Enhancing Breast Cancer Diagnosis in Ultrasound Imaging

Maryam Alshehyari, Bashayer Alsereidi, Naema Alkhzaimi
Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{Maryam.Alshehyari, Bashayer.Alsereidi, Naema.Alkhzaimi}@mbzuai.ac.ae
Group ID = G-19

Abstract

Breast cancer remains a critical health challenge, with accurate classification of breast lesions using ultrasound being particularly difficult in resource-limited settings where access to trained radiologists is often constrained. Current computational approaches face limitations in handling data imbalance and optimizing feature extraction, resulting in inconsistent classification performance. This project explores the use of deep learning models for breast ultrasound image classification, focusing on distinguishing between benign, malignant, and normal cases. DenseNet121 demonstrated the highest performance, achieving a validation accuracy of 93.98%. Key enhancements, including data augmentation, advanced preprocessing techniques, and the creation of intersection images emphasizing tumor regions, were implemented to address these challenges. The use of intersection images proved highly effective in enhancing model performance. Our findings underscore the importance of dataset diversity, preprocessing strategies, and model focus on tumor regions for improving breast cancer classification. DenseNet121 showed promise, but challenges related to dataset size and class imbalance remain. Future work will focus on expanding the dataset, optimizing models for clinical deployment, and integrating AI-assisted diagnosis into real-world healthcare environments. This research highlights the potential of AI in enhancing early breast cancer detection, especially in underserved regions.
Code is available at https://github.com/MaryamAlshehyari/breast-cancer-classification.git.

1 Introduction

Breast cancer is a significant global health issue, standing as the most commonly diagnosed cancer among women. In 2020, around 2.3 million new breast cancer cases were recorded, resulting in 685,000 deaths worldwide (1). Early detection is crucial for improving patient outcomes, as it dramatically increases survival rates and allows for less invasive treatments (2). Among diagnostic tools, ultrasound imaging is widely utilized due to its non-invasive nature and accessibility, especially valuable in resource-limited areas. However, conventional diagnostic methods, such as mammography and biopsy, can be slow, subjective, and prone to human error (3). These limitations point to the need for more efficient solutions to improve diagnostic accuracy and efficiency. This project addresses this need by employing computational techniques to classify breast cancer cases into benign, malignant, and normal categories using ultrasound images from the Al-Dhabyani et al. (2020) dataset (4). Our project offers a practical solution for breast cancer diagnosis using ultrasound, designed to be accurate, accessible, and easy to use. It improves diagnostic accuracy, reduces human error, and cuts down on the time and cost of traditional methods. This makes it especially useful in remote or resource-limited settings, where access to specialized radiologists may be

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

limited. By automating classification, even non-specialist healthcare workers can use it, expanding early detection to underserved areas. The model's flexibility means it could be adapted for other medical imaging needs, making it a valuable tool for broader healthcare applications. Ultimately, our project supports quicker, more reliable diagnoses and better patient outcomes, while helping bridge gaps in healthcare access.

Contributions

- Benchmarking to identify the best way to improve the classification of breast cancer.
- Exploring and testing various advanced augmentation techniques to enhance diagnostic accuracy.
- Aiming to achieve state-of-the-art results in classification accuracy.

