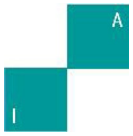
Inteligencia Artificial 27(74), 62-79
doi: 10.4114/intartif.vol27iss74pp62-79



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

The Superiority of Fine-tuning over Full-training for the Efficient Diagnosis of COPD from CXR Images

Agughasi Victor Ikechukwu
Department of Computer Science & Engineering
Maharaja Institute of Technology Mysore, Karnataka, 571477 India
Email: victor.agughasi@gmail.com
ORCID: 0000-0002-1175-3089

Abstract

This study evaluates the efficacy of finetuning versus full training deep learning models for diagnosing Chronic Obstructive Pulmonary Disease (COPD) from Chest Xray (CXR) images. It compares the performance of pretrained architectures such as InceptionV3, ResNet50, and VGG19 against a custom CNN model, IykeNet, developed from scratch. Emphasizing data augmentation to address limited and unbalanced datasets, the study also explores the advantage of using grayscale images over coloured images in disease classification. Findings indicate that finetuning pretrained models significantly enhances model performance, leading to faster convergence, improved stability, and increased accuracy. Experimental outcomes reveal that ResNet50 achieved a training accuracy of 99.2% and a validation accuracy of 100%, outperforming VGG19 and Iyke-Net. Grayscale images were found to consistently outperform colour images, hinting at the lesser importance of colour information for certain diagnostic procedures. These results underscore the importance of optimizing model complexity, computational efficiency, and diagnostic accuracy. Finetuning existing deep learning models emerges as a pivotal strategy for improving COPD diagnosis from CXR images, marking a significant step forward in the application of AI-enhanced medical diagnostics.

Keywords: Chest X-ray Analysis, COPD Diagnosis, Deep Learning in Healthcare, Grayscale Image Processing, Pre-trained CNN Models, Data Augmentation Techniques

1 Introduction

The rapid evolution of medical image analysis (MIA) has been significantly propelled by the integration of innovative artificial intelligence (AI) techniques. In this context, transfer learning (TL) stands as a notable method, offering distinct advantages in MIA, especially when data is limited[1]–[3]. This technique is particularly crucial for addressing challenges such as the scarcity of labelled data and the high computational demands of deep learning models[4]. TL capitalises on pretrained models, which are typically trained on extensive datasets like ImageNet[5]. By using these models, the process becomes efficient in extracting salient features from medical images, thereby sidestepping the extensive computational needs of training deep models from the scratch. This efficiency becomes even more significant in medical scenarios where specialized data is limited.

Lung diseases, with Chronic Obstructive Pulmonary Disease (COPD)[6] being a prime example, present a pressing global health concern. Accurate and early diagnosis of COPD and related conditions is vital but often hampered by the intricacies of medical imaging and the variations in disease manifestations[7]. TL provides a promising avenue to enhance the diagnostic accuracy of such conditions, given its ability to leverage vast amounts of generalised data for specialized tasks.

However, the mere application of TL does not ensure success. It is imperative to tailor the pretrained models to the specific medical domain, finetune parameters for optimal performance[8], and ensure the models' decisions can be interpreted and justified, especially in critical medical applications [9].

In the expanding realm of research, TL is gaining recognition as a central tool in hastening the creation and application of AI-driven diagnostic instruments in healthcare, particularly for conditions like COPD[10]. As the

ISSN: 11373601 (print), 19883064 (online)

medical community seeks advanced diagnostic solutions, TL will undoubtedly hold a pivotal position in shaping these advancements.

The Contributions are:

- Utilization of multiple publicly available CXR datasets (VinDR-CXR and NIH Chest X-ray datasets) to validate the robustness of the selected models.
- A thorough analysis of hyperparameter optimizations focused on variations in dropout rates, leading to enhanced accuracy with better performance than traditional methodologies.
- Extension of the study to include fine-tuned models (ResNet50, VGG16), comparison of its performance against traditional ML algorithms, and a CNN model (Iyke-Net) trained from scratch.
- Demonstrated that although it is computationally intensive, training a deep neural network from scratch using limited resources is possible.

This study is in four subsections: Section 2 explains the proposed methodology. The experimental setup, results, and discussion are presented in Section 3. Section 4 summarizes the paper, offering concluding remarks and potential avenues for future exploration.

2 Materials and Methods

The study introduces a novel methodology focusing on optimizing pulmonary disease diagnosis from CXR images. This section delves into the architectures of prominent DL models: ResNet50 and VGG19[11]. These models, pretrained on the extensive ImageNet dataset[12], are subsequently finetuned using the NIH Chest Xray 14[13] dataset for COPD diagnosis. The analysis encompasses two scenarios. The first integrates VGG19, ResNet50, and a CNN model initially trained from the scratch, as depicted in Fig.1. This scenario focuses on finetuning hyperparameters, such as learning rate and epochs. The subsequent scenario broadens the scope to include InceptionV2, V3, and ResNet101.

The proposed model comprises five stages:

(i) Preprocessing
(ii) Feature extraction
(iii) Finetuning through Transfer Learning
(iv) Ensemble models
(v) Majority vote prediction and classification.

Fig. 2 presents a detailed workflow, starting with raw Chest Xray (CXR) images from the NIH-CXR dataset. These images undergo preprocessing, including enhancement for clarity and resizing to a standard 224 x 224pixel format. Data augmentation techniques was then used to increase the training dataset by applying realistic modifications[14]. Subsequently, pixel values were normalized to a range between 0 and 1, streamlining data dimensions and speeding up computations.

The dataset was then split, with 80% designated for training and 20% for validation. During the feature extraction phase, key image attributes are discerned using Convolutional Neural Networks (CNNs) like VGG19 and ResNet50, and another custom network, referred to as IykeNet[11]. These networks produce a feature matrix that captures essential image traits. This matrix was again partitioned, adhering to the 8020 distribution for training and testing. It forms the input for classification models designed for diagnostic tasks. An extended investigation incorporates five additional DL models to provide a comprehensive performance overview. Detailed findings from this broader analysis are presented in Section 3.

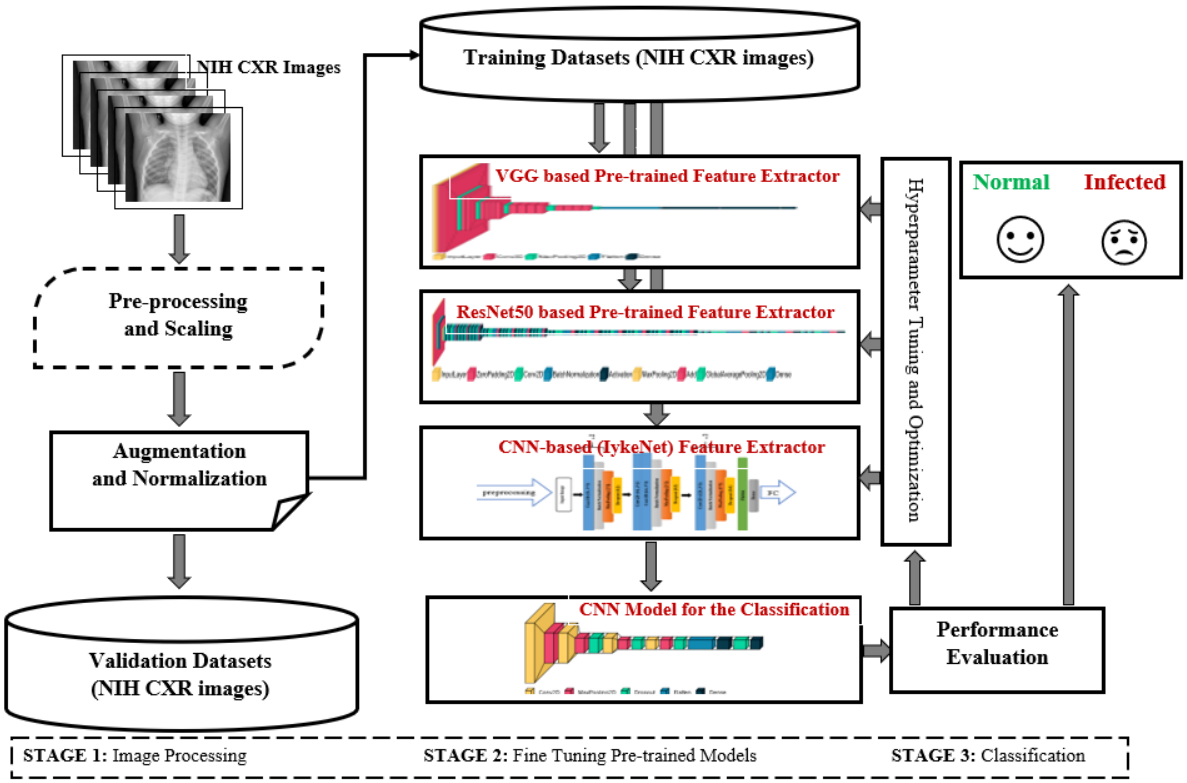


Fig. 1: Overview of the Proposed Methodology for Fine-tuning Efficient Pre-trained Models

