluation under conditions of class imbalance, a consideration often overlooked

in earlier studies.

Overall, the proposed research advances existing work by offering a modality-aware, deep learning–driven, and clinically relevant framework that demonstrates improved diagnostic performance and robustness, making it more suitable for practical deployment in computer-aided breast cancer diagnosis systems.

V. Conclusion

This study has presented a deep learning–based framework for automated breast cancer classification using ultrasound and mammography images. Multiple CNN architectures, including CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, were systematically evaluated to analyze modality-specific performance. The experimental results demonstrate that deep learning models effectively learn discriminative features directly from medical images, eliminating the need for handcrafted feature extraction and outperforming traditional machine learning approaches.

Mammography-based classification achieved superior performance compared to ultrasound imaging, with CNN and ResNet101 producing the highest accuracy and F1-scores, highlighting the effectiveness of deeper architectures in capturing complex structural patterns. The present study does not include robustness evaluation under challenging conditions such as noisy inputs, image corruption, or domain shift scenarios. As a result, the stability and generalization ability of the proposed models under real-world variations in medical imaging data remain unverified. Future research will focus on addressing these issues by incorporating robustness testing and domain adaptation techniques. Although ultrasound-based results were comparatively lower due to inherent image noise and low contrast, AlexNet and DenseNet exhibited stable learning behavior and competitive performance. Overall, the proposed framework offers a robust and clinically relevant computer-aided diagnostic solution that can aid radiologists in early breast cancer detection and informed decision-making.

Future research will focus on expanding the framework using larger multi-institutional datasets to improve generalization. Incorporating advanced data augmentation techniques, class imbalance handling methods, attention mechanisms, and tumor localization or segmentation modules may further enhance diagnostic accuracy and interpretability. Additionally, real-time deployment and clinical validation remain crucial directions for the practical adoption in healthcare settings.

References

[1] "Breast cancer," LCH, https://labaidcancer.com/cancer-details/Breast-Cancer (accessed Dec. 19, 2025). .

[2] H. Australia, "Breast cancer," Healthdirect Australia. Accessed: Dec. 20, 2025. [Online]. Available: https://www.healthdirect.gov.au/breast-cancer

[3] J. Ferdause, A. K. M. S. Kadir, and M. Haque, "Understanding diagnostic delays among newly diagnosed breast cancer patients at a tertiary cancer care center in a low-middle-income country like Bangladesh," *Medicine*, vol. 104, e41775, 2025, doi: 10.1097/MD.0000000000041775.

[4] S. Charan, J. Khan, and K. Khurshid, "Breast cancer detection in mammograms using convolutional neural network," in Proc. IEEE Int. Conf. Mathematics and Engineering Technologies (iCoMET), Mar. 2018, doi: 10.1109/ICOMET.2018.8346384.

[5] C. Easttom, S. Thapa, and J. Lawson, "A comparative study of machine learning algorithms for use in breast cancer studies," in Proc. 2020 IEEE Annu. Comput. Commun. Workshop Conf. (CCWC), Jan. 2020, pp. 412–416, doi: 10.1109/CCWC47524.2020.9031266.

[6] M. Ranjan, A. Shukla, K. Soni, S. Varma, and M. Kuliha, "Cancer prediction using random forest and deep learning techniques," in Proc. 2022 Int. Conf. Communication Systems and Network Technologies (CSNT), 2022, pp. 227–231, doi: 10.1109/CSNT54456.2022.9787608.

[7] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in Proc. 2018 Int. Conf. on Trends in Electronics and Informatics (CTEMS), 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.

[8] T. Jain, V. K. Verma, A. Yadav, and A. Jain, "Supervised machine learning approach for the prediction of breast cancer," in Proc. IEEE Int. Conf. System, Computation, Automation and Networking (ICSCAN), MVIT Puducherry, India, Jul. 3–4, 2020, doi: 10.1109/ICSCAN49426.2020.9262403.

[9] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in Proc. 2016 5th Int. Conf. Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, United Arab Emirates, Dec. 6–8, 2016, pp. 1–4, doi: 10.1109/ICEDSA.2016.7818560.

[10] A. Joshi and A. Mehta, "Mammogram image classification using various machine learning algorithms," in Proc. 2022 7th Int. Conf. on Computing, Communication and Security (ICCCS), Nov. 2022, pp. 106–110, doi: 10.1109/ICCCS55188.2022.10079398.

[11] B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," in Proc. 2016 IEEE Int. Conf. Computational Intelligence and Computing Research (ICCIC), Chennai, India, Dec. 2016, pp. 1–5, doi: 10.1109/ICCIC.2016.7919576.

