FusedMammoNet: Ensemble of diverse models for multi-class mammogram analysis

M. Gayathri[1], M. Sterena Lilly[2], N. Sri Lekha[3], G. Siva Bhavani[4]

[1] *Under Graduate, ECE, Vasireddy Venkatadri Institute Of Technology, Andhra Pradesh, India*
[2] *Under Graduate, ECE, Vasireddy Venkatadri Institute Of Technology, Andhra Pradesh, India*
[3] *Under Graduate, ECE, Vasireddy Venkatadri Institute Of Technology, Andhra Pradesh, India*
[4] *Under Graduate, ECE, Vasireddy Venkatadri Institute Of Technology, Andhra Pradesh, India*

ABSTRACT

Breast cancer remains a significant global health challenge, necessitating accurate and efficient detection methods to improve patient outcomes. Mammography serves as a cornerstone for early diagnosis, yet the interpretation of mammograms can be prone to errors, leading to both false positives and missed diagnoses. In response to this critical issue, this study focuses on harnessing the power of Convolutional Neural Networks (CNNs) for the automated detection of breast cancer in mammographic images. The research investigates a diverse range of deep learning techniques, including popular network architectures such as VGG19, ResNet152, InceptionV3, DenseNet121, MobileNetV2, and EfficientNetB0. Various factors crucial to model performance are explored, such as class weighting strategies, input image dimensions, preprocessing methodologies, transfer learning approaches, dropout rates, and the impact of different mammogram projections. Through a systematic and comprehensive analysis, this project aims to evaluate the effectiveness and efficiency of these deep learning methodologies in the context of breast cancer detection. By employing a divide-and-conquer approach, the study seeks to gain valuable insights into selecting the most suitable techniques for enhancing detection accuracy while minimizing the need for extensive trial and error experimentation. The ultimate goal of this research is to advance automated breast cancer screening by optimizing deep learning models for mammogram analysis. By understanding the nuances of various parameters and their impact on model performance, this study aims to contribute to improved diagnostic accuracy and ultimately enhance patient care in the realm of breast cancer detection. The proposed FusedMammoNet model achieved a test accuracy of 96%, recorded highest AUC-ROC ranged from 0.98-1.00 and both precision and recall ranging from 93% to 94%.

Keywords: - *Mammography, CNN, Deep Learning, EfficientNetB0, MobileNetV2, inceptionV3, Transfer Learning, ensemble model.*

1. INTRODUCTION

At 268,600 diagnoses and 41,760 deaths predicted for 2019, breast cancer is the second most common cause of cancer-related deaths among women in the US. Early detection is still essential in the fight against this common disease, and screening mammography, a low-dose X-ray test, is essential in detecting breast cancer in its early stages. However, because of the great quality of the pictures and the subtle character of asymptomatic cancer lesions, which are frequently small and sparsely dispersed, radiologists face a significant problem when interpreting screening mammograms.

Screening mammography has been shown to be effective in lowering the death rate from breast cancer, but it has several drawbacks. In a large majority of cases, false-positive recalls and later biopsies yield benign results. Convolutional neural networks (CNNs) have been used to help radiologists with mammography analysis in order to overcome these constraints. Studies now in progress frequently modify models intended for natural image analysis, like VGGNet and Faster R-CNN, to address detection issues with breast cancer.

Given its high mortality rate and status as the most funded malignancy in the US, breast cancer provides important insights into the evolution of tumors. Early identification is still crucial, particularly in nations where access to health care is restricted and fatality rates are greater. Biopsies, ultrasounds, mammograms, and MRI results are among the diagnostic techniques. The form and features of the tumor determine its nature, whether benign or

malignant, underscoring the significance of precise mass segmentation in computer-aided diagnosis systems for reliable categorization. Breast cancer is the most prevalent cancer in women worldwide, which emphasizes the need for early identification, research, and efficient treatment.

Taking into account the intrinsic distinctions between natural and mammography images, this work presents ensemble model. This model utilizes all the features of efficientnet b0, inceptionV3, mobilenet models. It helps to reduce the False Positives, False Negative and increases the accuracy.

2. LITERATURE SURVEY

Due to limited labeled data for mammography, transfer learning is crucial for effective model training. The study discusses the challenges of small public databases (e.g., DDSM mammography dataset) and the use of transfer learning to improve classification accuracy. In this study we have used three pre-trained model weights which is tested on ImageNet dataset. Aboutalib et al[4] used an incremental approach for a 3-class classification task on mammograms, achieving varying performance on benign and malignant cases and achieved AUC ranged from 0.77 to 0.96 for DDSM dataset while for the propsed model achieved AUC ranged from 0.98 to 100 on DDSM dataset and considered all the 5 classes. Mohiyuddin et al[3] achieved a record accuracy of 95.50% using YOLOv5, focusing on masses in mammograms while the propsed system achieved highest accuracy of 96% on the same DDSM dataset. Chun-ming et al[5] employed a Deep Cooperation Neural Network with two parallel CNNs for a 5-class classification task. Pre-annotated ROIs were used, resulting in an accuracy of 91% for negative-class classification and for the same ROIs our model achieved an accuracy of 96% for the same 5 class classification. Levy et al[6] utilized GoogleNet with pre-annotated ROIs and achieved an accuracy of 93% which is less than our achieved accuracy of 96%, precision of 92% and recall of 93% while the proposed model achieved precision and recall ranged 93% to 94%.

