

Development of an Ensemble Deep Learning Model for Breast Cancer Diagnosis

Musibau Adekunle Ibrahim

kunle_ibrahim2001@yahoo.com

Osun State University

Taye Abidakun

Osun State University

Patrick Ozoh

Osun State University

Article

Keywords: Breast Cancer, Deep Learning Algorithm, BI-RADS Classification, InceptionResNetV2 model, ResNet-18 model

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10349400/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Musibau Adekunle Ibrahim^{1*}, Taye Abidakun², Patrick Ozoh³

¹ Department of Computer Science, Osogbo, Osun State University, Nigeria

² Department of Computer Science, Osogbo, Osun State University, Nigeria

³ Department of Computer Science, Osogbo, Osun State University, Nigeria

Corresponding author: Musibau Adekunle Ibrahim

E-mail: kunle_ibrahim2001@yahoo.com

Development of an Ensemble Deep Learning Model for Breast Cancer Diagnosis

Abstract

Breast cancer remains a major cause of mortality among women, and early detection is essential for improving treatment outcomes. A mammogram is a low-dose X-ray image of the breast used to detect abnormalities, including early signs of breast cancer before symptoms become apparent. Although artificial intelligence (AI) has enhanced mammogram analysis but many existing models trained on foreign datasets that may not adequately represent local clinical environment. This study proposes an ensemble deep learning model for classifying mammogram images into Breast Imaging Reporting and Data System (BI-RADS) categories using locally sourced data. Mammogram images obtained from Total Diagnostic Center, Ibadan, Nigeria, were preprocessed to improve image quality and ensure consistency. Two deep learning models, InceptionResNetV2 and ResNet-18, were developed and combined using an ensemble approach to classify the images into BI-RADS categories. The performance of the ensemble model was evaluated using accuracy, precision, recall, and F1-score. The experimental results show that the proposed ensemble model achieved an accuracy of 99.26%, a precision of 99.27%, a recall of 99.25%, and an F1-score of 99.25%. The ensemble model also outperformed the individual InceptionResNetV2 and ResNet-18 models, demonstrating its effectiveness in accurately classifying mammogram

images into BI-RADS categories. The study demonstrates the importance of developing AI models using locally representative datasets.

Keywords: Breast Cancer, Deep Learning Algorithm, BI-RADS Classification, InceptionResNetV2 model, ResNet-18 model

1 INTRODUCTION

1.1 Background of the study

Breast cancer is prevalent in women and poses a serious risk when diagnosis is delayed. Early and accurate diagnosis improves prognosis (Khan et al., 2025). According to Chang et al. (2025), mammography is widely used and has proven to be effective in detecting breast cancer. Even so, Quinn et al. (2026) noted that different radiologists can examine the same mammogram image and still arrive at different conclusions. To reduce this kind of variation, the BI-RADS classification system was introduced to guide how mammograms are interpreted and reported in a more consistent way (Dai et al., 2025). Shahi and Mishra (2026) explained that the BI-RADS system also provides guidance on how patients should be managed and monitored over time. Interpreting mammograms is difficult and depends on the radiologist's experience and judgment, this can lead to variability in diagnosis (Wu et al., 2025). Some signs in the images are very subtle and can be missed when the cases are complex or when there are many images to review. The workload can also be heavy, and fatigue can affect concentration, which may lead to wrong diagnosis (Chen et al., 2023).

Wang et al. (2025) noted that the development of computer-aided systems for medical image analysis has accelerated due to advances in artificial intelligence, particularly deep learning. Convolutional neural networks can directly extract valuable features from images and perform well in image categorization. As a result, diagnosis consistency is increased and manual feature extraction is decreased (Zhao et al., 2024).

Despite these advancements, many deep learning models are developed using datasets obtained outside Nigeria and may not properly represent the characteristics of the local population. Franklin et al. (2024) explained that bias in machine learning algorithms often originates from the data used for training, and such biases may reinforce inequalities and healthcare disparities. As a result, there is often a difference between the performance reported during research and the performance observed in real clinical settings. Differences in imaging equipment, patient demographics, breast tissue characteristics, and disease presentation can influence how well these models perform across different environments (Gao et al., 2026). Therefore, there is a need for AI models trained with locally representative mammogram data from Nigeria to ensure that the models are reliable, accurate, and suitable for use within the local healthcare system.

For this reason, this study develops an ensemble deep learning model trained on local data for classifying mammogram images according to BI-RADS categories. This study uses mammogram data collected from a diagnostic centre in Ibadan, Nigeria, so the dataset reflects the people in this environment. The model integrates two deep learning models into an ensemble to improve classification accuracy and make the results more reliable.

2. Literature Review

Several studies have explored the application of deep learning models for BI-RADS-based mammogram classification. Tsai et al. developed a deep neural network model based on the EfficientNet architecture for the classification of mammograms into BI-RADS categories 0, 1, 2, 3, 4A, 4B, 4C, and 5. The model used block-based segmented mammogram images resized to 224×224 pixels as input. The study achieved an overall accuracy of 94.22%, sensitivity of 95.31%, specificity of 99.15%, and an AUC of 97.23%. The study demonstrated that deep learning models can effectively classify mammograms into detailed BI-RADS categories with high accuracy. However, the dataset used in the study consisted mainly of mammograms from Taiwanese women, which may limit generalization to other populations. In addition, the block-based segmentation approach may lead to loss of important global breast contextual information.

Similarly, Sabani et al. applied a deep Convolutional Neural Network (dCNN) for the classification of mammographic soft tissue opacities based on BI-RADS categories. The study used 1,744 mammograms from 438 patients and applied a sliding window technique to generate probability maps for lesion visualization. The developed model achieved a real-world classification accuracy of 92.6%, sensitivity of 84.0%, specificity of 100%, and an AUC of 92%. The study showed that deep learning models can support observer-independent classification of mammographic abnormalities. However, the study focused mainly on soft tissue opacities and grouped BI-RADS classes into broader categories, which may limit detailed multiclass BI-RADS classification.

In another related work, Schöenberger et al. proposed a deep convolutional neural network for BI-RADS-based classification of mammographic microcalcifications. The researchers used more than 56,000 preprocessed images extracted from mammograms and trained different dCNN models for BI-RADS classification. The developed models achieved validation accuracies above 98%, while the best real-world test accuracy reached 80.9%. The study also generated probability maps for visualization of microcalcification distributions. Although the study demonstrated the effectiveness of deep learning for microcalcification analysis, the approach focused

mainly on microcalcifications and did not fully address broader mammographic abnormalities required for comprehensive breast cancer diagnosis.

