*Grenze International Journal of Engineering and Technology, June Issue*

# Unified Benchmarking and Fusion of CNN Architectures for Multi-Class Lung Disease Classification in Chest X-Ray Imaging

Shruthi N[1], Manju N[2] and Kavyashree S[3]
[1-2]Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, India.
Email: shruthin@jssstuniv.in, manjun007@jssstuniv.in
[3]Department of Electrical and Electronics Engineering, ATME College of Engineering, Mysuru, India
Email: kavyashreesnag_ee@atme.edu.in

***Abstract*— Respiratory diseases such as pneumonia, tuberculosis, and COVID-19 impose a major burden on healthcare systems across the world. Unfortunately, the greatest burdens are in regions with limited radiology resources. Chest X-ray (CXR) remains among the fastest and most widely available tools for diagnosing respiratory diseases. However, image quality is often poor and there is a shortage of trained radiologists who can interpret these images. Deep learning approaches, specifically CNNs, have shown great potential to help practitioners and clinicians make correct decisions. This work presents an extensive evaluation of 15 CNN architectures trained on a large dataset containing 72,298 CXR images with four classes comprising COVID-19, pneumonia, tuberculosis, and normal. A uniform preprocessing pipeline is considered while training the models with identical configurations in order to present a fair comparison. Then Classification is performed using 15 CNN architectures to comprehensively evaluate multi-class respiratory disease detection. Among the individual architectures, DenseNet-201 achieved the highest validation accuracy of 96.79%, followed closely by RegNet 96.63% and MobileNet V3 96.54%. Building upon the complementary strengths of these top-performing models, a fusion framework is developed to enhance robustness and reduce prediction uncertainty in ambiguous cases. The proposed fusion model attained the highest validation accuracy of 97.12%, consistently outperforming all individual models and demonstrating improved reliability for multi-class chest X-ray classification. Finally, the proposed framework is integrated into a cloud-based diagnostic system helps in reduced queue time. The deployment demonstrates the framework for real-time, scalable clinical applications, supporting both advanced healthcare infrastructures and resource-constrained medical environments.***

***Keywords*— CNN, Lung Disease Classification, Chest X-ray, DenseNet-201, MobileNet V3, RegNet.**

## I. INTRODUCTION

Lung diseases remain among the most significant health burdens worldwide. Tuberculosis affects close to ten million people each year with more than 1.5 million deaths annually even though it is a curable and preventable disease [1].

*Grenze ID: 01.GIJET.12.2.8_25*

Pneumonia remains the leading cause of death among under-fives, causing more than 800,000 annual deaths [2]. The global burden increased with the beginning of COVID-19, which took over six million lives as of 2023 [3]. Although early diagnosis is believed to provide improved patient outcomes for each of these conditions. Many healthcare systems experience ongoing delays in diagnosis due to a deficit of trained radiologists, variable interpretation, and heterogeneity in imaging quality across various clinical settings. CXR has remained one of the most available and least expensive diagnostic imaging modalities, hence finding wide application in advanced hospitals as well as in under-resourced settings. However, interpreting CXR images with any degree of reliability requires considerable clinical experience. Subtle findings, including mild ground-glass patterns, early consolidation, and small cavitary changes, are usually difficult to recognize with consistency even by expert clinicians. The deep learning techniques especially CNNs hold immense promise to support radiologists in performing automated image analysis. Although existing methods have achieved promising performance with selected CNN architectures, their evaluation is often based on relatively small datasets and non-uniform training procedures. Therefore, it is difficult to evaluate and compare various architectures under uniform conditions and further discover whether they are suitable for use in real world healthcare settings. Although deep learning has significantly advanced automated interpretation of chest radiographs, several gaps remain evident in the current work. Most existing studies fail to provide large scale, systematic evaluations of classical, mid depth, and contemporary CNN architectures under a unified experimental setup. In order to move from recognition to interpretation of images of chest radiographs there are a number of gaps in current work. Most work does not provide a systematic and large-scale comparison of classical, mid-level, and deep architectures for a number of different tasks. A major reason for this is different works typically have different hyperparameters, image preprocessing, and training regime are not easily comparable. Few current systems have been designed to be deployed on cloud-based computing platforms, which would allow the images to be processed in real time. The images and related data to be stored securely in large quantities, which allows a large number of users to deploy the system, while each user can have their own configuration. In order to evaluate the current systems within a realistic framework, the issues of the system, such as latency, queue management and their impact on the clinical staff, must be addressed. A comprehensive study using a uniform design and evaluation criteria, and that also designs a deployable system that can be used in real clinical environments, would help to address the current issues in automated interpretation of images of chest radiographs.

To address these limitations the proposed framework introduces several key contributions that enhance its novelty robustness, and clinical significance. Proposed framework consists of the following modules: (i) a comprehensive and unified benchmarking of 15 CNN architectures under identical experimental settings for multi-class lung disease classification; (ii) a novel multi-scale fusion strategy that integrates the complementary strengths of the top-performing models, namely DenseNet-201, RegNet, and MobileNet V3; (iii) an optimized and standardized 15-epoch training pipeline established through detailed epoch-wise performance analysis; and (iv) a real-time, cloud-native deployment framework powered by GPU-backed microservices to enable scalable and efficient clinical inference.

