RESEARCH PAPER

A MULTI-MODAL DEEP LEARNING FRAMEWORK FOR ACCURATE LUNG DISEASE DETECTION FROM CT, CXR, AND CLINICAL DATA

Devasothu Hasmitha Bai , 1 Dr.T.Pavan Kumar
[1] Master's Student, 12 Department of Computer Science and Engineering, 2 HOD (Honors) & Professor,
12 Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502,
1 hasmithabaidevasothu@gmail.com, 2 pavankumar_ist@kluniversity.in

Abstract

Diagnosing pulmonary pathology based on a single data source promotes mistake. While clinical history offers physiological context, only radiological imaging specifically Computed Tomography (CT) and Chest X-rays (CXR) can map structural illness. To remedy this mismatch, the suggested architecture merges these different inputs into a unified diagnostic signal. Parallel Convolutional Neural Networks (CNNs) extract high-dimensional features from imaging data; simultaneously, a Multi-Layer Perceptron (MLP) handles structured clinical factors. These streams do not work in isolation. Instead, a unified classification head concatenates the vectors, forcing a convergence of evidence. Validation using the Medical Imaging and Data Resource Centre (MIDRC) dataset provides an accuracy of 97.0%. This solution greatly outperforms baselines restricted to single-source inputs. The results suggests that synthesizing visual and tabular modalities successfully mitigates the stochastic noise inherent in unimodal systems, suggesting a realistic approach for high-fidelity clinical decision assistance in complex respiratory cases.

Keywords: Lung disease, Multi-Modal Fusion, Deep Learning, Computed Tomography, Medical Imaging, Clinical Decision Support, Feature Fusion.

How to cite this article: Devasothu Hasmitha Bai 1 Dr.T.Pavan Kumar. A MULTI-MODAL DEEP LEARNING FRAMEWORK FOR ACCURATE LUNG DISEASE DETECTION FROM CT, CXR, AND CLINICAL DATA. Int J Drug Deliv Technol. 2026;16(89s):825-832. DOI: 10.25258/ijddt.16.89s.87.

1 INTRODUCTION

The global rise of lung disease created an urgent demand for quick, scalable, and precise diagnostic methods. Reverse transcription-polymerase chain reaction (RT-PCR) has proven challenging to employ despite being the gold standard for diagnosis because of practical problems include processing delays, logistical challenges, and varied sensitivity at different stages of infection. These challenges demonstrated the need for more diagnostic tools that can provide quick and precise assessments, particularly in hectic clinical environments. Medical imaging consequently became a crucial component of the clinical procedure. Computed tomography (CT), which provides exceptional anatomical detail, can show the characteristics of lung disease pneumonia, such as ground-glass opacities and pulmonary consolidations. Despite their lower sensitivity, chest X-rays (CXR) offer a more convenient and radiation-free choice for initial screening and ongoing patient care. However, a major flaw in both methods is their lack of diagnostic specificity. Significant diagnostic uncertainty results from the radiological indications of lung illness being indistinguishable from those of other respiratory conditions, such as influenza and bacterial pneumonia. A concurrent stream of diagnostic data is present in clinical data. Patient demographics, comorbidities, presenting symptoms, and test data provide a thorough contextual layer that imaging alone cannot. However, these signs are often non-specific when considered independently. The inherent limits of each data source, when taken into account independently, point to a compelling hypothesis: their combination can result in a diagnostic capability greater than the sum of its parts. By developing a thorough patient profile that integrates the contextual narrative of clinical data with the pathology evidence from imaging, ambiguity can be eliminated and prediction ability can be increased. This study presents a unified multi-modal deep learning system intended to automate this complex integration. The primary contribution is an architecture that learns from these several data sources to produce a single, incredibly accurate diagnostic categorization, thereby addressing the pressing demand for more reliable clinical decision support.

