Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 3, June 2024, pp. 1739~1752
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i3.pp1739-1752

A new deep learning model with interface for fine needle aspiration cytology image-based breast cancer detection

Manjula Kalita[1], Lipi B. Mahanta[2], Anup Kumar Das[3], Mananjay Nath[4]
[1]Department of Computer Science, Gauhati University, Guwahati, India
[2]Mathematical and Computational Sciences Division, Institute of Advanced Study in Science and Technology (IASST), Guwahati, India
[3]AryaWellness Centre, Guwahati, India
[4]Girijananda Chowdhury Institute of Management and Technology, Guwahati, India

Article Info

Article history:

Received Dec 20, 2023
Revised Feb 16, 2024
Accepted Feb 29, 2024

Keywords:

Breast cancer
CNN model
Deep learning
FNAC
Transfer learning

ABSTRACT

Cytological evaluation through microscopic image analysis of fine needle aspiration cytology (FNAC) is pivotal in the initial screening of breast cancer. The sensitivity of FNAC as a screening tool relies on both image quality and the pathologist's expertise. To enhance diagnostic accuracy and alleviate the pathologist's workload, a computer-aided diagnosis (CAD) system was developed. A comparative study was conducted, assessing twelve candidate pre-trained models. Utilizing a locally gathered FNAC image dataset, three superior models-MobileNet-V2, DenseNet-121, and Inception-V3-were selected based on their training, validation, and testing accuracies. Further, these models underwent evaluation in four transfer learning scenarios to enhance testing accuracy. While the outcomes were promising, they left room for improvement, motivating us to create a novel deep convolutional neural network (CNN). The newly proposed model exhibited robust performance with testing accuracy at 85%. Our research concludes that the most lightweight, high-accuracy model is the one we propose. We've integrated it into our user-friendly Android App, "Breast Cancer Detection System," in TensorFlow Lite format, with cloud database support, showcasing its effectiveness. Implementing an artificial intelligent (AI)-based diagnosis system with a user-friendly interface holds the potential to enhance early breast cancer detection using FNAC.

Corresponding Author:

Lipi B. Mahanta
Mathematical and Computational Sciences Division
Institute of Advanced Study in Science and Technology (IASST)
Guwahati, Assam, India
Email: lbmahanta@iasst.gov.in

1. INTRODUCTION

Breast cancer arises in the lining cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. Initially, the cancerous growth is confined to the duct or lobule ("in situ"), where it generally causes no symptoms and has minimal potential for spread (metastasis) [1], [2]. Breast cancer treatment can be highly effective, especially when the disease is identified early. Fine needle aspiration cytology (FNAC) is a reliable clinical method for detecting malignancy from breast cytopathology cell samples under a microscope for small breast lesions [3]. When a lump or bump is discovered in superficial areas of the body, such as the breast and neck, a test known as FNAC is recommended to determine whether the lump is cancerous. Though many clinical methods like mammograms, ultrasound, biopsy, and magnetic resonance

imaging (MRI) are available, FNAC has gained popularity due to its fast, easy, and inexpensive approach [4]. FNAC is commonly used for the diagnosis of breast cancer [5].

FNAC entails using a narrow gauge (25-22 G) needle to collect a lesion sample for microscopic examination. In the traditional approach, a visual assessment of the breast cytopathology cell samples is done under a microscope to evaluate the state of various cytological features. Therefore, there are many challenges in maintaining the consistency and reproducibility of findings. The advent of digital imaging and computational aids in diagnosis can improve diagnostic accuracy and reduce the effective workload of pathologists. Researchers and pathology practitioners have been using quantitative analysis for computer-aided diagnosis (CAD) of pathology samples, including breast. Both FNAC and core needle biopsy (CNB) are equally good in the assessment of breast lesions. However, FNAC is more suitable in developing countries like ours due to better turnaround time and cost-effectiveness. CNB can be used as the next step in assessment to minimize the chance of a missed diagnosis of breast cancer.

With recent developments in cost-effective and high-performance computer technology, medical diagnostic decision support systems (MDSSs) have become an established component of medical technology. Digital pathology has become amenable to applying quantitative analysis through decision support systems [6] and computer vision and machine learning techniques-based CAD systems. The convolutional neural network (CNN) stands out as a powerful and versatile tool for image analysis, demonstrating the capability to discern visual patterns directly from pixel data with minimal preprocessing requirements. Deep learning, especially CNNs, excels in FNAC image analysis due to its ability to automatically learn intricate patterns and features from the images. FNAC images often contain subtle details crucial for accurate diagnosis, which CNNs can effectively capture through hierarchical feature extraction. This allows for precise classification and detection of cancerous cells or abnormalities, aiding pathologists in making timely and accurate diagnostic decisions.

2. EARLIER WORK

Wolberg and Mangasarian [7] introduces a versatile mathematical method applicable to diverse medical diagnostic systems, showcasing its effectiveness in breast cytology analysis. Concurrently, the development of a CAD system for breast cancer using cytological images is explored. A review of breast cytology techniques, both manual and computer-aided breast cytology techniques, is also presented [8]. Das *et al.* [9] employ transfer learning with GoogLeNet architecture, achieving an impressive 94.67% accuracy in diagnosing breast histopathology samples across multiple magnifications. They used 5-fold cross-validation with 58 malignant and 24 benign cases of breast tumors to obtain the average accuracy.

Saikia *et al.* [5] compared deep CNN-based approaches for FNAC cell sample diagnosis, identifying GoogLeNet V3 as the optimal method based on training accuracy. They utilized an indigenous image dataset comprising 212 samples (99 benign and 113 malignant), which was subsequently augmented and cleansed to 2,120 images (990 benign and 1,130 malignant). Another study presented a meta-analysis [10], which affirmed the potential of neural network algorithms in aiding FNAC cancer diagnosis. At the same time [11] explored twenty-eight hybrid architectures combining various deep learning techniques for feature extraction (DenseNet 201, Inception V3, Inception ReseNet V2, MobileNet V2, ResNet 50, VGG16, and VGG19) with classifiers (multilayer perceptron (MLP), support vector machine (SVM), decision tree (DT), and K-nearest neighbor (KNN)) to advance binary classification of breast cytological images. Unfortunately, the dataset is wrongly cited, and they worked with the same dataset as in [5]. Also, the accuracy they claimed seems to be of training data only.

