

Enhancing early breast cancer diagnosis through automated microcalcification detection using an optimized ensemble deep learning framework

Jing Ru Teoh¹, Khairunnisa Hasikin^{1,2}, Khin Wee Lai¹, Xiang Wu³ and Chong Li⁴

¹ Biomedical Engineering Department, University of Malaya, Wilayah Persekutuan Kuala Lumpur, Malaysia

² Centre of Intelligent Systems for Emerging Technology (CISSET), Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

³ Institute of Medical Information Security, Xuzhou Medical University, Xuzhou, Jiangsu, China

⁴ Graduate School, Xuzhou Medical University, Xuzhou, Jiangsu, China

ABSTRACT

Background: Breast cancer remains a pressing global health concern, necessitating accurate diagnostics for effective interventions. Deep learning models (AlexNet, ResNet-50, VGG16, GoogLeNet) show remarkable microcalcification identification (>90%). However, distinct architectures and methodologies pose challenges. We propose an ensemble model, merging unique perspectives, enhancing precision, and understanding critical factors for breast cancer intervention. Evaluation favors GoogLeNet and ResNet-50, driving their selection for combined functionalities, ensuring improved precision, and dependability in microcalcification detection in clinical settings.

Methods: This study presents a comprehensive mammogram preprocessing framework using an optimized deep learning ensemble approach. The proposed framework begins with artifact removal using Otsu Segmentation and morphological operation. Subsequent steps include image resizing, adaptive median filtering, and deep convolutional neural network (D-CNN) development *via* transfer learning with ResNet-50 model. Hyperparameters are optimized, and ensemble optimization (AlexNet, GoogLeNet, VGG16, ResNet-50) are constructed to identify the localized area of microcalcification. Rigorous evaluation protocol validates the efficacy of individual models, culminating in the ensemble model demonstrating superior predictive accuracy.

Results: Based on our analysis, the proposed ensemble model exhibited exceptional performance in the classification of microcalcifications. This was evidenced by the model's average confidence score, which indicated a high degree of dependability and certainty in differentiating these critical characteristics. The proposed model demonstrated a noteworthy average confidence level of 0.9305 in the classification of microcalcification, outperforming alternative models and providing substantial insights into the dependability of the model. The average confidence of the ensemble model in classifying normal cases was 0.8859, which strengthened the model's consistent and dependable predictions. In addition, the ensemble models attained remarkably high performances in terms of accuracy, precision, recall, F1-score, and area under the curve (AUC).

Submitted 4 February 2024

Accepted 3 May 2024

Published 29 May 2024

Corresponding authors

Khairunnisa Hasikin,

khairunnisa@um.edu.my

Chong Li,

lichong1985@xzhmu.edu.cn

Academic editor

Jyotismita Chaki

Additional Information and
Declarations can be found on
page 17

DOI 10.7717/peerj-cs.2082

© Copyright

2024 Teoh et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Conclusion: The proposed model's thorough dataset integration and focus on average confidence ratings within classes improve clinical diagnosis accuracy and effectiveness for breast cancer. This study introduces a novel methodology that takes advantage of an ensemble model and rigorous evaluation standards to substantially improve the accuracy and dependability of breast cancer diagnostics, specifically in the detection of microcalcifications.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Neural Networks

Keywords Breast cancer, Early diagnosis, Deep learning, Microcalcification, Ensemble

INTRODUCTION

Breast cancer is the world leading cancer for females caused by the uncontrolled proliferation of cells in the breast and the accumulation of additional tissues known as a tumor. It is most frequently diagnosed cancer and the primary contributor to cancer-related deaths in the female population. The high incidence of breast cancer in Asia presents a substantial public health concern. The region experiences a significant prevalence of breast cancer cases, exhibiting varied patterns among its member countries. Asia exhibits the lowest age-standardized mortality rate (ASMR) and age-standardized incidence rate (ASIR) compared to other regions. However, the mortality-to-incidence ratio (M/I) in Asia, with a value of 0.32, surpasses the global average of 0.28, positioning it as the second-highest area worldwide in terms of M/I *Yip, Taib & Mohamed (2006)*. The breast cancer mortality rates in low- and middle-income countries are higher than in their high-income counterparts.

In Malaysia, a middle-income country, there were 21,634 cases of breast cancer discovered between 2012 and 2016, accounting for 34.1% of the cancers diagnosed in the country (*Azizah et al., 2019*). Given the potential severity of breast cancer, early detection and immediate implementation of appropriate treatment become essential in minimizing its impact. This emphasizes the critical significance of identifying the disease at an early stage and preventing its progression. Mammography screening is the most common and practical method of detecting breast cancer as it reduced breast cancer mortality by around 20% in women aged 50 to 70 (*Christiansen, Autier & Støvring, 2022*). The cornerstone of breast cancer diagnosis, mammography screening, plays a critical role in recognizing microcalcifications—tiny calcium deposits inside breast tissue. Although frequently an early indicator of breast abnormalities, these microcalcifications are not conspicuously apparent through symptoms. Mammography, on the other hand, greatly assists in the diagnosis of suspected early-stage breast cancers. With mammography lowering breast cancer mortality by roughly 20% in women aged 50 to 70, the focus on early diagnosis is directly related to identifying these microcalcifications.

By looking for microcalcification in the mammography images, breast cancer in its initial stages can be discovered. Microcalcification, a calcium deposit in the breast, appears as a scattering of white dots on a mammogram. They come in sizes ranging 0.1 to 1 mm (*Cai et al., 2019*). According to *Hakim, Prajitno & Soejoko (2021)*, microcalcification is an

indication of breast cancer occurring in between 12.7% and 41.2% of women who completed mammography screening. [Brahimetaj et al. \(2022\)](#) estimated that between 85% and 95% of cases of ductal carcinoma *in situ* (DCIS) were detected due to the existence of microcalcifications in roughly 55% of non-palpable breast cancers. Additionally, there is a strong correlation between a cluster of microcalcification as well as an increased probability of breast cancer ([Azam et al., 2021](#)). Invasive disease can be prevented by recognizing microcalcifications given that they are associated with premalignant and developing breast disease ([Logullo et al., 2022](#)). Therefore, the early intervention to detect breast cancer is identifying signs of microcalcification.

Yet it can be difficult and challenging to detect microcalcification because their distribution in the breast is unpredictable and their shapes are uncertain. Radiologists often missed or misdiagnosed microcalcification, therefore many healthcare professionals have had trouble spotting it. The microcalcifications can be observed as a cluster of several white dots and they can be identified in common benign lesions including fibrocystic alterations, inflammatory lesions, and breast abnormalities ([Brahimetaj et al., 2022](#)).