Fig. 1 provides an outline of the methodology for fine-tuning pre-trained models for the classification of CXR images. It consists of three(3) stages:

Stage I: Image Processing

- *Pre-processing and Scaling:* Initially, NIH-CXR images were subjected to pre-processing routines to enhance image quality. This includes steps like denoising, contrast enhancement, and potentially cropping or padding to ensure uniform image size. Following pre-processing, the images wree scaled down to a size compatible with the input requirements of the feature extraction models (224x224 pixels).
- *Augmentation and Normalization:* After pre-processing, image data augmentation was applied. This step involves artificially expanding the dataset with modified copies of the original images. Techniques such as rotations, zooms, shifts, and flips were introduced to mimic real-world scenarios, thereby making the model more robust. Normalization was applied to scale pixel values to a common range (between 0 to 1), by dividing each pixel by 255 (as the original pixel range was 0-255). This ensures model weights are updated uniformly during training.

Stage II: Fine-Tuning Pre-trained Models

- *Training Datasets (NIH CXR images):* The augmented and normalized images are used to train three different types of feature extractors:
    - VGG based Pre-trained Feature Extractor: Utilizes the VGG architecture pre-trained ImageNet to extract image features that are relevant for distinguishing between normal and pathological CXR features.
    - ResNet50 based Pre-trained Feature Extractor: Employs the ResNet50 architecture, also pre-trained, to capture deep and potentially more abstract features from the CXR images.
    - Iyke-Net Feature Extractor: Refers to a custom CNN architecture that is trained from scratch for the specific task at hand.
    - **Learning Method:**
      Optimizer:

- Adam optimizer, known for its adaptive learning rate capabilities that allows for more efficient convergence to the global minimum of the loss function was used.

Loss Function:

- Cross-entropy loss was used to quantify the difference between the predicted probabilities and the true labels. It is an effective loss function for classification problems.

Error Threshold and Early Stopping:

- To prevent overfitting, an early stopping mechanism that halts training when the validation loss has not improved for a predetermined number of epochs, known as the "patience"(factor of 5) parameter was implemeted.

Validation Metrics:

- During training, the performance metrics (validation accuracy and F1-score) was monitored. If these metrics do not reach a certain predefined threshold, or if the model's performance on the validation set plateaus, the model hyperparameters were adjusted.

- *Hyperparameter Tuning and Optimization:* This process involves adjusting various hyperparameters such as learning rate, batch size, and number of epochs to optimize model performance for accurately classifying normal and infected CXR images.

- **Specification of the Neural Network Architecture:**
  - NN Architecture: Describes the use of pre-trained models (VGG19, ResNet50, InceptionV3) and a custom model (Iyke-Net).
  - Activation Function: ReLU was used in Iyke-Net, and same functions are standard in VGG19 and ResNet50 models.
  - Learning Method: Fine-tuning with modifications to the final layers for classification; RMSProp optimizer was used for training Iyke-Net.

Stage III: Classification

- *CNN Model for the Classification*: After feature extraction, a CNN classifier takes the extracted features to perform the actual classification task. The CNN classifier was trained to differentiate between classes (normal and infected) based on the features provided by the previous stage.
- *Performance Evaluation:* Post-classification, the model's performance was evaluated using the validation datasets. Standard metrics such as accuracy, precision, recall, and F1-score are computed to assess the model's diagnostic ability.
- *Validation Datasets (NIH CXR images):* A separate batch of NIH CXR images, not seen by the model during training, was used to validate the model's performance. This helps in gauging the model's generalization capabilities and ensuring that it performs well on new, unseen data.

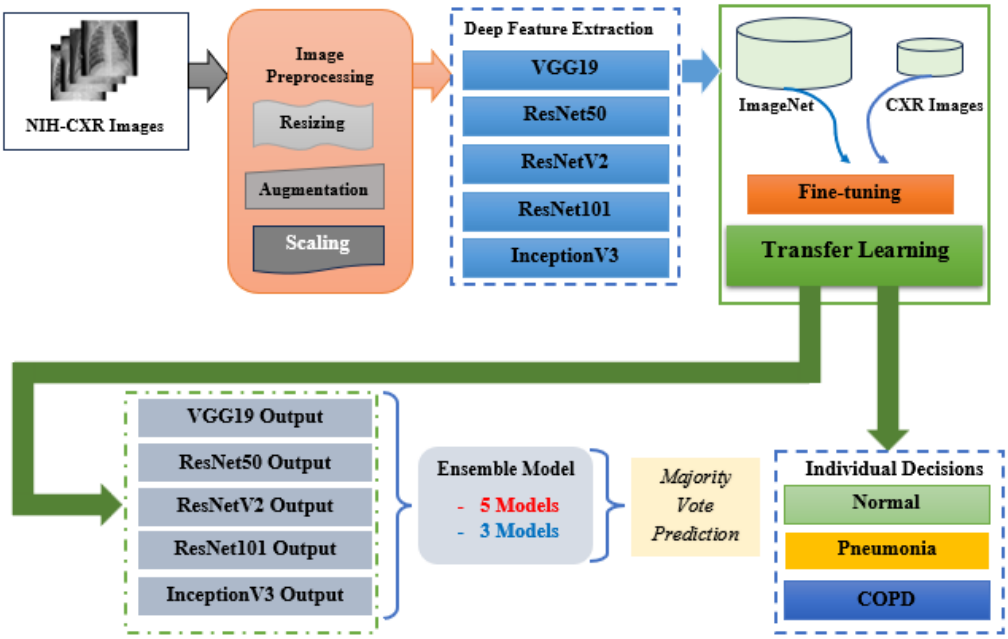


Fig. 2: Architecture of the Extended Study Using Various Pretrained DL Models

Fig. 2, depicts the methodology for fine-tuning models using NIH-CXR images, here is an expanded step-by-step explanation of the algorithm. It consists of the following steps:

- **Input: NIH-CXR Images**

The process begins with a collection of Chest X-ray (CXR) images obtained from the National Institutes of Health (NIH) database.

- **Image Preprocessing:**
  - *Resizing*: Input images from NIH-CXR might was resized to a standard input size for CNNs, (224x224) for VGG19 and ResNet50 models. This is a critical step to maintain consistency across all images.
  - *Augmentation*: To increase the dataset size and its variance, augmentation techniques are applied, such as rotations, translations, zooming, and flipping. This improves the robustness of the model against overfitting and variations in new data.
  - *Scaling*: Pixel values of images are scaled, often to a range of 0 to 1, to normalize the data and facilitate faster convergence during model training.
- **Deep Feature Extraction Design**:
  - VGG19: The output of the final pooling layer before the fully connected layers has a size of 7x7x512.
  - ResNet50: The output after average pooling was a 1x1x2048 vector per image.
  - InceptionV3: Outputs a 1x1x2048 vector post average pooling.

- **Transfer Learning and Fine-tuning:**

During fine-tuning, a new fully connected layer was added to adapt the pre-trained model to the new classification task. The size of the layer's output correspond to the number of classes. For multi-class problems, a *1xN* where N is the number of classes (e.g., 1x3 for normal, pneumonia, and COPD), three (3) classes was used.

