

Selective Ensemble and Transfer Learning for Breast Cancer Classification from Mammograms

Mayank Kapadia^{1,2} and Parita Oza^{2,*}

¹ Department of Applied Data Science, San Jose State University, San Jose, USA

² Institute of Technology, Nirma University, Ahmedabad, India

Email: mayank.kapadia@sjsu.edu and 19bce092@nirmauni.ac.in (M.K.); parita.prajapati@nirmauni.ac.in (P.O.)

*Corresponding author

Abstract—Breast cancer remains one of the most dangerous and prevalent diseases in women worldwide, so early and precise diagnostic methods are very important. Mammogram analysis using deep learning has been very promising, but single-model solutions are prone to poor generalization and suffer from overfitting, particularly when class-imbalanced or small medical datasets are involved. We present here a framework by integrating deep transfer learning with selective ensemble learning for improved classification of breast cancer from mammograms. We trained four state-of-the-art convolutional neural networks, i.e., DenseNet121, EfficientNetV2B0, MobileNetV2, and XceptionNet, on two datasets, the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSDM) and Digital Mammography Dataset for Breast Cancer Diagnosis (DMID). The models that performed best were then combined to work together, boosting overall accuracy. Our results demonstrate that the ensemble model outperforms the networks in isolation, achieving higher accuracy, lower loss, and improved generalization. The research also compares the energy consumption and carbon footprint of each model and determines EfficientNetV2B0 to be the most environmentally friendly architecture. Statistical inferences through Analysis of Variance (ANOVA) and Quantile-Quantile (QQ) plots were also performed on the models. The proposed method offers a clinically viable, efficient, and environmentally friendly approach to automatic breast cancer diagnosis.

Keywords—breast cancer, mammograms, deep learning, transfer learning, ensemble learning, carbon footprint

I. INTRODUCTION

One of the most prevalent and fatal diseases affecting women globally is breast cancer [1]. Early and accurate diagnosis is paramount in improving patient prognosis and reducing mortality rates. Mammography, the first-line imaging procedure for breast cancer screening, has significantly transformed early detection, but interpretation proves to be a laborious process due to subtle visual signals, tissue heterogeneity of the breasts, and the limitations of human judgment [2].

New developments in deep learning have opened up new prospects for medical image analysis. Convolutional

Neural Networks (CNNs), in particular, have demonstrated phenomenal ability to learn hierarchical feature representations from raw input directly [3]. They enable the automatic and accurate detection of abnormalities in mammograms, often identifying subtle patterns not visible using conventional image processing techniques [4]. Nonetheless, single-model approaches are prone to generalization, robustness, and overfitting problems, especially when handling small and class-imbalanced medical datasets [5, 6].

To address these challenges, transfer learning and ensemble learning proved to be effective solutions. Transfer learning enables using the features of large pre-trained networks and adapting them to domain-specific tasks. Ensemble learning is a method of aggregating different models to reduce variance and improve predictive stability [7, 8]. In this paper, an end-to-end mammogram classifier is presented that includes deep transfer learning and selective ensemble methods to improve classification accuracy and generalization.

Despite the fact that deep learning is proven highly efficient in the task of image classification, mammogram interpretation continues to be a problem. Training samples are not enough, and classes are imbalanced, leading models to overfit. Patient anatomy and image quality variability also lead to inconsistency. Single-model pipelines may fail to properly represent the complex and dynamic patterns of breast imaging, which can affect the reliability of diagnostic outcomes [9, 10].

Motivation and Challenges: The motivation for this work is to improve the diagnostic accuracy and reliability of breast cancer classification through mammogram-based methods with state-of-the-art deep learning approaches. Various challenges define the problem landscape:

- Limited and heterogeneous data: Mammographic datasets tend to be small in size with varied abnormalities, thus it will be challenging for a model to generalize well.
- Overfitting risks: Deep models learned from small sets can overfit, especially for noisy or unrepresentative data.

- Model variability: All deep networks are not created equal; some models perform better with specific imaging patterns or tasks.
- Clinical deployment needs: Clinical use in healthcare settings demands models to be accurate, efficient, and interpretable [11].

The resolution of these problems requires a multi-pronged solution involving model adaptation, augmentation, testing, and selective combination.

Research Contribution: The paper introduces a deep learning system that utilizes transfer learning and ensemble methods to improve mammogram classification. The contributions are:

- Utilization of pre-processing techniques like orientation correction, CLAHE, and artifact removal to improve image quality and homogeneity.
- Use of deep transfer learning models, DenseNet121, EfficientNetB0, MobileNetV2, XceptionNet and train on curated mammographic datasets.
- Use of the Ensemble approach to combine models to improve model performance.
- Calculation of carbon and energy footprint to measure the ecological cost of individual models.

II. LITERATURE REVIEW

Jabeen *et al.* [12] suggested a hybrid model, BC2NetRF, which marries enhanced deep learning features with feature selection by the Regula-Falsi method. While they experienced a boost in the classification performance, the computational burden induced by the equilibrium-Jaya optimization rendered it less viable for real-time clinical application. Similarly, Atban *et al.* [13] employed classical machine classifiers augmented with deep features, yet their framework was not stable when the system was tested on heterogeneous datasets.

Peirera *et al.* [14] proposed a dialectical optimization method towards thermographic image-based lesion classification. Though the approach achieved excellent accuracy, it suffered from convergence instability in the process of detecting optimal feature subsets. In contrast, Oza *et al.* [8] enhanced the state-of-the-art using transfer learning as well as ensemble methods on mammography images, incorporating diverse data augmentations for greater model generalizability. The article pointed out the synergy between pretrained networks and ensemble combination and provided a baseline for future comprehensive frameworks.

From the clinical support perspective, Pesapane *et al.* [15] developed an Artificial Intelligence (AI) software for microcalcification analysis with AlexNet, ResNet18, and ResNet34. They were trained on a database of nearly 2000 manually annotated mammograms and achieved high sensitivity, albeit for microcalcification-specific tasks. Complementing this, Sahu *et al.* [16] gave an overview of deep learning approaches for mammography and declared that Transfer Learning (TL) models largely outperform traditional Deep Learning models, especially in the small data regime.

In a multi-modal diagnostic setup, Amethiya *et al.* [17] compared Machine Learning models with biosensors for

the detection of early breast cancer. Their comparative analysis of Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Learning Machine-Radial Basis Function (ELM-RBF) proved the viability of integrating biological signal processing with image-based AI classification, though without ensemble learning integration.

Mohapatra *et al.* [18] compared the deep architectures like AlexNet, VGG16, and ResNet50 evaluated on the mini-DDSM dataset. They indicated the efficiency of transfer learning, especially in improving deeper model performances like ResNet50. Similarly, Ozsahin *et al.* [19] used Breast Imaging-Reporting and Data System (BIRADS) descriptors for interpreting mammograms using interpretable ML models. Their findings reflected that logistic regression gave the best Area Under Curve (AUC), while SVM provided more accuracy and precision.