Previous techniques have also been used for BI-RADS-based mammogram classification. Ozsahin et al. evaluated six stand-alone machine learning algorithms, including logistic regression, artificial neural network, Naïve Bayes, support vector machine, random forest, and K-nearest neighbor, for breast cancer classification. Among the evaluated models, support vector machine achieved the highest accuracy of 80.6% and precision of 76.8%, while logistic regression provided the best AUC of 84.6%. The result showed that machine learning algorithms can support breast cancer diagnosis and classification. However, traditional machine learning methods often rely on manually engineered features and may not effectively capture complex mammographic patterns compared to deep learning approaches. Although previous studies have demonstrated promising results in automated mammogram classification, several limitations still exist. Many studies focused on specific mammographic abnormalities such as microcalcifications or soft tissue opacities instead of utilizing full mammogram images. In addition, limited availability of locally representative datasets remains a challenge for developing models suitable for the Nigerian healthcare environment.

To address these limitations, the present study proposes an ensemble deep learning model for multiclass BI-RADS classification of mammogram images. The study combines the strengths of multiple convolutional neural networks to improve feature extraction, classification accuracy, and robustness. Unlike several previous studies that used segmented image patches or grouped BI-RADS categories, the proposed approach utilizes full mammogram images for more comprehensive learning of both local and global breast tissue characteristics. The developed model will also be trained using locally sourced mammogram data, thereby improving its relevance and applicability to the Nigerian population.

3. Materials and Methods

This section describes the materials, dataset, preprocessing techniques, and deep learning methods used for the development of the proposed ensemble model for BI-RADS classification of mammogram images. The study followed a systematic process beginning from data collection and preprocessing to model training, ensemble prediction, and performance evaluation. The data investigated consisted of mammogram images collected from Total Diagnostic Center, Ibadan, Nigeria. The images were categorized into BI-RADS classes based on radiologists' annotations and clinical reports. The dataset contained mammographic images representing different breast tissue conditions required for breast cancer diagnosis and classification. The mammogram images were grouped into the selected BI-RADS categories used in this study. The dataset was divided into training, validation, and testing sets to ensure proper model learning and evaluation. The training set was used for model learning, the validation set was used for monitoring model performance during training, while the testing set was used for final performance evaluation. The methodology adopted in this study involved several stages including data collection, data preprocessing, feature learning using deep learning models, ensemble learning, and model evaluation. The flowchart of the system is illustrated in Figure 1.

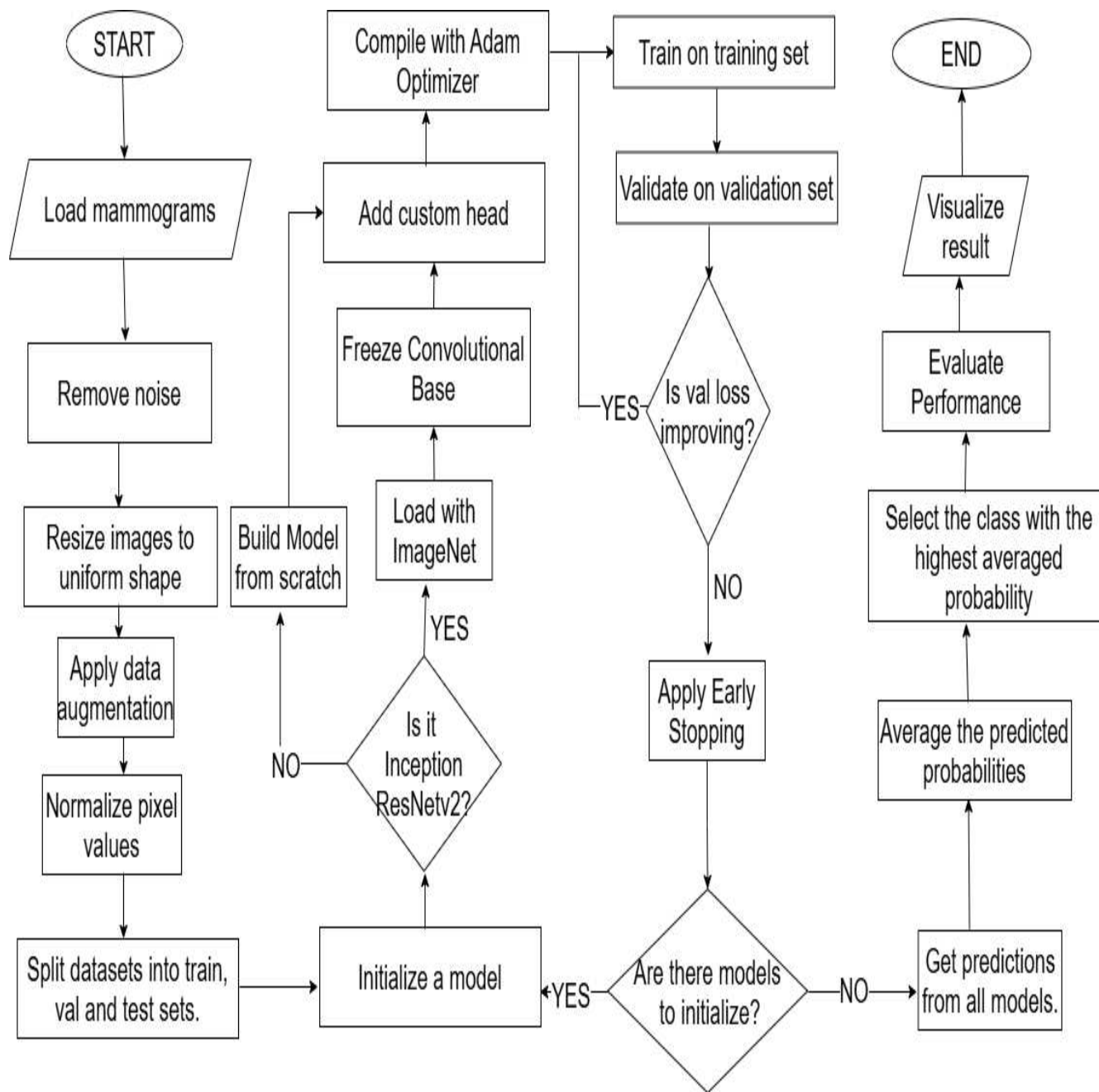


Figure 1: Flowchart of the ensemble Model

The methodology began with the collection of mammogram images from the medical diagnostic center. The acquired images were then subjected to preprocessing operations to improve image quality and prepare the dataset for deep learning classification. After preprocessing, the images were passed into the proposed ensemble deep learning model for BI-RADS classification. The final output generated by the model represented the predicted BI-RADS category of the mammogram image. Data preprocessing was carried out to improve the quality, consistency, and suitability of the mammogram images for deep learning classification. The preprocessing operations performed in this study are illustrated in Figure 2. The collected mammogram images were first organized according to their respective BI-RADS categories based on radiologists' annotations and clinical reports. The mammogram images were resized to a fixed image dimension compatible with the input requirements of the deep learning models. Resizing ensured that all images maintained a uniform shape and allowed efficient processing during training and classification. Image normalization was also performed to scale the pixel intensity values into a standardized range. This helped improve model convergence during training and reduced variations caused by differences in image brightness and contrast. For us to increase the level of robustness and generalization capability of the developed system, data normalization techniques were applied to the dataset. The augmentation process increased the number of mammogram images by a factor of 10, expanding the dataset from 184 images to 1840 images. The augmentation operations included image rotation, horizontal flipping, zooming, width shifting, and height shifting. These transformations generated multiple variations of the original mammogram images while preserving their diagnostic characteristics. The augmented images enabled the models to learn more discriminative and generalized features for BI-RADS classification. After augmentation and preprocessing, the dataset was divided into training, validation, and testing subsets. The training set was used for model learning, the validation set was used to monitor model performance during training, while the testing set was used for final evaluation of the proposed ensemble model. The preprocessed mammogram images were then supplied to the ensemble deep learning architecture for feature extraction and BI-RADS classification.

