Classification of COVID 19, Pneumonia and healthy lung on CXR images using Deep Transfer Learning Ensemble Framework with GRAD-CAM visualization

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—COVID-19 and pneumonia both affect the human respiratory system and can cause symptoms ranging from mild respiratory issues to severe conditions. Radiological imaging techniques such as X-ray is really effective to diagnose these diseases. Though these patterns can overlap on a chest X-ray, pneumonia usually appears as a dense area in one part of the lung, while COVID-19 often shows up as hazy, cloud-like areas and scattered spots in both lungs. In this paper, we worked with a customized dataset which contains 21,000 labeled chest X-ray images and for classification part several pre-trained networks such as MobileNet, ResNet (50 and 152 layers) and ViT were implemented. We used Gradient-weighted Class Activation Mapping (Grad-CAM) before fully connected layer to generate heat maps that show the most focused region of the image the model considers for its prediction. The results of the Ensemble framework of these transfer learning approaches outperformed most state-of-the-art models with an accuracy of 0.9956 which is reliable and robust for classifying thoracic diseases from chest X-ray images.**

***Index Terms*—Pneumonia and Covid-19, Pre-trained networks, Chest X-ray images, Grad-CAM**

I. Introduction

Pneumonia is a lung infection that can be serious and sometimes can be potentially life-threatening conditions specially for young children.According to the World Health Organization, 2.5 million pneumonia-related deaths were reported in 2019, with children between the ages of 0 and 5 making up 14% of these deaths. For the elderly, and individuals with weakened immune systems. The risks associated with pneumonia include severe respiratory distress, in some cases, long-term damage to the lungs[1].The mortality rate for pneumonia varies widely based on factors such as the patient's age, overall health, the type of pneumonia, and the quality and timeliness of medical care.

COVID-19, caused by the SARS-CoV-2 virus, primarily affects the lungs, leading to inflammation, fluid buildup, and lung tissue damage. SARS-CoV-2, a new strain of the coronavirus, poses a serious threat to global health and has resulted in significant loss of life. First identified in Wuhan, China, in 2019, COVID-19 had infected over 507 million people worldwide by April 2022, with a death toll surpassing 6 million. Countries like the United States, India, Brazil, and several European nations have been hit hard, due to factors like healthcare systems, public health measures, population density, and the emergence of more contagious variants [2].

Pneumonia, which can be caused by bacteria, viruses, or fungi, also leads to inflammation and fluid accumulation in the lungs' air sacs. In both conditions, the inflammation and fluid interfere with normal breathing and oxygen exchange, leading to respiratory symptoms like coughing, shortness of breath, and chest pain. However, in the early stages of COVID-19, symptoms usually include fever, cough but in severe cases symptoms can intensify, leading to shortness of breath. Ground-glass opacities (GGO) are common in COVID-19 but less frequent in regular pneumonia. In COVID-19, these spots usually appear on both sides of the lungs and near the edges, while in pneumonia, they are often found in one specific area or lobe of the lung. Another difference is that pneumonia tends to show thicker, more solid lung areas, while COVID-19 has a more scattered and spread-out appearance.

To detect COVID-19, PCR testing identifies the virus's genetic material from a sample, usually collected via a nasal or throat swab, by replicating the virus's RNA and transforming it into DNA. However, the process can be time-consuming, and some studies have shown a sensitivity of around 90.7% [3]. On the other hand detecting lung infection through X-ray images involves identifying specific patterns in the lungs that are indicative for both pneumonia and Covid-19. Deep learning models help detect pneumonia and COVID-19 from chest X-rays by learning to recognize patterns specific to each condition. These algorithms have gained immense popularity due to their ability to learn and extract intricate features from raw image data automatically. They are trained on labeled images, extract key features, and classify new X-rays as normal, pneumonia, or COVID-19. Radiologists and AI-based tools can assist in diagnosing these conditions faster from chest

X-rays.