2.1. Gaps in earlier work

We identified some potential gaps in the prior studies:

i). Limited generalizability: while the studies demonstrate promising results within their specific or limited datasets and methodologies, there may be limitations in generalizing these findings to broader populations or different medical settings.

ii). Dataset heterogeneity: the studies may not adequately address the heterogeneity of FNAC images and histopathology samples, which can vary significantly regarding quality, staining, and tissue composition. A more diverse and representative dataset could better assess model performance.

iii). Validation and clinical implementation: testing play a crucial role in the classification process by rigorous evaluation of model performance, assessing generalization capabilities, detecting overfitting, guiding parameter optimization, and facilitating model comparison. None of the prior work reports test accuracy.

iv). Robustness to variability: the robustness of the developed models to variations in imaging conditions, such as different magnifications or staining techniques, may not have been thoroughly evaluated. Robust models are crucial for reliable performance across diverse clinical scenarios.
v). Model weight: the model's weight is a significant gap, especially in limited computational resources or deployment constraints. Deep learning models, especially complex ones like CNNs, can have large numbers of parameters, resulting in heavy computational and memory requirements during training, inference, and deployment. This can be impractical in resource-constrained environments, such as mobile devices or edge computing systems.

Addressing these gaps could further strengthen the validity and utility of the studies' findings in advancing FNAC image-based breast cancer detection.

3. OBJECTIVE OF STUDY

With this motivation, this paper proposed a deep CNN-based classification model for diagnosing breast FNAC cell samples and presented a comparison study with various transfer learning pre-trained models. Transfer learning models used in this study are VGG-16, VGG-19, ResNet-50, GoogLeNet-V3 (aka Inception-V3), MobileNet, MobileNetV2, DenseNet-121,169,201, NASNetMobile, EfficientNet-B0, B1. As per our knowledge, this is the first paper that proposes a CNN model for FNAC image classification that shows better testing accuracy than various transfer learning models. Figure 1 summarizes the contribution of this paper, reflecting all the steps followed.

Our proposed CNN architecture has shown better testing accuracy than the transfer learning-based CNN models. By collecting more data samples, the models' accuracy can be increased in further research. The most lightweight, high-accuracy model is concluded to be the proposed one with less number of convolution layers. To make the classification model more user-friendly, one Android App, the 'breast cancer detection system,' is designed, and the proposed classification model is integrated into a TensorFlow Lite format with a centralized cloud database.

3.1. Challenges

Every research endeavor encounters its own set of hurdles. In our study:

- Our main challenge is database preparation. We collected FNAC slides from two sources to address the first two gaps and prepared the database with more image variation.
- The second challenge is proposing a CNN model that can give better testing accuracy on unseen data. This addresses the gaps 3 and 4.
- The third challenge is to strike the optimized balance between model testing accuracy and model weight to develop a user-friendly tool to do the FNAC test.

3.2. Contributions

The following points highlight some key aspects of our endeavors:

- Our primary focus and significant time investment revolve around data collection. This is a crucial aspect of our contribution. To enhance robustness, we meticulously sourced data from two distinct channels.
- In previous research, conclusions were drawn primarily based on validation accuracy. However, recognizing the significance of testing accuracy in assessing a model's ability to generalize to new data, our focus shifted toward improving it. There was a trade-off between training and validation accuracy in this pursuit to strike a balance and attain the highest testing accuracy possible.
- Our key contribution lies in pioneering the development of a user-friendly tool for FNAC testing. Notably, no accessible tool of this nature exists to date, highlighting the novelty and importance of our endeavor. By introducing this innovative tool, we aim to enhance accessibility and efficiency in FNAC testing procedures.

The paper is organized as follows: section 4 describes the materials and methods used, and section 5 presents the classification results obtained and discusses the findings. Section 6 offers a discussion of the study. Conclusions and future work for the study are presented in section 7.

4. MATERIALS AND METHODS

The present study is conducted on breast FNAC images by applying two approaches. Approach-1 uses pre-trained transfer learning CNN models, and approach-2 uses a proposed CNN model. The study is done in five stages. Figure 1 illustrates the diagrammatic representation of the study methodology followed.

4.1. FNAC image acquisition and ground-truth labeling

Hospital-based data collected from the real-world clinical setup is fundamental in medical research. FNAC slides were collected from two distinguished medical diagnostic centres of the North-Eastern regions of India, namely from Arya Wellness Centre Pvt. Ltd, Guwahati and Life Care Diagnosis, Guwahati. Experienced doctors collected the samples from patients and prepared the slides using standard laboratory protocol. Following all institutional ethical clearance protocols (No. IEC(HS)/IASST /1082/2021/5), slides were collected from the two diagnostic centres with patient consent along with the corresponding original reports of the patients, which were later used for labelling the images. We digitally captured images using a Leica ICC50 HD microscope at ×400 resolution using 24-bit color depth and with a 5-megapixel camera associated with the microscope. The captured images are of size 2,048×1,536. Experienced certified cytopathologists then reviewed the captured digitized images, and finally, a total of 427 images (152 benign and 275 malignant) were selected for this study. We give a glimpse of some microscopic images from Figure 2 (row I show 5 benign samples and row II shows 5 malignant samples). We are considering only Pap-stained images for the study.

4.2. Image augmentation

For a cell-level study of the samples with deep learning, we require more data than the current count to feed into the CNN models. Deep learning models rely on extensive training datasets to address overfitting and achieve network generalization. To enhance CNN network generalization during testing or deployment, data augmentation is preferred to generate a broader range of trainable data representing individual classes [12].

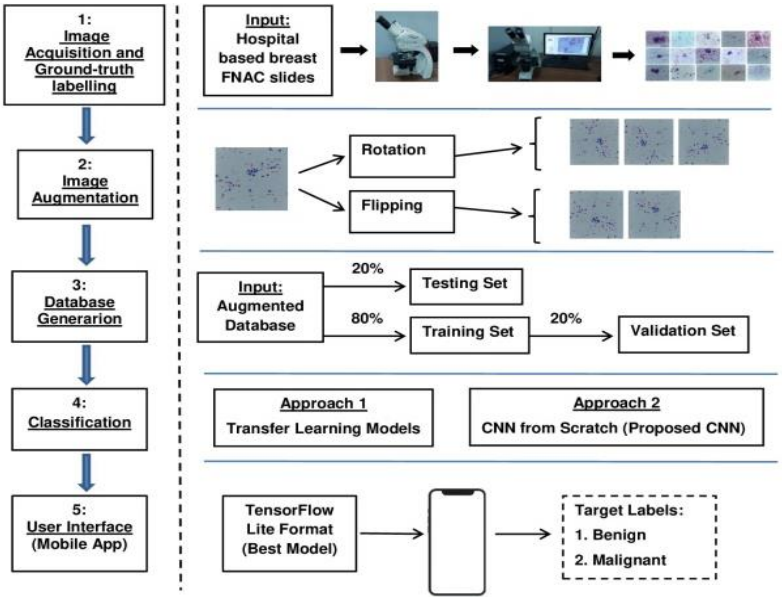


Figure 1. Workflow of the proposed method

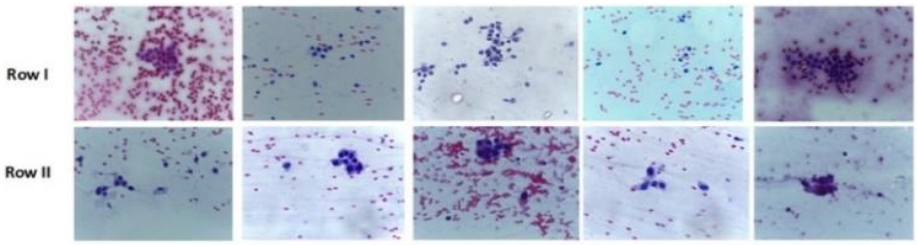


Figure 2. Papanicolaou stained breast's FNAC images

Many augmentation techniques, such as rotation, flipping, shearing, cropping, and zooming, are available. Hence, in the next step, appropriate data augmentation was the prime focus to be implemented on the acquired data. Here in our study, we have used rotation (90 degrees clockwise, 90 degrees counter-