When considering a middle-income country such as Malaysia, where financial resources are limited, the incorporation of deep learning models into the healthcare infrastructure can offer a substantial enhancement to mammography screening. Through the effective detection of microcalcifications that might otherwise evade detection, these models make a valuable contribution to the timely detection of potential breast abnormalities. This is in line with the overarching goal of minimizing the detrimental effects of breast cancer through early detection. The potential for significant improvement in the diagnosis of breast cancer, especially during the critical early phases, could be realized through the synergistic collaboration of deep learning and mammography screening. This would enable the implementation of targeted and timely intervention strategies.

ResNet, a novel deep learning (DL) framework unveiled in 2015, addresses complexities inherent in deep networks through the utilization of skip connections within residual blocks. ResNet-50, an instance of the ResNet architecture, has garnered considerable attention in domains such as mammogram analysis. Studies by [Shiri Kahnouei et al. \(2022\)](#) and [Leong et al. \(2022\)](#) demonstrated the effectiveness of ResNet-50 in mammogram-based calcification detection and segmentation, achieving impressive accuracies of 96.7% and 97.58%, respectively. This showcases ResNet-50's potential for advancing medical diagnosis and intervention through precise image analysis. [Montaha et al. \(2021\)](#) have built upon VGG16, developing models like BreastNet18 for early-stage breast cancer identification. GoogLeNet, a 22-layer architecture created by Google researchers, achieved high accuracy rates of 93.3% on the ImageNet dataset. [Jhang \(2018\)](#) explored the use of GoogLeNet with class activation mapping for detecting suspicious microcalcification regions in mammogram images. [Sharma & Mukherjee \(2020\)](#) also compared the classification performance of GoogLeNet and AlexNet in microcalcification detection.

The application of these deep learning models to the detection and segmentation of microcalcifications has been the primary focus of recent research. However, there remains a need for comprehensive comparative analyses to better understand the strengths and

limitations of these models, as well as their practical implementation in clinical settings. In this context, ensemble learning emerges as a promising approach to enhance the performance of deep learning model in breast cancer detection. Ensemble learning is a method that combines multiple base models to produce a more powerful model. There are several advantages in utilizing ensemble learning. First of all, ensemble models combine outcomes from multiple base models and produce higher performance outcomes. This is achieved by the ensemble methods in leveraging the strength of different algorithms resulting in a better overall outcome performance compared to individual models. In addition, ensemble models help in reducing the risk of overfitting by capturing irrelevant patterns from training data. Ensemble methods can reduce the model biases and variance and improved the generalization performance of the ensemble models (Mohammed & Kora, 2023). In our study, we aim to improve the detection of microcalcifications in mammogram images by developing an optimized ensemble model. Specifically, we utilize an averaging technique to combine two baseline deep learning models and form an optimized ensemble model, thereby enhancing the reliability and effectiveness of breast cancer detection.

MATERIALS AND METHODS

Our study proposes the implementation of an ensemble model that integrates the advantageous features of individual pre-trained convolutional neural network models to efficiently tackle the issue of detecting microcalcification. In order to address the existing research gap, this study undertakes an exhaustive comparative analysis to ascertain the merits, drawbacks, and particular suitability of each model in detecting these nuanced yet critical indicators of breast abnormalities in their early stages. Furthermore, the objective of the project is to investigate the viability and practical execution of incorporating these models into clinical environments to improve the precision of breast cancer detection and their practical implications in the healthcare sector.

Data description

The datasets utilized in this study were acquired from mini-MIAS mammography, INBreast, and CBIS-DDSM. The Mammographic Image Analysis Society supplied Mini-MIAS mammography, whereas the Breast Research Group, INSEC Porto from Portugal supplied INBreast, as detailed in Table 1. For the classification procedure, normal images were excluded from the Mini-MIAS and INBreast datasets *via* a filtration process. This study utilized 1,511 abnormal cases of breast cancers exhibiting calcification that were obtained from the aforementioned database. Every single image extracted from the mentioned database was a cranial caudal (CC) or mediolateral oblique (MLO) mammogram. The data was subsequently separated into their respective classes and labelled. Subsequently, the accumulated datasets were arbitrarily partitioned 70% for training purposes and 30% for testing. Specifically, the training dataset comprised 700 images of the normal class and 1,096 images of microcalcification. In the testing dataset, there were 450 microcalcification images and 300 images of normal class. The training

Table 1 Dataset used in this study.

Database	Features
Mini-MIAS	Obtained from: https://www.kaggle.com/datasets/kmader/mias-mammography . Consisted of total number of 161 cases (322 images) of normal and abnormal images. The abnormal images consisted of benign and malignant with different features such as calcification, masses, architectural distortion <i>etc.</i>
INBreast	Obtained from: https://www.kaggle.com/datasets/ramanathansp20/inbreast-dataset . The database comprised a total of 115 cases, consisting of 412 images. The dataset encompassed various types of mammograms, including normal cases, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, and images with multiple findings.
CBIS-DDSM	Obtained from: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629 . The database consisted of 1,566 cases (6,671 of scanned film mammograms). It only contained abnormal images of masses and calcification.

dataset was used for real training, whereas the testing dataset was used for evaluating the performance of the model.

Data preprocessing

Preparing the input dataset for the DL algorithms was a critical step in the pipeline. It involved a series of preprocessing steps to ensure the mammogram images are in optimal condition for further analysis. These steps encompassed various techniques and transformations aimed at enhancing the quality and suitability of the data. The process of preprocessing mammogram images played a vital role in facilitating accurate and effective analysis using DL algorithms.

(a) Artifact removal

In [Fig. 1](#), the dataset of normal images from mini-MIAS and INBreast underwent two crucial processes: the Otsu segmentation method and MorphologicalEx method. The Otsu segmentation method played a vital role in automatically identifying the optimal threshold value for segmenting the mammogram images. This segmentation process helped in separating the desired regions from the background noise. Subsequently, the MorphologicalEx method, also known as Morphological Erosion, was employed to effectively eliminate the labeled regions from the images. This meticulous step ensured that the presence of free labeling does not interfere with the subsequent image cropping process. By implementing these techniques, the dataset was prepared for further preprocessing steps.

(b) Image cropping and resizing

After the removal of artifacts, the images underwent cropping and resizing. This crucial step ensured that the images were appropriately sized at 224×224 , aligning them with the requirements of the DL algorithms for efficient classification. By standardizing the image size, the DL algorithms could effectively perform their classification operations with optimal accuracy and efficiency.