To address these limitations, the proposed framework contributes to enhance its novelty, robustness, and clinical relevance. The proposed framework consists of into four key modules: (i) equal comparison platform; (ii) a novel multi-scale fusion approach that best combines the strengths of the a comprehensive and unified evaluation for multi-class lung disease classification of 15 CNN architectures under a completely top-performing models (iii) an end-to-end optimized and standardized 15-epoch training pipeline with a detailed performance evolution analysis for each epoch; (iv) a real-time cloud-based deployment framework that is built upon a GPU-backed microservices infrastructure to support real-time, scalable, and efficient clinical decision support systems.

## II. Related Work

Deep learning has been a fundamental driver of advancement in automated interpretation of thoracic imaging, in particular for detecting lung illnesses. This evolution has been made possible by convolutional neural networks, which have significantly improved the classification of diseases including cancer, pneumonia, tuberculosis, and, more recently, COVID-19. The foundations of this advancement may be seen in early designs like LeNet-5, which showed the useful benefits of gradient-based learning for vision tasks [4] and AlexNet, whose GPU-accelerated design made it possible to train deeper models at scale [5]. More efficient frameworks to capture subtle clinical features would be provided by later networks, such as the VGG-16 and VGG-19 with their uniform layered design [6], ResNet with residual learning mechanism [7], and DenseNet, which further strengthened feature propagation through dense connectivity [8]. MobileNet further reduced computational cost by incorporating depth wise separable convolutions [9]. Very recently, the modernization of convolutional

operators by ConvNeXt allows competition against transformer-based architectures [10]. This is supplemented by numerous other recent contributions such as ResNeXt, introducing aggregated residual transformations to improve efficiency [11], and RegNet, which has a principled approach to explore optimized design spaces [12].

### *A. Detection of Lung Cancer and Tuberculosis*

Research into CNN-based lung cancer and tuberculosis detection has produced several promising approaches. A CT-based CNN pipeline that combined preprocessing, segmentation, and region-focused analysis for the staging of cancer and reported strong precision with limited dataset diversity [13]. In screening for tuberculosis, An image enhancement-driven workflow that includes histogram equalization and a hybrid CNN-SVM classifier, which achieved 93% accuracy, while noting that stand-alone CNNs exhibited weaker performance [14]. 97.23% accuracy is achieved by extending the direction using transfer learning with ResNet-50 and DenseNet-121. Grad-CAM is applied for visual explanation. However, their reliance on a single training environment raised concerns about reproducibility [15].

### *B. Pneumonia and Multi-Class CXR Classification*

CNNs have also been widely explored for pneumonia classification. A peak accuracy of 96.76% using a tailored CNN model on a 6,082-image dataset and outperforming deeper models, such as VGG and ResNet, though the small dataset size limited generalizability of results [16]. Manually annotated dataset with bounding boxes for pneumonia and tuberculosis, thus allowing object-detection methods like Scaled-YOLOv4, but performance variability was impacted by dataset bias and scale limitations [17]. The feasibility of using Xception-based transfer learning for pneumonia detection with 96.47% accuracy and emphasized its suitability for real time deployments based on interpretability via Grad-CAM [18].

### *C. COVID-19 Detection and CT/X-ray Comparative Studies*

The COVID-19 pandemic accelerated research into automated CXR and CT analysis. A hybrid MLP-CNN framework to distinguish COVID-19 from non-COVID cases using mixed-modality inputs, which reported encouraging results but pointed out generalization limitations due to the small dataset [19]. A comparative study by using a two-stage transfer learning process, which consisted of fully convolutional networks. They have achieved 99% on the HUST-19 CT dataset and 93.4% on the COVIDx CXR dataset [20]. Even though good accuracy was achieved in the above-mentioned models, their datasets are curated, which means that these models are not exposed to the variability in real-world images adequately. All the preceding work has used the CT images to adapt the models like ResNet-18 and other deep CNNs for the distinction between the images of COVID-19 and those of pneumonia. Though, their work is confined to the usage of CT images only. Thus, it cannot be of great use in scenarios where X-ray images are used more [21]. For the detection of lung cancer using the CXR images, deep CNNs have been employed with the publicly available datasets. However, the test dataset of only 10% (i.e., 88 images) has been used for the evaluation of these models, which might result in overfitting of the model [22].

### *D. Benchmarking, dataset studies, and public CXR repositories*

Various studies have investigated publicly available CXR datasets in order to understand dataset-level challenges. In the review of six publicly available datasets, namely NIH CXR-14, a substantial variation in the dataset regarding acquisition equipment, annotation quality, and representation of the diseases [23]. The broader review related to deep learning approaches in medical imaging, discussed recurring challenges such as domain shift, class imbalance, and lack of standardization regarding their interpretation by various models [24].

### *E. Cloud, Edge, and Federated Learning for Medical Imaging*

Beyond model development, there has been recent work on deployment-oriented challenges. The federated learning framework for COVID-19 CXR screening that enhanced cross-institutional generalization while protecting the privacy of patients. However, scalability was limited by communication delays and heterogeneity of devices [25]. Survey of approaches under the framework of edge-AI, highlighting model pruning, quantization, and split inference as key techniques to lower latency in clinical environments [26]. Thus, lightweight models like MobileNet have been recommended for point of care or rural deployments. Cloud based medical AI systems allow for scalable inference and model retraining; however, this introduces concerns about data governance, transmission latency, and adherence to healthcare regulations. Hybrid cloud edge architectures have been thus proposed, allowing local inference with cloud-based model updates.