Drives hybrid learning paradigms ahead, Jenefa *et al.* [20] introduced a hybrid of MobileNetV2 and LSTM networks for breast cancer classification. The model achieved an excellent classification accuracy of over 98%, validating the capability of combining spatial and sequential feature extraction mechanisms.

Lastly, Al-Antari *et al.* [21] proposed an integrated Computer-Aided Diagnosis (CAD) system integrating YOLO-based lesion detection with deep learning classifiers like CNN, ResNet-50, and InceptionResNet-V2. The proposed system obtained classification accuracy ranging from 88.74% to 97.50% and showed outstanding performance on different evaluation metrics on DDSM and IN breast databases.

Another study proposed a transfer learning-based approach for breast cancer classification using ultrasound images, combining pre-trained models such as VGG16, ResNet50, and MobileNetV2 with Long-Short Term Memory (LSTM). Their VGG16-based model achieved very high performance (F1-Score up to 99%) and further used Synthetic Minority Over-sampling Technique (SMOTE) Tomek for class balancing, along with Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME) for model explainability. They also validated their results using cross-validation and confidence interval analysis, reporting strong and stable performance compared to several state-of-the-art models [22]. One more study proposed a deep learning-based approach for lung cancer classification using Computed Tomography (CT) images from the Kaggle dataset. The authors modified the DenseNet201 model for feature extraction and applied feature selection methods followed by different machine learning classifiers to improve performance. Their system achieved very high accuracy (up to 100%) and showed strong results in 5-fold cross-validation, highlighting the potential of combining deep learning and machine learning for reliable lung cancer diagnosis [23].

Though these are significant contributions, the majority of existing frameworks utilize fixed TL models or simple ensemble approaches without model selection. The limitation of these is covered in the current work by

utilizing a selective ensemble methodology that, besides the utilization of multiple pre-trained Convolutional Neural Network (CNN) architectures, emphasizes strategic model selection based on validation performance. The resultant combined framework aims to enhance classification dependability, suppress overfitting, and promote flexibility across mammographic datasets.

III. MATERIALS AND METHODS

A. Datasets

The proposed approach was evaluated on two benchmark datasets employed in this paper: the recently published DMID dataset and the publicly published CBIS-DDSM dataset.

DMID (Digital Mammography Dataset for Breast Cancer Diagnosis Research) [24, 25] is a recently published dataset holding over 500 mammographic images covering a great diversity of breast abnormalities, including masses, microcalcifications, architectural distortions, and asymmetries. The data is thoughtfully organized into five directories: (i) Digital Imaging and Communications in Medicine (DICOM) Images (ii) TIFF Images with TIFF-formatted renditions thereof, (iii) Reports with annotations from radiologists, (iv) Pixel-level Annotation with high-quality lesion-level segmentation masks, and (v) Region of Interest (ROI) Masks with region-of-interest masks for targeted analysis. Furthermore, a *Metadata.csv* file rounds out the data with essential clinical and image-level descriptors, enabling structured data access and facilitating downstream analysis.

CBIS-DDSM [26] is one of the most established datasets widely used in breast cancer diagnosis research. It is a thoroughly curated subset of the original DDSM, specifically designed to help in the development and testing of Computer-Aided Detection (CAD) systems. CBIS-DDSM contains expertly curated and labeled mammographic cases, including 891 mass cases and 753 calcification cases. It also supplies each patient with pathologic labels and segmentation masks that make it a gold mine to train and test machine learning models for diagnostic imaging applications.

This study presented a binary classification task: distinguishing between benign and malignant lesions. CBIS-DDSM relied on pathology-confirmed labels (benign and malignant) from the dataset. In DMID, for every image, class labels are mentioned in a Comma-Separated Values (CSV) file called metadata. To ensure a fair comparison between the two datasets, we first aligned (harmonized) their label definitions; class 0 to benign lesion and class 1 to malignant lesion. Although the datasets were collected independently, some labels had slightly different names or category boundaries. We carefully reviewed the annotation guidelines for both datasets and created a unified label mapping.

The DMID dataset used in our study has approximately 47% correspond to malignant lesions and 53% to benign lesions, indicating a near-balanced distribution [25]. The CBIS-DDSM dataset consists of 3103 images with

abnormal structure, along with other types. Considering only malignant and benign cases used in our study, the distribution is roughly 44% malignant images, 43% benign [27]. To avoid data leaking, the dataset was divided into 70% training and 30% testing subgroups for each patient. In CBIS-DDSM, where numerous mammograms may belong to the same patient, all images from that patient were assigned solely to either the training or testing sets.

B. Image Preprocessing Pipeline

The DMID and CBIS-DDSM datasets were both utilized for training deep learning models after passing through a standardized and rigidly designed preprocessing pipeline. This is summarized in Fig. 1, where the raw mammographic images are progressively converted into clean and model-ready inputs.

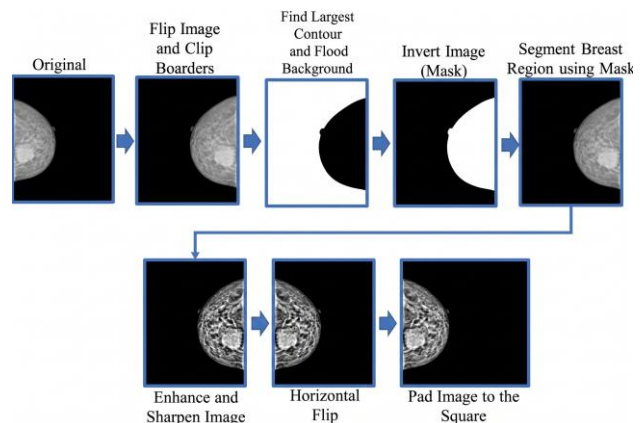


Fig. 1. Preprocessing pipeline for mammogram images including orientation correction, border cropping, CLAHE enhancement, mask-based segmentation, and resizing.

First, all images are conditionally flipped horizontally to ensure the breast area is always on the right-hand side. This spatial normalization enables proper identification of a flooding seed point, which is necessary for segmenting the anatomical region of interest. Border cropping is then employed to eliminate non-informative white borders. Specifically, 10% of the image is cropped from the left, 0% from the right, and 5% from the top and bottom borders.

The second step is finding the largest contour in the image as a surrogate for the breast area. A flood-fill operation on this contour is followed by inversion of the result to create a binary mask. The mask is applied to the original image to segment the breast area and remove anatomically irrelevant background structures and artifacts, increasing anatomical focus.

For increasing local contrast, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied with a clip limit of 2 and an 8×8 tile grid size. Both of these parameters were empirically selected for their ability to render diagnostic features more visible without introducing noise. A sharpening filter that enhances edges is subsequently applied following CLAHE for outlining tumor boundaries and rendering lesion contours more perceptible.