(c) Image filtering

Then, an adaptive median filter was employed to enhance the quality of the mini-MIAS and INBreast cropped images, as well as the ROI images containing microcalcification from CBIS-DDSM. This filter effectively mitigated noise, resulting in improved

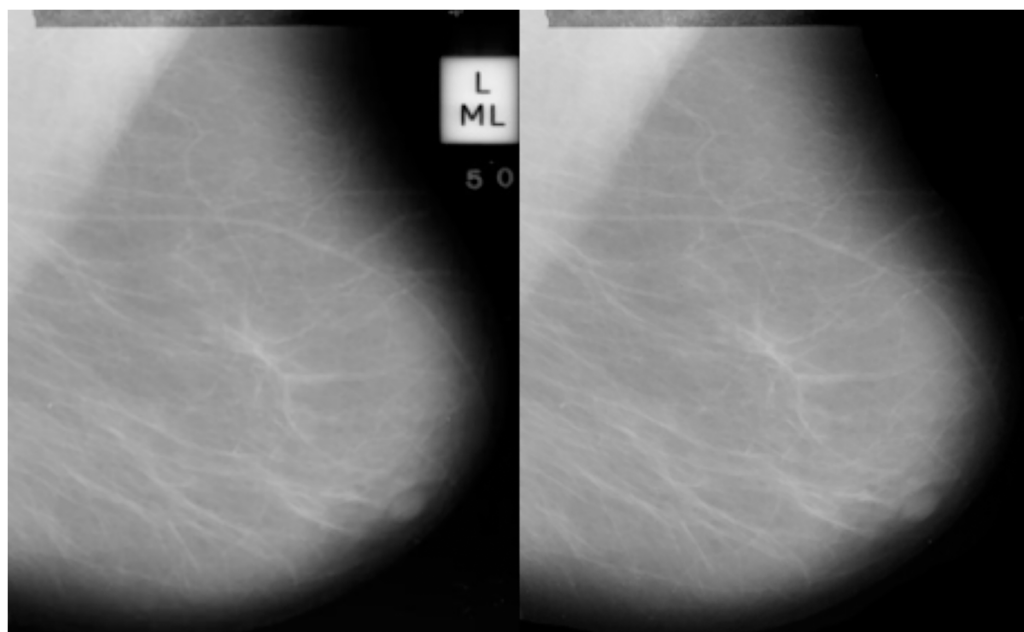



Figure 1 Before and after applying Otsu segmentation method and MorphologicalEx method where the image labelled were removed. Image source credit: (C) The mini-MIAS database of mammograms, <http://peipa.essex.ac.uk/info/mias.html>. Full-size  DOI: 10.7717/peerj-cs.2082/fig-1

mammography image quality. The application of the adaptive median filter demonstrated its efficacy in reducing noise and enhancing image clarity.

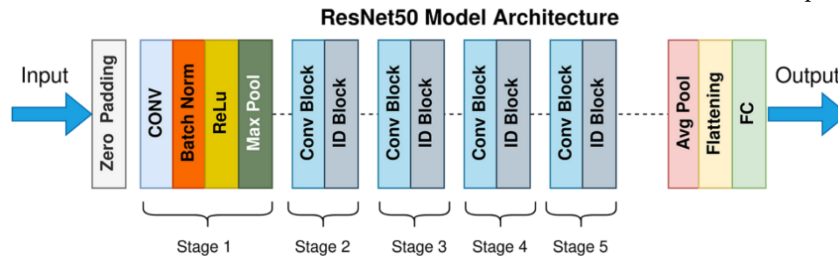
Classification model development using deep learning approach

A deep convolutional neural network (D-CNN) model is developed by finely tuning its hyperparameters to accurately detect microcalcifications in mammogram images. Instead of training the CNN model from scratch, transfer learning is applied, utilizing pre-trained CNN models such as ResNet-50, GoogLeNet, VGG16 and AlexNet architecture from the torchvision library. To achieve optimal results, critical hyperparameters such as the number of epochs, learning rate, batch size, and step size are carefully adjusted. The input images underwent normalization to the standard normalization parameters specified by ImageNet ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]). This normalization was performed because the transfer learning model employed was pre-trained using the ImageNet database. Normalization is necessary to restore the original color of the input cell regions that have been stained.

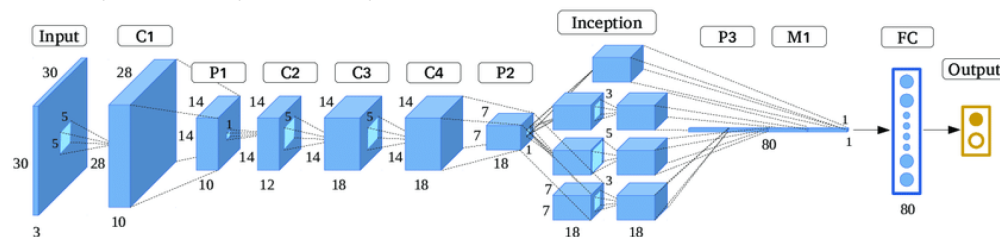
A comprehensive summary of each deep learning architecture can be found in [Table 2](#). Upon completion of the training process, the model's performance is evaluated using performance metrics (accuracy, precision, recall, F1-score, area under curve from true positive rate vs false positive rate graph) and confusion matrix as summarized in [Table 3](#). The confusion matrix considers true positives (accurate microcalcification detection), true negatives (correct classification of negative cases), false positives (incorrectly labeled as

Table 2 Overview of deep learning models in this study.

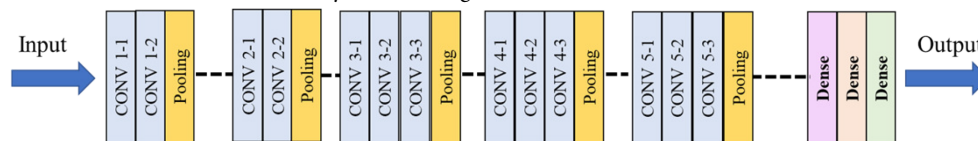
Model	Descriptions
ResNet-50	ResNet-50 is a variant of the ResNet model, which has 48 convolution layers along with 1 MaxPool and 1 Average Pool layer. ResNet-50 is an artificial neural network (ANN) that forms networks by stacking residual blocks. Identity connections take the input directly to the end of each residual block and learn the features to be classified as the desired output.



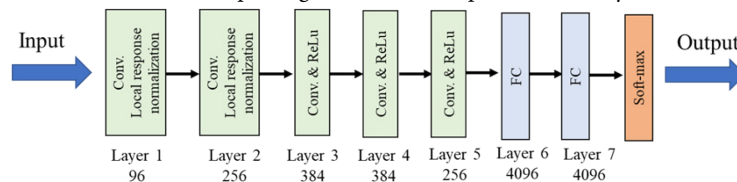
GoogLeNet The GoogLeNet network is an impressive deep convolutional neural network comprising 22 layers. The GoogLeNet architecture is designed to handle input images with a resolution of 224×224 pixels. It was specifically engineered to be a computational powerhouse, surpassing or competing with existing networks at the time of its creation.



VGG16 VGG-16 is a convolutional neural network architecture with convolutional layers of 3×3 filters with stride 1 and constant padding and maxpool layers of 2×2 filters with stride 2 used. This arrangement of convolutional and max pool layers is followed uniformly throughout the entire architecture. The end has 2 F.C. (fully connected layers) followed by a softmax for output. The number 16 in VGG16 indicates that it has 16 layers with weights.



