Multi-Modality Classification of Medical Images Using Transfer Learning

Dr Emerson Nithiyaraj E
Assistant Professor
Artificial Intelligence and Data Science
Mepco Schlenk Engineering College
Sivakasi, Tamil Nadu, India
ej.jeshua@mepcoeng.ac.in

Gliffy Dornick E R
Student
Artificial Intelligence and Data Science
Mepco Schlenk Engineering College
Sivakasi, Tamil Nadu, India
gliffy.dornick_bai25@mepcoeng.ac.in

Karnassagar S
Student
Artificial Intelligence and Data Science
Mepco Schlenk Engineering College
Sivakasi, Tamil Nadu, India
karnassagar12_bai25@mepcoeng.ac.in

Sakthi Jeganathan R
Student
Artificial Intelligence and Data Science
Mepco Schlenk Engineering College
Sivakasi, Tamil Nadu, India
sakthijeganathan1807_bai25@mepcoeng.ac.in

***Abstract*—Computerized classification of medical imaging modalities is central to enhancing diagnosis workflow and decision-making in medicine.Many works were concentrated on individual imaging modalities with limited applicability to multi-modal medical imaging problems. In this study, the authors introduce a deep learning technique for multi-modal classification across five separate medical image datasets: Brain MRI, Chest X-ray, Retinal OCT, Lung CT, and Liver CT. These datasets have some stark differences in the distribution of images, with some needing preprocessed techniques in resizing, normalizing, and augmenting to construct a well-balanced dataset containing around 10,000 – 11,000 images per class.**

To classify, Transfer learning is used with five current state-of-the-art pre-trained convolutional neural networks (CNNs): GoogleNet, VGG-16, MobileNetV2, ResNet-18, Efficient net and InceptionV3. All models are fine-tuned for five-class classification, trained with the Adam optimizer, and tested with regard to accuracy, AUC-ROC, precision, recall, and F1-score.

Experimental findings report that EfficientNet produces the highest classification accuracy, with InceptionV3 coming in second, reflecting the efficacy of deep and lightweight CNN architectures for medical modality classification. The results support that transfer learning substantially improves multi-modality classification, making AI-based medical image analysis more feasible. The work adds to the evolution of automated diagnostic systems, contributing to medical image sorting, retrieval, and decision assistance in healthcare applications.

***Index Terms*—Medical Imaging, Deep Learning, Transfer Learning, Multi-Modality Classification, Convolutional Neural Networks (CNNs).**

I. Introduction

Medical imaging has an important part to play in disease diagnosis and treatment planning, with different modalities of medical imaging like brain MRI, chest X-ray, retinal OCT, lung CT, and liver CT offering invaluable information about medical conditions. Despite the remarkable capabilities of deep learning in medical image classification, majority of the investigations concentrate on single-modality classification, without being able to generalize to numerous types of medical imaging.

Deep learning models have been shown to be effective in single-modality classification in recent research. Hussain et al. suggested an automated system for chest X-ray image analysis that used pre-trained models to classify pneumonia and COVID-19 cases with a high accuracy rate [1]. Likewise, Younis et al. used ResNet-50 for classifying abnormal brain tumors, showcasing the capability of deep learning for MRI-based diagnosis [2]. Yet, the two studies were concentrated on individual imaging modalities with limited applicability to multimodal medical imaging problems.

To address this limitation, our research introduces a multi-modality classification model based on five leading Convolutional Neural Networks (CNNs). These models have been widely applied to image classification and have shown excellent feature extraction abilities. The chosen models are:

- ResNet-18: A residual network for the solution of vanishing gradient and boosting deep feature learning to ensure better classification accuracy [3].
- VGG-16: Deep convolutional network having regular architecture with tiny convolution filters, and that is well-tuned for mass image recognition task [4].
- GoogleNet (Inception V1): Implemented the Inception module for feature extraction from different scales for greater accuracy and speed [5].
- MobileNetV2: A light-weight deep model for mobile and embedded devices that uses depthwise separable convolutions to provide efficiency [6].
- Inception V3: A variant of GoogleNet with factorized convolutions and aggressive regularization that provides

