An Experimental Investigation of Classifying Breast Cancer using Different CNN Models

Hirenkumar Kukadiya
Gandhinagar Institute of Computer Science and Applications, Gandhinagar University, (Gandhinagar), India

Divyakant Meva
Faculty of Computer Application, Marwadi University, (Rajkot), India

ABSTRACT
Breast cancer continues to be one of the most common cancers that affect women worldwide. Improving patient outcomes requires an early and precise diagnosis. The performance of many Convolutional Neural Network (CNN) designs for breast cancer image categorization is compared experimentally in this publication. We tested a number of cutting-edge CNN models, such as VGG16, ResNet50, DenseNet121, EfficientNet, and MobileNet, using a number of publically accessible datasets related to mammography and breast cancer histology. According to our tests, EfficientNet-B3 demonstrated the best trade-off between computational efficiency and performance, while DenseNet121 obtained the highest overall accuracy (94.8%) and F1-score (0.937). Additionally, we suggest a brand-new ensemble method that leverages the advantages of several CNN designs, improving classification accuracy by 2.3% over the top-performing single model. Our results offer important new information for the practical application of deep learning algorithms for the diagnosis of breast cancer.

Keywords
AI in Healthcare, Comparative Analysis, Experimental Study, Feature Extraction, Medical Image Analysis.

1. INTRODUCTION
Breast cancer is the most common cancer among women worldwide, with over 2.3 million new cases reported each year [1]. Early detection and accurate diagnosis are essential for lowering death rates and improving treatment results. Conventional diagnostic techniques mostly depend on clinical examination and the subjective and time-consuming interpretation of medical pictures by radiologists and pathologists.

In recent years, Convolutional Neural Networks (CNNs), a subset of deep learning techniques, have shown remarkable improvements in medical image processing [2]. From raw picture data, CNNs can automatically develop hierarchical feature representations, which could help medical practitioners diagnose patients more quickly and accurately. Although a number of CNN designs have been put out for the classification of breast cancer, there is currently little thorough evaluation of how well they perform on standardized datasets.

This study aims to fill this gap by conducting a systematic evaluation of various CNN architectures for breast cancer image classification. We investigate both well-established models such as VGG16 [3] and ResNet50 [4], as well as more recent architectures including DenseNet121 [5], EfficientNet [6], and MobileNet [7]. Our analysis spans multiple publicly available datasets, encompassing both histopathology images and mammograms, to provide a robust assessment of model performance across different imaging modalities.

This paper's primary contributions are:

1. A comprehensive performance evaluation of five state-of-the-art CNN architectures on breast cancer image classification.
2. An analysis of the computational efficiency and model complexity trade-offs for clinical implementation.
3. A novel ensemble approach that combines multiple CNN models to improve classification accuracy.
4. Insights into feature visualization and model interpretability to enhance clinical trust and adoption.

2. RELATED WORK
2.1 Classification of Breast Cancer Using Conventional Machine Learning
Breast cancer classification was frequently done using conventional machine learning techniques prior to the broad acceptance of deep learning. Usually, these techniques entailed manually extracting features, which were then followed by classification algorithms like k-Nearest Neighbors, Random Forests, and Support Vector Machines (SVM) [8]. Even though these methods produced respectable results, their capacity to identify intricate patterns in high-dimensional picture data was constrained, and they mostly depended on domain knowledge for feature engineering.

2.2 Deep Learning Approaches
The introduction of deep learning has changed medical image analysis. Numerous research have investigated the use of CNNs in the categorization of breast cancer. Araújo et al. [9] 87.9% accuracy was attained when a CNN was used to classify photos of breast cancer histology. Wang et al. [10] utilized a modified AlexNet architecture for mammogram classification, reporting an area under the curve (AUC) of 0.86.

Transfer learning has become a popular approach in medical imaging due to limited dataset sizes. Huynh et al. [11] achieved an accuracy of 85.7% in the identification of breast cancer in mammograms using transfer learning with a pre-trained ResNet50 model. Similarly, Choudhary and Hazra [12] optimized a previously trained VGG16 model using histopathological pictures, achieving 90.1% accuracy.