AlexNet AlexNet is made up of five convolutional layers, three max-pooling layers, two normalization layers, two fully connected layers, and one softmax layer. Convolutional filters and a nonlinear activation function ReLU are used in each convolutional layer. The pooling layers are used for maximum pooling. Because of the presence of fully connected layers, the input size is fixed.



positive), and false negatives (missed microcalcification detection). Analyzing these metrics allows for a robust evaluation of the model's overall accuracy.

The classification model optimization through ensemble model

The classification model was refined through the development of an ensemble architecture. Through a thorough evaluation, the two top-performing models distinguished by their high accuracy and minimal computational requirements, were selected. These models are

Table 3 Model evaluation metrics.

Metrics	Formula	Description
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	A fraction of correctly classified cases over the total samples.
Precision	$\frac{TP}{TP + FP}$	A proportion of positive classes were correctly classified.
Recall	$\frac{TP}{TP + FN}$	A ratio of all positive samples correctly predicted as positive.
F1-score	$2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$	A combination of precision and recall as their harmonic mean.

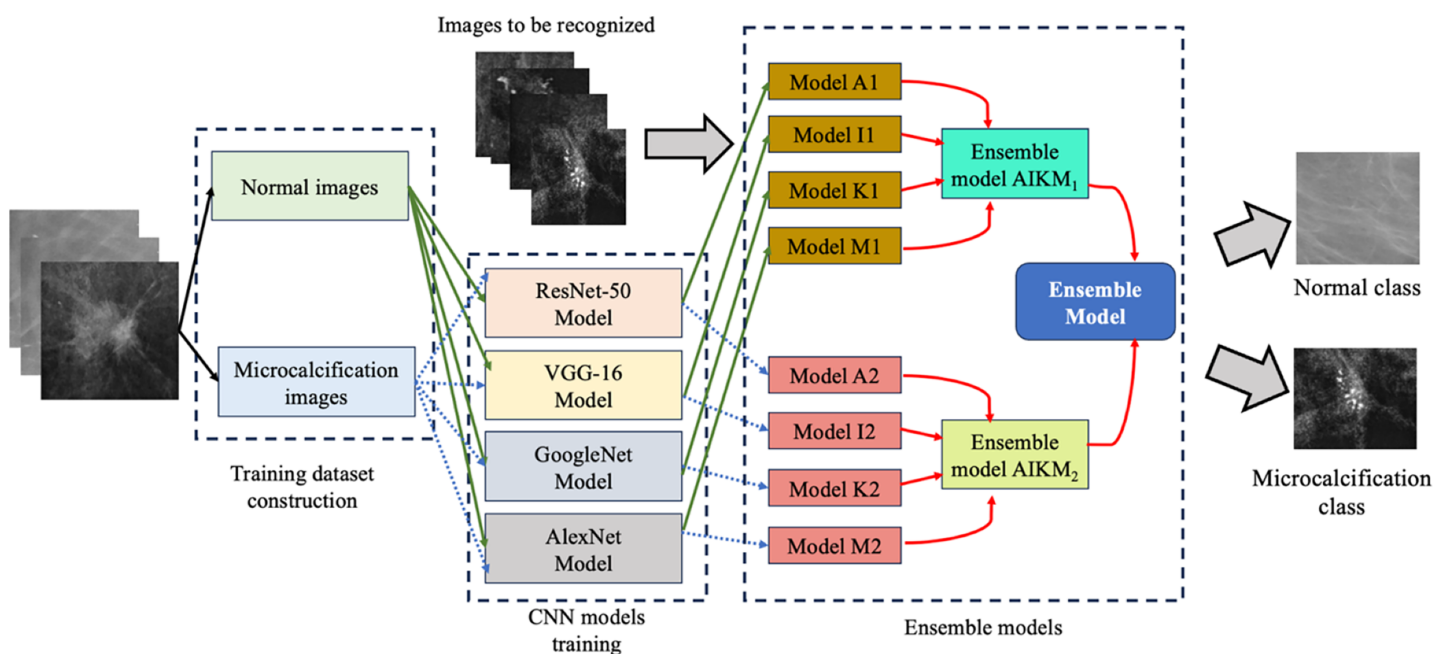


Figure 2 The development of final ensemble models for microcalcification identification. Image source credit: (C) The mini-MIAS database of mammograms, <http://peipa.essex.ac.uk/info/mias.html>. Full-size DOI: 10.7717/peerj-cs.2082/fig-2

then integrated into the ensemble architecture, employing a weighted averaging approach to combine their outputs. This process ensures a more resilient and accurate final classification outcome as illustrated in Fig. 2.

The developed DL models of AlexNet, GoogLeNet, VGG16, and ResNet-50 were constructed and tested for their capability to detect microcalcifications in mammogram images. The process of individual testing was initiated by training and evaluating each model on their accuracy and proficiency in discerning these subtle yet vital characteristics. Specifically, the following models were subjected to comprehensive evaluation underwent rigorous testing. The AlexNet model was chosen due to its innovative contributions to image classification, while VGG16, was compared due to its ability to be distinguished by its depth and resilient performance. GoogLeNet, which is renowned for its inception


```

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Load and preprocess the dataset
# Ensure labels for the dataset

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(data, labels, test_size=0.2, random_state=42)

# Initialize and train individual models
alexnet_model = AlexNet()
# Train the AlexNet model using the training data
alexnet_model.fit(X_train, y_train)

googlenet_model = GoogLeNet()
# Train the GoogLeNet model using the training data
googlenet_model.fit(X_train, y_train)

vgg16_model = VGG16()
# Train the VGG16 model using the training data
vgg16_model.fit(X_train, y_train)

resnet_model = ResNet50()
# Train the ResNet-50 model using the training data
resnet_model.fit(X_train, y_train)

# Evaluate individual models on the test set
alexnet_predictions = alexnet_model.predict(X_test)
alexnet_accuracy = accuracy_score(y_test, alexnet_predictions)

googlenet_predictions = googlenet_model.predict(X_test)
googlenet_accuracy = accuracy_score(y_test, googlenet_predictions)

vgg16_predictions = vgg16_model.predict(X_test)
vgg16_accuracy = accuracy_score(y_test, vgg16_predictions)

resnet_predictions = resnet_model.predict(X_test)
resnet_accuracy = accuracy_score(y_test, resnet_predictions)

# Select two best-performing models
best_model_1 = select_best_model(alexnet_accuracy, googlenet_accuracy, vgg16_accuracy, resnet_accuracy)
best_model_2 = select_second_best_model(alexnet_accuracy, googlenet_accuracy, vgg16_accuracy, resnet_accuracy)

# Combine the predictions of the two best models (ResNet-50 and GoogLeNet) for the ensemble
ensemble_predictions = combine_predictions(resnet_predictions, googlenet_predictions)
ensemble_accuracy = accuracy_score(y_test, ensemble_predictions)

```

Figure 3 Pseudocode of ensemble model development. Full-size  DOI: 10.7717/peerj-cs.2082/fig-3

modules; and ResNet-50, which utilizes skip connections to ensure robust training. The Pseudocode of the optimization through development of ensemble model as shown in Fig. 3.