clockwise, 180 degrees) and flipping (vertical, horizontal) only to retain the nucleus structure such that the augmentation technique will not disturb the differentiation of the main features of benign and malignant cell samples [13]. For e.g., applying zooming and shearing on a benign sample may look like a malignant sample. After augmentation, the new count increases to 2,562 images (912 benign and 1,650 malignant).

4.3. FNAC image database generation

The dataset is divided into training and testing sets at an 80:20 ratio according to the method proposed by Mohanty *et al.* [14]. Within the training set, 20% is reserved for validation to detect overfitting, while the remaining portion is used for training. The total counts for training, validation, and testing datasets are 1639, 410, and 513, respectively, with random sample selection.

4.4. Classification

4.4.1. Approach 1: pre-trained deep learning models

Pre-trained deep learning models expedite learning by reinitializing the last fully connected layers, acting as classification layers. Figure 3 schematically depicts the FNAC image classification via transfer learning with pre-trained networks. Approach 1 workflow is illustrated in Figure 4.

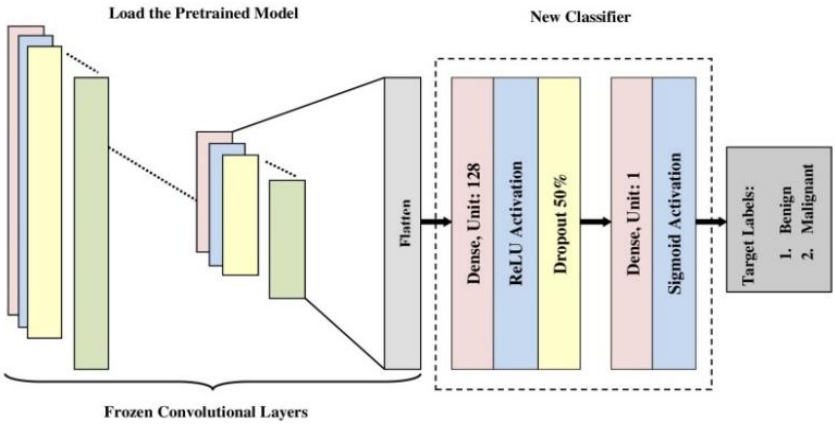


Figure 3. The fine-tuning approach using the pre-trained network

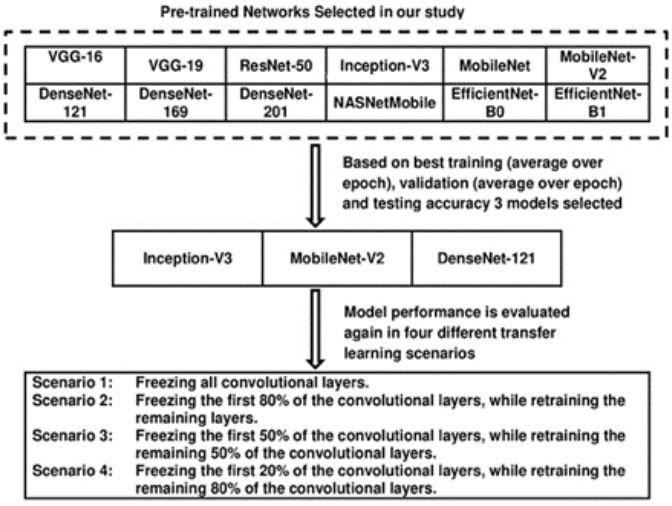


Figure 4. Approach 1–workflow

The candidate models underwent fine-tuning [15] by initializing them with pre-trained weights from the ImageNet dataset and adapting them to the specific target task through model remodeling. The original classification layer of these candidate models was truncated, and a new classifier was added that comprises two dense layers and one sigmoid layer. With two layers to classify the two categories (classes) of breast cancer, i.e., Benign and Malignant from FNAC images.

For fine-tuning classification, we chose candidate pre-trained networks, including VGG (16 and 19 layers), ResNet (50 layers), Inception-V3, MobileNet, MobileNet-V2, DenseNet (121, 169, and 201 layers), NASNetMobile, and EfficientNet (B0 and B1) from the ImageNet database to learn deep features and classify FNAC images into benign and malignant categories. The selection of these models for FNAC breast image classification is justified by their proven performance, robust architecture design, transfer learning capabilities, generalization ability, community support, and scalability. Descriptions of pre-trained models in brief are as follows:

a) VGG-16: VGG-16, short for visual geometry group (VGG) 16, is a CNN architecture designed for image classification. The inception of CNNs can be attributed to the VGG at the University of Oxford, as outlined in their seminar paper [16]. VGG16 is known for its simplicity and effectiveness. The VGG-16 architecture consists of 16 weight layers. It is characterized by a series of convolutional layers followed by max-pooling layers.
b) VGG-19: VGG-19 is a CNN architecture that belongs to the VGG family. It was introduced by researchers at the University of Oxford in 2014 [16]. The architecture is characterized by simplicity and uniformity, consisting of 19 layers (hence the name VGG-19) with a stack of convolutional layers and max-pooling layers. The VGG-19 architecture is known for its simplicity and has been widely used as a baseline in computer vision tasks. However, due to its deep structure and large number of parameters, it can be computationally expensive.
c) ResNet-50: ResNet-50, which belongs to the ResNet family, is another deep CNN architecture that was introduced by Microsoft Research in the paper titled "Deep residual learning for image recognition" by He *et al.* [17], which was presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). The key innovation of ResNet is the use of residual learning, which addresses the problem of vanishing gradients in very deep networks. The architecture introduces shortcut or skip connections, which allow the gradient to flow more easily through the network during training. ResNet-50 has been widely used in various computer vision tasks, particularly image classification.
d) Inception-V3: Inception V3 is a CNN architecture designed for image classification and object detection tasks. It was developed by Google researchers as part of the Inception project, aiming to improve the efficiency and accuracy of deep neural networks. Inception V3 has been widely used and has demonstrated strong performance on various image classification tasks [18], [19]. It balances model complexity and computational efficiency, making it suitable for research and practical applications.
e) MobileNet and MobileNet-V2 are both lightweight CNN architectures designed for efficient deployment on mobile and edge devices [20]. They aim to balance model size, computational efficiency, and accuracy well. Both MobileNet and MobileNet-V2 are designed for resource-constrained environments, but MobileNet-V2 introduces improvements in terms of model architecture, including inverted residuals and skip connections, to enhance accuracy and training efficiency.
f) DenseNet: DenseNet is a deep learning architecture introduced by Huang et al. [21] in their 2017 paper titled "Densely connected convolutional networks". The architecture addresses the vanishing gradient problem by establishing dense connections between layers, enabling each layer to receive gradients from all subsequent layers directly. DenseNet is characterized by its dense block structure, where each layer is connected to every other layer in a dense manner. The two popular variants are DenseNet-121 and DenseNet-169, both of which have 121 and 169 layers, respectively. There's also DenseNet-201, which has 201 layers. These characteristics make DenseNet an efficient and parameter-sharing architecture that performs well on various image classification tasks.
g) NASNetMobile: NASNet is a family of neural network architectures developed using neural architecture search (NAS) techniques. NASNetMobile is one specific variant designed for mobile devices with limited computational resources. NASNetMobile is tailored for mobile applications, aiming to provide a good trade-off between model accuracy and computational efficiency, making it suitable for deployment on devices with limited hardware capabilities. It is used on various image classification tasks [22].
h) EfficientNet-B0, B1: EfficientNet is a family of neural network architectures aiming to achieve better accuracy and efficiency than traditional deep learning models. The family includes models labelled from EfficientNet-B0 to EfficientNet-B7, with B0 being the smallest and B7 the largest [23], [24]. EfficientNet models are known for achieving state-of-the-art performance on various computer vision tasks while maintaining efficiency in terms of computational resources. The family is defined by a compound scaling method that simultaneously scales the network's depth, width, and resolution to find an optimal balance between model size and performance. These models have been widely used in tasks such as image classification, object detection, and segmentation, demonstrating their effectiveness across various applications. On our dataset, EfficientNet has not shown any improvement, and therefore, we have not gone to the larger architecture of the EfficientNet family and stopped experimenting with EfficientNet-B0, B1 only.

From the experimental study presented under the Results section, it is evident that we have selected the best three models, Inception-V3, MobileNet-V2, and DenseNet-121, based on best average training, average validation, and testing accuracy for further study. Next, the selected model performance is evaluated in four different transfer learning scenarios.

Varying the transfer learning scenarios (the portions of the retrained base networks) also demonstrates clear effects on model performance. The ImageNet dataset that the base network architecture is trained on is extensive compared to our target dataset. The effects of the scenarios are further reviewed.

Scenario 1: freezing all convolutional layers.

Scenario 2: freezing the first 80% of the convolutional layers while retraining the remaining layers.

Scenario 3: freezing the first 50% of the convolutional layers while retraining the remaining 50% of the convolutional layers.

Scenario 4: freezing the first 20% of the convolutional layers while retraining the remaining 80% of the convolutional layers.

But in all the cases selected, model testing performance has not improved. This may be because of our limited dataset or the complex structure of pre-trained models. So, we have found the need to design a CNN model from scratch that can give high testing accuracy to use the model in real scenarios when trained with a small dataset.

4.4.2. Approach 2: proposed CNN

A CNN model can best address the breast cancer image classification challenge by employing an optimized architecture with hyperparameter tuning to learn visual and spatial features from 224×224 pixel input images. Our proposed model, depicted in Figure 5, combines existing deep learning concepts with customized hyperparameter adjustments tailored to our problem domain. In this way, the proposed CNN sequentially transforms the original image into the targeted class, preparing a trained model for image recognition on test data. We further clarify the steps by adding the flowchart as shown in Figure 6 of our proposed architecture. A detailed description of the proposed CNN model, which consists of 13 layers, is shown in Table 1.