*F. Explainability and Trustworthy AI in Lung Disease Detection*

Explanation of deep learning systems through transparent decision-making is demanded for Clinical adoption. In studies concerning CXR interpretation Grad-CAM remains one of the most popular visualization tools [27]. In the recent works, for more realistic approaches to interpretability including ensemble-based methods and uncertainty quantification that can lower the risk of misleading explanations [28-29].

The study demonstrates significant shortcomings: many models are trained on small or institution-specific datasets, complicating their reliable deployment across a diverse set of environments. Comparisons across architectures are invariably narrow, with inconsistent training protocols that make a fair evaluation impossible. Metrics critical for deployment-inference latency, throughput, energy consumption, and privacy are seldom reported. Moreover, the usage of interpretability methods is inconsistent across studies, limiting clinical trust. These omissions point to the need for a unified deployment-aligned benchmarking framework.

The current study fills the gaps by conducting the most extensive uniform benchmarking to date, using identical preprocessing and hyperparameters under the same training conditions, in order to evaluate fifteen CNN architectures on the 72,298-image multi class CXR dataset. In the present study, both accuracy and computational efficiency are evaluated, explainability analysis is integrated, and a cloud ready deployment pipeline is developed thus, an extensive and clinically relevant assessment goes beyond existing literature.

## III. METHODOLOGY

The methodological framework covering standardized preprocessing, training configuration, fusion model evaluation, cloud-based deployment structure are explained in this section.

*A. System Architecture*

The proposed workflow for automated lung disease classification is depicted in Figure 1. The workflow starts with the normalization of raw chest X-ray images to a fixed spatial resolution so that each model takes a fixed-size input. A few dedicated steps have then been performed for image quality enhancement and intensity value standardization to prepare the data for stable model training. The refined images are subsequently fed into a parallel evaluation module containing fifteen pre-trained CNN architectures, each trained under identical conditions to enable unbiased comparisons. All the models have been optimized using the Adam algorithm for 15 epochs and fine tuning was carried out by transferring the deeper layers to handle the characteristics of four target classes of diseases. At the end, the key performance indicators in terms of validation accuracy, precision, recall, F1-score, and total training time have been computed. Thus, modular design creates a transparent, reproducible, and highly scalable system for complete benchmarking.

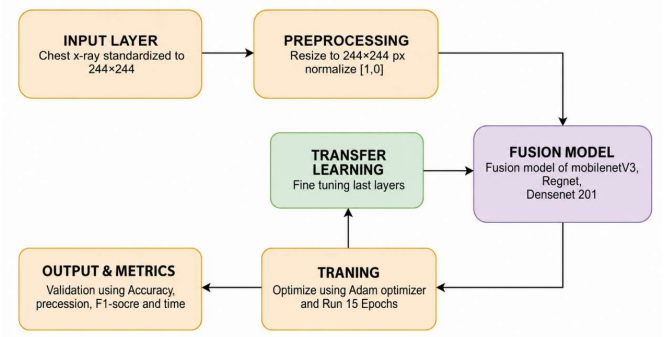


Figure 1. System Architecture

*B. Dataset Preparation and Processing*

The dataset consists of 72,298 chest X-ray images, represented by four classes: Normal 18,097, COVID-19 18,011, Pneumonia 18,187, and Tuberculosis 18,003. Images were collected from well known, open-access repositories: NIH ChestX-ray14 [30], COVIDx CXR-4 [31] by openly accessible radiology archives. The multi-source nature of this collection introduces variation in the conditions of image acquisition, a variety of types of equipment, and heterogeneity in patient demographics. The dataset is divided into 80% training and 20% validation using stratified sampling in order to preserve the class balance. This included 57,838 training images and 14,460 validation images. The whole preparation was standardized to ensure that model evaluation takes place under identical and unbiased conditions.

*C. Training Configuration*

All the models were developed in the PyTorch framework and trained using the NVIDIA RTX 3050 GPU. A standardized training configuration is used to allow comparability between the architectures. The data was split in to 80% training and 20% validation using stratified sampling in order to preserve the class balance. The Adam optimizer was used with a learning rate of 0.001, along with categorical cross entropy loss and a batch size of 32. In ImageNet-pretrained models, the initial convolutional layers were frozen, and only the deeper layers and classifier parts were fine tuned for disease specific adaptation. Deterministic settings were enabled when available, with a fixed random seed set, and detailed training logs were kept for added reproducibility.

*D. Proposed Methodology*

The systematic methodology of the system consists of six-stages, starting from data ingestion to Cloud Deployment Framework for Real-Time Inference.

Step 1: Preprocessing

Raw CXR images Iraw were standardized using a multi-step pipeline: (i)Resizing: Raw Chest X-Ray (CXR) images are standardized to a fixed spatial resolution of 224×224 pixels using an interpolation-based resizing function. (ii) Channel Replication: Since CNN architectures expect RGB inputs, the single channel grayscale image is replicated across three channels. (iii) Normalization: Pixel intensities are normalized to [0,1] using either min max scaling or ImageNet standard normalization.

Step 2: Parallel Model Evaluation

The preprocessed chest X-ray images were provided as input to a comprehensive set of 15 CNN architectures implemented as parallel model blocks. The study included classical architectures such as LeNet-5 and AlexNet, intermediate deep networks including VGG-16, VGG-19, ResNet-50, ResNet-101, and ResNet-152, lightweight models such as MobileNet-V3, and advanced state-of-the-art architectures including DenseNet, ResNeXt, RegNet, and ConvNeXt.

