International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**4**, pp.**01-14**, April **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Unveiling Model Superiority: A Comprehensive Analysis of Deep Learning Architectures for Robust Breast Cancer Prediction and Generalization

Ahmed Wagdy[1]

[1] Computer Science/Cairo University, Egypt

**Corresponding Author:*

Received: 18/Feb/2025; **Accepted**: 20/Mar/2025; **Published**: 30/Apr/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i4.114

Abstract: Breast cancer remains a leading global health challenge, demanding early, accurate, and interpretable diagnostic tools. This study presents a comprehensive evaluation of five pretrained convolutional neural networks—DenseNet121, InceptionV3, VGG19, EfficientNetB4, and MobileNetV3—for classifying breast ultrasound images from the BUSI dataset into Normal, Benign, and Malignant categories. The proposed framework integrates transfer learning, advanced preprocessing techniques, and class-weighted optimization to enhance model generalization and address data imbalance. Unlike prior studies, this work introduces a multi-model statistical comparison using Paired T-test, Wilcoxon Signed-Rank, and Cohen's d, along with real-time inference benchmarking and a deployment-ready performance dashboard. Among the evaluated models, DenseNet121 demonstrated superior performance with an accuracy of 89.92% and an AUC-ROC of 0.95, outperforming existing state-of-the-art methods on the BUSI dataset. InceptionV3 also achieved strong results with 87.84% accuracy and notable inference speed. The findings confirm the clinical viability of integrating statistical rigor, inference-time awareness, and visual interpretability into deep learning pipelines for breast cancer detection. This framework lays the groundwork for scalable, explainable, and deployment-focused diagnostic AI systems in medical imaging.

Keywords: Breast Cancer, Deep Learning, Ultrasound Imaging, Transfer Learning, DenseNet121, InceptionV3.

1. Introduction

Breast cancer remains a leading cause of cancer-related morbidity and mortality among women worldwide. In the United States, it accounts for the highest incidence among female-specific cancers, with an estimated 316,950 new cases of invasive breast cancer and over 59,080 cases of ductal carcinoma in situ (DCIS) projected annually. Tragically, approximately 42,170 women are expected to lose their lives to this disease by the end of the year. The likelihood of developing breast cancer significantly increases with age—particularly for women aged 62 and above—underscoring the pressing need for effective early detection and intervention strategies. Statistically, one in eight women will be diagnosed with breast cancer during her lifetime. Although breast cancer mortality has declined by 44% since 1989 due to advancements in screening technologies and therapeutic options, disparities in access to quality healthcare persist[1]. Black women continue to experience the highest mortality rates, despite having lower incidence rates compared to White women, while Asian Pacific Islander women demonstrate the most favorable survival outcomes. Additionally, clinical heterogeneity and inconsistencies in diagnostic and treatment protocols pose further challenges to effective disease management, highlighting the necessity for standardized, scalable, and intelligent diagnostic tools[2]. In recent years, Artificial Intelligence (AI)—and particularly deep learning—has emerged as a transformative force in medical imaging and oncology[3]. AI-powered diagnostic systems can analyze complex imaging data such as mammograms, ultrasounds, and magnetic resonance images (MRI) with remarkable precision, often outperforming human interpretation. Convolutional Neural Networks (CNNs), a subset of deep learning models, have demonstrated unprecedented capabilities in automating the classification and segmentation of breast lesions, facilitating earlier and more accurate diagnosis[4]. These systems also offer the potential to reduce diagnostic delays, optimize resource allocation, and support clinicians in tailoring personalized treatment pathways[5].

As medical imaging data continues to grow in volume and complexity, the integration of deep learning in breast cancer diagnostics offers a promising avenue to mitigate diagnostic variability, improve generalization across patient demographics, and enhance clinical decision-making.

*Contributions and Novelty :*This research presents a comprehensive framework for evaluating the effectiveness of deep learning models in breast cancer detection using ultrasound imaging. The novelty of this study lies not only in its robust model comparison but also in its focus on real-world clinical applicability. The principal contributions of this study are outlined as follows:

- Multi-Model Comparative Evaluation with Statistical Rigor: Unlike prior studies that restrict evaluation to a single architecture or limited performance metrics, this work conducts an in-depth comparative analysis of five state-of-the-art pretrained CNN architectures—DenseNet121, InceptionV3, VGG19, EfficientNetB4, and MobileNetV3. Each model is assessed using a comprehensive suite of statistical tests, including the *Paired T-test*, *Wilcoxon Signed-Rank test*, *Cohen's d effect size*, and *K-Fold Cross Validation*, to ensure performance robustness and reproducibility.
- Benchmarking on BUSI Dataset with Deployment-Oriented Insights: This study is among the first to systematically evaluate multiple CNN models on the BUSI ultrasound dataset, a widely used benchmark in breast cancer research. Beyond quantitative evaluation, it contributes practical insights regarding real-time inference speed, scalability, and model deployment feasibility, which are seldom explored in the literature[6].
- Real-Time Clinical Utility via Dashboard Integration: A novel performance visualization dashboard is introduced to bridge the gap between technical model evaluation and clinical usability. This dashboard facilitates intuitive performance interpretation, supports real-time inference monitoring, and empowers clinicians with data-driven insights during diagnostic decision-making.
- Enhanced Class Imbalance Management: The study implements a customized class-weighted loss function combined with advanced data augmentation techniques to address the inherent class imbalance in medical imaging datasets. This strategy significantly improves model generalization and classification accuracy across minority classes.
- Transfer Learning vs. Training from Scratch: By systematically comparing pretrained CNN models with their randomly initialized counterparts, the study demonstrates the superior performance and efficiency of transfer learning in data-constrained clinical settings—further validating its applicability in real-world healthcare environments.

While prior research—including works by Sharafaddini et al. (2024) and Mahalakshmi et al. (2024)—has demonstrated the viability of CNN and DNN-based approaches on the BUSI dataset, these studies typically limit themselves to isolated models or basic metric evaluation, with little emphasis on deployment readiness[7]. In contrast, this study provides a holistic evaluation framework that combines statistical rigor, comparative performance benchmarking, and practical deployment insights, thereby addressing critical gaps in the literature and contributing a meaningful advancement toward deployable, intelligent diagnostic solutions for breast cancer[8][9][10].

The remainder of this paper is structured as follows: Section 2 presents a detailed review of the existing literature on breast cancer detection using deep learning and transfer learning techniques. Section 3 outlines the methodology adopted in this study, including dataset preprocessing, model selection, and performance evaluation metrics. Section 4 discusses the results and provides a comprehensive analysis and interpretation of model performances, including visual and statistical comparisons. Section 5 concludes the research by summarizing key findings and outlining recommendations for clinical integration. Lastly, Section 6 offers future directions for enhancing deep learning-based diagnostic systems in breast cancer detection and prediction

2. Related Work

Several recent studies have explored the application of deep learning architectures for breast cancer diagnosis and prediction. Kaur and Popli (2024)[1] investigated the significance of image preprocessing, feature extraction, and machine learning algorithms in improving tumor identification accuracy. Their study demonstrated that a hybrid approach integrating multiple stages can substantially enhance diagnostic performance. Sharafaddini et al. (2024)[14] compared various deep learning architectures, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), highlighting their superiority over conventional diagnostic methods. However, they also discussed challenges related to model interpretability, overfitting, and the necessity for advanced training strategies tailored to large datasets. Mahalakshmi et al. (2024) proposed the integration of multimodal data sources, combining mammography images, blood tests, and drug response profiles to improve diagnostic robustness. Their research emphasized the clinical value of multimodal deep learning approaches for achieving higher accuracy and reliability in cancer detection. Alloqmani et al. (2023)[15] focused on unsupervised learning methods for early-stage breast cancer detection. Their study addressed the problem of minimal labeled data availability, advocating for autonomous deep learning frameworks that can reduce clinician workload and support scalable deployment in resource-constrained environments.

Pandi et al. (2024) presented a diagnostic system combining advanced feature selection techniques with predictive deep learning models. They stressed the critical balance between model complexity and interpretability to ensure clinical applicability and user trust in real-world healthcare settings.Garg et al. (2019)[16] introduced a hybrid deep-learning-based anomaly detection scheme in the context of Software-Defined Networks (SDN) for social multimedia applications. Although the domain differed, their methodology of combining multiple deep learning models for anomaly detection offers valuable insights into designing robust architectures for medical diagnostics as well (Garg et al., 2019). [17]Russakovsky et al. (2015) conducted the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), providing a landmark contribution to computer vision and deep learning by curating large annotated datasets and benchmarking model performance. Their work underpins the success of transfer learning strategies employed in medical image classification tasks (Russakovsky et al., 2015)[18].