Finally, all the images are padded and resized to a shared input resolution of 224×224 pixels. This renders them uniform in batches and aligns with the input layers of all the deep learning models used in the research.

C. Model Architectures

The taxonomy model presented here is built upon four state-of-the-art Convolutional Neural Network (CNN) models, each of which is transferred through adaptation via transfer learning. These DenseNet121, MobileNetV2, EfficientNetB0, and XceptionNet models were selected based on their proven capability in feature representation, computational efficiency, and scalability.

1) DenseNet121

DenseNet (Densely Connected Convolutional Networks) [28] counteracts the growing complexity of visual data by offering dense connectivity between layers. As opposed to conventional CNNs where each layer is linked to its immediate successor, as shown in Fig. 2 DenseNet connects each layer to all the following layers in a feed-forward manner. The dense connectivity is amenable to feature reuse, enhances gradient flow, and avoids the vanishing gradient issue. Layers within each dense block involve summing feature maps of all the preceding layers to achieve a diverse and rich representation. The design of architecture supports improved parameter updates throughout training as well as improved convergence in learning.

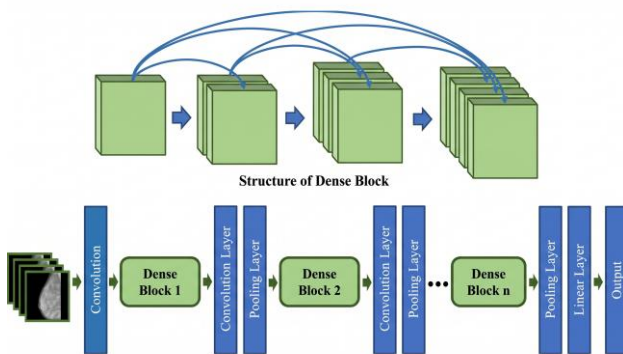


Fig. 2. Architecture of Densenet model with dense block.

2) MobileNetV2

As shown in Fig. 3, MobileNetV2 [29] is architected specifically for mobile and embedded platforms with high accuracy and low computational expense. Depthwise separable convolutions are the core of the architecture, where spatial filtering (depthwise convolution) and channel-wise feature mixture (pointwise convolution) are decoupled. This factorization drastically reduces the parameters and computation required. Inverted residual patterns with linear bottlenecks and projection layers are utilized in MobileNetV2, which enable fast inference with adequate representational capacity. Modularity of its design enables it to be deployable in resource-constrained environments.

3) EfficientNetB0

EfficientNetB0 [30] introduces a paradigm shift in scaling models through its compound scaling strategy that scales network depth, width, and resolution in a similar proportion. With this technology, in conjunction with the Mobile Inverted Bottleneck Convolution (MBConv) and squeeze-and-excitation modules, the architecture is capable of achieving higher accuracy using fewer parameters. EfficientNetB0 is the minimal variant of the EfficientNet series and is optimized for both performance and efficiency in real-world applications. Its architecture achieves a balance between expressiveness and compactness, making it suitable for scalable deployment scenarios.

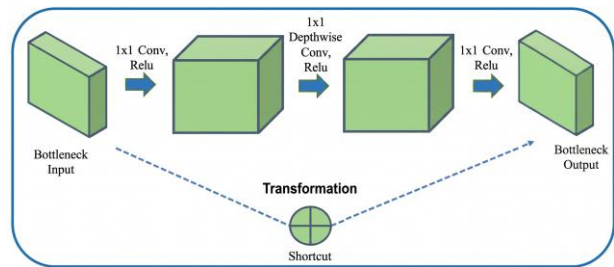


Fig. 3. Basic architecture of Mobilenet model.

The MBConv block constitutes the basic building block, linking light convolutions with dynamic feature re-balancing. As compared with standard scaling methods, EfficientNetB0 obtains better performance on image classification tasks with very reduced resource utilization, making it a landmark design in CNNs.

4) XceptionNet

Xception, short for “Extreme Inception” [31], is a Convolutional Neural Network (CNN) architecture proposed by Francois Chollet in 2017. It is a significant architectural advance from traditional CNNs and Inception modules by adopting full reliance on depthwise separable convolutions, as shown in Fig. 4. This design paradigm enables the model to learn intricate hierarchical features with improved computational efficiency.

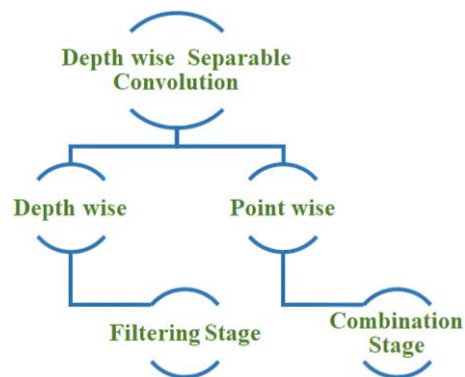


Fig. 4. Classification of layers in the Xception model.

Having its foundation on the conceptual design of the Inception architecture that uses a stack of parallel convolutional paths to generate multi-scale features, Xception replaces standard convolutions with an

alternative two-stage factorized process, as shown in Fig. 5. That is, it separates the convolution operation into (i) depthwise convolution that performs spatial filtering separately per channel, and subsequently (ii) pointwise convolution (11 kernel) that offers cross-channel interactions. Isolating concerns serves to greatly reduce the number of parameters and computational cost while keeping the model expressiveness intact.

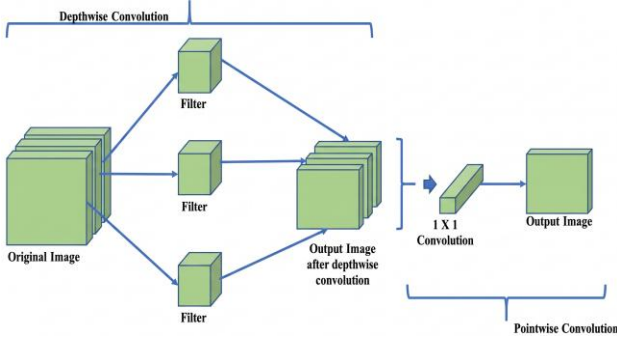


Fig. 5. Basic architecture of Xception model.

The Xception architecture is further enhanced by incorporating residual (skip) connections and a global average pooling layer. The residual paths ensure improved gradients flow across layers, which helps in overcoming the degradation problem commonly observed in deeper networks, and the pooling layer is helpful for spatial dimensionality reduction without reducing significant feature information.

From a performance perspective, Xception has demonstrated competitive and often state-of-the-art performance on several image classification benchmark problems. Its parameter-efficient and modular design also makes it particularly well-suited for deployment in settings where computational resources are scarce, such as mobile and embedded systems.