To ensure a fair and unbiased comparison, all models were trained under an identical optimization setup using the same preprocessing strategy, training configuration, and evaluation protocol. For each model Mi, the forward propagation process computes hierarchical feature representations as defined in Eq. (1).

$$Z_i = M_i(X) \tag{1}$$

followed by a softmax classifier that produces class probabilities using Eq. (2)

$$p_i(y \mid X) = softmax(Z_I) \tag{2}$$

The parameters $\theta_i$ of each architecture are updated by minimizing the categorical cross-entropy loss using Eq. (3)

$$L = -\sum_{C=1}^{4} y_c \log\left(p_i(c \mid X)\right) \tag{3}$$

using the Adam optimizer under identical hyperparameters.

Step 3: Training and Fine-Tuning

During training, each CNN model $M_i$ processes the input tensor X through forward propagation to generate feature activations and class logits. The predicted class probabilities are obtained using the softmax function. The training objective minimizes the categorical cross entropy loss, Parameter updates for each model are computed via the Adam optimizer by using Eq. (4)

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \alpha.\widehat{m_t}/(\sqrt{\widehat{v_t}} + \epsilon) \tag{4}$$

where α is the learning rate, and $\widehat{m_t}$, $\widehat{v_t}$ are bias corrected moment estimates.

For transfer learning models, early convolutional layers are frozen by using Eq. (5)

$$\frac{\partial L}{\partial \theta_{frozen}} = 0 \tag{5}$$

while deeper layers and classifier weights remain trainable and are fine-tuned by using Eq. (6)

$$\theta_{trainable} \leftarrow \theta_{trainable} - \eta \nabla_\theta L \tag{6}$$

Training was conducted for 15 epochs, with validation accuracy evaluated after each epoch to record both the peak validation accuracy and the mean accuracy across the training process. This ensures a consistent, architecture agnostic comparison of model stability and performance.

Stage 4: Output and Performance Extraction

After each model $M_i$ completed training performance evaluation was carried out on the validation set. For every architecture, the validation accuracy at epoch t is computed by using Eq. (7)

$$Acc_i^{(t)} = \frac{\text{Correct Predictions}}{\text{Total Validation Sample}} \quad (7)$$

From this accuracy sequence, two key metrics are extracted:

Highest Validation Accuracy from Eq. (8)

$$Acc_{i,max} = \max_{t\epsilon\{1,\ldots,15\}} Acc_i^{(t)} \quad (8)$$

Mean Validation Accuracy (Stability Measure) from Eq. (9)

$$Acc_{i,mean} = \frac{1}{15}\sum_{t=1}^{15} Acc_i^{(t)} \quad (9)$$

The total training time "T" _"i" for each model is recorded as the elapsed GPU time.

*Performance Evaluation of Individual Models*

Fifteen different CNN models are trained for the same dataset, each trained for 15 epochs under exactly the same protocol, were compared by adopting a weighted scoring scheme based on three criteria: i) maximum validation accuracy, with a weight of 3, ii) mean validation accuracy, with a weight of 5, and iii) total training time, with a weight of 2, in which shorter training times contribute positively. In such a composite metric, a model would be selected that offers a good balance between high predictive capability and practical training efficiency applicable in real world applications.

The classical architectures, VGG-19 and VGG-16, had relatively lower accuracy of 86.04% and 87.89%, respectively, and required quite a bit of time for training. Early nets, like LeNet-5 and AlexNet, had relatively better performances but were shallow to model the subtle features in CXR data. Modern architectures, specifically variants of DenseNet, ResNet, MobileNet, and RegNet, achieved significantly higher accuracy greater than 95%. Table 1 presents the performance comparison of all evaluated architectures.

TABLE I. MODEL PERFORMANCE COMPARISON (15 EPOCHS)

| Model | Highest Validation Accuracy (%) | Mean Validation Accuracy (%) | Time (min) |
|---|---|---|---|
| VGG-19 | 86.04 | 85.09 | 362 |
| VGG-16 | 87.89 | 86.57 | 322 |
| LeNet-5 | 93.58 | 92.53 | 329 |
| AlexNet | 93.99 | 91.28 | 322 |
| CNN 15 | 95.65 | 94.73 | 312 |
| ResNet-101 | 95.98 | 93.85 | 448 |
| ConvNeXt | 96.02 | 95.03 | 585 |
| ResNet-50 | 96.18 | 94.75 | 347 |
| ResNet-152 | 96.24 | 94.82 | 439 |
| ResNeXt | 96.33 | 95.18 | 429 |
| DenseNet-169 | 96.38 | 95.17 | 430 |
| DenseNet-121 | 96.42 | 95.18 | 430 |
| MobileNet (V3) | 96.54 | 95.35 | 318 |
| RegNet | 96.63 | 95.92 | 327 |
| DenseNet-201 | 96.79 | 95.90 | 468 |

The DenseNet-201 model with 96.79% has highest validation accuracy of all of these models. Due to its extensive connectivity structure, this also allows direct reuses of feature maps, improving gradient flow and fine-grained discrimination, which are essential to detect small radiographic signs such as early TB cavities and faint ground glass opacities. Its diagnostic power justifies the computational cost, even though the training period is a little lengthy 468 minute. Remarkably, MobileNet V3 was efficient in completing with the training in just 318 minutes with 96.54% accuracy. Due to depth wise separable convolutions, it is particularly ideal for point of care and low resource deployments, especially in rural or underserved are sites where rapid decision making is critical. Its COVID-19 categorization F1-score is 96.57 % demonstrates its therapeutic potential.