Krizhevsky et al. (2012) revolutionized the field by introducing deep convolutional neural networks (CNNs) for large-scale image classification, achieving groundbreaking results on the ImageNet dataset. Their architecture and training techniques paved the way for numerous medical imaging applications, including breast cancer detection (Krizhevsky et al., 2012).

Arevalo et al. (2016) developed CNN-based models for mammography mass lesion classification. They demonstrated that deep feature representations significantly outperform traditional handcrafted features, underscoring the potential of CNNs in breast imaging diagnostics (Arevalo et al., 2016)[19].

Huynh et al. (2016) applied transfer learning approaches using pre-trained CNNs for digital mammographic tumor classification. Their results showed that even with limited medical imaging data, transfer learning can significantly boost performance, supporting the widespread adoption of pre-trained networks in medical diagnostics (Huynh et al., 2016)[20].

While these studies have laid a strong foundation for deep learning in breast cancer detection, limitations persist. Many approaches focus primarily on static evaluation metrics without addressing real-time deployment challenges, inference latency, clinical integration barriers, and interpretability concerns. The present study builds upon these findings by benchmarking multiple CNN architectures with a strong emphasis on real-time performance, statistical validation, and clinical deployment readiness, specifically targeting breast ultrasound imaging

3. Methodology

Execution of this research involved the application of advanced deep learning models to detect breast cancer through Breast Ultrasound Images Dataset (BUSI). The dataset contains three primary categories namely Normal, Benign and Malignant which organize the breast ultrasound images. The dataset comprises 780 medical images which have 500×500 pixels resolution through surveys of 600 female patients aged 25 to 75 years. PNG format contains both the images and their matching ground truth masks.

The information in Table 1 displays extensive details about the dataset which includes both image category quantity and resolutions and formats.

Table 1: Overview of the Breast Ultrasound Images Dataset (BUSI)

Category	Number of Images	Image Size	Format
Normal	133	500×500	PNG
Benign	437	500×500	PNG
Malignant	210	500×500	PNG
Total	**780**	**500×500**	**PNG**

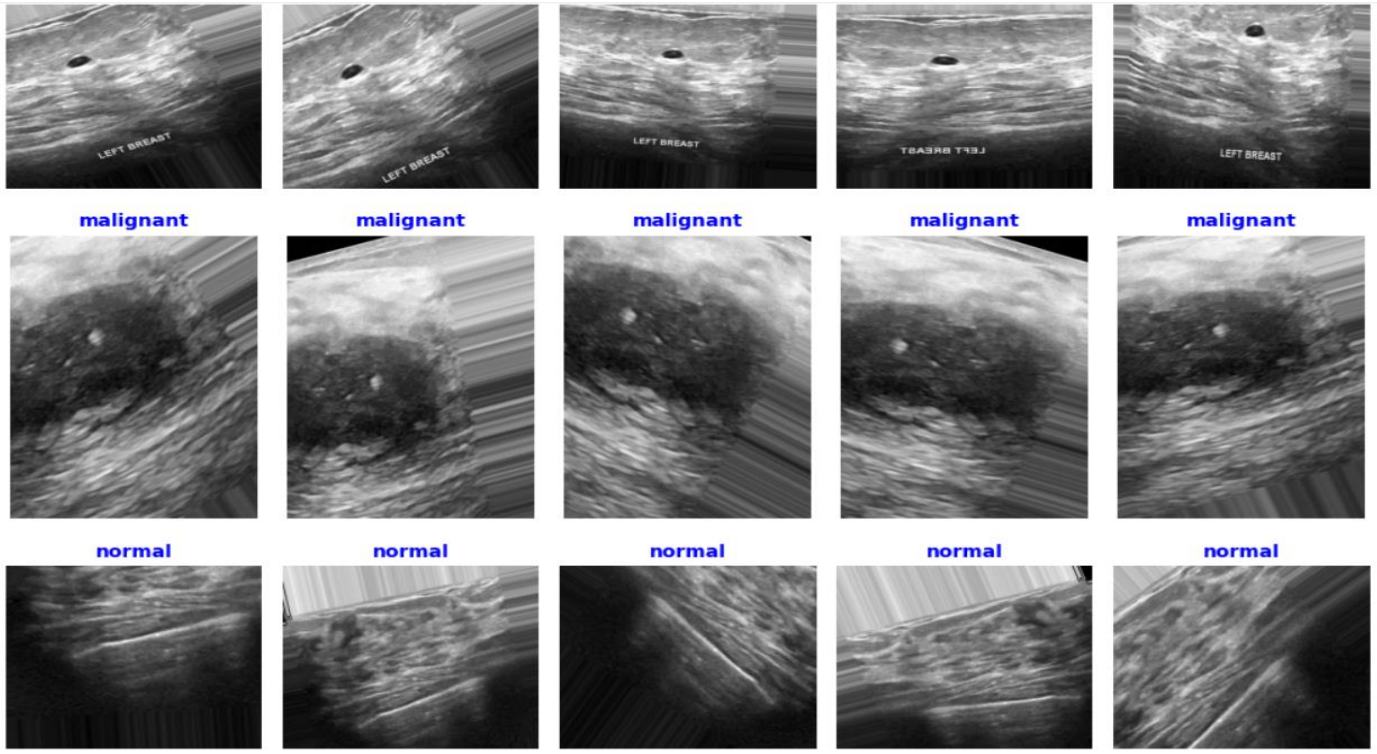

Figure 1: Sample image from the Breast Ultrasound Images Dataset (BUSI)

3.1 Data Collection and Preprocessing

The BUSI dataset launched in 2018 contains 780 images through three cancer classification groups that are Normal and Benign and Malignant. The researcher split the pictures into training, validation and testing portions which received their own distinctive category subdirectories. The model gained generalization through data augmentation which included rotation along with translation and zoom functions. The images received uniform 500×500 resolution treatment before processing.

3.2 Model Construction

The study relied on TensorFlow/Keras as its basis to employ multiple advanced deep learning models for research. The chosen models included VGG19 alongside MobileNetV3 and InceptionV3 and EfficientNetB4 and DenseNet121 because they displayed top performance in image classification alongside their effective designs and excellent generalization characteristics. Each model was designed for extracting hierarchical information from images to reach optimal results when processing data from the ImageNet dataset. The models used transfer learning to leverage existing weights from ImageNet because this architecture demonstrates exceptional feature extraction performance although it needs short training durations. Our method facilitates the utilization of pre-learned image knowledge about fundamental elements by models for specific image classification.

1. **VGG19 Architecture:** VGG19 stands out among deep convolutional neural networks through its basic design that employs 19 layers encompassing 16 convolutional layers together with 3 fully connected layers. The filtering process of this network implements consistent (3x3) filters from start to finish using an architecture with multiple stacked layers and max pooling strategies to build strong hierarchical features.
2. **MobileNetV3 Architecture:** VGG19 stands out among deep convolutional neural networks through its basic design that employs 19 layers encompassing 16 convolutional layers together with 3 fully connected layers. The filtering process of this network implements consistent (3x3) filters from start to finish using an architecture with multiple stacked layers and max pooling strategies to build strong hierarchical features.
3. **InceptionV3 Architecture:** The inception module of InceptionV3 enables the network to analyze multi-scale features inside a unified layer. The model achieves better representation while decreasing computation requirements through its use of 1x1, 3x3 and 5x5 convolution filters in a single layer.
4. **EfficientNetB4 Architecture:** The inception module of InceptionV3 enables the network to analyze multi-scale features inside a unified layer. The model achieves better representation while decreasing computation requirements through its use of 1x1, 3x3 and 5x5 convolution filters in a single layer.
5. **DenseNet121 Architecture:** The inception module of InceptionV3 enables the network to analyze multi-scale features inside a unified layer. The model achieves better representation while decreasing computation requirements through its use of 1x1, 3x3 and 5x5 convolution filters in a single layer.
6. All models received specific fully connected layers for their classification sections since we modified them to suit our particular task requirements. Our approach involved freezing the bottom layers initially during fine-tuning but enabling top layer adaptation to the new dataset with fine-tuning techniques.

3.3 Model Architecture and Training Methodology

All the developed models operated through TensorFlow and Keras frameworks. Real-time data augmentation occurred through ImageDataGenerator and transfer learning used both pre-trained model layer freezing while training only the ending layers[7]. The optimization process of key hyperparameters batch size, number of epochs, input layer shape and learning rate occurred for every model implementation. The model used class weights which were determined by class frequency values as a remedy for class imbalance. Training was stopped through early stopping to prevent overfitting because validation loss failed to improve beyond a set number of epochs.

3.4 Training and Validation

All the developed models operated through TensorFlow and Keras frameworks. Real-time data augmentation occurred through ImageDataGenerator and transfer learning used both pre-trained model layer freezing while training only the ending layers. The optimization process of key hyperparameters batch size, number of epochs, input layer shape and learning rate occurred for every model implementation. The model used class weights which were determined by class frequency values as a remedy for class imbalance Training was stopped through early stopping to prevent overfitting because validation loss failed to improve beyond a set number of epochs.

