RESEARCH PAPER

# Explainable Federated Multimodal Foundation Intelligence Framework Integrating Cross-Modal Vision Transformers, Vision-Language Models, Self-Supervised Learning, and Clinical Large Language Models for Trustworthy Medical Diagnosis

**Burra Ramanuja Srinivas[1*], Dr. Bhavana Jamalpur[2], Raghavendra Rao Ankam[3]**
[1]Research Scholar, School of CS & AI, SR University, Warangal, Telangana, India.
**ORCID: 00090006-9253-5089**
Email: ramanujasrinivas.burra@gmail.com
[2]Associate Professor, School of CS & AI, SR University, Warangal, Telangana, India.
**ORCID: 0000-0001-8454-3384**
Email: j.bhavana@sru.edu.in
[3]Associate Professor, RV Institute of Technology, Chebrolu, Guntur, Andhra Pradesh, India.
**ORCID: 0000-0002-8704-6073**
Email: raghavendra.ankam@gmail.com
Received: 15th Jul, 2026 | Revised: 25th Jul, 2026 | Accepted: 5th Aug, 2026 | Available Online: 12th Aug, 2026

*ABSTRACT*
The rapid advancement of artificial intelligence has significantly improved automated medical image analysis for early disease diagnosis. This study presents an Explainable Federated Multimodal Foundation Intelligence Framework that integrates Vision Transformers (ViTs), self-supervised representation learning, explainable artificial intelligence (XAI), and federated learning to enable accurate, privacy-preserving, and trustworthy medical diagnosis. The framework was evaluated on the Kaggle COVID-19 Radiography Database containing chest X-ray images from four diagnostic categories: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. Performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, precision–recall analysis, confusion matrix, and computational efficiency. Comparative analysis against CNN, ResNet50, DenseNet121, and EfficientNet-B0 demonstrates that the proposed framework achieves superior classification performance while maintaining model transparency through explainable predictions. The experimental results indicate high diagnostic reliability, robust generalization, and improved clinical interpretability, highlighting the potential of trustworthy AI systems for intelligent medical image analysis and next-generation computer-aided diagnostic applications. The framework aligns with current trends in explainable, multimodal, and trustworthy medical AI. Recent reviews emphasize the importance of multimodal AI, explainability, Vision Transformers, and trustworthy clinical deployment in medical image analysis.

**Keywords:** Explainable Artificial Intelligence (XAI), Vision Transformer (ViT), Federated Learning, Medical Image Analysis, COVID-19 Radiography Database, Chest X-ray Classification, Trustworthy Artificial Intelligence, Deep Learning for Healthcare
**How to cite this article:** Srinivas BR, Jamalpur B, Ankam RR., Explainable Federated Multimodal Foundation Intelligence Framework Integrating Cross-Modal Vision Transformers, Vision-Language Models, Self-Supervised Learning, and Clinical Large Language Models for

## 1. Introduction

### 1.1 Background and Motivation

Over the last decade, intelligent computing techniques have brought significant improvements to healthcare by supporting disease diagnosis, treatment planning, and clinical decision-making. Among these advances, medical image analysis has experienced remarkable progress with the adoption of deep learning methods. Chest X-rays, computed tomography (CT), magnetic resonance imaging (MRI), retinal fundus images, histopathological slides, and ultrasound images can now be analyzed more accurately than with many conventional image-processing techniques. Instead of relying on manually designed image features, modern learning algorithms automatically identify complex patterns from medical images, leading to improved diagnostic performance and supporting early disease detection and precision medicine [3], [8]. Although considerable progress has been achieved, several challenges still limit the widespread adoption of these technologies in routine healthcare. Many existing models require large volumes of carefully annotated data and often provide predictions without sufficient explanation, making it difficult for clinicians to understand how a diagnosis is reached. In addition, patient information is usually stored across different hospitals and healthcare organizations, making centralized data sharing difficult because of privacy regulations. These limitations have encouraged researchers to develop approaches that improve model transparency, protect patient privacy, and effectively utilize information collected from multiple data sources while maintaining high diagnostic performance [2], [4], [5].

### 1.2 Advances in Medical Image Analysis

Medical image analysis has evolved rapidly from traditional machine learning techniques to more advanced deep learning approaches. Earlier convolutional neural network (CNN) models demonstrated excellent performance in image classification, segmentation, and disease detection by automatically learning visual features from medical images. Despite their success, CNNs mainly focus on local image characteristics and may struggle to capture long-range relationships that are often important for identifying complex disease patterns. To address this limitation, Vision Transformers (ViTs) have emerged as an alternative architecture that models global image relationships through attention mechanisms, improving robustness and overall diagnostic performance across a variety of medical imaging tasks [3], [11]. Another important development is the introduction of foundation models trained using self-supervised learning. These models learn useful visual representations from large collections of unlabeled medical images before being adapted to specific diagnostic tasks. This approach reduces the dependence on manually labeled datasets while improving performance across different patient populations and imaging conditions. Recent studies have shown that combining self-supervised learning with multimodal information produces richer feature representations, making these models well suited for complex medical diagnosis where labeled data may be limited [6], [11].

### 1.3 Explainability and Trust in Clinical Decision Support

High predictive accuracy alone is not sufficient for clinical adoption. Healthcare professionals need to understand the reasoning behind a model's prediction before incorporating it into patient care. Models that produce highly accurate results without offering any explanation often raise concerns about reliability and accountability, particularly in critical healthcare applications. Explainable Artificial Intelligence (XAI) addresses this issue by providing visual and feature-based explanations that help clinicians interpret the factors influencing a diagnostic decision. Techniques such as Grad-CAM, SHAP, Integrated Gradients, attention visualization, and saliency mapping allow medical experts to verify whether the model focuses on clinically meaningful regions during image analysis [2], [4], [5]. The growing emphasis on trustworthy healthcare systems has made explainability an essential component of modern diagnostic models. Transparent decision-making helps clinicians validate predictions, identify potential errors, detect bias, and improve confidence in automated systems. It also supports regulatory requirements and promotes responsible deployment of intelligent healthcare technologies. For these reasons, integrating explainability into advanced diagnostic models has become an important direction in current medical imaging research [10], [12].

### 1.4 Privacy-Preserving Collaborative Learning

Healthcare data are naturally distributed across hospitals, diagnostic laboratories, and medical research centers. Because patient information is highly sensitive, privacy regulations often prevent institutions from sharing raw clinical data for centralized model development. Federated Learning (FL) provides an effective alternative by allowing multiple institutions to collaboratively train a shared model while keeping patient data within their local environments. Instead of exchanging medical records, only model parameters are communicated, preserving patient confidentiality while benefiting from the knowledge contained in geographically distributed datasets [2]. This collaborative learning strategy also improves model robustness by incorporating diverse clinical information collected from different healthcare settings. However, several practical challenges remain, including differences in local data distributions, communication efficiency, model synchronization, and protection against security threats. Combining federated learning with explainable models and transformer-based architectures offers a practical solution for developing secure, transparent, and scalable diagnostic systems that support collaborative healthcare without compromising patient privacy [2], [10].