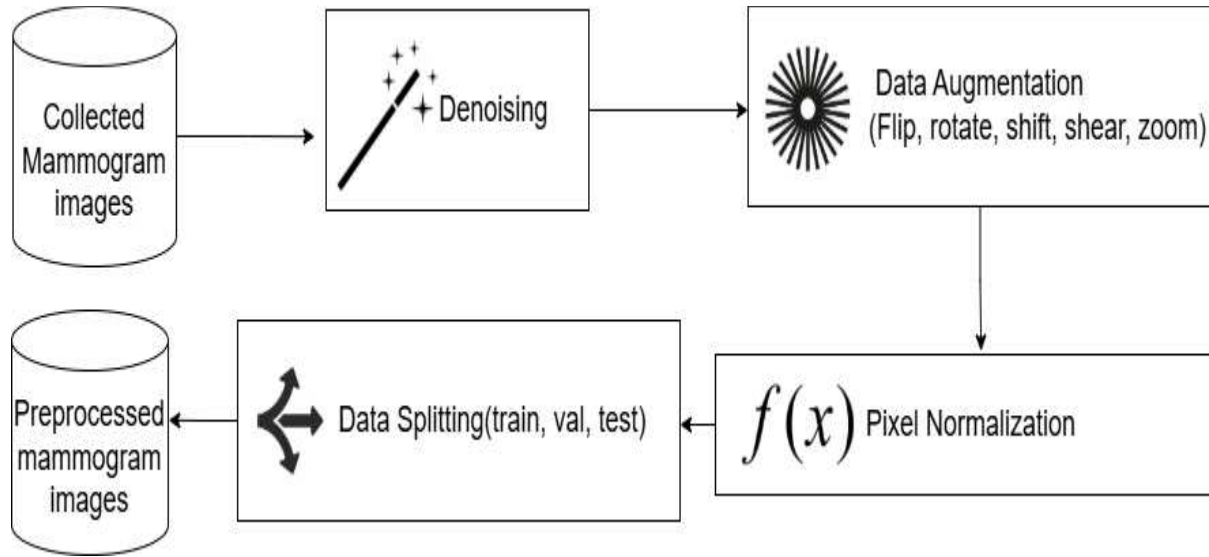


Figure 2: Data Preprocessing Architecture

The proposed ensemble architecture combined two convolutional neural network models, namely ResNet18 and InceptionResNetV2, for mammogram image classification. The architecture of the proposed ensemble model is shown in Figure 3. ResNet18 was used because of its residual learning capability, which helps in solving the vanishing gradient problem and enables efficient feature extraction from medical images. InceptionResNetV2 was selected due to its ability to capture complex image patterns through inception modules combined with residual connections. Each model independently received the preprocessed mammogram image as input and generated probability scores for the BI-RADS classes. The outputs from both models were then combined using a soft voting ensemble technique. In the soft voting approach, the prediction probabilities produced by the individual models were averaged to obtain the final classification result. The ensemble prediction can be expressed as the average of the probability outputs generated by the two models. The BI-RADS class with the best value combined probability score has been chosen as the best prediction output of the system.

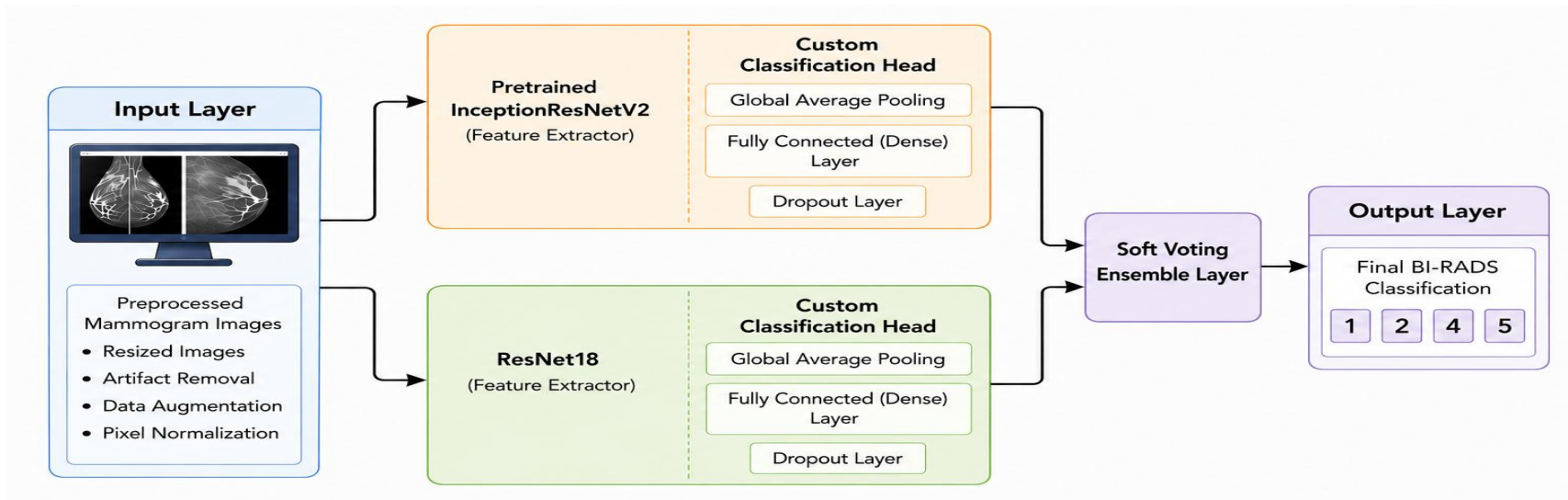


Figure 3: Model Architecture of the Ensemble Deep Learning Model

The deep learning models were trained using the training dataset while the validation dataset was used to monitor performance during training. Different training approaches were adopted for the two models used in the ensemble architecture. The InceptionResNetV2 model used transfer learning with pretrained weights obtained from the ImageNet dataset. The pretrained weights enabled the model to leverage previously learned image features, thereby improving feature extraction capability and reducing training time. In contrast, the ResNet18 model was trained from scratch using the mammogram dataset collected for this study. The model weights were randomly initialized and updated during training through backpropagation. Training the ResNet18 model from scratch allowed the network to learn features directly from the mammogram images without relying on previously learned representations from external datasets. The models were trained using the categorical cross-entropy loss function because the task involved multiclass classification.

