

Enhancing Breast Cancer Diagnosis with Ensemble Learning: Leveraging Convolutional Neural Networks and Pretrained Models through Averaged Predictions

Elia Setiana^{1*}, Reni Nursyanti², Nur Alamsyah³, Nayla Nurul Azkiya⁴

Informatics^{1*,2,4}
Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia^{1*,2,3,4}
<https://unibi.ac.id> ^{1*,2,3,4}
elia.setiana@unibi.ac.id^{1*}

Abstract. Breast cancer continues to be a serious health issue at the global level, and early detection can significantly improve patient outcomes. This research uses imaging techniques to examine the design of an improved classification model in breast cancer detection. This project uses deep learning approaches through Convolutional Neural Networks (CNN) and ensemble learning models to potentially improve classification accuracy. To further enhance performance while controlling for class imbalance and overfitting, we leverage several models, such as ResNet18 and VGG16, with data augmentation and pre-trained models. Our methods included standard preprocessing of medical images, splitting datasets into training, testing and validation sets, and training each model with the Adam optimizer. Performance measurement included accuracy, precision, recall, and F1 score metrics. Overall, prototypes recently created displayed clear advantages based on finding results achieved through an ensemble method, which demonstrated improved model stability and reduced significant misclassification errors, and model accuracy reached 0.96. This research is crucial while developing strong deep-learning models to aid in breast cancer detection, ultimately allowing us to set a base for developing better diagnostic inference systems in medical-based applications. These systems may help improve early detection and overall patient care.

Key words: breast cancer detection; deep learning; ensemble learning; convolutional neural networks; pre-trained models

1. INTRODUCTION

Breast cancer is still one of the most common and lethal diseases among women globally. Early detection is critical when dealing with breast cancer to provide patients with options for improving their outcomes, and earlier diagnosis leads to better opportunities and survival rates [1]. There has been a significant improvement in breast cancer detection methods as deep learning and ensemble learning have become beneficial and improved the field, particularly in the context of histopathology images [2]. Deep learning models, especially convolutional neural networks (CNNs), help analyze medical images and provide automated and accurate methods for detecting cancers in images [3]. There is continued variability in image quality class imbalance, requiring improved generalization for models [4]. Ensemble learning, which combines multiple models and predictions to gain greater accuracy and reduce overfitting, has developed into a strong approach to help solve these

problems [5]. The ability to capture and harness several components of several CNN models through an ensemble approach allows for enhancing the aspects of the data and enhancing predictive performance [6]. This has been used successfully in breast cancer diagnosis; for example, InceptionV3, EfficientNet, and ResNet have been used together to improve classification accuracy and provide stronger results [7]. Ensemble learning techniques have demonstrated potential in augmenting prediction accuracy and creating models that produce understandable outputs to the user, which includes interpretability and contributes to the understanding of the model used to make the determination [8].

Moreover, more and more studies have considered using pre-trained models and ensemble methods. The emergence of pre-trained models established and trained on large datasets gives a base for small domain-specific datasets to learn from, as in breast cancer diagnosis [9]. Combining a



pre-trained model and one's model assures that the system can generalize to new data while also learning to extract specific features in the domain [10]. The integration of deep learning and ensemble learning with pre-trained models can yield evidence for the accuracy and reliability of breast cancer detection systems [11].

This research investigates the development and evaluation of deep learning-based models to detect diseases, particularly breast cancer, from medical images. This research aimed to develop a disease classification model for medical images that achieved high accuracy with two classes labelled externally as 'normal' and 'sick'. In particular, we anticipate more classification performance than one model using Convolutional Neural Networks (CNNs), ResNet18 and VGG16 models using ensemble methods. This research also looks to mitigate any classification error through multi-model ensemble methods learned through training and explore data augmentation and pre-trained models to improve the model's ability to generalize to previously unseen data.

This research has provided several significant contributions to developing deep-learning classification methods for breast cancer detection. This study proposes a simple yet effective ensemble approach through a prediction averaging mechanism that combines a specially developed CNN model with two pretrained models, namely ResNet18 and VGG16. Unlike most previous studies that rely on complex ensemble architectures with multiple backbones or stacking strategies, this study demonstrates that a combination of three complementary models can improve classification stability while maintaining relatively low computational complexity. Furthermore, this study not only evaluates performance using classification metrics but also analyzes feature representation separation using t-SNE visualization to provide a better understanding of how the ensemble strategy affects feature distributions.

2. RELATED WORK

Deep learning has become one of the dominant approaches for breast cancer diagnosis because of its capability to automatically learn hierarchical image representations without requiring handcrafted feature extraction. Among various architectures, Convolutional Neural Networks (CNNs) have consistently demonstrated superior performance in analyzing histopathological and mammographic images by capturing complex spatial characteristics associated with benign and malignant tissues. Nevertheless, conventional CNN models often experience performance degradation when trained on limited medical datasets due to overfitting and insufficient generalization capability [12], [13].