II. Related Study

In the paper [4], Gayathri J.L. et al. used a variety of pre-trained architectures for classification and sparse encoders for feature selection when working with chest x-ray images to detect COVID-19. The combination of InceptionResnetV2 and Xception yielded the best results, with an accuracy of 0.9578 and an AUC of 0.9821. But due to insufficient number of images in the experiment it can easily fall into overfit, also didn't explore multi-class classification. In order to diagnose between COVID-19 and pneumonia from image data, Linh T. Duong et al. [5] experimented with lung CT and chest x-ray images. Differnt versions of Efficient and MixNet were employed but EfficientNet-B0 (Acc. 96.64%) and EfficientNet-B3 (Acc. 95.82%) outperformsfor other networks on two different X-ray datasets. Gaffari Celik examined on several transfer learning models by keeping certain parameters constant on the same dataset [6] where the proposed model combination of CovidDWNet+GB(Gradient Boosting) achieved 96.81% accuracy on chest X-ray images. Xingsi Xue et. al. [7] investigates various deep learning techniques, including ResNet152, VGG16, ResNet50, and DenseNet121, for detecting COVID-19, Pneumonia and Normal from CT and radiography images. An enhanced VGG16 has achieved 99% accuracy and average F1-score 95%, to recognize three classes of radiographic images. Using Xception and Visual Geometry Group (16 & 19), Deepak Kumar Jain [8] et al. conducted an experiment to classify regular, pneumonia, and normal X-ray pictures with an accuracy of 98%, which was obtained by the Xception model. Ameer Hamza et al. utilized a technique based on ancanonical correlation analysis (ICCA) to fuse the selected featurers after the hyperparameters were established through Bayesian optimization to categorize COVID-19 and pneumonia from MRI scan and X-ray images. Following additional tuning with the tree growth optimization algorithm, classification was accomplished with MobileNet (MNN) from the X-ray dataset with an accuracy of 99.6% utilizing ResNet50, InceptionV3, and MNN. For multi-class classification, Hassaan Malik [10] et al. introduced a new CNN network called Chest Disease Classification (CDC), which combines dilated convolution and residual network concepts. It outperforms other CNN models (ResNet50, VGG19, and InceptionV3) and obtained an AUC of 99.53% accuracy for five class classification on twelve distinct datasets related to chest diseases.

III. Methodology

A. Dataset and Experiment

Our dataset was generated by combining a multitude of publicly accessible datasets and repositories, each of which is dispersed and has a different format. A stringent quality control procedure guaranteed the dataset's quality by identifying and eliminating duplicates, images of extremely low quality, and overexposed images. X-rays have gradually become more accessible to the public.Details of different data sources are given below:

TABLE I
Details of Dataset 1 and Dataset 2

Dataset	Class	CXR/ Class	Dataset splitting		
			Train	Validation	Test
Dataset 1	Covid 19	7000	4900	1050	1050
	Pneumonia	7000	4900	1050	1050
	Normal	7000	4900	1050	1050
Dataset 2	Covid 19	1626	1138	244	244
	Pneumonia	1800	1249	271	270
	Normal	1802	1261	271	270

- COVID-19 Radiography Database [3.1, 3.2]: They have released 219 COVID-19, 1341 normal, and 1345 viral pneumonia chest X-ray (CXR) images in the initial release. The latest database expanded the images includes 3616 COVID-19 positive cases, 10,192 normal cases, 6012 lung opacity (Non-COVID lung infection), and 1345 viral pneumonia images and corresponding lung masks.
- Chest X-ray (Pneumonia & COVID-19) [3.3]: The chest X-ray images for this dataset (anterior-posterior) were selected from retrospective cohorts of pediatric patients aged one to five years at the Guangzhou Women and Children's Medical Center in Guangzhou. The data are divided into three folders (train, test, value) and includes subfolders for each image category (Pneumonia/Normal). There are 5,863 X-ray images (JPEG) and two categories (Normal/Pneumonia).
- 15K Chest X-Ray Images (COVID-19) dataset [3.4]: The dataset contains train and test images. There are 200 images each for testing Covid 19 and normal. And the training directory contains 2158 images for covid 19 and 13.8k images for normal.
- COVID-19+PNEUMONIA+NORMAL Chest X-Ray Image Dataset [3.5, 3.6]: The dataset is a medical image directory structure divided into three subfolders (COVID, NORMAL, and PNEUMONIA). It contains chest X-ray (CXR) images where COVID-19: 1626 images, NORMAl: 1802 images and 1800 images of PNEUMONIA.