The Adam optimizer was used for parameter optimization due to its efficiency and fast convergence properties. Model training was performed over multiple epochs with an appropriate batch size to achieve optimal classification performance. The performance of the proposed ensemble model was evaluated using standard classification metrics. These metrics included accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The evaluation metrics were calculated using the testing dataset to measure the effectiveness of the model in classifying mammogram images into the correct BI-RADS categories. Accuracy was used to measure the proportion of correctly classified mammogram images. Precision measured the ability of the model to correctly identify positive predictions. Recall evaluated the model's ability to detect relevant cases within each BI-RADS category. The F1-score was used to provide a balanced measure of precision and recall. The AUC-ROC metric was used to evaluate the classification capability of the model by measuring the trade-off between sensitivity and specificity across different classification thresholds. A higher AUC-ROC value indicated better discrimination performance of the model in distinguishing between the BI-RADS classes.

A confusion matrix was also used to analyze the classification performance of the model across the different BI-RADS categories. The evaluation results obtained from the ensemble model were compared with those of the individual deep learning models to determine the effectiveness of the ensemble approach in improving breast cancer diagnosis performance.

Ethics Approval and Consent Statement

I hereby confirm that all experiments were performed in accordance with relevant guidelines and regulations.

All experimental protocols were approved by the Biomedical Imaging and Computer Vision Research Group (BICVRG), Osun State University, Nigeria.

Furthermore, I confirm that:

- i. All methods were carried out in accordance with relevant guidelines and regulations.
- ii. All experimental protocols were approved by BICVRG, Osun State University.
- iii. Informed consent was obtained from all participants.
- iv. The study was conducted in accordance with the guidelines and regulations of Osun State University.

4. Results and Discussion

This section presents and discusses the performance of the developed deep learning models for BI-RADS classification of mammogram images. The models were evaluated using precision, recall, F1-score, accuracy, confusion matrix, and Area Under the Receiver Operating Characteristic Curve (AUC–ROC). The performance results are presented in Tables 1 to 3 and Figures 4 to 7. The classification report for the InceptionResNetV2 model is presented in Table 1. The model achieved an overall classification accuracy of 98.52%, with high precision, recall, and F1-score values across all BI-RADS categories. The results indicate that the model was highly effective in learning discriminative features from the mammogram images. The model achieved perfect recall values for BI-RADS 1 and BI-RADS 5, indicating that all mammograms belonging to these classes were correctly identified. Only one BI-RADS 2 image was incorrectly classified. BI-RADS 4 recorded a slightly lower recall value of 0.9552, suggesting that a few samples in this category were misclassified. Similarly, the class accuracy values demonstrate strong model performance, with BI-RADS 1 and BI-RADS 5 achieving perfect accuracies of 1.0000, while BI-RADS 2 and BI-RADS 4 achieved accuracies of 0.9853 and 0.9552, respectively. The confusion matrix shown in Figure 4 provides further insight into the model predictions. All 67 BI-RADS 1 images and all 68 BI-RADS 5 images were correctly classified. In BI-RADS 2, one image was misclassified as BI-RADS 4. For BI-RADS 4,

one image was incorrectly predicted as BI-RADS 1, while two images were predicted as BI-RADS 5. These findings suggest that BI-RADS 4 presents greater classification difficulty due to feature similarities shared with both benign and malignant breast tissue patterns. The AUC-ROC curve for the InceptionResNetV2 model is presented in Figure 5. The model achieved very high AUC values across all classes, with a macro-average AUC of 0.9991. This indicates excellent class separability and strong predictive capability.

Table 1: Classification Report for InceptionResNetV2

BI-RADS Class	Precision	Recall	f1-score	Accuracy
BI-RADS 1	0.9853	1.0000	0.9926	1.0000
BI-RADS 2	1.0000	0.9853	0.9926	0.9853
BI-RADS 4	0.9846	0.9552	0.9697	0.9552
BI-RADS 5	0.9714	1.0000	0.9855	1.0000
Overall Performance	0.9853	0.9851	0.9851	0.9852

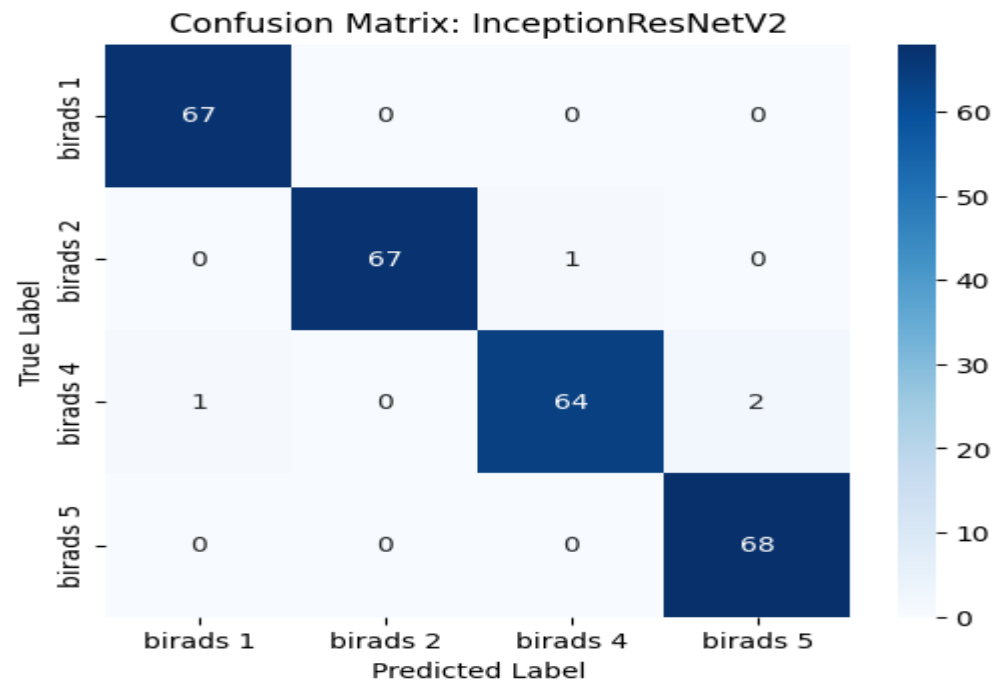


Figure 4 : Confusion Matrix for InceptionResNetV2 model.