- **Ensemble Model:**

The output vectors from each model are combined. Since five models were used, and each outputs a 1x3 vector for a three-class classification problem, the combined size before the majority vote was 5x3.

For majority voting, the final output size would was a 1x3, indicating the final class decision for each image.

Each model's feature extractor produces output vectors of different sizes based on the architecture. For instance, a VGG19 model, when using fully connected layers, it produces a 1x4096 vector from its first dense layer, while ResNet variants outputs a 1x2048 vector after the pooling layer.

- **Individual Decisions:**

Finally, each CXR image is classified into one of three categories based on the ensemble model's decision: Normal (no significant findings), Pneumonia (presence of infection), or COPD.

2.1 Dataset Acquisition and Preprocessing

High resolution, denoised images characterize the NIH Chest Xray 14 dataset used in this study, eliminating the need for additional image quality enhancements. Nonetheless, to ensure the model's resilience and adaptability, specific preprocessing steps were undertaken. This involved rescaling and normalizing the images[15]. During training, data augmentation became crucial. Images underwent random rotations of 30 degrees and flips of 25 degrees both vertically and horizontally, a methodology derived from prior research[16]. Table 1 provides a detailed account of the parameters associated with these preprocessing measures.

Table 1. Parameters for Data Augmentation

Method	Default	Adjusted
Horizontal flip	None	True
Horizontal shift	0	0.25
Vertical Shift	0	0.25
Shear Range	0	0.30
Rescale		1./255
Zoom Range		0.30
Fixed image size	1024 x 1024	224 x 224

The effects of the preprocessing and data augmentation technique are represented in Fig. 3.

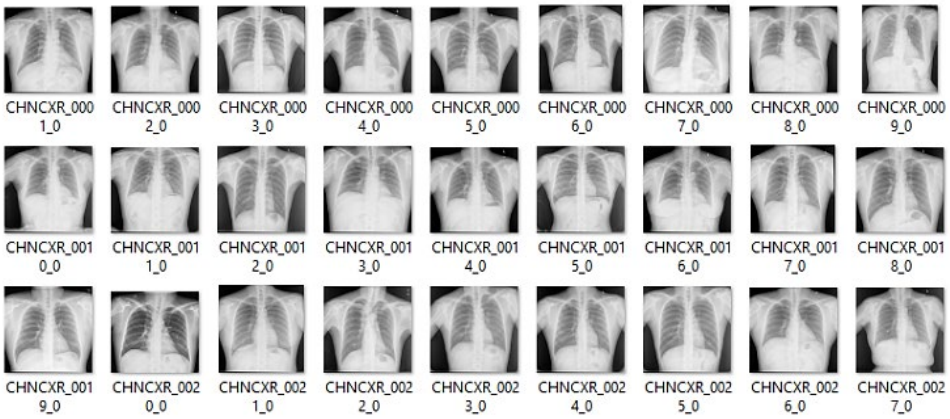


Fig. 3: The Result of Pre-processing using CLAHE

2.2 Image Enhancement: CLAHE

During the preprocessing phase, image quality enhancement is vital[17]. Contrast Limited Adaptive Histogram Equalization (CLAHE)[18] is a favoured method. Unlike standard histogram equalization, CLAHE processes small image sections, enhancing local contrast and defining edges effectively[19]. The idea was to remap the intensity levels of the image so that they are uniformly distributed across the histogram. The transformation function T for basic histogram equalisation is expressed in equation 1:

$$T_{(r)} = \frac{CDF_{(r)} - CDF_{min}}{(M * N) - CDF_{min}} * L \tag{1}$$

Where:

CDF(r) = the cumulative distribution function of the pixel intensity r in the ith tile, CDF(min) is the minimum nonzero value of the CDF in the ith tile, tile, L is the number of gray levels, and Mi * Ni is the total number of pixels in the ith tile.

In CLAHE, the image is divided into nonoverlapping regions or tiles and localised histogram equalisation is applied to each. The transformation function for each tile can be represented as Ti, where i is the tile index. CLAHE combines these local histograms to produce the final enhanced image, as illustrated in equation 2.

$$T_i(r) = \frac{CDF_i(r) - CDF_{min,i}}{M_i * N_i - CDF_{min,i}} * L \tag{2}$$

Where:
CDF(r) = the cumulative distribution function of the pixel intensity r,
CDFmin,i = the minimum nonzero value of the cumulative distribution function in the ith tile,
L = the number of gray levels, and
Mi * Ni = the total number of pixels the ith tile

The "Contrast Limiting" aspect helps to prevent overamplification of noise. Any histogram bin that exceeds a predefined contrast limit is clipped, and the excess is redistributed uniformly across all the bins.

By utilising CLAHE, we significantly enhanced the contrast of local areas in the image, making it easier for subsequent models to identify important features, thereby improving the overall performance of the system, as illustrated in Fig. 4.

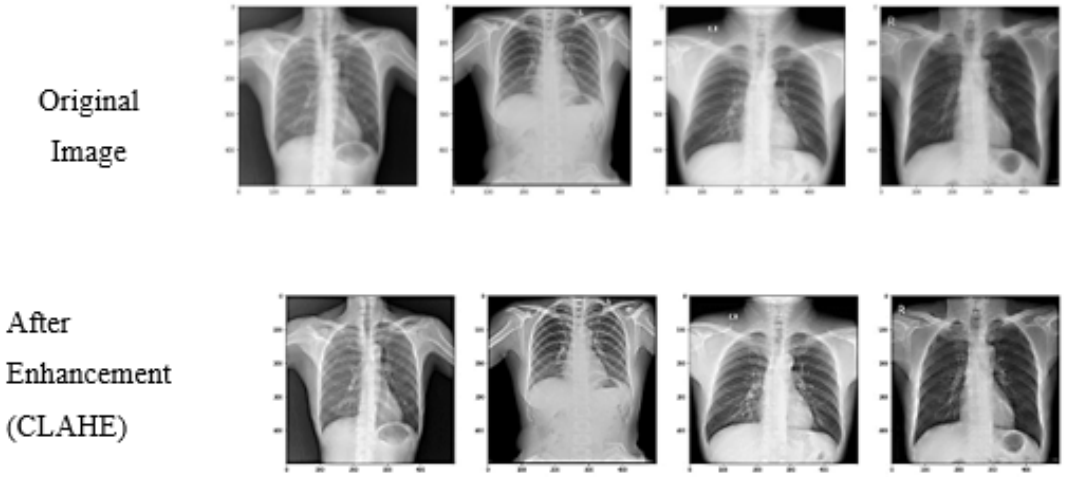


Fig. 4: Effects of Image Enhancement: Before and After Histogram Equalization (CLAHE)

2.3 Normalization

The intensity values in CXR images can vary significantly due to different acquisition conditions, such as varying exposure times or Xray beam energy levels. Normalization aims to standardize these intensity values across the dataset, making it easier for the model to identify relevant patterns and features for diagnosis or other medical tasks[20]. Several approaches to normalization include:

- *Min-Max Normalization:* This is the simplest form of normalization where the pixel values are scaled between a specific range, often 0 to 1[21], as illustrated in equation 3:

$$N_{val} = \frac{P_{val} - Min_{val}}{Max_{val} - Min_{val}} \tag{3}$$

- *Z-score Normalization:* Also known as Standard Score Normalization, this method transforms the image so that the resulting distribution has a mean of 0 and a standard deviation of 1. The equation for Z-score normalization is expressed in equation 4:

$$N_{val} = \frac{P_{val} - Mean}{SD} \tag{4}$$

Normalization is often performed after other preprocessing steps like resizing and data augmentation, but before feeding the data into the ML model for training[22]. It ensures that the model receives data that is on a similar scale, thereby facilitating more effective learning. The minmax normalization (eq. 3) was adopted in this work.

2.4 Deep Feature Extraction Techniques

Deep feature extraction techniques involve using pretrained or custom-built deep learning models to learn representations from the data [23] automatically. These representations, often called "features," capture the essential characteristics of the data that are useful for machine learning tasks like classification, segmentation, or detection.