3. DATASET

The dataset used in this study is a compilation of images sourced from the DDSM [1] and CBIS-DDSM [2] datasets. These images have undergone pre-processing, including the conversion into 299x299 dimensions by extracting Regions of Interest (ROIs). The dataset is stored in TensorFlow's tfrecords format.

In terms of composition, the dataset consists of 55,890 training examples. Among these, 14% are labeled as positive, while the remaining 86% are categorized as negative. The data is organized into 5 tfrecords files. Here we are using only 3 tfrecords for training, validation and testing datasets.

An important note regarding data division: The separation of data into training and test sets aligns with the CBIS-DDSM dataset's categorization. However, an unintended split occurred in the test files. Specifically, the test numpy files exclusively contain masses, while validation files exclusively contain calcifications. A correction involves merging these files to establish a balanced and comprehensive test dataset.

Moving on to pre-processing details, negative (DDSM) images underwent a two-step process involving tiling into 598x598 tiles and subsequent resizing to 299x299. Positive (CBIS-DDSM) images followed a distinct procedure. Regions of Interest (ROIs) were extracted using masks, with a slight padding for contextual information. Each ROI underwent three random crops into 598x598 images, including random flips and rotations. Finally, these images were resized down to 299x299.

In terms of labeling, the images are assigned two labels. The first, labeled as "label_normal," is set as 0 for negative instances and 1 for positive instances. The second, simply labeled as "label," follows a full multi-class system. Here,

1. negative
2. benign calcification
3. benign mass
4. malignant calcification
5. malignant mass.

4. PROCESSING AND AUGMENTATION

Since the DDSM dataset has data in the form of TFRecord files, a popular format for storing large amounts of data efficiently, the raw image data is decoded using TensorFlow's decode_raw function, converting it from a string to a NumPy array of unsigned 8-bit integers (tf.uint8). The image is resized to [224, 224] using OpenCV (cv2.resize). This is done because the neural network models used in the proposed model are efficientNet,

mobileNet which expects the input size [224,224]. For inceptionV3 model we trained the model with original size [299,299].

The image pixel values are normalized by dividing them by 255. This is a standard practice to scale pixel values to the range [0, 1] before feeding them into a neural network. The image is stacked along the first axis to create a 3-channel image. This is done using np.stack([image, image, image], axis=0). This step is necessary because neural networks commonly expect input images with three channels (e.g., RGB), and the original images are assumed to be single-channel. The resulting image is converted to a PyTorch tensor using torch.from_numpy(image).float() to ensure it has the correct data type for PyTorch. The label is also converted to a PyTorch tensor using torch.tensor(label, dtype=torch.long).

5. PROPOSED WORK

We have experimented this dataset on different models like efficientnet b0, Vision Transformer, InceptionV3, DenseNet, MobileNetV2, ResNet152, VGG19, GoogleNet models. Based on our needs the top 3 models, which performed better based on selecting suitable useful features, are EfficentNet, InceptionV3, MobileNet. Based on this we performed stacking ensemble on the three selected models.

MobileNetV2: A thin neural network architecture called MobileNetV2 is intended for mobile and edge devices. To keep performance high while lowering computational complexity, depthwise separable convolutions are used. Effective in real-time picture classification problems where computing resources are limited.

InceptionV3: InceptionV3, developed by Google, is part of the Inception family[8] of neural network architectures. Known for its deep architecture with inception modules that incorporate multiple filter sizes. Effective in capturing both local and global features, making it suitable for various computer vision tasks.

EfficientNetB0: EfficientNetB0 is part of the EfficientNet[7] family, known for achieving state-of-the-art performance with efficient model scaling. These models balance depth, width, and resolution to optimize model size and computational efficiency. Provides competitive accuracy with lower computational requirements compared to larger models.