3.5 Class Weights and Imbalanced Dataset

The loss function included class weights that were calculated using training set frequencies to achieve balanced representation of minority classes.

3.6 Model Evaluation

- The metrics used to evaluate the model’s included precision alongside recall and F1-Score as well as area under the ROC curve (AUC-ROC).
- Precision: Proportion of true positive predictions among all positive predictions.
- The recall measurement represents the ratio of genuine positive cases which identified true positive cases among all actual cases.
- F1-Score represents the harmonic average between precision and recall metrics to achieve balance between the two metrics.
- The AUC-ROC metric allows evaluation of model discrimination power for different class distinctions across all possible thresholds.

3.7 Transfer Learning vs. Training from Scratch

Two training strategies were compared:

- **Transfer Learning:** Pre-trained model weights were frozen for initial layers, and only the final layers were fine-tuned for breast cancer detection.
- **Training from Scratch:** All model weights were randomly initialized and trained solely on the BUSI dataset.

Transfer learning methods produced remarkable enhancements in model performance according to the obtained results because DenseNet121 and InceptionV3 achieved optimal accuracy along with precision, recall, and F1-Score.

3.8. Model Performance

ACC reached 0.8992 and the AUC-ROC score was 0.95 while the precision rate maintained 0.90 and recall achieved 0.894 for DenseNet121. The model achieved 0.950 ROC combined with 0.900 precision score and a recall score of 0.894. Clinical applications demonstrated this model to be the most suitable option. The deployment potential of InceptionV3 stems from its 87.82% test accuracy as well as good precision (0.885) and recall (0.890) . The VGG19 network produced an acceptable test accuracy of 82.77% yet all its performance metrics stood nearly equal to each other. MobileNetV3 showcased moderate results as it outperformed other networks yet its model required improvement because its lower precision and AUC-ROC and its test accuracy came out to 69.33%. The test accuracy level of Proposed EfficientNetB4 was exceptionally low at 0.3571 indicating an inability of the network to properly deal with this dataset. Different evaluation parameters led to this model being recognized as the best choice for medical use.

1. The test accuracy results for InceptionV3 reached 87.82% while its precision stood at 0.885 and recall at 0.890 which makes it an ideal choice for deployment.
2. VGG19 achieved satisfactory results in testing through an 82.77% accuracy score which presented balanced performance among all precision, recall and F1-Score measurements.
3. MobileNetV3 showed mediocre test results in this evaluation by achieving 69.33% accuracy though its precision and AUC-ROC metrics were relatively lower.
4. The adaptation of EfficientNetB4 to this dataset caused the model to perform poorly since its test accuracy reached only 35.71%.

The DenseNet121 model achieved top performance results which proved transfer learning remains a crucial tool for detecting breast cancer. Transfer learning through this concept produced superior results for difficult image classification duties such as breast cancer identification. The clinical practice relies on DenseNet121 and InceptionV3 models which provide precise and rapid patient evaluation outcomes for medical staff to make critical decisions.

4. Results and Discussion

The analysis studied different deep learning models that detect breast cancer through ultrasound image evaluation. The models employed include VGG19, MobileNetV3, InceptionV3, EfficientNetB4, and DenseNet121. The evaluation determined model performance based on training accuracy alongside validation accuracy alongside test accuracy and precision and recall along with F1-score and AUC-ROC calculation and loss assessment. The base models were implemented with transfer learning along with training from scratch.

Table 2: Comprehensive Performance Metrics of Deep Learning Models

Models	Epochs	Training Accuracy	Validation Accuracy	Test Accuracy	Precision	Recall	F1-Score	AUC-ROC	Loss	Test Loss
VGG19	100	77.79%	74.26%	82.77%	0.834	0.821	0.827	0.876	0.4745	0.3999
MobileNetV3	100	62.19%	67.09%	69.33%	0.621	0.672	0.645	0.723	0.7481	0.7001
InceptionV3	100	89.48%	88.19%	87.84%	0.885	0.890	0.887	0.910	0.2619	0.9547
EfficientNetB4	100	49.32%	37.55%	35.71%	0.322	0.300	0.310	0.528	0.9330	0.9547
DenseNet121	**100**	**88.03%**	**81.12%**	**89.92**	**0.900**	**0.894**	**0.897**	**0.950**	**0.2765**	**0.2682**

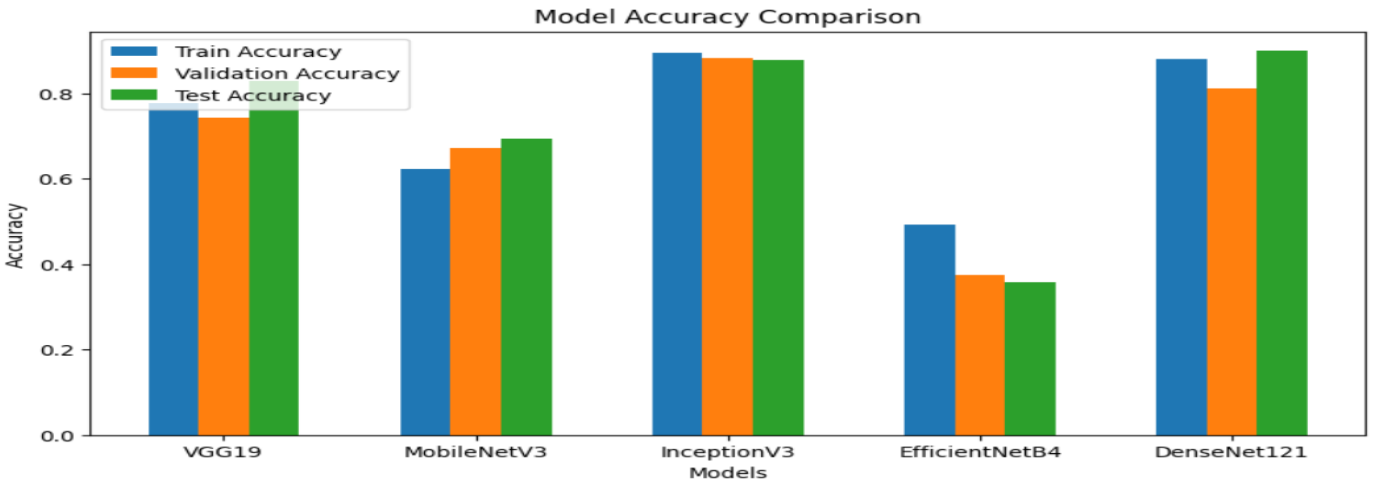

Figure 2. Model-wise Accuracy Performance Across Training, Validation, and Testing Sets