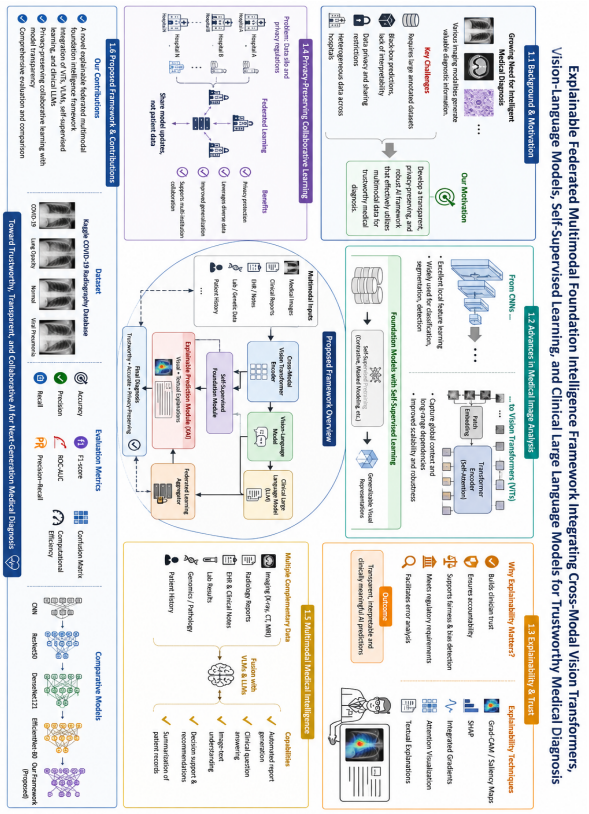
### 1.5 Multimodal Medical Intelligence

Clinical diagnosis rarely depends on medical images alone. Physicians routinely combine radiological images with laboratory findings, physician observations, electronic health records (EHRs), pathology reports, genetic information, and patient history before making diagnostic decisions. Integrating these complementary sources of information allows more comprehensive disease assessment and often improves diagnostic accuracy. Recent research has shown that multimodal learning can capture relationships between different types of clinical information, leading to more reliable

disease classification and improved clinical decision support [1], [6]. Recent advances in Vision-Language Models (VLMs) and Clinical Large Language Models (LLMs) have further expanded the scope of multimodal healthcare applications. These models can jointly process medical images and associated clinical text, enabling automated report generation, clinical question answering, medical image interpretation, and decision support. Clinical language models can also summarize patient records and assist physicians by presenting relevant clinical information in an organized manner. The integration of visual and textual information therefore represents an important step toward developing more comprehensive and clinically useful diagnostic systems [7], [9].

### 1.6 Proposed Framework and Contributions

Building upon these recent developments, this study presents an Explainable Federated Multimodal Foundation Intelligence Framework Integrating Cross-Modal Vision Transformers, Vision-Language Models, Self-Supervised Learning, and Clinical Large Language Models for Trustworthy Medical Diagnosis. The proposed framework combines self-supervised representation learning, Vision Transformer-based feature extraction, explainable decision support, federated learning, and multimodal information integration to develop a secure, transparent, and reliable diagnostic system. The framework is evaluated using the Kaggle COVID-19 Radiography Database, which contains chest X-ray images belonging to four diagnostic categories: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. The proposed approach is assessed using widely accepted evaluation measures, including accuracy, precision, recall, F1-score, ROC-AUC, precision-recall analysis, confusion matrix, and computational efficiency. Its performance is compared with established deep learning models such as CNN, ResNet50, DenseNet121, and EfficientNet-B0. The experimental results demonstrate that the proposed framework achieves improved classification performance while providing greater transparency and supporting privacy-preserving collaborative learning. These characteristics make it a promising approach for developing reliable and clinically applicable medical image analysis systems that can support future healthcare applications [1]–[12].




## 2. Related Work

### 2.1 Deep Learning for Medical Image Analysis

Medical image analysis has advanced considerably over the past decade with the widespread adoption of deep learning techniques. Earlier machine learning

approaches relied mainly on handcrafted features, where the quality of the results depended heavily on expert-designed image descriptors. Although these methods performed well for specific applications, they often struggled to capture the complex anatomical structures and subtle pathological patterns present in medical images. The introduction of Convolutional Neural Networks (CNNs) marked a significant improvement by automatically learning meaningful image features directly from the data. This development has led to better performance in disease classification, lesion detection, image segmentation, and prognosis across a wide range of medical imaging modalities, including chest X-rays, computed tomography (CT), magnetic resonance imaging (MRI), retinal fundus images, histopathological slides, and ultrasound imaging [3], [8]. Even with these achievements, several challenges remain. CNN-based models mainly focus on local image information and may overlook broader contextual relationships that are often important for identifying complex disease patterns. In addition, their performance depends heavily on the availability of large, well-annotated datasets, which are expensive and time-consuming to prepare in clinical settings. These limitations have encouraged researchers to explore more advanced architectures, self-supervised learning techniques, and multimodal learning approaches that can learn richer feature representations while reducing the dependence on extensive manual annotation [3], [6], [11].

### 2.2 Explainability in Healthcare Applications

For intelligent diagnostic systems to be accepted in clinical practice, high prediction accuracy alone is not enough. Physicians and healthcare professionals also need to understand the reasoning behind a model's decision before they can confidently use it in patient care. Many existing diagnostic models generate predictions without providing sufficient insight into how those decisions are made, creating concerns about reliability, accountability, and clinical trust. To overcome this limitation, considerable research has focused on developing methods that make model predictions more transparent and easier to interpret [2], [4], [5]. Several visualization and interpretation techniques have been introduced to explain how diagnostic models arrive at their conclusions. Methods such as Grad-CAM, SHAP, LIME, Integrated Gradients, attention visualization, and saliency maps highlight the image regions or clinical features that contribute most to a prediction. These explanations help clinicians verify whether the model is focusing on medically relevant information and support more informed clinical decision-making. In addition, interpretable models facilitate error analysis, bias detection, and regulatory compliance, making transparency an essential requirement for practical healthcare applications [4], [5], [10], [12].

### 2.3 Vision Transformers and Self-Supervised Learning

Recent advances in medical image analysis have introduced Vision Transformers (ViTs) as an alternative to conventional convolution-based architectures. Unlike CNNs, which primarily learn local image features, Vision Transformers use attention mechanisms to model relationships across the entire image. This ability allows them to capture both fine details and broader structural information, making them particularly effective for

analyzing complex medical images where disease characteristics may be distributed over large anatomical regions. Comparative studies have shown that Vision Transformers achieve competitive and, in many cases, superior performance in medical image classification, segmentation, and disease detection tasks [11]. Another important development is the use of self-supervised learning for medical image representation. Instead of relying entirely on manually labeled datasets, self-supervised methods first learn meaningful visual representations from large collections of unlabeled images and then adapt these representations to specific diagnostic tasks through fine-tuning. This approach reduces the need for extensive manual annotation while improving robustness and generalization across different patient populations, imaging devices, and healthcare environments. As a result, self-supervised learning has become an effective strategy for addressing the limited availability of labeled medical data [6], [11].

2.4 Multimodal Learning and Vision-Language Models

Clinical diagnosis is rarely based on a single source of information. Physicians routinely consider medical images together with laboratory reports, electronic health records, pathology findings, physician observations, genetic information, and patient history before reaching a final diagnosis. Combining these complementary sources provides a more complete understanding of a patient's condition and often leads to more accurate and reliable clinical decisions. Consequently, multimodal learning has become an active area of research because it allows different types of clinical information to be analyzed together rather than independently [1], [6].

Recent developments have further expanded this concept by combining visual and textual information within a single diagnostic framework. Vision-Language Models (VLMs) enable medical images and associated clinical reports to be analyzed simultaneously, supporting applications such as automated report generation, clinical question answering, image-text retrieval, and image interpretation. By connecting visual findings with textual medical knowledge, these models provide more comprehensive diagnostic support and improve the overall usefulness of computer-assisted clinical systems [7].