[12] N. Khuriwal and N. Mishra, "Breast cancer detection from histopathological images using deep learning," in Proc. 2018 3rd Int. Conf. and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Nov. 2018, pp. 1–4, doi: 10.1109/ICRAIE.2018.8710426.

[13] A. Kumar, R. Patra, and A. Ghosh, "Model selection for predicting breast cancer using supervised machine learning algorithms," in Proc. 2020 IEEE 1st Int. Conf. for Convergence in Engineering (ICCE), Kolkata, India, 2020, pp. 320–324, doi: 10.1109/ICCE50343.2020.9290578.

[14] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning," in Proc. 2018 37th Chinese Control Conference (CCC), July 2018, pp. 9428–9433, doi: 10.23919/ChiCC.2018.8483140.

[15] E. Halim, P. P. Halim, and M. Hebrard, "Artificial intelligent models for breast cancer early detection," in Proc. 2018 Int. Conf. Information Management and Technology (ICIMTech), Indonesia, 2018, pp. 517–521, doi: 10.1109/ICIMTech.2018.8528140.

[16] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in Proc. 2018 Int. Conf. on Computational Techniques in Engineering and Management (CTEMS), 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.

[17] M. R. Ahmed, M. A. Ali, J. Roy, S. Ahmed, and N. Ahmed, "Breast cancer risk prediction based on six machine learning algorithms," in Proc. 2020 IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, Dec. 2020, pp. 1–5, doi: 10.1109/CSDE50874.2020.9411572.

[18] V. A. Telsang and K. Hegde, "Breast cancer prediction analysis using machine learning algorithms," in Proc. 2020 Int. Conf. on Communication, Computing and Industry 4.0 (C2I4), Dec. 2020, pp. 1–5, doi: 10.1109/C2I451079.2020.9368911.

[19] S. Bhise, S. Gadekar, A. S. Gaur, S. Aswale, et al., "Detection of breast cancer using machine learning and deep learning methods," in Proc. 2022 3rd Int. Conf. Intelligent Engineering and Management (ICIEM), Apr. 2022, doi: 10.1109/ICIEM54221.2022.9853080.

[20] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast cancer detection using K-nearest neighbor machine learning algorithm," in Proc. 2016 9th Int. Conf. on Developments in eSystems Engineering (DeSE), Liverpool, United Kingdom, Aug.–Sep. 2016, pp. 35–39, doi: 10.1109/DeSE.2016.8.

[21] T. Mehejabin, F. Rahman, S. Yeasmin, M. Sarkar, and F. Rahman, "Identification of most relevant breast cancer miRNA using machine learning algorithms," in Proc. 2020 IEEE 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–6, doi: 10.1109/ICCCNT49239.2020.9225624.

[22] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in Data Mining and Big Data, Y. Tan, Y. Shi, and M. Tuba, Eds., Communications in Computer and Information Science, vol. 1234, Springer, Singapore, Jul. 2020, pp. 108–117, doi: 10.1007/978-981-15-7205-0-10.

[23] F. Rahman, T. Mehejabin, S. Yeasmin, and M. Sarkar, "A comprehensive study of machine learning approach on cytological data for early breast cancer detection," in Proc. 2020 IEEE 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–6, doi: 10.1109/ICCCNT49239.2020.9225448.

[24] D. Selvathi and A. Poornila, "Performance analysis of various classifiers on deep learning network for breast cancer detection," in Proc. 2017 Int. Conf. on Signal Processing and Communication (ICSPC), Coimbatore, India, Jan. 2017, pp. 359–363, doi: 10.1109/CSPC.2017.8305869.

[25] A. R. Vaka, B. Soni, and S. Reddy K., "Breast cancer detection by leveraging machine learning," ICT Express, vol. 6, no. 4, pp. 320–324, Dec. 2020, doi: 10.1016/j.icte.2020.04.009.

[26] X. Zhou, Y. Li, R. Gururajan, G. Bargshady, X. Tao, R. Venkataraman, P. D. Barua, and S. Kondalsamy-Chennakesavan, "A new deep convolutional neural network model for automated breast cancer detection," in Proc. 2020 7th IEEE Int. Conf. on Behavioural and Social Computing (BESC), Bournemouth, United Kingdom, 5–7 Nov. 2020, pp. 1–4, doi: 10.1109/BESC51023.2020.9348322.

[27] L. Qian, J. Bai, Y. Huang, D. Q. Zeebaree, A. Saffari, and D. A. Zebari, "Breast cancer diagnosis using evolving deep convolutional neural network based on hybrid extreme learning machine technique and improved chimp optimization algorithm," Biomedical Signal Processing and Control, vol. 87, p. 105492, Jan. 2024, doi: 10.1016/j.bspc.2023.105492.