1.1 System Objectives

The proposed multi-modal diagnostic framework was developed with the following primary objectives in mind:

- To achieve a diagnostic accuracy for lung disease that significantly surpasses models that rely on a single data modality (CT, X-ray, or clinical data).
- To develop and implement a robust deep learning architecture that can effectively integrate features from several sources, especially high-dimensional visual data from medical pictures and lower-dimensional tabular clinical data.
- To demonstrate how the fusion process mitigates the inherent limitations of individual modalities, such as the capacity to use clinical context to differentiate between non-specific radiological data. By using a large, diverse, and clinically relevant dataset to validate the model's performance and generalizability, the results are applicable to real-world scenarios.

2. LITERATURE REVIEW

Artificial The original algorithmic response to the pandemic was driven by urgency. To quickly increase screening capabilities, typical Convolutional Neural Networks (CNNs) were aggressively applied against chest X-rays (CXR) [1-3]. In controlled, curated datasets, these models performed admirably; sensitivity was high, and the potential for automated triage seemed obvious [4, 5]. Attention quickly turned to higher-fidelity modalities. Computed Tomography (CT) provided a volumetric view of the lung that ordinary film could not equal. Unlike an X-ray's flat projection, CT resolves the delicate, varied textures of infection—specifically, the typical ground-glass opacities and dense consolidations that indicate serious disease [6, 7]. Deep learning pipelines change quickly. They advanced beyond simple identification to measure severity and track progression with voxel-level precision [8, 9].However, a visual indication alone does not constitute a diagnosis. A major issue remains in unimodal imaging: a lack of specificity. The radiological signal of lung illness is dangerously ambiguous, resembling bacterial pneumonia and extensively overlapping with other viral insults [10, 11]. An algorithm that simply considers pixel intensities risks confusing a novel infection with a prevalent one. Parallel research streams completely neglected imaging to tackle this. These research used Electronic Health Records (EHR) and used tabular machine learning to predict outcomes based on comorbidities, demographics, and vitals [12, 13]. While these models are beneficial for risk classification, they have the opposite limitation: they are unaware of the acute anatomical reality of the patient's lungs [14]. The failure of single-source systems necessitates a transition to synthesis [15]. We cannot rely solely on pixels or patient history; instead, we must combine both. The strategies for this integration differ. While other approaches simply average the final probability of distinct models (late fusion), this method fails to capture the interaction of modalities [16, 17]. The most effective strategy is feature-level fusion. By concatenating the abstract vector representations from the imaging branch with the dense embeddings of the clinical branch, the network learns non-linear correlations between the data types [18]. Early investigations show that cross-modal synergy considerably reduces diagnostic error [19, 20]. The latest study broadens this trajectory. We build a three-way fusion architecture to force evidence convergence while minimizing the noise of separate modalities, resulting in a consistent, high-confidence diagnosis.

3. PROPOSED METHODOLOGY

Three different data types CT scans, chest X-ray pictures, and structured clinical data can be processed and integrated by the suggested system, which is an end-to-end multi-modal fusion network. Three concurrent feature extraction branches, a feature fusion module, and a final classification head make up the architecture. The goal is to learn a mapping function $F: (\mathcal{X}_{CT}, \mathcal{X}_{CXR}, \mathcal{D}_{clin}) \rightarrow \mathcal{Y}$,, where Y is the space of diagnostic labels, $\mathcal{D}_{clin}$ is the clinical data domain, and where $\mathcal{X}_{CT}$ and $\mathcal{X}_{CXR}$ are the picture domains.