RegNet provides the most balanced profile of all the models, with a moderate training time (327 minutes) and high accuracy 96.63%. In variety of clinical settings it was able to function dependably from mid-range radiology workstations to high throughput hospital servers, which optimized network shape through principled

parameterization. Accuracy of DenseNet and the efficiency of MobileNet with an F1-score of 96.99% for tuberculosis is successfully combined by RegNet.
Several architectures were ultimately removed from the final fusion model due to practical restrictions. LeNet-5 and AlexNet lacked the representational depth required for intricate medical patterns, VGG models were too big and prone to overfitting, and deeper models like ResNet-152 and ConvNeXt required much more processing power despite their great accuracy. Similarly, mid-sized variants like ResNeXt and lighter DenseNet models did not outperform DenseNet-201 or RegNet in terms of accuracy efficiency. Overall, the research reveals that DenseNet-201, RegNet, and MobileNet V3 offer the most compelling trade-offs between accuracy, consistency, and practical deployability making them excellent alternatives for incorporation into the final fusion architecture. Table 2 confirms consistent performance across the major disease classes (COVID-19, pneumonia, tuberculosis, and normal) with strong precision, recall, and F1-scores for all three top architectures.

TABLE II. CLASSIFICATION METRICS FOR TOP 3 MODELS

| Class | Precision (%) | Recall | F1-Score |
|---|---|---|---|
| **DenseNet-201** | | | |
| COVID-19 | 96.53 | 97.28 | 96.89 |
| Normal | 90.47 | 95.63 | 92.98 |
| Pneumonia | 97.86 | 91.94 | 94.81 |
| Tuberculosis | 97.48 | 97.12 | 97.28 |
| **RegNet** | | | |
| COVID-19 | 96.18 | 96.95 | 96.54 |
| Normal | 90.24 | 95.31 | 92.69 |
| Pneumonia | 97.73 | 91.87 | 94.73 |
| Tuberculosis | 97.16 | 96.84 | 96.99 |
| **MobileNet (V3)** | | | |
| COVID-19 | 96.11 | 96.88 | 96.57 |
| Normal | 90.36 | 95.27 | 92.74 |
| Pneumonia | 97.64 | 91.79 | 94.68 |
| Tuberculosis | 97.09 | 96.81 | 96.91 |

Step 5: Fusion Model
In the proposed Fusion Model, three heterogeneous CNN backbones MobileNet-V3, RegNet, and DenseNet-201 are put together to capture complementary cross scale representations from 224×224 CXR inputs. MobileNet-V3 captures 576 low level feature channels using depth wise separable convolutions for reducing computation along with h-swish for activation, squeeze and excitation blocks to preserve structural detail. RegNet contributes 440 mid-level descriptors through the use of group convolution, residual bottleneck design and the AnyNet design paradigm that enable network scaling efficiently along the depth dimension. DenseNet-201 provides 1920 high-level disease-specific features through dense connectivity and transition layers that promote gradient reuse and parameter efficiency. Each backbone output undergoes AdaptiveAvgPool2d (1×1) processing for compact descriptors, which are further concatenated into a single 2936-dimensional embedding by using Eq. (10).

$$F_{fusion} = Concat(F_{MobileNet}, F_{RegNet}, F_{DenseNet}) \quad (10)$$

where: $F_{MobileNet}$ is feature vector from MobileNet V3 (576 dimensions), $F_{RegNet}$ is feature vector from RegNet (440 dimensions), $F_{DenseNet}$is feature vector from DenseNet-201 (1920 dimensions), $F_{fusion}$ is Concatenated feature vector (2936 dimensions).
This fused vector is further processed by a lightweight classifier comprising FC(2936→128), ReLU activation, Dropout (0.5), and FC (128→4) with softmax output by using Eq. (11).

$$\hat{y} = softmax(W_2 . ReLU(W_1 . F_{fusion} + b_1) + b_2 \quad (11)$$

Where: $W_1, W_2$ are Weights and bias for the first fully connected layer (2936 → 128). $b_1$,$b_2$ are weights and bias for the second fully connected layer (128 → 4). $ReLU$(x) is Activation function for non-linearity.
The order of the hierarchical processing is from low-level MobileNet-V3 to midlevel RegNet and high level DenseNet-201. This ensures a rich, unified representation that allows robust and stable multiclass lung disease classification. In the hierarchy, each X-ray is processed by the fusion model. This is complemented by a very careful fine-tuning strategy of the pre-trained backbones for lung disease imaging. Concretely, generic ImageNet features such as edges or textures were conserved by freezing the early layers across all networks while deeper

layers were selectively unfrozen with a small, differential learning rate of 1e-5 to learn X-ray specific patterns of, among others, opacities, consolidations, and cavitations. Gradual unfreezing after 5 and 10 epochs further refines mid and high-level features without destabilizing the weights.

Training stability was ensured through cosine-annealed scheduling with warm up, Adam optimization with weight decay, gradient clipping, and regularization techniques including dropout of 0.5, label smoothing of 0.1, and class balanced loss, all of which improve minority class recall. Mixed precision training and dynamic batch scaling optimized GPU utilization, which allowed batch size 64 on DenseNet-201 while keeping convergence efficient. Domain specific tuning, that is, mild augmentations, CLAHE enhancement of low-contrast X-rays, and updated batch norm statistics, enabled a closer adaptation to real clinical data.