RESULTS

As shown in Table 4, the performance of each DL algorithm can be evaluated using the following performance metrics: accuracy, precision, recall, F1-score, AUC, and training time. These metrics offer valuable insights regarding the quality and effectiveness of the models utilized for microcalcification detection in mammogram images. The best results of each performance metric were made bold.

In overall, ResNet-50 was preferred in this study as it has high detection accuracy in detecting microcalcifications in mammogram images. The remarkable performance of

Table 4 Performance metrics and training time for each DL algorithm. The best results are shown in bold.

	ResNet-50	GoogLeNet	VGG16	AlexNet
Accuracy	0.9827	0.9560	0.9587	0.9507
Precision	0.9828	0.9586	0.9586	0.9524
Recall	0.9827	0.9560	0.9587	0.9507
F1-score	0.9826	0.9555	0.9586	0.9502
AUC	0.9800	0.9456	0.9556	0.9411
Training time	76 m 31 s	58 m 46 s	158 m 48 s	33 m 16 s

ResNet-50 in detecting microcalcifications in mammogram images can be attributed to its distinctive architecture, specifically the integration of skip connections. By establishing connections in this way, the model can effectively learn to preserve important information from the input layers through to the deeper layers. This prevents the loss of crucial details during training, which can happen in traditional deep neural networks due to the vanishing gradient problem. ResNet-50 tackles this issue by ensuring that gradients can flow smoothly through the network, allowing it to identify subtle patterns, such as those indicating microcalcifications, more accurately.

The particular benefit of ResNet-50 in mitigating the vanishing gradient issue has significant importance in the context of microcalcification detection. In mammography pictures, the identification of these indications typically presents as subtle, necessitating the model's ability to distinguish detailed patterns across several layers. The skip connections in ResNet-50 enable the preservation and transmission of important information, allowing for the effective collection and learning of small but significant traits associated with microcalcifications, even in the deeper levels of the network. The preservation of complex features is crucial for achieving precise detection, resulting in improved precision for recognizing small but diagnostically important abnormalities seen in mammography images.

The architectural superiority of ResNet-50 lies in its ability to preserve the accuracy of characteristics that are crucial for detecting microcalcifications across its network layers. This attribute plays a pivotal role in the exceptional performance of ResNet-50 in this task. Additional investigation aimed at maximizing the efficiency of the training period while maintaining the remarkable performance of this method has the potential to significantly improve its practical applicability in clinical environments. This would facilitate the detection process by reducing the time required, while ensuring that accuracy is not compromised.

The presentation of the confusion matrix for the DL models in Fig. 4 produced significant insights regarding their performance. A comparative analysis was undertaken to assess and contrast the performance of every deep learning algorithm with the objective of determining which model is superior to the others. The confusion matrix was of the utmost importance in providing a thorough comprehension of the performance of the

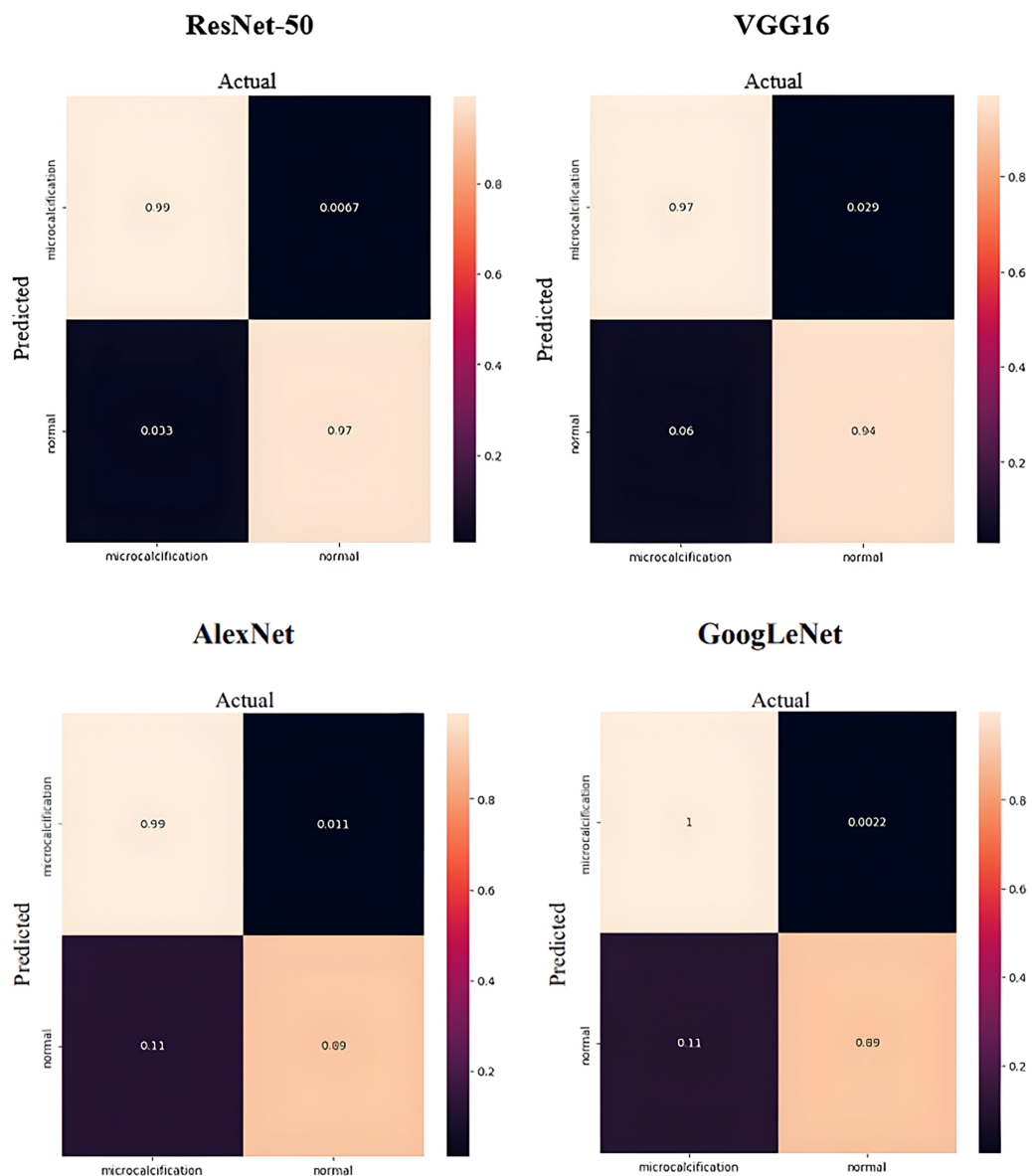


Figure 4 Confusion matrix for each DL algorithm. [Full-size !\[\]\(ba1b80118482ccef74a5d718ca4d7242_img.jpg\) DOI: 10.7717/peerj-cs.2082/fig-4](https://doi.org/10.7717/peerj-cs.2082/fig-4)

models. The information it furnished was crucial and can be utilized to assess the precision and efficacy of the models in detecting microcalcifications.