Therefore, we modified those datasets to create COVID-PNEUMOINA-MRI-21k chest X-ray [3.7], which comprises more than 21,000 CXR images from three distinct classes.

Fig. 1. Sample chest X-ray images from the Pneumonia_Covid_21k dataset for each class

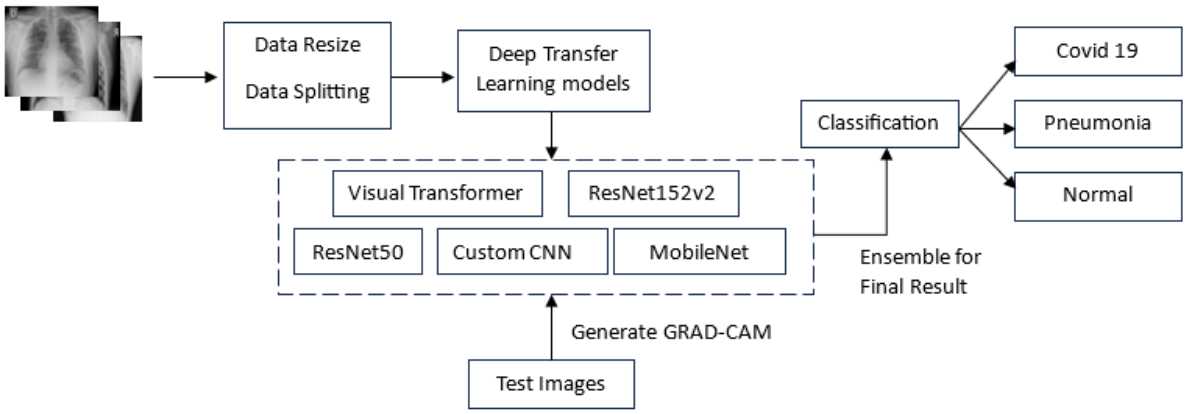

Fig. 2. Block diagram of the multi-class classification of proposed ensemble framework

B. Training Models

1) Deep Transfer Learning Models:

- MobileNetV2: The design of the MobileNetV2, a lightweight convolutional neural network (CNN), includes important features that make it both efficient and effective for image classification. In MobileNet V2, each block includes a 1x1 expansion layer along with depthwise and pointwise convolution layers. The expansion layer increases the number of channels based on a factor before sending the data to the depthwise convolution. Each layer uses batch normalization and ReLU as the activation function, but the output of the projection layer doesn't have an activation function. This version also includes a residual connection, and the full MobileNet V2 architecture has 17 bottleneck residual blocks [11]. The depthwise separable convolutions split the convolution into two distinct operations: depthwise convolution and pointwise convolution and reduce the architecture cost.

