A Diagnostic Pipeline for Multi-Class Classification of Chest X-Ray Images Using Knowledge Distillation and Semi-Supervised Segmentation

Ristian Uddin
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
ristian.uddin@northsouth.edu

Mohammod Abdullah Bin Hossain
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
bin.abdullah@northsouth.edu

Mohammed Nahid Hossain
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mohammed.hossain09@northsouth.edu

Tashfia Kashem Chowdhury
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
tashfia.chowdhury02@northsouth.edu

Shahnewaz Siddique
Associate Professor, Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
shahnewaz.siddique@northsouth.edu

***Abstract*—Artificial Intelligence (AI)-powered medical diagnosis systems can play a vital role in healthcare by enabling early detection of COVID-19, pneumonia, and tuberculosis. While chest X-ray diagnostics continue to be an essential tool in clinical assessment, timely identification is crucial as it facilitates prompt medical intervention. In this research, we present a novel AI-driven diagnostic system that combines three operational components: applying knowledge distillation methods with semi-supervised segmentation procedures and multi-class classification features to achieve better accuracy rates and operational effectiveness. Our research employed 20,000 four-class chest x-ray images for evaluation following preprocessing operations involving downsampling and augmented techniques. Using semi-supervised segmentation, the UNet model achieved a Dice score of 98% and an IoU of 97%, effectively isolating lung regions to create a refined dataset. We then trained both subsets of the dataset - segmented and original - on CNN models such as InceptionV3, VGG16, ResNet-50, DenseNet-121, and their Attention enhanced variants, followed by ensembled CNNs and CLIP-ViT-L/14. Of these, the ViT model demonstrated the highest score through training on segmented data at 97% accuracy, establishing it as the most suitable teacher model for knowledge distillation. This led to CLIP-KDViT, which distilled knowledge into the MobileNetV2 student model, achieving 99% accuracy on segmented images and classifying 3,000 unseen test images with 97.2% accuracy and a 2.8% misclassification rate. The optimized pipeline was then integrated into our web application, *Respire Check*, which incorporates Grad-CAM for explainability, demonstrating the full functionality of our AI-driven X-ray diagnosis system.**

***Index Terms*—UNet, Semi-Supervised, Dice, IoU, Ensemble, CLIP-KDViT, Knowledge Distillation (KD), Streamlit**

I. Introduction

The three major global health threats that are the leading causes of death worldwide are COVID-19, pneumonia, and tuberculosis (TB). Millions lose their lives each year to these diseases and posing substantial risks to vulnerable populations [1]. Recent health statistics demonstrate the critical need for advanced diagnostic methods because tuberculosis rates have risen since 2020, with annual fatalities reaching 1.5 million [2], pneumonia cases among 5 to 14-year patients increased by 73% [3], and COVID-19 caused 7 million deaths across 229 countries [4]. Despite these alarming figures, X-ray imaging stands as one of the primary diagnostic tools for respiratory conditions. However, its effectiveness relies on radiologists' expertise, and the complexity of interpreting diverse disease patterns can lead to long examination times and potential misdiagnoses [6]. This is where AI-driven technology demonstrates great potential to automate disease detection [5]. Due to issues with unequal class distributions and illness feature similarities, many deep learning models now used for lung disease classification require large amounts of labeled data to function well enough, albeit consuming substantial computational resources [7]. So, we have devised an approach where we deploy a model that improves classification accuracy through the segmentation process, as it solely allows the model to focus on the regions of interest [8] and then make the model efficient enough to make it globally accessible on any devices using knowledge distillation [9], so that people from rural and underserved areas can also use it through their smartphones, bridging the gap in healthcare accessibility and empowering communities with limited resources.

II. Related Works

A. Disease Detection and Classification Models

Medical image classification has experienced a revolutionary shift due to deep learning integration in medical fields. [10] led to a diagnosis accuracy of 98.05% for COVID-19 along with pneumonia and lung cancer. The research team of Ahmed et al. [11] achieved 98.72% success in pneumonia and tuberculosis diagnosis when working with limited medical resources. The clinical user interface from Narayana et al. [12] united VGG16 and SMOTE into one detection system

to identify eight lung diseases with 96.42% accuracy.Through an analysis of 145,202 images, Wang et al. [13] successfully classified pneumonia into viral, non-viral, and COVID-19 types at radiologist standards. The study by Kulkarni et al. [14] evaluated CNNs to determine their performance through AUC measures that reached 0.95 for COVID-19, 0.99 for TB, and 0.98 for pneumonia detection. The researchers at Muthaki et al. [15] established a diagnostic model ensemble that delivered 98.37% accuracy for thoracic disease detection.

B. Advanced models and optimization techniques

Deep learning recent developments improved the accuracy levels of detecting chest diseases through CXR examinations. The TB detection performance of DenseNet201 utilizing segmented lungs reached 98.6% according to Rahman et al. [16] while surpassing the 96.47% accuracy of ChexNet. According to Mamalakis et al. [17] DenResCov-19 accomplished 99.60% AUC-ROC performance by integrating DenseNet-121 with ResNet-50. DeepX-Ray by Chakraborty et al. [18] merged ResNet50-UNet to achieve 100% accuracy and 96.19% IoU. Hadhoud et al. [19] developed a TB detection system by uniting ResNet-50 with ViT-b16 which resulted in 98.97% accuracy. The research conducted by Chen et al. [20] applied modifications to ViT which yielded superior results than traditional CNN models with four-class recognition reaching 95.79% and three-class recognition at 99.57%. The union of VGG16 and VGG19 with attention modules for TB detection according to Kebache et al. [21] resulted in 99.78% accuracy.