To overcome these limitations, transfer learning has been widely adopted in medical image classification. Instead of training deep networks from scratch, pretrained architectures such as VGG16, VGG19, ResNet, DenseNet, EfficientNet, and Inception are initialized using large-scale

natural image datasets before being fine-tuned for breast cancer diagnosis. This strategy enables the models to retain generic visual representations while adapting to domain-specific pathological characteristics, leading to faster convergence and improved classification performance, particularly when annotated medical datasets are relatively small [13], [14].

Recent studies have further demonstrated that ensemble learning can improve diagnostic robustness beyond the capability of individual deep learning models. Rather than relying on a single architecture, ensemble approaches integrate predictions from multiple classifiers to reduce model variance and compensate for individual prediction errors. Hameed et al. (2020) proposed an averaging ensemble of fine-tuned VGG16 and VGG19 models for breast cancer histopathology classification and reported higher sensitivity and overall accuracy than the corresponding individual networks [12]. Their findings indicate that prediction averaging effectively exploits complementary feature representations extracted by different pretrained architectures.

The effectiveness of ensemble strategies has also been validated using more heterogeneous network combinations. Balasubramanian et al. (2024) developed an ensemble framework combining VGG16, ResNet34, and ResNet50 for breast cancer histopathology image analysis on the BACH and BreakHis datasets [13]. Their results demonstrated that integrating multiple CNN architectures substantially improved classification reliability while maintaining strong performance across different datasets and imaging conditions. These findings suggest that ensemble learning contributes not only to higher predictive accuracy but also to better model stability and generalization.

Beyond conventional ensemble CNNs, several researchers have explored more sophisticated hybrid frameworks. Munshi et al. (2024) introduced an optimized ensemble learning framework that combines transfer learning-based image analysis with handcrafted numerical features and machine learning classifiers supported by explainable artificial intelligence (XAI) [15]. Their framework integrates CNN-derived deep features with Random Forest and Support Vector Machine classifiers, illustrating that heterogeneous ensembles can effectively leverage complementary information from multiple feature spaces. The evolution of ensemble learning has also expanded toward integrating modern deep learning architectures. Recent investigations have combined EfficientNet, Vision Transformer (ViT), Inception, and other advanced backbone networks through stacking or averaging strategies, demonstrating further improvements in classification accuracy and robustness for breast cancer diagnosis. These studies consistently report that model diversity is an important factor in improving predictive performance because different architectures capture distinct visual characteristics from histopathological images [16].



Despite these encouraging developments, several research gaps remain. First, many ensemble studies employ relatively complex architectures involving numerous backbone networks or multi-stage optimization procedures, increasing computational cost and limiting practical deployment. Second, most studies primarily emphasize classification performance while providing limited visualization of feature separability after ensemble integration. Third, although pretrained CNNs such as ResNet and VGG have individually demonstrated strong discriminative capability, comparatively fewer studies investigate lightweight prediction-level averaging between a custom CNN and complementary pretrained networks while simultaneously evaluating class distribution through low-dimensional visualization techniques.

3. METHODS

This research study aims to develop and evaluate a deep-learning model to detect diseases in medical imaging, especially breast cancer. There are a lot of steps in the research process, including data preparation, model training, and testing the model. Below is a detailed description of each stage in Figure 1.

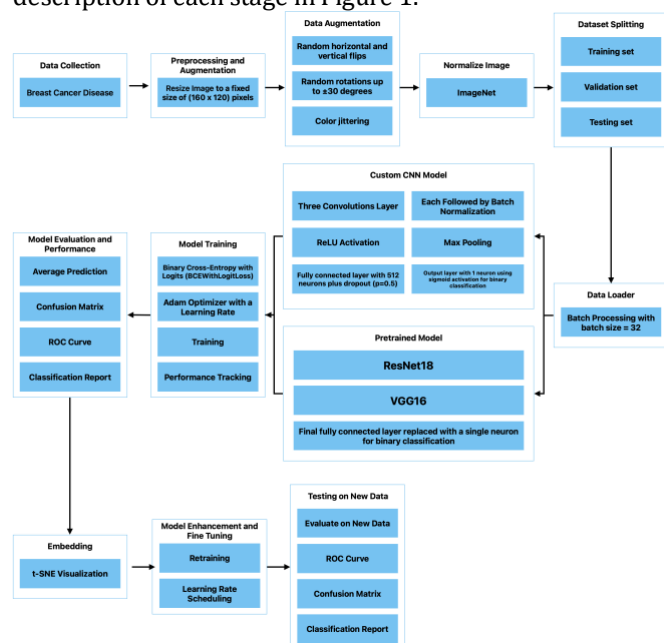


Fig.1. Proposed Method

3.1. Data Collection

The breast cancer image data set encoded comes from different data sources, mainly Kaggle, which provided preprocessed data for this study. The images in the dataset reflect individual patients' breast cancer conditions, and this information will be used to train the model. Figure 2 is the sample dataset. The dataset consists of two classes: "sick", which contains 101 images and "normal", which includes 163 images. The data quality is exceedingly

essential, and it will determine the extent to which the model can classify any pictures correctly. The good quality data led to better production of the models.