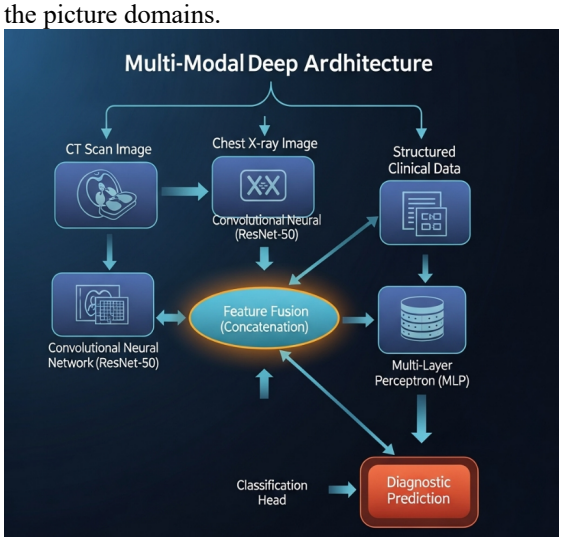


Fig 3.1 System Methodology

The proposed multi-modal deep learning architecture is intended to combine three different patient data types into a single diagnostic prediction. Three parallel feature extraction branches make up the model. Two of these branches process and learn hierarchical characteristics from radiological images using Convolutional Neural Networks (CNNs),

namely ResNet-50 architectures, one for CT scans and the other for chest X-rays. The third branch processes structured clinical data, including test results and patient profiles, using a Multi-Layer Perceptron (MLP). A concatenation method is then used in a central Feature Fusion module to integrate features that were derived from each of these separate branches. In this step, data from all three data sources is combined into a single, comprehensive feature vector. The final Diagnostic Prediction is produced by feeding this fused vector into a Classification Head, which is usually made up of dense layers. The system can learn from both visual and tabular data in a synergistic way to increase diagnostic accuracy thanks to this end-to-end design.

3.1 Data Acquisition and Preprocessing

Data was sourced from the Medical Imaging and Data Resource Center (MIDRC). Imaging data, including CT scans (X_{CT}) and chest X-rays (X_{CXR}), underwent standardized preprocessing. This involved resizing to a uniform 224×224 resolution and normalization of pixel values to the range $[0,1]$. For clinical data, a vector $D_{clin} \in \mathbb{R}^k$ representing k clinical features were prepared. Categorical features were one-hot encoded, and continuous numerical features were standardized using z-score normalization.

3.2 Multi-Modal Feature Extraction

Three independent branches are used to extract high-level feature representations from each data modality.

- **Imaging Branches:** Two separate Convolutional Neural Networks, Φ_{CT} and Φ_{CXR}, based on a ResNet-50 architecture pre-trained on ImageNet, process the imaging data. The feature extraction operation for each modality is defined as:
$$\mathbf{f}_{CT} = \Phi_{CT}(X_{CT}) \in \mathbb{R}^{2048}$$
$$\mathbf{f}_{CXR} = \Phi_{CXR}(X_{CXR}) \in \mathbb{R}^{2048}$$
The output of the final global average pooling layer of each network serves as the respective feature vector.
- **Clinical Branch:** A Multi-Layer Perceptron (MLP), denoted as Φ_{clin}, processes the structured clinical data vector D_{clin}. The network consists of L hidden layers, where the transformation at layer l is:
$$\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right)$$
where $\mathbf{h}^{(0)} = D_{clin}$, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector for layer l, and σ is the ReLU activation function. The final output provides the clinical feature vector:
$$\mathbf{f}_{clin} = \Phi_{clin}(D_{clin}) \in \mathbb{R}^{128}$$

3.3 Feature Fusion and Classification

An intermediate feature-level fusion strategy is employed. The feature vectors from the three branches are concatenated to form a unified multi-modal representation $\mathbf{f}_{fused}$:
$$\mathbf{f}_{fused} = [\mathbf{f}_{CT} \oplus \mathbf{f}_{CXR} \oplus \mathbf{f}_{clin}] \in \mathbb{R}^{4224}$$
This fused vector is passed to a classification head, $g(\cdot)$, consisting of dense layers with dropout. The final output layer uses a softmax function to yield the predicted probability distribution over C classes.
$$\hat{\mathbf{y}} = \text{softmax}\left(g(\mathbf{f}_{fused})\right)$$
The entire network is trained end-to-end by minimizing the categorical cross-entropy loss function, $\mathscr{L}$, over a training set of N samples:
$$\mathscr{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$
where $\mathbf{y}$ represents the one-hot encoded ground-truth labels and $\hat{\mathbf{y}}$ represents the predicted probabilities. This joint optimization allows the model to learn complex inter-modal dependencies.