2 Related Work

The application of artificial intelligence (AI) in breast cancer diagnosis using ultrasound imaging has shown promising advancements in recent studies, with numerous researchers exploring innovative methodologies and achieving notable results. A 2022 study introduced a grid-based deep feature generation approach, employing models such as ResNet101, MobileNetV2, and EfficientNetb0 to classify breast cancer cases into benign, malignant, and normal categories. This approach achieved a classification accuracy of 97.18%, highlighting the importance of effective feature extraction and selection in enhancing model performance for breast cancer classification (5). In a similar vein, Niu and colleagues, in 2020, developed a machine learning-based diagnostic method that leverages BI-RADS features along with morphological and texture-based attributes derived from ultrasound images. This study targeted the differentiation of benign and malignant lesions within the BI-RADS 4A category by utilizing support vector machines (SVMs), which led to an improvement in diagnostic precision. This work emphasized the utility of traditional machine learning techniques in refining ultrasound image analysis for more accurate diagnosis (6). Further highlighting the potential of deep learning, Shen et al. (2021) applied a deep learning model to a dataset containing over 5.4 million ultrasound images to classify breast cancer. Their model achieved an impressive area under the curve (AUC) score of 0.976 in distinguishing between benign and malignant lesions, illustrating the power of deep learning to process extensive datasets effectively and achieve high diagnostic accuracy. This study underscores the advantages of deep learning for precise differentiation in large-scale imaging datasets (7). Huang et al. (2020) provided a comprehensive review of deep learning approaches for breast cancer detection in both mammography and ultrasound imaging. This review discusses how convolutional neural networks (CNNs) and transfer learning techniques have contributed to improved diagnostic accuracy, despite challenges such as high variability and noise in ultrasound imaging. Huang's review highlights the potential of deep learning to improve both sensitivity and specificity in ultrasound-based breast cancer diagnosis, offering insights into the state-of-the-art methods in this area (8). In 2021, Li et al. explored the use of deep learning models combined with data augmentation and advanced feature extraction techniques for breast cancer classification in ultrasound images. By applying models such as ResNet50 and InceptionV3 along with augmentation methods, this study achieved over 90% accuracy. Li's work underscores the importance of data preprocessing and augmentation, which help to manage the variability inherent in ultrasound imaging, thereby improving model performance and robustness (9). Building upon these studies, our project focuses on classifying breast cancer into three categories—benign, malignant, and normal—using ultrasound images. Our work extends prior research by incorporating enhanced preprocessing techniques and optimizing feature extraction methods specifically tailored to ultrasound imaging. By improving the accuracy and reliability of classification models, our project aims to contribute meaningfully to AI-based breast cancer diagnosis in clinical applications.

3 Problem Statement

Despite advancements in medical imaging, accurately classifying breast lesions using ultrasound remains a challenge, particularly in resource-limited settings where access to trained radiologists may be limited. Current computational approaches often struggle with data imbalance and optimizing feature extraction, leading to inconsistent classification performance. There is a pressing need for a more robust solution that enhances diagnostic accuracy while maintaining reproducibility and efficiency. This project addresses these issues by developing a specialized classification model for ultrasound images, applying advanced preprocessing techniques to improve feature representation. The goal is to create a more reliable tool for breast cancer diagnosis that can be effectively utilized in various healthcare contexts, ultimately supporting early detection efforts where they are needed most.

- The primary objective of this work is to enhance diagnostic accuracy through advanced data augmentation and preprocessing techniques.
- The secondary objective of this work is to contribute to faster and more accurate early detection of breast cancer to support better patient outcomes.
- The tertiary objective of this project is to make the model accessible and usable for healthcare providers in resource-limited settings.

4 Methodology

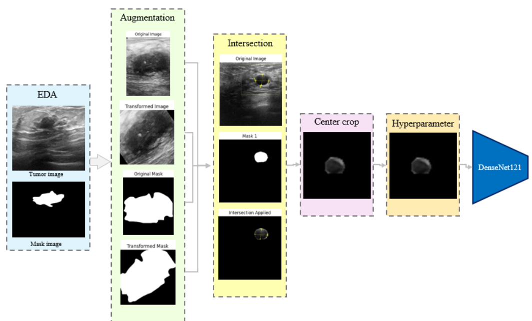

Figure 1: Workflow of the Proposed Methodology for Tumor Detection and Classification. This figure illustrates the complete workflow of the proposed methodology for tumor detection and classification. The pipeline begins with the input of tumor images and their corresponding mask images. It includes various stages such as data augmentation (green outlined section), intersection (yellow outlined section), centre cropping (pink outlined section), hyperparameter tuning, and the use of the DenseNet121 model for classification. Each section demonstrates how preprocessing, and augmentation techniques are applied to enhance the model's performance and focus on relevant regions of the images.