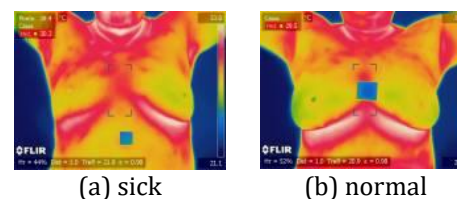


Fig. 2. Sample Dataset

3.2. Preprocessing and Augmentation

Following data collection, the dataset is preprocessed so all images are in standard form. Each image is resized to a standard size of 160x120 pixels so that the model can process the images more uniformly. Part of the preprocessed procedure allows for data augmentation, such as flipping images, rotating images, and adjusting colours. Allowing for data augmentation is to vary the training data and prevent overfitting of the model. Data augmentation is essential because we create variants of each image in the dataset, and the model can learn more features from the limited number of images in the data set.

3.3. Dataset Splitting

We then divide the preprocessed data into three parts: a training set (60%), a validation set (20%), and a test set (20%). The training set is used to train the model, the validation set is used to tune the model's parameters and avoid overfitting, and the test set is used to evaluate the performance of a trained model on new data. The distributions of the test set and training set and the validation set and test set are equal to 60:20 and 20:20, respectively. This relatively balanced splitting gives us confidence in the strength of the training model and the evaluation set against the trained model.

3.4. Data Loader

We use a DataLoader to process the data in batches. The DataLoader processes the data in batches of size 32, which is quicker for training and testing for data sets about 60 images in size. Using batches helps the model process images jointly while achieving some memory efficiency benefits. Batching is intended to compromise the memory efficiency aspect and training speed for images and improves model stability over individual image training experience.

3.5. Custom CNN Model

The mathematical model of the provided CNN architecture starts with an input image with shape (batch_size,3,160,120), where 3 is the three colour channels (RGB), and the image size is 160x120 pixels. The first operation is convolution, where a 2D-convolutional process



occurs for each convolutional layer. Formula (1) is the convolution formula, where x_j is the input to the previous layer, W_{ij} is the convolution filter and b_j is the bias.

$$y_i = \sum_j (x_j \cdot W_{ij} + b_i) \quad (1)$$

After convolution, the results are processed using the ReLU activation function as listed in formula (2), which introduces non-linearity to the model.

$$y_{\text{ReLU}} = \max(0, y) \quad (2)$$

After that, a max-pooling operation is performed to reduce the spatial dimension of the output, where this operation takes the maximum value of each local patch, i.e. in formula (3).

$$y_{\text{pool}} = \max(y_{\text{patch}}) \quad (3)$$

The output of the convolutional layer is then fed into the fully connected (FC) layer, which first converts the data into vector form and then performs the linear transformation contained in formula (4), where W is the weight matrix, and b is the bias vector.

$$y_{\text{fc}} = W \cdot x + b \quad (4)$$

Similarly, to mitigate overfitting in the first RF layer, we also employ dropout with probability $p = 0.5$, which is derived from formula (5), where the mask is a binary vector randomized during training.

$$y_{\text{dropout}} = y \cdot \text{mask} \quad (5)$$

The output finally passes through the sigmoid layer to produce a probability between 0 and 1 generated from formula (6), indicating the optimistic class prediction.

$$\hat{y} = \sigma(y_{\text{fc}}) = \frac{1}{1 + \exp(-y_{\text{fc}})} \quad (6)$$

During the training process, the model parameters are updated using Adam's optimizer with the update step in formula (7), where α is the learning rate and $\nabla_{\theta} L(\hat{y}, y)$ is the gradient of the loss function over the model parameters.

$$\theta \leftarrow \theta - \alpha \cdot \nabla_{\theta} L(\hat{y}, y) \quad (7)$$

3.6. Pretrained Models

This study's ResNet-18 and VGG-16 models have several main components involving convolution, activation, pooling, and fully connected layers. In ResNet-18, each residual block consists of two convolutional layers followed by a ReLU activation function and a shortcut connection, which can be seen in formula (8), where $F(x, W)$ is the

transformation result of the two convolutional layers, and x is the shortcut connection.

$$\hat{y} = F(x, W) + x \quad (8)$$

After several residual blocks, the output is flattened and passed through the fully connected layer calculated by formula (9) and finally processed by the sigmoid function in formula (10).

$$\hat{y}_{fc} = W_f \cdot \text{flatten}(x) + b_f \quad (9)$$

$$\hat{y} = \sigma(\hat{y}_{fc}) \quad (10)$$

VGG-16 uses several convolutional layers, a max pooling layer, and a fully connected layer. The convolutional layer applies convolution in formula (11), followed by pooling operation in formula (3). After the convolutional and pooling layers, the output is flattened and computed with a fully connected layer from formula (12), and the final output is calculated using the sigmoid function in formula (10).

$$y_i = \text{ReLU}(\text{Conv2D}(x, W_i) + b_i) \quad (11)$$

$$y_{fc1} = W_{fc1} \cdot \text{flatten}(x) + b_{fc1} \quad (12)$$

3.7. Model Training

Once the model had been created, training commenced using the Adam optimizer and BCEWithLogitsLoss as the loss function (13). The Adam optimizer adjusted the learning rate throughout training, allowing for a more rapid and stable convergence. The model was trained over several epochs, and each epoch consisted of a forward pass, performance measure, loss calculation, and backpropagation to update the model's weights.

$$L(\hat{y}, y) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (13)$$

3.8. Performance Tracking

While training, the model's performance is monitored by calculating accuracy and the loss at each epoch. Accuracy is the number of accurate predictions divided by the actual total number of predictions. Loss is the distance between the predicted value and the exact value. These metrics allow for identifying the model learning well or poorly and if some adjustment is necessary.