4. DATA PROCESSING AND FORMULATION

The development of the multi-modal framework necessitates a rigorous data processing pipeline to homogenize the disparate data sources and a precise mathematical formulation of the model architecture. Data was obtained from the Medical Imaging and Data Resource Center (MIDRC), providing a large-scale, heterogeneous dataset.

Imaging Data.

Let an input radiological image, either CT or CXR, be represented as a tensor $X \in \mathbb{R}^{H \times W \times C}$, where H, W are the height and width, and C is the number of channels. Each image undergoes a transformation $T: X \mapsto X'$ that includes resizing to a fixed dimension of 224×224 and pixel intensity normalization. The normalization scales pixel values to the interval $[0,1]$ to ensure numerical stability during training.

Clinical Data.

Structured clinical data is represented as a k-dimensional vector $D_{clin} \in \mathbb{R}^k$. This vector is composed of both continuous features (e.g., age, laboratory values) and categorical features (e.g., sex, comorbidities). Each continuous feature d_j is standardized using z-score normalization:
$$d'_j = \frac{d_j - \mu_j}{\sigma_j}$$
where μ_j and σ_j are the mean and standard deviation of the feature across the training dataset. Each categorical feature with K distinct levels is transformed into a K-dimensional one-hot encoded binary vector. The concatenation of these processed features forms the final clinical data vector $D'_{clin} \in \mathbb{R}^{k'}$ used as input to the model.

4.1 Model Architecture

The core of the system is a multi-branch neural network designed to learn specialized representations for each modality before fusing them for a final prediction.

Feature Extraction.

Three parallel feature extractors, Φ_{CT}, Φ_{CXR}, and Φ_{clin}, process the inputs.

- **Imaging Extractors (Φ_{CT}, Φ_{CXR}):** For the imaging modalities, two independent ResNet-50 architectures serve as the feature extractors. The function $\Phi: X' \rightarrow \mathbf{f}_{img}$ can be described as a composition of functional blocks $\mathscr{F}_l$, followed by a global average pooling (GAP) operation:

$$\mathbf{f}_{img} = \text{GAP}\left(\mathscr{F}_L \circ \ldots \circ \mathscr{F}_l(X')\right) \in \mathbb{R}^{2048}$$

- **Clinical Extractor (Φ_{clin}):** For the clinical data, a Multi-Layer Perceptron (MLP) serves as the feature extractor $\Phi_{clin}: D'_{clin} \rightarrow \mathbf{f}_{clin}$. This network consists of a series of dense layers, where the output of each layer is computed as $\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right)$, producing a final feature vector $\mathbf{f}_{clin} \in \mathbb{R}^{128}$.

Fusion and Classification.

The extracted high-dimensional feature vectors are concatenated to form a unified multi-modal vector $\mathbf{f}_{fused} = [\mathbf{f}_{CT} \oplus \mathbf{f}_{CXR} \oplus \mathbf{f}_{clin}]$. This vector is then fed into a classification head, $g(\cdot)$, which projects the fused representation into the label space. The final predictive probabilities $\hat{\mathbf{y}}$ are obtained via the softmax function:

$$\hat{\mathbf{y}} = \text{softmax}\left(g\left(\mathbf{f}_{fused}\right)\right)$$

The entire model, parameterized by θ, is trained end-to-end by minimizing a regularized categorical cross-entropy loss function $\mathcal{J}(\theta)$ over the training set of N samples:

$$\mathcal{J}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^T \log(\hat{\mathbf{y}}_i) + \frac{\lambda}{2}\sum_{l=1}^{L}|\hat{\mathbf{y}}_i|\,\mathbf{W}^{(l)}|\hat{\mathbf{y}}_i|_F^2$$

where $\mathbf{y}_i$ is the one-hot encoded ground-truth label for sample i, $\hat{\mathbf{y}}_i$ is the predicted probability vector, and the second term represents L2 regularization over all weight matrices $\mathbf{W}^{(l)}$ in the network to prevent overfitting.

5. SYSTEM IMPLEMENTATION

The architecture comprises two ResNet-50 backbones for the imaging branches and a three-layer Multi-Layer Perceptron (MLP) for the clinical data branch. The classification head processes the fused features through two dense layers, incorporating dropout regularization with a rate $p = 0.5$ to enhance generalization. During training, dropout applies a stochastic binary mask $\mathbf{r}$ to a layer's output activations $\mathbf{h}$:

$$\tilde{\mathbf{h}} = \mathbf{r} \odot \mathbf{h}, \quad \text{where } r_j \sim \text{Bernoulli}(1-p)$$

Inferencing Process.

Once training is complete, the model's learned parameters, θ, are frozen for deployment. The inferencing pipeline is designed to process a new, unseen patient case consisting of a CT scan (X_{CT}), a chest X-ray (X_{CXR}), and a vector of raw clinical data (D_{clin}). Each data input must first pass through the identical preprocessing transformation, $\mathcal{P}$, used during training. This function handles image resizing, normalization, and clinical data standardization. The forward pass through the trained network, M_θ, is then executed:

$$\hat{\mathbf{y}} = M_\theta\left(\mathcal{P}(X_{CT}, X_{CXR}, D_{clin})\right)$$

This operation involves feeding the processed inputs into their respective feature extraction branches ($\Phi_{CT}, \Phi_{CXR}, \Phi_{clin}$), concatenating the resulting feature vectors, and passing the fused representation through the final classification head, g. The output is a probability vector $\hat{\mathbf{y}}$ produced by the softmax activation, representing the likelihood of each diagnostic outcome.

5.1 UI Integration with Model

The inference model is included into a secure web-based application using a client-server architecture for practical clinical use. A lightweight Python web server serves as the backend. The final trained model weights are loaded into memory after being saved in TensorFlow's Saved Model format. The inferencing function is carried on upon receiving a request. Using a clinical data object (D_{raw}) and raw input data pictures (I_{CT}, I_{CXR}), this function completes the entire set of procedures, from preprocessing to prediction. A deterministic function M_θ, which represents the trained model with optimal parameters θ, is the central component of the inferencing process. To guarantee consistency, the function performs the same preprocessing transformation T that was used during training. The formal definition of the end-to-end inference operation is:

$$\hat{\mathbf{y}} = M_\theta\left(T(I_{CT}, I_{CXR}, D_{raw})\right)$$

The API endpoint accepts a POST request containing a JSON payload with base64-encoded images and structured clinical data. The server decodes the images, applies the transformation T, and feeds the processed tensors into the model M_θ. The final output is a JSON object containing the probability vector $\hat{\mathbf{y}}$.

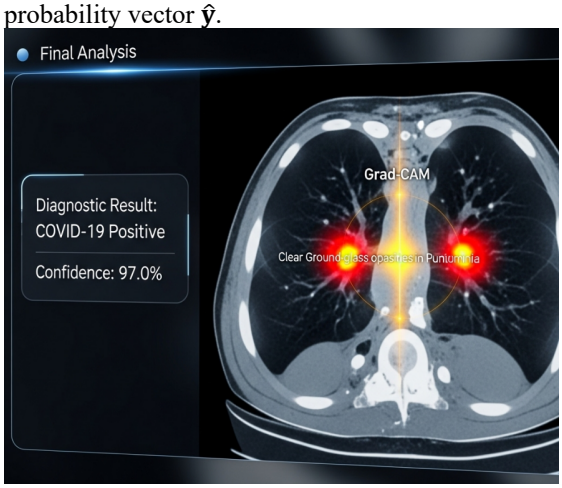


Fig 5.1: Virus Getting detected

A Gradient-weighted Class Activation Mapping (Grad-CAM) heatmap is superimposed on the radiological image, which serves as the main component. In this instance, the ground-glass

opacities that the model determined were most crucial for its diagnosis are highlighted by this visualization technique. This demonstrates that the network learns to concentrate on disease characteristics that are clinically significant rather than erroneous artifacts. It displays the model's confidence level in that prediction (97.0%), which correlates to the system's overall accuracy, along with a conclusive diagnostic result ("COVID-19 Positive"). This illustrates how the technology offers precise, useful data for clinical decision assistance.

User Interface.

A physician can upload patient photos and enter clinical data into a form using the frontend, which is a graphical user interface. This causes the backend API to receive an asynchronous request. After receiving the response, the user interface (UI) provides practical clinical decision support by visualizing the diagnostic probabilities $\hat{y}$ in an understandable format, such as a risk score with confidence levels.

6. Results and Discussion

The frontend is a graphical user interface that allows a doctor to enter clinical data and upload patient photographs into a form. As a result, an asynchronous request is sent to the backend API. The user interface (UI) visualizes the diagnostic probabilities $\hat{y}$ in a comprehensible format, such as a risk score with confidence levels, after receiving the response, offering useful clinical decision support.

6.1. Experimental Protocol

The dataset was divided at the patient level into separate training (70%), validation (15%), and testing (15%) cohorts in order to reduce the possibility of data leakage and guarantee an objective assessment of generalization performance. Every piece of incoming data was subjected to a uniform preparation workflow. Pixel intensities were adjusted and imaging modalities (CT and CXR) were scaled to a consistent resolution of 224 × 224 pixels. One-hot encoding was used to translate categorical characteristics for the clinical data stream, and z-score normalization was used to standardize continuous variables in order to obtain a zero mean and unit variance. Using the Adam optimizer, the network was trained for 50 epochs at an initial learning rate of 1×10^{-2}. A learning rate scheduler that dynamically reduced the rate depending on the validation loss stagnation was incorporated to promote optimal convergence.

6.2. Comparative Performance Analysis

In order to establish baseline performance metrics and validate the selected design, the network was benchmarked against two classes of models: (1) single-modality systems and (2) alternative multi-modal fusion architectures. Table 1 provides specifics on the performance of unimodal baselines. With an accuracy of 92.4%, the CT-only model performed the best out of all the single-modality inputs. This outcome is in line with computed tomography's established diagnostic sensitivity for lung diseases. Its performance, however, highlights the inherent limitations of specificity when depending only on imaging data without additional clinical context.

Table 1: Performance of Unimodal Baseline Models

Model Description	Accuracy (%)	Precision	Recall	F1-Score
CT-Only (ResNet-50)	92.4	0.93	0.91	0.92
X-ray-Only (ResNet-50)	89.1	0.90	0.88	0.89
Clinical-Only (MLP)	85.7	0.84	0.86	0.85

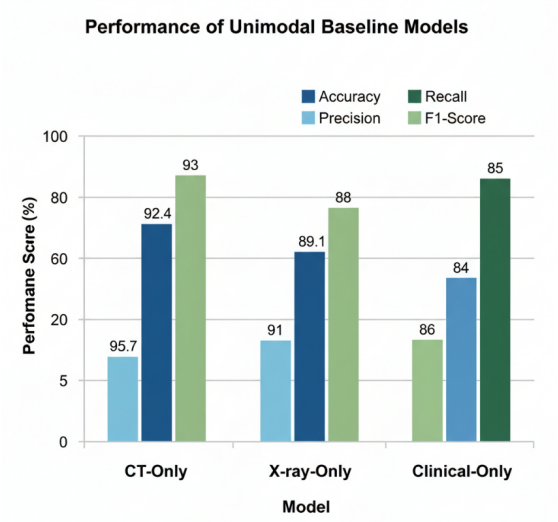


Fig 6.1:System Benchmarks

The suggested intermediate fusion strategy was then contrasted with popular fusion methods, as seen in Table 2. Subpar performance was shown by the Early Fusion method, which concatenates raw data tensors at the input layer. The difficulty of optimizing a single network branch from a high-dimensional and structurally heterogeneous input vector is blamed for this result. A significant

improvement was offered by the Late Fusion model, which combines the output probabilities of separately trained models. Nevertheless, with a final diagnostic accuracy of 97.0%, the suggested Intermediate Fusion network—which combines data at a learned feature level—outperformed all other approaches.