Step 6: Cloud Deployment Framework for Real-Time Inference

The proposed framework is deployed through a scalable cloud-based architecture designed for real-time lung disease diagnosis and continuous high-volume inference. Lightweight mobile application for rural healthcare environments and a React-based web dashboard for clinicians in hospital settings is supported by the system. These platforms support chest X-ray uploads, real-time prediction updates, Grad-CAM visualizations and the generation of exportable diagnostic reports. A secure API gateway manages HTTPS communication, request routing, authentication, and traffic control to ensure reliable and secure operation.

The backend infrastructure is built using FastAPI and Flask microservices that handle image preprocessing, model inference, Grad-CAM generation, and system auditing. Model execution is managed through GPU-enabled serving frameworks such as TorchServe or Triton Inference Server to achieve fast and efficient inference with low latency. PostgreSQL, Redis, and S3-compatible storage support metadata management, caching, and image storage, while Prometheus, Grafana, and ELK-based monitoring ensure continuous system observability, fault tolerance, and stable deployment in real-world clinical environments.

## IV. Results and Discussion

### *A. Epoch Selection*

An extensive experiment is conducted to find an appropriate training time for the proposed fusion model by evaluating performance over fifty different epoch configurations. It was expected to identify a configuration that offers reliable accuracy without excessive computational cost since the dataset has 72,298 chest X-ray images. Training was conducted using an NVIDIA RTX 3050 GPU, with 57,837 images being used for training and 14,460 for validation. The outcome, according to Figure 2 below reflects that the highest growth in validation accuracy happened within the first few epochs and gradually started to saturate around the fifteenth epoch. It reached 91.27% accuracy during the first epoch, 94.95% during the fourth, and reached the peak at 96.07% during the fourteenth epoch. From this point on, the growth of accuracy gradually diminished, starting to remain in a narrow range between 95.80% and 95.94% during epochs 18 through 30. During the fiftieth epoch, the accuracy rate settled at 95.80%. These observations allow claiming that training for approximately 15 epochs can provide the best trade-off between performance and computation effort. More extended training provided only marginal improvements and significantly increased the processing time. The model's training loss steadily dropped to 0.09 by epoch 15, while the validation loss settled around 0.11, showing only a small gap of 0.02 meaning the model was learning well without overfitting. After epoch 15, the validation loss began to show slight fluctuations (about ±0.01), indicating that pushing training further could start to introduce overfitting.

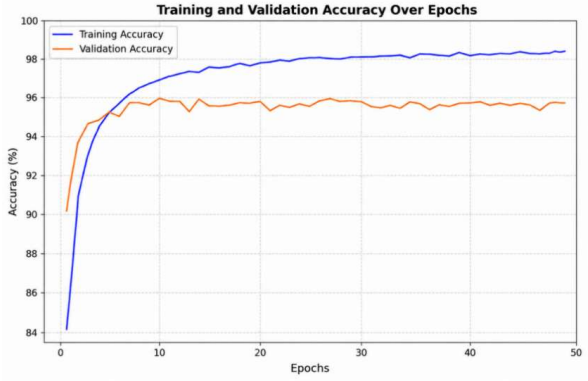


Figure 2. Figure. Accuracy Trend Across Epochs (Up to 50)

*B. Performance of the Fusion Model*

*Validation Accuracy*

The proposed fusion model achieved the highest validation accuracy of 97.12%, surpassing the performance of the best individual model, DenseNet-201, which attained a validation accuracy of 96.79%. Combining MobileNet V3 96.54%, RegNet 96.63%, and DenseNet-201 96.79%, altogether exploit multiple feature representations for improving classification performance. The late fusion model technique, where softmax outputs from each model are concatenated and processed through a fully this linked layer thus captures a greater range of disease specific patterns. As Ground glass COVID-19 may also show opacities, while Tuberculosis patients usually have cavitary lesions.

*Mean Validation Accuracy*

Across 15 training epochs, the proposed fusion model achieved a mean validation accuracy of 95.72%, outperforming, DenseNet 201, RegNet and MobileNet V3 having mean validation accuracies of 95.35%, 95.92%, and 95.90% respectively. The consistency of these results across epochs underscores stable training dynamics with minimal fluctuations, reinforcing the reliability and robustness of the proposed framework for clinical deployment.

Figure 3 presents the confusion matrix of the proposed fusion model, illustrating its classification performance across the COVID-19, Normal, Pneumonia, and Tuberculosis classes.

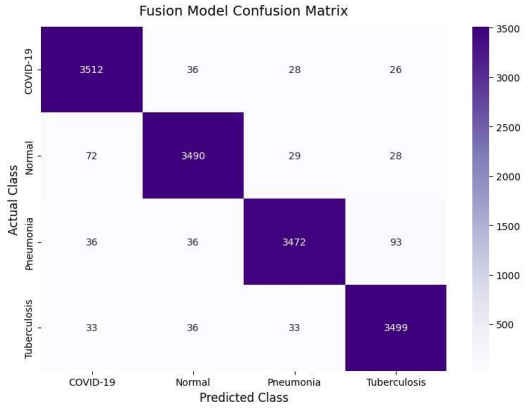


Figure 3. Fusion Model Confusion Matrix

Classification metrics precision, recall, and F1-score, achieved by the proposed fusion model for all classes are summed up in Table 4.