2.3 Ensemble Methods
Several models are combined in ensemble methods to enhance prediction performance. Khan et al. [13] proposed an

ensemble of three CNN models for breast cancer classification, reporting a 3.1% improvement over the best individual model. However, limited research exists on the systematic combination of diverse CNN architectures for breast cancer image analysis, which our study addresses.

3. MATERIALS AND METHOD

3.1 Dataset

Our study utilized three publicly available breast cancer image datasets:

1. Breast Cancer Histopathology (BreakHis) Dataset [14]: 9,109 microscopic pictures of breast tumor tissue taken from 82 patients with various magnification factors (40×, 100×, 200×, and 400×) make up this dataset. The pictures fall into two primary categories: benign and malignant.

2. CBIS-DDSM (Curated Breast Imaging Subset of DDSM) [15]: This dataset contains approximately 10,000 mammography images from 6,775 studies. Each study includes two images of each breast, along with annotations indicating the presence of masses, calcifications, or both, as well as their malignancy status.

3. BACH (Breast Cancer Histology) Dataset [16]: Four kinds of high-resolution microscopy images—normal, benign, in situ cancer, and aggressive carcinoma—make up this collection.

Training (70%), validation (15%), and testing (15%) sets were created from the datasets, ensuring that images from the same patient were not distributed across different sets to avoid data leakage.

3.2 Data Preprocessing and Augmentation

Every image was scaled to 224 × 224 pixels in order to ensure uniformity across various CNN architectures. For histopathology images, color normalization was applied using the method proposed by Macenko et al. [17] to reduce the variability in staining procedures. For mammograms, histogram equalization was performed to enhance contrast.

These data augmentation methods were used to rectify class imbalance and enhance model generalization: Random horizontal and vertical flips, Random rotations (±15 degrees), Random zoom (±10%), Random brightness and contrast adjustments (±10%)..

3.3 CNN Architecture

We evaluated the following CNN architectures:

1. VGG16: A 16-layer CNN architecture known for its simplicity and effectiveness, using small 3×3 convolutional filters stacked together.

2.ResNet50: Skip connections are used in a 50-layer deep residual network to solve the vanishing gradient issue in deep networks.

3.DenseNet121: A 121-layer densely connected convolutional network where each layer receives feature maps from all preceding layers, promoting feature reuse and reducing the number of parameters.

4. EfficientNet-B3: In order to obtain high accuracy with fewer parameters, a CNN design that balances network depth, width, and resolution through neural architecture search optimization

5. MobileNetV2: A lightweight CNN designed for mobile and edge devices, using depthwise separable convolutions to reduce computational cost.

The TensorFlow/Keras framework was used to implement each model, and the ImageNet dataset was used to pre-train the weights. The final classification layer was adjusted based on the distribution of classes in the dataset.

3.4 Transfer Learning and Fine-tuning

We employed a two-stage transfer learning approach:

1. Feature Extraction: Initially, the pre-trained convolutional base was frozen, and only the newly added fully connected layers were trained for 10 epochs.

2.Fine-tuning: Subsequently, the last few convolutional blocks were unfrozen, and the entire network was trained with a smaller learning rate (0.0001) for an additional 50 epochs.

3.5 Ensemble Method

We proposed a weighted ensemble approach combining the predictions of the top three performing models (DenseNet121, EfficientNet-B3, and ResNet50). Using a grid search technique, the weights for each model were established according to how well it performed on the validation set. The final prediction was calculated as:

$$P(y|x) = \sum(wi * Pi(y|x))$$

where $Pi(y|x)$ is the prediction probability of model i for class y given input x, and wi is the weight assigned to model i, with $\sum wi = 1$.

3.6 Evaluation Metrics

We used the following metrics to assess each model's performance: Accuracy,Precision, Recall, and F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Computational efficiency (inference time and number of parameters).

3.7 Implementation Details

An NVIDIA Tesla V100 GPU with 32GB of RAM was used for all of the trials. After the validation loss plateaued for five consecutive epochs, the models' initial learning rate of 0.001 was lowered by a factor of 0.1 using the Adam optimizer. Ten epochs of patience were used for early stopping. For each experiment, a batch size of 32 was used.