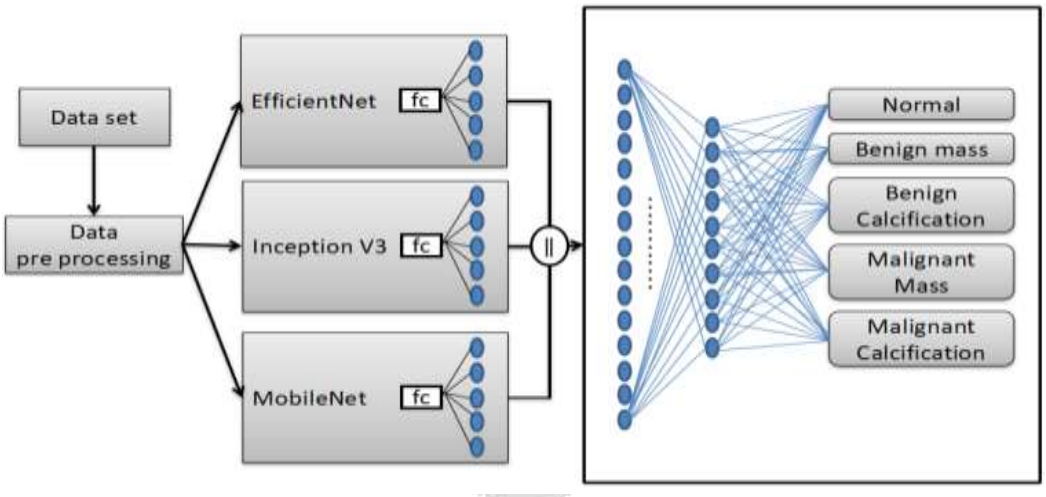

Fig-1 : Block diagram of FusedMammoNet model

Three pre-trained models (EfficientNetB0, InceptionV3, and MobileNetV2) are used as input to the ensemble model along with the output classes. Forward passes through each individual model (modelA, modelB, modelC). Concatenate the model outputs along the feature dimension. Two fully connected layers are used in this enseble model. First fully connected layer with input size 3 * input (concatenated feature dimension) and output size 10. Second fully connected layer with input size 10 and output size 'number of classes'. Softmax activation is applied to the output.

6. EXPERIMENT AND RESULT

6.1 Accuracy and Loss: Accuracy measures the proportion of correctly classified cases from the total number of objects in the dataset. To compute the metric, divide the number of correct predictions by the total number of predictions made by the model.

Accuracy=(TP+TN)/(TP+FP+FN+TN)

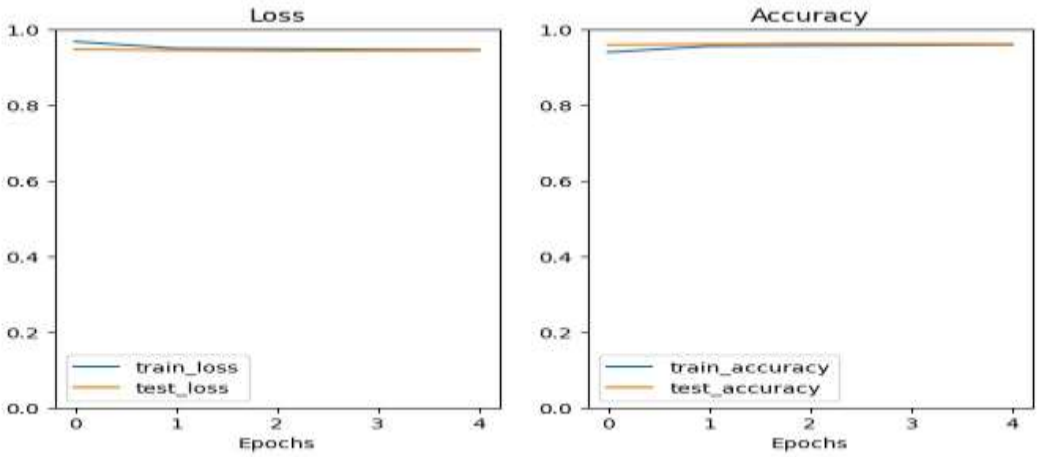

Fig-2 : Accuracy and loss curves

Values	Training	Validation
Accuracy	95.88	96.10
Loss	0.9458	0.9433

Table-1: Accuracy and loss values

Precision: In multiclass classification, precision is calculated for each class individually. Precision for a given class is defined as the fraction of instances correctly classified as belonging to that specific class out of all instances the model predicted to belong to that class. The formula for precision (P) for a class "C" can be expressed as:

Precision for class C = (True Positive for Class C)/(True Positive for Class C+False Positive for Class C)

Recall : Recall for a given class is defined as the fraction of instances in that class that the model correctly classified out of all instances belonging to that class. The formula for recall (R) for a class "C" can be expressed as:

Recall for class C = (True Positive for Class C)/(True Positive for Class C+False Negative for Class C)

F1-score: This is a harmonic mean between Precision and Recall, combining their importance into a single metric. It considers both false positives and false negatives, aiming for a balance between them.

Support: Number of actual occurrences of a particular class in the dataset.