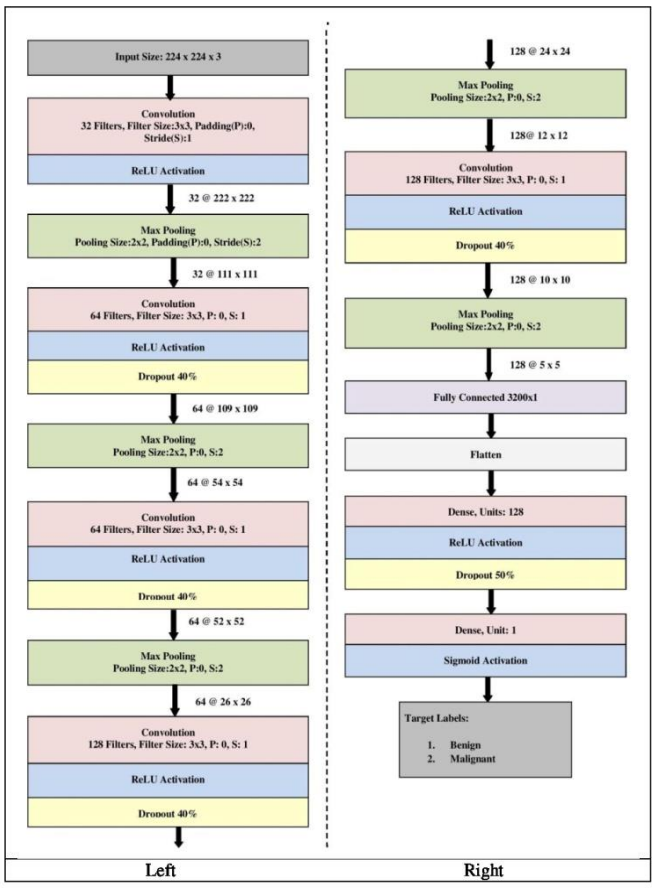


Figure 5. Approach 2 - proposed CNN model

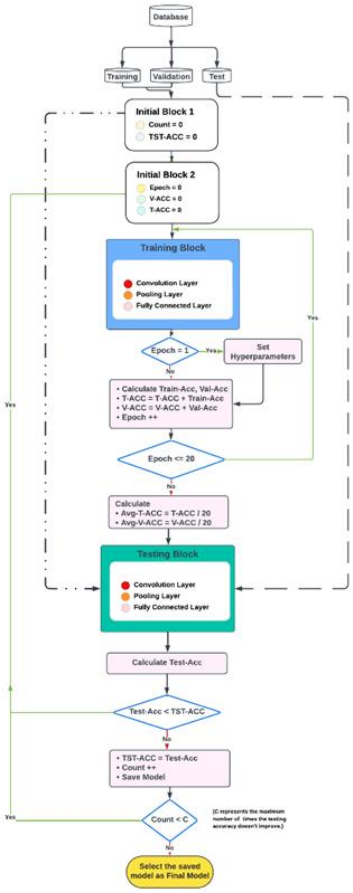


Figure 6. Flowchart of proposed CNN model

Table 1. Proposed CNN model layers

Layer	Details
0	Convolutional layer (ReLU activation), input: n×n=224×224, filter size f×f=3×3, padding p=0, stride s=1, filters=32, output: 32@222×222 where 32 is the number of filters used, and 222 comes from the formula ((n+2p−f)/s) +1.
1	Max pooling layer, input: 32@222×222, pool size=2×2, p=0, s=2, output: 32@111×111. 111 comes from the same formula ((n+2p−f)/s) +1. This layer has no activation function
2	Convolutional layer (ReLU activation), input: 32@111×111, f×f=3×3, p=0, s=1, filters=64, output: 64@109×109.
3	Max pooling layer, input: 64@109×109, pool size=2×2, p=0, s=2, output: 64@54×54.
4	Convolutional layer (ReLU activation), input: 64@54×54, f×f=3×3, p=0, s=1, filters=64, output: 64@52×52.
5	Max pooling layer, input: 64@52×52, pool size=2×2, p=0, s=2, output: 64@26×26.
6	Convolutional layer (ReLU activation), input: 64@26×26, f×f=3×3, p=0, s=1, filters=128, output: 128@24×24.
7	Max pooling layer, input: 128@24×24, pool size=2×2, p=0, s=2, output: 128@12×12.
8	Convolutional layer (ReLU activation), input: 128@12×12, f×f=3×3, p=0, s=1, filters=128, output: 128@10×10.
9	Max pooling layer, input: 128@10×10, pool size=2×2, p=0, s=2, output: 128@5×5.
10	Fully connected layer (Flattened), output: one-dimensional vector of size 3200.
11	Dense layer (ReLU activation), units: 128.
12	Dense layer (Sigmoid activation), units: 1

4.5. Implementation

The proposed CNN is implemented in Spyder version 5.1.5 environment using Python version 3.6 in Keras and TensorFlow 2.8.0 platform on Windows 11. All experiments were conducted in an Intel(R) Core(TM) i5-10300H CPU @2.50GHz with 8 GB RAM and 4 GB NVIDIA GeForceGTX 1650 Graphics card. Data preparation stands as the initial step in our approach. Images were stored in two different directories representing the two different classes. Images were arranged according to the reports of the patients collected and with proper consultation with the pathologist. Evaluating pre-trained models' performance on our dataset is the second step, as described in Figure 4. Building and compiling the proposed CNN model is the third step as shown in Figure 5. Defining the initialization parameters, such as batch size, the number of epochs, and learning rate, is very important to create the CNN model. The proposed CNN model's accuracy (training, validation and testing) is compared with other pre-trained models in the fourth step. The higher the testing accuracy, the higher the rate of truly classified classes.

4.6. Performance metrics

The testing accuracy of classification was evaluated by computing the number of correctly recognized benign class samples true positive (TP), the number of correctly identified malignant class samples true negative (TN), and samples that either were incorrectly assigned to the benign class false positive (FP) or were incorrectly assigned to the malignant class false negative (FN) [25]. Precision is the percentage of the positive predictions made that were correct. The recall is the percentage of actual positive predictions correctly classified by the classifier. F1-score is defined as the weighted harmonic mean of both recall and precision.

Accuracy: accuracy gives the correct predictions from all the classes (positive and negative) [26].

$$Accuracy\ (\%) = \frac{TP + TN}{TP + FN + FP + TN} \times 100 \quad (1)$$

Precision: the precision measures the model's accuracy in classifying a sample as positive.

$$Precision\ (\%) = \frac{TP}{TP + FP} \times 100 \quad (2)$$

Recall: the recall measures the model's ability to detect positive samples.

$$Recall\ (\%) = \frac{TP}{TP + FN} \times 100 \quad (3)$$

F1-score: precision-recall values are valuable for assessing the performance of individual algorithms and tailoring results to specific needs. Yet, comparing multiple algorithms trained on the same data solely using Precision-Recall values can be challenging. Therefore, a standardized metric is essential to make such comparisons more comprehensible.

$$\text{F1-score}\ (\%) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \quad (4)$$

4.7. Design user interface

An Android app is designed to assist pathologists using the latest Android Studio and Cloud database. This can help pathologists detect benign and malignant categories using a user-friendly interface. Detailed descriptions of app interfaces are shown in the result section.

5. RESULTS

Max pooling with a 3×3 kernel and batch size 16 consistently outperformed in all experiments. After 20 epochs, no further improvement was observed, leading us to fix the number of epochs at 20 with these settings for subsequent runs, ensuring stability without compromising accuracy. Results of hyper-parameters tuning of the pre-trained models and proposed CNN are shown in the below sections.

5.1. Assessing pre-trained networks with varied optimizers and learning rates

Utilizing optimizers is crucial for enhancing model optimization and improving training accuracy by minimizing loss. In this experiment, three well-known optimizers-RMSprop, Adam, and SGD-were employed with various learning rates (LR) to identify the most suitable pre-trained model for our classification task. Table 2 represents the best hyperparameter setting of the pre-trained models that give the best accuracy on our dataset.

From the experimental study explained in Table 2, we have found that Inception-V3 (using optimizer RMSprop, LR 0.00001), MobileNet-V2 (using optimizer Adam, LR 0.0001), and DenseNet-121 (using optimizer Adam, LR 0.0001) have shown the better performance among all and therefore can be chosen for our further study. In the preliminary stage of the results analysis, we did not use the four scenarios of freezing layers as they would be computationally expensive. After obtaining the three best-performing pre-trained models, four scenarios were tested, as shown in Table 3, to ensure a proper replicable evaluation of the model. A bold value indicates the best scenario and evaluation score for each model in Table 3. MobileNet-V2 (using Adam optimizer, LR 0.0001 on Sc 1) can be considered better among all the transfer learning models as it performs well on unseen data. However, in all cases, the testing performance of pre-trained models is not so satisfactory.