The methodology for enhancing model accuracy is depicted in **Figure 1**, which provides a comprehensive visual representation of each step. The following subsections elaborate on the specific techniques applied at each stage.

1- Exploratory Data Analysis (EDA)
The methodology began with Exploratory Data Analysis (EDA) to gain a comprehensive understanding of the dataset. Through visualizations to examined feature distributions, as showed in **Figure 2**. This analysis revealed a class imbalance, with a higher frequency of benign instances compared to malignant and normal cases. These observations, illustrated indirectly in the blue section of **Figure 1**, guided the selection of strategies to address class imbalance and improve model performance.

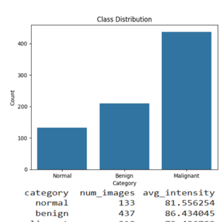

Figure 2: Feature distribution

2- Finding Baseline Model

To establish a benchmark, we tested a baseline convolutional neural network model. Without advanced preprocessing, this model achieved 89.1% accuracy on the test dataset. Additionally, an initial implementation of DenseNet121 (blue triangle in **Figure 1**) without augmentation or advanced preprocessing achieved 87.18% accuracy. This baseline was trained on the same Kaggle dataset, ensuring consistency for comparisons. The results of this baseline highlighted the potential for enhancements, particularly through data augmentation and preprocessing techniques, to improve the model's performance. This baseline model provided a point of reference to assess the impact of the enhancements introduced in this work.

3- Best Model Selection

Various architectures were evaluated, including ResNet18, VGG19, InceptionV3, EfficientNet-B0, DenseNet121, ConvNeXt-Tiny, and ConvNeXt-Small. DenseNet121, depicted at the end of the pipeline in **Figure 1**, demonstrated the highest validation accuracy of 82.69%, making it our chosen model for refinement. Its dense connections allowed for efficient feature propagation, proving advantageous for the dataset.

4- Applying Data Augmentation

Data augmentation, illustrated in the green outlined section of **Figure 1**, was applied to address class imbalance and improve robustness. Techniques such as rotation, flipping, and scaling were implemented to create varied perspectives of the images. This enriched the dataset, enabling the model to learn invariant features and generalize better. The augmented dataset improved validation accuracy to 93.98%, surpassing the baseline.

5- Implementing MONAI

Using the MONAI framework, preprocessing techniques were applied to improve image quality and relevance. Transformations such as median smoothing and UltrasoundConfidenceMapTransform were used to focus on regions of higher tumor likelihood. However, this method. reduced accuracy to 92.11%, and thus, was not included in the final setup.

6- Applying Intersection

The intersection step, shown in the yellow outlined section of **Figure 1**, emphasized tumor regions by combining ultrasound images with their corresponding masks. This approach minimized distractions and improved the model's ability to differentiate between classes, particularly malignant and benign cases.

7- Center Cropping

Depicted in the pink outlined section of **Figure 1**, was applied to focus on the tumor region by centering it within each image. This standardization improved feature extraction and classification accuracy by ensuring the model prioritized the most relevant areas.

8- Hyperparameter Tuning Process

Hyperparameter tuning was conducted to optimize the model's performance. The adjustments, performed on the DenseNet121 model as shown at the end of **Figure 1**, included learning rate, batch size, optimizer type, regularization parameters, and number of epochs. The best configuration utilized Adam with a learning rate of 0.0001 and a batch size of 32, achieving a stable and high validation accuracy of 98.87%. Although additional experiments with regularization and learning rate schedulers were attempted, they reduced accuracy slightly and were excluded from the final configuration.