	Precision	Recall	F1-score	Support
Class 0	0.938	0.943	0.940	5830
Class 1	0.942	0.941	0.941	265
Class 2	0.933	0.936	0.934	231
Class 3	0.944	0.945	0.944	174
Class 4	0.939	0.937	0.938	207
Accuracy			0.961	6707
Macro avg	0.939	0.940	0.939	6707
Weighted avg	0.962	0.962	0.962	6707

Table-2: Evaluation metrics for each class

Specificity: Specificity in multiclass classification is the ability of a model to correctly identify instances that do not belong to a particular class, out of all instances that do not belong to that class. Specificity gives information about a model's capacity to avoid misclassifying instances as belonging to a particular class when they do not. The formula for specificity for a class "C" is given by:

Specificity for class C= (True Negative for Class C)/(True Negative for Class C+False Positive for Class C)

Sensitivity: Sensitivity in multiclass classification is the ability of a model to correctly identify instances of a specific class out of all instances belonging to that class. Sensitivity provides insights into how well a model is capturing instances of a particular class among all the instances that actually belong to that class. The formula for sensitivity for a particular class "C" is given by:

Sensitivity for class C= (True Positive for Class C)/(True Positive for Class C+False Negative for Class C)

Class	Specificity	Sensitivity
0	0.941847	0.995540
1	0.992859	0.664151
2	0.993669	0.809524
3	0.987601	0.643678
4	0.990154	0.700483

Table- 3: Specificity and Sensitivity for each class

Confusion Matrix: The confusion Matrix gives a comparison between actual and predicted values. It is used for the optimization of machine learning models. The confusion matrix is a N x N matrix, where N is the number of classes or outputs.

Fig-3: Confusion Matrix

AUC-ROC curve: AUC-ROC ranges from 0 to 1, where 0.5 indicates random classification, and 1.0 indicates perfect classification. A higher AUC-ROC score suggests better discrimination and ordering of classes based on their predicted probabilities. AUC-ROC can be useful for assessing the model's ability to distinguish between different classes, even in a multiclass setting.

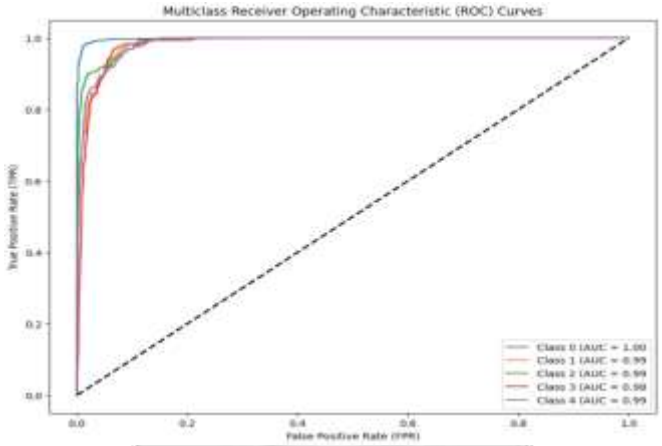

Fig-4 : AUC-ROC plot

7. REFERENCES

[1]. The Digital Database for Screening Mammography, Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore and W. Philip Kegelmeyer, in Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218, Medical Physics Publishing, 2001. ISBN 1-930524-00-5.

[2]. Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi , Daniel Rubin (2016). Curated Breast Imaging Subset of DDSM. The Cancer Imaging Archive.

[3]. Mohiyuddin Aqsa and Basharat Asma and Ghani Usman and Abbas Sidra and Naeem Osama Bin et al., Breast tumor detection and classification in mammogram images using modified YOLOv5 net_x0002_work, Computational and Mathematical Methods in Medicine, (2022), Hindawi. https://doi.org/10.1155/

[4]. Aboutalib S.S., Mohamed A.A., Berg W.A., Zuley M.L., Sumkin J.H. and Wu S., 2018. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clinical Cancer Research, 24(23), pp.5902–5909. https://doi.org/10.1158/1078-0432.CCR-18-1115 PMID: 30309858

[5]. Chun-ming T.A.N.G., Xiao-mei C.U.I., Xiang Y.U. and Fan Y.A.N.G., 2019. Five Classification of Mam_x0002_mography Images Based on Deep Cooperation Convolutional Neural Network. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 57(1), pp.10–21. https://www.asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/4942

[6]. Le´vy, D. and Jain, A., 2016. Breast mass classification from mammograms using deep convolutional neural networks. arXiv preprint arXiv:1612.00542. https://arxiv.org/abs/1612.00542

[7]. Mingxing Tan, Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. https://doi.org/10.48550/arXiv.1905.11946 or https://arxiv.org/abs/1905.11946v5

[8]. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. https://doi.org/10.48550/arXiv.1512.00567 or https://arxiv.org/abs/1512.00567

[9]. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. https://doi.org/10.48550/arXiv.1801.04381 or https://arxiv.org/abs/1801.04381v4