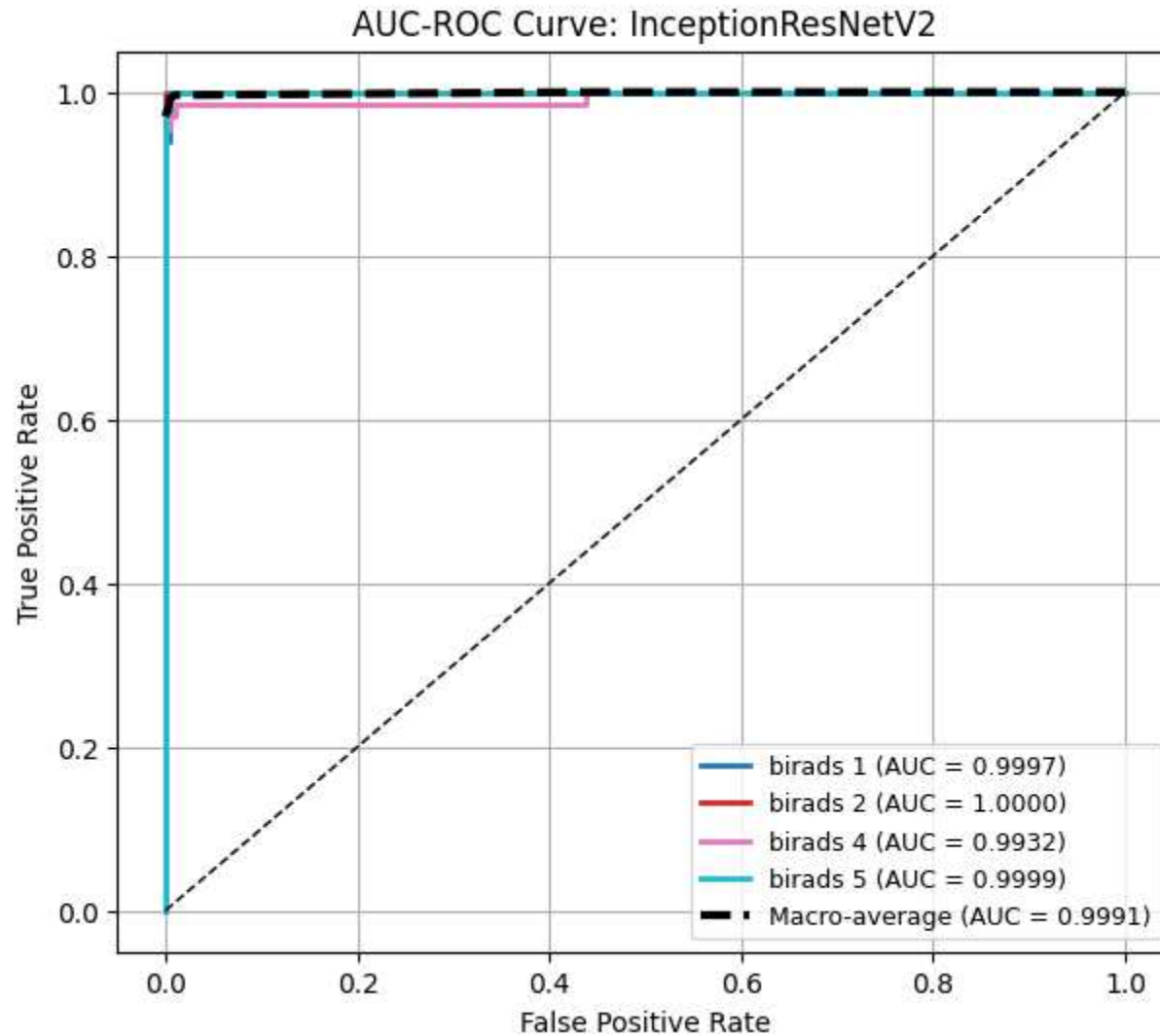


Figure 5: AUC-ROC Curve for InceptionResNetV2 model.

Table 2 presents the classification performance of the ResNet18 model. The model achieved an overall accuracy of 92.22%. Although the model demonstrated good performance across the BI-RADS categories, its performance was lower than that of InceptionResNetV2. BI-RADS 5 recorded the lowest precision value of 0.8421, indicating that some predictions for this category were

incorrect. The class accuracy values for BI-RADS 1, BI-RADS 2, BI-RADS 4, and BI-RADS 5 were 0.9254, 0.9265, 0.8955, and 0.9412, respectively. These values further reflect the comparatively weaker performance of the ResNet18 model. The confusion matrix shown in Figure 4.3 reveals that several misclassifications occurred across all BI-RADS classes. Five images were misclassified in BI-RADS 1, five in BI-RADS 2, seven in BI-RADS 4, and four in BI-RADS 5. The larger number of classification errors suggests that ResNet18 had greater difficulty learning complex mammographic features compared to InceptionResNetV2. The AUC-ROC curve presented in Figure 7 shows that the ResNet18 model still achieved strong discriminative performance, with a macro-average AUC value of 0.9953 across all BI-RADS classes.

Table 2: Classification Report for ResNet18

BI-RADS Class	Precision	Recall	F1-score	Accuracy
BI-RADS 1	0.9538	0.9254	0.9394	0.9254
BI-RADS 2	0.9265	0.9265	0.9265	0.9265
BI-RADS 4	0.9836	0.8955	0.9375	0.8955
BI-RADS 5	0.8421	0.9412	0.8889	0.9412
Overall Performance	0.9265	0.9221	0.9231	0.9222

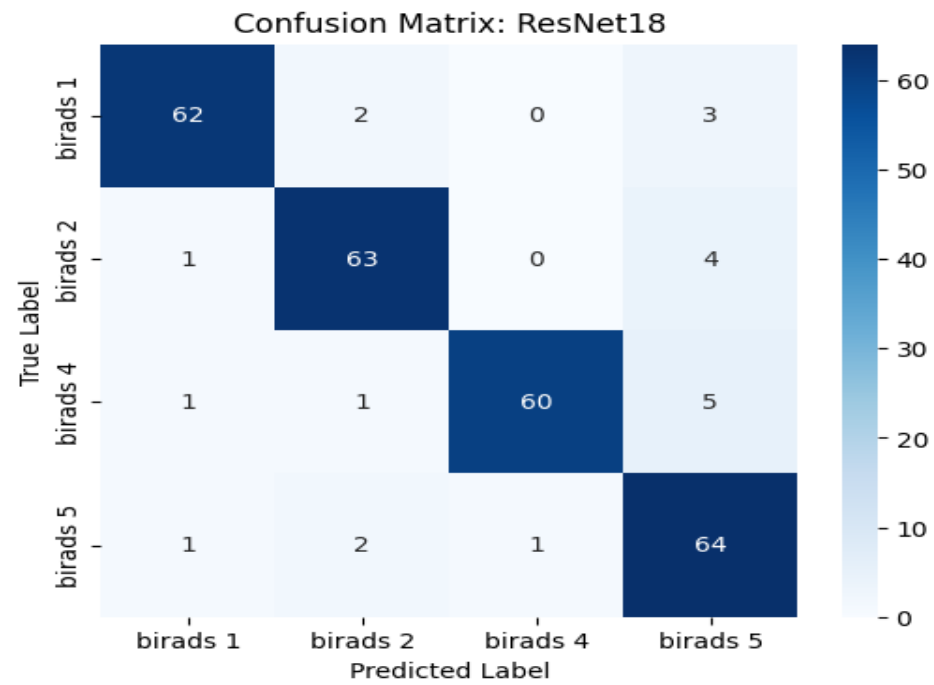


Figure 6: Confusion Matrix for ResNet18 model.