C. Knowledge Distillation, segmentation, interpretability and Explainable AI

Medical imaging processes have improved through knowledge distillation and advanced segmentation models that allow efficient diagnosis of COVID-19 and pneumonia and tuberculosis. Real-time mobile and cloud diagnoses were measured at 97% accuracy according to Kabir et al. [22] BabaAhmadi et al. [23] derived knowledge from VGG19 and ResNet50V2 to apply it in MobileNetV2 with a precision rate of 98.8%. Akhter et al. [24] built MLCAK utilizing ViT for making low-resolution CXR image classifications. The integration of Xception with UNet and UNet+ achieved 97.45% accuracy according to Nillmani et al. [25]. Ou et al. [26] implemented a U-Net ensemble to segment TB lesions while reporting 1.0 accuracy and 0.70 IoU. The research team of Panwar et al. [27] managed to detect COVID-19 within two seconds through their implementation of transfer learning and Grad-CAM method.

III. Methodology

A. System Diagram

The overall workflow of our proposed approach is illustrated in Fig. 1. All the stages will be explained in the following sections.

B. Dataset Acquisition

We leveraged multiple datasets to ensure our segmentation approaches are robust across diverse medical imaging domains.

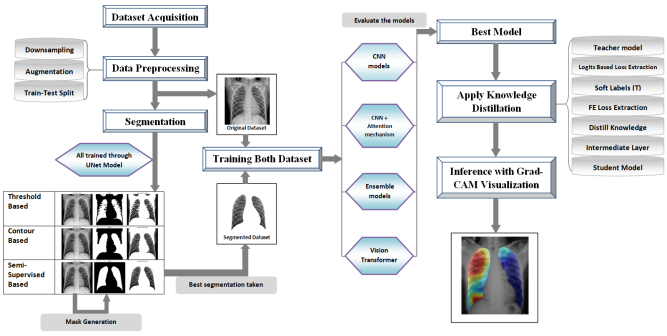

Fig. 1. System Diagram

1) NIAID TB Dataset: The NIAID TB dataset [28] includes 10,500 TB-positive CXR images from 6,000 cases in PNG/JPEG formats, shared via academic access for research purposes.

2) Open Access Mendeley Dataset: This dataset combines two open sources [29] [30], providing approximately around 4,100 images per class (Normal, COVID-19, Pneumonia) after merging and balancing.

3) NIDCH Test Dataset: We collected 28 real-time lung X-rays from NIDCH Hospital, verified by a radiology specialist there, for testing on unseen data.

4) Merging The Datasets: We merged the Open Access Mendeley [29], [30] and NIAID TB [28] datasets, removing duplicate COVID-19 images and adding custom images for Normal and Pneumonia. Final counts: Normal (4,115), Pneumonia (4,110), COVID-19 (3,214), and selected 4,053 high-quality TB images from NIAID, ensuring a balanced dataset for model training.

C. Preprocessing Techniques

To enhance model performance, we applied preprocessing techniques including downsampling and augmentation. Images were resized to 256 by 256 pixels for consistency, reducing file sizes and improving training efficiency. Augmentation methods such as rotations, shifts, brightness adjustments, zooming, and flips were employed to mitigate class imbalance and overfitting, expanding each class to 5,000 images, as shown in Table I and Fig. 3.

TABLE I
SAMPLES OF EACH CLASS BEFORE AND AFTER AUGMENTATION

Class Name	Before	After
NORMAL	4115	5000
PNEUMONIA	4110	5000
TUBERCULOSIS	4053	5000
COVID	3214	5000

D. Segmentation

A crucial preprocessing step for isolating lung regions in chest X-rays. Using a pre-trained U-Net model (Fig. 4), we generate masks to extract lung areas, ensuring the model focuses on relevant features during training. This study explores three segmentation approaches: threshold-based, contour-based, and semi-supervised segmentation.

1) U-Net Architecture: U-Net is a CNN designed for medical image segmentation, as shown in Fig. 2, consisting of an encoder-decoder structure, where x is the input, W represents the convolutional kernel, and b is the bias.

$$F(x) = W * x + b \tag{1}$$

The downsampling process is achieved using max pooling, where R defines the pooling region: $y = \max_{i,j \in R} x_{i,j}$

To restore spatial resolution, upsampling is performed using transposed convolution, where W^T is transposed weight matrix. Skip connections used further refine segmentation.

$$y = W^T * x \tag{2}$$

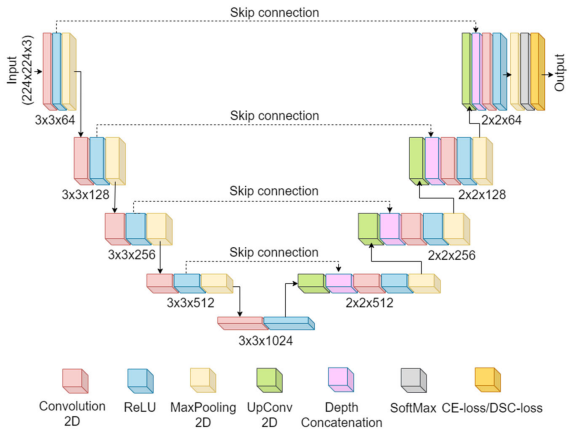

Fig. 2. U-Net architecture Diagram [25]

2) Threshold Based Segmentation: The threshold-based segmentation used pixel intensity limit to separate the lung region from the whole picture.