3.9. Average Prediction

Once the training is complete, the three models, i.e. CNN, ResNet18, and VGG16, are combined using average prediction to improve the classification importance. Ensemble learning combines the three models to utilize the strengths of each model to produce more accurate predictions by reducing the bias and variance of individual



models. Each model outputs logits, which produce probabilities between 0 and 1 after applying a sigmoid function. After calculating the probability from each model, the next step is to average the probabilities for all the models. The average for each image instance is calculated as formula (14)

$$\hat{p}_{avg} = \frac{1}{N} \sum_{i=1}^N \sigma(M_i(x)) \quad (14)$$

where p_i is the probability for the model M_i , and N is the total number of models. Thus, the final prediction is the average of the probability output by each model, and the goal is to harness the collective strength of numerous models to reduce uncertainty and maximize accuracy prediction.

3.10. Model Evaluation and Performance

The classification report lists the model's performance for each class and contains several metrics, such as accuracy, precision, recall, F1 score, and support. A confusion matrix assesses the accuracy of a model by comparing the predictions from the model for each class with the actual labels. It shows how many predictions were correctly and incorrectly labelled for each class. A confusion matrix gives some idea of the classes that are often confused. If there are some particular classes we might want to try to improve, it can also show which classes are. We will present the confusion matrix as a heatmap because it conveys the model's successes and errors due to misclassifications. The ROC curve visually depicts the relationship between the False Positive Rate (FPR) and True Positive Rate (TPR) to assess the model's ability to discriminate between classes.

3.11. Embedding and t-SNE Visualization

We use t-SNE to embed and visualize the data to understand better how the model maps it into feature space. t-SNE is a way to reduce the number of dimensions in data and show it in two dimensions. This will help us figure out how the model divides the classes.

3.12. Testing on New Data

At the concluding stage, the trained model is evaluated using entirely new data that was never utilized in any part of the training or validation. This allows one to assess the generalization of the model for unseen data. The new data is evaluated exactly as in the tri-stage of a model performance evaluation using a confusion matrix, ROC curve, and classification report to assess the model's performance for the new data and measure how well generalization to new data can be applied.

4. RESULTS AND DISCUSSIONS

4.1. Model Training

Figure 3 depicts the evaluation graph during training on the three models - CNN, ResNet18, and VGG16. There are notable differences in convergence speed and model performance consistency.

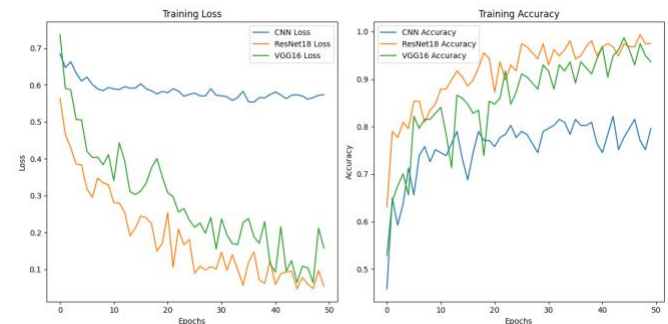


Fig. 2. Model Training

The CNN had an initial training loss of 0.6850 in the first epoch. The CNN model initially struggled to learn the data, but the training was a stepwise decline in loss. There were variances in loss for some of the 50 epochs. The model's accuracy started from 0.4586 at the first epoch to 0.7962 at the 50th epoch. While there is good improvement, the accuracy of this model could not reach the accuracy of the other models, indicating that simple CNN architecture is possibly not the best architecture for handling more complex data. Although the overall improvement in accuracy, 0.3376, shows a positive advance, the variances in accuracy do lend to an absence of consistency.

In contrast, ResNet18, which utilizes a deep learning architecture with RESIDUAL connections, performed better. The model commenced training at a starting loss of 0.5637, then decreased that loss significantly faster, reaching a loss of 0.0536 at the 50th epoch. The increase in accuracy is also very substantial, as it started at an accuracy of 0.6306 in the first epoch and concluded at 0.9363 in the last epoch. The loss and accuracy numbers quickly give a clear viewpoint of ResNet18. As a result of using residual connections, the model is learning deeper and faster in contrast to CNN. This model is also more stable and efficient learning than CNN and is the best option. The steady and stable 0.3057 increase in accuracy also shows that ResNet18 learned better and prevented larger fluctuations in its training.

On the other hand, VGG16 did perform reasonably well even though it was not as effective as ResNet18. It began training with a loss of 0.7370, higher than the other two models. While it ended training with a lower loss of 0.1576, VGG16's larger fluctuations in loss imply that it is more volatile/fits the training data better. It likely needs more hyperparameter settings or tuning to smooth out performance. The accuracy value for VGG16 starts at 0.5287 and goes to 0.9363, with a gain of 0.4076. However,



the more significant fluctuations of the accuracy graph are indicative again of VGG16 needing more regularisation or data augmentation of other techniques to build performance consistency. The accuracy gain of 0.4076 is significant. However, the fluctuation of the accuracy graph that VGG16 appears to experience will make it perform less stable than ResNet18.