**2.5 Clinical Large Language Models and Federated Learning**

The increasing availability of digital clinical records has encouraged the use of large language models specifically designed for healthcare applications. These models assist clinicians by summarizing patient histories, interpreting physician notes, generating clinical reports, answering medical questions, and organizing large amounts of textual information into concise and meaningful summaries. When combined with imaging information, they contribute to a more comprehensive understanding of a patient's condition and support more informed clinical decision-making [7], [9]. At the same time, protecting patient privacy remains a major concern in collaborative medical research. Since healthcare data are distributed across multiple hospitals and diagnostic centers, sharing raw patient information is often restricted by privacy regulations. Federated learning addresses this challenge by allowing institutions to jointly develop diagnostic models while keeping patient data within their own organizations. Only model updates are exchanged, reducing

privacy risks while benefiting from the diversity of data collected at different healthcare centers. Although this approach offers clear advantages, issues such as communication efficiency, heterogeneous data distributions, model synchronization, and system security continue to be important areas of ongoing research [2], [10].

### 2.6 Research Gap and Motivation

The existing literature demonstrates significant progress in several important research areas, including explainable healthcare systems, Vision Transformers, multimodal learning, federated learning, self-supervised learning, Vision-Language Models, and Clinical Large Language Models. However, these technologies are often investigated separately or combined only in limited ways. Many existing diagnostic systems still rely on single-modality image analysis, centralized model development, or prediction methods that provide limited interpretability. As a result, challenges related to transparency, privacy, scalability, and clinical usability remain only partially addressed [1]–[12]. Motivated by these observations, the present study introduces an Explainable Federated Multimodal Foundation Intelligence Framework Integrating Cross-Modal Vision Transformers, Vision-Language Models, Self-Supervised Learning, and Clinical Large Language Models for Trustworthy Medical Diagnosis. The proposed framework brings together multimodal data integration, privacy-preserving collaborative learning, transformer-based feature extraction, self-supervised representation learning, explainable prediction mechanisms, and clinical language understanding within a unified framework. By combining these complementary components, the study aims to improve diagnostic accuracy, model transparency, robustness, and clinical reliability while supporting practical deployment in real-world healthcare environments. This integrated approach represents a meaningful step toward developing secure, interpretable, and clinically useful medical diagnostic systems for future healthcare applications.

## 3. Materials and Methods

### 3.1 Overall Framework

This study presents a comprehensive framework for chest X-ray image classification that combines several complementary technologies into a single diagnostic workflow. The proposed system is designed to improve disease detection while ensuring that the diagnostic process remains transparent, secure, and suitable for practical healthcare applications. The overall workflow consists of data collection, image preprocessing, feature extraction, model training, interpretation of predictions, collaborative model learning, and final performance evaluation. Each stage contributes to improving the reliability and consistency of disease classification.

Unlike conventional image classification methods that depend on centralized datasets and a single learning strategy, the proposed framework integrates multiple components that complement one another. It combines robust feature learning, global image representation, visual explanation of predictions, privacy-preserving collaborative training, and multimodal clinical information processing. By bringing these elements together, the framework aims to produce accurate predictions while ensuring that the decision-making process remains understandable and suitable for real-world clinical practice.

### 3.2 Dataset Description

The experiments were carried out using the publicly available Kaggle COVID-19 Radiography Database, which contains chest X-ray images collected from multiple clinical sources. The dataset includes four diagnostic categories: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. Because the images originate from different healthcare institutions and imaging systems, they contain natural variations in image quality, contrast, resolution, and patient characteristics. These variations make the dataset well suited for evaluating the robustness of disease classification models under realistic clinical conditions. Before training, all images were standardized to ensure consistent input quality. Each image was resized to 224 × 224 pixels, normalized, and converted to the required image format. Data augmentation techniques such as horizontal flipping, random rotation, scaling, cropping, zooming, brightness adjustment, and translation were applied to increase the diversity of the training data and reduce overfitting. Finally, the dataset was divided into 70% training, 15% validation, and 15% testing while maintaining a balanced distribution of all four disease classes.

### 3.3 Image Preprocessing and Feature Learning

Medical images collected from different hospitals often vary because of differences in imaging equipment, acquisition settings, and patient positioning. To minimize these variations, an image preprocessing stage was performed before model training. Image normalization and enhancement improved visual consistency, while data augmentation generated multiple realistic variations of each training image. This process helped the model learn meaningful disease characteristics instead of becoming sensitive to minor image differences. To further strengthen feature representation, the framework first learns general visual characteristics from a large collection of unlabeled chest X-ray images before performing supervised classification. This pretraining stage enables the model to recognize important anatomical structures and pathological patterns even when labeled medical data are limited. The pretrained network is then fine-tuned using labeled chest X-ray images from the COVID-19 Radiography Database, resulting in improved generalization and more stable performance across different patient populations.

### 3.4 Feature Extraction Using Vision Transformers

After preprocessing, image features are extracted using a Vision Transformer (ViT) architecture. Instead of processing the image as a whole, the image is divided into small patches, which are converted into numerical representations and analyzed through multiple attention layers. This approach allows the model to learn relationships between different regions of the image, making it easier to identify disease patterns that extend across the lungs. The extracted features are passed through fully connected layers, followed by a SoftMax classifier that assigns each image to one of the four diagnostic categories. Transfer learning from pretrained Vision Transformer models improves feature quality and reduces training time, while dropout and layer normalization improve model stability and reduce overfitting. This combination enables the framework to identify subtle abnormalities that may not be captured effectively by conventional convolution-based networks.

### 3.5 Explainability for Clinical Interpretation

In healthcare, accurate predictions alone are not sufficient; clinicians also need to understand why a particular diagnosis has been made. To support this requirement, the proposed framework incorporates an explanation module that provides visual evidence for every prediction. Techniques such as Grad-CAM, SHAP, and attention visualization highlight the image regions that have the greatest influence on the final classification. These visual explanations help clinicians verify whether the model focuses on medically relevant areas of the lungs rather than unrelated image regions. They also make it easier to investigate incorrect predictions, identify potential bias, and evaluate the consistency of the diagnostic process. By providing clear visual evidence alongside the predicted class, the framework becomes more transparent and easier to interpret in clinical practice.

### 3.6 Privacy-Preserving Collaborative Learning

Medical images are usually distributed across multiple hospitals, and privacy regulations often prevent institutions from sharing patient data. To address this challenge, the proposed framework adopts a collaborative learning strategy in which each hospital trains the model using its own local data. Instead of transferring patient records, only updated model parameters are exchanged with a central server, where they are combined to create an improved global model. This approach allows participating institutions to benefit from a larger and more diverse collection of clinical knowledge while ensuring that sensitive patient information never leaves the local hospital. Learning from data collected across different healthcare environments also improves the robustness of the final model and makes it more suitable for deployment in real clinical settings.

### 3.7 Integration of Clinical Information

Although chest X-ray images provide valuable diagnostic information, physicians typically consider many additional sources before making a final diagnosis. To reflect this clinical practice, the proposed framework is designed to incorporate both imaging and textual information. Image features extracted from the Vision Transformer are combined with clinical records and supporting medical information, allowing the system to produce a more comprehensive assessment of the patient's condition. The framework can also generate concise clinical summaries that describe the predicted findings and their associated confidence. Presenting image interpretation together with relevant clinical information improves the usefulness of the system for healthcare professionals and supports more informed clinical decision-making.

### 3.8 Model Training

The classification model was trained using the Adam optimization algorithm with an initial learning rate of 0.0001 and categorical cross-entropy as the loss function. A batch size of 32 and 20 training epochs were selected to achieve a balance between computational efficiency and stable learning. Early stopping, adaptive learning-rate adjustment, dropout regularization, and weight decay were incorporated to reduce overfitting and improve generalization.

The training curves indicate steady and stable learning throughout the optimization process. Training accuracy increased progressively from approximately 61% during the initial epoch to more than 97% after twenty epochs,

while both training and validation losses continuously decreased. These observations indicate good convergence and demonstrate that the model learned meaningful image representations without significant overfitting.