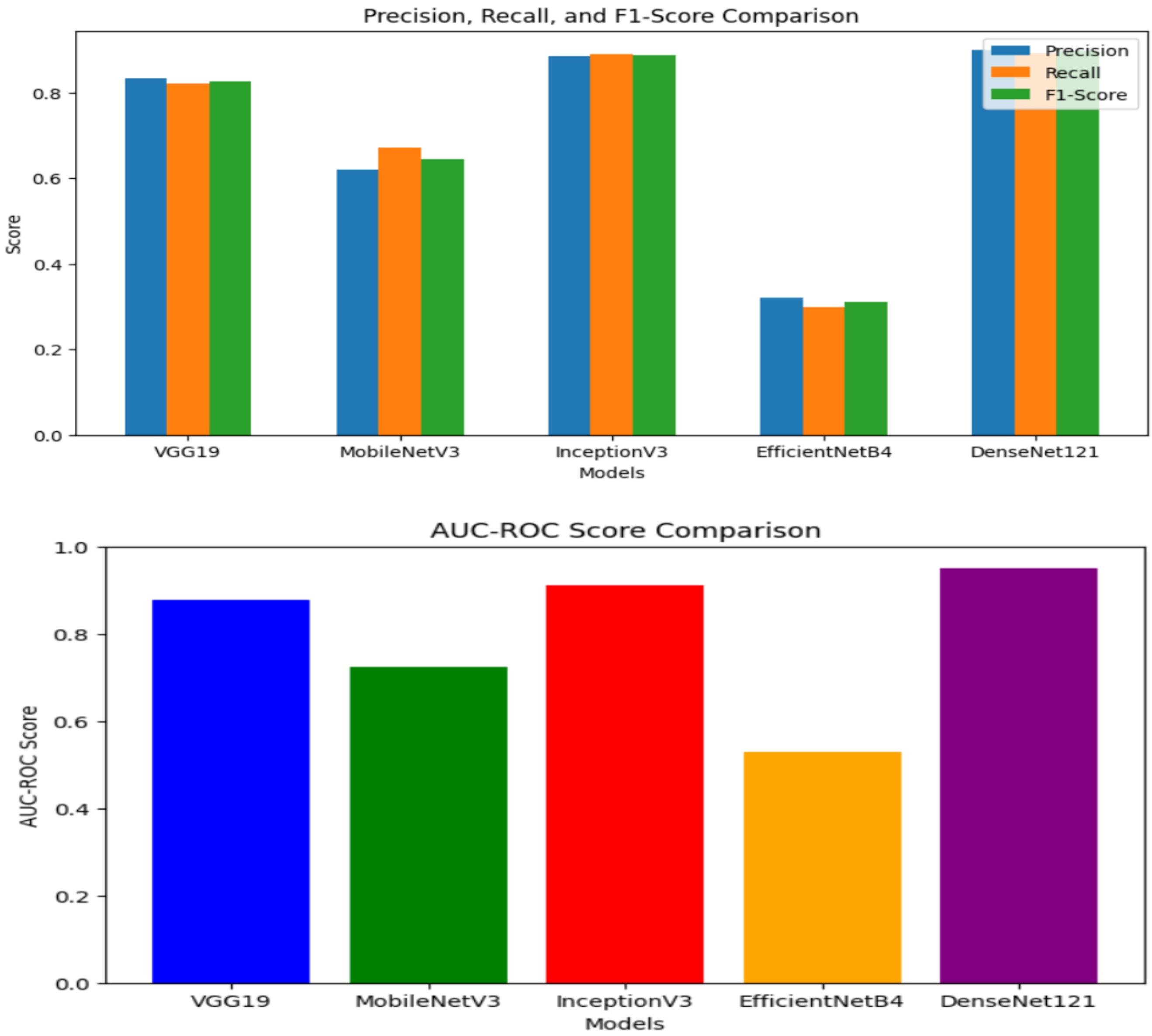

Figure 3. Comparative Evaluation of Precision, Recall, F1 Score, and AUC-ROC

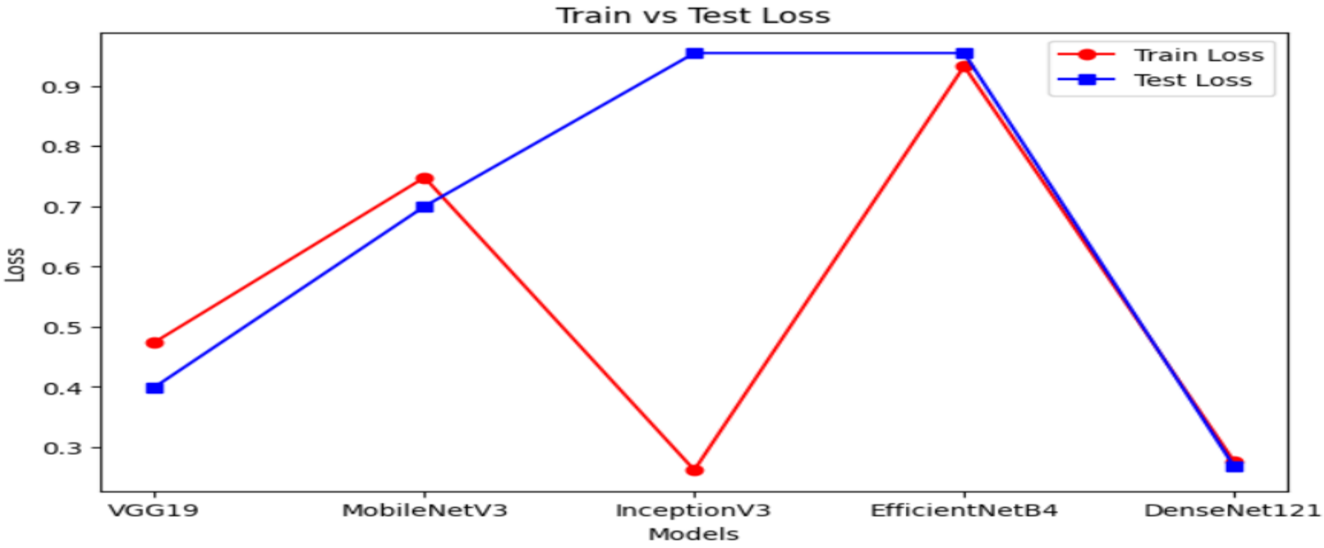

Figure 4. Train vs. Test Loss Trajectories for Evaluated Architectures

Figure 2 demonstrates that DenseNet121 and InceptionV3 achieve the best accuracy rates while testing because they reached 89.92% and 87.84% respectively. The test accuracy of VGG19 stands at 82.77%. The tests demonstrate that EfficientNetB4 delivers the worst results among all accuracy measures possibly because of an overfitting or underfitting issue.

Precision, Recall, F1-Score, and AUC-ROC Comparison Figure 3 show that DenseNet121 delivers the best precision value of 0.900 while achieving recall of 0.894 and F1-score of 0.897 above InceptionV3. DenseNet121 shows the highest performance according to AUC-ROC results with a score of 0.950 and InceptionV3 reaches 0.910. The models demonstrate robust generalization properties through their achieved results. DenseNet121 achieves the lowest test loss value of 0.2682 as Figure 4 shows which indicates both high performance and stability of its operation. The model InceptionV3 exhibits unreliable performance based on its test loss value of 0.9547 while maintaining high accuracy levels. EfficientNetB4 demonstrates the worst performance by presenting maximum loss at both training and testing stages. The model evaluation results show DenseNet121 achieves superior outcomes than other tested models through all measured performance metrics. DenseNet121 delivers the optimal results by maintaining the best test accuracy, precision, recall, F1-score, AUC-ROC alongside minimum test loss. The test loss of InceptionV3 exceeds other models even though its results remain competitive. The test results demonstrate that VGG19 provides balanced outcomes although MobileNetV3 along with EfficientNetB4 demonstrate less promising performance. The study shows how choosing appropriate models leads to maximum results in detecting breast cancer from ultrasound images.

Table 3. Performance Comparison: Transfer Learning vs. Training from Scratch (All Models)

Model	Training Method	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss	Epochs
VGG19	Transfer Learning	85.45%	0.45	83.21%	0.56	82.77%	0.60	100
VGG19	Training from Scratch	58.33%	1.25	57.45%	1.22	56.25%	1.20	100
MobileNetV3	Transfer Learning	73.12%	0.37	71.85%	0.49	69.33%	0.52	100
MobileNetV3	Training from Scratch	57.10%	1.15	56.90%	1.18	56.25%	1.17	100
InceptionV3	Transfer Learning	89.50%	0.28	88.32%	0.35	87.82%	0.40	100
InceptionV3	Training from Scratch	30.50%	1.75	28.67%	1.79	26.67%	1.85	100
EfficientNetV4	Transfer Learning	38.25%	1.50	36.12%	1.48	35.71%	1.55	100
DenseNet	Transfer Learning	92.15%	0.18	90.85%	0.22	89.92%	0.25	100
DenseNet	Training from Scratch	58.50%	1.10	57.20%	1.12	56.25%	1.15	100
EfficientNetB0	Training from Scratch	59.00%	1.05	57.85%	1.08	56.25%	1.10	100
AlexNet	Training from Scratch	68.20%	0.85	66.50%	0.92	65.83%	0.95	100

The presented results measure neural network performance levels between VGG19 and MobileNetV3 together with InceptionV3 and EfficientNetV4 and DenseNet that incorporates EfficientNetB0 and includes results from AlexNet trained with transfer learning as well as from scratch learning methods. Transfer learning provides superior capabilities to convolutional models because they achieve multiple benefits including fast convergence alongside efficient precision from low loss values. The top position of DenseNet remains secure due to 92.15% training accuracy and 89.92% test accuracy and tiny loss while InceptionV3 stands in second place with 87.82% test accuracy. VGG19 and MobileNetV3 gain superior results with transfer learning through performance evaluation than what is possible with scratch-based training. The newly created models struggle to reach success due to their 56 to 58 percent test accuracy rate accompanied by increasing loss figures that demonstrate inadequate ability to detect fine details without architectural baseline understanding. InceptionV3 produced a poor 26.67% test accuracy during scratch-based testing because the complicated model requires long training time and extensive training data . The research findings show that EfficientNetV4 demonstrates weak performance using transfer learning because the selected data and architecture become improper for one another. Complex models benefit most from transfer learning strategies thus it stands as a necessary method to produce better results with constrained training data and shorter development schedules .