5 Experimental Setup

Our approach will involve a combination of personal computing devices and university computational resources. The distribution of resources is as follows:

- **Personal Devices:** These will be utilized for initial model development, testing, and small-scale training. Personal devices, such as laptops or desktops equipped with mid-range GPUs (e.g., NVIDIA GTX 1650 or RTX 3050) and at least 16GB of RAM, will handle tasks like preprocessing data, debugging code, and running lightweight experiments.
- **University Computers:** We will leverage university-provided workstations equipped with high-performance GPUs, such as NVIDIA RTX 3090 or A100, and at least 64GB of RAM for computationally intensive tasks. These systems include processors like the AMD Ryzen Threadripper PRO 3955WX with 16 cores, 62 GiB of memory, and NVIDIA GeForce RTX 4090 GPUs with 24 GB of VRAM. The storage capacity is 894.3 GB (device: sda), and the systems run on Ubuntu 22.04.3 LTS. These resources will facilitate training on the full

dataset, hyperparameter optimization, and fine-tuning models. Additionally, access to high-speed network storage and multi-GPU setups will ensure efficient handling of large-scale machine learning workflows.

Hyperparameters and Training Setup

Model Architecture

- We use DenseNet121 as the base model for its efficient feature propagation and ability to reuse features through dense connections, addressing the vanishing gradient problem. Pre-trained on ImageNet, it benefits from transfer learning, enabling faster training and improved performance on ultrasound images. The final layer is modified to classify into three classes (normal, benign, malignant) using a custom linear layer.

Data Loading

- Batch Size (Training): A batch size of 32 is chosen for training, balancing between memory constraints and gradient update frequency, which generally results in faster convergence while fitting well within GPU memory limits.
- Batch Size (Validation): A batch size of 32 is also used for validation, matching the training batch size for consistency in memory usage and processing speed.

Optimizer

- Type: We use the Adam optimizer, which is widely preferred for deep learning tasks due to its adaptive learning rate capabilities. Adam combines the benefits of momentum and RMSProp, making it suitable for handling sparse gradients in our complex classification task.
- Learning Rate (LR): A learning rate of 0.001 is set, balancing between fast learning and stability. This value is commonly effective with Adam and fine-tuned for DenseNet121 on our dataset.

Loss Function

- Type: CrossEntropyLoss is used as the loss function, appropriate for multi-class classification tasks. It measures the difference between the predicted and actual class probabilities, effectively guiding the model to minimize misclassification.

Training Setup

- Number of Epochs: The model is trained for 100 epochs, a sufficient duration to allow convergence while monitoring for potential overfitting.
- Device: Training is conducted on GPU if available, which significantly accelerates the processing speed for deep learning tasks.

Validation Metrics

- Validation Loss: The loss on the validation set is monitored each epoch to track the model's generalization performance and detect any signs of overfitting.
- Validation Accuracy: This metric provides a direct measure of the model's classification performance on unseen data, helping us assess the model's effectiveness in real-world scenarios.

Model Saving

- Saving Condition: The model is saved only when the validation accuracy improves, ensuring that we keep the best-performing version on the validation set. This approach helps prevent overfitting by saving models that generalize better.

5.1 Datasets

The dataset consists of 780 breast ultrasound images (10). It is moderately sized and classified into three categories: normal, benign, and malignant. Below is a summary of the dataset characteristics:

Class	Number of Cases	Modality	Mask Included
Normal	133	Ultrasound	Yes
Benign	437	Ultrasound	Yes
Malignant	210	Ultrasound	Yes
Total	780	-	-

Table 1: Dataset Overview for Breast Ultrasound Images.

5.2 Evaulation Metric(s)

The model's capacity to correctly categorize breast ultrasound images into three groups—normal, benign, and malignant—is the main indicator of our project's success. We use a variety of important measures to evaluate overall performance. The most important of them is classification accuracy, which gives a broad indication of how frequently each class is properly identified by the model. However, because of the dataset's imbalance, where benign instances are far more prevalent than malignant and normal cases, classification accuracy alone is insufficient. To combat this, we additionally use validation accuracy to track the model's performance on unobserved data, which aids in assessing the model's generalizability.