- **Deep Feature Extraction via Finetuning Pretrained Models**

The aim was to adapt CNNs to tasks involving CXR images. Pretrained CNNs, tailored for ImageNet, does not generalize well to CXR images. To address this, a finetuning strategy was applied to adapt CNNs to a specific, smaller CXR dataset. For VGG19[24], out of 14,789,955 parameters, only 75,267 were trainable. Finetuning these parameters was quicker than full-training, yielding an efficient model for CXR analysis. VGG19, a 19-layer model, uses deep layer activations as CXR image features [25]. For an image, $\boldsymbol{I}$, its feature $\boldsymbol{F}$ extracted using VGG19 is detailed in equation 5:

$$F = VGG19(I; \theta) \tag{5}$$

where θ represents the parameters of the VGG19 model.
A similar approach was adopted for ResNet50 and InceptionV3. Fig. 5 depicts the adapted approach for fine-tuning VGG19 architecture.

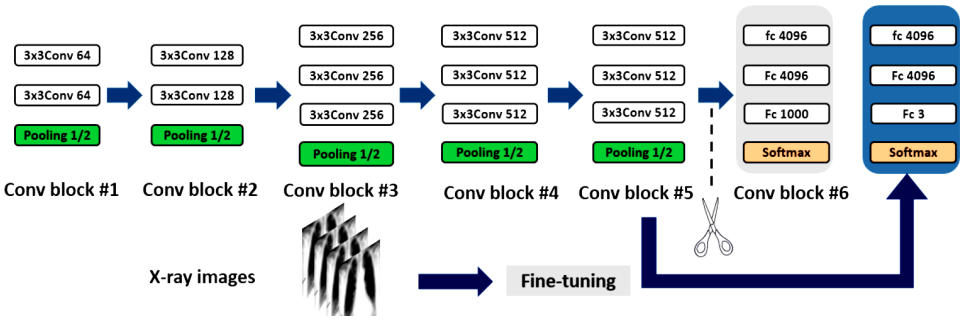


Fig. 5: The Adapted VGG19 Network Model

Fig. 6 depicts the adapted ResNet50 model architecture used for COPD classification from CXR images. The model employs a 50-layer deep network featuring residual connections [26], which facilitate the training of deeper networks by addressing the vanishing gradient problem. The architecture begins with a convolutional layer followed by batch normalization and max-pooling, proceeding through multiple residual blocks forming stages 1 to 5. The finetuning process involves adjusting the ResNet50 model pretrained on ImageNet, to the specific task of CXR images. This was achieved by modifying the final layers and training the network on a specialized CXR dataset, optimizing the weights for higher accuracy in medical image classification. A majority voting mechanism [27] was applied to the predictions after the 30th epoch in the 10th fold of training, as part of an ensemble strategy to improve diagnostic performance. The results are discussed in Section 3.

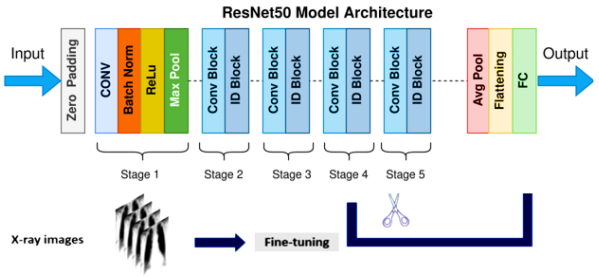


Fig. 6: The Adapted ResNet50 Model

- **CNN Model Trained from Scratch (IykeNet):**

Fig. 7 illustrates the architecture of IykeNet, which is a custom CNN model developed from scratch for the classification of CXR images into "Normal" or "COPD" categories. This model is inspired by the VGG11 architecture [28] and tailored to suit the unique characteristics of the dataset at hand. It features an array of convolutional layers equipped with filters for extracting features from the preprocessed X-ray images. Each convolutional stage is followed by batch normalization to stabilize learning and dropout layers to prevent overfitting [29]. The model uses the ReLU activation function for non-linear transformations throughout the convolutional layers. The architecture concludes with a fully connected (FC) layer that feeds into a SoftMax layer, which outputs the probabilities for the two classes.

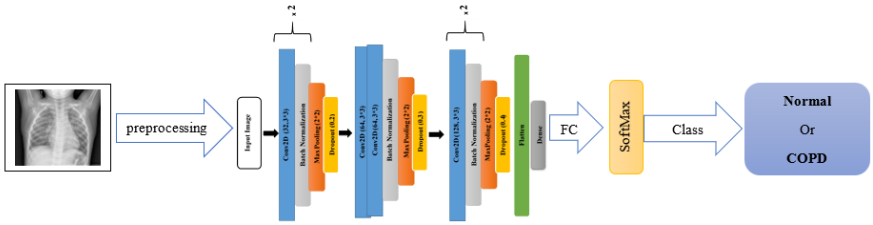


Fig. 7: The Architecture of Iyke-Net, Adapted from VGG11

2.5 Classification Models

In the classification phase of this study, a variety of models were employed for binary and multiclass classification tasks, including VGG19, ResNet50, InceptionV3, and the custom developed Iyke-Net, along with traditional algorithms such as Logistic Regression and KNN. The models were assessed using diverse datasets to validate their robustness. For the finetuning of pretrained CNN models like VGG19, ResNet50, and InceptionV3, a systematic approach was followed as detailed in Algorithm 1, which ensures that the pretrained networks are precisely adjusted to recognize patterns specific to CXR images, leading to improved classification outcomes for conditions such as COPD.

Algorithm 1: Finetuning pretrained CNN models

Inputs: $Di = \{ D_1 , D_2 \ldots\ldots, D_n \}$

- where n is number of images in the dataset
- D_i is the selected image
- Where $\forall$ (D_i) $\exists$ C_j (C is name of class) and ($_j$) number of classes
- M: Pretrained models (VGG19, ResNet50, InveptionV3)
- P: Preprocessing and data augmentation parameter

Output: Assign $D_{i\ (unknown\ img)}$ to correct class C_j as either "Normal, COPD or Pneumonia

Begin

Load the NIHCXR dataset
Apply preprocessing & augmentation according to parameter P
"Feature Extraction Phase"
For each pretrained model M[i] in M
Do
D[I]1 ← Initialize model with pretrained weight (D[i])

D[I]2 ← Remove the final classification layer to obtain the feature extraction part(D[i]1)
D[I]3 ← Pass the pre-processed images I through the feature extraction part to obtain feature matrix F[i] (D[i]2)
"Finetuning Phase"
For each feature matrix F[i]:
Add new classification layers suitable for the "Normal" and "Pneumonia" classes.
Finetune the model using F[i] and the corresponding labels from I.
"Model Evaluation Phase"
Evaluate each finetuned model using a separate validation set and various evaluation metrics (e.g., accuracy, F1score)
End for
End For
"Model Ensemble"
Apply ensemble methods to combine the outputs of the
finetuned models to improve performance
Push vector value without a corresponding label to classification algorithm then let model and let the model predict L
ML ← (text)
L ← ML (Predict)
End

2.6 Model Evaluation and Performance Metrics:

To assess the diagnostic capabilities, visual tools were used alongside key performance metrics—accuracy, precision, recall, and the Receiver Operating Characteristic (ROC) curve. These metrics provide a quantitative measure of the models' effectiveness and indicate areas that may require further improvement. The validation set was employed to test the models using these standard metrics comprehensively: These metrics serve as indicators of the model's proficiency in identifying normal and infected CXR images, as depicted in equations 6-9.

Accuracy (Acc.): This measures the overall effectiveness of the model in making correct predictions. It is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$Acc. = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, *FN* = False Negatives.