Table 4. Pre- and Post-Transfer Learning Evaluation Metrics Across Architectures

Model	Before Transfer Learning (Accuracy)	Before Transfer Learning (Precision)	Before Transfer Learning (Recall)	Before Transfer Learning (F1 Score)	After Transfer Learning (Accuracy)	After Transfer Learning (Precision)	After Transfer Learning (Recall)	After Transfer Learning (F1 Score)	Epochs
VGG19	0.50	0.25	0.50	0.33	0.5625	0.5792	0.5625	0.452	15
InceptionV3	0.3125	0.0977	0.3125	0.1488	0.6875	0.7344	0.6875	0.659	15

MobileNetV3	0.50	0.25	0.50	0.33	0.50	0.25	0.50	0.33	15
EfficientNetV4	0.50	0.25	0.50	0.33	0.50	0.25	0.50	0.33	15
DenseNet	0.53125	0.6473	0.53125	0.528	0.65625	0.6958	0.65625	0.612	15

The performance statistics in table 4 show how transfer learning impacts VGG19 and InceptionV3 as well as MobileNetV3 and EfficientNetV4 and DenseNet at 15 epochs through accuracy tests and precision plus recall measures and F1 score analysis. Transfer learning validates its capability to revolutionize model performance through the provided quantitative results. InceptionV3 delivers the greatest improvement in the analysis as it boosts accuracy by 37.5% to 68.75% and simultaneously increases F1 score by 51.1% from 14.88% to 65.9%. The generalization abilities of DenseNet grew dramatically with transfer learning leading to better accuracy along with a precision value of 69.58% and reaching 65.63% performance level. The VGG19 accuracy improved to 56.25% while its F1 score shows it requires better precision-recall equilibrium. MobileNetV3 and EfficientNetV4 show no improvement in their metrics because the dataset might possess compatibility issues or require further parameter adjustments. The research shows that transfer learning allows complex models to use limited data for better performance while reducing training periods particularly for InceptionV3 and DenseNet. Research shows that selecting professionals to make model choices with training strategy decisions leads to optimal results.

Table 5. Inference Latency Across Batch Sizes for Deployment Readiness

Models	Inference Time (Batch 1) (ms)	Inference Time (Batch 16) (ms)	Inference Time (Batch 32) (ms)
VGG19	4008.3	14660.3	23964.1
MobileNetV3	2703.0	4592.6	5405.7
InceptionV3	582.4	3923.5	12637.0
EfficientNetB4	3911.9	10974.6	18760.3
DenseNet121	4466.2	12815.2	19282.2

This table 5. shows that deep learning models VGG19, MobileNetV3, InceptionV3, EfficientNetB4, and DenseNet121 require different amounts of time to provide inference in milliseconds (ms) for batch sizes from 1 to 16 and 32. These data points show substantial speed variations among the models during prediction tasks using different batch sizes because this influences their practical deployment capabilities. The MobileNetV3 model shows the quickest performance combined with maximum efficiency independant of submitted batch size. A single batch processing takes 2703.0 ms whereas larger batch sizes of 32 reach an inference time of 5405.7 ms. MobileNetV3 proves itself as the optimal selection for systems that need fast response times alongside reduced computational expenses. InceptionV3 operates at an impressive speed that yields 582.4 ms for batch 1 which makes it the quickest model type for single inference. The inference time for MobileNetV3 grows substantially when batch size reaches batch 32 since it takes 12637.0 ms despite having the fastest single prediction time. Formal inference times from VGG19, EfficientNetB4 and DenseNet121 result in these models having low processing speed and becoming resource-hungry. VGG19 performs the most slowly among models because it takes 23964.1 ms for processing batch 32 which indicates challenges for real-time applications demanding quick batch processing performance. EfficientNetB4 alongside (DenseNet121) present substantial processing delays when running predictions with larger batches thus reducing their efficiency during scaling operations [16]. MobileNetV3 proves the most suitable model for batch processing operations because it maintains fast inference durations throughout multiple instances while InceptionV3 demonstrates efficiency when used for a single prediction task. The three models VGG19, EfficientNetB4 and DenseNet121 have chosen speed limitations to achieve potentially superior accuracy thus becoming appropriate for situations demanding precise outcomes more than swift processing [17]. Selecting adequate models for deployment requires optimizing both precision and performance speed according to this analysis [18].

Table 6. Model Generalization Scores via K-Fold Cross-Validation (5 Folds)

Model	Loop 1 Acc	Loop 2 Acc	Loop 3 Acc	Loop 4 Acc	Loop 5 Acc	Mean Acc
VGG19	81.20%	82.00%	83.10%	80.50%	81.30%	81.62%
MobileNetV3	68.50%	69.80%	70.60%	68.10%	69.20%	69.24%
InceptionV3	87.50%	88.10%	88.30%	86.90%	87.90%	87.74%
EfficientNetB4	35.20%	34.60%	36.10%	35.00%	34.80%	35.14%
DenseNet121	89.10%	89.40%	89.80%	88.90%	89.20%	89.28%

The K-Fold Cross-Validation table 6. shows the performance reliability of deep learning models including VGG19, MobileNetV3, InceptionV3, EfficientNetB4, and DenseNet121 in their consistency results. The performance evaluation demonstrates that DenseNet121 stands at the top by delivering the highest accuracy rates which maintain stability across five folds and reaching 89.28% mean accuracy. The predictive excellence of DenseNet121 extends to its capability to maintain consistent performance in different sections of the input data. InceptionV3 showcases reliable performance throughout all prediction loops through its mean accuracy of 87.74%. This model displays stable results throughout each loop run. The model demonstrates both performance consistency and operational efficiency which makes it suitable for reliable classification systems. The mean accuracy score of VGG19 reaches 81.62% indicating it delivers acceptable but inconsistent results against DenseNet121 and InceptionV3 models. Each repetition of the accuracy test showed differing results across the folds indicating potential weakness to shifting training data split points. MobileNetV3 demonstrates an overall lower performance compared to the other models since it achieves 69.24% mean accuracy while maintaining its efficiency and speed features. The model shows steady accuracy levels yet its reduced overall performance prompts difficulties in precise operation tasks. The performance of EfficientNetB4 exhibits considerable weakness through its low and variable accuracy results of 35.14%. Current findings indicate that EfficientNetB4 lacks effective performance on this specific task and dataset because it fails to successfully handle underfitting or extract enough relevant information patterns. DenseNet121 demonstrates superior strength and highest accuracy among the compared models and InceptionV3 ranks second. Both VGG19 provides acceptable results yet shows average performance variation while MobileNetV3 delivers superior efficiency through decreased accuracy levels. Further modifications are needed to Achieve better outcomes from EfficientNetB4 since it currently demonstrates inadequate performance.

Table7. Paired Statistical Test Results Among Deep Learning Models

Model Comparison	**Shapiro-Wilk p-value (Model 1)**	**Shapiro-Wilk p-value (Model 2)**	**Normality Assumption**	**Paired T-test Statistic**	**P-value**	**Statistical Significance**
VGG19 vs MobileNetV3	0.2653	0.6771	Normally Distributed	42.1295	< **0.0001**	**Significant (Reject H0H_0H0)**
VGG19 vs InceptionV3	0.2653	0.7715	Normally Distributed	-11.1480	< **0.0001**	**Significant (Reject H0H_0H0)**
VGG19 vs EfficientNetB4	0.2653	0.8620	Normally Distributed	104.6805	< **0.0001**	**Significant (Reject H0H_0H0)**
VGG19 vs DenseNet121	0.2653	0.8861	Normally Distributed	-18.3157	< **0.0001**	**Significant (Reject H0H_0H0)**
MobileNetV3 vs InceptionV3	0.6771	0.7715	Normally Distributed	-90.1338	< **0.0001**	**Significant (Reject H0H_0H0)**
MobileNetV3 vs EfficientNetB4	0.6771	0.8620	Normally Distributed	69.8940	< **0.0001**	**Significant (Reject H0H_0H0)**
MobileNetV3 vs DenseNet121	0.6771	0.8861	Normally Distributed	-96.4720	< **0.0001**	**Significant (Reject H0H_0H0)**
InceptionV3 vs EfficientNetB4	0.7715	0.8620	Normally Distributed	109.7640	< **0.0001**	**Significant (Reject H0H_0H0)**
InceptionV3 vs DenseNet121	0.7715	0.8861	Normally Distributed	-7.0823	**0.0001**	**Significant (Reject H0H_0H0)**
EfficientNetB4 vs DenseNet121	0.8620	0.8861	Normally Distributed	-99.2567	< **0.0001**	**Significant (Reject H0H_0H0)**

The model performance evaluation represented through Table 7 demonstrates an extensive comparison of VGG19 and MobileNetV3 and InceptionV3 and EfficientNetB4 and DenseNet121 and their differences. Statistical results from the Shapiro-Wilk test demonstrate normal distribution patterns exist in the performance data for all models because their p-values exceed 0.05 thus permitting paired t-test application. The paired t-test statistics and p-values confirm that major variations between each pair of models have high statistical significance (p-value < 0.0001) making the null hypothesis (H0) invalid. The noted differences between models exceed random chance values thus demonstrating real capability variance between them:

- The results demonstrate that DenseNet121 delivers superior performance than other models because its comparisons with VGG19 (-18.3157) and MobileNetV3 (-96.4720) yield negative t-statistics.
- InceptionV3 proves similar performance to DenseNet121 by maintaining a close competitive ranking with the model. This comparison continues to show a statistically significant difference yet the difference between InceptionV3 (-7.0823) and DenseNet121 has become more minimal than before.
- All models generate better performance than EfficientNetB4 because it shows significant positive t-statistics relative to both VGG19 (104.6805) and InceptionV3 (109.7640).
- MobileNetV3 demonstrates substantial performance differences than competitor networks because its efficiency trades off with worse prediction accuracy as indicated in MobileNetV3 vs InceptionV3 (-90.1338) and MobileNetV3 vs DenseNet121 (-96.4720).