6 Results and Discussion

Model Performance

Our approach aimed to improve breast ultrasound image classification by addressing key challenges such as class imbalance and feature representation. The following summarizes the results at each stage:

Validation Accuracy Results		
Base Model	87.18%	
Enhancement Stage	**Accuracy**	**Improvement**
Augmentation	93.98%	6.8%
MONI	92.11%	-1.87%
Intersection	98.87%	6.76%
Hyperparameter	98.87%	0%

Table 2: Validation accuracy results through the enhancement process.

Baseline Model

The baseline DenseNet121 model achieved a test accuracy of 89.10%, providing a reference point for evaluating further improvements. Its initial strong performance confirmed that DenseNet121's architecture is well-suited for complex feature extraction tasks in ultrasound imaging.

Model Selection

We tested multiple architectures, including ResNet18, VGG19, InceptionV3, EfficientNet-B0, DenseNet121, ConvNeXt-Tiny, and ConvNeXt-Small. DenseNet121 achieved the highest validation accuracy of 82.69%, standing out for its dense connections that enhance feature reuse and alleviate vanishing gradients. The comparative results are presented below:

Model	Validation Accuracy	Validation Loss	Training Loss
ResNet18	81.41%	0.7663	0.0001
VGG19	64.10%	4.9920	0.0901
InceptionV3	73.72%	0.6973	0.4727
EfficientNet-B0	76.92%	0.7942	0.0293
DenseNet121	**82.69%**	**0.6133**	**0.0010**
ConvNeXt-Tiny	60.90%	2.2112	0.0112
ConvNeXt-Small	59.62%	1.9557	0.0329

Table 3: Different models accuracy comparison table.

Enhancement Experiments

Data Augmentation: Rotations, flips, and scaling increased the diversity of the dataset, helping the model generalize to unseen data. Validation accuracy improved to 93.98%, demonstrating the significance of augmentation in addressing class imbalance and enhancing robustness. Advanced Preprocessing with MONAI: Techniques such as median smoothing and UltrasoundConfidenceMapTransform were tested to reduce noise and direct focus on regions of interest. While these methods improved image quality, the validation accuracy slightly declined to 92.11%, indicating that these techniques might not align optimally with the dataset's characteristics. Intersection Images: Combining ultrasound images with their respective masks helped the model focus explicitly on tumor regions, minimizing distractions. Hyperparameter Tuning: Optimizing the learning rate, batch size, and other hyperparameters ensured stable convergence and efficient training. The iterative tuning process maximized the model's potential and improved its generalization. The study highlights DenseNet121 as the most effective architecture for classifying breast ultrasound images into benign, malignant, and normal categories. Its dense connectivity ensures better feature reuse, making it ideal for complex patterns in medical imaging. The improvement in validation accuracy through data augmentation underscores the value of a balanced and diverse training dataset. Addressing class imbalance with transformations like flipping and rotation enhanced the model's robustness, a key requirement for real-world applications where data diversity is significant. Although MONAI preprocessing showed potential, its lower accuracy suggests that highly specific preprocessing may not always yield improvements. This finding emphasizes the importance of tailoring methods to the dataset rather than adopting general-purpose solutions. Intersection images emerged as a breakthrough, allowing the model to concentrate on the most relevant features. This technique aligns with clinical practices where radiologists focus on key regions during diagnosis, making it a valuable addition to the pipeline.