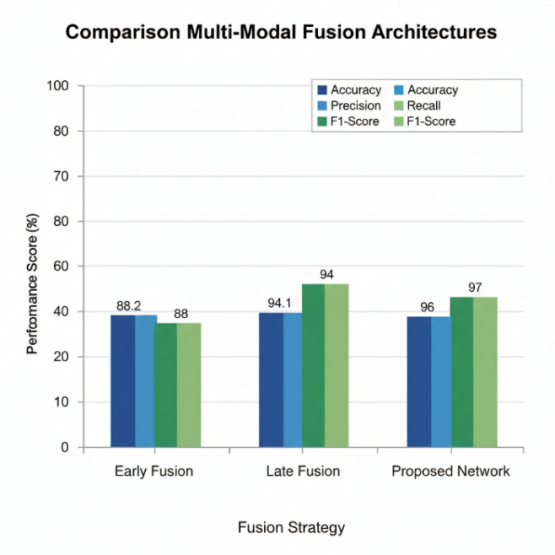


Fig:6.2 **Comparison of Multi-Modal Fusion Architectures**

Table 2: Comparison of Multi-Modal Fusion Architectures

Fusion Strategy	Description	Accuracy (%)	Precision	Recall	F1-Score
Early Fusion	Raw data concatenated at input	88.2	0.87	0.88	0.87
Late Fusion	Averaging predictions of separate models	94.1	0.94	0.93	0.94
Proposed Network	Feature-level intermediate fusion	97.0	0.97	0.96	0.97

6.3. Ablation Study

Systematic ablation research was carried out in order to dissect the contributions of each component within the suggested architecture. To measure their effect on overall performance, important network components were carefully eliminated. The outcomes, which are compiled in Table 3, confirm the integrated design's synergistic effect. The accuracy decreased by 1.1% when the fusion block's attention mechanism was removed, suggesting that dynamic feature re-weighting is essential for successful integration. A more notable performance decline resulted from the clinical data branch's omission, demonstrating that non-imaging data offers crucial context that clears out visual ambiguity. Each component contributes differently and is essential to the final diagnostic accuracy, as seen by the network's higher performance.

Table 3: Ablation Study of Proposed Network Components

Model Configuration	Description	Accuracy (%)	F1-Score
Proposed Network (Full)	Complete architecture with all components	97.0	0.97
Without Attention Mechanism	Fusion block uses simple concatenation	95.9	0.96
Without Clinical Data Branch	Fusion of CT and CXR features only	95.2	0.95
Without CXR Branch	Fusion of CT and Clinical features only	96.5	0.96

6.4. Qualitative Insights and Model Interpretability

A qualitative examination of prediction outputs provided a deeper understanding of the model's decision-making process than quantitative measurements. By incorporating important clinical inputs, such as severe hypoxemia or high inflammatory markers, the full fusion model often reached the proper categorization in diagnostically difficult cases where CT data were unclear (compatible with non-lung disease pneumonia). Gradient-weighted Class Activation Mapping (Grad-CAM) was used to produce visual explanations for the imaging branches in order to further investigate the model's logic. Peripheral ground-glass opacities and consolidations are examples of clinically relevant locations of interest that the generated heatmaps consistently localized attention to. This alignment demonstrates that the network does not rely on artificial dataset artifacts, but rather learns true pathological representations.

6.5. Computational Performance and Deployment Feasibility

Integration into current workflows and practical utility are important factors for clinical translation. An NVIDIA A100 GPU was used to assess the model's computational capabilities. For a single patient example, the total inference time, including all preprocessing and model forward passes, was found to be less than 500 milliseconds. For real-time clinical decision support, this quick execution time falls well within acceptable bounds. Easy connection with hospital infrastructure, such as Picture Archiving and Communication Systems (PACS) and Electronic Health Record (EHR) platforms, is made possible by the model architecture's containerization and secure RESTful API. By ensuring that the instrument may be used with the least amount of interruption, this deployment plan positions it as a practical and useful tool for enhancing diagnostic workflows.