In brief, Xception shows a very effective architectural solution through unbinding spatial and channel-wise processing. That it can always be extremely accurate at using a lot fewer parameters makes it a general-purpose and computationally elegant solution for a multitude of various computer vision problems.

D. Transfer Learning Framework

In order to develop an efficient clinical decision support system for breast mass classification from mammograms, we adopt the transfer learning-based method that leverages the representational power of a number of state-of-the-art Convolutional Neural Networks (CNNs). Our approach blends public and private datasets to generate a diverse training corpus.

Four pre-trained models, DenseNet121, MobileNetV2, XceptionNet, and EfficientNetB0, are selected for initial testing because of their architectural heterogeneity and proven efficacy in medical imaging tasks. The models are adapted to the classification task by substituting the final output layer with a task-specific dense layer and retraining the network end-to-end or partially frozen based on its convergence behavior.

The entire transfer learning process is illustrated in Fig. 6, realizing the sequence from image preprocessing to model fine-tuning. All input images are normalized into a spatial size of 224×224 pixels to be the same as the input size of selected CNN architectures. Preprocessing is the removal of high- and low-intensity artifacts, image noise suppression, and contrast enhancement, as explained in the preprocessing section.

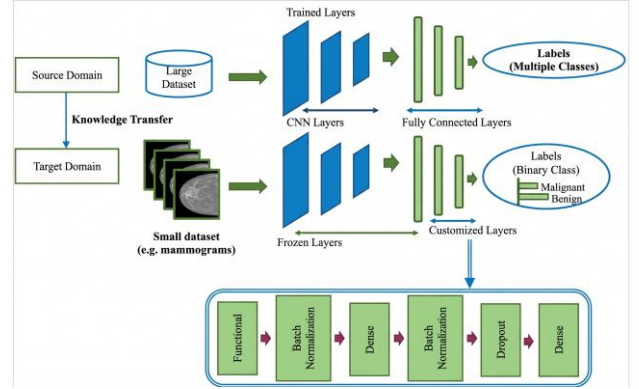


Fig. 6. Transfer learning workflow.

E. Training Setup and Hyperparameters

Every model is trained for 100 epochs with a batch size of 64. The Adam optimizer with a learning rate of 1×10^{-5} is utilized. The binary cross-entropy loss function is employed for classification because the task is binary in nature. The dataset is split randomly into a 70% training and 30% testing ratio, with an internal validation set carved out from the training subset to monitor learning and prevent overfitting.

For enhancing generalization performance and mitigating data scarcity, intensive data augmentation methods are used while training. These include horizontal flipping following clockwise rotation, and random image shifting and zooming procedures. These procedures enable the models to acquire features that are resistant to minor spatial distortions. Hyperparameters for training are given in Table I and were chosen empirically based on validation performance. Validation loss monitoring and early stopping are used during training to avoid underfitting, in addition to overfitting.

TABLE I. TRAINING CONFIGURATION AND SELECTED HYPERPARAMETERS

Parameter	Value
Optimizer	Adam
Learning Rate	1×10^{-5}
Batch Size	64
Epochs	100
Loss Function	Binary Cross-Entropy
Data Split	70% Training/30% Testing
Input Size	224×224
Augmentations	Flip, Rotation, Shift, Zoom

After preliminary testing, ensemble learning techniques, such as simple averaging, weighted averaging, and feature-level concatenation, are employed to merge

predictions of top-performing base models. In the feature-level concatenation approach, the final classification layers of the selected backbone networks were removed, and deep feature embeddings were extracted from their penultimate layers. These feature representations were concatenated into a unified feature vector, which was then passed to a newly introduced fully connected classification layer.

This final classifier was fine-tuned jointly, allowing the combined feature space to adapt to the binary classification task. This approach differs fundamentally from output-level ensemble techniques such as simple averaging or weighted averaging, which combine predicted class probabilities without enabling interaction between internal feature representations. Feature-level fusion allows complementary spatial and semantic features from different architectures to interact before classification, thereby enhancing discriminative power and robustness. Feature-level fusion improves discriminative capacity and robustness by allowing joint representation learning and interaction between complementary feature spaces, as opposed to output-level ensembles (simple or weighted averaging) that only combine prediction probabilities without learning cross-model interactions. The ensembles are then tested for accuracy, robustness, and convergence behavior gains.

F. Evaluation Metrics

The performance of the proposed deep learning models was evaluated using a systematic experimental setup. The data was split into 70% for training and 30% for testing. From the training set, a part was reserved for validation to monitor the model performance during training as well as against overfitting due to the application of early stopping measures.

To measure classification performance, several performance metrics were recorded at train and test time, including:

- **Accuracy:** The number of instances accurately predicted divided by the total number of samples.
- **Loss:** Computed using binary cross-entropy as an indicator of confidence of class probabilities from predictions.
- **Validation Metrics:** Accuracy and loss are tracked during training for every model as metrics to watch out for convergence and to determine overfitting.

After single-model training and testing, the models with higher validation accuracy were selected for ensemble combination. Ensembles were constructed using schemes such as simple averaging, weighted averaging, and feature concatenation. The ensemble models were then tested using the same metrics to ascertain improvement in generalization and predictability robustness.

Moreover, statistical validation was carried out on individual models using ANOVA for variance examination, and model performance consistency changes in significance. These included analysis of variance among model groups as well as residual visualization through boxplots and QQ plots.

The models were also assessed for their energy consumption and carbon emissions during training and inference, to evaluate their environmental sustainability and computational efficiency.

IV. EXPERIMENTAL RESULTS

A. Single Model Performance

Each of the deep learning models, individually DenseNet121, EfficientNetV2B0, MobileNetV2, and XceptionNet, was evaluated on the standard classification metrics across the training, validation, and test datasets. The training dynamics, as graphed in Figs. 7 and 8, show the accuracy and loss progression throughout the training epochs.

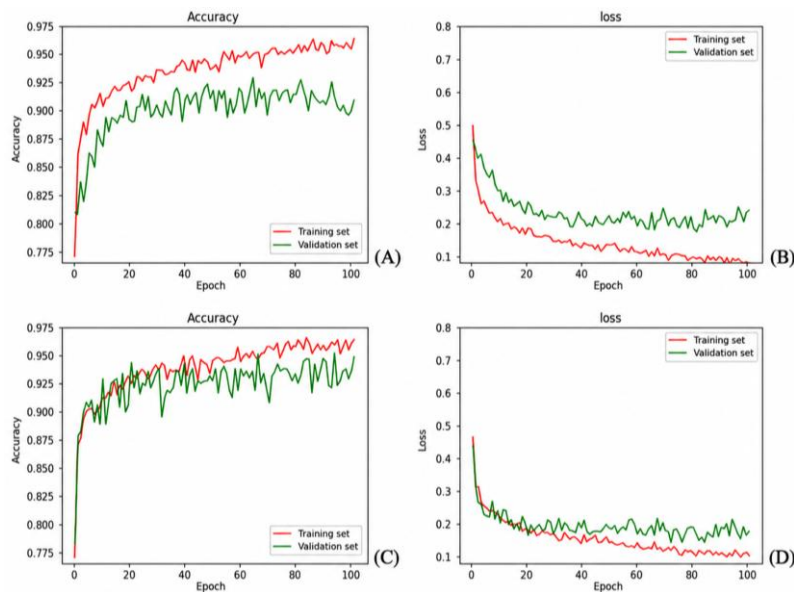


Fig. 7. Accuracy and loss curves for DenseNet and EfficientNet. (A) accuracy curves for DenseNet; (B) loss curves for DenseNet; (C) accuracy curves for EfficientNet; (D) loss curves for EfficientNet.