3) Contour Based Segmentation: Contour-based segmentation detected lung boundaries by identifying the edges, where the intensity shows a great change.

4) Semi Supervised (Manually Annotated Mask) Segmentation: This technique involved using already prepared manually annotated masks. For this, we utilized a dataset from [31] containing 3,000 original images (1,000 each from Normal, COVID-19, and Pneumonia classes), along with their 3000 ground truth masks, which were trained through our U-Net model, just like the other two methods.

Determining the best segmentation method, we generated masks for the entire dataset, isolating lung regions for a fully segmented version dataset.

E. Train-Test-Validation Split

To ensure robust model evaluation, both original and segmented dataset was split into three subset ratio: 0.7 for training, 0.15 for testing, and 0.15 for validation, which allows the model to learn from a large portion of the data during training while being assessed on unseen data in both the validation and testing phases.

F. Applied Models

Our research incorporated deep-learning architectural strategies by utilizing CNNs to extract spatial data and employing ensemble techniques to improve predictive tasks. The integration of Transformer-based models allowed the detection of complex relationships within the data to explore various patterns and enhance total performance.

1) CNN Models: Following pre-trained CNN architectures were deployed to extract spatial features from images:

VGG-16 applies 3x3 convolutions with max pooling and fully connected layers for classification. **InceptionV3** integrates multi-scale convolutions (1x1, 3x3, 5x5) with auxiliary classifiers for efficiency. **ResNet-50** introduces residual learning with skip connections to combat vanishing gradients. **DenseNet-121** enhances gradient flow using dense connections where each layer receives inputs from all previous layers. **EfficientNet-B2** utilizes compound scaling for an optimal balance of accuracy and efficiency, making it suitable for ensemble learning, which is deployed later.

CNNs apply convolution, activation, and pooling, as shown in Fig. 3, mathematically represented as:

$$\hat{y} = \text{softmax}(W_n F_n + b_n) \tag{3}$$

where $\hat{y}$ is the predicted class probability, W_n and b_n are learned parameters, and F_n is the extracted feature representation.

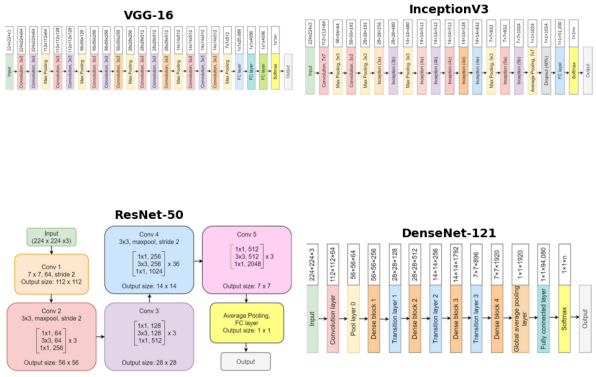

Fig. 3. CNN Architectures Overview [25]

2) CNN Models with Attention Mechanism: To enhance feature extraction, we integrated the **Squeeze-and-Excitation (SE) Block** into base CNN architectures. The SE Block consists of three steps:

Squeeze: Global average pooling reduces the spatial dimensions of feature maps.

Excitation: Two fully connected layers learn channel-wise dependencies:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{4}$$

where W_1 and W_2 are weights, δ is ReLU, σ is sigmoid, and $\mathbf{z}$ is the squeezed feature map.

Scale: The excitation weights are applied to the feature map through channel-wise multiplication.

3) Ensemble Models: We implemented four ensemble models combining two CNNs each, addressing input size differences through resizing and adaptive pooling. The ensembles included: **(1)** EfficientNet-B2 and ResNet-50, **(2)** EfficientNet-B2 and DenseNet-121, **(3)** VGG-16 and DenseNet-121, **(4)** VGG-16 and EfficientNet-B2.

Final predictions were averaged across individual models:

$$y_{\text{ensemble}} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{5}$$

where y_i is the output of the i-th model, and N is the number of models in the ensemble.

4) Vision Transformer Model (CLIP ViT-L/14): The CLIP ViT-L/14 model uses the Vision Transformer (ViT) architecture, which divides images into fixed-size patches, embeds them, and processes them through self-attention layers [32]. It aligns visual features with text through contrastive learning, where the attention matrix captures relationships between image patches. The image encoder generates embeddings from a classification token, while the text encoder generates embeddings from tokenized text, so we freeze text encoder to avoid training complexity. For tasks like classification, freezing text encoder allows model to focus on training the image encoder, enabling it to learn task-specific visual patterns without the complexity of retraining the text encoder. The attention matrix is calculated as follows:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

That is how, CLIP ViT-L/14, as shown in Fig. 4, is employed in this learning, enhancing classification performance.

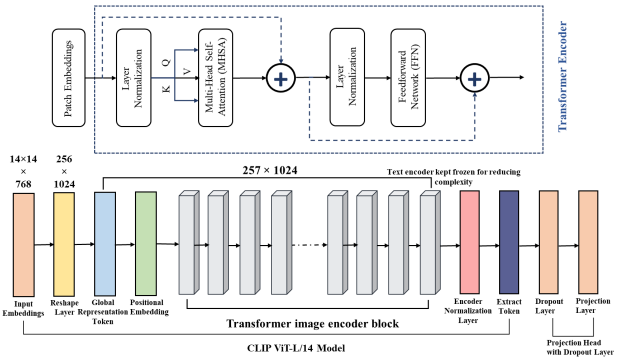

Fig. 4. Clip ViT-L/14 model architecture

G. Knowledge Distillation

We propose **CLIP-KDViT**, a *Knowledge Distillation (KD)* framework that transfers knowledge from **CLIP ViT-L/14** to a compact **MobileNetV2** student model, optimizing accuracy and efficiency for mobile and edge devices. The teacher's robust feature extraction enhances the student's generalization and classification performance. Both models extract feature representations for loss computation, with the student trained on original and segmented datasets using *soft labels* to capture fine-grained class similarities. Training follows an iterative **AdamW** optimization process, selecting the best student model based on validation performance.To ensure effective knowledge transfer, we optimize multiple loss functions:

- **Logit-based KD Loss** (*KL Divergence*): Aligns soft labels.
- **Feature-based KD Loss** (*feat*): Matches feature representations.
- **Attention-based KD Loss** (*Attn*): Aligns attention maps for focus.
- **Cross-Entropy Loss** (*CE*): Learns from true labels.
- **Intermediate Layer Matching** (*MSE*): Aligns feature maps.

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{feat} + \gamma\mathcal{L}_{attn} + \delta\mathcal{L}_{logit} + \varepsilon\mathcal{L}_{int} + \frac{\tau}{T}\mathcal{L}_{KD} \tag{7}$$

An **Attention Mechanism** refines feature alignment, ensuring the student mimics the teacher's focus on key image regions. *Dynamic Temperature* (τ) and *Alpha* (α) are adaptively adjusted for optimal teacher-student supervision balance. By integrating advanced KD techniques, **CLIP-KDViT** achieves a bridge between accuracy and efficiency, making it ideal for edge devices and low-resource environments [33] [34].

H. Explainable AI for Heatmap Generation

To improve interpretability, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) to highlight critical image regions influencing model predictions. Grad-CAM computes gradients of the target class score with respect to the final convolutional layer feature maps , determining feature importance. The class activation map is computed as:

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{8}$$

I. Matrices Used for Result Evaluation

We evaluated model performance using Accuracy, IoU, Dice, Precision, Recall, and F1-score. The final results were averaged across all four classes rather than reported individually. The mathematical equations for each matrix are given in the equation below [35] [16]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$\text{IoU/Jaccard Index} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \tag{10}$$

$$\text{Dice Coefficient} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FN} + \text{FP}}. \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are the fundamental components of the confusion matrix.

IV. Results and Discussion

A. Results of Segmentation

The Threshold-based produced low-accuracy grayscale masks, making it ineffective for clear lung isolation. While better than thresholding, Contour-based included background noise, reducing precision. The most accurate was shown by Semi-Supervised, where the model precisely learned to capture lung shapes, as shown in Fig. 5.

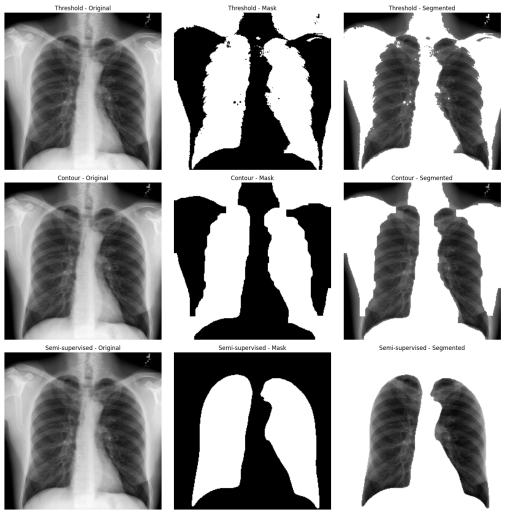

Fig. 5. U-Net Generated Masks and Segmented Images from all Segmentation methods.

Table II and Fig. 6 shows that the best segmentation results were demonstrated by the Semi-Supervised method, with an accuracy of 99.1 percent, a Dice score of 0.98 and IoU of 0.97. These metrics indicate outstanding performance, making it the most effective approach for segmentation.

TABLE II
Performance Metrics for Different Segmentation Types

Segmentation Types	Accuracy	Loss	Precision	Recall	Dice	IoU
Threshold	92.6	0.11	0.86	0.95	0.90	0.85
Contour	92.8	0.18	0.93	0.94	0.93	0.87
Semi-Supervised	**99.1**	**0.02**	**0.98**	**0.98**	**0.98**	**0.97**

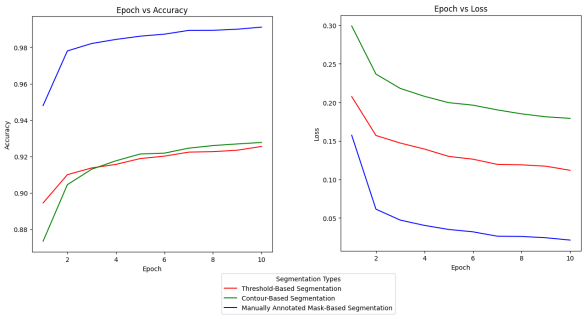

Fig. 6. Accuracy and Loss Plots of All Segmentation Methods

B. Results of CNN Models

CNN models exhibited distinct trends across original and segmented datasets (Table III, Fig. 7). VGG-16 maintained stable performance, while InceptionV3 with attention on segmented data achieved the highest accuracy of 0.96. ResNet-50 showed improvement with segmentation, suggesting noise reduction, whereas DenseNet-121 performed best on original data, indicating reliance on fine-grained details lost during segmentation. The SE attention mechanism consistently enhanced all the model's performance.

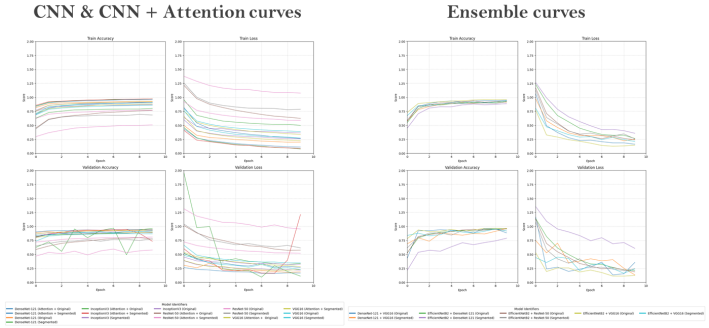

Fig. 7. Accuracy and Loss Plots of All CNN and Ensemble Models

C. Results of Ensemble Models

The ensemble models further improved classification outcomes (Table IV, Fig. 7). VGG-16 + EfficientNet-B2 (Segmented) achieved the highest accuracy of 0.96 with validation accuracy reaching 0.97. Segmentation also benefited VGG-16 + DenseNet-121 and EfficientNet-B2 + ResNet-50, both showing noticeable accuracy gains, reinforcing the role of segmentation in enhancing robustness.

D. Results of CLIP ViT-L/14 Model

Moving beyond CNNs, the ViT model demonstrated superior generalization (Table VI). Accuracy increased from 0.96 to 0.97 on segmented data, with lower training and validation losses. Precision, recall, and F1-score all reached 0.97, highlighting the benefits of segmentation in focusing on key features.

E. Efficiency of Knowledge Distillation with CLIP ViT-L/14

After applying knowledge distillation using a much smaller and less complex model, as shown in Table V, the student model outperformed the teacher (Table VI, Fig. 8), reaching 0.97 accuracy on original and 0.99 on segmented data. The distilled student model demonstrated minimal test loss (0.07) and near-perfect precision, recall, and F1-score (0.98), marking the highest performance among all pipelines. MobileNetV2 effectively absorbed ViT's knowledge, underscoring the advantages of compression for medical imaging.

F. Discussion

The results highlight a structured progression towards achieving optimal performance. While our CNN-based approach performed comparably to existing work [25], the

TABLE III
PERFORMANCE OF CNN MODELS WITH AND WITHOUT ATTENTION MECHANISM IN BOTH DATASETS

CNN Models	Accuracy	Loss	Val Accuracy	Val Loss	Precision	Recall	F1-Score
VGG16 (Original)	0.90	0.27	0.92	0.23	0.92	0.92	0.92
VGG16 (Segmented)	0.86	0.39	0.87	0.35	0.87	0.87	0.86
VGG16 (Original + Attention)	0.92	0.23	0.92	0.23	0.92	0.92	0.92
VGG16 (Segmented + Attention)	0.88	0.34	0.89	0.28	0.90	0.90	0.90
InceptionV3 (Original)	0.91	0.26	0.92	0.22	0.93	0.93	0.93
InceptionV3 (Segmented)	0.87	0.36	0.89	0.32	0.89	0.89	0.89
InceptionV3 (Original + Attention)	0.96	0.08	0.94	0.11	0.96	0.95	0.95
InceptionV3 (Segmented + Attention)	**0.96**	**0.09**	**0.96**	**0.16**	**0.96**	**0.96**	**0.96**
ResNet50 (Original)	0.51	1.07	0.58	0.95	0.62	0.60	0.60
ResNet50 (Segmented)	0.70	0.78	0.78	0.62	0.76	0.75	0.75
ResNet50 (Original + Attention)	0.77	0.62	0.77	0.57	0.79	0.79	0.79
ResNet50 (Segmented + Attention)	0.77	0.59	0.81	0.52	0.80	0.80	0.80
DenseNet121 (Original)	0.93	0.20	0.94	0.18	0.94	0.94	0.94
DenseNet121 (Segmented)	0.80	0.50	0.89	0.32	0.88	0.88	0.88
DenseNet121 (Original + Attention)	0.95	0.11	0.94	0.17	0.94	0.94	0.94
DenseNet121 (Segmented + Attention)	0.91	0.26	0.93	0.21	0.92	0.92	0.92

TABLE IV
PERFORMANCE OF ENSEMBLE MODELS IN BOTH DATASETS

Ensemble Models	Accuracy	Loss	Val Accuracy	Val Loss	Precision	Recall	F1-Score
EfficientNet-B2 + ResNet-50 (Original)	0.92	0.26	0.95	0.20	0.92	0.92	0.92
EfficientNet-B2 + ResNet-50 (Segmented)	0.94	0.24	0.96	0.18	0.96	0.96	0.96
EfficientNet-B2 + DenseNet-121 (Original)	0.89	0.36	0.79	0.60	0.88	0.88	0.86
EfficientNet-B2 + DenseNet-121 (Segmented)	0.93	0.26	0.94	0.21	0.95	0.95	0.95
VGG-16 + DenseNet-121 (Original)	0.92	0.22	0.96	0.13	0.94	0.94	0.94
VGG-16 + DenseNet-121 (Segmented)	0.95	0.16	0.94	0.13	0.95	0.95	0.95
VGG-16 + EfficientNet-B2 (Original)	0.94	0.20	0.95	0.16	0.91	0.90	0.90
VGG-16 + EfficientNet-B2 (Segmented)	**0.96**	**0.10**	**0.97**	**0.09**	**0.96**	**0.96**	**0.96**

TABLE V
PERFORMANCE METRICS AND PARAMETERS OF TEACHER AND STUDENT MODEL IN KNOWLEDGE DISTILLATION

Metric	Teacher (Original)	Student (Original)	Teacher (Segmented)	Student (Segmented)
Model Size (MB)	1631.24 MB	9.24 MB	1631.24 MB	9.24 MB
Trainable Parameters	427.62M	2.03M	427.62M	2.03M
Training Time/Epoch (s)	840s	360s	660s	270s
Total Training Time (s)	8400s	3600s	6600s	2700s
Inference Time/Image (s)	4s	2s	3s	1s
GPU Utilization (%)	85-90%	35-40%	75-80%	20-30%
CPU Utilization (%)	60-70%	15-25%	45-50%	10-25%

integration of ensemble learning further refined accuracy. However, the most substantial improvement came from the ViT model, which outperformed CNNs on segmentation tasks (Table VI). By leveraging knowledge distillation, the student model not only retained ViT's accuracy but exceeded it, achieving 0.99 accuracy in training and 0.972 on unseen images (Fig. 9). With a misclassification rate well below regulatory thresholds (510(k) FDA guidelines [36]), this system demonstrates strong potential for clinical deployment.

G. Grad-CAM Visualization

Lastly, we deployed our best evaluated model pipeline into our own web-based application, **Respire Check**, which is developed using Streamlit in a Python environment. The application combines segmentation and classification along with Grad-CAM visualizations to improve transparency in decision-making, as shown in Fig. 10.

H. Real-Time Test Data Evaluation

Our model was also evaluated using chest X-Ray images provided by NIDCH Hospital in Bangladesh. Patient identification data was anonymized before being provided to us. Of the 28 provided images, 24 were correctly classified (86%) in the web app's final test, which is shown in Fig. 11.

V. LIMITATIONS AND FUTURE WORKS

Despite having promising results, this study has limitations. The model's performance depends on data quality and differentiability, risking biases if it is not carefully curated. While MobileNetV2 improves deployment efficiency, it also leads to decreased accuracy levels relative to larger model versions. AI integration in clinical workflows concerns data privacy, liability, and trust. The future research agenda includes Android deployment, various datasets that enhance generalization

TABLE VI
PERFORMANCE COMPARISON OF TEACHER AND STUDENT MODEL IN BOTH DATASETS

Models	Accuracy	Loss	Val Accuracy	Val Loss	Test Accuracy	Test Loss	Precision	Recall	F1-Score
Teacher (Original)	0.96	0.15	0.94	0.19	0.95	0.15	0.95	0.94	0.94
Student (Original)	0.97	0.12	0.97	0.05	0.96	0.13	0.96	0.96	0.96
Teacher (Segmented)	0.97	0.09	0.98	0.08	0.96	0.10	0.97	0.97	0.97
Student (Segmented)	**0.99**	**0.06**	**0.98**	**0.03**	**0.97**	**0.07**	**0.98**	**0.98**	**0.98**

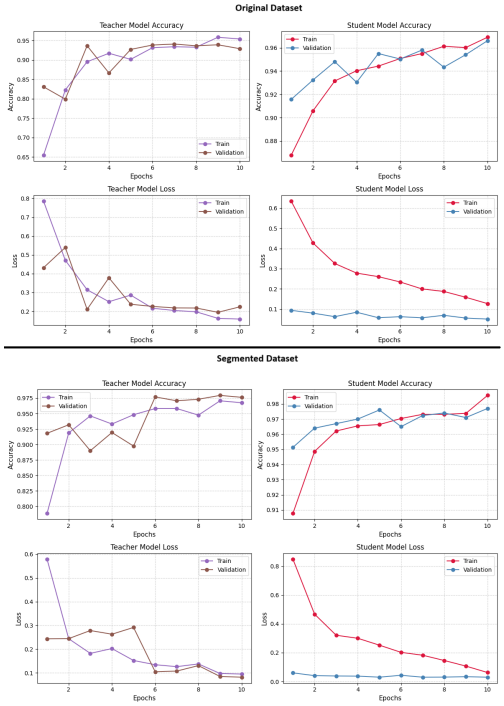

Fig. 8. Accuracy and Loss Plots for Teacher and Student Model on Both Original and Segmented Dataset

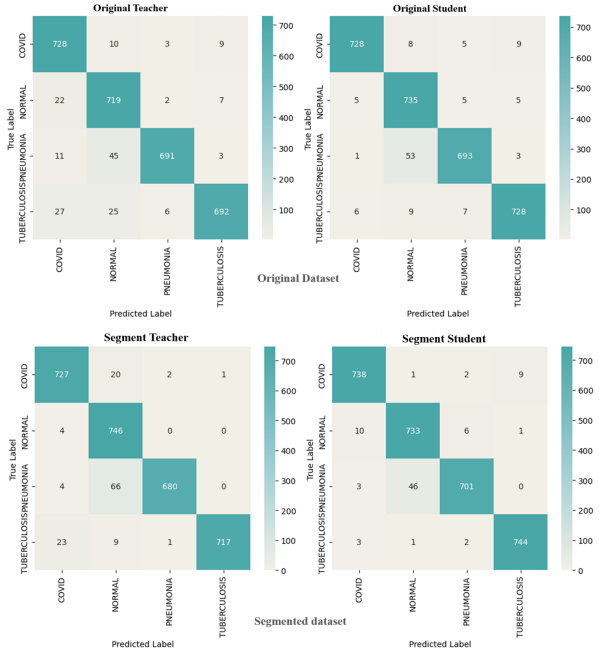

Fig. 9. Confusion Matrix of both Teacher and Student Model for Original and Segmented Dataset

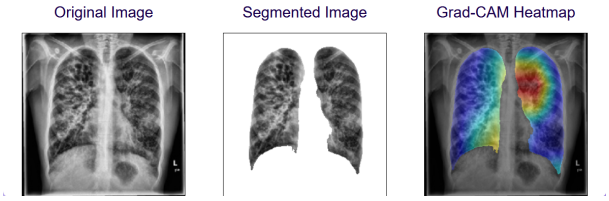

Fig. 10. Grad-CAM Heatmap Visualization

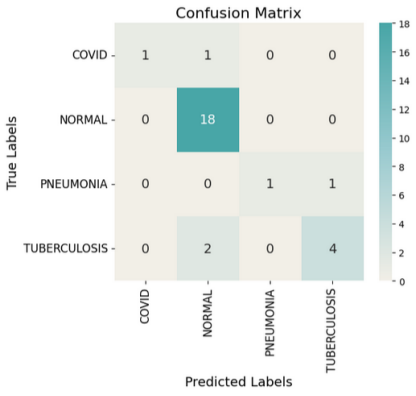

Fig. 11. Confusion Matrix of NIDCH Dataset

abilities, and real-time analysis for newly arising respiratory conditions.

VI. CONCLUSSION

Several recent studies emphasize the ability of AI-powered medical imaging platforms to enhance lesion detection. Our multi-class X-ray diagnostic framework integrates fluency in semantic, multi-class classification, and knowledge distillation with a tractable semi-supervised approach to segmentation, providing a high-performance as well as interpretable but computationally-capable solution that is feasible in resource-poor settings. It connects AI diagnostics to actual medical outcomes by combining easy AI models with explanatory techniques. The framework will be scaled up and fine-tuned for real-world usage in future studies.

ACKNOWLEDGMENT

The authors sincerely thank Dr. Shahnewaz Siddique for his invaluable guidance and acknowledge Ristian and Mohammod Abdullah for their key contributions here. Additionally, we thank Karlynn Noble (NIAID) for granting academic access to the TB dataset and the Radiology Department at NIDCH

for providing data to test our model in real-world conditions. Their support was also crucial to this research.

REFERENCES

[1] WHO, World health statistics 2022, in: Monitoring Health for the SDGs Sustainable Development Goals, 2022.

[2] L. Worthington,"What to know about rising tuberculos is cases in the U.S.,"National Geographic,Jan.29,2025.Available: www.nationalgeographic.com/science/article/us-tuberculosis-tb-rates-rising

[3] C. Trobajo-Sanmartín, M. E. Portillo, A. Navascués, I. Martínez-Baz, C. Ezpeleta, and J. Castilla, "Unusually high incidence of pneumonia in Navarre, Spain, 2023–2024," Enfermedades Infecciosas y Microbiología Clínica, vol. 43, no. 2, pp. 93–96, 2025, doi: 10.1016/j.eimc.2024.08.008.

[4] https://www.worldometers.info/coronavirus/

[5] Yu Z, Wang K, Wan Z, Xie S, Lv Z. Popular deep learning algorithms for disease prediction: a review. Cluster Comput. 2023;26(2):1231-1251. doi: 10.1007/s10586-022-03707-y. Epub 2022 Sep 13. PMID: 36120180; PMCID: PMC9469816.

[6] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2097–2106.

[7] Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest X-ray classification. Scientific Reports, 9(1), 1–10.

[8] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241.

[9] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[10] D.Ibrahim,N.Elshennawy,and A.Sarhan,"Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Computers in Biology and Medicine*, vol. 132, p. 104348, 2021. doi:10.1016/j.compbiomed.2021.104348.

[11] M. Ahmed et al., "Joint Diagnosis of Pneumonia, COVID-19, and Tuberculosis from Chest X-ray Images: A Deep Learning Approach," *Diagnostics*, vol. 13, 2023. doi: 10.3390/diagnostics13152562.

[12] B. Narayana, G. Reddy, S. Kosaraju, and S. Choudhary, "Integrated Hybrid Model for Lung Disease Detection Through Deep Learning," *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, 2024. doi: 10.35784/iapgos.6081.

[13] Wang,G.,Liu,X.,Shen,J.,Wang,C. Li,Z.,Ye,L.,et al.(2021).A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. Nature Biomedical Engineering, 5, 509 - 521. https://doi.org/10.1038/s41551-021-00704-1.

[14] Kulkarni, A., Parasnis, G., Balasubramanian, H., Jain, V., Chokshi, A., & Sonkusare, R. (2023). Advancing Diagnostic Precision: Leveraging Machine Learning Techniques for Accurate Detection of COVID-19, Pneumonia, and Tuberculosis in Chest X-Ray Images. ArXiv, abs/2310.06080. https://doi.org/10.48550/arXiv.2310.06080.

[15] Muthaki, T., Masuk, S., Maksud, A., Rafi, M., & Sakib, N. (2023). An Approach to Detect Cardiomegaly, COVID-19, Pneumonia, Pneumothorax and Tuberculosis from CXR Images Using Ensembles of Deep Learning. 2023 26th International Conference on Computer and Information Technology (ICCIT), 1-6. https://doi.org/10.1109/ICCIT60459.2023.10441082.

[16] Rahman, T.; Khandakar, A.; Kadir, M.A.; Islam, K.R.; Islam, K.F.; Mazhar, R.; Hamid, T.; Islam, M.T.; Kashem, S.; Mahbub, Z.B.; et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access 2020, 8, 191586–191601.

[17] Mamalakis, M., Swift, A., Vorselaars, B., Ray, S., Weeks, S., Ding, W., et al. (2021). DenResCov-19: A deep transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from X-rays. Computerized Medical Imaging and Graphics, 94, 102008 - 102008. https://doi.org/10.1016/j.compmedimag.2021.102008.

[18] Chakraborty, A., Nowrin, A., & Pathak, A. (2024). DeepX-Ray: A Comparative Study of Deep Learning-Based Classification and Segmentation Techniques for Automated Detection and Diagnosis of COVID-19 from Chest X-ray Images. Indonesian Journal of Computer Science. https://doi.org/10.33022/ijcs.v13i2.3827.

[19] Hadhoud, Y., Mekhaznia, T., Bennour, A., Amroune, M., Kurdi, N. A., Aborujilah, A. H., & Al-Sarem, M. (2024). From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images. Diagnostics, 14(23), 2754. https://doi.org/10.3390/diagnostics14232754.

[20] Chen, T., Philippi, I., Phan, Q. B., Nguyen, L., Bui, N. T., daCunha, C., & Nguyen, T. T. (2024). A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images. Healthcare Analytics, 5, 100332. https://doi.org/10.1016/j.health.2024.100332.

[21] Kebache, R., Laouid, A., Guia, S., Kara, M., & Bouadem, N. (2023). Tuberculosis Detection Using Chest X-Ray Image Classification by Deep Learning. Proceedings of the 7th International Conference on Future Networks and Distributed Systems. https://doi.org/10.1145/3644713.3644759.

[22] Kabir, M. M., Mridha, M. F., Rahman, A., Hamid, M. A., & Monowar, M. M. (2024). Detection of COVID-19, pneumonia, and tuberculosis from radiographs using AI-driven knowledge distillation. Heliyon, 10(5), e26801. https://doi.org/10.1016/j.heliyon.2024.e26801.

[23] BabaAhmadi, A., Khalafi, S., Shariatpanahi, M., & Ayati, M. (2023). Designing an improved deep learning-based model for COVID-19 recognition in chest X-ray images: a knowledge distillation approach. Iran Journal of Computer Science, 1-11. https://doi.org/10.48550/arXiv.2301.02735.

[24] Akhter, Y., Ranjan, R., Singh, R., & Vatsa, M. (2024). Low-Resolution Chest X-Ray Classification Via Knowledge Distillation and Multi-Task Learning. 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 1-5. https://doi.org/10.1109/ISBI56570.2024.10635737.

[25] Nillmani, Sharma, N., Saba, L., Khanna, N., Kalra, M., Fouda, M., & Suri, J. (2022). Segmentation-Based Classification Deep Learning Model Embedded with Explainable AI for COVID-19 Detection in Chest X-ray Scans. Diagnostics, 12. https://doi.org/10.3390/diagnostics12092132.

[26] Ou, C., Chen, I., Chang, H., Wei, C., Li, D., Chen, Y., & Chang, C. (2024). Deep Learning-Based Classification and Semantic Segmentation of Lung Tuberculosis Lesions in Chest X-ray Images. Diagnostics, 14. https://doi.org/10.3390/diagnostics14090952.

[27] Panwar, H., Gupta, P., Siddiqui, M., Morales-Menéndez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. Chaos, Solitons, and Fractals, 140, 110190 - 110190. https://doi.org/10.1016/j.chaos.2020.110190.

[28] National Institute of Allergy and Infectious Diseases, "TB Portals Program," [Online]. Available: https://tbportals.niaid.nih.gov/. [Accessed: AUGUST. 28, 2024].

[29] Asraf, Amanullah; Islam, Zabirul (2021), "COVID19, Pneumonia and Normal Chest X-ray PA Dataset", Mendeley Data, V2, doi: 10.17632/mxc6vb7svm.2

[30] Asraf, Amanullah; Islam, Zabirul (2021), "COVID19, Pneumonia and Normal Chest X-ray PA Dataset", Mendeley Data, V1, doi: 10.17632/jctsfj2sfn.1

[31] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, 2021, Art. no. 104319, doi: 10.1016/j.compbiomed.2021.104319.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proc. 31st Conf. Neural Inf.Process.Syst.(NeurIPS), Long Beach,CA,USA,2017.

[33] B. Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., & Li, Y. (2024). ViTKD: Feature-based Knowledge Distillation for Vision Transformers. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1379-1388. https://doi.org/10.1109/CVPRW63382.2024.00145.

[34] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós and J. C. S. Miguel, "Attention-Based Knowledge Distillation in Scene Recognition: The Impact of a DCT-Driven Loss," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 9, pp. 4769-4783, Sept. 2023, doi: 10.1109/TCSVT.2023.3250031

[35] Jain, R.; Gupta, M.; Taneja, S.; Hemanth, D.J. Deep learning based detection and analysis of COVID-19 on chest X-ray images. Appl. Intell. 2021, 51, 1690–1700.

[36] Komatsu, M.; Sakai, A.; Dozen, A.; Shozu, K.; Yasutomi, S.; Machino, H.; Asada, K.; Kaneko, S.; Hamamoto, R. Towards Clinical Application of Artificial Intelligence in Ultrasound Imaging. Biomedicines 2021, 9, 720.