Figure 4 illustrates that GoogLeNet exhibited the highest percentage value of TP, suggesting that it is the most effective model for accurately identifying microcalcifications in mammogram images. AlexNet and ResNet-50 both had comparable percentage values of TP; however, ResNet-50 exhibited a smaller FP value in comparison to AlexNet. This suggested that AlexNet exhibited a propensity to incorrectly classify non-microcalcifications as microcalcifications on a more frequent basis. In addition, ResNet-50 acquired a greater value of TN and a lesser value of FN in comparison to VGG16. The findings of this study demonstrated that ResNet-50 exhibited a high degree of accuracy in

identifying mammogram images devoid of microcalcifications, with no instances of misclassification involving the microcalcifications themselves. Overall, ResNet-50 performed the best in the confusion matrix, as its proportion of TN was the highest and its proportion of FN was the lowest.

DISCUSSION

The proposed ensemble model performance comparison

The exceptional performance shown by several deep learning models, including ResNet-50, VGG16, GoogLeNet, and AlexNet, is evident in their ability to accurately identify microcalcifications in mammography images, with accuracies over 90%. Nevertheless, these models exhibit unique architectural structures and training procedures, effectively reflecting many aspects of the characteristics that are indicative of microcalcifications. Therefore, an ensemble model is a smart combination of the different and complex points of perspective of the individual models. When considering the detection of microcalcifications, this methodology presents the potential for enhanced precision, heightened resilience and a more holistic comprehension of the complex characteristics that are vital for the timely identification and intervention of breast cancer.

Nevertheless, following an extensive battery of tests, the ensemble model's final selection was reached upon GoogLeNet and ResNet-50. Not only did these two models demonstrate exceptional accuracy, but they also possessed unique nuances and strengths in their respective methodologies. Their selection was significantly influenced by the implementation of skip connections in ResNet-50 and the inception modules of GoogLeNet. In addition to displaying exceptional performance on an individual basis, these models also implemented a variety of techniques to capture the intricate patterns that are essential for the detection of microcalcification. Although AlexNet achieved higher accuracy than GoogLeNet, it needed longer training period. Thus, GoogLeNet is chosen as it has a shorter training period and lower computational cost.

The deliberate incorporation of GoogLeNet and ResNet-50 into the ensemble model signifies the synthesis of varied capabilities and methodologies to produce a solution that is more all-encompassing and resilient. By capitalizing on their distinct capabilities and perspectives, they combined their predictive prowess in a manner that ensured greater precision and dependability, thereby ultimately improving the clinical detection and diagnosis of microcalcifications in mammogram images.

The performance comparison of the developed predictive models prior to and subsequent to the construction of the ensemble model is presented in [Table 5](#). The ensemble model under consideration was constructed by selecting the model with the highest performance among two previously developed models, GoogLeNet and ResNet-50. To ensure the best performance was achieved and generalizability of the developed model, the hyperparameters were tuned and tested. We compare the performance of the proposed ensemble model with the individual developed model in terms of average confidence score in normal and microcalcification classifications. The term "average confidence" pertains to the degree of surety or reliance that a model has on its predictions regarding the detection of microcalcifications in mammogram images. It signifies the degree of confidence the

Table 5 Performance comparison of the ensemble model and the developed pre-trained models.

Tested models	Average confidence	
	Normal	Microcalcification
Ensemble model 1: ResNet-50 + GoogleNet Batch size = 20 Learning rate = 0.00001 Epoch = 100 Step size = 100	0.8322	0.9176
Ensemble model 2: ResNet-50 + GoogleNet Batch size = 7 Learning rate = 0.00001 Epoch = 100 Step size = 100	0.8852	0.9179
Ensemble model 3: ResNet-50 + GoogleNet Batch size = 7 Learning rate = 0.00001 Epoch = 150 Step size = 100	0.8859	0.9305
Resnet-50 Batch size = 7 Learning rate = 0.000001 Epoch = 200 Step size = 200	0.8433	0.8097
GoogleNet Batch size = 21 Learning rate = 0.00001 Epoch = 300 Step size = 200	0.9058	0.7590
VGG-16 Batch size = 6 Learning rate = 0.000001 Epoch = 200 Step size = 200	0.8621	0.8296
AlexNet Batch size = 12 Learning rate = 0.000001 Epoch = 200 Step size = 200	0.9358	0.9062

model possesses in its prognostications, specifying whether microcalcifications are present or absent in a given image. A high confidence score on average across predictions indicates that the model identifies these features with consistent certainty and confidence.

Regarding the classification of normal cases, AlexNet attained the highest average confidence score of 0.9358. In contrast, the ensemble model comprising ResNet-50 and GoogleNet (Ensemble model 3) achieved the highest average confidence value of 0.9305, indicating its superior performance in the classification of microcalcification. Given that the primary objective of this study is to detect microcalcifications, we have opted for the ensemble model comprising ResNet-50 and GoogleNet over the alternative models. This model achieved an average confidence level of 0.8859, which is notably high even for the

Table 6 Table of comparisons with other published works.