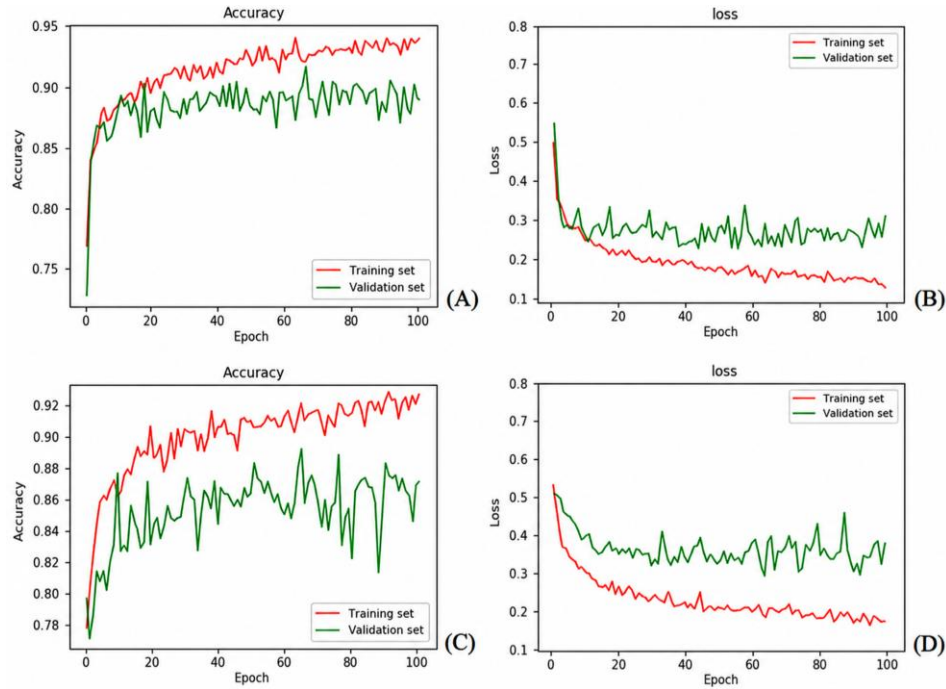


Fig. 8. Accuracy and loss curves for MobileNet and XceptionNet. (A) accuracy curves for MobileNet; (B) loss curves for MobileNet; (C) accuracy curves for XceptionNet; (D) loss curves for XceptionNet.

A detailed comparative summary is presented in Tables II and III. Table II lists accuracy results on test, validation, and training sets for final and early-stopped (optimized) models. Table III adds the corresponding loss values and the number of training epochs.

TABLE II. MODEL PERFORMANCE: ACCURACY AND LOSS METRICS (FINAL VS. OPTIMIZED)

Metric	Model	Final	Optimized
Test Accuracy	DenseNet121	0.9256	0.9167
	EfficientNetV2B0	0.8958	0.9018
	MobileNetV2	0.8482	0.8571
	XceptionNet	0.8542	0.8482
Validation Accuracy	DenseNet121	0.9103	0.9290
	EfficientNetV2B0	0.9196	0.9346
	MobileNetV2	0.8972	0.9047
	XceptionNet	0.8729	0.8935
Training Accuracy	DenseNet121	0.9654	-
	EfficientNetV2B0	0.9556	-
	MobileNetV2	0.9402	-
	XceptionNet	0.9266	-

According to Fig. 9, EfficientNetV2B0 consistently topped the performance, achieving a test accuracy of 90.18%, validation accuracy of 93.46%, and best validation loss (0.1862) while being consistent in both training and testing sessions. MobileNetV2 followed closely, with comparable accuracy and generalization. XceptionNet and DenseNet121, while possessing high train accuracy, manifested larger disparities between train and test measures, indicating poor generalization and possible overfitting on the given datasets. To achieve optimal model performance and prevent overfitting, an early stopping strategy was employed during training. This

technique continuously monitors the model's performance on the validation set and automatically halts the training process when no further improvement is observed, ensuring that the model generalizes well without wasting unnecessary computational resources.

TABLE III. MODEL PERFORMANCE: LOSS AND TRAINING EPOCHS

Metric	Model	Final	Optimized
Test Loss	DenseNet121	0.2262	0.2307
	EfficientNetV2B0	0.2710	0.2764
	MobileNetV2	0.3876	0.3505
	XceptionNet	0.3897	0.4131
Validation Loss	DenseNet121	0.2556	0.1926
	EfficientNetV2B0	0.2371	0.1862
	MobileNetV2	0.3144	0.2282
	XceptionNet	0.3694	0.2880
Training Loss	DenseNet121	0.0949	-
	EfficientNetV2B0	0.1212	-
	MobileNetV2	0.1422	-
	XceptionNet	0.1829	-
Epochs	DenseNet121	100	81
	EfficientNetV2B0	100	82
	MobileNetV2	100	56
	XceptionNet	100	81

B. Statistical Analysis

The ANOVA analysis shows no statistically significant differences among the evaluated models ($p > 0.05$) (see Table IV). While one model achieved marginally higher average performance, the difference is not statistically significant. This suggests that all models perform comparably under the tested conditions. Future studies with larger datasets or repeated trials may provide clearer evidence of performance differences.

TABLE IV. ONE-WAY ANOVA TEST RESULTS COMPARING PERFORMANCE ACROSS MODEL GROUPS

Source	Sum of Squares	df	F-value	p-value
C(Models)	0.004615	3	0.0104	0.9985
Residual	2.953438	20	-	-

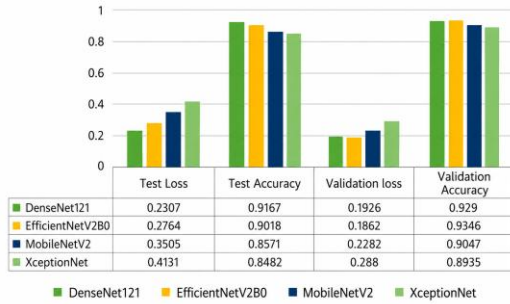


Fig. 9. Comparison of test and validation metrics across models.