Table 2. Performance of pre-trained models over 20 epochs (using the best hyper-parameters setting)

Pre-trained model	Learning rate	Optimizer	Training accuracy (Avg)	Validation accuracy (Avg)	Testing accuracy
VGG-16	0.0001	Adam	93.78	90.35	62
VGG-19	0.0001	Adam	90.25	85.86	61
ResNet-50	0.00001	Adam	70.58	73.11	58
Inception-V3	**0.00001**	**RMSprop**	**96.15**	**93.82**	**63**
MobileNet	0.0001	Adam	98.17	97.46	43
MobileNet-V2	**0.0001**	**Adam**	**98.21**	**98.18**	**71**
DenseNet-121	**0.0001**	**Adam**	**97.49**	**95.89**	**64**
DenseNet-169	0.0001	Adam	93.32	96.63	45
DenseNet-201	0.0001	Adam	95.96	96.93	51
NASNetMobile	0.0001	Adam	93.44	94.26	60
EfficientNetB0	0.0001	Adam	63.43	64.29	44
EfficientNetB1	0.0001	Adam	64.38	64.39	46

Table 3. Evaluation scores of Inception-V3, MobileNet-V2, and DenseNet-121 in different scenarios

Model	Scenario	Trainable parameters	Non-trainable parameters	Training accuracy (Avg)	Validation accuracy (Avg)	Testing accuracy
Inception-V3 (RMSprop Optimizer, LR 0.00001)	Sc 1	6,553,857	21,802,784	96.15	93.82	63
	Sc 2	**17,668,737**	**10,687,904**	**96.27**	**94.17**	**65**
	Sc 3	23,343,297	5,013,344	96.13	93.17	57
	Sc 4	27,617,521	739,120	96.33	94.68	44
MobileNet-V2 (Adam Optimizer, LR 0.0001)	**Sc 1**	**8,028,417**	**2,257,984**	**98.21**	**98.18**	**71**
	Sc 2	9,646,977	639,424	98.17	87.00	36
	Sc 3	10,091,905	194,496	98.12	92.00	65
	Sc 4	10,232,753	53,648	98.20	74.35	36
DenseNet-121 (Adam Optimizer, LR 0.0001)	Sc 1	6,422,785	7,037,504	97.49	95.89	64
	Sc 2	8,143,233	5,317,056	95.71	95.29	64
	Sc 3	**11,054,657**	**2,405,632**	**98.44**	**99.21**	**67**
	Sc 4	12,725,249	735,040	98.17	98.46	64

5.2. Hyper-parameters tuning of the proposed CNN

In this section, the experimental performance of the proposed CNN model is demonstrated. The experiments were done with three renowned optimizers- RMSprop, Adam, and SGD with different learning rates and numbers of epochs to recognize the most suitable configuration of our proposed model for our classification. Table 4 shows the performance of the proposed CNN with different hyperparameter settings. Bold values indicate the best parameter setting. As the model's performance is not improving after 20 epochs with different hyper-parameters settings, we have fixed the epoch at 20 for further experiments of the model.

Table 4. Proposed CNN performance with different hyper-parameters setting

Epoch	Optimizer	Learning rate	Training accuracy (Avg)	Validation accuracy (Avg)	Testing accuracy
20	**Adam**	**0.001**	**90.00**	**90.61**	**85**
20	RMSprop	0.001	86.97	89.09	82
20	SGD	0.001	65.04	63.70	64
25	Adam	0.001	90.46	91.17	74
25	RMSprop	0.001	87.19	87.98	75
30	Adam	0.001	91.27	90.27	76
20	RMSprop	0.0001	78.95	71.84	73
20	Adam	0.0001	77.14	75.67	78
25	Adam	0.01	64.36	64.45	64
20	Adam	0.01	64.41	64.45	64
20	RMSprop	0.01	64.45	64.35	64

5.3. Comparative assessment of pre-trained networks and proposed CNN with various performance metrics

From Table 5, we can observe that the testing accuracy of the proposed model is 85%, which is higher than the testing accuracy of all the transfer learning models when tested with an unseen dataset (test dataset). Further, the confusion matrix in Table 5 and training accuracies in Table 2 reveal that transfer learning models are overfitting on the training dataset [26]. Testing accuracy is crucial to check the performance of a model. A model can be considered good if it performs well when tested with an unseen dataset (test dataset).

Table 5. Testing performance comparisons of the proposed CNN with transfer learning models (bold values indicating the best results)

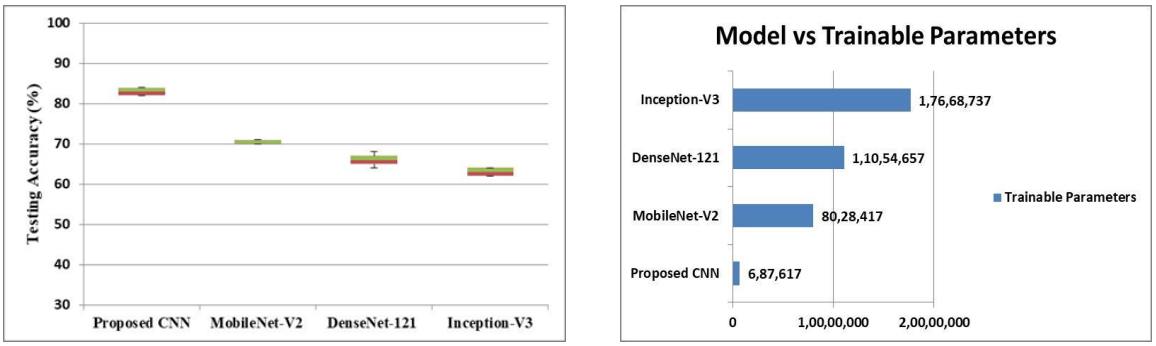
Model	Class	Precision	Recall	F1 score	Confusion matrix	Testing accuracy
Proposed CNN (Adam optimizer, LR=0.001)	**Benign**	**0.94**	**0.61**	**0.75**	**[[112 71]**	**85**
	Malignant	**0.82**	**0.98**	**0.89**	**[7 323]]**	
MobileNet-V2 (Adam optimizer, LR 0.0001, freezing all layer)	Benign	0.68	0.36	0.47	[[65 118]	71
	Malignant	0.72	0.91	0.80	[30 300]]	
DenseNet-121 (Adam optimizer, LR=0.0001, freezing top 50% layer)	Benign	0.77	0.11	0.19	[[20 163]	67
	Malignant	0.67	0.98	0.79	[6 324]]	
Inception-V3 (RMSprop optimizer, LR=0.00001, freezing top 80% layer)	Benign	0.65	0.06	0.11	[[11 172]	65
	Malignant	0.65	0.98	0.78	[6 324]]	

For further cross-validation of our model, we have designed four more datasets containing training, validation, and testing following the rule mentioned in section 4.3. The experimental results of the proposed CNN and three selected transfer learning models on five datasets are represented in Table 6. Data set 2 is used for all the results mentioned above in the analyses.

Table 6. Training (Tr), validation (Vl) and testing (Ts) result of five datasets

Data Set ⇩	Proposed CNN (Adam Optimizer, LR 0.001)			MobileNet-V2 (Adam Optimizer, LR 0.0001, Freezing all layer)			DenseNet-121 (Adam Optimizer, LR=0.0001, Freezing top 50% layer)			Inception-V3 (RMSprop Optimizer, LR=0.00001, Freezing top 80% layer)		
	Tr	Vl	Ts	Tr	Vl	Ts	Tr	Vl	Ts	Tr	Vl	Ts
DS1	92	91	82	99	97	70	98	98	66	96	94	65
DS2	90	91	85	98	98	71	98	99	67	96	94	65
DS3	89	89	83	98	96	70	98	98	65	95	95	64
DS4	91	91	85	97	96	71	97	97	65	96	95	63
DS5	91	91	84	99	98	70	99	98	68	96	94	62

Performance comparisons of the proposed CNN and three selected transfer learning models are shown in Figure 7. As Figure 7(a) shows, the box plot displays that the median testing accuracy of the proposed CNN is higher than that of the other networks. Bar graph representation of trainable parameters, Figure 7(b), reveals that the proposed CNN is the most lightweight model. The experimental study concludes that our proposed CNN is the most accurate and lightweight model.