Author// year	Image dataset	Algorithms	Performance metrics	Main findings
<i>Leong et al. (2022)</i>	CBIS-DDSM	VGG16, ResNet34, AlexNet and ResNet50	Accuracy	The study reviewed the performance of four difference DL techniques for detecting benign and malignant microcalcifications. ResNet50 had the highest accuracy of 97.58%, subsequently followed by ResNet34 with 97.35%, VGG16 with 96.97%, and AlexNet with 83.06%.
<i>Kumar Singh et al. (2022)</i>	CBIS-DDSM	InceptionResNetV2 model	Sensitivity, specificity, accuracy, and area under the curve (AUC)	The author utilized a pretrained model of InceptionResNetV2 model and compared with different optimizers to compare their performance on microcalcification categorization. The study revealed remarkable results with the ADADelta optimizer, showcasing an impressive training rate of 98% and a validation accuracy of 94% while AUC and sensitivity are 96% and 97% respectively.
<i>Hossain (2022)</i>	CBIS-DDSM	U-Net	F-measure, Dice score, Jaccard index, accuracy, sensitivity, and precision	U-Net is used to segment the microcalcification region from mammogram images. The performance of U-Net is indicated by the value achieved in its sensitivity, precision, accuracy, F-measure, Dice score, and Jaccard index. The average value of each category that U-Net achieved are 98.4%, 94.7%, 98.2%, 98.5%, 97.8%, and 97.4% respectively.
<i>Rehman et al. (2021)</i>	CBIS-DDSM	computer-vision-based FC-DSCNN along with the DCNN	Sensitivity, specificity, recall, F1-score, and AUC curve	The FC-DSCNN model was evaluated using two distinct datasets, and their performances were meticulously compared. DDSM dataset yielded specificity, accuracy, F1-score, precision, and recall values of 0.82, 0.90, 0.85, 0.89, and 0.82, respectively while the PINUM dataset exhibited specificity, accuracy, F1-score, precision, and recall values of 0.79, 0.84, 0.80, 0.86, and 0.79, respectively.
<i>Sharma & Mukherjee (2020)</i>	CBIS-DDSM	AlexNet, GoogleNet	Sensitivity, specificity, accuracy, and area under the curve (AUC)	Another author compared two DL transfer learning techniques to develop an automated segmentation and classification model for microcalcification of mammogram on CC and MLO views.
<i>Hakim, Prajitno & Soejoko (2021)</i>	INBreast	U-Net	Accuracy, Sensitivity, precision, dice score, Mean Squared Logarithmic Error (MSLE)	The U-Net model that was used in this study gave the best MSLE loss value of 0.05 followed by sensitivity of 88.14%, precision of 91.6%, dice score of 90% and accuracy of 90.3%.
<i>Jakhar, Gupta & Mrityunjay (2022)</i>	BreakHis, WBCD	Extra tree classifier, random forest, adaboost, gradient boosting and 9-nearest neighbor (KNN9)	Accuracy, precision, sensitivity, specificity, F1-score, ROC, MCC	The author proposed a stacked-based ensemble learning framework called SELF model. Five base learners are chosen to form ensemble model which are extra tree classifier, random forest, adaboost, gradient boosting and KNN9. The proposed model achieved high accuracy, precision, sensitivity, F1-score, ROC and MCC of 94.35%, 92.45%, 95.96%, 82.87%, 94.17%, 89.41%, and 80.81% respectively in BreakHis dataset. For the other dataset of WBCD, the proposed model achieved high accuracy of 98.8%, AUC of 99.06%, MCC of 97.45%, sensitivity, precision, and F1-score of 99.09%.

Table 6 (continued)

Author// year	Image dataset	Algorithms	Performance metrics	Main findings
Our proposed model	CBIS- DDSM INBreast Mini- MIAS	Ensemble model ResNet-50 + GoogleNet	Accuracy, precision, recall, F1-score, AUC, Average confidence	In our study, we proposed ensemble model by using ResNet-50 with GoogleNet. The model achieved high accuracy, recall, and F1-score of 99.07%, and precision of 99.08% and AUC of 99.06%. The average confidence of our proposed model is 88.59% for normal class and 93.05% for microcalcification class.

classification of normal cases. The microcalcification class, which possesses the highest average confidence metric, offers significant insights into the reliability and surety of the model's predictions.

When the performance of the ensemble models with different hyperparameters is examined, some parameters are critical to obtaining better outcomes in identifying cases of normal and microcalcification. Particularly, the number of training epochs and batch size have obvious impacts on the performance of the model. Models with smaller batch sizes, such as Ensemble Model 2 and Ensemble Model 3, tend to yield slightly higher average confidence scores for both normal and microcalcification cases compared to those with larger batch sizes, as seen in Ensemble Model 1. This suggests that reducing the batch size can potentially enhance the model's discriminative capabilities. In addition, Ensemble Model 3, which has been trained with 150 epochs, has shown a slight improvement in performance with respect to those models trained for 100 epochs. This demonstrates the critical need for intensive training during which the model will be capable of better understanding the underlying patterns within the data.

Comparative performance analysis with the other published works

Additionally, we conducted a comparative analysis of the performance of the ensemble model we proposed with that of other recently published works that utilized a similar dataset (Table 6). We select seven most significant works that were published from 2020 to 2022. Among the seven published works, only one study proposed an ensemble learning model using five base-learners in breast cancer classification (Jakhar, Gupta & Singh, 2023). Their method involved stacking Extra Tree, Random Forest, Adaboost, Gradient Boosting, and KNN9 models to create an ensemble model. In contrast, our approach distinguishes itself by focusing on mammogram images for breast cancer classification. Although both studies employ ensemble learning techniques, the datasets utilized, and the imaging modalities differ significantly. Our study's emphasis on mammogram images underscores the unique challenges and considerations inherent in this imaging modality for breast cancer detection.

In addition, our research made use of a comprehensive dataset, which consisted of three different datasets: CBIS-DDSM, INBreast, and mini-MIAS. By means of this comprehensive combination, our predictive model was able to acquire information regarding a wide range of patterns and variations, thereby augmenting its capability to

handle practical clinical situations. In addition to guaranteeing a thorough training process, the larger and more varied dataset enhanced our model's ability to capture variations that are commonly encountered in clinical environments.

Furthermore, in contrast to the results compiled in [Table 6](#) of related studies, our research prioritized the crucial parameter of average confidence scores. In these earlier works, the explicit classification of normal and microcalcification classes was not addressed. This focal point examines the model's fundamental dependability in differentiating microcalcifications from healthy tissues. Comprehending this difference is critical in the field of medical diagnosis and indispensable in facilitating well-informed clinical judgement.

By employing a distinctive methodology that involves a careful examination of the mean confidence scores within these classes, we enhance the comprehensiveness of our research. This provides a more subtle understanding of the model's dependability and effectiveness in crucial classification, thereby ultimately bolstering the model's accuracy and practicality in the field of medicine.

CONCLUSIONS

Our research represents a significant progression in the accurate detection and classification of microcalcifications in mammogram images *via* the construction and implementation of the proposed ensemble model comprising optimized GoogLeNet and ResNet-50. The model's precision and dependability were enhanced due to their outstanding individual performances and distinctive architectural details, which led to this decision. The integration of these models into an ensemble not only combined a wide range of capabilities but also facilitated increased resilience and a more comprehensive comprehension of the complex attributes that are crucial for prompt intervention in breast cancer. Moreover, our research is distinguished by its focus on average confidence scores, specifically in the classification of normal and microcalcification cases. The proposed model attained the average confidence scores of 0.9305 and 0.8859 in classifying microcalcification and normal cases respectively. This particular aspect enhances the comprehension of the model's dependability, thereby making a substantial contribution to its precision and applicability in the field of clinical diagnosis. Our research is distinguished by the thorough integration of datasets and the in-depth analysis of average confidence scores within classes. This establishes a standard for future models in the critical field of breast cancer detection, which can be more refined and dependable. In comparison to microcalcification images from CBIS-DDSM, the datasets for normal images from mini-MIAS and INBreast are comparatively limited, which is the primary limitation of this study. The resultant distribution of the datasets for each class is imbalanced. By incorporating a broader range of datasets, one can obtain a more substantial and representative sample, which may enhance the performance and generalizability of deep learning algorithms, ultimately leading to more dependable and superior model performance. Further research could investigate automated approaches to cropping mammogram images in order to facilitate the development of an end-to-end classification system.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Jiangsu Province Social Science Application Research Boutique Engineering Project (grant number 23SYB-010). The Xuzhou Science and Technology Project under Grant No. KC23310 provided the funding for the APC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Jiangsu Province Social Science Application Research Boutique Engineering: 23SYB-010.
Xuzhou Science and Technology: KC23310.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jing Ru Teoh conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Khairunnisa Hasikin conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Khin Wee Lai conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Xiang Wu conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Chong Li conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The MIAS Mammography dataset is available at Kaggle: <https://www.kaggle.com/datasets/kmader/mias-mammography>.