In the end, ResNet18 was the best model for fast loss reduction and systematic accuracy improvement, making it the most suitable for more challenging classified tasks. VGG16 was still a solid model but also required more attention in terms of regularisation and stability. CNN was also a simpler classifier and garnered decent results, but CNN accuracy and stability results could not match trained deep learning models.

4.2. Model Evaluation and Performance

Table 1 presents the classification report, and the model demonstrates perfect precision, recall, and F1-score across both classes. Precision for the Normal class is 1.0, meaning there are no false positives in this class, and precision for the Sick class is 0.94, which is the rate of correct predictions to the number of optimistic predictions. Recall for the Sick class is also very high and is equal to 1, which means that all Sick examples had been correctly identified. However, recall for the Normal class is 0.89. While the prediction is perfect for the Normal classes, it does reflect that some typical examples were not predicted as usual. The F1-score for both classes is also very high; the Normal F1-score equals 0.94, and the Sick equals 0.97, meaning that precision and recall are at a good balance. Overall, the model performed well, with a 0.96 prediction accuracy for both the majority and minority classes. Overall, the model is a very effective and efficient classification model, which has a weakness in recall for the 'Normal' class; however, the model's performance on the 'Sick' class was exceptional in that recall for the Sick class was perfect with high precision. The model has an accuracy of 0.96 and a solid F1-Score for both classes, making it ideally suited to applications where the detection of 'Sick' conditions is required with a high level of sensitivity.

TABLE 1. Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	1.00	0.89	0.94	19
Sick	0.94	1.00	0.97	34
Accuracy			0.96	53
Macro avg	0.97	0.95	0.96	53
Weighted avg	0.96	0.96	0.96	53

Figure 3 is the confusion matrix. The model works well at differentiating between the classes Normal and Sick. The model correctly identified 34 sick cases and 19 normal cases; therefore, the model was able to distinguish both classes with fairly accurate predictions. However, although

the model accuracy is relatively high, some Normal cases showed as Sick, which is part of the precision of 0.94 for the 'Sick' class and recall of 0.89 for the 'Normal' class. This indicates that we have false positives and negatives in predictions, particularly in the Normal class. So, while the model works very well, we can improve the false negatives in the 'Normal' class and false positives in the 'Sick' class.

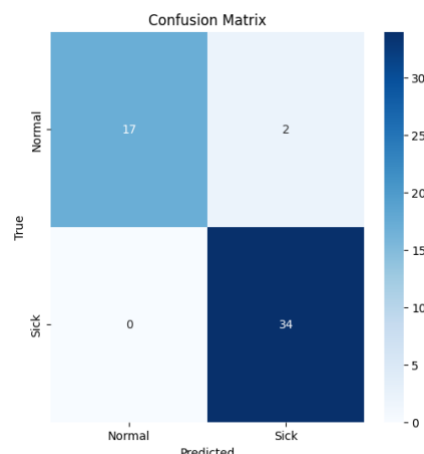


Fig. 3. Confusion Matrix

Figure 4 shows an ROC curve that indicates how well the model can tell the difference between the "Sick" and "Normal" classes. The Area Under the Curve (AUC) is 0.97, which means the model is better at telling the difference between the two classes. An AUC close to 1 means the model can distinguish between positive and negative classes. The model is quite good at finding "Sick" cases and makes very few mistakes when categorizing "Normal" cases as "Sick." This is shown in the graph, where the True Positive Rate or Recall increases significantly while the False Positive Rate increases. This suggests that the model is good at finding "Sick" circumstances and doesn't make many mistakes when it says something is "Normal." With an AUC of 0.97, the ROC curve reveals that this model is suitable for applications that must accurately and sensitively identify the "Sick" class.

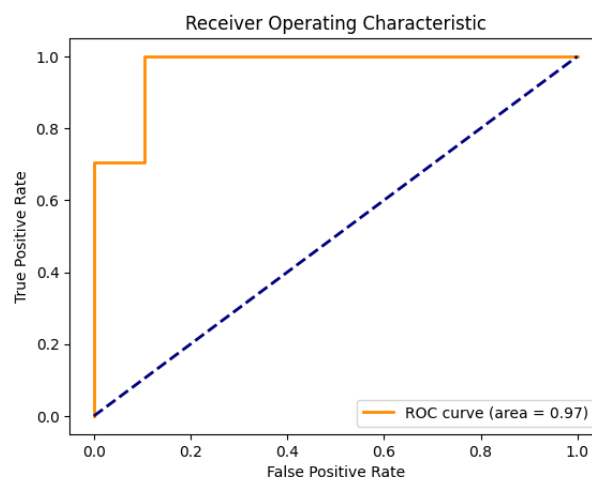


Fig. 4. ROC Curve

Figure 5 is a t-SNE graph that displays the mapping of the combined models' predictions in a lower dimensional space. It is based on the results from the ensemble of the three models using the average prediction method. To better understand if the ensemble learning successfully improves the separation between the 'Normal' and 'Sick' classes, the t-SNE is used to visualize the clusters of predictions from the combined models.