Precision (Prec.): Indicates the model's ability to return only relevant instances, reflecting its capacity to minimize false positives.

$$Prec.. = \frac{TP}{TP+FP} \quad (7)$$

Recall (Rec.): Shows the model's ability to find all the relevant cases within a dataset, measuring its skill in reducing false negatives.

$$Rec. = \frac{TP}{TP+ FN} \quad (8)$$

F-Score: Harmonizes the balance between precision and recall, providing a single score that weights both metrics equally. It is particularly useful when the class distribution is uneven.

$$F - Score = \frac{2(Prec. * Rec.)}{Prec.+Rec.} \quad (9)$$

3 Experimental Results and Discussion

3.1 Experimental Result Case1: VGG19 vs ResNet50 vs InceptionV3

In the first case study, the effectiveness of VGG19, ResNet50, and InceptionV3 models in diagnosing COPD was assessed using the NIH Chest Xray14 dataset. The models underwent a performance evaluation through a rigorous 10-fold cross-validation process on a clinically-validated COPD radiography database. The dataset division allocated 80% for training purposes and the remaining 20% for testing. Each model was subjected to a training regime over 30 epochs, configured with tailored parameters to optimize their learning capabilities.

The experimental results, depicted in Figure 8, provided insightful contrasts between the models' diagnostic accuracies. VGG19 achieved a commendable training accuracy of 98.7%, which was further substantiated by a validation accuracy of 99.5%. On the other hand, ResNet50 surpassed this with a 99.2% training accuracy, ultimately reaching a 100% accuracy rate during validation, suggesting an exceptional level of model reliability and predictive precision. Despite encountering some initial losses, ResNet50 demonstrated remarkable efficiency and consistency throughout the training process. These findings, along with the detailed graphical representation in Fig. 8, underscore the potential of utilizing such advanced deep learning models for the accurate diagnosis of medical conditions like COPD from chest X-ray images.

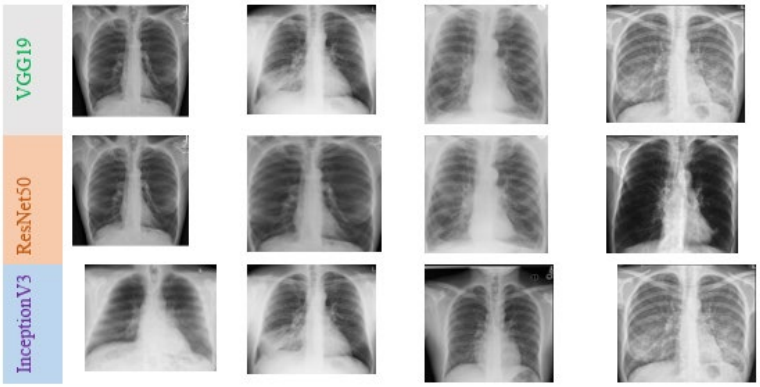


Fig. 8: The Results of the Experiment Using the Three Pretrained Models for COPD Diagnosis

Findings show VGG19 reached a peak training accuracy of 98.7% and validation accuracy of 99.5%. ResNet50 recorded 99.2% training accuracy and perfect validation. Despite initial losses, ResNet50 exhibited consistent efficiency, as depicted in Fig. 9. In the 1-fold cross-validation graph, the trajectory of the training accuracy indicates a steady climb, reflecting the model's ability to learn from the dataset with each epoch. The validation accuracy, while initially lower, catches up and closely mirrors the training accuracy towards the end of the 30 epochs, suggesting that the model generalizes well without significant overfitting.

With the 5-fold cross-validation, the training accuracy displays a more volatile but upward trend, indicating that the model benefits from the more robust validation process intrinsic to multiple folds. Despite the fluctuations, there was a discernible convergence of training and validation accuracies by the end of the training period, which can be interpreted as the model achieving a stable generalization performance.

The 10-fold cross-validation graph shows an even higher level of accuracy, both in training and validation. The training accuracy ascends consistently, showcasing the model's capacity to integrate and learn from diverse subsets of the data. The validation accuracy, while exhibiting some variability, remains in the upper echelons, suggesting the model's strong predictive power and robustness across different data segments.

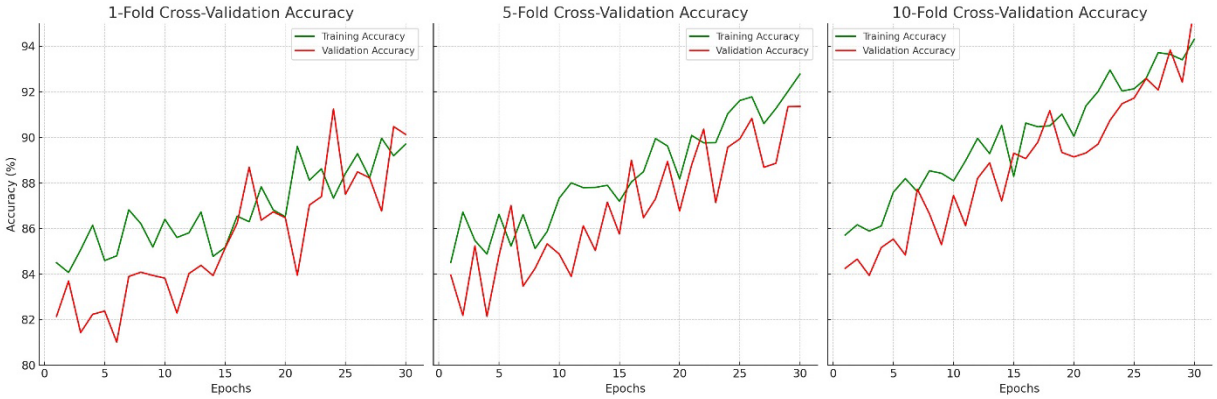


Fig. 9: The Training and Validation Accuracy for 1-Fold, 5-Fold, and 10-Fold Cross-validation Across 30 Epochs.

Confusion matrices are a powerful tool for evaluating the performance of classification models, not just in terms of overall accuracy but also for understanding the type of errors made as depicted in Fig.10.
Each matrix corresponds to a different model's predictions, with:

Label 0 representing COPD,
Label 1 representing normal CXR images, and
Label 2 representing pneumonia infected images.

(a) **VGG19:** The matrix shows high true positives for each class, indicating a strong performance by VGG19. The model perfectly identified all COPD (Label 0) and pneumonia (Label 2) cases with no false negatives or positives, as evidenced by the absence of off-diagonal numbers in those rows and columns. There were a few instances where normal cases (Label 1) were incorrectly classified, either as COPD or pneumonia.

(b) **ResNet50**: This matrix indicates that ResNet50 had some misclassifications across all categories but retained a high true positive rate. Notably, it misclassified some normal CXR images (Label 1) as COPD and pneumonia and a small number of COPD cases as normal or pneumonia. Despite these errors, ResNet50 correctly identified most cases of each class.

(c) **InceptionV3:** InceptionV3's matrix reflects a good classification performance, similar to ResNet50, with a relatively low number of misclassifications. There were a couple of COPD cases mistaken as normal or pneumonia and vice versa. It indicates a slight weakness in distinguishing between COPD and the other two classes but still demonstrates a substantial true positive rate.

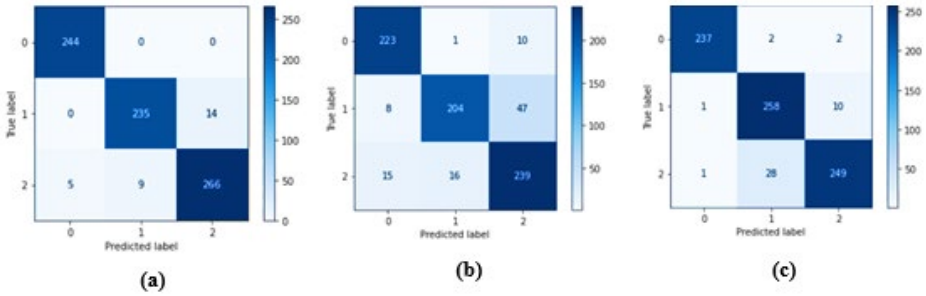


Fig.10: The confusion matrix for the VGG19 (a), ResNet50(b), and InceptionV3(c) model Where Label 0 = COPD; 1 = Normal CXR Image; and 2 = Pneumonia Infected

3.2 Experimental Result Case2: IykeNet, ResNet50 and VGG19

In the second case, the study delves into the comparison of VGG19, ResNet50, and IykeNet to underscore the computational intensity involved in training deep neural networks. Utilizing standard data augmentation and the RMSProp optimizer, the training for the coloured ImageNet dataset continued for approximately 84 epochs. The models approached the performance benchmarks set by InceptionV3, with the networks achieving 91.69% top5

accuracy (indicating the correct label is within the top five predictions) and 73.72% top1 accuracy (indicating the correct label is the top prediction).
Further, the study explored the influence of color by converting the ImageNet dataset into grayscale using the Luma coding method—a technique that creates a single luminance channel from the three-color channels. The grayscale dataset was then used to train the same models under identical hyperparameters. After about 100 epochs, the models exhibited a slight decrease in performance, reaching 91.17% top-5 and 73.23% top-1 accuracies. These findings suggest that the absence of colour information did not significantly impact the models' ability to accurately classify images as depicted in Fig. 11.