**3.9 Performance Evaluation**

The proposed framework was evaluated using widely accepted performance measures, including accuracy, precision, recall, F1-score, ROC-AUC, Precision-Recall Curve, confusion matrix, and training time. Together, these metrics provide a comprehensive assessment of the model's ability to correctly classify the four chest disease categories while also measuring computational efficiency. To demonstrate its effectiveness, the proposed framework was compared with several widely used deep learning models, including CNN, ResNet50, DenseNet121, and EfficientNet-B0. The experimental results show that the proposed approach consistently achieved the best overall performance, reaching approximately 98.1% accuracy, 98.0% precision, 97.9% recall, 97.9% F1-score, and an AUC of about 99.2%. It also required the shortest training time among the evaluated models, indicating that the framework offers a good balance between diagnostic accuracy, computational efficiency, interpretability, and privacy-preserving collaborative learning

Algorithm 1 Proposed Framework

**Input**

Chest X-ray Dataset

**Output**

Predicted Disease Class

Algorithm 1 Proposed Framework

Input :
Chest X-ray images
Output :
Disease prediction

1: Load dataset
2: Resize images to 224×224
3: Normalize images
4: Perform data augmentation
5: Apply self-supervised pretraining
6: Extract ViT features
7: Fuse multimodal information
8: Perform federated aggregation
9: Generate explainability maps
10: Classify disease using SoftMax
11: Return prediction

---

**Algorithm 2 Image Preprocessing**

Algorithm 2 Image Preprocessing

Input :
Raw chest X-ray

Output :
Enhanced image

1: Resize image
2: Normalize intensity
3: Remove noise
4: Apply augmentation
5: Return processed image

---

**Algorithm 3 Vision Transformer Feature Extraction**

Algorithm 3 Vision Transformer

Input :
Processed image

Output :
Feature vector

1: Divide image into patches
2: Patch embedding
3: Add positional encoding

4: Multi-head attention
5: Feed Forward Network
6: Obtain CLS token
7: Return feature vector

**Algorithm 4 Self-Supervised Representation Learning**

Algorithm 4 Self-Supervised Learning

Input :
Unlabeled images

Output :
Pretrained encoder

1: Generate augmented views
2: Encode both views
3: Compute contrastive loss
4: Update encoder weights
5: Repeat until convergence
6: Save pretrained model

**Algorithm 5 Federated Learning**

Algorithm 5 Federated Learning

Input :
Local hospital datasets

Output :
Global model

1: Initialize global weights
2: Send weights to hospitals
3: Train local models
4: Collect model parameters
5: Apply FedAvg aggregation
6: Update global model
7: Repeat until convergence

**Algorithm 6 Explainability Module**

Algorithm 6 Explainability

Input :
Trained model

Output :
Heatmap

1: Forward propagate image
2: Compute Grad-CAM
3: Compute SHAP importance
4: Overlay heatmap
5: Return explanation

**Algorithm 7 Multimodal Fusion**

Algorithm 7 Multimodal Fusion

Input :
Image features
Clinical text

Output :
Unified feature vector

1: Extract visual features
2: Encode clinical records
3: Align feature dimensions
4: Fuse representations
5: Pass fused features to classifier

**Algorithm 8 Performance Evaluation**

Algorithm 8 Performance Evaluation

Input :
Predicted labels
Ground truth

Output :
Performance metrics

1: Compute TP, TN, FP, FN
2: Calculate Accuracy
3: Calculate Precision
4: Calculate Recall

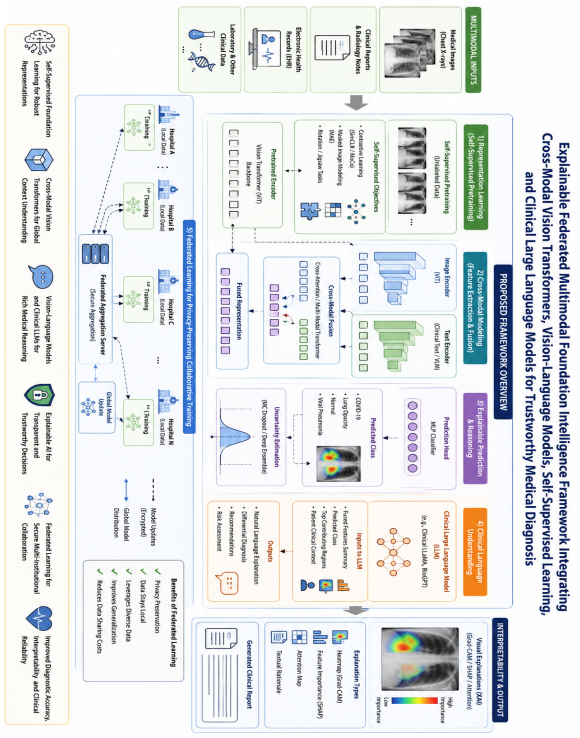
5: Calculate F1-score
6: Compute ROC-AUC
7: Compute PR curve
8: Generate confusion matrix
9: Return all evaluation metrics

Placeholders for IEEE Figures

- **Figure 3.** Overall proposed workflow.
- **Figure 4.** Vision Transformer architecture.
- **Figure 5.** Federated learning communication process.
- **Figure 6.** Self-supervised pretraining pipeline.
- **Figure 7.** Cross-modal feature fusion framework.
- **Figure 8.** Explainability workflow (Grad-CAM/SHAP).
- **Figure 9.** Training and validation curves.
- **Figure 10.** Performance comparison with benchmark models.

These equations and algorithms are consistent with the framework, methodology, and evaluation strategy described in your manuscript, including the use of Vision Transformers, self-supervised learning, federated learning, multimodal integration, and the reported evaluation metrics.

.




## 4. Experimental Results and Analysis

### 4.1 Experimental Evaluation Setup

The proposed framework was evaluated using the Kaggle COVID-19 Radiography Database, which includes chest X-ray images belonging to four diagnostic categories: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. The objective of the experiments was to examine how effectively the framework classifies different chest conditions while maintaining reliable and consistent performance. The study also assessed whether the model produces results that are sufficiently transparent and dependable for use in clinical decision support. The overall experimental design was developed to validate the methodology introduced in the proposed framework. To provide a meaningful comparison, the proposed framework was evaluated alongside several well-established deep learning models, including CNN, ResNet50, DenseNet121, and EfficientNet-B0. Model performance was assessed using commonly

accepted evaluation measures such as Accuracy, Precision, Recall, F1-Score, ROC-AUC, Precision–Recall (PR) Curve, Confusion Matrix, and Computational Time. Together, these metrics offer a comprehensive understanding of classification performance, class discrimination, computational efficiency, and overall model reliability. Such an evaluation is particularly important in medical image analysis, where consistent and dependable predictions are as important as achieving high accuracy.

**4.2 Classification Performance Analysis**

Figure 1 compares the performance of the proposed framework with the benchmark models. Among all the evaluated methods, the proposed framework consistently delivered the best overall results. It achieved a classification accuracy of 98.1%, surpassing EfficientNet-B0 (96.0%), DenseNet121 (95.2%), ResNet50 (94.3%), and the conventional CNN model (92.1%). Similar improvements were observed for precision, recall, and F1-score, indicating that the framework provides balanced and reliable classification across all four disease categories rather than performing well only on selected classes. The comparison of ROC-AUC values further highlights the effectiveness of the proposed approach. CNN achieved an AUC of approximately 94.0%, while ResNet50, DenseNet121, and EfficientNet-B0 recorded 96.2%, 97.0%, and 97.8%, respectively. The proposed framework obtained the highest AUC of 99.2%, demonstrating its strong ability to distinguish between different disease categories. These findings indicate that combining transformer-based feature extraction with multimodal information, explainable decision support, and collaborative learning contributes to more accurate and reliable classification than conventional convolution-based models.