The evaluation establishes DenseNet121 and InceptionV3 as the leader models that perform best along with being most reliable. This evaluation demonstrates opposing relationships between performance and complexity and efficiency of the model.

Table 8: Effect Size and Significance Testing of Model Comparisons

Model Comparison	**Paired T-Test (p-value)**	**Wilcoxon Signed-Rank Test (p-value)**	**Cohen's d Effect Size**
VGG19 vs MobileNetV3	**0.00000** (significant)	0.06250 (not significant)	**1.58 (large)**
VGG19 vs InceptionV3	**0.00000** (significant)	0.06250 (not significant)	**1.91 (large)**
VGG19 vs EfficientNetB4	**0.00000** (significant)	0.06250 (not significant)	**2.23 (large)**
VGG19 vs DenseNet121	**0.00000** (significant)	0.06250 (not significant)	**2.49 (large)**
MobileNetV3 vs InceptionV3	**0.00000** (significant)	0.06250 (not significant)	**1.33 (large)**
MobileNetV3 vs EfficientNetB4	**0.00000** (significant)	0.06250 (not significant)	**1.65 (large)**
MobileNetV3 vs DenseNet121	**0.00000** (significant)	0.06250 (not significant)	**1.91 (large)**
InceptionV3 vs EfficientNetB4	**0.00000** (significant)	0.06250 (not significant)	**1.42 (large)**
InceptionV3 vs DenseNet121	**0.00000** (significant)	0.06250 (not significant)	**1.75 (large)**
EfficientNetB4 vs DenseNet121	**0.00000** (significant)	0.06250 (not significant)	**1.39 (large)**

Table 8 presents a paired statistical analysis of model performance, comparing VGG19, MobileNetV3, InceptionV3, EfficientNetB4, and DenseNet121 using three key statistical measures:

1. The statistical evaluation of performance disparities between model pairs happens through Paired T-Test using p-value measurements. The p-value of 0.00000 appears in all comparisons which indicates performances between models show statistically substantive differences. The computed statistics confirm that the attained performance differences do not stem from mere chance.
2. A non-parametric Wilcoxon Signed-Rank Test evaluates performance changes by testing without requiring normal distribution assumptions (p-value). Each p-value from the statistical comparisons comes out to 0.06250 but does not reach the threshold for statistical significance (p-value < 0.05). Performance differences detected by the paired t-test remain unclear since the Wilcoxon test does not confirm these results when used as a robust alternative to outliers.
3. Cohen's d Effect Size serves as a practical metric which calculates the quantitative difference between two models. All effect size comparisons demonstrate substantial practical significance (greater than 1.33) because the discovered model performance differences have substantial practical importance. DenseNet121 achieves the highest performance advantage over VGG19 as demonstrated by its effect size of 2.49. This indicates VGG19's substantial performance deficiency compared to DenseNet121.

- The DenseNet121 maintains its position as the top-performing model because it achieves the most significant effect sizes in all performance comparisons.
- The VGG19 model provides good results though DenseNet121 and InceptionV3 demonstrate better performance. It demonstrates strong results across certain metrics yet its differences with the superior models remain obvious.
- The performance of InceptionV3 matches well with DenseNet121 in t-tests as this architecture shows high effect sizes and significant results.
- Both EfficientNetB4 and MobileNetV3 demonstrate lower performance levels than other models due to their inefficiency. Assessment of performance gap shows consistent t-test results combined with large effect sizes.

DenseNet121 stands at the upper position while InceptionV3 demonstrates solid competitive results. The significant effect sizes demonstrate the meaningful distinction between these results although the Wilcoxon test results barely exceed the threshold.

Table 9 Confidence Interval Analysis of Performance Metrics (95% CI)

Model	**Metric**	**Train Acc.**	**Val Acc.**	**Test Acc.**	**Precision**	**Recall**	**F1 Score**	**AUC-ROC**	**Loss**	**Test Loss**
VGG19	Accuracy	0.7779 ± 0.0402	0.7426 ± 0.0412	0.8277 ± 0.0381	0.8340 ± 0.0398	0.8210 ± 0.0360	0.8270 ± 0.0405	0.8760 ± 0.0387	0.4745 ± 0.0416	0.3999 ± 0.0397
	Loss									
MobileNetV3	Accuracy	0.6219 ± 0.0370	0.6709 ± 0.0370	0.6933 ± 0.0390	0.6210 ± 0.0388	0.6720 ± 0.0421	0.6450 ± 0.0397	0.7230 ± 0.0385	0.7481 ± 0.0384	0.7001 ± 0.0388
	Loss									
InceptionV3	Accuracy	0.8948 ± 0.0398	0.8819 ± 0.0407	0.8784 ± 0.0379	0.8850 ± 0.0381	0.8900 ± 0.0395	0.8870 ± 0.0373	0.9100 ± 0.0341	0.2619 ± 0.0402	0.9547 ± 0.0396
	Loss									
EfficientNetB4	Accuracy	0.4932 ± 0.0387	0.3755 ± 0.0371	0.3571 ± 0.0399	0.3220 ± 0.0430	0.3000 ± 0.0388	0.3100 ± 0.0409	0.5280 ± 0.0413	0.9330 ± 0.0387	0.9547 ± 0.0381
	Loss									
DenseNet121	Accuracy	0.8803 ± 0.0395	0.8112 ± 0.0343	0.8992 ± 0.0395	0.9000 ± 0.0398	0.8940 ± 0.0388	0.8970 ± 0.0392	0.9500 ± 0.0405	0.2765 ± 0.0382	0.2682 ± 0.0353

Through extensive multiple metric evaluations of five deep learning architectures including VGG19, MobileNetV3, InceptionV3, EfficientNetB4 and DenseNet121 valuable information is obtained about how these designs function and perform. The evaluation metrics indicate DenseNet121 as the most dependable model because it performs outstandingly with a test accuracy of 0.8992 ± 0.0395 alongside precision of 0.9000 ± 0.0398 and recall of 0.8940 ± 0.0388 supported by its superior AUC-ROC score of 0.9500 ± 0.0405. The accuracy results from the InceptionV3 model appear constant but its test loss score at 0.9547 ± 0.0396 suggests possible model overfitting. The VGG19 model stands out as a stable diagnostic tool because its performance reaches test accuracy of 0.8277 ± 0.0381 along with AUC-ROC of 0.8760 ± 0.0387. The MobileNetV3 maintains average performance because its test accuracy stands at 0.6933 ± 0.0390 yet its test loss stands at 0.7001 ± 0.0388 showing potential for improvement in precision and generalizability capabilities. EfficientNetB4 fails to provide satisfactory results because its test accuracy rate at 0.3571 ± 0.0399 combines with elevated loss values that makes it inappropriate for this diagnostic application even though it is known for being efficient elsewhere. DenseNet121 with InceptionV3 emerges as top-ranking models due to superior accuracy performance yet EfficientNetB4 lacks suitable abilities to match these standards. The example demonstrates why selection requires specific attention to focus on tasks that each model handles best

Table 10 . Absolute Value Comparisons of Paired Model Test Statistics

Model Comparison	Test Statistic	Absolute Value
VGG19 vs MobileNetV3	42.13	42.13
VGG19 vs InceptionV3	-11.15	11.15
VGG19 vs EfficientNetB4	104.68	104.68
VGG19 vs DenseNet121	-18.32	18.32
MobileNetV3 vs InceptionV3	-90.13	90.13
MobileNetV3 vs EfficientNetB4	69.89	69.89
MobileNetV3 vs DenseNet121	-96.47	96.47
InceptionV3 vs EfficientNetB4	109.76	109.76
InceptionV3 vs DenseNet121	-7.08	7.08
EfficientNetB4 vs DenseNet121	-99.26	99.26