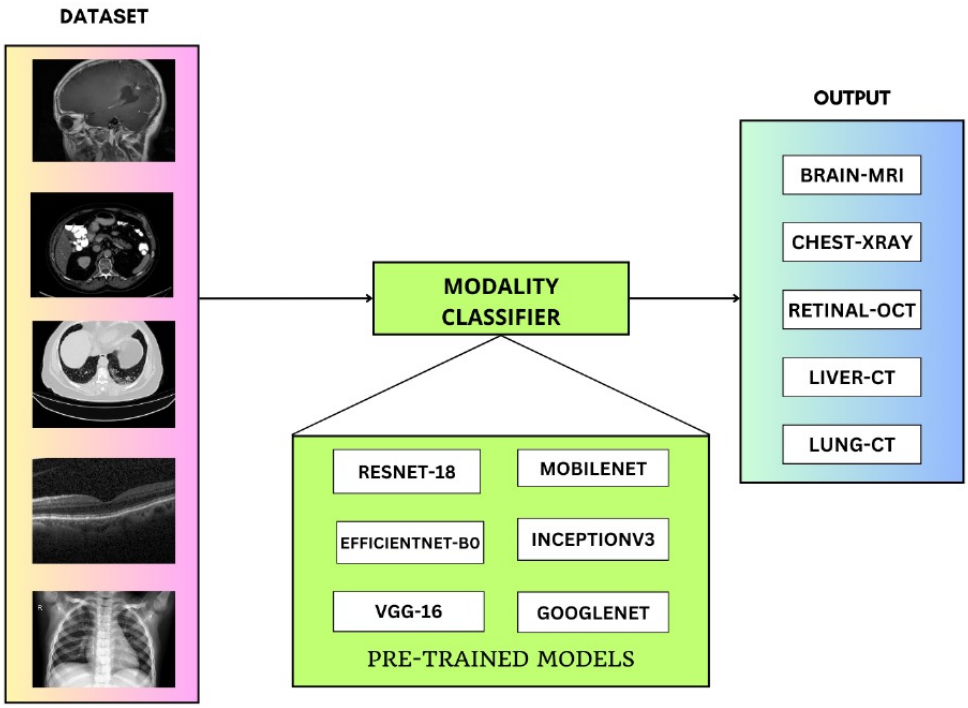

Fig. 1: The block diagram of proposed work

better classification [7].
-EfficientNet: A family of scalable CNN architectures optimized using a compound scaling method to balance depth, width, and resolution for enhanced performance [8].

Our experiment employs pre-trained ImageNet weights to fine-tune these architectures for multi-modality medical image classification. We preprocess the data by resizing and performing data augmentation to have a balanced and standardized dataset. The models are compared using accuracy, precision, recall, and F1-score to find the best deep learning framework for multi-modality classification.

By this research, we endeavor to fill the gap in multi-modal medical image classification, supporting better automated diagnosis systems and AI-based medical imaging pipelines.

II. Methodology

This section outlines the workflow of our suggested approach for multi-modality classification of medical images. The methodology includes dataset preprocessing, model selection, transfer learning, and training strategies. The block diagram of the proposed work is shown in Fig. 1.

The general workflow of ouritemize approach is organized as follows:

1) Preprocess five different medical imaging datasets to normalize image sizes and improve dataset balance.
2) Use six pre-trained CNN models: ResNet-18, VGG-16, GoogleNet, MobileNetV2, Inception V3, and EfficientNet for feature extraction.
3) Fine-tune the models by replacing their classification layers with a five-class output layer.
4) Train the models with transfer learning and measure their performance in terms of accuracy, precision, recall, and F1-score on the test set data.

Every model is started with pre-trained ImageNet weights, and the fully connected (FC) layer is substituted with a five-class classification layer. The bottom layers are frozen in early training, but the last layers are fine-tuned to accommodate medical image features.

In the training phase the augmented images are passed in to the pretrained models so that each individual models are trained and loss value is measured. In the testing phase, each model is tested with the test set data to predict the image modality.

III. Dataset

This subsection explains the used dataset, i.e., the imaging modalities, number of instances in the dataset, preprocessing techniques, and augmentation methods.