Fig. 10 presents the box plot of validation accuracy obtained from multiple runs. The boxplot illustrates the median accuracy, interquartile range, and variability for each model. DenseNet121 and EfficientNetV2B0 demonstrate relatively stable and higher median accuracy compared to MobileNetV2 and XceptionNet, indicating more consistent performance. A limitation of this study is that Fig. 10 shows only validation accuracy and does not include loss values. While accuracy reflects overall model performance, loss can provide additional insight into model learning and stability. Future work could include both accuracy and loss distributions to give a more complete picture of model behavior.

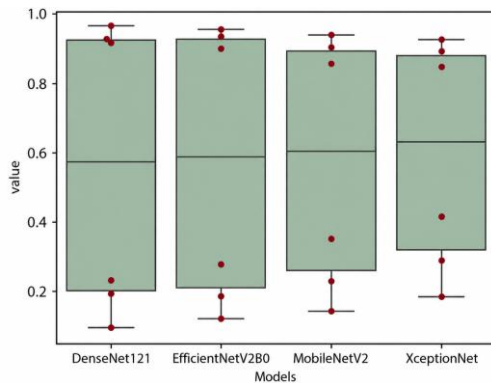


Fig. 10. Boxplot showing performance distribution by model type.

In addition, a QQ-plot was used to visually examine the normality assumption of residuals (see Fig. 11). The QQ

plot reveals noticeable deviations from the theoretical normal reference line, including a mild S-shaped pattern and clustering of residuals. These features suggest that the assumption of normality is not fully satisfied. Although the residuals are generally confined within ± 1.2 and no extreme outliers are observed, the presence of structured deviations indicates potential non-normality. Therefore, the model results should be interpreted with caution, and the robustness of the findings should not be attributed solely to the QQ plot assessment.

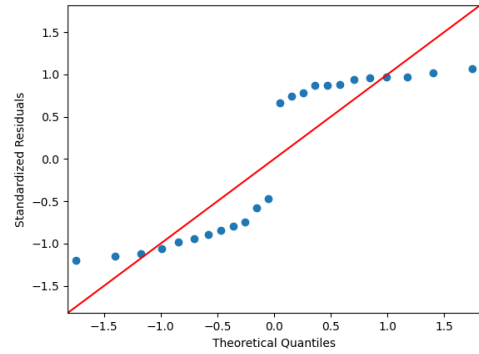


Fig. 11. QQ-plot for normality of residuals.

A. Energy & Carbon Footprint Trade-offs

In addition to performance, energy efficiency and environmental sustainability were also gauged. All the models' emission and power values are presented in Table V. Energy usage and emissions were measured using the CodeCarbon Python library, an open-source tool designed to track power consumption and estimate carbon footprint during machine learning/deep learning experiments. CodeCarbon monitors CPU power, Random-Access Memory (RAM) power, CPU energy, RAM energy, and energy usage during runtime. CodeCarbon automatically detected system specifications, geographic location, and grid carbon intensity. Energy consumption was calculated as the integral of measured power over execution time. Carbon where the emission factor corresponds to the electricity carbon intensity of the detected country/region. All experiments were conducted on a Linux system (Kernel 3.10.0-514.el7.x86_64) running Python 3.8.12, equipped with an Intel Xeon Gold 6139 processor operating at 2.30 GHz with 36 CPU cores and 92.9 GB of RAM. Model training and evaluation were performed using CPU-based computation.

TABLE V. CARBON FOOTPRINT AND ENERGY CONSUMPTION BY DL MODELS

Model	Duration (s)	Emissions	Emission Rate	CPU Power	RAM Power	CPU Energy	RAM Energy	Total Energy	Carbon Footprint
DenseNet121	7874.59	0.3266	0.0415	269.61	34.84	0.5646	0.0762	0.6408	0.2093
EfficientNetV2B0	1632.33	0.0397	0.0243	151.82	34.84	0.0621	0.0158	0.0779	0.0031
XceptionNet	6993.70	0.2056	0.0294	156.14	34.84	0.3357	0.0676	0.4033	0.0829
MobileNetV2	1681.87	0.0691	0.0411	259.20	34.84	0.1192	0.0163	0.1355	0.0094

EfficientNetV2B0 was the most sustainable model, consuming the lowest energy (0.078 kWh) and generating the lowest carbon footprint (0.0031 kgCO₂).

DenseNet121, though most accurate, consumed the highest resource cost. These results support the use of

model architectures that yield a quality trade-off between accuracy and energy efficiency.

With quantitative outcomes, Figs. 12 and 13 also present energy consumption curves of all models. Fig. 12 presents a hierarchical breakdown of carbon emissions, CPU/RAM energy, and overall power consumption. Fig. 13 even highlights model running times and power consumption. These graphs corroborate that EfficientNetV2B0 is the most environmentally friendly choice, and DenseNet121 incurs the highest carbon cost.

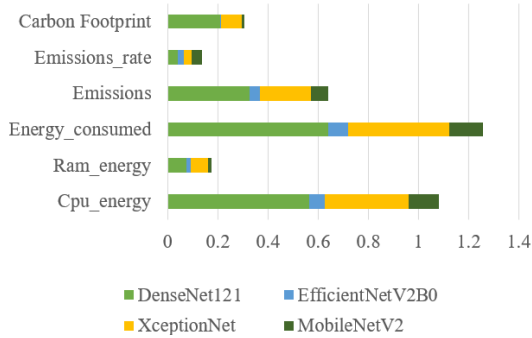


Fig. 12. Energy and carbon metrics across models.

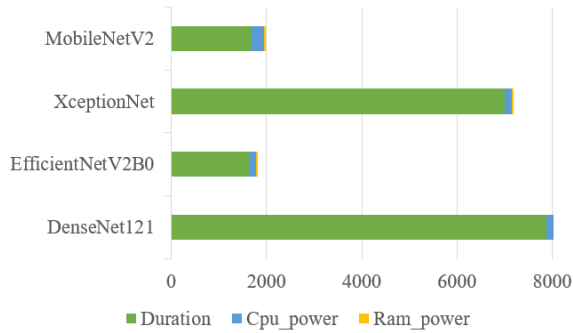


Fig. 13. Model duration and power consumption comparison.

B. Ensemble Strategy & Results

In order to leverage the performance of various models, ensemble learning methods were attempted. The best performing ensemble was one created by the concatenation of features of MobileNetV2 and EfficientNetV2B0. The selection of top-performing models for ensemble construction was guided by both predictive performance and sustainability considerations.

Specifically, models were evaluated based on test and validation accuracy, consistency across performance metrics, computational complexity, and environmental impact, including energy consumption and carbon footprint. This ensemble provided a validation accuracy of 90.97% and an accuracy of 91.19% in the test set, better than all the individual base models.