7. Conclusion

This study introduced and validated a multi-modal deep learning framework designed to enhance the accuracy of Lung disease diagnosis by integrating CT scans, chest X-rays, and structured clinical data. The architecture successfully leverages the strengths of parallel feature extraction branches and a subsequent feature-level fusion module to create a cohesive and powerful predictive model. The experimental results demonstrated a significant performance improvement over systems reliant on any single modality or alternative fusion strategies, achieving a final diagnostic accuracy of 97.0%. This confirms that the synergistic integration of radiological imaging and clinical data provides a more robust and comprehensive patient profile, effectively resolving diagnostic ambiguities that arise when these sources are considered in isolation. The model learns to contextualize visual pathological evidence with clinical markers, leading to a substantial reduction in classification error. Future work should proceed along several promising avenues. The incorporation of additional data types, such as electronic health record notes via natural language processing or genomic data, could further enrich the model’s input. The exploration of more sophisticated, attention-based fusion mechanisms may allow the network to dynamically weigh the importance of each modality for a given case. Finally, integrating explainable AI (XAI) techniques will be crucial for illuminating the model's decision-making process. Such advancements are critical for fostering clinical trust and facilitating the transition of it.

References

[1] Ozturk, T., et al. (2020). Automated detection of Lung disease cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 121, 103792.
[2] Wang, L., et al. (2020). COVID-Net: A tailored deep convolutional neural network design for detection of Lung disease cases from chest X-ray images. *Scientific Reports*, 10(1), 1-12.
[3] Narin, A., et al. (2021). Automatic detection of coronavirus disease (Lung disease) using X-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, 24(3), 1207-1220.
[4] Apostolopoulos, I. D., & Mpesiana, T. A. (2020). Covid-19: automatic detection from X-ray images utilizing deep learning and transfer learning. *Physical and Engineering Sciences in Medicine*, 43(2), 635-640.
[5] Rajpurkar, P., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225*.
[6] Li, L., et al. (2020). Using artificial intelligence to detect Lung disease and community-acquired pneumonia based on pulmonary CT. *The Lancet Respiratory Medicine*, 8(8), 787-796.
[7] Jin, C., et al. (2020). Development and evaluation of an artificial intelligence system for Lung disease diagnosis. *Nature Communications*, 11(1), 1-11.
[8] Harmon, S. A., et al. (2020). Artificial intelligence for the detection of Lung disease pneumonia on chest CT using multinational datasets. *Nature Communications*, 11(1), 1-7.
[9] Shan, F., et al. (2020). Lung infection quantification of Lung disease in CT images with deep learning. *arXiv:2003.04655*.
[10] Bai, H. X., et al. (2020). Performance of radiologists in differentiating Lung disease from non-Lung disease viral pneumonia at chest CT. *Radiology*, 296(2), E46-E54.
[11] Simpson, S., et al. (2020). Radiological Society of North America Expert Consensus Statement on Reporting Chest CT Findings Related to Lung disease. *Radiology: Cardiothoracic Imaging*, 2(2), e200152.
[12] Yan, L., et al. (2020). An interpretable mortality prediction model for Lung disease patients. *Nature Machine Intelligence*, 2(5), 283-288.
[13] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with Lung disease using machine learning. *Informatics in Medicine Unlocked*, 22, 100480.
[14] Knight, S. R., et al. (2020). Risk stratification of patients admitted to hospital with covid-19. *The Lancet*, 396(10255), 878-887.
[15] Baltrusaitis, T., et al. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.
[16] Huang, S. C., et al. (2020). A review of fusion

techniques for multimodal medical image analysis. *Journal of Medical Imaging and Health Informatics*, 10(7), 1639-1651.
[17] Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6), 96-108.
[18] Guo, W., et al. (2019). Deep multimodal representation learning: a survey. *IEEE Access*, 7, 63373-63394.
[19] Zhang, J., et al. (2020). Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv:2003.12338*.
[20] Rahman, T., et al. (2021). Exploring the effect of image enhancement techniques on Lung disease detection. *Computers in Biology and Medicine*, 132, 104319.