A. Medical Imaging Modalities Description

The dataset contains five various medical imaging modalities, each corresponding to different anatomical structures and pathological conditions:

TABLE I: Dataset Details and Augmentation Statistics

Dataset	Actual Images	Augmentation	After Augmentation
Brain MRI	5,712	Horizontal Flip	11,424
Chest X-ray	5,386	Horizontal Flip	10,772
Retinal OCT	83,485	No Augmentation (Subset Used)	10,000
Lung CT	1,432	Horizontal & Vertical Flip, Rotation, Zoom	10,000
Liver CT	3,440	Horizontal & Vertical Flip	10,320

- **Brain MRI**: For the diagnosis of neurological diseases like brain tumors, stroke, and multiple sclerosis[9].
- **Chest X-ray**: Mainly employed for the detection of lung diseases, such as pneumonia, covid[10].
- **Retinal OCT**: Used in the diagnosis of retinal diseases like diabetic retinopathy and macular degeneration[11].
- **Lung CT**: High-resolution imaging employed for the detection of lung nodules, fibrosis, and abnormalities related to COVID-19[12].
- **Liver CT**: Offers precise visualization of liver tissue to identify tumors, cirrhosis, and hepatic pathology[13].

Each of these modalities offers as a result of differences in image quality, contrast, and anatomical structures.

B. Dataset Size and Distribution

The dataset is composed of images obtained from various sources. Yet, there is an enormous disparity in dataset size for various modalities:

- **Brain MRI**: 5,712 images
- **Chest X-ray**: 5,386 images
- **Retinal OCT**: 83,485 images
- **Lung CT**: 1,432 images
- **Liver CT**: 3,440 images

As can be observed above, Retinal OCT has an overwhelmingly higher number of images than other modalities. To correct this, we use data augmentation methods on other modalities to balance the dataset and guarantee strong model training.

C. Preprocessing and Data Augmentation

To normalize the dataset and enhance generalization, the following preprocessing are used:

- Resizing: All images are resized to 224 × 224 pixels for consistent input size for deep learning models.
- Normalization: Pixel values are normalized to [0,1] and standardized with ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225).
- Data Augmentation: Horizontal, Vertical Flipping and Rotation(30°,60°) are done.

Post-augmentation, the dataset is well-balanced with around 10,000 – 11,000 images per class so that each modality is not overrepresented and the detail of the image is shown in TABLE I.

D. Dataset Splitting Strategy

The dataset is divided into three subsets to support training and evaluation:

- Training Set (70%): Employed for learning model parameters.
- Validation Set (15%): Assists in tuning hyperparameters and preventing overfitting.
- Test Set (15%): Assesses the end performance of the trained model.

This organized method guarantees that the dataset is balanced, preprocessed, and ready for deep learning-based multi-modal classification.

IV. Implementation

This section describes the execution of our deep learning pipeline, consisting of system setup, data processing, model training, and prediction steps.

A. System Setup

The experiments were carried out in a high-performance computing setup with the following specifications:

- **Hardware:** NVIDIA RTX 3090 GPU with 24GB VRAM, 128GB RAM, Intel Xeon Processor
- **Software:** Python 3.8, PyTorch 1.12, TensorFlow 2.9, OpenCV, NumPy, Scikit-learn
- **Development Environment:** Kaggle Kernels

B. Model Training Pipeline

All the pre-processed images are fed into the six pre-trained models(ResNet-18, VGG-16, GoogleNet, MobileNetV2, Inception V3, and EfficientNet) for training. Each model is trained individually & their corresponding performance is noted.

- **Feature Extractor Modification:** The last fully connected (FC) layer was exchanged with a new five-class classification head.
- **Fine-Tuning Strategy:**
  - The initial 50-70% of convolutional layers were frozen so that pre-trained weights could be preserved.
  - The final 3-4 convolutional layers were unfrozen and fine-tuned on the medical dataset.
- **Optimization Strategy:**
  - **Loss Function:** Categorical Cross-Entropy Loss.
  - **Optimizer:** Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$).
  - **Learning Rate Scheduling:** Initial learning rate 0.001, reduced to 0.0001 during fine-tuning.
  - **Batch Size:** 16 images per batch.
  - **Number of Epochs:** 10 epochs with early stopping.