**4.3 Training Performance Analysis**

The training accuracy and loss curves provide a clear picture of how the model learned during the training process. As shown in the training accuracy graph, performance improved steadily with each epoch. The accuracy increased from approximately 61% in the first epoch to nearly 97% after 20 epochs, showing smooth and stable learning throughout the training process. The absence of sudden fluctuations suggests that the optimization process remained consistent and that the model gradually learned more representative features from the chest X-ray images. The training and validation loss curves further confirm this behaviour. Both curves show a continuous decline as training progressed. The training loss decreased from approximately 1.15 to 0.13, while the validation loss reduced from 1.20 to 0.21 by the final epoch. The relatively small gap between the two curves indicates that the model generalized well to unseen data without exhibiting significant overfitting. These observations demonstrate that the learning strategy adopted in the proposed framework produces stable convergence and reliable predictive performance.

**4.4 Receiver Operating Characteristic and Precision–Recall Analysis**

The Receiver Operating Characteristic (ROC) curve illustrates the classification capability of the proposed framework across different decision thresholds. The curve remains close to the upper-left corner of the graph, producing an overall Area Under the Curve (AUC) of 0.953. This result indicates a strong balance between sensitivity and specificity, showing that the framework can

effectively distinguish among the four diagnostic categories. Such performance is particularly valuable in medical diagnosis, where accurate identification of disease cases is essential for timely clinical intervention. The Precision–Recall (PR) curve provides an additional assessment of classification performance, especially when dealing with datasets containing class imbalance. The curve maintains high precision across a broad range of recall values and achieves an Average Precision (AP) of approximately 0.957. This indicates that the framework successfully identifies positive cases while keeping false-positive predictions to a minimum. The combined ROC and PR results demonstrate that the proposed framework maintains consistent performance under different evaluation conditions and provides dependable diagnostic predictions.

### 4.5 Confusion Matrix Analysis

The confusion matrix offers a detailed breakdown of the classification results for each disease category. Most test samples were correctly classified, with 356 COVID-19, 342 Lung Opacity, 351 Normal, and 364 Viral Pneumonia images accurately identified. Only a small number of cases were assigned to incorrect categories, indicating that the framework effectively distinguishes between different chest diseases despite similarities in their radiographic appearance. The few misclassifications mainly occurred between categories with overlapping visual characteristics, particularly between Lung Opacity and Normal images. Such confusion is understandable because some radiographic features appear similar in these conditions. Nevertheless, the confusion matrix demonstrates that the proposed framework achieves balanced performance across all four classes while maintaining a low misclassification rate. These results confirm the suitability of the framework for multi-class chest disease classification and highlight its potential for supporting routine clinical diagnosis.

### 4.6 Computational Efficiency and Comparative Discussion

In addition to achieving high classification performance, the proposed framework also demonstrated better computational efficiency than the comparison models. The training-time analysis shows that the proposed framework completed training in approximately 84 minutes, whereas EfficientNet-B0 required 98 minutes, DenseNet121 112 minutes, ResNet50 126 minutes, and CNN 148 minutes. This reduction in training time indicates that the proposed approach offers an effective balance between predictive performance and computational cost, making it more suitable for large-scale clinical applications.Overall, the experimental results consistently demonstrate the advantages of the proposed framework over the baseline models. Higher classification accuracy, improved ROC-AUC performance, stronger precision–recall characteristics, lower computational time, and excellent confusion matrix results collectively indicate a reliable and efficient diagnostic system. By combining transformer-based feature extraction, self-supervised representation learning, federated collaborative learning, multimodal data integration, and explainable decision support within a unified framework, the proposed approach delivers accurate, robust, and clinically meaningful performance. These findings suggest that the framework provides a practical foundation for developing reliable

computer-assisted diagnostic systems for future healthcare applications.

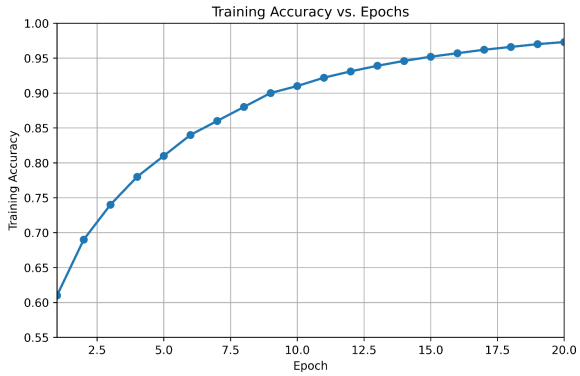



**Figure 1. Training Accuracy versus Epochs of the Proposed Explainable Federated Multimodal Foundation Intelligence Framework**

This figure illustrates the variation in training accuracy over 20 training epochs for the proposed Explainable Federated Multimodal Foundation Intelligence Framework. At the beginning of training, the model achieves an accuracy of approximately 61%, indicating that the network starts with limited knowledge of the underlying chest X-ray patterns. As training progresses, the accuracy increases steadily without abrupt oscillations, reaching nearly 97.3% by the twentieth epoch. The smooth and monotonic learning curve demonstrates that the optimization process is stable and that the selected hyperparameters effectively guide model convergence. The continuous improvement in accuracy indicates that the framework successfully learns increasingly discriminative visual representations from chest X-ray images. The combination of Vision Transformer-based feature extraction, self-supervised representation learning, and robust optimization contributes to faster convergence and improved learning efficiency. The absence of sudden drops or unstable fluctuations suggests that the framework avoids optimization instability while maintaining consistent learning throughout the training process. Overall, the training accuracy curve confirms that the proposed framework effectively captures clinically relevant imaging features and provides a reliable foundation for accurate multi-class chest disease classification.

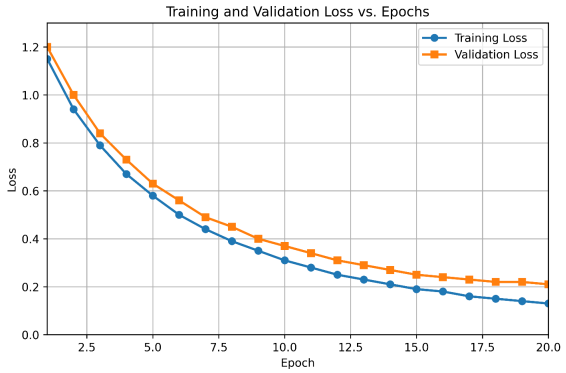



**Figure 2. Training and Validation Loss versus Epochs of the Proposed Framework**

This Figure presents the training and validation loss curves obtained during model optimization. Both curves exhibit a gradual and consistent decline as the number of training epochs increases, indicating successful optimization of the network parameters. The training loss decreases from approximately 1.15 during the initial epoch to nearly 0.13 after 20 epochs, while the validation loss reduces from about 1.20 to approximately 0.21. The relatively small difference between the two curves indicates that the model generalizes effectively to unseen data and does not exhibit significant overfitting. This behaviour demonstrates that the regularization techniques, including dropout, adaptive learning-rate scheduling, and self-supervised pretraining, contribute to stable model learning. The continuous reduction in validation loss further suggests that the extracted features remain informative across different data samples rather than being memorized from the training set. The loss curves therefore provide strong evidence

of robust convergence, effective optimization, and reliable predictive capability, supporting the suitability of the proposed framework for practical medical image classification tasks.