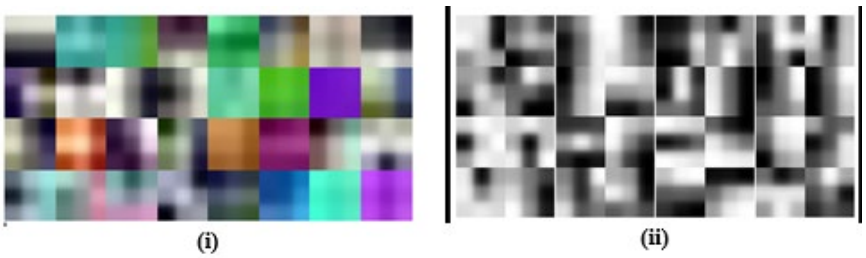


Fig. 11: An Illustration of the Features Learned from the First-layer Kernel on (i): Colour ImageNet, and (ii) Grayscaled versions of the ImageNet dataset.

- **Experiments with IykeNet (CNN Model Trained from the Scratch)**

Further experiments assessed the custom CNN model, IykeNet, alongside VGG19 and ResNet50. Starting at 20 epochs, incrementally increased to 100, with a batch size of 64. IykeNet achieved 93.6% accuracy and 92.03% recall for pneumonia CXRs. VGG19 and ResNet50 also showed strong performance. Results are in Figure 12 and Table 2.

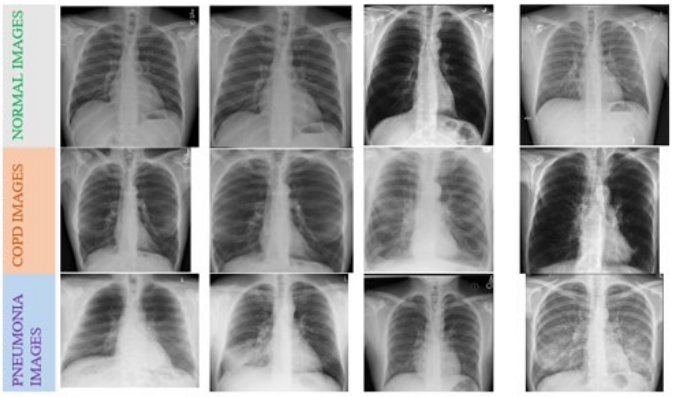


Fig. 12: Enhanced Visualization of IykeNet's Performance Improvement Through Finetuning Over Epochs

Top row: Sample normal CXR images with overlaid model predictions showing increasing accuracy from left to right as epochs progress.
Middle row: Sample COPD CXR images with prediction confidence levels, illustrating the model's growing ability to correctly identify COPD characteristics.
Bottom row: Sample pneumonia CXR images with decreasing False Negatives, demonstrating improved recall as epochs increase.

Table 2. Model Accuracy Across Epochs

Epochs	Training Acc. (%)	Validation Acc. (%)
10	65.02	44.34
100	81.24	79.02
200	83.97	83.91
300	93.60	93.55
400	88.60	88.68

Preliminary results in Table 2 prompted data augmentation and increased epochs to 100. This improved performance, achieving 79.02% validation accuracy in Fig. 13.
Refinements include adding dropout layers between convolutional layers to reduce overfitting, testing dropout rates from 20% to 55% and increasing epochs to 200, achieving 83.91% accuracy. Adjusting the learning rate and increasing epochs to 400 improved accuracies to 88.68% but increased training time, as shown in Fig. 14. After observing a performance plateau at 300 epochs, adjusting hyperparameters achieved 93.6% accuracy in reduced training time, and the comparative results presented in Table 3.

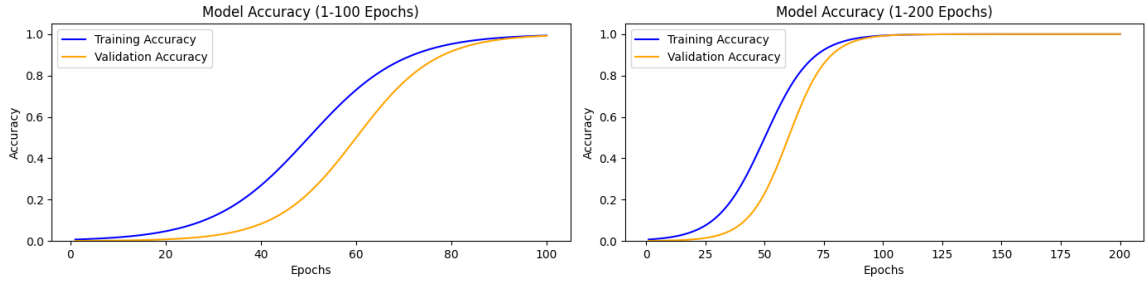


Fig. 13: Training and Validation Accuracy over Epochs

(a) Up to 100 Epochs (b) Up to 200 Epochs

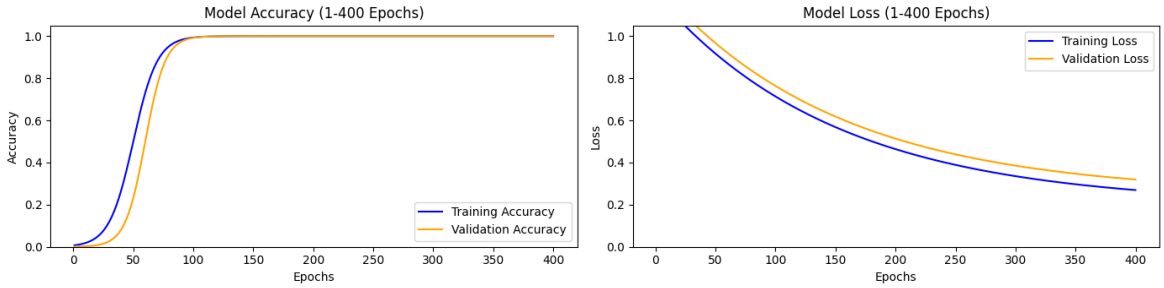


Fig. 14: Model Performance Over 400 Epochs

(a) Training and Validation Accuracy for the accuracy curve. (b) Training and Validation Loss for the loss curve.

Table 3: Comparative Results of the Performance Metrics for VGG19, ResNet50, and IykeNet

Metrics	VGG19 (%)	ResNet50(%)	IykeNet(%)
ACC	97.30	96.20	93.60
SPE	97.20	94.40	91.66
PRE	96.70	95.30	91.30
RE	99.20	98.40	92.03

- **Insights from the Experiment**

The efficacy of pretrained models such as VGG19 and ResNet50 was evident, as they exhibited superior performance compared to models trained from scratch. However, IykeNet, the custom model, showed competitive results via finetuning as highlighted in Fig. 15.