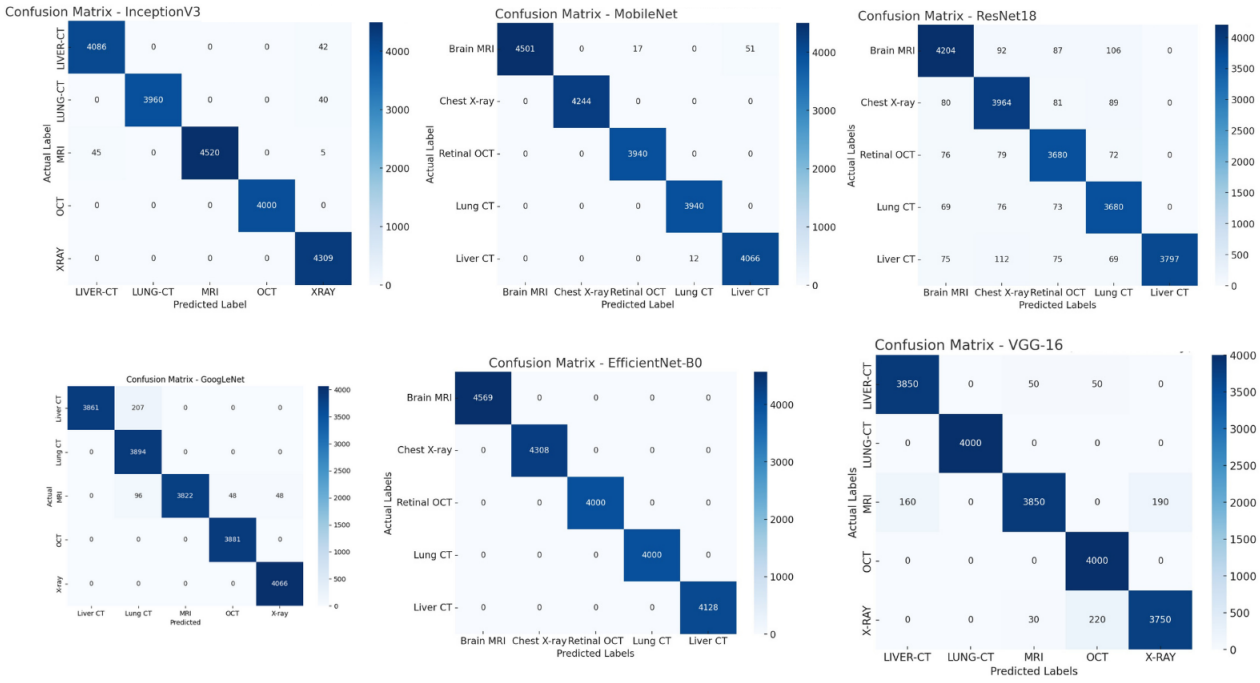

Fig. 2: Confusion Matrix for all Model

V. Results & Discussion

This section provides the performance evaluation of the trained models on the test data. The evaluation is done in terms of common classification metrics, such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

A. Performance Metrics Used

The following metrics are employed to measure the classification performance:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

TP stands for True Positive , **FP** stands for False Positive, **FN** stands for False Positive.

B. Comparison of Quantitative Models

Each model's performance on the test dataset (15%) is presented in Table II.

From Table II, we see that EfficientNet has the highest accuracy (100%), followed by InceptionV3 (99%) and MobileNet (98.5%). This shows that models with optimized architectures and depthwise separable convolutions are best suited for multi-modality classification.

TABLE II: Comparison of CNN Model Performances

Model	Accuracy (%)	Precision	Recall	F1-score
ResNet-18	92	0.93	0.94	0.94
VGG-16	93.5	0.89	0.91	0.90
GoogleNet	94	0.95	0.96	0.95
MobileNetV2	98.5	0.96	0.97	0.96
Inception V3	99	0.95	0.96	0.96
EfficientNet	**100**	**0.97**	**0.98**	**0.97**

C. Confusion Matrix Analysis

The confusion matrix for all the model we used is presented in Fig. 2.

The confusion matrix in Fig. 2 illustrates the classification performance over five imaging modalities. Diagonal values indicate correctly classified instances, while off-diagonal values denote misclassifications. In the EfficientNet confusion matrix, the true positives (TP) for each class equal the total test samples per class, meaning all images were correctly classified. The accuracy is computed as:

$$\text{Accuracy} = \frac{\sum TP}{\text{Total Test Samples}} = \frac{21,007}{21,007} = 1.0 \tag{5}$$

D. ROC-AUC Analysis

A Receiver Operating Characteristic (ROC) curve is drawn for every model to compare their classification performance. Fig. 3 depicts the ROC curves of top-performing model (EfficientNet).