The INbreast Dataset is available at Kaggle: <https://www.kaggle.com/datasets/ramanathansp20/inbreast-dataset>.

The datasets generated and/or analyzed during the current study are available in the Cancer Imaging Archive repository: <https://www.cancerimagingarchive.net/collection/cbis-ddsm>.

The code is available at Zenodo: Teoh, J. R., Hasikin, K., Lai, K. W., Wu, X., & Li, C. (2024). Enhancing Early Breast Cancer Diagnosis Through Automated Microcalcification Detection Using an Optimized Ensemble Deep Learning Framework. Zenodo. <https://doi.org/10.5281/zenodo.10566835>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj-cs.2082#supplemental-information>.

REFERENCES

- Azam S, Eriksson M, Sjölander A, Gabrielson M, Hellgren R, Czene K, Hall P. 2021. Mammographic microcalcifications and risk of breast cancer. *British Journal of Cancer* 125(5):759–765 DOI 10.1038/s41416-021-01459-x.
- Azizah AM, Hashimah B, Nirmal K, Siti Zubaidah AR, Puteri NA, Nabihah A, Sukumaran R, Balqis B, Nadia SMR, Sharifah SSS, Rahayu O, Nur Alham O, Azlina AA. 2019. Malaysia national cancer registry report (MNCR) 2012–2016. Available at [https://www.moh.gov.my/moh/resources/Penerbitan/Laporan/Umum/2012-2016%20\(MNCRR\)/MNCR_2012-2016_FINAL_\(PUBLISHED_2019\).pdf](https://www.moh.gov.my/moh/resources/Penerbitan/Laporan/Umum/2012-2016%20(MNCRR)/MNCR_2012-2016_FINAL_(PUBLISHED_2019).pdf).
- Brahimetaj R, Willekens I, Massart A, Forsyth R, Cornelis J, Mey JD, Jansen B. 2022. Improved automated early detection of breast cancer based on high resolution 3D micro-CT microcalcification images. *BMC Cancer* 22(1):162 DOI 10.1186/s12885-021-09133-4.
- Cai H, Huang Q, Rong W, Song Y, Li J, Wang J, Chen J, Li L. 2019. Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. *Computational and Mathematical Methods in Medicine* 2019(4):1–10 DOI 10.1155/2019/2717454.
- Christiansen SR, Autier P, Støvring H. 2022. Change in effectiveness of mammography screening with decreasing breast cancer mortality: a population-based study. *European Journal of Public Health* 32(4):630–635 DOI 10.1093/eurpub/ckac047.
- Hakim ANR, Prajitno P, Soejoko DS. 2021. Microcalcification detection in mammography image using computer-aided detection based on convolutional neural network. In: *AIP Conference Proceedings*. Vol. 2346.
- Hossain MS. 2022. Microcalcification segmentation using modified U-net segmentation network from mammogram images. *Journal of King Saud University-Computer and Information Sciences* 34:86–94 DOI 10.1016/j.jksuci.2019.10.014.
- Jakhar A, Gupta A, Mrityunjay S. 2022. SELF: a stacked-based ensemble learning framework for breast cancer classification. 17:1341–1356.
- Jakhar AK, Gupta A, Singh M. 2023. SELF: a stacked-based ensemble learning framework for breast cancer classification. *Evolutionary Intelligence* 7(3):104863 DOI 10.1007/s12065-023-00824-4.
- Jhang J. 2018. Localization of microcalcification on the mammogram using deep convolutional neural network. Electronic theses and dissertations. 2957. Available at <https://openprairie.sdstate.edu/etd/2957>.
- Kumar Singh K, Kumar S, Antonakakis M, Moirogiorgou K, Deep A, Kashyap KL, Bajpai MK, Zervakis M. 2022. Deep learning capabilities for the categorization of microcalcification. *International Journal of Environmental Research and Public Health* 19 DOI 10.3390/ijerph19042159.
- Leong YS, Hasikin K, Lai KW, Mohd Zain N, Azizan MM. 2022. Microcalcification discrimination in mammography using deep convolutional neural network: towards rapid and early breast cancer diagnosis. *Frontiers in Public Health* 10:145 DOI 10.3389/fpubh.2022.875305.
- Logullo AF, Prigenzi KCK, Nimir C, Franco AFV, Campos M. 2022. Breast microcalcifications: past, present and future (review). *Molecular and Clinical Oncology* 16(4):81 DOI 10.3892/mco.2022.2514.

- Mohammed A, Kora R. 2023.** A comprehensive review on ensemble deep learning: opportunities and challenges. *Journal of King Saud University—Computer and Information Sciences* 35(2):757–774 DOI [10.1016/j.jksuci.2023.01.014](https://doi.org/10.1016/j.jksuci.2023.01.014).
- Montaha S, Azam S, Rafid AKMRH, Ghosh P, Hasan M, Jonkman M, De Boer F. 2021.** BreastNet18: a high accuracy fine-tuned VGG16 model evaluated using ablation study for diagnosing breast cancer from enhanced mammography images. *Biology* 10(12):1347 DOI [10.3390/biology10121347](https://doi.org/10.3390/biology10121347).
- Rehman KU, Li J, Pei Y, Yasin A, Ali S, Mahmood T. 2021.** Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network. *Sensors (Basel)* 21 DOI [10.3390/s21144854](https://doi.org/10.3390/s21144854).
- Sharma K, Mukherjee S. 2020.** A web based MATLAB solution for classifying micro-calcification on mammograms. *International Journal of Innovative Technology and Exploring Engineering* 9(9):2439–2446 DOI [10.35940/ijitee.D2108.029420](https://doi.org/10.35940/ijitee.D2108.029420).
- Shiri Kahnouei M, Giti M, Akhaee MA, Ameri A. 2022.** Microcalcification detection in mammograms using deep learning. *Iranian Journal of Radiology* 19(1):e120758.
- Yip CH, Taib NA, Mohamed I. 2006.** Epidemiology of breast cancer in Malaysia. *Asian Pacific Journal of Cancer Prevention* 7:369–374.