(a) (b)

Figure 7. Comparing performance of different models (a) testing performance comparisons and (b) trainable parameters comparisons

5.4. User interface

As a high-accuracy and lightweight model, the proposed CNN model in TensorFlow Lite format is integrated into our "Breast Cancer Detection System" with a user-friendly interface and centralized cloud database. Some screenshots are displayed in Figures 8 and 9. Admin login and patient registration process are shown in Figure 8. Figure 9 depicts the picture of a patient diagnosed as negative (Malignant case) and a patient diagnosed as positive (Benign case) who has undergone an FNAC test.

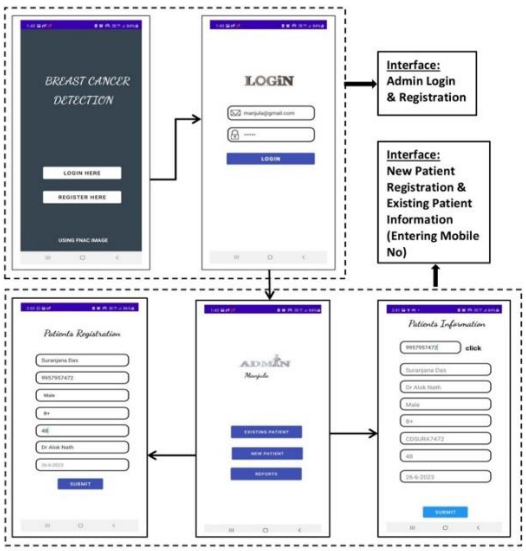


Figure 8. Admin login and patient registration

Figure 9. Patient FNAC test

6. DISCUSSION

The experimental results present a comprehensive analysis of various aspects of breast cancer detection models, including the evaluation of pre-trained networks, hyper-parameter tuning of the proposed CNN, and comparisons based on different performance metrics. Several points stand out for discussion as follows:

a) Evaluation of pre-trained networks: the study conducted an in-depth analysis of pre-trained models using different optimizers and learning rates. Notably, Inception-V3, MobileNet-V2, and DenseNet-121 emerged as the top-performing models, exhibiting promising accuracy on our datasets. Further investigation through scenario-based testing shed light on the impact of freezing layers on model performance. By leveraging these models, clinicians can make more reliable and timely diagnoses, leading to better patient outcomes and reduced misdiagnosis rates.
b) Hyper-parameter tuning of proposed CNN: the proposed CNN model underwent extensive experimentation to identify the optimal hyper-parameter settings. Results demonstrated the sensitivity of model performance to variations in optimizer, learning rate, and number of epochs. Adam optimizer with a learning rate 0.001 consistently yielded competitive performance across different epochs. This knowledge can guide future researchers and practitioners in designing efficient and effective deep-learning models for medical image analysis tasks.
c) Evaluation based on performance metrics: comparative analysis based on precision, recall, and F1 score highlighted the robustness of the proposed CNN model, showcasing superior performance compared to transfer learning models across multiple datasets. The evaluation also revealed insights into the potential of overfitting observed in some transfer learning models, emphasizing the importance of testing accuracy for model validation. This approach can serve as a blueprint for evaluating and benchmarking other machine learning models in the domain of medical imaging, fostering the development of more reliable and widely applicable diagnostic tools.
d) User interface integration: the successful integration of the proposed CNN model into the "Breast Cancer Detection System" app underscores its practical utility and accessibility for end-users. The user-friendly interface, coupled with centralized cloud database support, enhances the usability and scalability of the application, facilitating seamless breast cancer detection. Such applications empower healthcare providers with accessible and intuitive breast cancer screening and diagnosis tools, potentially expanding access to healthcare services and improving patient outcomes, especially in resource-constrained settings.

7. CONCLUSION AND FUTURE WORK

In this study, we have showcased the development of a CNN model created from the ground up, designed specifically for precisely classifying FNAC images related to breast cancer. This novel CNN model exhibits a remarkable testing accuracy, consistently outperforming other established deep-learning Keras models, with an impressive maximum testing accuracy of 85%. Furthermore, our CNN model outperforms various transfer learning classifiers on our dataset. To further improve accuracy, we suggest expanding the database and optimizing CNN parameters, including the number of convolution layers and hyper-parameters. Our lightweight, high-accuracy model has been seamlessly integrated into the "Breast Cancer Detection System" Android App. This tool streamlines breast cancer diagnosis via FNAC images, offering efficiency and accuracy. In the future, we plan to compare the proposed CNN approach with machine learning classifiers and some hybrid deep learning networks. In conclusion, our study highlights the importance of ongoing innovation and collaboration in the field of medical image analysis. By fostering interdisciplinary collaboration between computer scientists, medical professionals, and industry partners, we can accelerate the development and adoption of advanced technologies for breast cancer detection, ultimately saving lives and improving public health.

ACKNOWLEDGEMENTS

The authors express their gratitude to the Head of the Department of CSE, GU, Guwahati, India, and the Institute of Advanced Study in Science and Technology Director, Guwahati, India, for giving them the platform to carry forward this work. They would like to thank the Principal, GIMT-Guwahati, India, for allowing them to pursue a PhD. They would also like to thank Dr. Ahmed Omar Sherif, Laboratory Director, Life Care Diagnostics, Guwahati, India, for helping them with data collection. This work has the consent of all the co-authors and authorities of the institute where this study has been carried out, and no conflict of interest exists anywhere.

REFERENCES

[1] World Health Organization, "Breast cancer," *World Health Organization*, Jul. 12, 2023, https://www.who.int/news-room/fact-sheets/detail/breast-cancer. (accessed Feb. 26, 2024).
[2] "What is ductal carcinoma in situ (DCIS) - Sunnybrook Hospital," *sunnybrook.ca*. https://sunnybrook.ca/content/?page=cancer-ductal-carcinoma-info. (accessed Feb. 26, 2024).