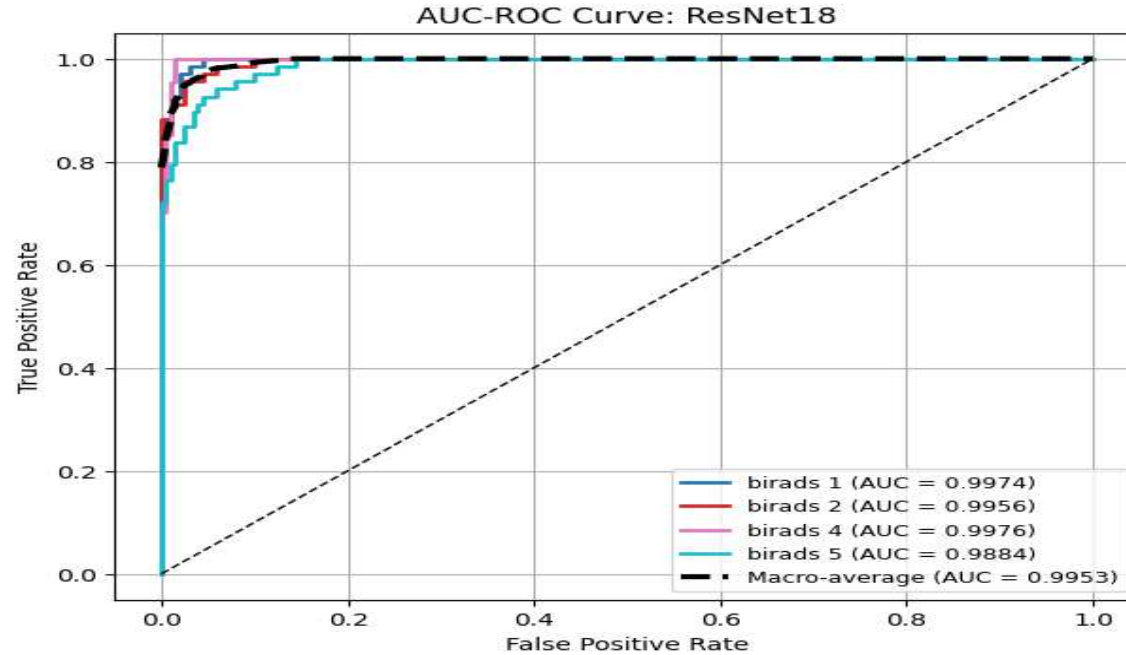


Figure 7: AUC-ROC Curve for ResNet18 model.

The classification report of the ensemble model is presented in Table 3. The ensemble model achieved the best overall performance among all evaluated models, with an overall accuracy of 99.26%. The model also achieved consistently high precision, recall, and F1-score values across all BI-RADS categories. BI-RADS 1 and BI-RADS 5 achieved perfect recall values, while BI-RADS 2 achieved perfect precision, recall, and F1-score values. The class accuracy results further confirm the effectiveness of the ensemble approach. BI-RADS 1, BI-RADS 2, and BI-RADS 5 achieved perfect accuracies of 1.0000, while BI-RADS 4 achieved an accuracy of 0.9701. The confusion matrix presented in Figure 8 shows that all samples belonging to BI-RADS 1, BI-RADS 2, and BI-RADS 5 were correctly classified. Only two BI-RADS 4 images were misclassified. This demonstrates that the ensemble model significantly reduced the classification errors observed in the individual models. The AUC-ROC curve for the ensemble model is shown in Figure

9. The model achieved perfect AUC values of 1.0000 across all BI-RADS categories, indicating outstanding classification performance.

Table 3: Classification Report for the Ensemble Model

BI-RADS Class	Precision	Recall	F1-score	Accuracy
BI-RADS 1	0.9853	1.0000	0.9926	1.0000
BI-RADS 2	1.0000	1.0000	1.0000	1.0000
BI-RADS 4	1.0000	0.9701	0.9848	0.9701
BI-RADS 5	0.9855	1.0000	0.9927	1.0000
Overall Performance	0.9927	0.9925	0.9925	0.9926

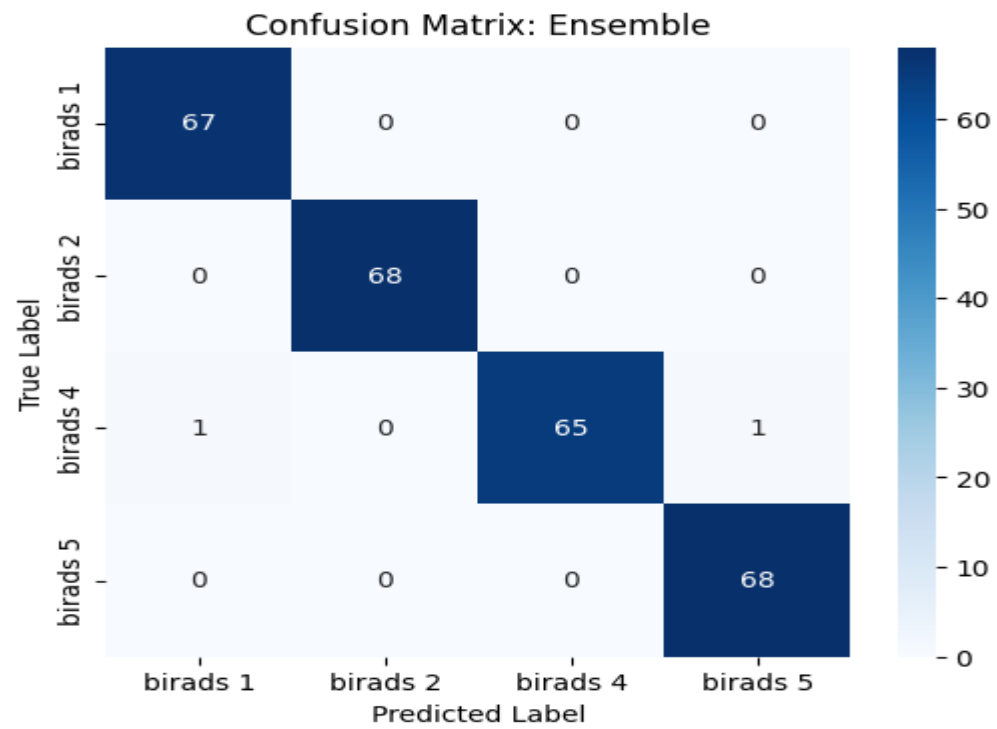


Figure 8: Confusion Matrix for the ensemble model.