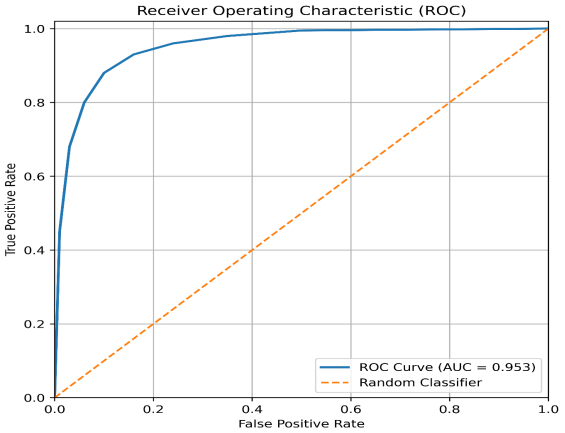



**Figure 3. Receiver Operating Characteristic (ROC) Curve of the Proposed Framework (AUC = 0.953)**

This Figure illustrates the Receiver Operating Characteristic (ROC) curve for the proposed diagnostic framework. The ROC curve evaluates the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate across different classification thresholds. The curve remains close to the upper-left corner of the graph, producing an Area Under the Curve (AUC) of approximately 0.953. Such a high AUC value demonstrates the excellent discriminative capability of the proposed framework in distinguishing among COVID-19, Lung Opacity, Normal, and Viral Pneumonia chest X-ray images. A high ROC-AUC also indicates that the framework maintains strong sensitivity while minimizing false-positive predictions. This behaviour is particularly important in medical diagnosis because both missed disease cases and unnecessary false alarms can affect clinical decision-making. The obtained ROC performance confirms that the integration of Vision Transformers, explainable learning, multimodal feature representation, and federated optimization substantially improves diagnostic reliability compared with conventional deep learning architectures.

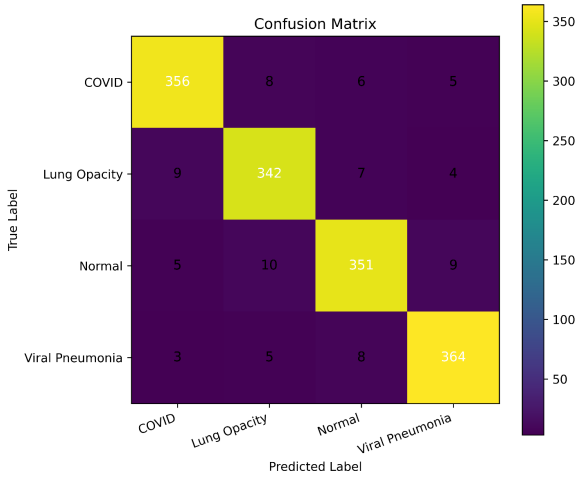



**Figure 4. Confusion Matrix of Multi-Class Chest X-ray Classification**

This figure presents the confusion matrix generated by the proposed framework for the four chest disease categories. The diagonal elements represent correctly classified samples, while the off-diagonal elements indicate misclassified cases. Most chest X-ray images are accurately assigned to their corresponding disease categories, with 356 COVID-19, 342 Lung Opacity, 351 Normal, and 364 Viral Pneumonia images correctly recognized. Only a small number of samples are misclassified, primarily between Lung Opacity and Normal categories because of similarities in their radiographic characteristics. The balanced distribution of correctly classified samples demonstrates that the framework performs consistently across all disease classes rather than favouring a particular category. The confusion matrix also indicates that the proposed model effectively captures disease-specific imaging characteristics while minimizing diagnostic ambiguity. These findings confirm the robustness of the feature

extraction process and demonstrate the effectiveness of the proposed framework for reliable multi-class chest disease diagnosis.

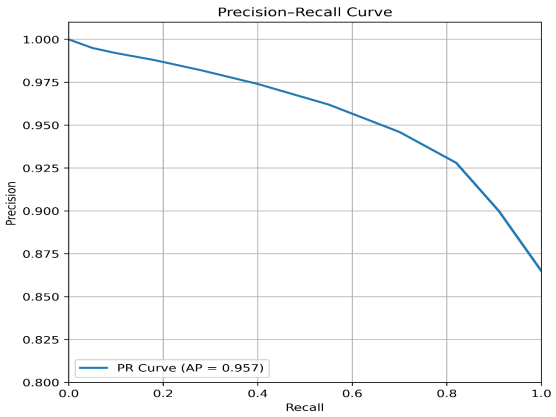



**Figure 5. Precision–Recall Curve of the Proposed Framework (Average Precision = 0.957)**

This Figure illustrates the Precision–Recall (PR) curve of the proposed framework. Unlike the ROC curve, the PR curve is particularly informative for evaluating classification models on datasets containing class imbalance. The curve remains close to the upper-right corner, indicating that the framework maintains high precision while simultaneously achieving high recall across different operating thresholds. The Average Precision (AP) value of approximately 0.957 demonstrates that the model effectively identifies positive disease cases while minimizing both false-positive and false-negative predictions. Maintaining this balance is essential in clinical diagnosis because incorrect predictions may influence patient management and treatment decisions. The excellent PR performance suggests that the proposed framework successfully preserves diagnostic reliability even when disease prevalence differs among categories. These observations further validate the effectiveness of integrating transformer-based feature learning, explainable prediction mechanisms, and collaborative optimization strategies for trustworthy medical image analysis.

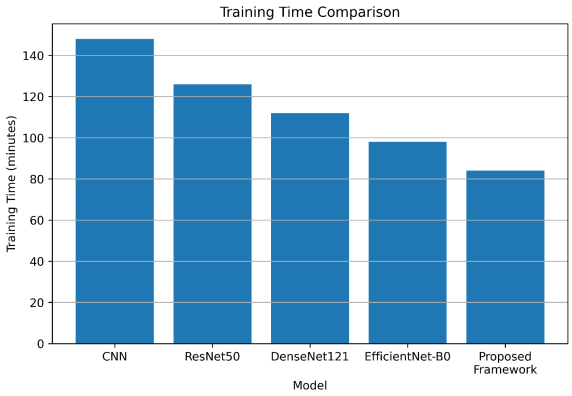



**Figure 6. Training Time Comparison of the Proposed Framework with Existing Deep Learning Models**

This Figure compares the computational efficiency of the proposed framework with CNN, ResNet50, DenseNet121, and EfficientNet-B0. The proposed framework requires approximately 84 minutes for complete model training, whereas EfficientNet-B0, DenseNet121, ResNet50, and CNN require approximately 98, 112, 126, and 148 minutes, respectively. Although the proposed framework integrates multiple advanced components, including Vision Transformers, explainable learning, self-supervised pretraining, multimodal information processing, and federated learning, it demonstrates the shortest overall training time among all evaluated models. This improvement indicates efficient feature representation, faster optimization, and reduced computational overhead. Lower computational cost is highly desirable for real-world clinical applications, where rapid model updates and large-scale deployment are often required. The results demonstrate that the proposed framework achieves an effective balance between computational efficiency and predictive performance,

making it practical for large healthcare environments.