4. RESULT

4.1 Comparing the Performance of Different CNN Models

Table1 displays the five CNN architectures' performance metrics on the three datasets' test sets.

Table 1: Performance comparison of CNN architectures on test sets

Model	BreakHis			CBIS-DDSM			BACH			Average		
	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC
VGG16	89.2%	0.88	0.92	85.7%	0.84	0.90	86.5%	0.86	0.91	87.1%	0.86	0.91
ResNet 50	91.5%	0.91	0.94	87.3%	0.86	0.92	90.2%	0.89	0.94	89.7%	0.89	0.37
DenseNet121	94.8%	0.94	0.97	89.1%	0.88	0.93	93.5%	0.93	0.96	92.5%	0.92	0.95
EfficientNet-B3	93.7%	0.93	0.96	88.4%	0.88	0.92	92.8%	0.92	0.96	91.6%	0.91	0.95
MobileNetV2	87.9%	0.87	0.91	84.6%	0.83	0.89	85.3%	0.84	0.90	85.9%	0.85	0.90
Ensemble	96.3%	0.96	0.98	91.2%	0.90	0.95	95.1%	0.94	0.97	94.2%	0.94	0.97

DenseNet121 achieved the highest overall performance with an average accuracy of 92.5%, followed by EfficientNet-B3 (91.6%) and ResNet50 (89.7%). MobileNetV2, despite having the lowest accuracy (85.9%), demonstrated the fastest inference time, making it suitable for resource-constrained environments.The proposed ensemble method outperformed all individual models, achieving an average accuracy of 94.2% across the three datasets, representing a 1.7% improvement over the best individual model (DenseNet121).

4.2 Computational Efficiency Analysis

Table 2 gives a comparison of the various models' computing efficiency.

Table 2: Computational efficiency comparison

Model	Parameters (M)	Model Size (MB)	Inference Time (ms)	FLOPs (G)
VGG16	138.4	528	22.5	15.5
ResNet50	25.6	98	18.7	4.1
DenseNet121	8.0	31	27.3	2.8
EfficientNet-B3	12.2	47	23.1	1.8
MobileNetV2	3.5	14	10.2	0.3
Ensemble	N/A	N/A	68.9	N/A

EfficientNet-B3 showed an excellent balance between performance and computational efficiency, requiring significantly fewer floating-point operations (FLOPs) compared to VGG16 while achieving better accuracy. MobileNetV2, with only 3.5 million parameters, demonstrated the highest efficiency, making it suitable for deployment on mobile and edge devices despite its lower accuracy.

4.3 Effect of Data Augmentation

To assess the effect of data augmentation on model performance, we carried out ablation investigations. Table 3 displays DenseNet121's accuracy using various augmentation methods on the BreakHis dataset.

Table 3: Effect of data augmentation techniques on DenseNet121 performance (BreakHis dataset)

Augmentation Technique	Accuracy	Improvement
No Augmentation	89.5%	-
Horizontal/Vertical Flips	91.2%	+1.7%
Rotation	90.8%	+1.3%
Zoom	90.3%	+0.8%
Brightness/Contrast	91.5%	+2.0%
All Combined	94.8%	+5.3%

The results indicate that combining all augmentation techniques led to a substantial improvement in accuracy (+5.3%), emphasizing the value of data augmentation in medical picture classification jobs, where there is frequently a lack of labeled data.

4.4 Feature Visualization and Interpretability

To enhance model interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized [18] to show the regions of the input images that most affected the predictions made by the model.

The visualizations revealed that DenseNet121 and EfficientNet-B3 consistently focused on clinically relevant regions of the images, such as cellular structures in histopathology images and suspicious masses in mammograms. In contrast, VGG16 occasionally attended to irrelevant background regions, which may explain its lower performance.

5. DISCUSSION

Our comprehensive evaluation of different CNN architectures for breast cancer image classification yields several important insights. DenseNet121 consistently outperformed other models across all datasets, suggesting that its dense

connectivity pattern is particularly effective for capturing the complex patterns in breast cancer images. The dense connections allow for feature reuse and implicit deep supervision, which may contribute to its superior performance.

EfficientNet-B3 is a promising option for clinical deployment when computational resources may be scarce because it showed a great balance between accuracy and computational efficiency. For medical image processing tasks, its compound scaling strategy, which strikes the ideal balance between network depth, width, and resolution, seems to work well.

The proposed ensemble method further improved classification performance, achieving a 1.7% higher accuracy compared to the best individual model. This improvement, while seemingly modest, could translate to a significant reduction in false positives and false negatives in a clinical setting, potentially improving patient outcomes.

The ablation studies on data augmentation highlight the importance of addressing data limitations in medical imaging. The substantial improvement achieved by combining multiple augmentation techniques underscores the value of data augmentation in enhancing model generalization, particularly when dealing with limited labeled data.

Model interpretability remains a crucial aspect for the clinical adoption of deep learning systems. Our Grad-CAM visualizations demonstrated that DenseNet121 and EfficientNet-B3 focus on clinically relevant regions, which can help build trust among healthcare professionals. However, further research is needed to develop more sophisticated interpretability methods that align with clinical decision-making processes**.**

6. LIMITATIONS AND FUTURE WORK

Not with standing the encouraging outcomes, our study includes a number of drawbacks. First, even though the databases are publically accessible, they could not accurately reflect the variety of situations that are encountered in clinical practice. These models should be validated in future research using bigger and more varied datasets from several clinical facilities.

Second, we did not include other clinical data, such as patient demographics, medical history, and genetic information, in our analysis; instead, we only looked at image-based classification. Integrating multiple data modalities through multimodal learning approaches could potentially improve diagnostic accuracy.

Third, while we employed Grad-CAM for model interpretability, more advanced explainability methods could be explored to provide clinicians with more detailed insights into the model's decision-making process.

Future research directions include:

1. Developing lightweight architectures specifically optimized for breast cancer image analysis
2. Investigating methods for self-supervised and unsupervised learning to take advantage of unlabeled medical pictures.
3. Investigating the integration of clinical knowledge into the model architecture through attention mechanisms or graph neural networks

Conducting prospective clinical validation studies to assess the real-world impact of these models on patient care.

7. CONCLUSION

This paper presented a comprehensive experimental study comparing the performance of various CNN architectures for breast cancer image classification. Our results demonstrate that DenseNet121 achieved the highest overall accuracy, while EfficientNet-B3 offered the best balance between performance and computational efficiency. The proposed ensemble approach further improved classification accuracy, highlighting the potential of combining multiple models for enhanced diagnostic performance.

The study's findings provide useful information for developing and utilizing deep learning systems for breast cancer diagnosis. While these models show promising results, further validation in clinical settings and continued improvement in model interpretability are essential for their successful integration into clinical practice. With ongoing advancements in deep learning and increasing availability of medical imaging data, CNN-based approaches have the potential to become valuable tools in assisting healthcare professionals in breast cancer diagnosis, ultimately improving patient outcomes.

8. ACKNOWLEDGMENTS

We would like to thank the providers of the public datasets used in this study for making their data available for research purposes. We also acknowledge the computational resources provided by Marwadi University that made this research possible.

9. REFERENCES

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

[2] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[6] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114).

[7] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[8] Aličković, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, 28(4), 753-763.

[9] Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., & Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks. PloS One, 12(6), e0177544.

[10] Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., & Li, L. (2016). Discrimination of breast cancer with microcalcifications on mammography by deep learning. Scientific Reports, 6, 27327.

[11] Huynh, B. Q., Li, H., & Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of Medical Imaging, 3(3), 034501.

[12] Choudhary, A., & Hazra, A. (2019). Breast cancer detection and classification using deep learning approaches: A comprehensive review. Biomedical Signal Processing and Control, 68, 102625.

[13] Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. P. C. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters, 125, 1-6.

[14] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 63(7), 1455-1462.

[15] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data, 4, 170177.

[16] Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al. (2019). BACH: Grand challenge on breast cancer histology images. Medical Image Analysis, 56, 122-139.

[17] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., & Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (pp. 1107-1110).

[18] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[19] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

[20] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., et al. (2020). Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Transactions on Medical Imaging, 39(4), 1184-1194.