From the resulting t-SNE plots, there is a more evidently separation between the two classes following the ensemble; this demonstrates that by averaging the three models by their predictions, the ensemble successfully reduced the overlap between the "Normal" and "Sick" classes, the evidence of this is the separation of the clusters when comparing the ensemble representation with the prediction results of each of the individual models (CNN, ResNet18, or VGG16 individually).

Although there was some overlap, the ensemble of models successfully reduced misclassification errors that often occur in single models, such as false positives and false negatives, which were previously more pronounced in the CNN and VGG16 models. This suggests that ensemble techniques with prediction averaging can help improve model consistency and accuracy in more complex classification tasks.

The t-SNE plots derived from this ensemble show that combining the predictions improved class separation and enhanced the models' ability to detect a 'Sick' case. It improved the stability visible in the individual models. The more apparent separation in the current plot demonstrates the ensemble has succeeded in improving performance, with a better overall improvement.

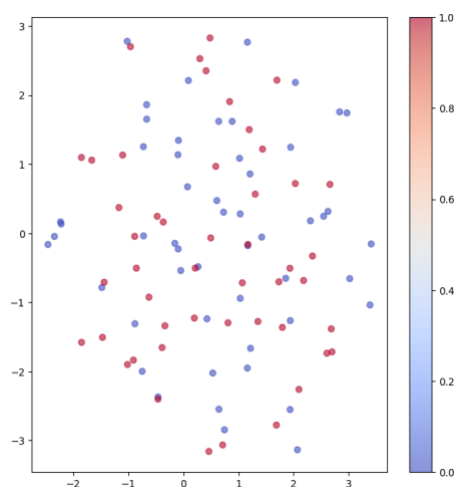


Fig. 5. t-SNE Visualization

4.3. Testing on New Data

The next stage of model evaluation utilized the new testing dataset of 131 images, which resulted in some critical insights into the model's performance, which is

summarised in Table 2. The model's accuracy decreased slightly from 0.96 on the original training dataset to 0.95 on the testing dataset but illustrates the model's ability to perform well nonetheless with relatively little degradation when generalizing to previously unseen data. There were some interesting changes in precision and recall values. The precision for the 'Normal' class increased from 1 to 0.98, suggesting that the error level was in classifying 'Sick' as 'Normal'. The recall for the 'Normal' class was 0.90, indicating that some 'Normal' cases were not detected well. For the 'Sick' case, the precision remained the same, while for the recall, it increased significantly from 0.94 to 0.99, indicating that almost all of the 'Sick' cases were detected very well. The F1 score for the 'Normal' class is 0.94, indicating a good tradeoff between precision and recall, even if it was slightly less than the value in the training dataset. The F1 score for the 'Sick' class was 0.96, illustrating an excellent performance for detecting the 'Sick' condition very well; nonetheless, there is a slight drop in the 'Normal' class for recall, which can still be improved.

TABLE 2. Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.98	0.90	0.94	50
Sick	0.94	0.99	0.96	81
Accuracy			0.95	131
Macro avg	0.96	0.94	0.95	131
Weighted avg	0.96	0.95	0.95	131

Figure 6 depicts the confusion matrix obtained from testing with the new dataset, and the matrix indicates that the model has done a great job of predicting the two classes, 'Normal' and 'Sick'. The model correctly detected the 50 Normal cases; 45 were classified as Normal (True Negatives), but five Normal cases were predicted as sick (False Positives). The model was correct on the 81 Sick cases, of which 80 were predicted as sick (True Positives), and only one ill case was misclassified as a Normal case (False Negative). This shows that the model does an outstanding job detecting the Sick class despite the small error for classifying Normal.



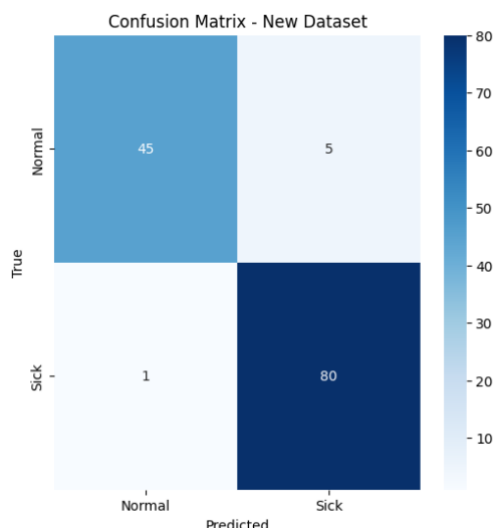


Fig. 6. Confusion Matrix on New Dataset

Figure 7 illustrates the ROC curve obtained from testing the new dataset with an increase in the AUC score of 0.98 from 0.97, which means the model can distinguish between the 'Normal' and 'Sick' classes. An AUC of one means that the model accurately separates the classes. This ROC curve instance also illustrates that the TPR sharply increases despite only a slight increase in the FPR, which indicates the model is very successful at identifying the 'Sick' condition and minimizes errors. Therefore, the model is very effective

at identifying 'Sick' and has a high sensitivity with only a slight prediction error in the 'Normal' class.