[3] J. A. T. de Cursi, M. E. A. Marques, C. A. C. de A. C. Castro, F. C. Schmitt, and C. T. Soares, "Fine-needle aspiration cytology (FNAC) is a reliable diagnostic tool for small breast lesions (≤1.0 cm): a 20-year retrospective study," *Surgical and Experimental Pathology*, vol. 3, no. 1, Dec. 2020, doi: 10.1186/s42047-020-00081-0.
[4] A. Berner and T. Sauer, "Fine-needle aspiration cytology of the breast," *Ultrastructural Pathology*, vol. 35, no. 4, pp. 162–167, Jun. 2011, doi: 10.3109/01913123.2011.576327.
[5] A. R. Saikia, K. Bora, L. B. Mahanta, and A. K. Das, "Comparative assessment of CNN architectures for classification of breast FNAC images," *Tissue and Cell*, vol. 57, pp. 8–14, Apr. 2019, doi: 10.1016/j.tice.2019.02.001.
[6] D. West and V. West, "Model selection for a medical diagnostic decision support system: a breast cancer detection case," *Artificial Intelligence in Medicine*, vol. 20, no. 3, pp. 183–204, Nov. 2000, doi: 10.1016/s0933-3657(00)00063-4.
[7] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences*, vol. 87, no. 23, pp. 9193–9196, Dec. 1990, doi: 10.1073/pnas.87.23.9193.
[8] M. Saha, R. Mukherjee, and C. Chakraborty, "Computer-aided diagnosis of breast cancer using cytological images: a systematic review," *Tissue and Cell*, vol. 48, no. 5, pp. 461–474, Oct. 2016, doi: 10.1016/j.tice.2016.07.006.
[9] K. Das, S. P. K. Karri, A. Guha Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, VIC, Australia, 2017, pp. 1024-1027, doi: 10.1109/ISBI.2017.7950690.
[10] J. Huang, D. Wang, and J. Da, "Automated classification of cancer from fine needle aspiration cytological image use neural networks: a meta-analysis," *Diagnostic Cytopathology*, vol. 48, no. 11, pp. 1027–1033, Jun. 2020, doi: 10.1002/dc.24520.
[11] H. Zerouaoui, A. Idri, F. Z. Nakach, and R. E. Hadri, "Breast fine needle cytological classification using deep hybrid architectures," *Lecture Notes in Computer Science*, pp. 186–202, Jan. 2021, doi: 10.1007/978-3-030-86960-1-14.
[12] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018*, pp. 117–122, 2018, doi: 10.1109/IIPHDW.2018.8388338.
[13] L. B. Mahanta, M. Kalita, and A. Das, "A review of automated digital clinical system of breast cancer detection using fine needle aspiration cytology images," *Annals of Oncology Research and Therapy*, vol. 1, no. 1, pp. 10–10, Jan. 2021, doi: 10.4103/aort.aort_6_21.
[14] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, Sep. 2016, doi: 10.3389/fpls.2016.01419.
[15] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, Mar. 2018, doi: 10.1016/j.compag.2018.03.032.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, Available: http://arxiv.org/abs/1409.1556.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Jun. 2016, doi: 10.1109/cvpr.2016.90.
[18] N. Dong, L. Zhao, C. H. Wu, and J. F. Chang, "Inception v3 based cervical cell classification combined with artificially extracted features," *Applied Soft Computing*, vol. 93, p. 106311, 2020.
[19] C. Wang *et al.*, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146533–146541, 2019.
[20] A. Michele, V. Colin, and D. D. Santika, "MobileNet convolutional neural networks and support vector machines for palmprint recognition," *Procedia Computer Science*, vol. 157, pp. 110–117, Jan. 2019, doi: 10.1016/j.procs.2019.08.147.
[21] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Jul. 2017, doi: 10.1109/cvpr.2017.243.
[22] A. Ghodekar and A. Kumar, "LightLeafNet: Lightweight and efficient NASNetmobile architecture for tomato leaf disease classification," *Available at SSRN 4570347*, Jan. 2023, doi: 10.2139/ssrn.4570347.
[23] "EfficientNet [model architecture]," *OpenGenus IQ: Computing Expertise and Legacy*, Nov. 05, 2022, https://iq.opengenus.org/efficientnet/. (accessed Feb. 26, 2024).
[24] K. Team, "Keras documentation: EfficientNet B0 to B7," *keras.io*. https://keras.io/api/applications/efficientnet/#efficientnetb0-function. (accessed Feb. 26, 2024).
[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[26] P. Shivaprasad, "Understanding confusion matrix, precision-recall, and F1-score," *Medium*, Oct. 19, 2020, https://towardsdatascience.com/understanding-confusion-matrix-precision-recall-and-f1-score-8061c9270011. (accessed Feb. 26, 2024).

BIOGRAPHIES OF AUTHORS

Manjula Kalita is a Research Scholar at the Department of Computer Science, Gauhati University (GU), Guwahati, Assam, India, and an Assistant Professor at Girijananda Chowdhury University-Guwahati, Assam, India. She holds an M.Tech. Degree in Information Technology from Central University, Tezpur, Assam, and an M.Sc. in Computer Science Department from Gauhati University, Guwahati, Assam. Her research interests include medical image analysis, machine learning, deep learning, data mining, and intelligent systems. She has experience in handling B.Tech. and M.Tech. Projects on different topics of computer science. She has good paper publications in IEEE, reputed peer-reviewed journals, and conferences. She has long teaching experience with other responsibilities like SIRO coordinator and NBA coordinator. She can be contacted at email: manjula_cse@gcuniversity.ac.in.

Dr. Lipi B. Mahanta an Associate Professor (II) at the Mathematical and Computational Sciences Division of the Institute of Advanced Study in Science and Technology, Guwahati, Assam, India (https://iasst.gov.in/), specializes in medical image analysis using artificial intelligence, epidemiology, and operations research techniques, particularly in cancer research. She obtained her Ph.D. in Statistics in 2004 from Gauhati University, Guwahati, Assam. With expertise in computer vision and pattern recognition, including domains such as ballistics and biometrics, she has authored over 90 papers in national and international journals. Her entrepreneurial venture focuses on translating advanced AI and ML algorithms into practical software solutions. She can be contacted at email: lbmahanta@iasst.gov.in.

Dr. Anup Kumar Das received his MBBS degree from Gauhati Medical College, Guwahati, Assam, India, and an MD in Pathology from All India Institute of Medical Sciences, New Delhi, India. He submitted a thesis on "Fine needle aspirates of breast carcinoma" during MD. He has twelve years of experience as a pathologist with a particular interest in histopathology and cytopathology and nine years of experience in an administrative role in planning, setting up, and running independent diagnostic centers. He has good research experience with 15 publications in reputed journals. He can be contacted at email: dranupdas@gmail.com.

Mananjay Nath received his B.Tech degree in Computer Science and Engineering from Girijananda Chowdhury Institute of Management and Technology, Guwahati, Assam, India. He also holds a diploma in computer engineering from Nowgong Polytechnic, Assam, India. During his B.Tech. studies, he worked on many projects using machine learning, Java, and PHP technologies. Now, he works as a web developer at Xeotec eService. His job is making websites that are good and easy for people to use. He can be contacted at email: mananjaynath98@gmail.com.