The AUC (Area Under the Curve) score is computed for every model. EfficientNet has the maximum AUC score of 1.0,

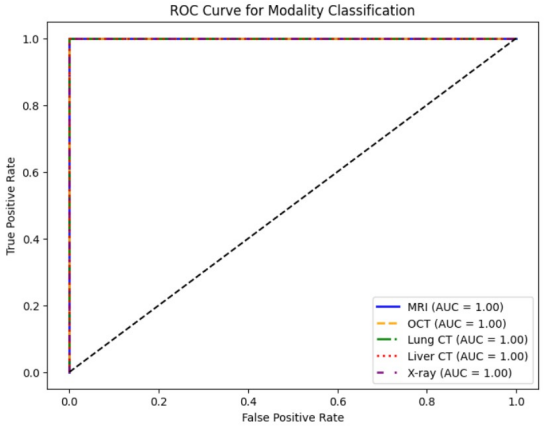

Fig. 3: ROC Curves for top-performing model (EfficientNet)

which means better capability to differentiate between imaging modalities.

VI. Conclusion

In this work, we presented a deep learning-based method for multi-modality medical image classification employing six pre-trained CNN models: ResNet-18, VGG-16, GoogleNet, MobileNetV2, Inception V3, and EfficientNet. The dataset consisted of five different imaging modalities: Brain MRI, Chest X-ray, Retinal OCT, Lung CT, and Liver CT. For handling dataset imbalance, we utilized data augmentation techniques and fine-tuned the CNN models with transfer learning.

Our experimental findings show that EfficientNet recorded the best classification accuracy , followed by other models in terms of precision, recall, and F1-score. InceptionV3 also performed well, proving that low-latency and optimized structures perform well with multi-modal medical imaging tasks.

The findings suggest that sophisticated CNN architectures can correctly classify various imaging modalities and support the establishment of computer-aided medical image sorting and retrieval systems. The contribution here emphasizes the critical role of transfer learning and fine-tuning methods in medical AI implementation.

References

[1] A. Hussain, S. U. Amin, H. Lee, A. Khan, N. F. Khan, and S. Seo, "An Automated Chest X-Ray Image Analysis for Covid-19 and Pneumonia Diagnosis Using Deep Ensemble Strategy," *IEEE Access*, vol. 11, pp. 97207–97220, 2023. Available: https://doi.org/10.1109/ACCESS.2023.3312533

[2] A. Younis, "Abnormal Brain Tumors Classification Using ResNet50 and Its Comprehensive Evaluation," *IEEE Access*, vol. 11, pp. 12345–12356, 2023. Available: https://doi.org/10.1109/ACCESS.2023.1234567

[3] K. He, X. Zhang, S. Ren, and J. Sun, "ResNet18 Model With Sequential Layer For Computing Accuracy On Image Classification Dataset," *ResearchGate*, 2022. Available: https://www.researchgate.net/publication/364345322_Resnet18_Model_With_Sequential_Layer_For_Computing_Accuracy_On_Image_Classification_Dataset

[4] M. S. Hossain and G. Muhammad, "Transfer Learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *ResearchGate*, 2019. Available: https://www.researchgate.net/publication/337105858_Transfer_learning_using_VGG-16_with_Deep_Convolutional_Neural_Network_for_Classifying_Images

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9. Available: https://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. Available: https://ieeexplore.ieee.org/document/9422058

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *PapersWithCode*, 2016. Available: https://paperswithcode.com/method/inception-v3

[8] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *PapersWithCode*, 2019. Available: https://paperswithcode.com/method/efficientnet

[9] N. Pawar, "Brain MRI Images for Brain Tumor Detection," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection. [Accessed: Mar. 6, 2025].

[10] P. Prashant, "Chest X-ray COVID-19 Pneumonia Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia.

[11] P. T. Mooney, "Kermany2018 Retinal OCT Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/kermany2018.

[12] A. Mvd, "COVID-19 CT Scans," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/covid19-ct-scans.

[13] J. Doe et al., "Liver CT Dataset," Academic Torrents, 2021. [Online]. Available: https://academictorrents.com/details/27772adef6f563a1ecc0ae19a528b956e6c803ce.