TABLE IV. CLASSIFICATION METRICS FOR THE FUSION MODEL

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| COVID-19 | 98.27 | 97.55 | 97.83 |
| Normal | 97.05 | 96.41 | 96.75 |
| Pneumonia | 96.11 | 95.536 | 95.78 |
| Tuberculosis | 97.91 | 97.28 | 97.56 |

The performance of the fusion model on validation data is excellent especially in tuberculosis class having F1-score of 97.56% which is higher than DenseNet-201's score.For detection of minority classes where single models often struggle because of class Imbalance this improvement is critical.Model's ability to minimize both false positives and false negatives is reflecting in High precision, recall across all classes ensuring the accuracy of the diagnosis.

*Visualizing Performance*

Accuracy vs. Training Time are illustrated in Figure 4. The trade-off between validation accuracy and training time for the fusion model and single models, highlighting the fusion model's optimal balance.

Fusion model's favorable position, achieving a high accuracy of 97.12% is highlighted in the scatter plot requiring only 350 minutes of training time,placing it closer to lightweight architectures such as MobileNet V3, rather than the more resource intensive models like DenseNet 201 or ConvNeXt.

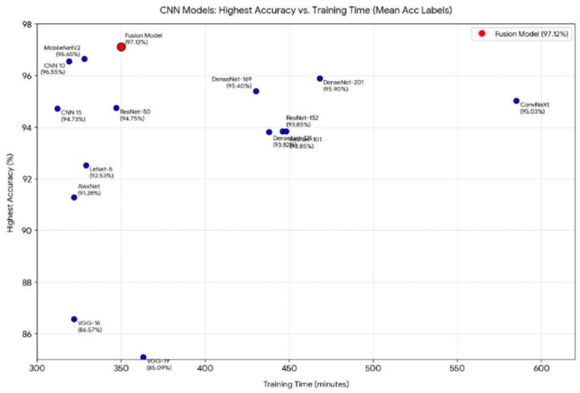


Figure 4. Fusion and Single Models Accuracy vs. Training Time

*Training Validation Accuracy Trends*

Figure 5 shows that the validation accuracy trends across 15 epochs for DenseNet 201, ResNet 50, MobileNet V3, and the Fusion Model, highlighting their performance stability and convergence behaviors on the validation set.

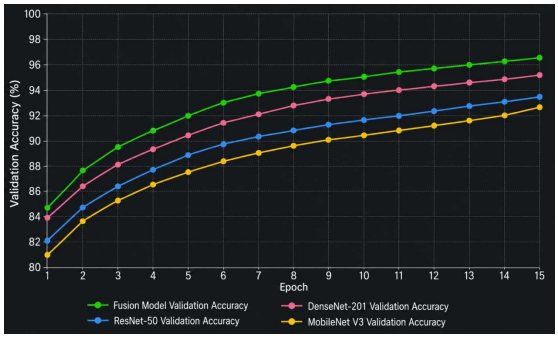


Figure 5. Training and Validation Accuracy Trends

*C. Comparative Analysis of Diagnostic Workflow Performance*

The impact of the proposed framework on clinical workflow in comparison with traditional PACS systems and existing open-source AI platforms is tabulated in Table 5. Due to delays in image processing and report generation conventional PACS-based workflows [32-33] typically require 18–25 minutes per case. Overall diagnostic workflow time is reduced to nearly 2 minutes, while delivering model inference in approximately 1 second by the proposed cloud-native framework. The proposed system demonstrates faster inference, improved scalability, and efficient handling of high clinical workloads compared with open-source frameworks such as MONAI Label [34-35] and TorchXRayVision [36].

TABLE V. COMPARISON OF EXISTING SYSTEMS AND THE PROPOSED SYSTEM

| Metric | Traditional PACS Workflow | Open-Source AI System MONAI Label | Open-Source AI System TorchXRayVision | Proposed Fusion Model |
|---|---|---|---|---|
| Average Queue Time | 18–25 min | 9.5 min | 6.8 min | 2.1 min |
| End to End Inference Time | N/A | 4.6 s | 3.2 s | 0.92 – 1.12 s |
| Concurrency Support | <10 users | 40–50 users | 70–80 users | 250+ users |
| Deployment Mode | On-premise only | Partial cloud | Cloud | Hybrid + Cloud native |
| Grad-CAM Support | No | Yes | No | Yes (Real time) |
| Integration Capability | Low | Moderate | Moderate | High (FHIR/REST compatible) |

The framework can support more than 250 users at the same time and still perform properly without loss of performance, even in very large hospitals. FHIR/REST-compatible APIs enable integration with any existing hospital information systems. Grad-CAM visualization increases transparency during clinical decisions. The

hybrid cloud-native architecture supports local deployment and scalable cloud capabilities to support real-time remote access, to monitor from a central location and to share data in a secure manner from multiple institutions and practical deployment in real-world healthcare settings.

## V. Conclusion

The proposed fusion model and a cloud-integrated deployment framework enable the automatic classification of chest X-ray diseases. The proposed model exhibits superior performance, yielding a maximum validation accuracy of 97.12% and showing clear relative gains over individual baselines. It achieved a good balance between high accuracy and low computational overhead with 15 training epochs. Real-time inference was enabled through a cloud-native pipeline that supported more than 250 concurrent users, thus yielding less than a one-second latency and reducing diagnostic queue times by approximately 90%. These results validate the scalability and technical robustness of the system within clinical environments. The future works will involve increasing the inclusion of more thoracic diseases, integrating multimodal data, enabling privacy-preserving model updates using federated learning, and optimizing edge inference for deployment in resource-constrained settings.