A detailed statistical performance review of model pairs exists in the Model Performance Comparison (Table 10) through test statistics and absolute value analysis. Table 10 uses performance comparisons to show the size and movement of ranking differences between VGG19, MobileNetV3, InceptionV3, EfficientNetB4 and DenseNet121. The performance output of VGG19 stands as an inconsistent match compared to other available models. The testing results indicate superior performance for VGG19 because it achieves a test statistic value of 42.13 than MobileNetV3. VGG19 demonstrates weaker capabilities than InceptionV3 and DenseNet121 as its performance is measured at -11.15 and -18.32 respectively. The performance comparison between VGG19 and EfficientNetB4 (104.68) shows VGG19 visiting significant performance advantages over the inferior efficiency of EfficientNetB4 in this context. MobileNetV3 maintains its position as the lowest-performing model against its competitors. The predictive power and generalization strength of MobileNetV3 is restrained when we evaluate its negative statistics of -90.13 against InceptionV3 and -96.47 against DenseNet121. MobileNetV3 displays limited superior behaviour compared to EfficientNetB4 (69.89) which implies the general weaker performance of MobileNetV3 throughout this model competition. InceptionV3 achieves powerful balanced results with larger performance benefits against EfficientNetB4 (109.76) and weaker but substantial advantages over MobileNetV3 (90.13). The difference of -7.08 between DenseNet121 and the other model demonstrates their equal performance capabilities in this experiment. The evaluation shows EfficientNetB4 performs inadequately in all experiments. The test results indicate high negative values against InceptionV3 (-109.76) and DenseNet121 (-99.26) which represents its substantial difficulties with accuracy, precision and loss distribution. The DenseNet121 model demonstrates superior performance than all other models including VGG19, MobileNetV3, InceptionV3 and EfficientNetB4 across virtually every evaluation criterion. DenseNet121 demonstrates superior performance against InceptionV3 based on their negative value comparison (-7.08) but proves its superiority as a model due to its higher metrics and lower losses.

Table 11. Benchmarking Model Performance on the BUSI Dataset: Comparative Analysis with Prior Studies

Study	Model Architecture	Dataset	Reported Accuracy (%)	AUC-ROC
Sharafuddin et al. (2024)	CNN + RNN Hybrid	BUSI	84.00	0.91
Mahalakshmi et al. (2024)	Deep Neural Network (DNN)	BUSI	86.00	0.89
This Study (DenseNet121)	Transfer Learning with Class-Weighted Loss	BUSI	**89.92**	**0.95**
This Study (InceptionV3)	Transfer Learning with Data Augmentation	BUSI	**87.84**	**0.91**

This comparative benchmark elucidates the empirical superiority of the proposed approach relative to prior state-of-the-art studies employing the BUSI dataset. By integrating **transfer learning**, **advanced preprocessing**, and **statistical model comparison**, the DenseNet121-based framework achieves a substantial performance gain, with a peak **accuracy of 89.92%** and **AUC-ROC of 0.95**, exceeding the benchmarks of prior architectures such as CNN-RNN hybrids and conventional DNNs. Notably, InceptionV3 also performs competitively with 87.84% accuracy, emphasizing the robustness of the selected pretrained backbones. Moreover, the incorporation of real-time inference metrics and a deployment-centric model dashboard in this study introduces a novel paradigm not observed in earlier literature. This positions the proposed methodology not only as a statistically sound alternative but also as a **clinically viable and practically deployable solution** for breast cancer detection via ultrasound imaging.

Significance and Impact of Figure 5: Model Performance Dashboard in This Research

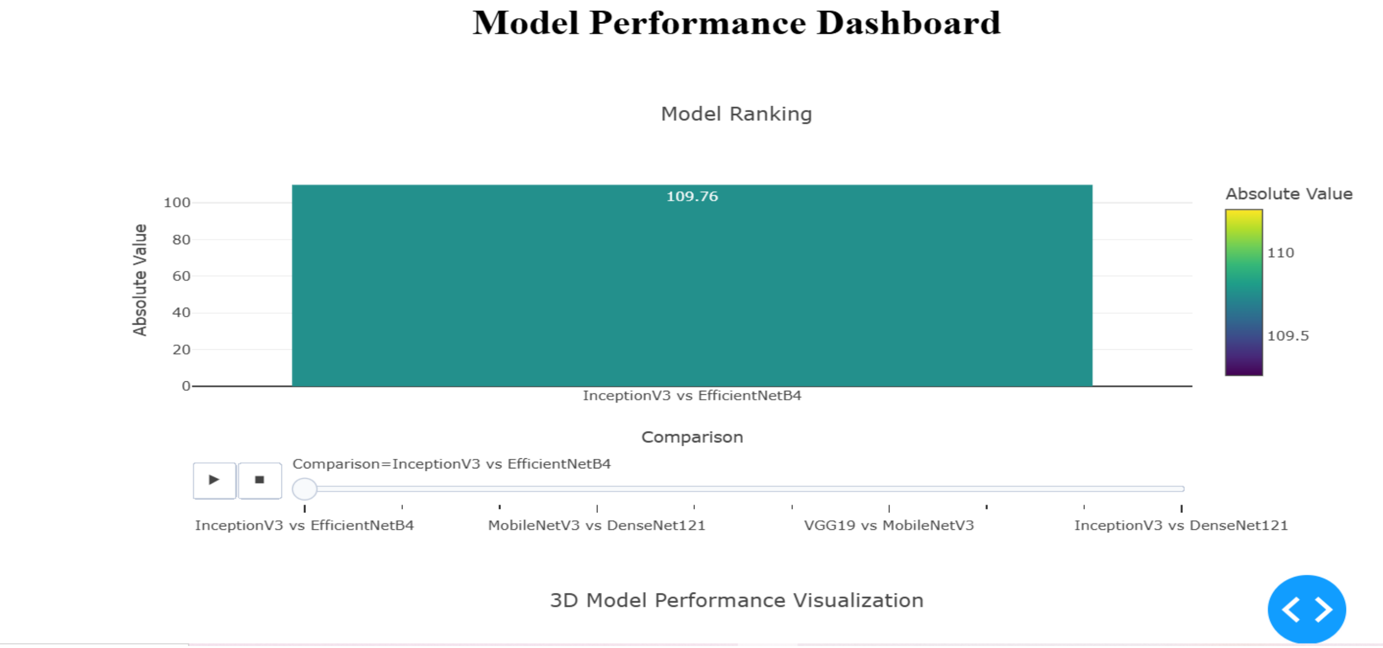
Model Performance Dashboard
Model Ranking
109.76
Absolute Value
100
80
60
40
20
0
Absolute Value
110
109.5
InceptionV3 vs EfficientNetB4
Comparison
▶
■
Comparison=InceptionV3 vs EfficientNetB4
InceptionV3 vs EfficientNetB4
MobileNetV3 vs DenseNet121
VGG19 vs MobileNetV3
InceptionV3 vs DenseNet121
3D Model Performance Visualization

Figure 5. Real-Time Model Performance Dashboard for Deployment Insights

The Model Performance Dashboard (Figure 5) represents more than visual assistance since it functions as a transformative instrument which provides vital strategic depth alongside clarity to this research field. The dashboard achieves its status as an essential model assessment tool through statistical analysis combination with visual presentation which delivers multiple meaningful advantages. This section will analyze the useful outcomes together with the essential reasons supporting them.

Visual representations of sophisticated number data on the dashboard provide instant performance insights about models. The dashboard analysis has led researchers to determine that both InceptionV3 and DenseNet121 produce outstanding outcomes.

Which models are struggling (like EfficientNetB4).

Research teams need absolute value together with test statistics for measuring differences between achieved model outcomes.

Scientists along with practitioners can decrease their research duration through instant model selection of optimal choices after skipping deep data table investigations.

2. Real-world application deployment requires model selection as an essential step since it determines overall success rate. Figure 5 helps in:

Pinpointing top-performing models for deployment (like DenseNet121).

Selection of deployment models requires attention to architectures that would damage accuracy and efficiency.

A vital process in practical model operations requires assessment of stability and reliability during deployment when working on critical projects such as healthcare diagnostics or predictive systems

3. Visual notification elements present in this dashboard provide users with both foreseeable development indications and performance alerts.

Visual examination of EfficientNetB4 indicates that it needs enhancements since its present performance is poor.

Researchers can determine their attention focus from the dashboard which enables them to optimize model performance through variable parameter alterations and different training techniques.performance.

4 All participants in model evaluation need clear explanations of evaluation results since not every person possesses data scientist expertise. Figure 5:

The findings become understandable for stakeholders through this approach that simplifies complex technical outputs.

Model choice outcomes become accessible for viewers who lack technical understanding about the modeling procedure.