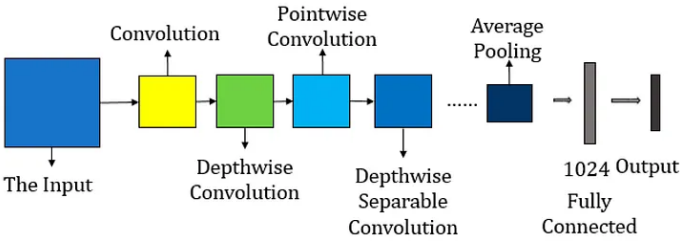

Fig. 3. MobileNet Architecture

- Residual Neural Network: In deeper CNN layers, a common issue called the vanishing gradient occurs. ResNet networks solve this by using residual connections. These connections act as shortcuts, letting information skip over some layers and go directly to the output. The key idea is using identity mapping, where the network learns to fit the residuals by jumping over layers using shortcut connections[12]. This approach helps ensure that adding more layers doesn't hurt performance, allowing for much deeper networks like ResNet, which can go up to 152 layers. In terms of depth, ResNet50 has 50 layers, while ResNet152 has 152 layers. ResNet50 has 16 blocks, whereas ResNet152 contains 50 blocks. From Fig. 4, during the first few training steps, the model skips the layers with 512 filters and passes through the X connection. When needed, it uses the layers with 512 filters to capture more complex features. It then combines the data from the X connection and the weighted layer FX.

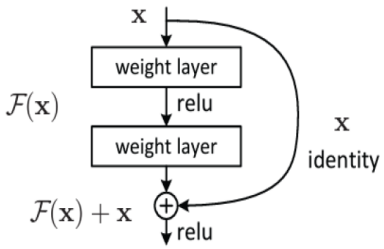

Fig. 4. Residual connection [12]

$$H(x) = F(x) + x \tag{1}$$

Where, x is the input to the set of layers, $F(x)$ is the residual function, and $H(x)$ is the mapping function from input to output.

2) ViT: The Vision Transformer (ViT) applies the principles of transformer models, originally used in natural language processing, to image classification. The Vision Transformer (ViT) uses ideas from transformer models, which were first used for language tasks, to classify images. It works by breaking an image into small square patches, turning them into flat data, and organizing them into a sequence. This sequence is then fed into the transformer, where attention mechanisms help the model understand how different patches of the image are related. After going through several layers, the output helps

classify the image, with one final piece of data summarizing the whole image for the classifier to predict its category. ViT is powerful because it can focus on both the small details and the bigger picture of the image.

C. Hyperparameters

In Convolutional Neural Networks (CNNs), key settings, or hyperparameters, include things like the number of convolutional layers, the size of the filters, stride, padding, activation functions (such as ReLU), learning rate, batch size, number of training rounds (epochs), and dropout rate. These settings are important because they influence how well the model learns and performs. The learning rate decides how fast the model adjusts its internal parameters, and dropout helps prevent overfitting by turning off some neurons randomly during training. In table 2 the hyperparameters are listed which are worked in the experimented models.

TABLE II
MODELS PARAMETERS FOR INPUT AND CLASSIFICATION STAGE

Parameters	Approch	
	CNN Models	Vision Transformer
Input shape	100 × 100	72 × 72
No. of epochs	100	100
Batch Sizes	32	16
Activation Function	Softmax	Softmax
Learning rate	0.0001	0.0001
patch_size	-	3
num_patches	-	(72 / 3) × 2
transformer_layers	-	8

D. Experimental Setup

The hardware for this experiment comprises 8GB Nvidia GEFORCE RTX 4060 and 16 GB RAM. We built the pre-trained networks using NumPy, the TensorFlow framework, Sklearn, the Matplotlib graph charting tool, the Seaborn data visualization tool, Python 3.10 with the TensorFlow framework.

E. Evaluation Criteria

Accuracy is needed to measure generalperformance of a network. It is easily interpretable and widely used metric for model evaluation.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2}$$

Precision evaluates how accurately the model identifies positive cases. It refers to the ratio of correct positive predictions to the total number of positive predictions made by the model.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3}$$

While precision shows how good the model is at making correct positive predictions, recall calculates how many actual positive cases the model was able to find.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4}$$

IV. Result and Discussion

GRAD-CAM (Gradient-weighted Class Activation Mapping) is a method used in image processing to visualize which areas of an image a Convolutional Neural Network (CNN) pays attention to when making predictions. By calculating a weighted combination of the feature maps from earlier layers, GRAD-CAM produces a heatmap that highlights the key regions of the image that influence the model's decision.