The ensemble model also demonstrated improved generalization with reduced test loss of 0.267 and reduced training loss of 0.1035, and high training accuracy of 97.14% (see Table VI). These results validate the hypothesis that model aggregation promotes robustness and classification fidelity in breast cancer diagnosis.

TABLE VI. PERFORMANCE COMPARISON OF EFFICIENTNETV2B0, MOBILENETV2, AND ENSEMBLE MODEL

Metric	EfficientNetV2B0	MobileNetV2	Ensemble Model
Test Loss	0.271	0.3876	0.267
Test Accuracy	0.8958	0.8482	0.9119
Validation Loss	0.2371	0.3144	0.2378
Validation Accuracy	0.9196	0.8972	0.9097
Training Loss	0.1212	0.1422	0.1035
Training Accuracy	0.9556	0.9402	0.9714

V. DISCUSSION

Comparison between models revealed that EfficientNetV2B0 was most consistent in performance, with a balance of accuracy and stability across all the metrics. MobileNetV2 had good performance with the lowest computational complexity. DenseNet121 and XceptionNet had good performance, but had a slightly higher variation between the validation and test phases.

Parameter complexity, as shown in Table VII, also played a role in the behavior of the models. XceptionNet, with over 33 million parameters, exhibited overfitting behavior. EfficientNetV2B0, however, reached a balance between model size (14 million parameters) and accuracy, showing the benefit of compound scaling. In contrast to previous work, such as Oza *et al.* [8], Our study introduces a feature-level concatenation strategy that combines representations from multiple deep learning models to improve classification performance. Additionally, we implement a customized image preprocessing pipeline, which enhances data quality and model generalization. Unlike the models evaluated in Ref. [8], we selected a distinct set of architectures (DenseNet121, EfficientNetV2B0, XceptionNet, and MobileNetV2) to explore diverse feature extraction capabilities.

TABLE VII. MODEL PARAMETER SUMMARY: TRAINABLE, NON-TRAINABLE, AND TOTAL PARAMETER COUNTS

Model	Trainable Parameters	Non-trainable Parameters	Total Parameters
DenseNet121	4,197,377	7,040,064	11,237,441
EfficientNetV2B0	8,195,585	5,922,384	14,117,969
XceptionNet	13,112,321	20,866,088	33,978,409
MobileNetV2	8,195,585	2,261,056	10,456,641

Despite the limited size of the DMID dataset, several measures were used to reduce overfitting and bias, in addition to typical data augmentation. Transfer learning reuses generalized feature representations from big datasets, lowering the risk of overfitting to restricted samples. Early stopping was used to prevent excessive training convergence. The selected ensemble approach improved robustness by combining model predictions, resulting in lower variance compared to single-model pipelines.

Preprocessing was used to handle variations in medical images, such as differences in contrast, lighting, and background. CLAHE improved local contrast to highlight subtle features, cropping focused on important regions, and sharpening enhanced edges and structures. These steps helped make the images consistent and supported reliable model training.

Our framework performed well on the datasets used, but we did not test how it generalizes across different datasets (training on one and testing on another). Assessing its performance on varied imaging sources could give useful insights into its robustness.

In this study, our main goal was to evaluate how well the models perform and how sustainable they are in terms of efficiency and resource use. We did not include explainability techniques such as Grad-CAM or SHAP because our focus was on comparing performance rather than preparing the system for direct clinical use. We understand that explainability is important in clinical settings to build trust and support decision-making. Adding interpretability methods is an important next step, and we plan to explore this in future work to make the system more suitable for real-world clinical applications.

While ensemble models do require more computational resources than individual models, their use is justified by the consistent improvement in overall performance and robustness. Even if the accuracy gain appears modest, ensembles help reduce variability and make predictions more reliable, which is particularly important in critical applications such as medical image analysis. In practice, the additional computational cost can be managed by parallel processing or optimized deployment, making the trade-off between performance and resources reasonable.

We recognize that minimizing false negatives is especially important in clinical practice, as missed cases can have serious consequences. Although explicit false negative rates were not recorded in this study, the model's high test and validation accuracy scores (91.19% and 90.97%, respectively) suggest that false negatives were limited in the evaluated dataset. This quantitative evidence supports our qualitative observation that the false negative control performance was satisfactory. Moreover, by combining strong model performance with expert feedback, the system can be further designed to reduce the likelihood of false negatives and better support safe decision-making in real-world clinical settings.

In this study, we mainly looked at how well individual models and combinations of the best models perform, focusing on getting reliable and accurate results. We also checked how environmentally friendly these models are by measuring their energy use and carbon footprint. During the experiments, we only tracked accuracy and did not include other measures like confusion matrices, AUC, or F1-Score.

VI. CONCLUSION AND FUTURE WORK

We have presented detailed research on the diagnosis of breast cancer from mammogram images by fusing the merits of deep transfer learning and model ensemble fusion. By comparing four pre-trained CNN architectures (DenseNet121, EfficientNetV2B0, MobileNetV2, and XceptionNet) on two typical mammographic datasets (DMID and CBIS-DDSM), we demonstrate the utility of transfer learning in enhancing the classification accuracy under small data scenarios.

Among the individual models, EfficientNetV2B0 achieved the most optimal trade-off between classification

accuracy and energy efficiency. EfficientNetV2B0 and MobileNetV2's ensemble model continued to boost predictive performance with the best test and validation accuracy, reduced overfitting, and improved generalization.

The study also emphasizes the requirement for sustainable AI by comparing the energy and carbon footprints of the architectures employed, with EfficientNetV2B0 being the most resource-efficient alternative.

In this study, we focused on binary classification (benign vs. malignant) to reflect the primary clinical goal of identifying lesions that require further diagnostic intervention. In future work, we plan to extend this framework to multi-class classification, such as distinguishing BI-RADS categories or specific lesion subtypes, to enable more detailed diagnostic assessment. To improve the model's clinical dependability, we propose to include extensive evaluation metrics such as AUC, F1-Score, and confusion matrices.

Energy, carbon footprint, and computing power measurements were recorded from a single run for each model. Repeating experiments to report variability was not feasible due to the high computational requirements and long training times of the selected models. Nevertheless, all models were evaluated under identical controlled conditions, making the comparative analysis valid. Future work could include multiple runs to provide more robust and reproducible results.

Future directions include extending the framework to multi-class breast abnormality classification, integrating explainability techniques (e.g., SHAP, Grad-CAM) for enabling clinical interpretability, and inference pipeline optimization for real-time deployment in resource-constrained environments. Further investigation will also include exploring the application of transformer-based models and federated learning strategies for privacy-preserving diagnostics between institutions.

The current framework was developed and tested on standard 2D mammograms, but its underlying deep learning architecture is flexible and can be extended to higher-resolution images or 3D tomosynthesis data. Scaling to larger or volumetric data may require adjustments in preprocessing, memory management, and model input handling, but it does not necessitate a complete redesign. With appropriate optimizations, the framework can be adapted to handle more complex imaging modalities while maintaining its core functionality.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

M.K. wrote and edited the manuscript text. P.O. conceptualized the study, conducted the experiments, wrote, edited, and reviewed the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGEMENT

The authors thank Nirma University, Ahmedabad, for providing the computing resources required to carry out this research.