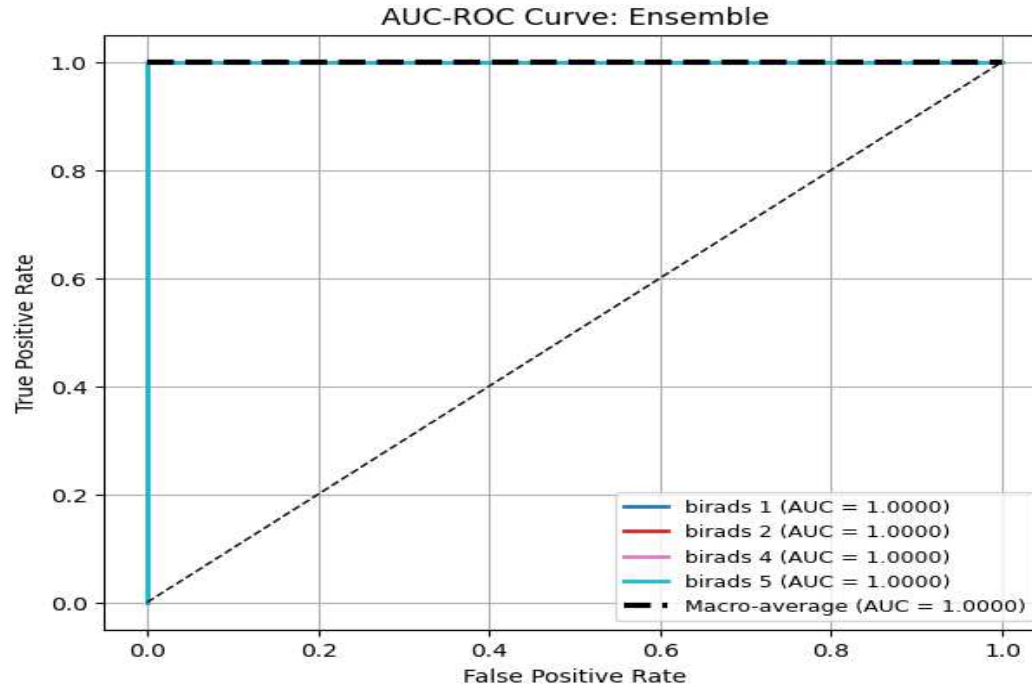


Figure 9: AUC-ROC Curve for the ensemble model.

Figure 10 compares the performances of InceptionResNetV2, ResNet18, and the ensemble model across all evaluation metrics. The ensemble model consistently outperformed both individual models in terms of accuracy, precision, recall, F1-score, and AUC. Among the individual models, InceptionResNetV2 performed better than ResNet18. The stronger performance of InceptionResNetV2 may be attributed to its deeper architecture and improved feature extraction capability through inception modules and residual connections. In addition, the use of transfer learning may have contributed significantly to its performance. Since the model was initialized with pretrained ImageNet weights, it already possessed prior knowledge of important image features such as edges, shapes, and textures, enabling more effective learning from the mammogram images.

In contrast, ResNet18 was trained entirely from scratch. As a result, the model had to learn all relevant image features directly from the dataset, which may have limited its ability to achieve the same level of performance as InceptionResNetV2. The superior performance of the ensemble model demonstrates the effectiveness of combining multiple deep learning models for medical image classification. By integrating the strengths of both architectures through soft voting, the ensemble model reduced individual model weaknesses and produced more reliable predictions. Across all models, BI-RADS 4 remained the most challenging category to classify. This category consistently recorded the highest number of misclassifications. This finding is expected because BI-RADS 4 cases often contain suspicious features that overlap with both benign and malignant tissue characteristics, making them more difficult to distinguish accurately. The results demonstrate that deep learning techniques are highly effective for mammogram image classification. The findings further show that ensemble learning can significantly improve classification performance and prediction reliability in BI-RADS-based breast cancer diagnosis.

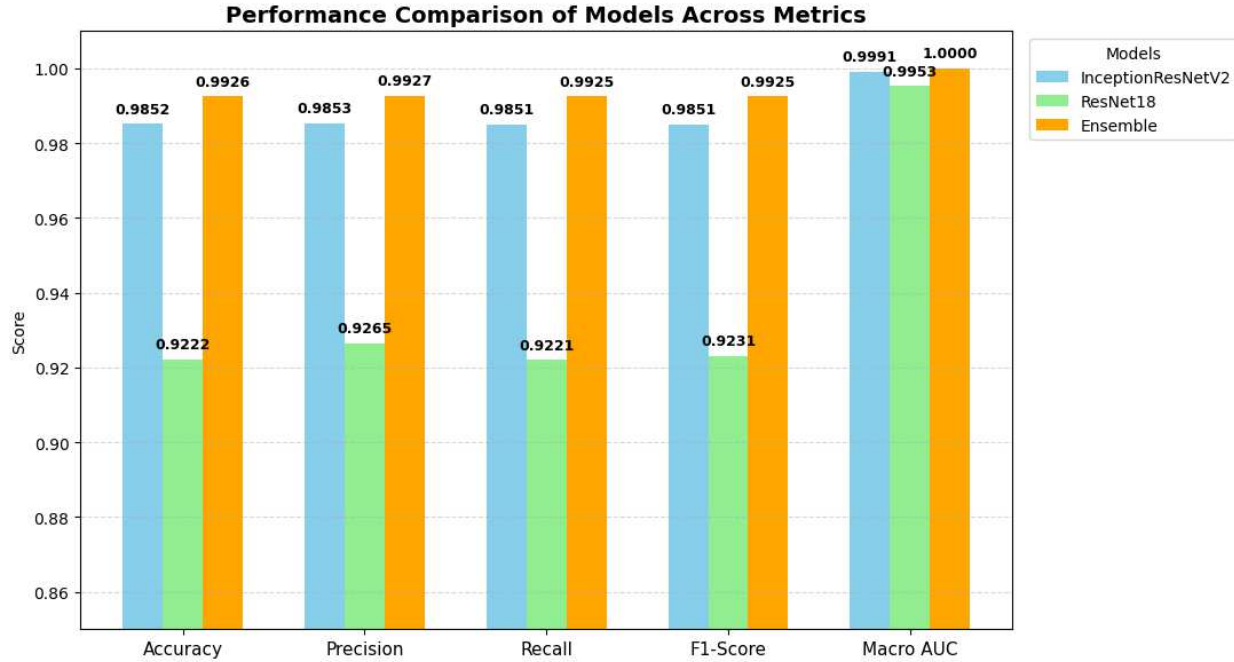


Figure 10: Performance comparison of the three models.