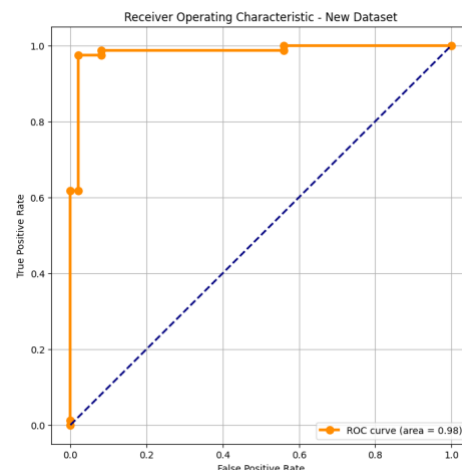


Fig. 7. ROC on New Dataset

4.4. Discussion

The comparison of the current study to the three other studies shows that deep learning models, particularly ensemble learning models, continue to evolve and improve as a systematic way to identify breast cancer. Table 3 above demonstrates how each study has a specific method and approach that differ slightly from each other to increase the sensitivity and accuracy of identifying malignancies in medical imaging, specifically breast cancer.

TABLE 3. Comparison of Breast Cancer Detection Studies Using Deep Learning Models

Aspect	This Study	[1]	[2]	[3]
Research Objective	Breast cancer detection using custom CNN, ResNet18, and VGG16.	Improving invasive breast cancer detection with an ensemble of CNN models (InceptionV3).	Addressing class imbalance in breast cancer data for improved detection accuracy.	Breast cancer diagnosis with an ensemble of advanced CNNs: EfficientNet, AlexNet, ResNet, and DenseNet.
Data Collection Method	Kaggle dataset: 101 "sick" and 163 "normal" images.	Internal and TCGA datasets, 587 training images, 118 test images.	Over-sampling to balance minority data in breast cancer dataset.	Combining four CNN models for breast cancer diagnosis.
Models Used	CNN, ResNet18, VGG16.	10 InceptionV3 ensemble models.	REEL ensemble model using Random Forest and particle swarm optimization.	Ensemble CNN: EfficientNet, AlexNet, ResNet, and DenseNet with optimization and tuning.
Augmentation and Preprocessing	Resize 160x120 px, image flipping, rotation, colour adjustments.	Image augmentation for data variety.	Over-sampling to increase minority data.	Augmentation and transfer learning applied.
Accuracy	0.96 on training data and 0.95 on new data. AUC = 0.97.	0.92 on the internal test set and 0.87 on the TCGA dataset.	High accuracy on breast cancer dataset and others.	0.946 accuracy, 0.924 sensitivity, AUC 0.98.
t-SNE Visualization	Shows a more precise separation between the 'Normal' and 'Sick' classes after using the ensemble model.	X	X	X
Model Strengths	Ensemble (CNN, ResNet18, VGG16) improves accuracy and reduces misclassification.	The ensemble performs better than a single CNN model.	REEL improves accuracy and addresses class imbalance.	The ensemble model provides the best results for early breast cancer detection.



This study exhibited great results for breast cancer detection, achieving 0.96 accuracy on the training data and an AUC of 0.97. There is potential to improve recall for the 'Normal' class; however, the overall results indicate using ensemble learning with a portfolio of custom CNN, ResNet18, and VGG16 is very beneficial for breast cancer detection. Several studies have demonstrated the benefits of an ensemble model, whether that model contains InceptionV3, REEL, or other advanced CNN approaches with high accuracy and good sensitivity.

The performance improvement achieved through the ensemble approach can be explained in terms of the diversity of feature representations. The CNN developed in this study tends to learn local characteristics of images, while ResNet18 utilizes residual learning to preserve the gradient flow in deeper networks. On the other hand, VGG16 produces more consistent visual representations through a simple yet effective convolutional architecture. These three models produce prediction errors that do not always occur on the same samples. Therefore, the prediction averaging process reduces the variance of individual predictions and yields a more robust final decision compared to using a single model.

The t-SNE visualization also shows that the distributions of the two classes became more distinct after applying the ensemble. This indicates that the prediction ensemble strategy not only improves accuracy but also enhances the quality of feature representation, thereby reducing the overlap between classes.

This study has several limitations. The dataset used is relatively small and consists of only two classes, so it does not fully represent the complexity of actual clinical cases. Furthermore, the evaluation was conducted on a single dataset; therefore, the model's generalizability to different imaging modalities and patient populations requires further investigation. Future research will focus on the use of multi-center datasets, external validation, and the exploration of more adaptive ensemble strategies to improve the model's reliability.

5. CONCLUSIONS

This study assessed and designed a successful deep-learning model for breast cancer detection based on medical image data. In this study, an ensemble approach was taken by averaging predictions made by several models using CNN, ResNet18, and VGG16 and subsequently improved classification with greater clarity, as shown in the t-SNE visualization. An ensemble model reduced the misclassification errors produced by an individual model (i.e., have false positives and/or false negatives) and had more stable results overall.