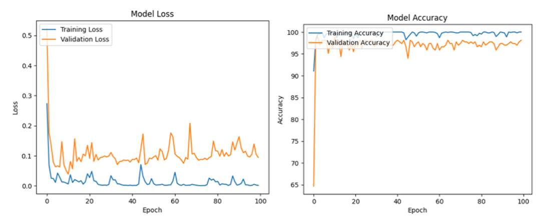

Figure 3: Results for Model loss and Model accuracy

7 Limitations

Despite promising results, several challenges remain. Notably, transformer-based architectures, which have shown great potential in medical imaging tasks, were not tested in this study. Incorporating such models in future research could provide deeper insights and potentially improve classification performance. Additionally, challenges related to dataset size and class imbalance persist, emphasizing the need for larger, more diverse datasets to enhance model generalization and reliability.

- Class Imbalance: The dataset's uneven distribution may lead to biases toward more frequent classes. Augmentation partially mitigates this issue, but acquiring more samples for underrepresented classes (e.g., malignant cases) would further improve performance.
- Dataset Size: With only 780 samples, the dataset limits the model's ability to generalize. Expanding the dataset with diverse imaging conditions and demographics would enhance robustness.
- Deployment Feasibility: The computational resources required for DenseNet121 may limit its use in resource-constrained settings. Future work should explore lightweight alternatives without sacrificing accuracy.

Future Directions

- Dataset Expansion: Including a larger and more diverse dataset would help the model handle variations across populations and imaging conditions.
- Advanced Balancing Techniques: Methods like Synthetic Minority Oversampling (SMOTE) or generative adversarial networks (GANs) could address class imbalance more effectively.

- Model Optimization: Exploring pruned or quantized versions of DenseNet121 could make the model suitable for edge devices and low-resource environments.
- Integration with Clinical Systems: Incorporating the model into a clinical workflow and evaluating its performance in real-world scenarios will validate its utility.
- The results demonstrate significant strides in using AI for breast ultrasound classification. With further refinements and expansions, the model holds great promise for improving early cancer detection, especially in underserved regions

8 Conclusion

Our project achieved significant improvements in breast ultrasound image classification, with DenseNet121 proving highly effective in distinguishing between benign, malignant, and normal cases. By leveraging data augmentation, advanced preprocessing, and intersection image creation, we addressed challenges like class imbalance and improved feature representation, achieving a validation accuracy of 93.98%. While advanced preprocessing techniques showed mixed results, the success of intersection images in highlighting tumor regions aligns with clinical practices, underscoring the model's potential for real-world deployment. Future work will focus on expanding datasets, refining balancing techniques, and integrating AI models into clinical workflows to enhance early breast cancer detection, particularly in underserved areas.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2021 May;71(3):209-249. DOI: 10.3322/caac.21660.

[2] World Health Organization. Breast cancer: prevention and control. WHO, 2021. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[3] Raghu M, Schmidt H, Bengio S. Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer. National Academies Press (US); 2001. DOI: 10.17226/10030.

[4] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. A Dataset of Breast Ultrasound Images. *Data in Brief*. 2020 Feb;28:104863. DOI: 10.1016/j.dib.2019.104863.

[5] Niu S, Huang J, Li J, et al. Application of ultrasound artificial intelligence in the differential diagnosis between benign and malignant breast lesions of BI-RADS 4A. *BMC Cancer*. 2020;20:959. DOI: 10.1186/s12885-020-07413-z.

[6] Fleury E, Marcomini K. Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images. *European Radiology Experimental*. 2019;3:34. DOI: 10.1186/s41747-019-0112-7.

[7] Shen Y, et al. Artificial intelligence system reduces false-positive findings in the 1-26023-2.

[8] Huang S, Yang J, Fong S, Zhao Q. Deep Learning for Breast Cancer Interpretation of Breast Ultrasound Exams. *Nature Communications*. 2021;12:5645. DOI: 10.1038/s41467-02ncer.

[9] Li X, Zhang S, Zhang Q, Wei X. Ultrasound Image Classification of Breast Cancer Using Deep Learning with Augmentation and Feature Extraction Techniques. *IEEE Access*. 9, 29845-29852. DOI: 10.1109/ACCESS.2021.3058410.

[10] "Breast Ultrasound Images Dataset," https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset.