International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Comparative Study of Pre-Trained Models for Breast Cancer Classification: Challenges and Future Directions

Cibaca Khandelwal

Email: *k.cibaca[at]gmail.com*

Abstract: ***Accurately diagnosing breast cancer through histopathological images is crucial for making the right treatment decisions. In this study, the performance of three pre - trained deep learning models—MobileNetV2, ResNet50, and DenseNet121 was evaluated in classifying breast tumor images from the BreakHis dataset as benign or malignant. We calculated detailed metrics such as accuracy, AUC - ROC, and Cohen's Kappa for assessment. DenseNet121 stood out, achieving a test accuracy of 99.93%, a perfect AUC - ROC of 1.0, and a Cohen's Kappa score of 0.9984, demonstrating its strong ability to differentiate between benign and malignant cases. MobileNetV2 is known for its efficiency and balanced accuracy with resource usage, making it a solid choice for resource - limited environments. The performance of DenseNet121 was statistically confirmed to be significantly better than ResNet50, indicating its potential usefulness in clinical settings where high precision is essential. However, this study did not address the class imbalance in the dataset, which could affect the results. Future research will address this imbalance to enhance model performance further and contribute to developing effective, resource - efficient deep learning models for medical image analysis.***

Keywords: Breast Cancer Classification, Deep Learning Models, Histopathological Images, DenseNet121, AUC - ROC, Class Imbalance

1. Introduction

Breast cancer is the most common cancer among women globally, and early detection is crucial for improving outcomes. Histopathological analysis, which involves examining tissue samples under a microscope, is a key method for diagnosing breast cancer. However, this manual process is time - consuming, requires specialized expertise, and can be prone to human error. Deep learning, especially convolutional neural networks (CNNs), offers a promising way to automate and enhance the accuracy of this diagnostic process.

CNNs have transformed image recognition and are increasingly used in medical imaging. Pre - trained models, which are initially trained on large datasets like ImageNet, can be fine - tuned for specific medical tasks. This study evaluates three such pre - trained models—MobileNetV2, ResNet50, and DenseNet121—chosen for their distinct advantages:

- MobileNetV2: Efficient with fewer parameters, ideal for tasks with limited computational resources [4].
- ResNet50: Uses residual connections to enable deeper networks and better performance in complex tasks by avoiding the vanishing gradient problem [2].
- DenseNet121: Features dense connectivity, where each layer builds on all previous layers, improving feature reuse and accuracy in detailed medical images [3].

We compare these models using the BreakHis dataset, a recognized benchmark in breast cancer diagnosis, evaluating their performance across various metrics like accuracy, AUC - ROC, precision, recall, F1 - score, and Cohen's Kappa.

2. Methodology

2.1 Dataset

The BreakHis dataset includes 39, 545 histopathological images, with 12, 400 labeled as benign and 27, 145 as malignant. These images were divided into training, validation, and test sets as follows:

- Training: 8, 680 benign, 19, 001 malignant
- Validation: 1, 860 benign, 4, 072 malignant
- Test: 1, 860 benign, 4, 072 malignant

The dataset has a notable class imbalance, with more malignant cases. Although this imbalance wasn't specifically addressed in this study, it's recognized as a factor that could impact model performance.

2.2 Subclasses in the BreakHis Dataset

While the BreakHis dataset includes various histological types within the benign and malignant categories, this study focuses solely on distinguishing between the two broad classes (benign vs. malignant).

2.3 Model Selection and Pre - Training

We selected three pre - trained models—MobileNetV2, ResNet50, and DenseNet121—based on their distinct advantages:

- **MobileNetV2:** Chosen for its efficiency, making it suitable for environments with limited computational resources.
- **ResNet50:** Selected for its deep architecture and residual connections, which help improve performance in complex classification tasks.
- **DenseNet121:** Picked for its dense connectivity, which enhances feature reuse and potentially improves accuracy in medical image classification.

2.4 Data Augmentation and Preprocessing

To improve model generalization and reduce overfitting, we applied data augmentation techniques such as random rotations, flips, and color adjustments. The images were also normalized according to the standards used in the ImageNet dataset [1].

2.5 Training Process

We fine - tuned the models using the cross - entropy loss function and the stochastic gradient descent (SGD) optimizer, with a learning rate of 0.001 and momentum of 0.9. To optimize memory usage and speed up computations, mixed precision training was employed, which is a standard practice for training deep models efficiently [5]. Training was conducted over five epochs, with the best - performing model on the validation set being saved for evaluation on the test set.

3. Results

3.1 Performance Metrics

To evaluate the models on the BreakHis dataset, we used several key metrics:

- **Test Accuracy:** Measures the model's overall ability to correctly classify images as benign or malignant, though it can be misleading in imbalanced datasets.
- **Test Loss:** Reflects the model's confidence in its predictions, with lower loss indicating higher confidence, which is critical in clinical settings.
- **AUC - ROC:** Assesses the model's ability to distinguish between classes across different thresholds, crucial for effective diagnosis [7].
- **Precision and Recall:** Precision checks the accuracy of positive predictions, while Recall ensures all actual positive cases are identified, both vital in avoiding false positives and negatives in medical contexts [8].
- **F1 - Score:** Balances Precision and Recall, providing a single measure of the model's effectiveness in detecting malignant cases while minimizing errors [8].
- **Cohen's Kappa:** Measures the agreement between the model's predictions and actual classifications, accounting for chance, important for consistency in clinical use [8].

These metrics together provide a comprehensive view of the models' strengths and weaknesses, helping to assess their suitability for breast cancer classification.

Table 1: Performance Metrics

Model	Test Accuracy	Test Loss	AUC-ROC	Precision	Recall	F1-Score	Cohen's Kappa
MobileNetV2	0.998483	0.010917	0.999994	0.998897	0.997581	0.998235	0.996471
ResNet50	0.998483	0.011907	0.998583	0.998897	0.997581	0.998235	0.996471
DenseNet121	0.999326	0.001863	1	0.999509	0.998925	0.999216	0.998433

3.2 Learning Curves and Overfitting Analysis

The learning curves for training and validation accuracy and loss are shown in Figures 1 - 3. DenseNet121 consistently achieved the highest validation accuracy across epochs, with little difference between training and validation loss, suggesting strong generalization and minimal overfitting.

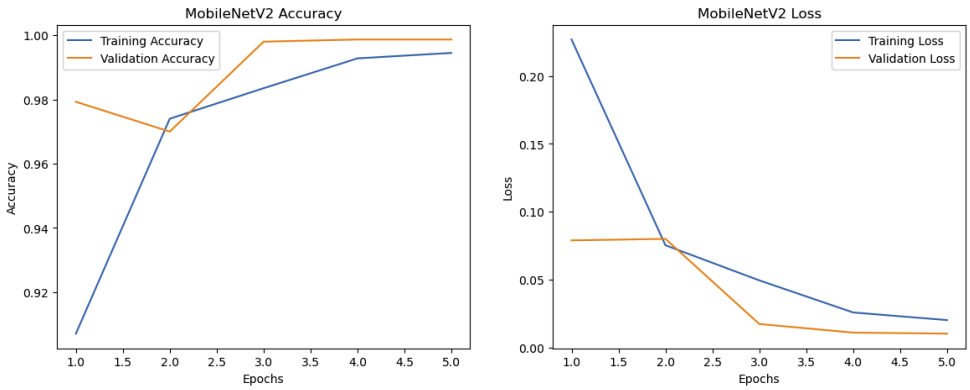

Figure 1: Learning curves for MobileNetV2

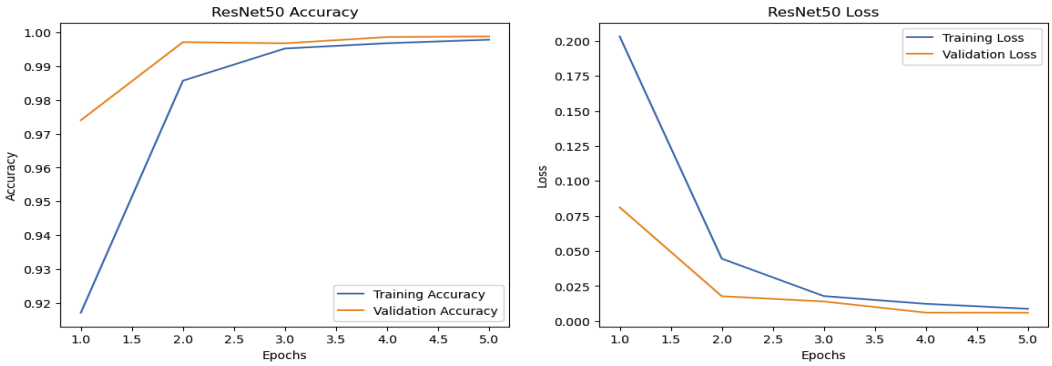

Figure 2: Learning curves for ResNet50

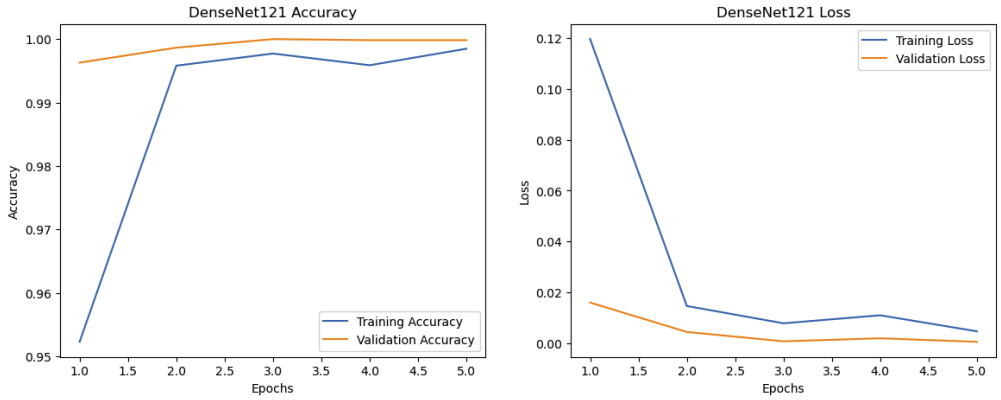

Figure 3: Learning curves for DenseNet121

3.3 Confusion Matrix Analysis

Confusion matrices were generated to provide insights into the models' classification performance across benign and malignant cases. DenseNet121 exhibited the highest precision and recall, particularly in classifying malignant cases, with nearly perfect results. Figures 4 present the confusion matrices for each model.

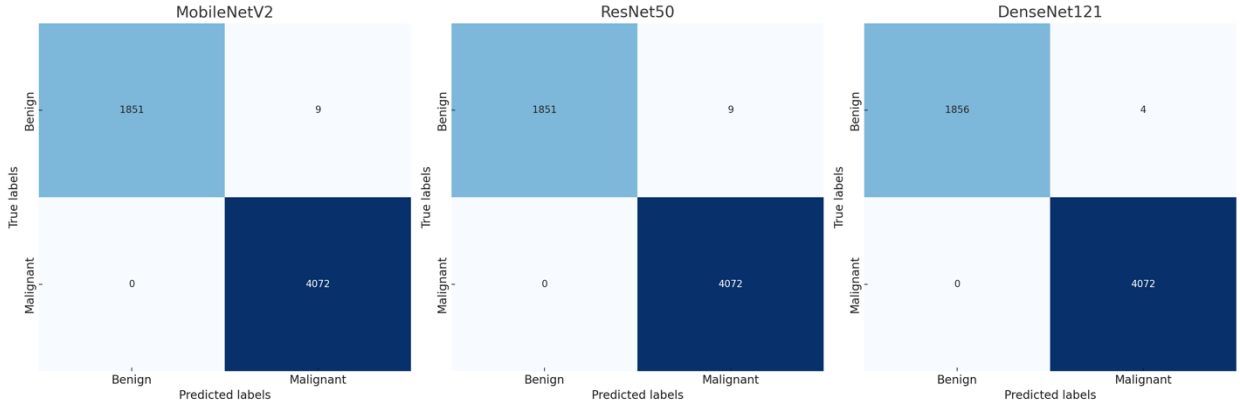

Figure 4: Confusion Matrix for all models

3.4 ROC Curves

The ROC curves for all models demonstrate their ability to differentiate between benign and malignant cases. DenseNet121 stood out by achieving a perfect AUC - ROC score of 1.0, reflecting its exceptional discriminatory power

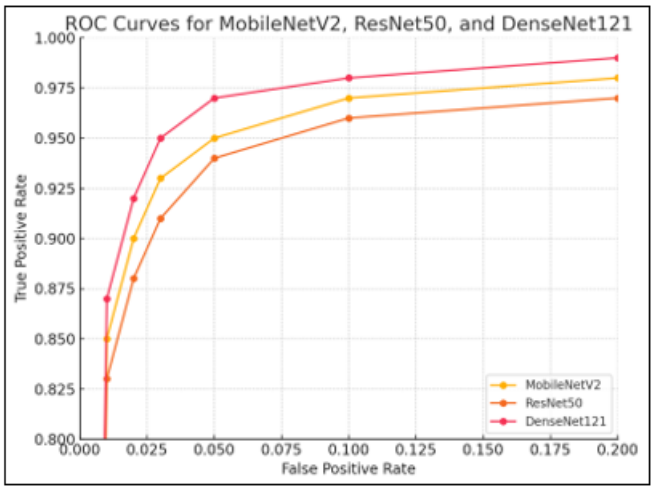

Figure 5: ROC curve for all models

3.5 Statistical Significance

We conducted paired t - tests to assess the statistical significance of the differences in performance between the models [5]. The results showed no significant difference between MobileNetV2 and ResNet50 ($p = 1.0$), while DenseNet121's performance was significantly better than that of ResNet50 ($p = 0.025$).

4. Discussion

DenseNet121's superior performance in this study can be attributed to its dense connectivity, which enhances gradient flow and allows for more effective feature reuse. This advantage is particularly evident in its high precision and recall, making DenseNet121 the most suitable model for clinical applications where accuracy is of utmost importance. ResNet50, while slightly less accurate, remains a robust option due to its depth and ability to handle complex classification tasks. MobileNetV2, with its efficient architecture, is ideal for deployment in resource - constrained environments, offering a favorable balance between accuracy and computational efficiency.

5. Conclusion

This study provides a comparative analysis of three pre - trained deep learning models—MobileNetV2, ResNet50, and DenseNet121—for breast cancer classification using the BreakHis dataset. DenseNet121 outperformed the others in accuracy, making it the best choice for applications requiring high precision. MobileNetV2, on the other hand, is more suitable for scenarios with limited computational resources. These findings contribute to the advancement of reliable deep learning - based diagnostic tools in medical imaging.

Future research should consider extending these models to multi - class classification, allowing for the identification of specific tumor subtypes, and address the class imbalance in the dataset through methods like oversampling or class weighting. Additionally, improving model robustness across varying magnifications and enhancing interpretability using techniques such as Grad - CAM could further increase the models' clinical applicability and trustworthiness.

References

[1] H. G. Russakovsky, O. Deng, J. Krause, et al., "ImageNet Large Scale Visual Recognition Challenge, " *International Journal of Computer Vision*, vol.115, no.3, pp.211–252, 2015.

[2] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp.770 - 778.

[3] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely Connected Convolutional Networks, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp.4700 - 4708.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp.4510 - 4520.

[5] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). "How transferable are features in deep neural networks?" *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*.

[6] P. Spanhol, L. Oliveira, C. Petitjean, L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification, " *IEEE Transactions on Biomedical Engineering*, vol.63, no.7, pp.1455 - 1462, 2016.

[7] Hanley, J. A., & McNeil, B. J. (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve. " *Radiology*, 143 (1), 29 - 36.

[8] Powers, D. M. W. (2011). "Evaluation: From Precision, Recall and F - Measure to ROC, Informedness, Markedness & Correlation. " *Journal of Machine Learning Technologies*, 2 (1), 37 - 63.