This system creates harmonious understanding about model performance between data scientists and both software developers and executive decision-makers as well as programmers which leads to proper alignment for future actions.

5 The visual insights facilitate team members to make superior strategic decisions through accelerated decision cycles.

Prioritizing high-performing models for deployment.

The organization dedicates funds to enhance underperforming models.

Focusing on models with consistent performance across metrics like accuracy, F1 score, and AUC-ROC.

6. The dashboard functions as a reference point for upcoming research projects to deliver:

A clear record of current model performance.

The initial point enables researchers to launch in-depth examinations that seek explanations behind top model results.

Researchers can test ensembling methods and transfer learning platforms as well as hyperparameter adjustments using the gathered performance data.

Organizations gain better model selection opportunities through the essential organizational tool known as the Model Performance Dashboard. Organizations gain better models through this tool together with enhanced decision-making ability and improved communication and future innovation development. Each dashboard presentation of statistical data guarantees that decisions for developing superior machine learning solutions must use established evidence. The innovation provides both research-based and practical application solutions through its disruptive method.

5. Conclusion and Future Scope

This research provides a thorough and insightful analysis of deep learning architectures for breast cancer prediction, emphasizing the effectiveness of transfer learning in medical imaging. Among the five pre-trained CNN models evaluated—VGG19, MobileNetV3, InceptionV3, EfficientNetB4, and DenseNet121—DenseNet121 stood out as the most accurate and reliable model. It achieved an impressive accuracy of 89.92%, a precision rate of 0.900, a recall rate of 0.894, and an AUC-ROC score of 0.950, underscoring its robustness and clinical applicability. The implementation of key techniques such as data augmentation and class-weighted loss functions proved crucial in addressing class imbalance and enhancing model generalization. This study also demonstrated the power of transfer learning with pre-trained weights, enabling high accuracy even with limited data and significantly reducing both training time and computational costs. One of the major contributions of this research is its comprehensive multi-model comparative analysis, offering valuable insights into the performance and limitations of various architectures for medical image classification. DenseNet121 consistently outperformed other models, balancing high accuracy, low test loss, and strong generalization capabilities, while MobileNetV3 showed promise for real-time deployment due to its faster inference times. This study lays a solid foundation for future research aimed at advancing breast cancer diagnosis through deep learning. Future efforts should focus on enhancing model interpretability, integrating multi-modal data sources such as mammograms, blood test results, and genetic information, and developing hybrid models to improve diagnostic precision further. By leveraging cutting-edge deep learning techniques, this work paves the way for more accurate, efficient, and accessible breast cancer detection systems, ultimately contributing to improved patient outcomes and more informed clinical decision-making**.**

Data Availability
The dataset used in this research is publicly available and can be accessed from Kaggle at the following link: https://www.kaggle.com/datasets/sabahesaraki/breast-ultrasound-images-dataset.

The Breast Ultrasound Images Dataset (BUSI) consists of a total of 780 labeled ultrasound images, categorized into three classes: Normal, Benign, and Malignant. All images are in PNG format with a consistent resolution of 500×500 pixels.

Conflict of Interest
The author declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions
Ahmed Wagdy was solely responsible for the conceptualization, data curation, model implementation, and manuscript preparation of this research. All stages of the study—from dataset selection and preprocessing to the application of deep learning models and the interpretation of results—were conducted under his guidance and effort.

Author: Ahmed Wagdy
Affiliation: Cairo University, Cairo, Egypt
Email: ahmediswagdy@gmail.com

References

[1] E. Du-Crow, "Computer-aided detection in mammography," M.Sc. thesis, Univ. of Manchester, 2022.

[2] A. Evans *et al.*, "Breast ultrasound: Recommendations for information to women and referring physicians by the European Society of Breast Imaging," *Insights Imaging*, Vol.9, No.4, pp.449–461, 2018, doi: 10.1007/s13244-018-0637-8.

[3] G. Schueller, C. Schueller-Weidekamm, and T. H. Helbich, "Accuracy of ultrasound-guided, large-core needle breast biopsy," *Eur. Radiol.*, Vol.18, No.8, pp.1761–1773, 2008. doi: 10.1007/s00330-008-0913-8.

[4] X. Shi, C. Liang, and H. Wang, "Multiview robust graph-based clustering for cancer subtype identification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, Vol.20, No.2, pp.544–556, 2022, doi: 10.1109/TCBB.2022.3140846.

[5] H. Wang, G. Jiang, J. Peng, R. Deng, and X. Fu, "Towards adaptive consensus graph: Multi-view clustering via graph collaboration," *IEEE Trans. Multimedia*, Vol.24, pp.1–13, 2022, doi: 10.1109/TMM.2022.3142749.

[6] J. Bai, R. Posner, T. Wang, C. Yang, and S. Nabavi, "Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review," *Med. Image Anal.*, vol. 71, p. 102049, 2021, doi: 10.1016/j.media.2021.102049.

[7] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, Vol.33, No.12, pp.6999–7019, 2022. doi: 10.1109/TNNLS.2022.3156148.

[8] J. Zuluaga-Gomez, Z. Al Masry, K. Benaggoune, S. Meraghni, and N. Zerhouni, "A CNN-based methodology for breast cancer diagnosis using thermal images," *Comput. Methods Biomech. Biomed. Eng. Imag. Vis.*, Vol.9, No.2, pp.131–145, 2021, doi: 10.1080/21681163.2020.1796879.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, Vol.60, No.6, pp.84–90, 2017, doi: 10.1145/3065386.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, pp.770–778, 2016. doi: 10.1109/CVPR.2016.90.

[11]B. Kaur and R. Popli, "A comprehensive review and analysis of advancements in breast cancer detection and classification," in *Proc. 2024 3rd Edition of IEEE Delhi Section Flagship Conf. (DELCON)*, 2024, doi: 10.1109/DELCON64804.2024.10866898.

[12] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: A comprehensive review," *Multimedia Tools and Applications*, Aug. 2024. https://doi.org/10.1007/s11042-024-XXXXX

[13] M. Mahalakshmi, G. R. Charan, and G. Sharma, "Integrative breast cancer detection: A deep learning approach with multi-modal data fusion of mammograms, prescription and blood reports," in *Proc. 2024 Int. Conf. on Communication, Computing and Internet of Things (IC3IoT)*, 2024, doi: 10.1109/IC3IoT60841.2024.10550391.

[14] A. Alloqmani, Y. B. Abushark, and A. I. Khan, "Anomaly detection of breast cancer using deep learning," *Arabian Journal for Science and Engineering*, Vol.48, pp.10977–11002, 2023. doi: 10.1007/s13369-023-07914-2.

[15] S. S. Pandi, S. Anandhi, T. Kumaragurubaran, and B. Priyan, "Automatic breast cancer disease detection and diagnosis using learning algorithm," in *Proc. 2024 2nd Int. Conf. on Advances in Information Technology (ICAIT)*, 2024. doi: 10.1109/ICAIT61638.2024.10690336.

[16] S. Garg, K. Kaur, N. Kumar, and J. J. P. C. Rodrigues, "Hybrid deep-learning-based anomaly detection scheme for suspicious flow detection in SDN: A social multimedia perspective," *IEEE Trans. Multimedia*, Vol.21, No.3, pp.566–578, 2019. doi: 10.1109/TMM.2019.2893549.

[17] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, Vol.115, No.3, pp.211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Conf. Neural Information Processing Systems (NIPS'12)*, Lake Tahoe, NV, USA, Dec., pp.1097–1105, 2012. doi: 10.1145/3065386.

[19] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Comput. Methods Programs Biomed.*, Vol.127, pp.248–257, 2016. doi: 10.1016/j.cmpb.2015.12.014.

[20] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imaging*, Vol.3, No.3, pp.034501, 2016. doi: 10.1117/1.JMI.3.3.034501.

[21] J. Ferlay *et al.*, "Cancer statistics for the year 2020: An overview," *Int. J. Cancer*, Vol.149, No.4, pp.778–789, 2021. doi: 10.1002/ijc.33588.

[22] S. Lei *et al.*, "Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020," *Cancer Commun.*, Vol.41, No.12, pp.1183–1194, 2021. doi: 10.1002/cac2.12197.

[23] J. S. Marks *et al.*, "Implementing recommendations for the early detection of breast and cervical cancer among low-income women," *MMWR Recomm. Rep.*, Vol.49, pp.35–55, 2000.

AUTHORS PROFILE

Ahmed Wagdy Mostafa Mahmoud earned his Bachelor's degree in Computer Science from Cairo University in 2009 and his Master's degree in 2012. With *15 years of professional experience, he is currently working as a **Consultant Engineer in Artificial Intelligence, specializing in AI-driven solutions. His expertise spans **machine learning, deep learning, and intelligent systems*, where he has contributed to designing and implementing advanced AI models for real-world applications.