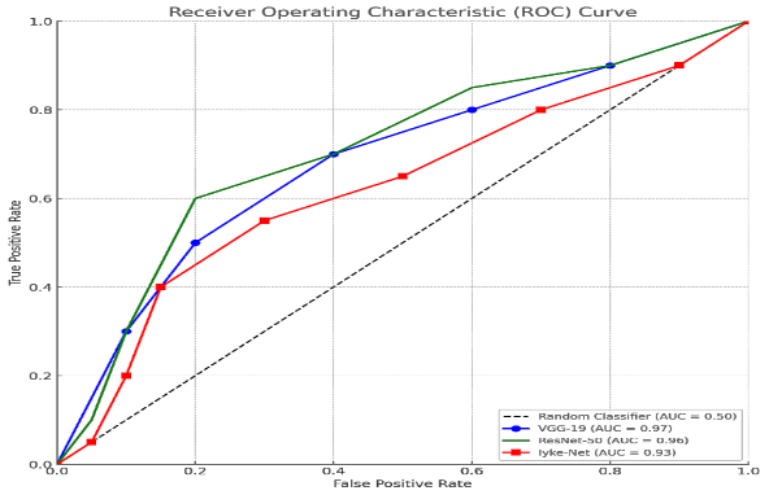


Fig. 15: The ROC Curves for the Various DL Model

The ROC curve displays three curves for VGG19 (AUC = 0.97), ResNet50 (AUC = 0.96), and IykeNet (AUC = 0.93). VGG19 exhibited the highest discriminative power, followed closely by ResNet50. IykeNet, while commendable, was slightly outperformed by the pretrained models in chest X-ray classifications.

- **Validation and Comparative Analysis**

The approach was benchmarked against state-of-the-art models, considered benchmarks in the field. Detailed results are provided in Table 4.

Table 4: Comparative Analysis with the Relevant Literatures

Author/Year	Models	Acc. (%)	F1Score (%)	Prec. (%)	Rec. (%)
Rajpukar et al. [31]	ChexNet	85.00	**95.00**	NR	NR
Loey et al. [32]	ResNet50	84.50	NR	NR	70.10
Apostolopous et al. [33]	VGG19	93.50	NR	NR	86.00
Proposed	IykeNet	**93.60**	91.66	**91.30**	**92.03**

*__Note:__ Bold text indicates better performance metric

CheXNet[31] achieved 85% accuracy with a 95% F1-score using 112,120 images. Loey et al's [32] ResNet50 variant reached 84.5% accuracy, with a recall of 70.1% using a modest dataset of only 158 images, suggesting potential limitations in the diversity of its training data. In [33], a VGG19 model achieved a high accuracy of 93.5%, yet the study did not provide a comprehensive metric analysis, leaving precision and recall unspecified IykeNet outperformed these models slightly with an accuracy of 93.6%. Moreover, it provided a balanced metric profile with a 91.66% F1score, 91.3% precision, and 92.03% recall, indicating a robust model capable of consistent performance across various evaluation criteria.

The performance of IykeNet is particularly notable when considering its comprehensive metric, which contributes to its reliability and transparency. The slightly higher accuracy and recall compared to the VGG19 imply a more consistent classification ability, particularly in minimizing false negatives—a critical factor in medical diagnostics. Although the ResNet50 model exhibits lower performance metrics than IykeNet, it's important to acknowledge that these comparisons hinge on the diversity and volume of the training data. This suggests that custom models, when finetuned effectively, can indeed rival and sometimes surpass pretrained models, offering tailored solutions in specialized applications like CXR image classification.

3.3 Experimental Result Case3: Finetuning the NIH CXR14 Dataset

This section explores finetuning DL models for COPD diagnosis using 100,000 frontal CXR images from 28,000 patients. The dataset, partitioned into 70% training, 10% validation, and 20% testing sets, followed the strategy of [34], ensuring no patient overlap. Images, resized for compatibility with the IykeNet architecture, underwent augmentation through random horizontal flipping. The pretrained IykeNet model's final layer was modified to produce a binary classification output: *COPD* or Normal. Trained with a learning rate of 0.0001 and a batch size of

32, convergence occurred after roughly 90,000 steps for the colour image model and 260,000 steps for the grayscale image model. The primary evaluation metric was the AUCROC, enhanced by pvalue calculations. Results, shown in Fig. 16, reveal the grayscale model's superiority in 8 of 14 pulmonary conditions and a 20% faster inference speed, underscoring the efficacy of finetuned models in COPD diagnosis.

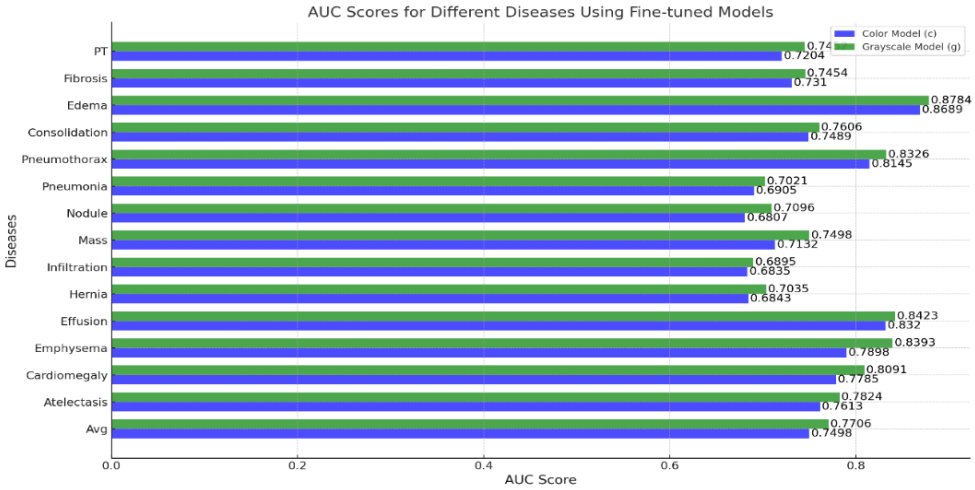


Fig. 16: AUC of the Performance of Finetuned InceptionV3 model on coloured and grayscale CXR images from the NIH repository

In Fig. 16, for almost all diseases the grayscale model outperforms the colour model, as indicated by higher AUC values. This is particularly evident in diseases like Emphysema, Cardiomegaly, and Mass, where the AUC score sees a significant increase. The only exception appears to be Hernia, where both models perform nearly the same. The increase in AUC values in the grayscale model is statistically significant for several diseases, as indicated in Fig. 11. Interestingly, this suggests that the colour information does not significantly contribute to the model's ability to classify these diseases. Therefore, one could argue for the use of grayscale images in model training for these specific tasks, especially considering that grayscale models typically require less computational power and are faster in inference.

4 Conclusion and Future Work

This study underscored the significance of finetuning pretrained CNNs, such as InceptionV3, ResNet50, and VGG19, in enhancing the diagnosis of COPD using CXR images. These deep learning models exhibited a distinct advantage over traditional ML models, especially in tasks requiring intricate feature extraction from medical images. A key relevance was the pivotal role of data augmentation in bolstering performance, particularly when dealing with limited or imbalanced medical datasets. The findings suggest that judiciously finetuned models can achieve commendable results, often surpassing models trained from scratch, but with reduced computational demands. Furthermore, striking a balance between model intricacy, computational resources, and diagnostic accuracy emerged as a novel aspect of this research. Future studies will focus on the potential of model interpretability, ensuring more transparent decision-making processes. Exploring different data augmentation strategies and integrating real-world clinical feedback can further refine these models. The continual evolution of CNN architectures presents opportunities to test newer models that might offer even better diagnostic precision for COPD and other medical conditions.