REFERENCES

- [1] M. Ataollahi, J. Sharifi, M. Paknahad, and A. Paknahad, "Breast cancer and associated factors: a review," *Journal of Medicine and Life*, vol. 8, no. 4, 6, 2015.
- [2] P. Oza, "Ai in breast imaging: Applications, challenges, and future research," in *Computational Intelligence and Modelling Techniques for Disease Detection in Mammogram Images*, Elsevier, 2024, pp. 39–54.
- [3] P. Oza, P. Sharma, S. Patel, and P. Kumar, "Deep convolutional neural networks for computer-aided breast cancer diagnostic: A survey," *Neural Computing and Applications*, vol. 34, no. 3, pp. 1815–1836, 2022.
- [4] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine*, vol. 103, no. 3, e36905, 2024.
- [5] P. Oza, P. Sharma, S. Patel, and A. Bruno, "A bottom-up review of image analysis methods for suspicious region detection in mammograms," *Journal of Imaging*, vol. 7, no. 9, 190, 2021.
- [6] P. Oza, P. Sharma, S. Patel, F. Adedoyin, and A. Bruno, "Image augmentation techniques for mammogram analysis," *Journal of Imaging*, vol. 8, no. 5, 141, 2022.
- [7] A. Darakh, A. Shah, and P. Oza, "Exploring the benefits of data augmentation for breast cancer classification using transfer learning," in *World Conference on Information Systems for Business Management*, Springer, 2023, pp. 509–520.
- [8] P. Oza, P. Sharma, and S. Patel, "Deep ensemble transfer learning based framework for mammographic image classification," *The Journal of Supercomputing*, vol. 79, no. 7, pp. 8048–8069, 2023.
- [9] A. Zizaan and A. Idri, "Evaluating and comparing bagging and boosting of hybrid learning for breast cancer screening," *Scientific African*, vol. 23, e01989, 2024.
- [10] R. M. Munshi, L. Cascone, N. Alturki, O. Saidani, A. Alshardan, and M. Umer, "A novel approach for breast cancer detection using optimized ensemble learning framework and xai," *Image and Vision Computing*, vol. 142, 104910, 2024.
- [11] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, 105151, 2022.
- [12] K. Jabeen, M. A. Khan, J. Balili, M. Alhaisoni, N. A. Almujaal, H. Alrashidi, U. Tariq, and J.-H. Cha, "Bc2netrf: Breast cancer classification from mammogram images using enhanced deep learning features and equilibrium-jaya controlled regula falsi-based features selection," *Diagnostics*, vol. 13, no. 7, 1238, 2023.
- [13] F. Atban, E. Ekinici, and Z. Garip, "Traditional machine learning algorithms for breast cancer image classification with optimized deep features," *Biomedical Signal Processing and Control*, vol. 81, 104534, 2023.
- [14] J. M. S. Pereira, M. A. de Santana, C. L. de Lima, R. d. C. F. de Lima, S. M. L. de Lima, and W. P. dos Santos, "Feature selection based on dialectical optimization algorithm for breast lesion classification in thermographic images," in *Biomedical Computing for Breast Cancer Detection and Diagnosis*, IGI Global, 2021, pp. 47–71.
- [15] F. Pesapane, C. Trentin, F. Ferrari, G. Signorelli *et al.*, "Deep learning performance for detection and classification of microcalcifications on mammography," *European Radiology Experimental*, vol. 7, no. 1, 69, 2023.
- [16] A. Sahu, P. K. Das, and S. Meher, "Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms," *Physica Medica*, vol. 114, 103138, 2023.
- [17] Y. Amethiya, P. Pipariya, S. Patel, and M. Shah, "Comparative analysis of breast cancer detection using machine learning and biosensors," *Intelligent Medicine*, vol. 2, no. 2, pp. 69–81, 2022.
- [18] S. Mohapatra, S. Muduly, S. Mohanty, J. Ravindra, and S. N. Mohanty, "Evaluation of deep learning models for detecting breast cancer using histopathological mammograms images," *Sustainable Operations and Computers*, vol. 3, pp. 296–302, 2022.
- [19] I. Ozsahin, B. Uzun, M. T. Mustapha, N. Usanese, M. Yuvali, and D. U. Ozsahin, "Bi-rads-based classification of breast cancer mammogram dataset using six stand-alone machine learning algorithms," in *Artificial Intelligence and Image Processing in Medical Imaging*, Elsevier, 2024, pp. 195–216.
- [20] A. Jenefa, A. Lincy, and V. E. Naveen, "A framework for breast cancer diagnostics based on mobilenetv2 and lstm-based deep learning," in *Computational Intelligence and Modelling Techniques for Disease Detection in Mammogram Images*, Elsevier, 2024, pp. 91–110.
- [21] M. A. Al-Antari, S.-M. Han, and T.-S. Kim, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms," *Computer Methods and Programs in Biomedicine*, vol. 196, 105584, 2020.
- [22] M. G. Lanjewar, K. G. Panchbhavi, and L. B. Patle, "Fusion of transfer learning models with lstm for detection of breast cancer using ultrasound images," *Computers in Biology and Medicine*, vol. 169, 107914, 2024. <https://www.sciencedirect.com/science/article/pii/S0010482523013793>
- [23] M. G. Lanjewar, K. G. Panchbhavi, and P. Charanarur, "Lung cancer detection from CT scans using modified densenet with feature selection methods and ml classifiers," *Expert Systems with Applications*, vol. 224, 119961, 2023. <https://www.sciencedirect.com/science/article/pii/S0957417423004633>
- [24] P. Oza, R. Oza, U. Oza, P. Sharma, S. Patel, P. Kumar, and B. Gohel. (November 2023). Digital mammography Dataset for Breast Cancer Diagnosis Research (DMID). [Online]. Available: <https://figshare.com/articles/dataset/bDigitalmammographyDatasetforBreastCancerDiagnosisResearchDMIDbDMIDrar/24522883>
- [25] P. Oza, U. Oza, R. Oza, P. Sharma, S. Patel, P. Kumar, and B. Gohel, "Digital mammography dataset for breast cancer diagnosis research (DMID) with breast mass segmentation analysis," *Biomedical Engineering Letters*, vol. 14, no. 2, pp. 317–330, 2024.
- [26] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, no. 1, pp. 1–9, 2017.
- [27] D. Ninja. (2026). Cbisddsm dataset on dataset. [Online]. Available: <https://datasetninja.com/cbis-ddsm>
- [28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv Preprint, arXiv:1704.04861*, 2017.
- [30] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning*, 2019, pp. 6105–6114.
- [31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).