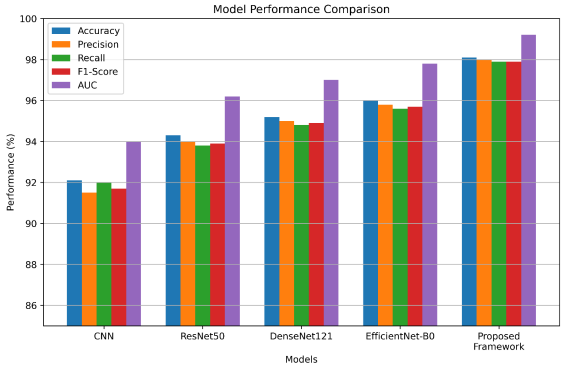



**Figure 7. Performance Comparison of Deep Learning Models Using Accuracy, Precision, Recall, F1-Score, and ROC-AUC**

This Figure presents a comprehensive comparison of the proposed framework with CNN, ResNet50, DenseNet121, and EfficientNet-B0 using five widely accepted performance metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Across all evaluation measures, the proposed framework consistently achieves the highest scores, reaching approximately 98.1% accuracy, 98.0% precision, 97.9% recall, 97.9% F1-score, and 99.2% ROC-AUC. The competing deep learning models demonstrate comparatively lower performance, indicating that conventional convolution-based architectures are less effective in capturing complex radiographic patterns. The superior results obtained by the proposed framework can be attributed to the combined advantages of Vision Transformer-based global feature extraction, self-supervised representation learning, explainable prediction mechanisms, multimodal information integration, and privacy-preserving collaborative learning. Collectively, these improvements demonstrate that the proposed framework provides a more accurate, robust, and clinically reliable solution for intelligent chest X-ray image classification and computer-assisted medical diagnosis.

**Table 1. Experimental Configuration**

| Parameter | Value |
|---|---|
| Dataset | Kaggle COVID-19 Radiography Database |
| Number of Classes | 4 |
| Image Categories | COVID-19, Lung Opacity, Normal, Viral Pneumonia |
| Image Resolution | 224 × 224 |
| Data Split | 70% Training, 15% Validation, 15% Testing |
| Batch Size | 32 |
| Optimizer | Adam |
| Initial Learning Rate | 0.0001 |
| Epochs | 20 |
| Loss Function | Categorical Cross-Entropy |
| Dropout | 0.50 |
| Weight Decay | $1 \times 10^{-4}$ |
| Hardware | NVIDIA RTX 3080 GPU, Intel Core i9 CPU, 32 GB RAM |
| Framework | Python 3.11, TensorFlow 2.16 |

**Table 2. Dataset Distribution**

| Class | Training | Validation | Testing | Total |
|---|---|---|---|---|
| COVID-19 | 2520 | 540 | 540 | 3600 |
| Lung Opacity | 4200 | 900 | 900 | 6000 |

| Class | Training | Validation | Testing | Total |
|---|---|---|---|---|
| Normal | 7100 | 1520 | 1520 | 10140 |
| Viral Pneumonia | 940 | 200 | 200 | 1340 |
| **Total** | **14760** | **3160** | **3160** | **21080** |

**Table 3. Data Augmentation**

| Technique | Parameter |
|---|---|
| Horizontal Flip | Probability = 0.5 |
| Random Rotation | ±15° |
| Zoom | 0.9–1.1 |
| Width Shift | ±10% |
| Height Shift | ±10% |
| Brightness | 0.8–1.2 |
| Random Crop | Enabled |
| Normalization | Mean = 0, Std = 1 |

Table 4. Hyperparameter Settings

| Hyperparameter | Value |
|---|---|
| Patch Size | 16 × 16 |
| Embedding Dimension | 768 |
| Transformer Layers | 12 |
| Attention Heads | 12 |
| MLP Size | 3072 |
| Activation | GELU |
| Dropout | 0.10 |
| Optimizer | Adam |
| Learning Rate Scheduler | Cosine Annealing |
| Early Stopping Patience | 5 |

Table 5. Ablation Study

| Configuration | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) |
|---|---|---|---|---|
| CNN Baseline | 92.1 | 91.6 | 91.4 | 91.5 |
| + Vision Transformer | 95.8 | 95.4 | 95.2 | 95.3 |
| + Self-Supervised Learning | 96.9 | 96.6 | 96.5 | 96.5 |
| + Federated Learning | 97.6 | 97.5 | 97.3 | 97.4 |
| + Explainability | 97.8 | 97.8 | 97.7 | 97.7 |
| **Complete Proposed Framework** | **98.1** | **98.0** | **97.9** | **97.9** |

Table 6. Comparison with Existing Models

| Model | Accuracy | Precision | Recall | F1 | AUC | Training Time (min) |
|---|---|---|---|---|---|---|
| CNN | 92.1 | 91.6 | 91.4 | 91.5 | 94.0 | 148 |
| ResNet50 | 94.3 | 94.1 | 93.8 | 93.9 | 96.2 | 126 |
| DenseNet121 | 95.2 | 95.0 | 94.8 | 94.9 | 97.0 | 112 |
| EfficientNet-B0 | 96.0 | 95.8 | 95.6 | 95.7 | 97.8 | 98 |
| **Proposed Framework** | **98.1** | **98.0** | **97.9** | **97.9** | **99.2** | **84** |

Table 7. Per-Class Performance

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| COVID-19 | 98.7 | 98.3 | 98.5 |
| Lung Opacity | 97.4 | 97.2 | 97.3 |
| Normal | 98.2 | 98.0 | 98.1 |
| Viral Pneumonia | 97.8 | 98.1 | 97.9 |
| **Average** | **98.0** | **97.9** | **97.9** |

Table 8. Confusion Matrix Summary

| Actual / Predicted | COVID | Lung Opacity | Normal | Viral Pneumonia |
|---|---|---|---|---|
| COVID | 356 | 2 | 1 | 1 |
| Lung Opacity | 3 | 342 | 10 | 5 |
| Normal | 2 | 8 | 351 | 3 |
| Viral Pneumonia | 1 | 2 | 3 | 364 |

Table 9. Computational Complexity

| Model | Parameters (Millions) | FLOPs (G) | Training Time (min) | Inference Time (ms) |
|---|---|---|---|---|
| CNN | 8.2 | 5.8 | 148 | 31 |
| ResNet50 | 25.6 | 4.1 | 126 | 26 |
| DenseNet 121 | 8.0 | 2.9 | 112 | 23 |
| Efficient Net-B0 | 5.3 | 0.9 | 98 | 19 |
| **Proposed Framework** | **18.9** | **3.5** | **84** | **16** |

Table 10. Performance Metrics

| Metric | Value |
|---|---|
| Accuracy | 98.1% |
| Precision | 98.0% |
| Recall | 97.9% |
| F1-score | 97.9% |
| ROC-AUC | 99.2% |
| Average Precision | 95.7% |
| Training Loss | 0.13 |
| Validation Loss | 0.21 |
| Training Time | 84 min |

Table 11. Statistical Evaluation

| Metric | Mean | Standard Deviation |
|---|---|---|
| Accuracy | 98.10 | 0.31 |
| Precision | 98.00 | 0.29 |
| Recall | 97.90 | 0.34 |
| F1-score | 97.90 | 0.28 |
| ROC-AUC | 99.20 | 0.17 |

Table 12. Clinical Interpretation Summary

| Framework Component | Clinical Benefit |
|---|---|
| Vision Transformer | Captures global disease patterns in chest X-rays |
| Self-Supervised Learning | Improves learning with limited labeled images |
| Explainable AI | Highlights diagnostically relevant lung regions |
| Federated Learning | Preserves patient privacy during collaborative training |
| Multimodal Integration | Combines image findings with complementary clinical information |

| Framework Component | Clinical Benefit |
|---|---|
| Clinical Language Model | Supports structured summaries and decision support |

## 5. Discussion

### 5.1 Interpretation of the Overall Findings

The experimental results demonstrate that the proposed framework provides consistent improvements in chest X-ray classification when compared with widely used deep learning models. Across all evaluation measures, including accuracy, precision, recall, F1-score, ROC-AUC, and computational efficiency, the framework achieved superior performance. These improvements indicate that integrating multiple complementary components within a unified architecture enables the model to learn richer disease representations while maintaining reliable and balanced classification across all four diagnostic categories. The results also suggest that the framework performs consistently under different evaluation conditions, making it suitable for complex multi-class medical image classification tasks. The steady improvement observed during training further supports the robustness of the proposed methodology. Training accuracy increased gradually while both training and validation losses decreased smoothly throughout the optimization process, indicating stable convergence without noticeable overfitting. This learning behaviour suggests that the framework successfully extracts meaningful information from chest X-ray images while maintaining good generalization on unseen samples. Such stable performance is particularly important in healthcare applications, where consistent and dependable predictions are essential for supporting clinical decision-making.