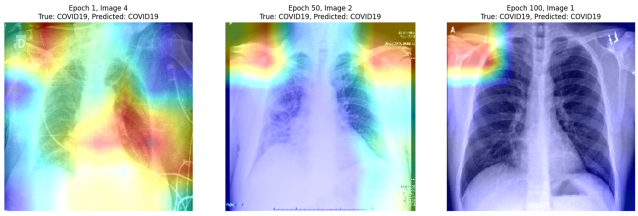

Fig. 5. **********

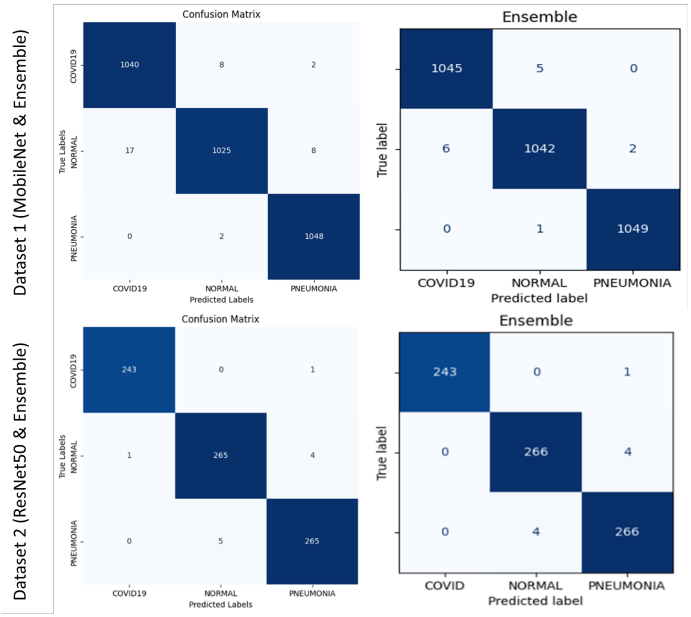

Fig. 6. Confution Matrix Best Model and Ensemble

V. Conclusion

Early detection is helpful to prevent spreading Covid-19 and seriousness in pneumonia. Automatic AI based networks and real-time detection is getting popular but the main limitation is the limited number of labeled images. To solve the data imbalance, we build a customized balanced dataset from four different sources. The experiment result shows the ensemble of multiple pre-trained and CNN networks has been implemented which outperforms single pre-trained or ViT networks. This experiment implemented different models with both balanced and imbalanced datasets. From the analysis it is shown that the performances of different architectures are examined by keeping certain parameters constant on the two different datasets

TABLE III
COMPARISON OF MODEL PERFORMANCE ON DATASET 1 AND DATASET 2

Model	Class	Dataset 1				Dataset 2			
		Precision	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Mobile Net	Covid-19	0.99	0.99	0.99	0.9914	1.00	0.99	1.00	0.9758
	Pneumonia	0.99	1.00	1.00		0.99	0.94	0.97	
	Normal	0.99	0.98	0.99		0.94	1.00	0.97	
ResNet 50	Covid-19	0.98	0.99	0.99	0.9883	1.00	1.00	1.00	0.9860
	Pneumonia	0.99	1.00	0.99		0.98	0.98	0.98	
	Normal	0.99	0.98	0.98		0.98	0.98	0.98	
ResNet 152	Covid-19	0.99	0.98	0.98	0.9876	1.00	1.00	1.00	0.9783
	Pneumonia	1.00	1.00	1.00		0.97	0.97	0.97	
	Normal	0.98	0.98	0.98		0.97	0.97	0.97	
Custom CNN	Covid-19	0.95	0.96	0.96	0.9638	0.98	0.99	0.99	0.9630
	Pneumonia	0.99	0.99	0.99		0.96	0.94	0.95	
	Normal	0.95	0.95	0.95		0.95	0.96	0.96	
Ensemble (Except ViT)	Covid-19	0.9943	0.9952	0.9948	0.9956	1.0000	0.9959	0.9979	0.9885
	Pneumonia	0.9981	0.9990	0.9986		0.9815	0.9852	0.9834	
	Normal	0.9943	0.9924	0.9933		0.9852	0.9852	0.9852	
Visual Transform (ViT)	Covid-19	0.94	0.95	0.94	0.9524	0.94	0.95	0.94	0.9640
	Pneumonia	0.98	0.99	0.98		0.98	0.99	0.98	
	Normal	0.94	0.92	0.93		0.94	0.92	0.93	