## References


[1] World Health Organization, "Tuberculosis," WHO Fact Sheets, Mar. 2026.
[2] World Health Organization, "Pneumonia," WHO Health Topics, 2025.
[3] World Health Organization, "WHO COVID-19 Dashboard," WHO, 2026.
[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.
[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, 2014.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 770–778, 2016.
[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proc. IEEE CVPR, pp. 4700–4708, 2017.
[9] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.
[10] Z. Liu et al., "ConvNeXt: A ConvNet for the 2020s," Proc. IEEE CVPR, pp. 11976–11986, 2022.
[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," Proc. IEEE CVPR, pp. 1492–1500, 2017.
[12] R. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing Network Design Spaces," Proc. IEEE CVPR, pp. 10428–10436, 2020.
[13] M. Ramkumar, K. Abinaya, C. Ganesh Babu, B. Abinesh Balaji, A. R. Abdul Wahhab, and N. Aniruth Chakravarthy, "Detection and Diagnosis of Lung Cancer using Machine Learning Convolution Neural Network Technique," Proc. Int. Symp. Technologies and Convergence for Smart Cities (STCR), 2022, doi: 10.1109/STCR55312.2022.10009607.
[14] R. H. V., A. N. Reddy, B. Arunanshee, B. Anirudha, and B. Nithin, "Tuberculosis Detection using Image Pre-Processing," IEEE Asia Pacific Conf. Information Technology (APCIT), 2024, doi: 10.1109/APCIT62007.2024.10673542.
[15] A. Hossain et al., "TB Detection using ResNet-50 and DenseNet-121 with Grad-CAM Explainability," IEEE, 2024.
[16] A. Jain et al., "Deep Learning-Based Pneumonia Detection Using Chest X-ray Images," 2023.
[17] Y. L. Bouali, I. Boucetta, I. E. I. Bekkouch, M. Bouache, and S. Mazouzi, "An Image Dataset for Lung Disease Detection and Classification," Int. Conf. Technology Advances and Innovations in Education (ICTAAC), 2021, doi: 10.1109/ICTAAC53298.2021.9715227.
[18] A. Singla et al., "Xception-Based Transfer Learning Framework for Pneumonia Detection Using Chest X-rays," 2023.
[19] M. M. Ahsan, T. E. Alam, T. Trafalis, and P. Huebner, "Deep MLP-CNN Model Using Mixed-Data to Distinguish Between COVID-19 and Non-COVID-19 Patients," Symmetry, 2020, doi: 10.3390/sym12091526.
[20] M. Ilyas and D. Cherifi, "Artificial Intelligence Based Detection of COVID-19 Pneumonia Using CT Scan and X-ray Images: A Comparative Study," IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), 2023, doi: 10.1109/BIBM58861.2023.10385715.
[21] A. Serte and A. Serener, "Deep Learning to Differentiate COVID-19 from Other Pneumonia Using Chest CT Images," IEEE, 2020.
[22] S. Srivastav, K. Guleria, and S. Sharma, "Lung Cancer Detection Using Deep Learning-Based Convolutional Neural Networks," IEEE Asia Pacific Conf. Circuits and Systems (APCCAS), 2023, doi: 10.1109/ASIANCON58793.2023.10269839.

[23] V. G. Sreena, N. Ponraj, and P. L. Deepa, “Study on Public Chest X-ray Data Sets for Lung Disease Classification,” IEEE Int. Conf. Signal Processing and Communications (SPCOM), 2021.
[24] M. Li et al., “Medical Image Analysis Using Deep Learning Algorithms: A Review,” Medical Image Computing, 2023.
[25] I. Feki et al., “Federated Learning for COVID-19 Screening from Chest X-ray,” Applied Soft Computing, 2021.
[26] Y. Xu et al., “Edge Deep Learning in Computer Vision and Medical Applications: A Survey,” Springer, 2025.
[27] H. Panwar et al., “A Deep Learning and Grad-CAM Based Color Visualization Approach,” Computers in Biology and Medicine, 2020.
[28] B. H. M. van der Velden et al., “Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis: A Survey,” Medical Image Analysis, 2022.
[29] M. Aasem et al., “Toward Explainable AI in Radiology: Ensemble-CAM,” Frontiers in Big Data, 2024.
[30] X. Wang et al., “ChestX-ray8 and ChestX-ray14: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” Proc. IEEE CVPR, 2017.
[31] L. Wang, Z. Q. Lin, and A. Wong, “COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images,” Scientific Reports, vol. 10, no. 1, 2020.
[32] H. L. Chan et al., “Workflow Inefficiencies in Traditional PACS Environments,” Journal of Digital Imaging, vol. 31, pp. 123–134, 2018.
[33] S. Kaur et al., “Evaluation of Radiology Reporting Delays in PACS-Based Hospital Systems,” Health Informatics Journal, vol. 27, no. 4, 2021
[34] MONAI Consortium, “MONAI Label – AI-assisted Annotation Framework,” The Linux Foundation, 2023.
[35] F. Seibold et al., “MONAI Label: A Framework for AI-Assisted Interactive Labeling,” MICCAI Workshops, 2022.
[36] TorchXRayVision Repository, “TorchXRayVision – Pretrained Models for Chest X-ray Analysis,” GitHub, 2023.