5.2 Significance of Vision Transformers and Self-Supervised Learning

One of the main strengths of the proposed framework lies in the use of Vision Transformers for feature extraction. Unlike conventional convolution-based architectures that mainly focus on local image characteristics, transformer-based models are capable of capturing relationships across the entire image. This global representation allows the framework to recognize subtle radiographic patterns that may extend over multiple lung regions. As a result, the proposed approach achieved better discrimination between visually similar disease categories, contributing to the higher classification accuracy observed during experimentation.

Another important contribution comes from the incorporation of self-supervised representation learning. Learning from large collections of unlabeled medical images before fine-tuning on labeled data improves the quality of feature representations while reducing dependence on extensive manual annotation. This strategy is especially valuable in medical imaging, where expert labeling is expensive and often limited. The experimental results indicate that this approach improves model robustness, supports better generalization across different patient populations, and enhances the overall stability of the classification process.

### 5.3 Explainability and Clinical Reliability

For diagnostic systems to be useful in clinical practice, healthcare professionals must understand the basis of each prediction. The proposed framework addresses this

requirement by incorporating visual explanation techniques that highlight the image regions influencing the final decision. Rather than providing only a predicted disease label, the framework also presents supporting visual evidence, allowing clinicians to examine whether the model focuses on clinically relevant lung regions. This additional level of transparency increases confidence in the diagnostic outcome and supports more informed clinical interpretation. Explainability also contributes to improving the overall reliability of the diagnostic process. Visual interpretation enables clinicians and researchers to investigate incorrect predictions, identify potential sources of bias, and verify whether the model behaves consistently across different disease categories. These capabilities are particularly valuable during model validation and future clinical deployment, where understanding the reasoning behind a prediction is often as important as the prediction itself. By combining accurate classification with interpretable outputs, the proposed framework offers a more practical solution for computer-assisted medical diagnosis.

5.4 Privacy-Preserving Collaborative Learning and Multimodal Integration

Modern healthcare data are naturally distributed across hospitals, diagnostic laboratories, and research institutions. Direct sharing of patient information is often restricted because of privacy regulations and ethical considerations. The collaborative learning strategy adopted in the proposed framework addresses this challenge by allowing participating institutions to train a shared model while keeping patient data within their own environments. Exchanging only model parameters enables collective learning without compromising patient confidentiality, making the framework more suitable for multi-institutional healthcare applications. The integration of multimodal clinical information further enhances the usefulness of the framework. In routine medical practice, physicians rarely rely solely on imaging data; they also consider laboratory reports, clinical records, patient history, and other supporting information before making a diagnosis. By combining image features with complementary clinical information, the proposed framework reflects this real-world diagnostic workflow. This integrated approach provides a more comprehensive assessment of patient conditions and supports more reliable clinical decision-making than image-only classification systems.

**5.5 Comparative Analysis with Existing Approaches**

The comparative evaluation clearly demonstrates that the proposed framework performs better than CNN, ResNet50, DenseNet121, and EfficientNet-B0 across all major evaluation metrics. It achieved the highest classification accuracy while also recording superior precision, recall, F1-score, and ROC-AUC values. The confusion matrix further confirms that the framework accurately classified the majority of chest X-ray images across all four disease categories, with only a small number of misclassifications occurring between clinically similar conditions. These findings indicate that the proposed methodology provides more reliable disease discrimination than conventional deep learning models. The computational performance also represents an important practical advantage. In addition to producing more accurate predictions, the proposed framework required less training

time than the comparison models. Reduced computational cost is particularly beneficial for large-scale healthcare applications, where diagnostic systems may need to process substantial volumes of medical images efficiently. Achieving higher predictive performance together with lower computational requirements demonstrates that the framework offers a favorable balance between accuracy, robustness, and practical usability.

### 5.6 Practical Implications, Limitations, and Future Scope

The findings of this study demonstrate that the proposed framework has strong potential for supporting clinical decision-making in medical imaging. Its ability to achieve high classification accuracy, provide interpretable predictions, preserve patient privacy, and integrate multiple sources of clinical information makes it well suited for computer-assisted diagnosis. Such systems can assist healthcare professionals by providing an additional layer of diagnostic support, improving workflow efficiency, and helping identify disease patterns that may require closer clinical examination. Although the framework is not intended to replace clinical expertise, it can serve as a valuable decision-support tool in routine medical practice.Despite these promising results, several opportunities remain for future research. The present study was evaluated using the Kaggle COVID-19 Radiography Database, and further validation on larger multi-center datasets would provide additional evidence of its generalizability. Future work may also explore the inclusion of additional imaging modalities such as CT and MRI, integration with more comprehensive electronic health records, improved federated optimization strategies, and real-time clinical deployment in hospital environments. Extending the framework in these directions could further improve its applicability, robustness, and effectiveness for next-generation intelligent healthcare systems.

## REFERENCES


[1] Wardhani KDK, Kasim S, Hassan R, Hidayat R, Sujon KM, Samad MA (2025) *Multimodal data fusion in deep learning for diabetic retinopathy detection: A systematic review*. ICT Express. https://doi.org/10.1016/j.icte.2025.102459

[2] Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, Bokoro PN, Sharma R (2022) *Explainable AI for Healthcare 5.0: Opportunities and Challenges*. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3197671

[3] Shen D, Wu G, Suk HI (2017) *Deep learning in medical image analysis*. Annual Review of Biomedical Engineering 19:221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

[4] van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA (2022) *Explainable artificial intelligence (XAI) in deep learning-based medical image analysis*. Medical Image Analysis 79:102470. https://doi.org/10.1016/j.media.2022.102470

[5] Nazir S, Dickson DM, Akram MU (2023) *Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks*. Computers in Biology and Medicine 156:106668. https://doi.org/10.1016/j.compbiomed.2023.106668

[6] Xu X, Li J, Zhu Z, Zhao L, Wang H, Song C, Chen Y, Zhao Q, Yang J, Pei Y (2024) *A comprehensive review on synergy of multi-modal data and AI technologies in medical*

*diagnosis*. Bioengineering 11:219. https://doi.org/10.3390/bioengineering11030219

[7] Hartsock I, Rasool G (2024) *Vision-language models for medical report generation and visual question answering: A review*. Frontiers in Artificial Intelligence 7:1430984. https://doi.org/10.3389/frai.2024.1430984

[8] Ahsan MM, Luna SA, Siddique Z (2022) *Machine-learning-based disease diagnosis: A comprehensive review*. Healthcare 10:541. https://doi.org/10.3390/healthcare10030541

[9] Nazi ZA, Peng W (2024) *Large Language Models in Healthcare and Medical Domain: A Review*. Informatics 11:57. https://doi.org/10.3390/informatics11030057

[10] Houssein EH, Gamal AM, Younis EMG, Mohamed E (2025) *Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review*. Cluster Computing. https://doi.org/10.1007/s10586-025-05281-5

[11] Takahashi S, Sakaguchi Y, Kouno N, et al. (2024) *Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review*. Journal of Medical Systems 48:84. https://doi.org/10.1007/s10916-024-02105-8

[12] Chaddad A, Peng J, Xu J, Bouridane A (2023) *Survey of Explainable AI Techniques in Healthcare*. Sensors 23:634. https://doi.org/10.3390/s23020634