TABLE IV
COMPARISON OF THE PROPOSED ENSEMBLE MODEL WITH OTHER DEEP LEARNING TECHNIQUES USING CHEST X-RAY IMAGES

Reference	No. of images			Scanning	Total Class	Used Model	Acc	Precision	Recall	F1 Score
	Covid-19	**Pneumonia**	**Normal**							
[8]	504	504 (non-Covid 19)		X-ray	3	IceptionNet Resnet V2 Xception	0.9578	0.9563	0.9563	0.9563
[8]	108	6041	8851	X-ray	3	Efficient B3	0.9664	0.968	0.978	0.973
	327	7386	10192			Efficient B0	0.9582	0.968	0.958	0.950
[8]	3616	1345	10192	X-ray	4	COVIDWNET GB	0.9681	0.980	0.970	0.980
[8]	1281	1300	1481	X-ray	3	Xception	0.980	0.990	0.930	0.960
[8]	2371	3867	2749	X-ray	5	CDC_Net	0.9939	0.9942	0.9813	0.9826
This Paper	7000	7000	7000	X-ray	3	Ensemble	0.9956	-	-	-

and the best result came from the implemented ensemble with custom-made COVID-PNEUMOINA-CXR-21k chest X-ray dataset. .

Early detection plays a crucial role in preventing the spread of COVID-19 and the severity of pneumonia. AI-based automatic networks and real-time detection systems are becoming more common, but they face a significant challenge due to the limited number of labeled images available. To address this data imbalance, we created a custom, balanced dataset from four different sources. The results of the experiment show that an ensemble of multiple pre-trained models and CNN networks performed better than using a single pre-trained model or ViT networks. Various models were tested on both balanced and imbalanced datasets, and the analysis reveals that the best performance came from the ensemble using the custom dataset COVID-PNEUMOINA-CXR-21k chest X-ray, with certain parameters kept constant during the comparison.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[9] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

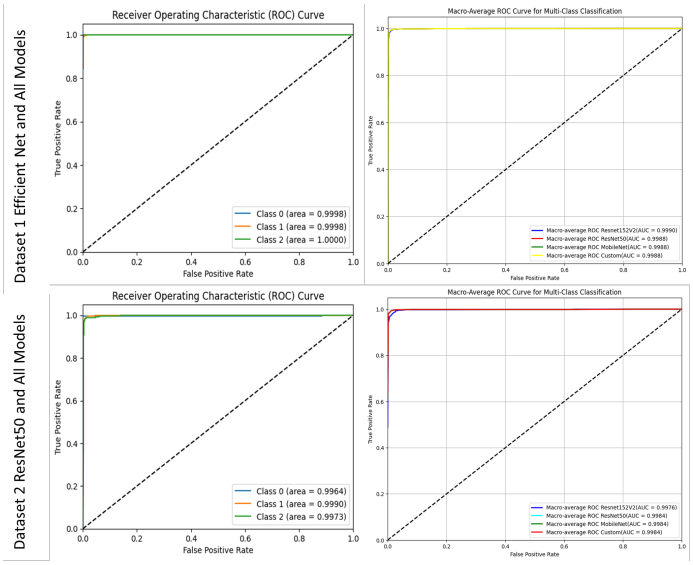

Fig. 7. ROC curve of all models and ensemble model

[10] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[11] K. Elissa, "Title of paper if known," unpublished.

[12] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[13] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[14] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.