5 Conclusion

This study developed an ensemble deep learning model for the categorization of images into BI-RADS categories using locally sourced data. Two deep learning models, ResNet18 and InceptionResNetV2, were developed and evaluated for BI-RADS classification. The outputs of both models were then combined using an ensemble soft voting approach to improve classification performance. The results obtained from the experiments showed that all the models achieved strong classification performance across the BI-RADS categories. However, the ensemble model achieved the best overall performance, with higher accuracy, precision, recall, F1-score, and AUC values than the individual models. The findings of this study demonstrate that deep learning techniques can effectively support mammogram image analysis and breast cancer diagnosis. The results also show that model architecture plays an

important role in classification performance. InceptionResNetV2 performed better than ResNet18 probably due to its deeper architecture, stronger feature extraction capability, and the use of transfer learning. Furthermore, the ensemble approach improved prediction reliability by combining the strengths of both models and reducing classification errors. This study also contributes to knowledge in the areas of medical image analysis, computer-aided diagnosis, and deep learning. The research demonstrates the effectiveness of ensemble learning for multiclass mammogram classification and highlights the importance of using locally representative datasets for developing AI systems intended for healthcare applications in Nigeria. By training the models on locally sourced mammogram images, the developed system becomes more relevant to the target clinical population and environment. Future studies should focus on collecting larger and more diverse mammogram datasets from multiple hospitals and diagnostic centers across different regions of Nigeria. Future work should also include the BI-RADS 3 category to create a more complete classification system.

Funding: There is no funding for this study

Data availability statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. Additionally, all data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests: The author(s) declare no competing interests.

References

Chang, Y. W., Ryu, J. K., An, J. K., Choi, N., Park, Y. M., Ko, K. H., & Han, K. (2025). Artificial intelligence for breast cancer screening in mammography (AI-STREAM): Preliminary analysis of a prospective multicenter cohort study. *Nature Communications*, 16, 2248. <https://doi.org/10.1038/s41467-025-57469-3>.

Chen, Y., Sudin, E. S., Partridge, G. J., Taib, A. G., Darker, I. T., Phillips, P., James, J. J., Satchithananda, K., Sharma, N., & Michell, M. J. (2023). Measuring reader fatigue in the interpretation of screening digital breast tomosynthesis (DBT). *The British Journal of Radiology*, 96(1143), 20220629. <https://doi.org/10.1259/bjr.20220629>.

Dai, X., Ke, M., Xie, D., Mei, M., Wei, S., Dai, Y., & Yan, R. (2025). Performance of ChatGPT-4o, Claude 3 Opus, and DeepSeek-R1 in BI-RADS category 4 classification and malignancy prediction from mammography reports: Retrospective diagnostic study. *JMIR Medical Informatics*, 13. <https://doi.org/10.2196/80182>.

Franklin, G., Stephens, R., Piracha, M., Tiosano, S., Lehouillier, F., Koppel, R., & Elkin, P. L. (2024). The sociodemographic biases in machine learning algorithms: A biomedical informatics perspective. *Life*, 14(6), 652. <https://doi.org/10.3390/life14060652>.

Gao, S., Liu, J., Li, L., et al. (2026). Application of deep learning technology in breast cancer: A systematic review of segmentation, detection, and classification approaches. *BioMedical Engineering OnLine*, 25, 19. <https://doi.org/10.1186/s12938-025-01502-5>.

Khan, A. Q., Touseeq, M., Rehman, S., Tahir, M., Ashfaq, M., Jaffar, E., & Abbasi, S. F. (2025). Advances in breast cancer diagnosis: A comprehensive review of imaging, biosensors, and emerging wearable technologies. *Frontiers in Oncology*, 15, Article 1587517. <https://doi.org/10.3389/fonc.2025.1587517>.

Ozsahin, I., Uzun, B., Mustapha, M. T., Usanese, N., Yuvali, M., & Uzun Ozsahin, D. (2024). BI-RADS-based classification of breast cancer mammogram dataset using six stand-alone machine learning algorithms. In W. A. Zgallai & D. Uzun Ozsahin (Eds.), *Artificial intelligence and image processing in medical imaging* (pp. 195–216). Academic Press. <https://doi.org/10.1016/B978-0-323-95462-4.00008-X>.

Quinn, L., Jenkinson, D., Taylor-Phillips, S., Takwoingi, Y., & Sitch, A. (2026). Interobserver variability of recall decisions between mammography readers in the English NHS breast screening programme: A comparison of interobserver variability measures. *European Journal of Radiology*, 197, 112723. <https://doi.org/10.1016/j.ejrad.2026.112723>.

Sabani, A., Landsmann, A., Hejduk, P., Schmidt, C., Marcon, M., Borkowski, K., Rossi, C., Ciritsis, A., & Boss, A. (2022). BI-RADS-based classification of mammographic soft tissue opacities using a deep convolutional neural network. *Diagnostics*, 12(7), 1564. <https://doi.org/10.3390/diagnostics12071564>.

Schönenberger, C., Hejduk, P., Ciritsis, A., Marcon, M., Rossi, C., & Boss, A. (2021). Classification of mammographic breast microcalcifications using a deep convolutional neural network: A BI-RADS-based approach. *Investigative Radiology*, 56(4), 224–231. <https://doi.org/10.1097/RLI.0000000000000729>.

Shahi, A., & Mishra, S. S. (2026). Recent advances in breast imaging reporting: A review of the BI-RADS® sixth edition. *International Journal of Community Medicine and Public Health*, 13(3), 1543–1548. <https://doi.org/10.18203/2394-6040.ijcmph20260347>.

Tsai, K. J., Chou, M. C., Li, H. M., Liu, S. T., Hsu, J. H., Yeh, W. C., Hung, C. M., Yeh, C. Y., & Hwang, S. H. (2022). A high-performance deep neural network model for BI-RADS classification of screening mammography. *Sensors*, 22(3), 1160. <https://doi.org/10.3390/s22031160>.

Wang, L., Zhang, S., Xu, N., He, Q., Zhu, Y., Chang, Z., Wu, Y., Wang, H., Qi, S., Zhang, L., Shi, Y., Qu, X., Zhou, X., & Song, J. (2025). Role of artificial intelligence in medical image analysis. *Chinese Medical Journal*, 138(22), 2879–2894. <https://doi.org/10.1097/CM9.00000000000003824>.

Wu, H., Li, M., Chen, J., Fan, X., He, M., Ji, Z., Wang, H., & Yin, X. (2025). Reflections on training and management of junior radiologists through analysis of correction rates in emergency radiology reports. *BMC Medical Education*, 25(1), 879. <https://doi.org/10.1186/s12909-025-07573-1>.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57, 99. <https://doi.org/10.1007/s10462-024-10721-6>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Datasets.zip](#)
- [BreastImage.zip](#)