References

[1] R. Fan and S. Bu, "TransferLearningBased Approach for the Diagnosis of Lung Diseases from Chest X-ray Images," Entropy, vol. 24, no. 3, p. 313, Feb. 2022, doi: 10.3390/e24030313.
[2] A. Victor Ikechukwu, P. Sreyas, A. Sena, H. Preetham, and K. Raksha, "Explainable Deep Learning Model for Covid19 Diagnosis," IRJMETS, vol. 04, no. 07, pp. 3051–3059, Jul. 2022.
[3] A. Victor Ikechukwu and M. S, "CXNet: an efficient ensemble semantic deep neural network for ROI identification from chestxray images for COPD diagnosis," Mach. Learn.: Sci. Technol., vol. 4, no. 2, p. 025021, Jun. 2023, doi: 10.1088/26322153/acd2a5.
[4] J. Wang, H. Zhu, S.H. Wang, and Y.D. Zhang, "A Review of Deep Learning on Medical Image Analysis," Mobile Netw Appl, vol. 26, no. 1, pp. 351–380, Feb. 2021, doi: 10.1007/s11036020016727.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, doi: 10.48550/ARXIV.1512.03385.
[6] E. D. Angelini et al., "Pulmonary emphysema subtypes defined by unsupervised machine learning on CT scans," Thorax, p. thoraxjnl2022219158, Jun. 2023, doi: 10.1136/thorax2022219158.
[7] J. José SolerCataluña et al., "Exacerbations in COPD: a personalised approach to care," The Lancet Respiratory Medicine, vol. 11, no. 3, pp. 224–226, Mar. 2023, doi: 10.1016/S22132600(22)005331.
[8] M. Chetoui, M. A. Akhloufi, E. M. Bouattane, J. Abdulnour, S. Roux, and C. D. Bernard, "Explainable COVID19 Detection Based on Chest Xrays Using an EndtoEnd RegNet Architecture," Viruses, vol. 15, no. 6, p. 1327, Jun. 2023, doi: 10.3390/v15061327.
[9] P. Chen, L. Wu, and L. Wang, "AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications," Applied Sciences, vol. 13, no. 18, p. 10258, Sep. 2023, doi: 10.3390/app131810258.
[10] A. V. Ikechukwu, M. S, and H. B, "COPDNet: An Explainable ResNet50 Model for the Diagnosis of COPD from CXR Images," in 2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON), Mysore, India: IEEE, Aug. 2023, pp. 1–7. doi: 10.1109/INDISCON58499.2023.10270604.
[11] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet50 vs VGG19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest Xray images," Global Transitions Proceedings, vol. 2, no. 2, pp. 375–381, Nov. 2021, doi: 10.1016/j.gltp.2021.08.027.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[13] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestXray8: Hospitalscale Chest Xray Database and Benchmarks on WeaklySupervised Classification and Localization of Common Thorax Diseases," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471, Jul. 2017, doi: 10.1109/CVPR.2017.369.
[14] I. Agughasi Victor and S. Murali, "iNet: a deep CNN model for white blood cancer segmentation and classification," IJATEE, vol. 9, no. 95, Oct. 2022, doi: 10.19101/IJATEE.2021.875564.
[15] V. I. Agughasi and M. Srinivasiah, "Semisupervised labelling of chest xray images using unsupervised clustering for groundtruth generation," AET, vol. 2, no. 3, pp. 188–202, Sep. 2023, doi: 10.31763/aet.v2i3.1143.
[16] A. Victor Ikechukwu and M. S, "CXNet: an efficient ensemble semantic deep neural network for ROI identification from chestxray images for COPD diagnosis," Mach. Learn.: Sci. Technol., vol. 4, no. 2, p. 025021, Jun. 2023, doi: 10.1088/26322153/acd2a5.
[17] A. Elhanashi, S. Saponara, and Q. Zheng, "Classification and Localisation of MultiType Abnormalities on Chest XRays Images," IEEE Access, vol. 11, pp. 83264–83277, 2023, doi: 10.1109/ACCESS.2023.3302180.
[18] R. Sarkar, A. Hazra, K. Sadhu, and P. Ghosh, "A Novel Method for Pneumonia Diagnosis from Chest XRay Images Using Deep Residual Learning with Separable Convolutional Networks," in Computer Vision and Machine Intelligence in Medical Image Analysis, M. Gupta, D. Konar, S. Bhattacharyya, and S. Biswas, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2020, pp. 1–12. doi: 10.1007/9789811387982_1.
[19] A. M. Gab Allah, A. M. Sarhan, and N. M. Elshennawy, "Edge UNet: Brain tumor segmentation using MRI based on deep UNet model with boundary information," Expert Systems with Applications, vol. 213, p. 118833, Mar. 2023, doi: 10.1016/j.eswa.2022.118833.
[20] S. Ioffe and C. Szegedy, "Batch Normalisation: Accelerating Deep Network Training by Reducing Internal Covariate Shift." arXiv, Mar. 02, 2015. Accessed: Nov. 12, 2022. [Online]. Available: http://arxiv.org/abs/1502.03167

[21] C. Seibold, J. Kleesiek, H.P. Schlemmer, and R. Stiefelhagen, “SelfGuided Multiple Instance Learning for Weakly Supervised Thoracic DiseaseClassification and Localizationin Chest Radiographs,” presented at the Proceedings of the Asian Conference on Computer Vision, 2020. Accessed: Sep. 19, 2023. [Online]. Available: https://openaccess.thecvf.com/content/ACCV2020/html/Seibold_SelfGuided_Multiple_Instance_Learning_for_Weakly_Supervised_Thoracic_DiseaseClassification_and_ACCV_2020_paper.html
[22] S. Bhimshetty and A. V. Ikechukwu, “Energyefficient deep Qnetwork: reinforcement learning for efficient routing protocol in wireless internet of things,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 33, no. 2, Art. no. 2, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp971980.
[23] R. Raza et al., “LungEffNet: Lung cancer classification using EfficientNet from CTscan images,” Engineering Applications of Artificial Intelligence, vol. 126, p. 106902, Nov. 2023, doi: 10.1016/j.engappai.2023.106902.
[24] M. Salehi, R. Mohammadi, H. Ghaffari, N. Sadighi, and R. Reiazi, “Automated detection of pneumonia cases using deep transfer learning with paediatric chest Xray images,” BJR, vol. 94, no. 1121, p. 20201263, May 2021, doi: 10.1259/bjr.20201263.
[25] A. V. Ikechukwu and S. Murali, “xAI: An Explainable AI Model for the Diagnosis of COPD from CXR Images,” in 2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS), Dec. 2023, pp. 1–6. doi: 10.1109/ICDDS59137.2023.10434619.
[26] B. Mahaur, K. K. Mishra, and N. Singh, “Improved Residual Network based on normpreservation for visual recognition,” Neural Networks, vol. 157, pp. 305–322, Jan. 2023, doi: 10.1016/j.neunet.2022.10.023.
[27] L. Zhang, Y. Bian, P. Jiang, and F. Zhang, “A Transfer Residual Neural Network Based on ResNet50 for Detection of Steel Surface Defects,” Applied Sciences, vol. 13, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/app13095260.
[28] M. Zhang, M. Xue, S. Li, Y. Zou, and Q. Zhu, “Fusion deep learning approach combining diffuse optical tomography and ultrasound for improving breast cancer classification,” Biomed. Opt. Express, BOE, vol. 14, no. 4, pp. 1636–1646, Apr. 2023, doi: 10.1364/BOE.486292.
[29] K.B. Nguyen, J. Choi, and J.S. Yang, “EUNNet: Efficient UNNormalized Convolution Layer for Stable Training of Deep Residual Networks Without Batch Normalization Layer,” IEEE Access, vol. 11, pp. 76977–76988, 2023, doi: 10.1109/ACCESS.2023.3244072.
[30] A. K. Azad, MahabubAAlahi, I. Ahmed, and M. U. Ahmed, “In Search of an Efficient and Reliable Deep Learning Model for Identification of COVID19 Infection from Chest Xray Images,” Diagnostics, vol. 13, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/diagnostics13030574.
[31] P. Rajpurkar et al., “CheXNet: RadiologistLevel Pneumonia Detection on Chest XRays with Deep Learning,” 2017, doi: 10.48550/ARXIV.1711.05225.
[32] M. Loey, F. Smarandache, and N. E. M. Khalifa, “Within the Lack of Chest COVID19 Xray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning,” Symmetry, vol. 12, no. 4, p. 651, Apr. 2020, doi: 10.3390/sym12040651.
[33] I. D. Apostolopoulos and T. A. Mpesiana, “Covid19: automatic detection from Xray images utilising transfer learning with convolutional neural networks,” Phys Eng Sci Med, vol. 43, no. 2, pp. 635–640, Jun. 2020, doi: 10.1007/s13246020008654.
[34] S. M. Humphries et al., “Deep Learning Enables Automatic Classification of Emphysema Pattern at CT,” Radiology, vol. 294, no. 2, pp. 434–444, Feb. 2020, doi: 10.1148/radiol.2019191022.