The application of pre-trained models such as ResNet18 and VGG16 has been proven to expedite training processes and increase classification efficiency due to the models being able to improve a certain degree on previously unexamined data. The evaluation results show a high

accuracy of 0.96 from the model, demonstrating considerable detection of the 'sick' classes with high sensitivity. However, there is a minor lack of recall for the 'normal' class. The model had an Area Under the Curve nearly equal to 1, indicating their ability to distinguish between the 'normal' and 'sick' classes was strong. The results of the study indicate that the ensemble approach has potential as a method to support medical image classification for breast cancer detection. However, further validation using multi-institutional datasets, larger sample sizes, and evaluation in real-world clinical settings is still needed before the model can be considered for implementation as a clinical decision support system.

REFERENCES

- [1] L. Solorzano, S. Robertson, B. Acs, J. Hartman, and M. Rantalainen, 'Ensemble-based deep learning improves detection of invasive breast cancer in routine histopathology images', *Heliyon*, vol. 10, no. 12, p. e32892, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32892.
- [2] Z. Wang, J. Xie, and J. Zhang, 'A Robust Enhanced Ensemble Learning Method for Breast Cancer Data Diagnosis on Imbalanced Data', *IEEE Access*, vol. 12, pp. 189776–189788, 2024, doi: 10.1109/ACCESS.2024.3516376.
- [3] D. Shah, M. A. U. Khan, M. Abrar, and M. Tahir, 'Optimizing Breast Cancer Detection With an Ensemble Deep Learning Approach', *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 5564649, 2024, doi: 10.1155/2024/5564649.
- [4] B. Zolfaghari, L. Mirsadeghi, K. Bibak, and K. Kavousi, 'Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning', *ACM Comput. Surv.*, vol. 55, no. 12, p. 262:1–262:34, Mar. 2023, doi: 10.1145/3580218.
- [5] S. Majumdar, P. Pramanik, and R. Sarkar, 'Gamma function based ensemble of CNN models for breast cancer detection in histopathology images', *Expert Systems with Applications*, vol. 213, p. 119022, Mar. 2023, doi: 10.1016/j.eswa.2022.119022.
- [6] E. K. Jadoon, F. G. Khan, S. Shah, A. Khan, and M. ElAffendi, 'Deep Learning-Based Multi-Modal Ensemble Classification Approach for Human Breast Cancer Prognosis', *IEEE Access*, vol. 11, pp. 85760–85769, 2023, doi: 10.1109/ACCESS.2023.3304242.
- [7] S. Sharmin, T. Ahammad, Md. A. Talukder, and P. Ghose, 'A Hybrid Dependable Deep Feature Extraction and Ensemble-Based Machine Learning Approach for Breast Cancer Detection', *IEEE Access*, vol. 11, pp. 87694–87708, 2023, doi: 10.1109/access.2023.3304628.
- [8] F. Azour and A. Boukerche, 'An Efficient Transfer and Ensemble Learning Based Computer Aided Breast Abnormality Diagnosis System', *IEEE Access*, vol. 11, pp. 21199–21209, 2023, doi: 10.1109/ACCESS.2022.3192857.
- [9] J. Ahmad *et al.*, 'Deep learning empowered breast cancer diagnosis: Advancements in detection and classification', *PLOS ONE*, vol. 19, no. 7, p. e0304757, Jul. 2024, doi: 10.1371/journal.pone.0304757.
- [10] S. Zarif, H. Abdulkader, I. Elaraby, A. Alharbi, W. S. Elkilani, and P. Pławiak, 'Using hybrid pre-trained models for breast cancer detection', *PLOS ONE*, vol. 19, no. 1, p. e0296912, Jan. 2024, doi: 10.1371/journal.pone.0296912.
- [11] U. Naseem *et al.*, 'An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers', *IEEE Access*, vol. 10, pp. 78242–78252, 2022, doi: 10.1109/ACCESS.2022.3174599.
- [12] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. Maria Vanegas, 'Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models', *Sensors*, vol. 20, no. 16, p. 4373, Aug. 2020, doi: 10.3390/s20164373.



- [13] A. A. Balasubramanian *et al.*, 'Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology', *Cancers*, vol. 16, no. 12, p. 2222, Jan. 2024, doi: 10.3390/cancers16122222.
- [14] M. I. Mahmud, M. Mamun, and A. Abdelgawad, 'A Deep Analysis of Transfer Learning Based Breast Cancer Detection Using Histopathology Images', 2023, *arXiv*. doi: 10.48550/ARXIV.2304.05022.
- [15] R. M. Munshi, L. Cascone, N. Alturki, O. Saidani, A. Alshardan, and M. Umer, 'A novel approach for breast cancer detection using optimized ensemble learning framework and XAI', *Image and Vision Computing*, vol. 142, p. 104910, Feb. 2024, doi: 10.1016/j.imavis.2024.104910.
- [16] A. Alotaibi *et al.*, 'ViT-DeiT: An Ensemble Model for Breast Cancer Histopathological Images Classification', 2022, *arXiv*. doi: 10.48550/ARXIV.2211.00749.

